Combinatorics, Probability and Computing (2010) 19, 561–578. © Cambridge University Press 2010
doi:10.1017/S0963548309990630

On the Subtree Size Profile of Binary Search trees

FLORIAN DENNERT and RUDOLF GRÜBEL

Institut für Mathematische Stochastik, Leibniz Universität Hannover,
Postfach 6009, D-30060 Hannover, Germany
(e-mail: {dennert,rgrubel}@stochastik.uni-hannover.de)

Received 11 June 2009; revised 10 November 2009; first published online 22 January 2010

For random trees T generated by the binary search tree algorithm from uniformly distributed input we consider the subtree size profile, which maps $k \in \mathbb{N}$ to the number of nodes in T that root a subtree of size k. Complementing earlier work by Devroye, by Feng, Mahmoud and Panholzer, and by Fuchs, we obtain results for the range of small k-values and the range of k-values proportional to the size n of T. In both cases emphasis is on the process view, *i.e.*, the joint distributions for several k-values. We also show that the dynamics of the tree sequence lead to a qualitative difference between the asymptotic behaviour of the lower and the upper end of the profile.

1. Introduction

By a binary tree T we mean a finite, prefix-stable subset of the set $N := \{0,1\}^\star$ of finite words with letters 0 and 1. The empty word represents the root node, and a non-empty finite sequence $u = (u_1, \dots, u_k) \in N$ identifies a node u of the tree with its route, starting at the root node and moving to the left if $u_i = 0$ and to the right if $u_i = 1$, $i = 1, \dots, k$ (see also Figure 1(a)). A labelled binary tree is a pair (T, ϕ), with T a binary tree and ϕ a function on T with values in some set; in our case we may take this to be the set of real numbers. The binary search tree (BST) algorithm transforms a sequence $(x_n)_{n\in\mathbb{N}}$ of pairwise distinct real numbers into a sequence $(T_n, \phi_n)_{n\in\mathbb{N}}$ of labelled binary trees, where T_n has n nodes. We start with the tree $T_1 = \{\emptyset\}$ that consists of the root node only, with $\phi_1(\emptyset) = x_1$. In order to obtain (T_{n+1}, ϕ_{n+1}) from (T_n, ϕ_n) we compare x_{n+1} to x_1, moving to the left if $x_{n+1} < x_1$ and to the right if $x_{n+1} > x_1$, repeating this with the next node and its label (content) until an empty node for x_{n+1} is found.

Our basic object in this paper is the sequence $(T_n)_{n\in\mathbb{N}}$ of random binary trees that results if we apply the BST algorithm to a sequence $(\xi_n)_{n\in\mathbb{N}}$ of independent random variables that are all uniformly distributed on the unit interval. As the trees depend on the order of the input values only, we may replace the uniform distribution by any other distribution

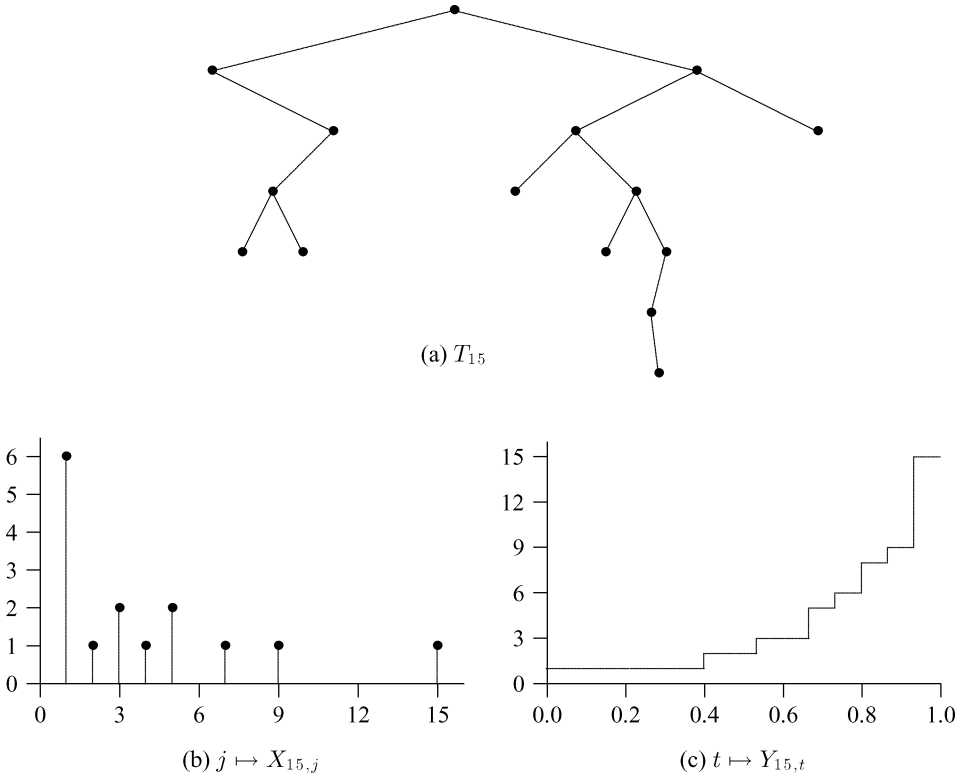

Figure 1. Binary tree (a) with associated subtree size profile (b) and cumulative big subtree counts (c)

that assigns the value 0 to individual numbers (has no atoms). We write BST(n) for the distribution of T_n; it is well known that this is *not* the uniform distribution on the set of binary trees with n nodes.

Binary trees and the BST algorithm are standard objects of discrete mathematics and theoretical computer science. Many authors have considered the above random input model; see, *e.g.*, [10], [12], and the references given there. For example, the node depth profile of T_n, which maps $k \in \mathbb{N}$ to the number of nodes $u \in T_n$ with depth k (where in the above representation the depth of a node $u = (u_1, \dots, u_k)$ is the word length k) has been investigated in [2], [5], [9], and elsewhere.

In the present paper we consider the subtree size profile, which maps $k \in \mathbb{N}$ to the number $X_{n,k}$ of nodes $u \in T_n$ that are the root of a subtree of T_n with size k. Here the subtree $T(u)$ of T associated with $u = (u_1, \dots, u_k) \in T$ consists of all $v = (v_1, \dots, v_l) \in \{0,1\}^\star$ with the property that $(u_1, \dots, u_k, v_1, \dots, v_l) \in T$. Figure 1 shows a tree with 15 nodes; two of the nodes, (0) and $(1,0,1)$, root a subtree of size 5, so that $X_{15,5} = 2$. Whereas the node depth profile is based on the number of ancestors of a node, the subtree size profile considers the number of its offspring. The first results we are aware of for the subtree size counts of random binary trees generated by the BST algorithm from

uniformly distributed random permutations are due to Devroye [4], who used a central limit theorem for m-dependent random variables to prove asymptotic normality for the standardized counts $(X_{n,k} - EX_{n,k})/\sqrt{\mathrm{var}(X_{n,k})}$ for fixed k as $n \to \infty$. Very recently Feng, Mahmoud and Panholzer [7] have obtained results for the case that $k = k_n$ varies with n, proving that asymptotic normality holds whenever $k_n/\sqrt{n} \to 0$ as $n \to \infty$. These authors also showed that the limit distribution of X_{n,k_n} is Poisson with mean $2/t^2$ if $k_n/\sqrt{n} \to t$ for some $t > 0$. Shortly thereafter, Fuchs [8] obtained a Berry–Esseen bound in connection with asymptotic normality, and a Poisson approximation result under the sole condition that $k_n \to \infty$. Whereas Devroye's approach is basically probabilistic, the later authors heavily rely on analytic machinery.

In the present paper we complement these earlier results by considering the stochastic processes $X_n = (X_{n,k})_{k\in\mathbb{N}}$ as $n \to \infty$, which implies that we obtain results on the dependencies of the subtree size counts for various k-values. Figure 1(b) shows a realization of the process X_{15}. We will deal with the 'lower' end, where k remains bounded, and the 'upper' end, where $k = k_n$ varies with n such that k_n/n tends to a positive value. For the lower end our main tool is the contraction method, which has become one of the standard techniques in this area. In connection with node depth profiles, this method has already been used in [9] and [5]. It turns out that a variant of the method introduced in [11] leads to convergence of the standardized processes $n^{-1/2}(X_{n,k} - EX_{n,k})_{k\in\mathbb{N}}$ to a discrete time Gaussian process $(X_{\infty,k})_{k\in\mathbb{N}}$ as $n \to \infty$, where convergence refers to weak convergence of the finite-dimensional distributions of the processes. We obtain an explicit description of the limiting second-order structure. We also show that there is a genuine reason for the fact that the result is on weak convergence only.

At the upper end, where $k = k_n$ varies with n such that $\lim_{n\to\infty} k_n/n = 1 - t$, $0 \leqslant t < 1$, it is known that the individual random variables X_{n,k_n} converge to 0 in probability. We show that a non-trivial limit arises if we pass to the partial sums

$$Y_{n,t} := \sum_{j \geqslant (1-t)n} X_{n,j}, \tag{1.1}$$

and we then obtain a limit process $Y_\infty = (Y_{\infty,t})_{0\leqslant t<1}$ for the processes $Y_n = (Y_{n,t})_{0\leqslant t<1}$ as $n \to \infty$. Figure 1(c) shows a path of Y_{15}; the two nodes with subtree size 5 correspond to a jump of size 2 at $t = 2/3$. Convergence refers to the usual Skorokhod metric on the space of cadlag functions on the interval $[0, t_0]$, for all $t_0 < 1$. It turns out that at this end of the subtree size profile we even have convergence almost surely. Previous work has concentrated on the distributions; taking into account the dynamics of the sequence we see that there is a major qualitative difference between the asymptotics of the two ends.

The results are given in the next section, and proofs are in Section 3. In a final section we collect some remarks on related results and problems. We write 1_A for the indicator function associated with a set A and $\mathcal{L}(X)$ for the distribution of a random quantity X, sometimes abbreviating $\mathcal{L}(X) = P$ to $X \sim P$. Also, '$=_{\mathrm{distr}}$' means equality in distribution and '$\to_{\mathrm{distr}}$' denotes convergence in distribution. Billingsley's classic [1] is our basic reference for convergence in distribution.

2. Results

Clearly, $X_{n,n} \equiv 1$ and $X_{n,k} \equiv 0$ if $k > n$. General formulas for the mean and variance of $X_{n,k}$ have already been obtained by [7, p. 178]

$$EX_{n,k} = \frac{2(n+1)}{(k+1)(k+2)} \quad \text{for } n \geqslant k+1, \tag{2.1}$$

$$\operatorname{var}(X_{n,k}) = \frac{2k(4k^2+5k-3)(n+1)}{(k+1)(k+2)^2(2k+1)(2k+3)} \quad \text{for } n \geqslant 2k+2. \tag{2.2}$$

As explained in the Introduction, our main interest here is in the 'process view', which means that we need a similar result for the covariances.

Proposition 2.1.

(a) *For* $1 \leqslant j < k$,

$$\operatorname{cov}(X_{j+k+2,j}, X_{j+k+2,k}) = -\frac{4j(j+2k+3)}{(j+k+1)(j+k+2)(k+1)(k+2)}.$$

(b) *For all* $n > j+k+2$,

$$\operatorname{cov}(X_{n,j}, X_{n,k}) = \frac{n+1}{j+k+3} \operatorname{cov}(X_{j+k+2,j}, X_{j+k+2,k}).$$

Our first theorem deals with the number of small subtrees.

Theorem 2.2. *Let* $(X_{\infty,k})_{k\in\mathbb{N}}$ *be a centred Gaussian process with covariance matrix* $\Sigma = (\Sigma(j,k))_{j,k\in\mathbb{N}}$ *given by*

$$\Sigma(j,k) = \begin{cases} \dfrac{2k(4k^2+5k-3)}{(k+1)(k+2)^2(2k+1)(2k+3)}, & \text{if } j = k, \\ -\dfrac{4j(j+2k+3)}{(j+k+1)(j+k+2)(j+k+3)(k+1)(k+2)}, & \text{if } j < k. \end{cases} \tag{2.3}$$

Then $(n^{-1/2}(X_{n,k} - EX_{n,k}))_{k\in\mathbb{N}}$ *converges to* $(X_{\infty,k})_{k\in\mathbb{N}}$ *as* $n \to \infty$ *in the sense that, for any fixed* $k \in \mathbb{N}$, *the* k*-dimensional random vectors*

$$n^{-1/2}(X_{n,1} - EX_{n,1}, \ldots, X_{n,k} - EX_{n,k}) \tag{2.4}$$

converge in distribution to the k*-dimensional random vector* $(X_{\infty,1}, \ldots, X_{\infty,k})$ *as* $n \to \infty$.

In [4], [7] and [8], limit results were obtained for the one-dimensional random variables $X_{n,k}$, where k may depend on n. Our result complements this, as the vector version also provides information about the dependencies of the random variables. We could view this as a partial answer to the question of what happens 'across k'. Our second main point in the present paper is the discussion of what happens 'across n': Taking the dynamics of the whole sequence $(X_n)_{n\in\mathbb{N}}$ into account, it is natural to ask for the 'best' mode of convergence. It turns out that we have almost sure convergence at the big subtrees end of

the profile, but that the weak convergence result in Theorem 2.2 cannot be strengthened. Rather than investigating this latter problem in some general abstract context of tail σ-fields, we give the following result on the number $X_{n,1}$ of 'leaf nodes', which may be of interest in its own right.

Theorem 2.3. *$P(X_{3m-1,1} = m$ for infinitely many $m \in \mathbb{N}) = 1$.*

Theorem 2.3 and formula (2.1) together imply that

$$\liminf_{n\to\infty} |X_{n,1} - EX_{n,1}| = 0 \quad \text{with probability 1,} \tag{2.5}$$

whereas Theorem 2.2 gives

$$n^{-1/2}(X_{n,1} - EX_{n,1}) \to_{\text{distr}} X_{\infty,1}, \tag{2.6}$$

where $X_{\infty,1}$ has a normal distribution with mean 0 and variance $2/45$. Clearly, (2.5) and (2.6) together imply that we do not have almost sure convergence for the random vectors in (2.4).

For the big subtrees we need to set up a suitable state space for the Y-processes first. Let D be the set of all cadlag (*i.e.*, right continuous with left limits) functions $f : [0,1) \to \mathbb{R}$. We say that $f_n \to f$ in D if the restrictions to every interval $[0,t_0]$, $t_0 < 1$, converge with respect to the Skorokhod topology; see [1, Chapter 3]. This defines a topology on D; we equip D with the associated Borel σ-field. Then $t \mapsto Y_{n,t}$ with $Y_{n,t}$ as defined in (1.1) is an element of this space, for every n, with $Y_{n,0} \equiv 1$ and $Y_{n,1-} \equiv n$.

Theorem 2.4. *In the space D, $Y_n = (Y_{n,t})_{0\leqslant t<1}$ converges almost surely to a limit process $Y_\infty = (Y_{\infty,t})_{0\leqslant t<1}$ as $n \to \infty$.*

In the course of the proof we will give a relatively explicit construction of Y_∞, based on the recursive structure of the family BST(n), $n \in \mathbb{N}$. This construction can also be used to obtain the mean and variance function of the limit process.

Theorem 2.5.

$$EY_{\infty,t} = \frac{1+t}{1-t} \quad \textit{for all } t \in [0,1),$$

$$\operatorname{var}(Y_{\infty,t}) = \begin{cases} 2 + \dfrac{2}{1-t} - \dfrac{4}{(1-t)^2} - \dfrac{8\log(1-t)}{1-t}, & \textit{for } t \in [0,1/2), \\ \dfrac{8\log 2 - 5}{1-t}, & \textit{for } t \in [1/2,1). \end{cases}$$

Note that $t \mapsto \operatorname{var}(Y_{\infty,t})$ is continuous but not differentiable at $t = 1/2$. Theorem 2.5 also shows that the limit in Theorem 2.4 is non-degenerate.

3. Proofs

3.1. Proof of Proposition 2.1

Suppose that $1 \leqslant j < k < \infty$. We begin with the case $n = j + k + 1$. Then nodes contributing to $X_{n,j}$ or $X_{n,k}$ must be elements of the left or right subtree of T_n. The size I_n of the left subtree is uniformly distributed on $\{0, \dots, n-1\}$ and, given $I_n = i$, the left and right subtrees are independent, with distributions BST(i) and BST($n-1-i$) respectively. In particular, we have the symmetry property

$$E[X_{n,j}X_{n,k}|I_n = i] = E[X_{n,j}X_{n,k}|I_n = n-1-i] \quad \text{for } i = 0, \dots, n-1. \tag{3.1}$$

We now consider different ranges for the size of the left subtree separately.

If $I_n = j$, then the left subtree contributes the fixed value 1 to $X_{n,j}$ and all other subtrees of size j must be contained in the right subtree; also, $X_{n,k} \equiv 1$, as the right subtree is the only subtree with k nodes. This gives

$$E[X_{n,j}X_{n,k}|I_n = j] = E[X_{n,j}X_{n,k}|I_n = k] = 1 + EX_{k,j}.$$

If $I_n \in \{j+1, \dots, k-1\}$, then $X_{n,k} = 0$, as neither subtree has enough nodes to accommodate a subtree of size k. Hence $E[X_{n,j}X_{n,k}|I_n = i] = 0$ for this range of i-values. If $I_n = i \in \{0, \dots, j-1\}$, then each subtree of size j must be a subtree of one possible subtree of size k of the right subtree with $n-1-i$ nodes, so that

$$E[X_{n,j}X_{n,k}|I_n = i] = E[X_{n,j}X_{n,k}|I_n = n-1-i] = EX_{n-1-i,k}\, EX_{k,j}.$$

Here we have used the simple fact that, for any two subtrees of a tree, either one of them is contained in the other, or they are disjoint; also, given that a node spawns a subtree of size k, the distribution of this subtree is BST(k). Using the known formula for $EX_{n,k}$, we obtain

$$\begin{aligned} EX_{n,j}X_{n,k} &= \frac{2}{n}\left(1 + \frac{2(k+1)}{(j+1)(j+2)}\right) + \frac{2}{n}\sum_{i=0}^{j-1} \frac{2(n-1-i+1)}{(k+1)(k+2)} \frac{2(k+1)}{(j+1)(j+2)} \\ &= \frac{2j^2k + 8j^2 + 14jk + 24j + 16k + 16 + 4k^2}{(j+k+1)(j+1)(j+2)(k+2)} \quad \text{for } n = j+k+1. \end{aligned}$$

We now turn to the case of interest, where $n = j + k + 2$. Again, we have $X_{n,k} = 0$ if $I_n \in \{j+2, \dots, k-1\}$, so that

$$\begin{aligned} EX_{n,j}X_{n,k} = {} & \frac{2}{n}E[X_{n,j}X_{n,k}|I_n = 0] + \frac{2}{n}\sum_{i=1}^{j-1} E[X_{n,j}X_{n,k}|I_n = i] \\ & + \frac{2}{n}E[X_{n,j}X_{n,k}|I_n = j] + \frac{1}{n}\big(2 - 1_{\{j+1=k\}}\big)E[X_{n,j}X_{n,k}|I_n = j+1]. \end{aligned} \tag{3.2}$$

Here we used (3.1); the indicator function ensures that we do not count $I_n = k$ twice if j and k differ by 1 only.

For the first term on the right-hand side of (3.2) we use

$$E[X_{n,j}X_{n,k}|I_n = 0] = EX_{j+k+1,j}X_{j+k+1,k},$$

which can now be evaluated with the formula for the case $n = j + k + 1$ from the first part of the proof. For $i \in \{1, \dots, j-1\}$ we get

$$E[X_{n,j}X_{n,k}|I_n = i] = EX_{n-1-i,k}\, EX_{k,j} = \frac{2(n-i)}{(k+1)(k+2)}\,\frac{2(k+1)}{(j+1)(j+2)}.$$

Further,

$$\begin{aligned} E[X_{n,j}X_{n,k}|I_n = j] &= (1 + EX_{k,j})EX_{k+1,k} \\ &= \left(1 + \frac{2(k+1)}{(j+1)(j+2)}\right)\frac{2(k+2)}{(k+1)(k+2)}, \end{aligned}$$

and, if $j + 1 < k$,

$$\begin{aligned} E[X_{n,j}X_{n,k}|I_n = j+1] &= EX_{j+1,j} + EX_{k,j} \\ &= \frac{2(j+2)}{(j+1)(j+2)} + \frac{2(k+1)}{(j+1)(j+2)}, \end{aligned}$$

whereas, for $j + 1 = k$,

$$E[X_{n,j}X_{n,k}|I_n = j+1] = 2\big(EX_{j+1,j} + EX_{k,j}\big)$$

in view of $X_{n,k} = 2$ on $I_n = j + 1$. Note that this difference between the two cases $j + 1 < k$ and $j + 1 = k$ cancels with the modification $2 - 1_{\{j+1=k\}}$ in (3.2). Put together, this leads to

$$EX_{n,j}X_{n,k} = \frac{4g(j,k)}{(j+k+2)(j+k+1)(j+1)(j+2)(k+1)(k+2)}$$

with

$$\begin{aligned} g(j,k) :=& 18 + 21j^2k + 54jk + 33j + 39k + 18j^2 + 29k^2 \\ &+ 6j^2k^2 + 27jk^2 + 9k^3 + 3j^3 + 2j^3k + 4jk^3 + k^4, \end{aligned}$$

so that finally

$$\begin{aligned} \operatorname{cov}(X_{j+k+2,j}, X_{j+k+2,k}) &= EX_{j+k+2,j}X_{j+k+2,k} - EX_{j+k+2,j}\, EX_{j+k+2,k} \\ &= -\frac{4j(j+2k+3)}{(j+k+2)(j+k+1)(k+1)(k+2)}. \end{aligned}$$

For the proof of (b) we may regard j and k as being fixed. It is evidently enough to show that, for all $n \geqslant j + k + 2$,

$$(n+1)a_{n+1} - (n+2)a_n = 0, \quad \text{with } a_n := \operatorname{cov}(X_{n,j}, X_{n,k}).$$

For any two random variables X and Y with finite second moment and a third random variable Z, we have the conditional covariance formula

$$\operatorname{cov}(X, Y) = E\big(\operatorname{cov}[X, Y|Z]\big) + \operatorname{cov}\big(E[X|Z], E[Y|Z]\big).$$

Together with the basic distributional split property of BST(n), this readily leads to the following recursion:

$$a_n = \frac{2}{n}\sum_{i=0}^{n-1} a_i + b_n, \quad \text{with } b_n := \operatorname{cov}\big(E[X_{n,j}|I_n], E[X_{n,k}|I_n]\big).$$

The recursion can easily be solved, resulting in

$$(n+1)a_{n+1} - (n+2)a_n = (n+1)b_{n+1} - nb_n,$$

so it remains to show that the right-hand side vanishes for $n \geqslant j+k+2$.

Again, nodes contributing to $X_{n,j}$ or $X_{n,k}$ must be elements of the right or left subtree, so that

$$E[X_{n,j}|I_n = i] = EX_{i,j} + EX_{n-1-i,j}, \tag{3.3}$$

and similarly with k instead of j. We now simply calculate nb_n for $n \geqslant j+k+2$. Using (3.3) we obtain

$$nb_n = \sum_{i=0}^{n-1} (EX_{i,j} + EX_{n-1-i,j})(EX_{i,k} + EX_{n-1-i,k}) - nEX_{n,j}EX_{n,k}. \tag{3.4}$$

With (2.1) and $EX_{i,j} = 0$ for $i < j$, $EX_{j,j} = 1$, and $n \geqslant j+k+2$, this leads to

$$\begin{aligned}
&(j+1)(j+2)(k+1)(k+2)nb_n \\
&\quad = 4(k+1)^2(k+2) + 8\sum_{i=k+1}^{n-1}(i+1)^2 + 4(n-j)(j+1)(j+2) \\
&\qquad + 4(n-k)(k+1)(k+2) + 8\sum_{i=j+1}^{n-2-k}(i+1)(n-i) - 4n(n+1)^2 \\
&\quad = -\frac{1}{3}j(8+4j+4j^2),
\end{aligned}$$

which indeed does not depend on n. (Some of the above computations were carried out with the help of the computer algebra system Maple.)

3.2. Proof of Theorem 2.2

We require some more notation. Throughout, we fix the dimension k and regard vectors as column vectors or $k \times 1$ matrices; A^{t} is the transpose of the matrix A, and we write $\delta_{i,j}$ for Kronecker's delta, which is 1 for $i = j$ and 0 otherwise. We use the Euclidean norm on $\mathbb{R}^k$ and the operator norm on the space $\mathbb{R}^{k\times k}$ of $k \times k$-matrices, writing $\|\cdot\|$ in both cases. Convergence of random vectors and random matrices refers to the respective L^3-norm. For example, $Y_n \to 0$ for a sequence $(Y_n)_{n\in\mathbb{N}}$ of k-dimensional random vectors means that $\lim_{n\to\infty} E\|Y_n\|^3 = 0$. Finally, $\mathrm{Id} = (\delta_{i,j})_{i,j=1}^k$ denotes the $k \times k$ unit matrix.

Suppose that $Y_n := (X_{n,1}, \ldots, X_{n,k})^{\mathrm{t}}$, $a_n := EY_n$ and $\Sigma_n := \mathrm{cov}(Y_n)$. We shall show that $\Sigma_n^{-1/2}(Y_n - a_n)$ converges in distribution to a d-dimensional standard normal random vector. As k was arbitrary, this together with the structure of the Σ_n given in Proposition 2.1 implies the statement of the theorem.

Splitting the tree into its left and right subtree as in Section 3.1, we obtain the following basic distributional recursion for the Y-vectors:

$$Y_n =_{\mathrm{distr}} Y_{I_n} + Y'_{n-1-I_n} + b_n. \tag{3.5}$$

Here Y_k' is an independent copy of Y_k for each $k \in \mathbb{N}$, I_n is independent of $(Y_n)_{n\in\mathbb{N}}$ and $(Y_n')_{n\in\mathbb{N}}$ and is uniformly distributed on $\{0,\dots,n-1\}$, and $b_n = (\delta_{n,1},\dots,\delta_{n,k})$. As this does not change the distributions, we may assume that $I_n = \lfloor nU \rfloor$ for all $n \in \mathbb{N}$, with U uniformly distributed on the unit interval.

For $n > k$ the 'toll terms' b_n in (3.5) disappear, and then, for the rescaled random vectors

$$Z_n := \Sigma_n^{-1/2}(Y_n - a_n), Z_n' := \Sigma_n^{-1/2}(Y_n' - a_n),\ n \in \mathbb{N},$$

the recursion (3.5) translates into

$$Z_n =_{\text{distr}} A_{n,I_n} Z_{I_n} + A_{n,n-1-I_n} Z'_{n-1-I_n} + v_{n,I_n}, \tag{3.6}$$

with

$$A_{n,i} := \Sigma_n^{-1/2}\Sigma_i^{1/2}, \quad v_{n,i} := \Sigma_n^{-1/2}(a_i + a_{n-1-i} - a_n) \quad \text{for } i = 0,\dots,n-1.$$

We need the asymptotic behaviour of the random vectors v_{n,I_n} and the random matrices A_{n,I_n}, $A_{n,n-1-I_n}$.

Lemma 3.1.

(a) *For all* $j \in \mathbb{N}$,

$$\lim_{n\to\infty} E1_{\{I_n \leqslant j\}} \|A_{n,I_n}\|^3 = 0, \quad \lim_{n\to\infty} E1_{\{n-1-I_n \leqslant j\}} \|A_{n,n-1-I_n}\|^3 = 0.$$

(b)

$$\lim_{n\to\infty} E\|v_{n,I_n}\|^3 = 0.$$

(c)

$$\lim_{n\to\infty} E\|A_{n,I_n} - \sqrt{U}\,\mathrm{Id}\|^3 = 0, \quad \lim_{n\to\infty} E\|A_{n,n-1-I_n} - \sqrt{1-U}\,\mathrm{Id}\|^3 = 0.$$

Proof. Throughout, we may assume that $n > 2k+2$. In particular, by Proposition 2.1(b),

$$\Sigma_n = (n+1)\Lambda, \tag{3.7}$$

with some fixed matrix Λ.

(a) By symmetry it is enough to prove the first part, and for this it is enough to use (3.7) and to note that

$$\sup_{i=0,\dots,j} \|\Sigma_i\| < \infty \quad \text{for all } j \in \mathbb{N}.$$

(b) From (2.1) we obtain that $a_i + a_{n-1-i} = a_n$ for $i = k+1,\dots,n-2-k$, so that

$$E\|v_{n,I_n}\|^3 1_{\{k<I_n<n-1-k\}} = 0. \tag{3.8}$$

For $i \leqslant k$ we have, for the jth component of $a_i + a_{n-1-i} - a_n$,

$$|(a_i + a_{n-1-i} - a_n)_j| \leqslant \left|\delta_{ij} - \frac{2(i+1)}{j+1}\right| \leqslant k+1,$$

so that

$$E\|a_{I_n} + a_{n-1-I_n} - a_n\|^3 1_{\{I_n \leqslant k\}} \leqslant (k+1)^{9/2} \frac{k+1}{n},$$

which together with (3.7) implies

$$\lim_{n\to\infty} E\|v_{n,I_n}\|^3 1_{\{I_n \leqslant k\}} = 0. \tag{3.9}$$

By symmetry,

$$\lim_{n\to\infty} E\|v_{n,I_n}\|^3 1_{\{I_n \geqslant n-k-1\}} = 0, \tag{3.10}$$

and part (b) of the lemma is now immediate from (3.8), (3.9) and (3.10).

(c) Again it is enough to prove the first part, and it is easy to see that we may neglect the range $I_n \in \{0,\dots,2k+2, n-2k-3,\dots,n-1\}$ asymptotically. Outside this range (3.7) gives $A_{n,I_n} = (\sqrt{I_n+1}/\sqrt{n+1})\mathrm{Id}$, so the assertion follows from the construction of the sequence $(I_n)_{n\in\mathbb{N}}$. □

Letting $n \to \infty$ in (3.6), we formally obtain the fixed-point equation

$$Z_\infty =_{\text{distr}} \sqrt{U}\, Z_\infty + \sqrt{1-U}\, Z'_\infty \tag{3.11}$$

for the prospective limit Z_∞. In view of

$$E\big(\|\sqrt{U}\mathrm{Id}\|^3 + \|\sqrt{1-U}\mathrm{Id}\|^3\big) = \frac{4}{5} < 1, \tag{3.12}$$

the right-hand side of (3.11) defines a contraction with respect to the Zolotarev ζ_3-metric on the space of k-dimensional distributions with mean 0 and variance Id, and it is easily seen that the k-dimensional standard normal distribution solves (3.11) and hence is the unique fixed point in this space.

This is made rigorous in [11]. The statements in the above lemma together with (3.12) validate the assumptions of Theorem 4.1 in [11], which provides the desired convergence to the fixed point, *i.e.*, asymptotic normality. We mention in passing that we only need a special case; with the notation used in [11] we have that $s = 3$, we only have two summands, and our I_n has a very special form.

3.3. Proof of Theorem 2.3

A tree T_n with n nodes has $n+1$ nodes u that are external in the sense that $T_{n+1} := T_n \cup \{u\}$ is a tree with $n+1$ nodes. Any node with maximal depth must be a leaf node, and any leaf node is ancestor to two external nodes. This leads to the (tight) bounds

$$X_{n,1} \geqslant 1, \quad 2X_{n,1} \leqslant n+1, \quad \text{for all } n \in \mathbb{N}. \tag{3.13}$$

In the BST sequence, T_{n+1} arises from T_n by choosing u uniformly at random from the external nodes of T_n. The new node may either increase $X_{n,1}$ by 1 or it may leave $X_{n,1}$ invariant; if $X_{n,1} = k$ these two possibilities have probabilities $(n+1-2k)/(n+1)$ and $2k/(n+1)$ respectively.

Let $Y = (Y_m)_{m\in\mathbb{N}}$ be defined by $Y_m := X_{3m-1,1} - m$ for all $m \in \mathbb{N}$. This is a non-homogeneous Markov chain with $Y_1 \equiv 0$, and some standard calculations show that

its transition probabilities are given by

$$\begin{aligned}
P(Y_{m+1} = k-1 | Y_m = k) &= \frac{8(m+k)^3}{3m(3m+1)(3m+2)},\\
P(Y_{m+1} = k+1 | Y_m = k) &= \frac{2(m+k)(m+1-2k)(m-2k)}{3m(3m+1)(3m+2)}\\
&\quad + \frac{2(m-2k)^2(m+k+1)}{3m(3m+1)(3m+2)}\\
&\quad + \frac{2(m-2k)(m-2k-1)(m+k+2)}{3m(3m+1)(3m+2)},\\
P(Y_{m+1} = k+2 | Y_m = k) &= \frac{(m-2k)(m-2k-1)(m-2k-2)}{3m(3m+1)(3m+2)},
\end{aligned}$$

and $P(Y_{m+1} \in \{k-1, k, k+1, k+2\} | Y_m = k) = 1$. Note that the restrictions (3.13) translate into

$$k \geqslant 1-m, \ 2k \leqslant m.$$

Now let $Z = (Z_m)_{m\in\mathbb{N}}$ be a random walk on $\mathbb{Z}$ with $P(Z_1 = 0) = 1$ that moves from k to $k-1$, $k+1$ and $k+2$ with probabilities $8/27$, $6/27$ and $1/27$ respectively, and again $P(Z_{m+1} \in \{k-1, k, k+1, k+2\} | Z_m = k) = 1$ for all $m \in \mathbb{N}$. It is straightforward to show that, for $k > 0$,

$$\begin{aligned}
P(Y_{m+1} = k-1 | Y_m = k) &\geqslant P(Z_{m+1} = k-1 | Z_m = k),\\
P(Y_{m+1} = k+1 | Y_m = k) &\leqslant P(Z_{m+1} = k+1 | Z_m = k),\\
P(Y_{m+1} = k+2 | Y_m = k) &\leqslant P(Z_{m+1} = k+2 | Z_m = k),
\end{aligned}$$

whereas for $k < 0$ all these inequalities hold with $\geqslant$ replaced by $\leqslant$ and *vice versa*. In words: on $\mathbb{N}$, the conditional increment of Y is stochastically bounded from above by the conditional increment of Z; on $-\mathbb{N}$, it is the other way round. We also have

$$P(Y_2 = -1) \leqslant P(Z_2 = -1), \quad P(Y_2 \geqslant 1) = P(Y_2 = 1) \leqslant P(Z_2 \geqslant 1).$$

Given a sequence $(\xi_n)_{n\in\mathbb{N}}$ of independent random variables, all uniformly distributed on the unit interval, we can construct Y and Z via $Y_1 = Z_1 \equiv 0$ and

$$Z_{m+1} = Z_m + f(\xi_m), \ Y_{m+1} = Y_m + g(m, Y_m, \xi_m) \quad \text{for all } m \in \mathbb{N},$$

where f and $g(m,k,\cdot)$ are the quantile functions associated with the distributions of the conditional increments. This construction yields a bivariate chain on $\mathbb{Z}\times\mathbb{Z}$ that has marginals Y and Z and is such that, for all $m \in \mathbb{N}$,

$$\begin{aligned}
0 < Y_m \leqslant Z_m &\implies 0 \leqslant Y_{m+1} \leqslant Z_{m+1},\\
Z_m \leqslant Y_m < 0 &\implies Z_{m+1} \leqslant Y_{m+1} \leqslant 0
\end{aligned}$$

and with the further property that $Y_2 Z_2 \geqslant 0$. This means that the return time distributions to 0 of Y are stochastically bounded from above by the distribution of the return time to 0 of Z. Note that for Y the return times to 0 are independent but not identically distributed: we do have a Markov chain, but it is not homogeneous in time. The random

walk $(Z_m)_{m\in\mathbb{N}}$ is null recurrent in view of the fact that its step distribution has mean 0, which means that its return time is finite with probability 1. Hence Y returns to 0 infinitely often.

3.4. Proof of Theorem 2.4

The subtree size profile can be regarded as an inverse to the standardized subtree size counts. We show that the counts converge, and that the Y-processes can be written as an almost surely continuous function of these counts. In order to make this precise, we define the empirical subtree size functional Ψ as a function that associates with a non-empty tree T the function $\Psi(T) : N \to [0,1]$ defined by

$$\Psi(T)(u) := \#T(u)/\#T, \quad u \in N.$$

Now let $(\eta_u)_{u\in N}$ be a family of independent random variables, all uniformly distributed on the unit interval. For each $u = (u_1,\dots,u_k) \in N$, let

$$\Phi_\infty(u) := \prod_{j=1}^{k-1} \eta_{(u_1,\dots,u_j)}^{(1-u_{j+1})} \cdot (1-\eta_{(u_1,\dots,u_j)})^{u_{j+1}}. \tag{3.14}$$

We interpret an empty product as 1, *i.e.*, $\Phi_\infty(\emptyset) = 1$. This defines a random element of $[0,1]^N$. The following proposition seems to belong to the folklore of the subject, but we have not been able to find it in the literature in the form required here. The result can be put into the wider context of boundary theory for transient Markov chains; see [6].

Proposition 3.2. *As $n \to \infty$, $\Psi(T_n)$ converges with probability 1 in the space $[0,1]^N$, endowed with the product topology. The distribution of the limit Ψ_∞ is the same as the distribution of Φ_∞.*

Proof. Let $u = (u_1,\dots,u_l) \in N$. We may assume that $u \neq \emptyset$. It is known that the fill level of the T_n converges to ∞ almost surely, so that

$$\tau := \min\{n \in \mathbb{N} : u \in T_n\} < \infty \quad \text{with probability 1.}$$

Let $(\xi_n)_{n\in\mathbb{N}}$ be the input sequence that generates the sequence of trees as explained in the Introduction. The order statistics $0 < \xi_{(\tau:1)} < \xi_{(\tau:2)} < \cdots < \xi_{(\tau:\tau)} < 1$ associated with $\xi_1,\dots,\xi_\tau$ form a partition of the unit interval. Let k be such that $\xi_\tau = \xi_{(\tau:k)}$ and put $\xi_{(\tau:0)} = 0$, $\xi_{(\tau:\tau+1)} = 1$. The sequence $(\xi_{\tau+n})_{n\in\mathbb{N}}$ is independent of the initial segment $(\xi_1,\dots,\xi_\tau)$ and again consists of independent random variables, all uniformly distributed on $[0,1]$. The subsequence of those that land in the interval $I := (\xi_{(\tau:k-1)}, \xi_{(\tau:k+1)})$ and thus contribute to the subtree rooted at u is again i.i.d., now uniformly distributed on I, conditionally on $(\xi_1,\dots,\xi_\tau)$, which means that in the limit the relative subtree sizes at $u0 := (u_1,\dots,u_l,0)$ and $u1 := (u_1,\dots,u_l,1)$ will be $(\xi_{(\tau:k)} - \xi_{(\tau:k-1)})/(\xi_{(\tau:k+1)} - \xi_{(\tau:k-1)})$ and $(\xi_{(\tau:k+1)} - \xi_{(\tau:k)})/(\xi_{(\tau:k+1)} - \xi_{(\tau:k-1)})$ respectively. Hence, if we have convergence of $\Psi(T_n)(u)$ then convergence also holds for $\Psi(T_n)(u0)$ and $\Psi(T_n)(u1)$. In view of $\Psi(T_n)(\emptyset) \equiv 1$ for all $n \in \mathbb{N}$, this proves almost sure convergence of the standardized subtree size functional in the product topology.

The distributional statement now follows immediately from the basic distributional recursion of the family BST(n), $n \in \mathbb{N}_0$. □

We need two properties of the limit function.

Lemma 3.3.

(a) *With probability* 1, *all values* $\Psi_\infty(u)$, $u \in N$, *are different*.
(b) *With probability* 1, $\#\{u \in N : \Psi_\infty(u) \geqslant t\}$ *is finite for all* $t > 0$.

Proof. In view of Proposition 3.2 we may consider Φ_∞, defined in (3.14), instead of Ψ_∞.

(a) For the proof of the first statement, let $u, v \in N$ with $u \neq v$ and let s be the last common ancestor of u and v. We first assume that $s \notin \{u, v\}$. If s is the direct ancestor to u and v, *i.e.*, $u = s0, v = s1$ or $u = s1, v = s0$, then

$$\Phi_\infty(u) = \eta\Phi_\infty(s), \quad \Phi_\infty(v) = (1-\eta)\Phi_\infty(s), \tag{3.15}$$

with η uniformly distributed on $(0,1)$ and independent of $\Phi_\infty(s)$. Clearly, this implies that $P(\Phi_\infty(u) = \Phi_\infty(v)) = 0$. If $u \notin \{s, s0, s1\}$ then a representation analogous to (3.15) would contain an additional factor $\tilde{\eta}$ for $\Phi_\infty(u)$, where $\tilde{\eta}$ has an absolutely continuous distribution and is independent of η and $\Phi(s)$. Again, this implies that $P(\Phi_\infty(u) = \Phi_\infty(v)) = 0$. By symmetry the same holds if $v \notin \{s, s0, s1\}$, and the remaining cases $s = u$ and $s = v$ can be handled similarly.

(b) We first note that the probability that $\Phi_\infty(u) \geqslant t$ for a specific node $u = (u_1, \ldots, u_k)$, $k \in \mathbb{N}$, can be written as the probability of the event $\eta_1 \cdots \eta_k \geqslant t$, with $\eta_1, \ldots, \eta_k$ independent and uniformly distributed on $[0,1]$. By a standard argument, using the fact that the variables $-\log \eta_i$, $i = 1, \ldots, k$, are exponentially distributed with mean 1,

$$\begin{aligned} P(\eta_1 \cdots \eta_k \geqslant t) &= P\left(s(\log\eta_1 + \cdots + \log\eta_k) \geqslant s\log t\right) \\ &\leqslant e^{-s\log t}\left(Ee^{s\log\eta_1}\right)^k \\ &\leqslant \frac{1}{t^s(1+s)^k} \quad \text{for all } s > 0. \end{aligned}$$

We have 2^k nodes of depth k. Hence, with $s = 3/2$,

$$\begin{aligned} E\#\{u \in N : \Phi_\infty(u) \geqslant t\} &= 1 + \sum_{k=1}^{\infty} E\#\{u \in N : |u| = k,\ \Phi_\infty(u) \geqslant t\} \\ &\leqslant 1 + \sum_{k=1}^{\infty} 2^k \frac{2^k}{t^{3/2}5^k} < \infty. \end{aligned}$$

This proves that $\#\{u \in N : \Phi_\infty(u) \geqslant t\}$ is finite with probability 1 for each individual $t > 0$. Using monotonicity in t, it is easy to construct a set of probability 1 that works for all $t > 0$ simultaneously. □

Suppose now that A is such that $P(A) = 1$ and such that $\Psi_\infty(\omega)$ has the properties described in Lemma 3.3 whenever $\omega \in A$. Because of Proposition 3.2, we may further

assume that on A we also have $\Psi(T_n)(u)(\omega) \to \Psi_\infty(u)(\omega)$ for all $u \in N$ as $n \to \infty$. We now claim that, for $\omega \in A$, $Y_n(\omega)$ converges in D to the limit $Y_\infty(\omega) = (Y_{\infty,t}(\omega))_{0 \leqslant t < 1}$ given by

$$Y_{\infty,t}(\omega) := \#\{u \in N : \Psi_\infty(u)(\omega) \geqslant 1 - t\}.$$

With $Y_\infty \equiv 0$ on A^c this would show that $Y_n \to Y_\infty$ with probability 1.

Let $\omega \in A$ be fixed; below, we omit the argument ω. Let $t_0 < 1$ be given and choose $\epsilon > 0$ such that $1 - 2\epsilon > t_0$. Then the number of nodes u with $\Psi_\infty(u) \geqslant \epsilon$ is finite, and these nodes $u_1, \ldots, u_k$ may be ordered such that $\Psi_\infty(u_{i+1}) < \Psi_\infty(u_i)$, $i = 1, \ldots, k-1$. Further, for each of these nodes, $\Psi(T_n)(u_i) \to \Psi_\infty(u_i)$.

Now consider the functions

$$f_n : [0, t_0] \to \mathbb{N}, \quad t \mapsto \#\{u \in N : \Psi(T_n)(u) \geqslant 1 - t\},$$

$n \in \mathbb{N}$, and

$$f_\infty : [0, t_0] \to \mathbb{N}, \quad t \mapsto \#\{u \in N : \Psi_\infty(u) \geqslant 1 - t\}.$$

Clearly, these are the restrictions of the Y_n- and Y_∞-path, respectively, to the interval $[0, t_0]$. For any given $\delta > 0$ we can find an $n_0 \in \mathbb{N}$ such that $|\Psi(T_n)(u_i) - \Psi_\infty(u_i)| \leqslant \delta$ for all $n \geqslant n_0$ and all $i \in \{1, \ldots, k\}$. We may further assume, by increasing n_0 if necessary, that the number of nodes in T_n, $n \geqslant n_0$, that have subtree size at least $1 - t_0$, does not exceed k. All these functions are then increasing, take their values in $\{1, \ldots, k\}$, have jumps of size 1 only (if δ is small enough) and the position of the ith jump of f_n converges to the position of the ith jump of f. Taken together, this implies that $f_n \to f$ as $n \to \infty$ with respect to the Skorokhod topology on the space of cadlag functions on $[0, t_0]$.

3.5. Proof of Theorem 2.5

Using the representation (3.14), we define $\Phi^L_\infty, \Phi^R_\infty : N \to [0, 1]$ by

$$\Phi^L_\infty(u) := \Phi_\infty(0u), \quad \Phi^R_\infty(u) := \Phi_\infty(1u),$$

where $0u = (0, u_1, \ldots, u_k)$, $1u = (1, u_1, \ldots, u_k)$ for all $u = (u_1, \ldots, u_k) \in N$. Let

$$Y^L_\infty(t) := \#\{u \in N : \Phi^L_\infty(u) \geqslant 1 - t\}, \quad Y^R_\infty(t) := \#\{u \in N : \Phi^R_\infty(u) \geqslant 1 - t\}.$$

Then we obtain from (3.14), with $\eta := \eta_\emptyset$,

$$Y_{\infty,t} =_{\text{distr}} 1 + 1_{[1-t,1)}(\eta)\, Y^L_\infty\left(\frac{\eta - 1 + t}{\eta}\right) + 1_{(0,t]}(\eta)\, Y^R_\infty\left(\frac{t - \eta}{1 - \eta}\right). \tag{3.16}$$

Clearly, $EY^L_\infty(t) = EY^R_\infty(t) = EY_{\infty,t}$, hence (3.16) implies that $f(t) := EY_{\infty,t}$ satisfies the integral equation

$$f(t) = 1 + 2\int_0^t f\left(\frac{t - s}{1 - s}\right) ds.$$

This is uniquely solved by $f(t) = (1 + t)/(1 - t)$, $0 \leqslant t < 1$. (We can use (2.1) to guess the solution, but note that almost sure convergence in D does not imply convergence of the first moments.)

In the above derivation of the mean function we have implicitly used that $EY_{n,t} < \infty$ for $0 \leqslant t < 1$, which follows from the argument given at the end of the proof of Lemma 3.3.

This argument can easily be extended to prove the existence of higher moments; in particular, $EY_{\infty,t}^2 < \infty$ for $0 \leqslant t < 1$.

To obtain the variance function $g(t) := \operatorname{var}(Y_{\infty,t})$ we once again make use of the conditional variance formula,

$$g(t) = \operatorname{var}\big(E[Y_{\infty,t}|\eta]\big) + E\big(\operatorname{var}[Y_{\infty,t}|\eta]\big),$$

with η as in (3.16). From (3.16) we obtain, with f again the mean function,

$$E[Y_{\infty,t}|\eta] = 1 + 1_{[1-t,1)}(\eta)\, f\left(\frac{\eta - 1 + t}{\eta}\right) + 1_{(0,t]}(\eta)\, f\left(\frac{t-\eta}{1-\eta}\right).$$

Using our formula for f we are thus led to

$$\operatorname{var}\big(E[Y_{\infty,t}|\eta]\big) = E\big(E[Y_{\infty,t}|\eta]\big)^2 - (EY_{\infty,t})^2 = h(t),$$

with

$$h(t) := \begin{cases} \dfrac{2t(t^2-6t+3)}{3(1-t)^2}, & 0 \leqslant t \leqslant 1/2, \\ \dfrac{2(1-t)}{3}, & 1/2 < t < 1. \end{cases}$$

Further, as Y_∞^L and Y_∞^R are independent given η,

$$\operatorname{var}[Y_{\infty,t}|\eta] = 1_{[1-t,1)}(\eta)\, g\left(\frac{\eta - 1 + t}{\eta}\right) + 1_{(0,t]}(\eta)\, g\left(\frac{t-\eta}{1-\eta}\right),$$

which leads to

$$E\big(\operatorname{var}[Y_{\infty,t}|\eta]\big) = 2\int_0^t g\left(\frac{t-s}{1-s}\right) ds = 2(1-t)\int_0^t \frac{1}{(1-s)^2}\, g(s)\, ds.$$

Putting this together, we obtain an integral equation for the variance function,

$$g(t) = 2(1-t)\int_0^t \frac{g(s)}{(1-s)^2}\, ds + h(t). \tag{3.17}$$

In particular, $g(0) = 0$ (which is also obvious from $Y_{\infty,0} \equiv 1$), g is continuous on $[0,1)$, and g is differentiable on $(0,1/2) \cup (1/2,1)$. Standard techniques, such as taking the derivative on both sides and solving the resulting differential equations inside the subintervals, can be used to show that (3.17) is uniquely solved by the function given in the theorem.

4. Comments

4.1. Contractions at the big end

We have used a variant of the contraction method to obtain asymptotic normality for the number of small subtrees. By design, a method that takes distributions as its basic objects will lead to weak convergence only, where in fact, by Theorem 2.4, the 'true' mode of convergence for the cumulative counts of large subtrees is convergence with probability 1. Nevertheless, it is interesting to see the contraction method at work at this end too. We refer the reader to the first author's doctoral thesis [3] for details, and simply give an overview.

Let D now be the set of all weakly increasing and right continuous functions $f : [0,1) \to \mathbb{N}$ with the property that $f(0) = 1$ and

$$\|f\| := \int_0^1 (1-t)\,|f(t)|\,dt < \infty.$$

This is a closed subset of the L^1-space associated with the measure $v(dt) = (1-t)\,dt$ on the unit interval. Let $\mathcal{B}(D)$ be the associated Borel σ-field. Then Y_n converges in distribution in the space D as $n \to \infty$, and the limit distribution is the unique fixed point of a suitably defined functional $\Phi : \mathbb{M} \to \mathbb{M}$, with $\mathbb{M}$ the set of all probability measures P on $(D, \mathcal{B}(D))$ that satisfy the condition $\int \|f\|\,P(df) < \infty$. We define a family $\{\phi(s,.) : 0 < s < 1\}$ of functions $\phi(s,.) : D \to D$ by

$$\phi(s,f)(t) := 1 + 1_{[1-t,1)}(s) \cdot f\left(\frac{s-1+t}{s}\right) + 1_{(0,t]}(s) \cdot f\left(\frac{t-s}{1-s}\right),$$

and then let $\Phi(P)$ be the distribution of $\phi(\eta, X)$, with η and X independent, η uniformly distributed on the unit interval, and P the distribution of the D-valued random variable X; see also (3.16). In fact, Φ turns out to be a strong contraction with respect to the metric

$$d(P,Q) := \inf\{E\|X - Y\| : X \sim P,\ Y \sim Y\}$$

on $\mathbb{M}$, and we have the following upper bound for the distance between the distribution of Y_n and the distribution of the limit:

$$d\big(\mathcal{L}(Y_n), \mathcal{L}(Y_\infty)\big) \leqslant 6(1 + \log n)/n \quad \text{for all } n \in \mathbb{N}.$$

4.2. The middle range

Given that we have found functional limits at the big and the small end of the subtree size functional, it is natural to ask what happens 'in the middle'. We know from the results of [7] and [8] that, for any individual $t \in (0,\infty)$, the counts X_{n,k_n} converge in distribution to a limit that is Poisson with mean $2/t^2$ if $k_n \sim tn^{1/2}$. It is easy to see that we cannot possibly have convergence almost surely in this situation, as this would mean that with probability 1 the random variable X_{n,k_n} does not change its value from some n onwards.

Regarding the joint distribution for more than one t-value, we conjecture that the associated counts are independent in the limit (Proposition 2.1 shows that the covariances tend to 0), but we do not have a proof. Use of the contraction method seems to require the construction of an appropriate accompanying sequence. Neininger and Rüschendorf [11] were able to carry this out in situations where asymptotic normality holds, as in our Theorem 2.2. For this it was important that for normal distributions there are two parameters that can be adjusted; also, the ideal metric ζ_3 used in connection with asymptotic normality does not seem to have an obvious analogue for distributions concentrated on $\mathbb{N}_0$. We plan to deal with this problem in a separate paper.

4.3. Use of subtree sizes

Passing from a binary tree to one of its characteristics entails some loss of information, but the intention is of course that the characteristic distils the features of the tree that

are of relevance to the application of interest. As with the node depth profile, the subtree size profile captures to some extent the balancedness of tree. For example, the sequence $(1, 1, \ldots, 1)$ would not completely specify the tree, but it would show that each node has exactly one direct child; the tree is essentially a linked list, and only the left–right structure of the tree's only path is lost when passing from the tree T_n to its subtree size profile X_n. Other characteristics of T_n can be read off from X_n: for example, the internal path length P_n, which is the sum of the heights of all nodes in T_n, and the Wiener index W_n, which is the sum of all distances between unordered pairs of nodes in T_n, can be written as

$$P_n = \sum_{j=1}^{n-1} jX_{n,j}, \quad W_n = \sum_{j=1}^{n-1} j(n-j)X_{n,j}.$$

Another use of subtree sizes appears in connection with the reconstruction of a sample $\xi_1, \ldots, \xi_n$ from the associated labelled binary tree (T_n, ϕ_n) produced by the BST algorithm. Where within the range of knowing the full sample and knowing the ordered sample lies (T_n, ϕ_n)? In the step from the order statistics to the original sample all $n!$ permutations are possible, and have equal likelihood. Given the labelled tree, it is clear that the first value ξ_1 of the sample is the label of the root node, but the permissible permutations associated with the left $L(T_n)$ and right subtree $R(T_n)$ of T_n may be put together in an arbitrary manner. This implies that, with $\psi(T_n)$ the number of permutations that are compatible with the outcome T_n, we have

$$\psi(T_n) = \binom{\#T_n - 1}{\#L(T_n)} \psi(L(T_n))\, \psi(R(T_n)).$$

This can easily be solved, resulting in

$$\psi(T_n) = \prod_{u \in T_n} \binom{\#T_n(u) - 1}{\#L(T_n(u))} = n! \prod_{u \in T_n} \frac{1}{\#T_n(u)} = n! \prod_{j=1}^{n-1} j^{-X_{n,j}},$$

which depends on the tree only via the associated subtree size profile. For example, with $n = 15$, 768768 different permutations lead to the tree T_{15} in Figure 1(a). Of course, as conditioning turns uniform distributions into uniform distributions, this also follows from the known formula for the probability of a specific tree under BST(n): see, *e.g.*, [12, Theorem 6.1].

References

[1] Billingsley, P. (1968) *Convergence of Probability Measures*, Wiley, New York.

[2] Chauvin, B., Drmota, M. and Jabbour-Hattab, J. (2001) The profile of binary search trees. *Ann. Appl. Probab.* **11** 1042–1062.

[3] Dennert, F. (2009) Zufällige binäre Bäume: Algorithmen, Asymptotik und Statistik. Dissertation, Leibniz Universität Hannover.

[4] Devroye, L. (1991) Limit laws for local counters in random binary search trees. *Random Struct. Alg.* **2** 303–315.

[5] Drmota, M., Janson, S. and Neininger, R. (2008) A functional limit theorem for the profile of search trees. *Ann. Appl. Probab.* **18** 288–333.

[6] Evans, S., Grübel, R. and Wakolbinger, A. (2009) Boundary theory for randomly growing binary trees. In preparation.
[7] Feng, Q., Mahmoud, H. M. and Panholzer, A. (2008) Phase changes in subtree varieties in random recursive and binary search trees. *SIAM J. Discrete Math.* **22** 160–184.
[8] Fuchs, M. (2008) Subtree sizes in recursive trees and binary search trees: Berry–Esseen bounds and Poisson approximations. *Combin. Probab. Comput.* **17** 661–680.
[9] Fuchs, M., Hwang, H.-K. and Neininger, R. (2006) Profiles of random trees: Limit theorems for random recursive trees and binary search trees. *Algorithmica* **46** 367–407.
[10] Mahmoud, H. M. (1992) *Evolution of Random Search Trees*, Wiley, New York.
[11] Neininger, R. and Rüschendorf, L. (2004) A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.* **14** 378–418.
[12] Sedgewick, R. and Flajolet, P. (1996) *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading.