Profiles of random trees: correlation and width of random recursive trees and binary search trees

MICHAEL DRMOTA
Institut für Diskrete Mathematik und Geometrie
Technische Universität Wien
Wiedner Hauptstrasse 8-10/118
1040 Wien
Austria

HSIEN-KUEI HWANG[1]
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

October 25, 2004

Abstract

We derive asymptotic approximations to the correlation coefficients of two level sizes in random recursive trees and binary search trees, which undergo sharp sign-changes when one level is fixed and the other one is varying. An asymptotic estimate for the expected width is also derived.

1 Introduction

This paper is a sequel to Drmota and Hwang (2004) and Fuchs et al. (2004)[1] in which we addressed the limit distributions of profiles (number of nodes at the levels) in random recursive trees and binary search trees. In addition to the many intriguing phenomena unveiled there, we show in this paper that the correlation coefficients of two level sizes in both classes of trees exhibit sharp sign-changes. The method of proof for deriving the uniform estimates for covariances will be useful in obtaining asymptotics of the expected widths for which only almost-sure results but no moment estimates were previously known.

Random recursive trees. Recursive trees of n nodes are non-plane, rooted, labelled trees with labels $\{1, \dots, n\}$ (at nodes) such that the labels along any path from the root form a strictly increasing sequence. By random recursive trees, we assume that all recursive trees of n nodes are equally likely. An alternative way of constructing a random recursive tree of n nodes is as follows. We start from a single node with label 1; then at the i-th insertion step, the new label i chooses any of the previous $i-1$ nodes equally likely to be its parent (and link them by an edge), and the same procedure continues until the tree contains n nodes. This procedure also implies that there are $(n-1)!$ such trees.

Recursive trees (following Meir and Moon, 1974) also appeared in other fields under different names: "concave node-weighted trees" in Tapia and Myers (1967), "growing trees" in Na and Rappoport (1970),

[1] Partially supported by a Research Award of the Alexander von Humboldt Foundation and by National Science Council under the Grant NSC-92-2118-M-001-019.

[1] This paper will be referred to as FHN throughout this paper due to frequent reference.

"pyramid scheme" in Gastwirth (1977), "heap-ordered trees" in Grossman and Larson (1989). They have been introduced as simple growing models for several real-life networks like social systems (Na and Rapoport, 1970), sales-distribution networks (Moon, 1974), and the Internet; see FHN for more references. Their simple tree representations also found applications in many linear tree algorithms; see Mitchell et al. (1979).

Profile of random recursive trees. We consider the number of nodes, denoted by $Y_{n,k}$, at distance k from the root in a random recursive tree of n nodes. Many properties of $Y_{n,k}$ are known. We briefly summarize the interesting phenomena exhibited by $Y_{n,k}$ as follows; see Drmota and Hwang (2004) and FHN for more information.

- For large, fixed n, the mean of $Y_{n,k}$ is asymptotically unimodal in k, but the variance is asymptotically *bimodal*.
- The normalized random variables $Y_{n,k}/\mathbb{E}(Y_{n,k})$ *converges in distribution* to some limit law $Y(\alpha)$ when $k \geq 1$ and $\alpha := \lim_{n\to\infty} k/\log n \in [0, e)$.
- *Convergence of all moments* of $Y_{n,k}/\mathbb{E}(Y_{n,k})$ to $Y(\alpha)$ holds only for $\alpha \in [0,1]$ but not for α outside the unit interval.
- If $\alpha = 0$ (and $k \geq 1$), then the sequence of the centered and normalized random variables $(Y_{n,k} - \mathbb{E}(Y_{n,k}))/\sqrt{\mathbb{V}(Y_{n,k})}$ converges in distribution to the standard normal law.
- If $\alpha = 1$ and $|k - \log n| \to \infty$, then $(Y_{n,k} - \mathbb{E}(Y_{n,k}))/\sqrt{\mathbb{V}(Y_{n,k})}$ converges in distribution (and with all moments) to $Y'(1)$, the same limit law as the total path length $\sum_k kY_{n,k}$.
- If $k = \log n + O(1)$, then $(Y_{n,k} - \mathbb{E}(Y_{n,k}))/\sqrt{\mathbb{V}(Y_{n,k})}$ does not converge to a fixed limit law.

Covariance of $Y_{n,k}$ and $Y_{n,h}$. The results derived in our previous papers dealt with stochastic behaviors of a fixed level size. We examine in this paper the asymptotics of the correlation coefficient of two level sizes, which turns out to undergo a sharp sign-change at $\alpha = 1$ (when the other level is fixed and not near $\log n$).

To state our results, we first introduce some notation. Define

$$f(u,v) := \frac{1}{\Gamma(u+v)(u+v-uv)} - \frac{1}{\Gamma(u+1)\Gamma(v+1)}, \tag{1}$$

where Γ is the Gamma function and

$$p(s,t) := c_2 st + c_1(s+t) + c_0, \tag{2}$$

with the coefficients given by

$$\begin{cases} c_2 := f''_{uv}(1,1) = 2 - \frac{\pi^2}{6}, \\ c_1 := -\frac{1}{2} f'''_{uv^2}(1,1) = c_2(1-\gamma) - \zeta(3) + 1, \\ c_0 := \frac{1}{4} f^{(4)}_{u^2v^2}(1,1) = c_2\left(1 + 2\gamma - \gamma^2\right) + 2c_1(1-\gamma) - \frac{\pi^4}{360}. \end{cases} \tag{3}$$

Also define

$$\begin{cases} c_3 := f'_y(\alpha,1) = -\frac{\psi(\alpha+1)+\gamma-\alpha}{\Gamma(\alpha+1)}, \\ c_4 := -\frac{1}{2} f''_{y^2}(\alpha,1) = -\frac{(\psi(\alpha+1)+1-\alpha)^2+(\alpha-1)^2-(1-\gamma)^2-\psi'(\alpha+1)-1+\pi^2/6}{2\Gamma(\alpha+1)}. \end{cases}$$

Let $k, h \geq 1$, $\alpha_{n,k} := k/\log n$, $\beta_{n,h} := h/\log n$ and α and β be their limit ratio, respectively, if exists (when n tends to infinity).

Theorem 1. *If $\alpha, \beta \in [0, 2)$, then the correlation coefficient of $Y_{n,k}$ and $Y_{n,h}$ satisfies*

$$\rho(Y_{n,k}, Y_{n,h}) \sim \begin{cases} \dfrac{\sqrt{(2k-1)(2h-1)}}{k+h-1}, & \text{if } \alpha = \beta = 0; \\ 0, & \text{if } \alpha = 0, \beta \neq 0; \\ \dfrac{f(\alpha,\beta)}{\sqrt{f(\alpha,\alpha)f(\beta,\beta)}}, & \text{if } \alpha, \beta \neq 1; \\ \dfrac{c_3 t_{n,h} + c_4}{\sqrt{f(\alpha,\alpha)p(t_{n,h}, t_{n,h})}}, & \text{if } \alpha \neq 1, \beta = 1; \\ \dfrac{p(s_{n,k}, t_{n,h})}{\sqrt{p(s_{n,k}, s_{n,k})p(t_{n,h}, t_{n,h})}}, & \text{if } \alpha = \beta = 1, \end{cases} \tag{4}$$

where $s_{n,k} := k - \log n$ and $t_{n,h} := h - \log n$.

By symmetry, all cases when $\alpha, \beta \in [0, 2)$ are covered. In particular, the result here also implies the estimates we derived for $\mathbb{V}(Y_{n,k})$ in previous papers. A comparison of the different approaches used so far for $\mathbb{V}(Y_{n,k})$ is given in the last section.

Corollaries and intuitive interpretations.

Corollary 1. *The correlation coefficient of $Y_{n,k}$ and $Y_{n,h}$ is asymptotic to zero if $k = o(\log n)$ and $k = o(h)$, where $0 \leq \beta < 2$.*

Thus the sizes at the first few levels ($k = o(\log n)$) are *asymptotically independent* of those at levels that are $\gg k$.

Corollary 2. *The correlation coefficient of $Y_{n,k}$ and $Y_{n,h}$ is asymptotic to 1 if (i) $\alpha = \beta \neq 1$ $(0 \leq \alpha, \beta < 2)$; or (ii) both $s_{n,k}, t_{n,h} \to \infty$ (not necessarily at the same rate) when $\alpha = \beta = 1$.*

The first case is intuitively clear, but the second case less transparent.

Corollary 3. *The correlation coefficient $\rho(Y_{n,k}, Y_{n,h})$ exhibits asymptotically a sharp sign-change at $\beta = 1$ when $\alpha \in (0, 2)$ is fixed and β is varying from 0 to 2.*

A few plots of the asymptotic correlation coefficient are given in Figures 1, 2, 3, highlighting in particular the discontinuous sign-change at 1.

Intuitively, the sizes of neighboring levels are expected to have positive correlation. The sharp sign-change at 1 is roughly because of the property that *almost all nodes in a random tree lie at the levels $k = \log n + O(\sqrt{\log n})$, each having about $n/\sqrt{\log n}$ nodes*, (by the estimate

$$\mathbb{E}(Y_{n,k}) \sim \frac{(\log n)^k}{k!\Gamma(\alpha_{n,k}+1)} \qquad (1 \leq k = O(\log n)),$$

and the bimodal behavior of the variance near these levels; see Drmota and Hwang, 2004). This implies that if one level k with, say $k/\log n < 1$ has more nodes, then (i) levels near $\log n$ are likely to have more nodes, and (ii) levels with $h/\log n > 1$ have fewer nodes available; this also roughly explains why $Y_{n,k}$ and $Y_{n,h}$ are negatively correlated (see Figure 1).

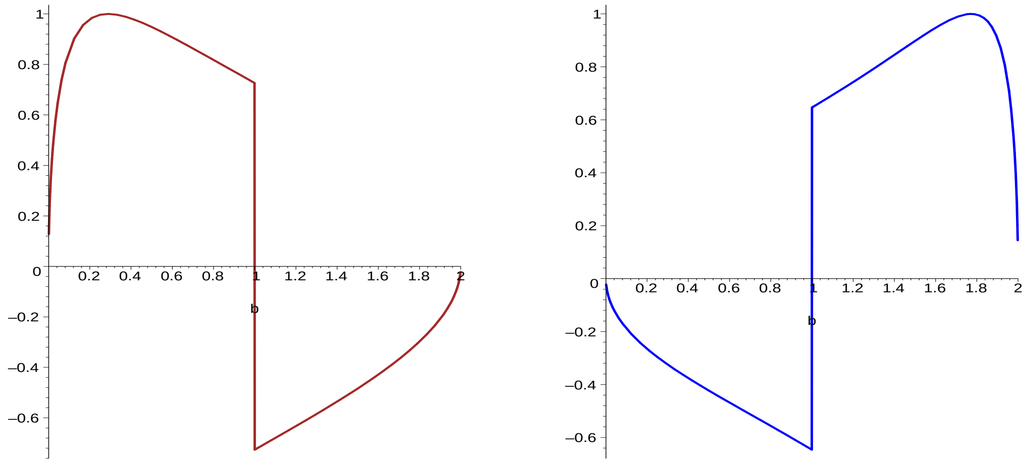

Figure 1: *Asymptotic correlation coefficient of the number of nodes at two levels. The discontinuity of sign at* 1 *is visible from both figures. Here* $\alpha = \gamma/2 \approx 0.28$ *(left) and* $\alpha = \sqrt{\pi} \approx 1.77$*, and* $\beta \in (0, 2)$.

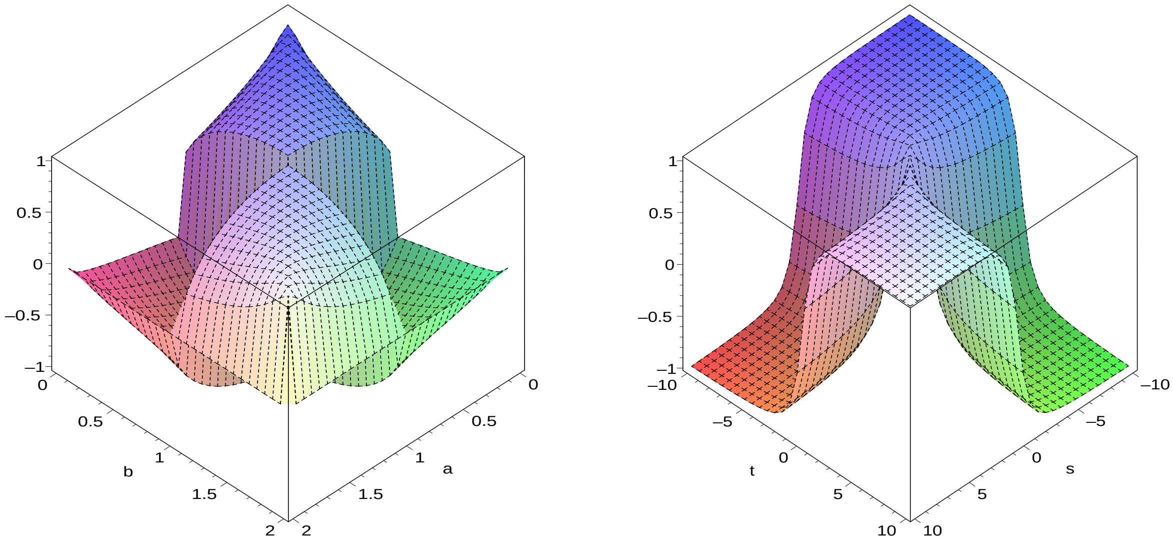

Figure 2: 3*-dimensional renderings of the limiting correlation coefficients:* $f(\alpha,\beta)/\sqrt{f(\alpha,\alpha)f(\beta,\beta)}$ *(left) and* $p(s,t)/\sqrt{p(s,s)p(t,t)}$ *(right).*

Our method of proof starts from the relation

$$\sum_{k,h} \mathbb{E}(Y_{n,k}Y_{n,h})u^k v^h = \frac{n}{u+v-uv}\left(\binom{n+u+v-1}{n} - \binom{n+uv-1}{n}\right); \tag{5}$$

see below for a self-contained proof or van der Hofstad et al. (2002). Then (4) is derived by a uniform estimate for the function on the right-hand side in the u, v plane (by applying the singularity analysis of Flajolet and Odlyzko, 1990) and then by extending the saddle point method used in Hwang (1995).

Width. Our analytic approach is also useful in deriving a uniform estimate for $\mathbb{E}\left((Y_{n,k} - Y_{n,h})^2\right)$, which turns out to be the crucial step for proving an asymptotic approximation to the expected width, defined to

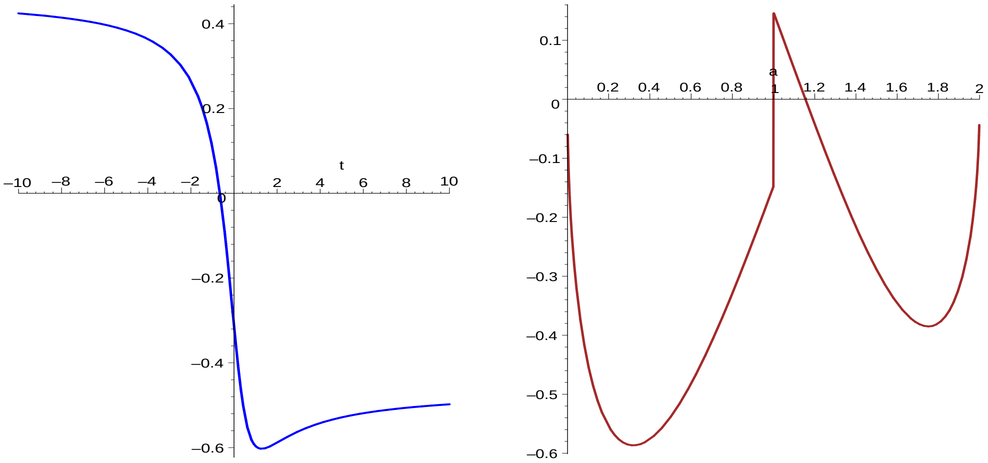

Figure 3: *Asymptotic correlation coefficient when $\beta = 1$: $\alpha = 0.1$ and t varies (left) and $t = \gamma/2$ and α varies (right).*

be $W_n := \max_k Y_{n,k}$.

Theorem 2. *The width W_n satisfies*

$$\frac{W_n}{n/\sqrt{2\pi \log n}} \to 1, \tag{6}$$

almost surely, and

$$\mathbb{E}(W_n) = \frac{n}{\sqrt{2\pi \log n}} \left(1 + O\left((\log n)^{-1/4} \log\log n\right)\right), \tag{7}$$

for any $\varepsilon > 0$.

The almost sure convergence is proved by modifying the martingale arguments used in Chauvin et al. (2001) for random binary search trees. Such arguments, based on considering the random polynomial $\sum_k Y_{n,k} z^k$, also provide a natural interpretation of the result (see FHN) that the sequence of random variables $(Y_{n,k} - \mathbb{E}(Y_{n,k}))/\sqrt{\mathbb{V}(Y_{n,k})}$ converges to the same limit law as the total path length $T_n := \sum_k kX_{n,k}$ when $k \sim \log n$ and $|k - \log n| \to \infty$; see Section 3 for more details.

Binary search trees. Binary search trees (abbreviated as BSTs) are rooted, labelled binary trees with the *search property*: all labels in the left (right) subtree of any node x are smaller (larger) than the label of x. Given a sequence of numbers, one can construct the BST by placing the first element at the root, and then by directing successively all smaller (larger) numbers to the left (right) branch. Both subtrees, if nonempty, are recursively constructed by the same procedure and are themselves BSTs; see Figure 4.

BSTs were first introduced in the early 1960's by Windley (1960), Booth and Colin (1960), Hibbard (1962), and are one of the simplest prototypical data structures; see Knuth (1997), Mahmoud (1992).

Random binary search trees. Assume that all $n!$ permutations of n elements are equally likely. Given a random permutation, we call the BST constructed from the permutation *a random BST*. We distinguish between two types of nodes: *internal nodes* are nodes holding labels and *external nodes* are virtual nodes added so that all internal nodes are of outdegree two; see Figure 4.

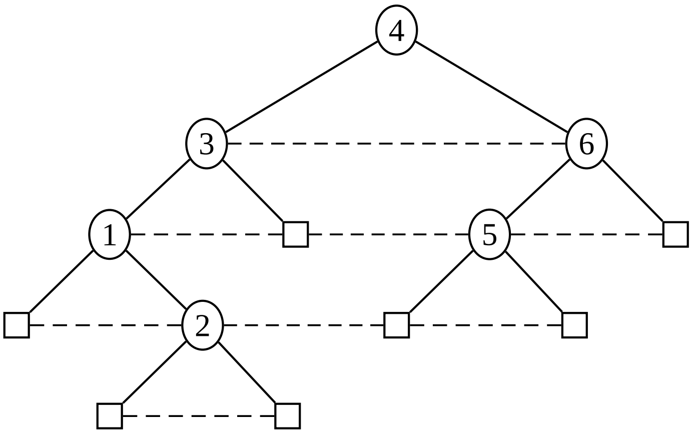

Figure 4: *The binary search tree constructed from the sequence* $\{4,3,1,6,5,2\}$. *Internal nodes are marked by circles and external nodes by squares.*

Denote by $X_{n,k}$ ($I_{n,k}$) the number of external (internal) nodes at level k in a random BST of n internal nodes, the root being at level zero. Distributional properties of both types of profile $X_{n,k}$ and $I_{n,k}$ are similar to those of $Y_{n,k}$; see FHN for details.

The interesting property here for the covariance of two level sizes is that while the limiting correlation coefficients of $I_{n,k}$ and $I_{n,h}$ exhibit a sharp sign-change at $\alpha = 2$, the limiting correlation coefficients of $X_{n,k}$ and $X_{n,h}$ exhibit two sharp sign-changes at $\alpha = 1$ and $\alpha = 2$. An intuitive interpretation will be given in Section 4.

Organization of the paper. We prove in the next section Theorem 1 on the asymptotic estimates of the correlation coefficients of two level sizes in random recursive trees. The width and related quantities are addressed in Section 3. Results for random BSTs are given in Section 4 without proof. We then conclude the paper with a brief comparative discussion of the methods of proof used to derive asymptotic estimates for the variances.

2 Correlation of two level sizes

We prove Theorem 1 in this section. Note that the L_2-convergence of $Y_{n,k}/\mu_{n,k}$ (see FHN) can also be applied to prove (4) in the case when $\alpha, \beta \notin \{0,2\}$, we give here a uniform approach applicable to all cases.

Recurrence of $Y_{n,k}$ and $\mathbb{E}(Y_{n,k})$. All our results are based on the recurrence relation satisfied by $Y_{n,k}$

$$Y_{n,k} \overset{\mathscr{D}}{=} Y_{\mathsf{uniform}[1,n-1],k-1} + Y^*_{n-\mathsf{uniform}[1,n-1],k} \qquad (n \geq 2; k \geq 1), \tag{8}$$

with the initial values $Y_{n,0} = \delta_{n,1}$, the Kronecker symbol, where the random variable $\mathsf{uniform}[1, n-1]$ takes values $\{1, \dots, n-1\}$ with equal probability, and $Y^*_{n,k}$ is an independent copy of $Y_{n,k}$; see FHN or van der Hofstad et al. (2002).

Let $\mu_{n,k} := \mathbb{E}(X_{n,k})$. From (8), it follows that the mean satisfies

$$\begin{aligned}\mu_{n,k} &= [u^k]\binom{n+u-1}{n-1}\\ &= \frac{(\log n)^k}{k!\Gamma(\alpha_{n,k}+1)}\left(1+O\left((\log n)^{-1}\right)\right), \end{aligned} \tag{9}$$

where $[u^k]F(u)$ denotes the coefficient of u^n in the Taylor expansion of F and the O-term holds uniformly for $1 \leq k = O(\log n)$; see Hwang (1995).

Proof of (5). We now prove (5). By (8), we have the recurrence

$$\begin{aligned}\mathbb{E}(Y_{n,k}Y_{n,h}) &= \frac{1}{n-1}\sum_{1\leq j<n}\left(\mathbb{E}(Y_{j,k-1}Y_{j,h-1}) + \mathbb{E}(Y^*_{j,k}Y^*_{j,h})\right)\\ &\quad + \frac{1}{n-1}\sum_{1\leq j<n}\left(\mu_{j,k-1}\mu_{n-j,h} + \mu_{j,k}\mu_{n-j,h-1}\right).\end{aligned}$$

Let $F_n(u,v) := \sum_{k,h}\mathbb{E}(X_{n,k}X_{n,h})u^kv^h$. Then $F_1(u,v) = 1$ and

$$F_n(u,v) = \frac{1+uv}{n-1}\sum_{1\leq j<n}F_j(u,v) + \frac{u+v}{n-1}\sum_{1\leq j<n}\binom{j+u-1}{j-1}\binom{n-j+v-1}{n-j-1}, \tag{10}$$

for $n \geq 2$. The last sum is equal to

$$\frac{u+v}{n-1}[z^n]z^2(1-z)^{-u-v-2} = \frac{u+v}{n-1}\binom{n+u+v-1}{n-2}.$$

The recurrence (10) is then either solved by considering $nF_{n+1} - (n-1)F_n$ and then iterating the resulting first-order difference equation or solved by considering the differential equation satisfied by $\sum_n F_{n+1}z^n$. This proves (5). ▮

An asymptotic expansion for the covariance. We now derive an asymptotic expansion for $\mathrm{Cov}(Y_{n,k}, Y_{n,h})$.

First, by singularity analysis (see Flajolet and Odlyzko, 1990), we have

$$n\binom{n+w-1}{n} = n[z^n](1-z)^{-w} = \frac{n^w}{\Gamma(w)}\left(1+O\left(|w|^2n^{-1}\right)\right),$$

the O-term holding uniformly for finite complex w. Thus, by (5) and (9),

$$\mathrm{Cov}(Y_{n,k}, Y_{n,h}) = C_{k,h}(n)\left(1+O\left(n^{-1}\right)\right) + O\left(\delta_{k,h}\frac{(\log n)^k}{k!}\right), \tag{11}$$

uniformly for $1 \leq k, h \leq K\log n$, where

$$C_{k,h}(n) := [u^kv^h]f(u,v)n^{u+v},$$

with f defined in (1).

If $\alpha+\beta \neq 0$, we apply the saddle point method used in Hwang (1995) by first expanding f as follows

$$f(u,v) = \sum_{\ell,r\geq 0} f_{\ell,r}(u-\alpha_{n,k})^{\ell}(v-\beta_{n,h})^{r},$$

where $f_{\ell,r} := f^{(\ell+r)}_{u^{\ell}v^{r}}(\alpha_{n,k},\beta_{n,h})/(\ell!r!)$; and then integrating term by term gives the formal expansion

$$C_{k,h}(n) \sim \sum_{\ell,r\geq 0} f_{\ell,r}\Xi_{\ell}(n,k)\Xi_{r}(n,h), \tag{12}$$

where

$$\begin{aligned}\Xi_{\ell}(n,k) &:= [u^k](u-\alpha_{n,k})^{\ell}n^{u}\\ &= \frac{(\log n)^k}{k!}\sum_{0\leq j\leq \ell}\binom{\ell}{j}(-\alpha_{n,k})^{\ell-j}\frac{k\cdots(k-j+1)}{(\log n)^j} \qquad (\ell\geq 0).\end{aligned}$$

The first few values of Ξ_r are as follows.

$$\begin{aligned}&\Xi_0(n,k)=1, \quad \Xi_1(n,k)=0, \quad \Xi_2(n,k) = -\frac{k}{(\log n)^2},\\ &\Xi_3(n,k) = \frac{2k}{(\log n)^3}, \quad \Xi_4(n,k) = \frac{3k(k-2)}{(\log n)^4}.\end{aligned}$$

Since $\Xi_r(n,k)$ equals $(\log n)^{-r}$ times a polynomial in k of degree $\lfloor r/2\rfloor$, the double sum on the right-hand side of (12) can be regrouped and gives an asymptotic expansion when $k = O(\log n)$. The error analysis is similar to those in Hwang (1995, 1997), and we obtain that (12) holds uniformly for $1\leq k,h\leq 2\log n - \omega_n\sqrt{\log n}$, ω_n being any sequence tending to infinity. The error term $[u^kv^h]C_{k,h}(n)O(n^{-1})$ appearing in (11) is handled similarly and is asymptotically negligible.

By the explicit forms of the Ξ_{ℓ}'s, we obtain the expansion

$$\begin{aligned}C_{k,h}(n) = \frac{(\log n)^{k+h}}{k!h!}\Bigg\{ & f_{0,0} - \frac{1}{\log n}(f_{2,0}\alpha_{n,k}+f_{0,2}\beta_{n,h})\\ &+\frac{1}{(\log n)^2}\left(3(f_{4,0}\alpha_{n,k}^2+f_{0,4}\beta_{n,h}^2)+f_{2,2}\alpha_{n,k}\beta_{n,h}+2(f_{3,0}\alpha_{n,k}+f_{0,3}\beta_{n,h})\right)\\ &+O\left((\log n)^{-3}\right)\Bigg\},\end{aligned} \tag{13}$$

which is sufficient for our use.

Special cases. Assume that $0\leq \alpha,\beta<2$. If $\alpha,\beta\notin\{0,1\}$, then

$$f_{0,0} = f(\alpha_{n,k},\beta_{n,h}) \to f(\alpha,\beta)\neq 0,$$

and we obtain

$$\mathrm{Cov}(Y_{n,k},Y_{n,h}) \sim f(\alpha,\beta)\frac{(\log n)^{k+h}}{k!h!},$$

uniformly for $1\leq k,h\leq 2\log n-\omega_n\sqrt{\log n}$. This proves Theorem 4 when $\alpha,\beta\notin\{0,1\}$. It also implies that

$$\mathbb{V}(X_{n,k}) \sim f(\alpha,\alpha)\frac{(\log n)^{2k}}{k!^2} \qquad (1\leq k\leq 2\log n-\omega_n\sqrt{\log n}).$$

If $\alpha = \beta = 1$, then, by (13) and the following approximations

$$f_{0,0} \sim c_2 \frac{s_{n,k} t_{n,h}}{(\log n)^2}, \qquad f_{0,2} \sim -c_1 \frac{s_{n,k}}{\log n}, \qquad f_{2,0} \sim -c_1 \frac{t_{n,h}}{\log n}, \qquad f_{2,2} \sim c_0,$$

where $k = \log n + s_{n,k}$, $h = \log n + t_{n,h}$ and the coefficients c_j's are defined in (3), we obtain

$$\mathrm{Cov}(Y_{n,k}, Y_{n,h}) \sim \frac{p(s_{n,k}, t_{n,h})}{(\log n)^2} \cdot \frac{(\log n)^{k+h}}{k!h!},$$

where p is given in (2). This also implies that $\mathbb{V}(X_{n,k}) \sim p(s_{n,k}, s_{n,k})(\log n)^{2k-2}/k!^2$.

If $\alpha = 0$ and $\beta \in (0,1)$, then, similarly as above, we have

$$\begin{aligned} &\mathrm{Cov}(Y_{n,k}, Y_{n,h}) \\ &\sim \begin{cases} -\dfrac{(\log n)^{k+h-1}}{(k-1)!h!\Gamma(\beta+1)} \left(\psi(\beta+1) - 1 + \gamma\right), & \text{if } \beta \neq 1; \\ \dfrac{(\log n)^{k+h-2}}{(k-1)!h!\Gamma(\beta+1)} \left(\left(1 - \dfrac{\pi^2}{6}\right) s_{n,h} + 2 - \gamma - \zeta(3) - \dfrac{\pi^2}{4} + \dfrac{\pi^2\gamma}{6} \right), & \text{if } \beta = 1, \end{cases} \end{aligned}$$

so that $\rho(Y_{n,k}, Y_{n,h}) \to 0$ in both cases.

The case when $\beta = 1$ and $\alpha \neq 1$ is treated similarly.

A change of variables $u \mapsto wv$ is useful for the remaining case when $\alpha = \beta = 0$; then a similar analysis gives

$$\mathrm{Cov}(Y_{n,k}, Y_{n,h}) \sim \frac{(\log n)^{k+h-1}}{(k-1)!(h-1)!(k+h-1)}. \tag{14}$$

Alternatively, we can use the exact expression (see van der Hofstad et al., 2002)

$$\mathbb{E}(Y_{n,k} Y_{n,h}) = \sum_{0 \le j \le k} \binom{2j+h-k}{j+h-k} [w^{j+h+1}] \binom{n-1+w}{n-1},$$

obtained from expanding the right-hand side of (5), and then proceed similarly as above (the two terms with indices $j = k-1, k$ suffice for obtaining (14)). ▮

Proof of Corollary 3. When $\alpha, \beta \in (0,2)$, $\alpha \neq 1$, we have, by (4),

$$\lim_{\beta \to 1} \frac{f(\alpha,\beta)}{\sqrt{f(\alpha,\alpha) f(\beta,\beta)}} = \mathrm{sign}(1-\beta) \frac{\psi(\alpha+1) - \alpha + \gamma}{\sqrt{c_2 f(\alpha,\alpha)}\, \Gamma(\alpha+1)};$$

thus the sign-change follows. The case when $\alpha = 1$ can also be checked similarly. ▮

The proof of other corollaries to Theorem 1 is straightforward and omitted.

3 Profile and width

Profiles of trees are closely related to many other shape parameters. We discuss briefly in this section the connection between profile and width, starting from deriving an asymptotic estimate for the expected width, namely from the proof of (7). Then we consider the level polynomial $M_n(z) := \sum_k Y_{n,k} z^k$, which will be seen to be a convenient tool for proving (6) and for bridging the limit properties of the profile and those of the total path length (and other weighted path lengths).

The expected width. Since the width is defined by $W_n = \max_k Y_{n,k}$, we have, by the estimate (9),

$$\mathbb{E}(W_n) \geq \max_k \mathbb{E}(Y_{n,k}) = \frac{n}{\sqrt{2\pi \log n}} \left(1 + O\left((\log n)^{-1/2}\right)\right).$$

However, it is less clear how to obtain a tight upper bound. The arguments introduced in Chauvin et al. (2001) can be used to prove the almost sure convergence result (6) (see below for a sketch of proof), but do not lead to an effective upper bound for the expected width. We introduce a new argument, reducing the upper bound to estimating the mean and the variance of some differences between level sizes, and show that the lower bound is indeed tight.

We start with a probabilistic lemma.

Lemma 1. *Let $Z(t)$ be a sequence of stochastic processes on the space of continuous functions on $[0,1]$. Assume that there exist constants $\lambda \geq 0$ and $\theta > 1$ such that*

$$\mathbb{P}\left(|Z(t_1) - Z(t_2)| \geq \delta\right) = O\left(|t_1 - t_2|^\theta \delta^{-\lambda}\right), \tag{15}$$

uniformly for all $t_1, t_2 \in [0,1]$. Then we have

$$\mathbb{P}\left(\max_{|s-t|\leq\varepsilon} |Z(s) - Z(t)| \geq \delta\right) = O\left(\varepsilon^{\theta-1}\delta^{-\lambda}\right). \tag{16}$$

Proof. We modify the proof of Theorem 12.3 in Billingsley (1968). First, the assumption (15) is exactly (12.50) from Billingsley (1968) with $F(t) = t$. It follows that for $\varepsilon > 0$ (and $1/\varepsilon$ is an integer; compare with (12.57) there)

$$\sum_{j<1/\varepsilon} \mathbb{P}\left(\sup_{j\varepsilon\leq s\leq(j+1)\varepsilon} |Z(s) - Z(j\varepsilon)| \geq \delta\right) = O\left(\varepsilon^{\theta-1}\delta^{-\lambda}\right).$$

Similarly, we obtain

$$\sum_{j<1/\varepsilon} \mathbb{P}\left(\sup_{j\varepsilon\leq s\leq(j+1)\varepsilon} |Z(s) - Z((j+1)\varepsilon)| \geq \delta\right) = O\left(\varepsilon^{\theta-1}\delta^{-\lambda}\right).$$

Now, suppose that there exist $s, t \in [0,1]$ with $|s-t| \leq \varepsilon$ and $|Z(s) - Z(t)| \geq \delta$. Then there exists $j < 1/\varepsilon$ with $\max(|s - j\varepsilon|, |t - j\varepsilon|) < \varepsilon$ and $\max(|Z(s) - Z(j\varepsilon)|, |Z(t) - Z(j\varepsilon)|) \geq \delta/2$. Consequently

$$\begin{aligned}\mathbb{P}\left(\max_{|s-t|\leq\varepsilon} |Z(s) - Z(t)| \geq \delta\right) &\leq \sum_{j<1/\varepsilon} \mathbb{P}\left(\sup_{j\varepsilon\leq s\leq(j+1)\varepsilon} |Z(s) - Z(j\varepsilon)| \geq \delta/2\right) \\ &\quad + \sum_{j<1/\varepsilon} \mathbb{P}\left(\sup_{j\varepsilon\leq s\leq(j+1)\varepsilon} |Z(s) - Z((j+1)\varepsilon)| \geq \delta/2\right) \\ &= O\left(\varepsilon^{\theta-1}\delta^{-\lambda}\right).\end{aligned}$$

This proves (16) for all ε such that $1/\varepsilon$ is an integer. However, the general case also follows from the O-estimate. ▮

Lemma 2. *Let $\Delta := h - k$ and $\overline{Y}_{n,k} := Y_{n,k} - \mathbb{E}(Y_{n,k})$. Then, uniformly for $k, h = \log n + o(\log n)$,*

$$\mathbb{E}\left((\overline{Y}_{n,k} - \overline{Y}_{n,h})^2\right) = O\left(n^2\Delta^2(\log n)^{-3}\right). \tag{17}$$

Proof. We may apply the results in previous section for the covariance of $Y_{n,k}$ and $Y_{n,h}$ in some ranges, but they do not lead to a uniform estimate in terms of $|k-h|^2$ in the whole range when $\alpha = \beta = 1$.

We give here a self-contained proof of (17). Assume, without loss of generality, that $h \geq k$. By (11), we have

$$\begin{aligned}\mathbb{E}\left((\overline{Y}_{n,k} - \overline{Y}_{n,h})^2\right) &= \left([u^k v^k] - 2[u^k v^h] + [u^h v^h]\right) f(u,v) n^{u+v} \left(1 + O(n^{-1})\right) \\ &\quad + O\left(\delta_{k,h} n (\log n)^{-1/2}\right).\end{aligned}$$

It suffices to find upper bounds for the dominant term

$$\begin{aligned}J &:= \left([u^k v^k] - 2[u^k v^h] + [u^h v^h]\right) f(u,v) n^{u+v} \\ &= \frac{1}{(2\pi)^2} \iint_{\mathcal{D}} e^{-ikx-iky} \left(1 - 2e^{-i\Delta y} + e^{-i\Delta(x+y)}\right) f\left(e^{ix}, e^{iy}\right) n^{e^{ix}+e^{iy}} \,\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

where $\mathcal{D} := [-\pi, \pi]^2$. Now

$$\begin{aligned}1 - 2e^{-i\Delta y} + e^{-i\Delta(x+y)} &= \left(1 - e^{-i\Delta y}\right)^2 + e^{-i\Delta y}\left(e^{-i\Delta x} - 1 + i\Delta x\right) \\ &\quad - e^{-i\Delta y}\left(e^{-i\Delta y} - 1 + i\Delta y\right) + e^{-i\Delta y}\left(i\Delta y - i\Delta x\right) \\ &=: Q_1 + Q_2 - Q_3 + Q_4.\end{aligned}$$

Let

$$J_m := \frac{1}{(2\pi)^2} \iint_{\mathcal{D}} Q_m e^{-ikx-iky} f\left(e^{ix}, e^{iy}\right) n^{e^{ix}+e^{iy}} \,\mathrm{d}x\,\mathrm{d}y \qquad (m = 1, \dots, 4).$$

By the elementary inequalities $|e^{iw} - 1| \leq |w|$ for real w and $1 - \cos w \geq c_5 w^2$ for $|w| \leq \pi$, where $c_5 := 2/\pi^2$, we have

$$|J_1| \leq \frac{n^2\Delta^2}{(2\pi)^2} \iint_{\mathcal{D}} y^2 \left|f\left(e^{ix}, e^{iy}\right)\right| n^{-c_5(x^2+y^2)} \,\mathrm{d}x\,\mathrm{d}y.$$

This, together with the uniform bound

$$\left|f\left(e^{ix}, e^{iy}\right)\right| = O(|xy|),$$

for $x, y \in \mathcal{D}$, yield

$$\begin{aligned}|J_1| &= O\left(n^2\Delta^2 \iint_{\mathcal{D}} |x||y|^3 n^{-c_5(x^2+y^2)} \,\mathrm{d}x\,\mathrm{d}y\right) \\ &= O\left(n^2\Delta^2 (\log n)^{-3}\right).\end{aligned}$$

Similarly, by the inequality $|e^{iw} - 1 - iw| \leq |w|^2/2$ for real w, we have

$$|J_2|, |J_3| = O\left(n^2\Delta^2(\log n)^{-3}\right).$$

For the last integral J_4, we use the expansion

$$f\left(e^{ix}, e^{iy}\right) = c_2 i^2 xy + O\left(|xy||x+y|\right),$$

and obtain $J_4 = J_5 + J_6$, where

$$\begin{aligned}J_5 &:= \frac{c_2\Delta}{(2\pi)^2} \iint_{\mathcal{D}} i^3 (y-x) xy e^{-ikx-iky-i\Delta y} n^{e^{ix}+e^{iy}} \,\mathrm{d}x\,\mathrm{d}y, \\ J_6 &= O\left(n^2|\Delta| \iint_{\mathcal{D}} |xy|(|x|+|y|)^2 n^{-c_5(x^2+y^2)} \,\mathrm{d}x\,\mathrm{d}y\right) \\ &= O\left(n^2|\Delta|(\log n)^{-3}\right).\end{aligned}$$

For J_5, we use the expansion

$$e^{-i\Delta y} = 1 + O(|\Delta||y|),$$

and the relation

$$\iint_{\mathcal{D}} (y-x)xye^{-ik(x+y)}n^{e^{ix}+e^{iy}}\,\mathrm{d}x\,\mathrm{d}y = 0,$$

(by symmetry), so that

$$\begin{aligned} J_5 &= O\left(n^2\Delta^2 \iint_{\mathcal{D}} |x||y|^2(|x|+|y|)n^{-c_5(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y\right) \\ &= O\left(n^2\Delta^2(\log n)^{-3}\right), \end{aligned}$$

uniformly for $k, h = \log n + o(\log n)$. This completes the proof of (17). ▮

Lemma 3. *Uniformly for* $k, h = \log n + o(\log n)$*,*

$$|\mathbb{E}(Y_{n,k} - Y_{n,h})| = O\left(n|\Delta|(\log n)^{-1}\right); \tag{18}$$

and uniformly for $k = \log n + O(1)$ *and* $h = \log n + o((\log n)^{2/3})$*,*

$$|\mathbb{E}(Y_{n,k}) - \mathbb{E}(Y_{n,h})| \sim \frac{n}{\sqrt{2\pi \log n}}\left(1 - e^{-(k-h)^2/(2\log n)}\right). \tag{19}$$

Proof. Assume that $|k - \log n| \le |h - \log n|$. By Cauchy's integral formula

$$\mathbb{E}(Y_{n,k}) - \mathbb{E}(Y_{n,h}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ikx}\left(1 - e^{-i\Delta x}\right)\frac{n^{e^{ix}}}{\Gamma(1+e^{ix})}\left(1 + O(n^{-1})\right)\,\mathrm{d}x.$$

In the first case when $k, h = \log n + o(\log n)$, we apply the the inequality $|1 - e^{-i\Delta x}| \le |\Delta x|$ and the same arguments as above, yielding

$$|\mathbb{E}(Y_{n,k}) - \mathbb{E}(Y_{n,h})| = O\left(|\Delta| n(\log n)^{-1}\right),$$

uniformly in k, h. This proves (18).

The approximation (19) follows from a straightforward application of the usual saddle-point method. ▮

An upper bound for the expected width. Let $k_0 = \lfloor \log n \rfloor$. Take

$$\Lambda := \lfloor (\log n)^{1/4+\xi_n} \rfloor \quad \text{and} \quad J := \lfloor (\log n)^{1/4} \rfloor,$$

where $\xi_n \in (0, 1/2)$ will be specified below. We use the following upper bound

$$\begin{aligned} W_n &\le \max_{0\le|j|\le J} Y_{n,k_0+j\Lambda} + \max_{|k-h|\le\Lambda} |\overline{Y}_{n,k} - \overline{Y}_{n,h}| \\ &\quad + \max_{|k-h|\le\Lambda} |\mathbb{E}(Y_{n,k} - Y_{n,h})| + \sum_{|k-k_0|\ge(\log n)^{1/2+\xi_n}} Y_{n,k} \\ &=: W_n^{(1)} + W_n^{(2)} + W_n^{(3)} + W_n^{(4)}. \end{aligned}$$

We show that, when taking expectation, the term Y_{n,k_0} in $W_n^{(1)}$ is dominant and all other terms are asymptotically of smaller order than $\mathbb{E}(Y_{n,k_0})$.

We start from $W_n^{(4)}$. By (9),

$$\mathbb{E}(W_n^{(4)}) = O\left(\sum_{|k-k_0|\geq(\log n)^{1/2+\xi_n}} \frac{(\log n)^k}{k!}\right)$$
$$= O\left(ne^{-(\log n)^{2\xi_n}/2}(\log n)^{-\xi_n}\right);$$

see Hwang (1997). If we choose

$$\xi_n := \frac{\log\log\log n}{\log\log n},$$

then

$$\mathbb{E}(W_n^{(4)}) = o\left(n(\log n)^{-1}\right).$$

For $W_n^{(3)}$, we have, by (18) for $k, h = \log n + O(J\Lambda)$ and by (9) for k, h outside this range,

$$\max_{|h-k|\leq\Lambda} |\mathbb{E}(Y_{n,h} - Y_{n,k})| = O\left(n\Lambda(\log n)^{-1}\right)$$
$$= O\left(n(\log n)^{-3/4+\xi_n}\right).$$

We then apply Lemmas 1 and 2 to prove that

$$\mathbb{E}(W_n^{(2)}) = O\left(\frac{n}{\sqrt{\log n}}\left(\frac{\Lambda}{(\log n)^{3/2-\xi_n}}\right)^{1/2}\right). \tag{20}$$

We first define $Y_n(t)$, $-1 \leq t \leq 1$, by

$$Y_n(t) = \overline{Y}_{n,k_0+t(\log n)^{1/2+\xi_n}} \frac{(\log n)^{1-\xi_n}}{n},$$

when $t(\log n)^{1/2+\xi_n}$ is an integer, and by linear interpolation otherwise. By Lemma 2, we have

$$\mathbb{E}((Y_n(s) - Y_n(t))^2) = O\left((s-t)^2\right),$$

uniformly for $s, t \in [-1, 1]$. By Chebyshev inequality,

$$\mathbb{P}(|Y_n(s) - Y_n(t)| \geq w) = O\left((s-t)^2 w^{-2}\right).$$

Take $\eta_n := \Lambda(\log n)^{-1/2-\xi_n}$. It follows, by Lemma 1, that

$$\mathbb{P}\left(\max_{|s-t|\leq\eta_n} |Y_n(s) - Y_n(t)| \geq w\right) = O\left(\eta_n w^{-2}\right),$$

and, consequently,

$$\mathbb{E}\left(\max_{|s-t|\leq\eta_n} |Y_n(s) - Y_n(t)|\right) = \int_0^\infty \mathbb{P}\left(\max_{|s-t|\leq\eta_n} |Y_n(s) - Y_n(t)| \geq w\right)\, \mathrm{d}w$$
$$= O\left(\eta_n^{1/2}\right).$$

This and the definition of $Y_n(t)$ imply (20), which can be written as

$$\mathbb{E}(W_n^{(2)}) = O\left(n(\log n)^{-9/8+2\xi_n}\right).$$

Thus it remains to find an upper bound for $W_n^{(1)}$. By Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}(W_n^{(1)}) &\le \mathbb{E}\sum_{|j|\le J} Y_{n,k_0+j\Lambda}\cdot \mathbf{1}_{[Y_{n,k_0+j\Lambda}=W_n^{(1)}]}\\ &\le \sum_{|j|\le J}\left(\mathbb{E}(Y_{n,k_0+j\Lambda}^2)\right)^{1/2}\mathbb{P}(Y_{n,k_0+j\Lambda}=W_n^{(1)})^{1/2}\\ &\le \frac{n}{\sqrt{2\pi\log n}}+O\left(n(\log n)^{-1}\right)\\ &\qquad + O\left(\frac{n}{\sqrt{\log n}}\sum_{1\le|j|\le J}\mathbb{P}(Y_{n,k_0+j\Lambda}=W_n^{(1)})^{1/2}\right).\end{aligned}$$

Here we used the estimates

$$\left(\mathbb{E}(Y_{n,k_0}^2)\right)^{1/2}=\frac{n}{\sqrt{2\pi\log n}}+O\left(n(\log n)^{-1}\right),$$

and $\left(\mathbb{E}(Y_{n,k}^2)\right)^{1/2}=O(n/\sqrt{\log n})$; see Drmota and Hwang (2004).

Set $D_j := Y_{n,k_0}-Y_{n,k_0+j\Lambda}$ for $1\le|j|\le J$. Then we have

$$\begin{aligned}\mathbb{P}(Y_{n,k_0+j\Lambda}=W_n^{(1)}) &\le \mathbb{P}(D_j\le 0)\\ &\le \mathbb{P}\left(|D_j-\mathbb{E}(D_j)|\ge \mathbb{E}(D_j)\right)\\ &\le \frac{\mathbb{V}(D_j)}{(\mathbb{E}(D_j))^2},\end{aligned}$$

by Chebyshev inequality.

By Lemma 2, we have

$$\mathbb{V}(D_j)=O\left(\frac{n^2}{(\log n)^3}j^2\Lambda^2\right).$$

This and (19) imply that

$$\mathbb{P}(D_j\le 0)^{1/2}=O\left(\frac{|j|\Lambda}{\log n}\left(1-e^{-j^2\Lambda^2/(2\log n)}\right)^{-1}\right),$$

for $1\le|j|\le J$; and it follows that

$$\begin{aligned}\sum_{1\le|j|\le J}\mathbb{P}(D_j\le 0)^{1/2} &= O\left(\frac{\Lambda}{\log n}\int_1^J x\left(1-e^{-x^2\Lambda^2/(2\log n)}\right)^{-1}\mathrm{d}x\right)\\ &= O\left(\Lambda^{-1}(\log n)^{2\xi_n}\right).\end{aligned}$$

Thus

$$\mathbb{E}(W_n^{(1)})\le\frac{n}{\sqrt{2\pi\log n}}+O\left(n(\log n)^{-3/4+\xi_n}\right).$$

Collecting these estimates gives

$$\mathbb{E}W_n\le\frac{n}{\sqrt{2\pi\log n}}\left(1+O\left((\log n)^{-1/4}\log\log n\right)\right),$$

which proves (7).

A possible refinement of the error term in (7). If we had the estimates

$$\mathbb{E}\left((\overline{Y}_{n,k}-\overline{Y}_{n,h})^{2m}\right)=O\left(n^{2m}\Delta^{2m}(\log n)^{-3m}\right),$$

for $m\geq 2$, then the error term $O(\log n)^{-1/4}\log\log n)$ in the approximation to the expected width would be improved to $O((\log n)^{-1/2+\varepsilon})$ for some $\varepsilon>0$, which is, up to $(\log n)^{\varepsilon}$, expected to be the right-order. A proof of these moment estimates would be to apply induction and the approach used in FHN, but the details would be very messy.

Asymptotics of the level polynomials. The proof of the almost sure convergence (6) follows from the same martingale arguments introduced in Chauvin et al. (2001). Thus we only sketch a few steps of the proof here.

We observe first that the normalized random function $\bar{M}_n(z):=M_n(z)/\mathbb{E}(M_n(z))$ (where $M_n(z):=\sum_k Y_{n,k}z^k$) is a martingale. Roughly, this reflects the construction that the new-coming key has the same probability of being attached to any of the existing nodes. Also by (9)

$$\mathbb{E}(M_n(z))=\binom{n-1+z}{n-1}.$$

By the martingale convergence theorem (see Hall and Heyde, 1980), $\bar{M}_n(\alpha)$ converges almost surely to a limit $M(\alpha)$ for any finite $\alpha>0$. Then by the recursive definition (8) of $Y_{n,k}$, we deduce, similar to contraction method (see FHN), that

$$M(\alpha)\overset{\mathscr{D}}{=}\alpha U^{\alpha}M(\alpha)+(1-U)^{\alpha}M(\alpha)^*,$$

where $M(\alpha)^*\overset{\mathscr{D}}{=}M(\alpha)$ and $M(\alpha),M(\alpha)^*,U$ are independent. This implies that $M(\alpha)\overset{\mathscr{D}}{=}Y(\alpha)$ for every $\alpha>0$.

Interestingly, this limit relation also extends to complex values of α.

Lemma 4. *For any compact set in $\{z\in\mathbb{C}:|z-1|<1\}$, the martingale $\bar{M}_n(z)$ converges almost surely, uniformly and in L_2 to its limit $M(z)$ (which is also an analytic function).*

The key step of the proof is to use an explicit expression for $\mathbb{E}(M_n(z_1)M_n(z_2))$ (see (5)), and to use Kolmogorov's criterion, coupling with vector martingale theorems. Finally, one recovers $Y_{n,k}$ almost surely (and uniformly for $1-\varepsilon\leq k/\log n\leq 1+\varepsilon$ for some $\varepsilon>0$) via Cauchy's integral formula

$$\begin{aligned}Y_{n,k}&=\frac{1}{2\pi i}\oint_{|z|=\alpha_{n,k}}M_n(z)z^{-k-1}\,\mathrm{d}z\\&\sim\frac{1}{2\pi i}\oint_{|z|=\alpha_{n,k}}M(z)\,\mathbb{E}(M_n(z))z^{-k-1}\,\mathrm{d}z\\&\sim M(\alpha_{n,k})\,\frac{1}{2\pi i}\oint_{|z|=\alpha_{n,k}}\mathbb{E}(M_n(z))z^{-k-1}\,\mathrm{d}z\\&\sim M(\alpha)\,\mathbb{E}(Y_{n,k}).\end{aligned}$$

We omit all technical details. Note that the radius $\alpha_{n,k}:=k/\log n$ in the contour integration is a natural choice because it is the saddle point of the integrand $\mathbb{E}(M_n(z))z^{-k-1}$. Since $M(z)$ is almost surely an analytic function and $M(1)=1$, it follows that

$$W_n=\max_k Y_{n,k}\sim\max_k\mathbb{E}(Y_{n,k})\sim\frac{n}{\sqrt{2\pi\log n}},$$

almost surely. This completes the proof of (6). ▮

Total path length.

Corollary 4. *Let T_n denote the total path length in a random recursive tree of n nodes. Then $\bar{M}_n'(1)$ is a martingale and*

$$\bar{M}_n'(1) = \frac{T_n - \mathbb{E}(T_n)}{n} \xrightarrow{\mathscr{D}} Y'(1),$$

almost surely and in L_2.

Proof. Since $T_n = \sum_k kY_{n,k}$, we have $\bar{M}_n'(1) = (T_n - \mathbb{E}(T_n))/n$ by the definition of $M_n(z)$. From Lemma 4, it follows that

$$\begin{aligned}\bar{M}_n'(1) &= \frac{1}{2\pi i}\int_{|z-1|=\delta<1} z^{-2}\bar{M}_n(z)\,\mathrm{d}z\\ &\to \frac{1}{2\pi i}\int_{|z-1|=\delta<1} z^{-2}M(z)\,\mathrm{d}z\\ &= M'(1) = Y'(1),\end{aligned}$$

almost surely. ∎

The result is already known; see Mahmoud (1991) and Dobrow and Fill (1999). However, our approach also gives

$$\bar{M}_n^{(m)}(1) \to M^{(m)}(1) \qquad (m \geq 1),$$

almost surely and in L_2. In particular, when $m = 2$, we have

$$\frac{1}{n}\sum_k k(k-1)(Y_{n,k} - \mu_{n,k}) - \frac{2}{n}\mathbb{E}(T_n)(T_n - \mathbb{E}(T_n)) \to M''(1).$$

Note that $M_n^{(m)}(1)$ is also a martingale for $m \geq 1$.

4 Profile of random binary search trees

We give in this section the corresponding results for the profiles of random BSTs. The proofs are similar to those for random recursive trees and are thus omitted. Recall that $X_{n,k}$ and $I_{n,k}$ denote the number of external nodes and internal nodes, respectively, at level k in a random BST of n elements.

4.1 External nodes

It is known since Lynch (1965) that

$$\sum_k \mathbb{E}(X_{n,k})u^k = \binom{n+2u-1}{n} \qquad (n \geq 0);$$

see also Françon (1977) or Mahmoud (1992).

Lemma 5. *For $n \geq 0$*

$$\begin{aligned}\sum_{k,h}\mathbb{E}(X_{n,k}X_{n,h})u^k v^h = &\frac{2uv}{2u+2v-2uv-1}\binom{n+2u+2v-2}{n}\\ &+ \frac{2u+2v-4uv-1}{2u+2v-2uv-1}\binom{n+2uv-1}{n}.\end{aligned}$$

This simplifies Lemma 4 in Chauvin et al. (2001).

From this lemma, we deduce, by singularity analysis (see Flajolet and Odlyzko, 1990), that

$$\mathbb{E}(X_{n,k}X_{n,h}) = 2^{k+h}[u^k v^h]\phi(u,v)n^{u+v-2}\left(1+O\left(n^{-1}\right)\right) + O\left(\delta_{k,h}\frac{(2\log n)^k}{k!n}\right),$$

uniformly for $\alpha, \beta \in [2-\sqrt{2}+\varepsilon, 2+\sqrt{2}-\varepsilon]$ for any $\varepsilon > 0$, where

$$\phi(u,v) := \frac{uv}{(2u+2v-uv-2)\Gamma(u+v-1)} - \frac{1}{\Gamma(u)\Gamma(v)}. \tag{21}$$

Theorem 3. *For $\alpha, \beta \in (2-\sqrt{2}, 2+\sqrt{2})$, the correlation coefficient $\rho(X_{n,k}, X_{n,h})$ is asymptotic to*

$$\begin{cases} \dfrac{\phi(\alpha,\beta)}{\sqrt{\phi(\alpha,\alpha)\phi(\beta,\beta)}}, & \textit{if } \alpha,\beta \notin \{1,2\}; \\ \dfrac{\phi'_v(\alpha,\beta)t_{n,h} - \frac{1}{2}\phi''_{v^2}(\alpha,\beta)}{\sqrt{\phi(\alpha,\alpha)p(\beta,\beta;t_{n,h},t_{n,h})}}, & \textit{if } \alpha \notin \{1,2\}, \beta \in \{1,2\}; \\ \dfrac{p(\alpha,\beta;s_{n,k},t_{n,h})}{\sqrt{p(\alpha,\alpha;s_{n,k},s_{n,k})p(\beta,\beta;t_{n,h},t_{n,h})}}, & \textit{if } \alpha,\beta \in \{1,2\}, \end{cases}$$

where

$$p(j,\ell;s,t) := \phi''_{uv}(j,\ell)st - \frac{1}{2}\left(j\phi'''_{u^2v}(j,\ell)t + \ell\phi'''_{uv^2}(j,\ell)s\right) + \frac{j\ell}{4}\phi^{(4)}_{u^2v^2}(j,\ell).$$

Unlike the profile of recursive trees, the limiting correlation coefficients of $\rho(X_{n,k}, X_{n,h})$ undergo two sharp sign-changes at 1 and 2; see Figures 5 and 6.

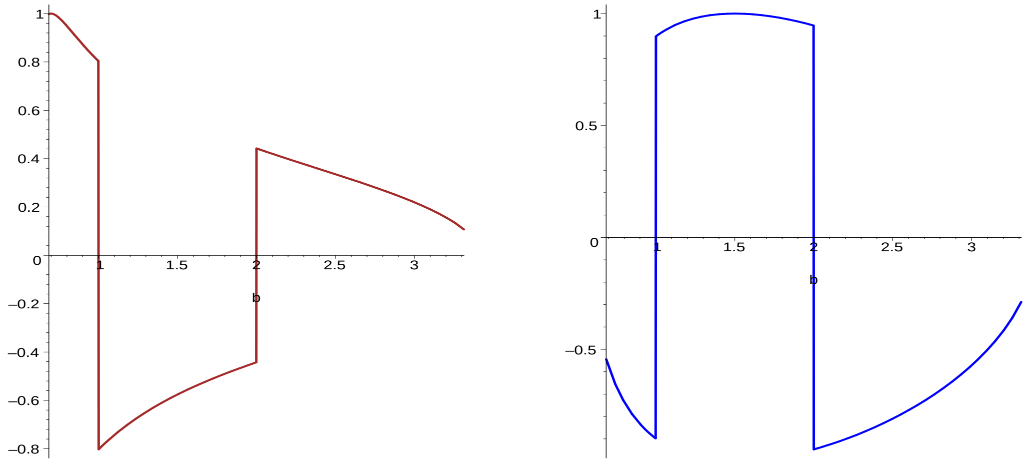

Figure 5: *Two sharp sign-changes for $\phi(\alpha,\beta)/\sqrt{\phi(\alpha,\alpha)\phi(\beta,\beta)}$: $\alpha = 0.7$ (left) and $\alpha = 1.5$ (right).*

Width. The same arguments as above lead to

$$\mathbb{E}(\max_k X_{n,k}) = \frac{n}{\sqrt{4\pi\log n}}\left(1+O\left((\log n)^{-1/4}\log\log n\right)\right).$$

This result is new. The corresponding almost sure convergence was established in Chauvin et al. (2001).

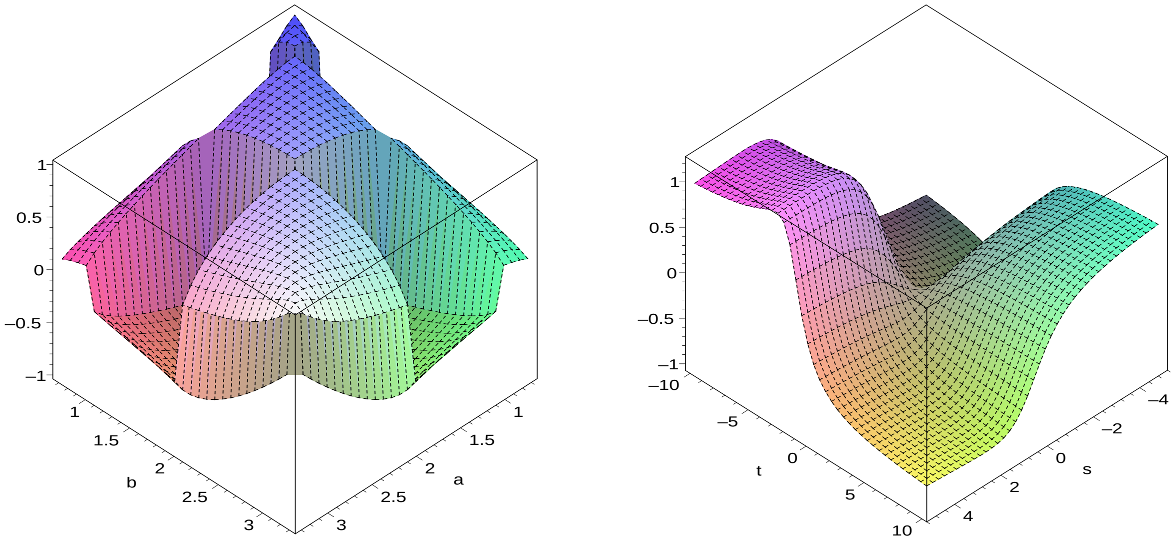

Figure 6: 3-*dimensional renderings of the limiting correlation coefficients:* $\alpha, \beta \in (2-\sqrt{2}, 2+\sqrt{2}) \setminus \{1, 2\}$ *(left) and* $\alpha = 1, \beta = 2$ *(right).*

4.2 Internal nodes

For internal nodes, we have

$$\mathbb{E}(I_{n,k}) = [u^k]\frac{1 - \binom{n+2u-1}{n}}{1-2u} = 2^k[u^k]\frac{1 - \binom{n+u-1}{n}}{1-u};$$

see Brown and Shubert (1984) or Mahmoud (1992).

Lemma 6. *For* $n \geq 0$

$$\begin{aligned}\sum_{k,h} \mathbb{E}(I_{n,k}I_{n,h})u^k v^h = &\frac{1}{(1-2u)(1-2v)}\left(1 - \binom{n+2u-1}{n} - \binom{n+2v-1}{n}\right) \\ &+ \frac{2uv}{(1-2u)(1-2v)(2u+2v-2uv-1)}\binom{n+2u+2v-2}{n} \\ &- \frac{1}{2u+2v-2uv-1}\binom{n+2uv-1}{n}.\end{aligned}$$

From this lemma, it follows, again by singularity analysis, that

$$\mathbb{E}(X_{n,k}X_{n,h}) = 2^{k+h}[u^k v^h]\varphi(u,v)n^{u+v-2}\left(1 + O\left(n^{-1}\right)\right) + O\left(\frac{(2\log n)^k}{k!n} + \frac{(2\log n)^h}{h!n}\right),$$

uniformly for $\alpha, \beta \in [2-\sqrt{2}+\varepsilon, 2+\sqrt{2}-\varepsilon]$ (for any $\varepsilon > 0$), where

$$\varphi(u,v) := \frac{\phi(u,v)}{(1-u)(1-v)},$$

ϕ being defined in (21).

Theorem 4. *For $\alpha, \beta \in (2-\sqrt{2}, 2+\sqrt{2})$, the correlation coefficient $\rho(X_{n,k}, X_{n,h})$ is asymptotic to*

$$\begin{cases} \dfrac{\varphi(\alpha,\beta)}{\sqrt{\varphi(\alpha,\alpha)\varphi(\beta,\beta)}}, & \textit{if } \alpha,\beta \notin \{2\}; \\ \dfrac{\varphi'_v(\alpha,2)t_{n,h} - \frac{1}{2}\varphi''_{v^2}(\alpha,2)}{\sqrt{\varphi(\alpha,\alpha)q(t_{n,h},t_{n,h})}}, & \textit{if } \alpha \neq 2, \beta = 2; \\ \dfrac{q(s_{n,k},t_{n,h})}{\sqrt{q(s_{n,k},s_{n,k})q(t_{n,h},t_{n,h})}}, & \textit{if } \alpha = \beta = 2, \end{cases}$$

where

$$q(s,t) := \varphi''_{uv}(2,2)st - (\varphi'''_{uv^2}(2,2)s + \varphi'''_{u^2v}(2,2)t) + \varphi^{(4)}_{u^2v^2}(2,2).$$

Figure 7 depicts the single sign-change of the limiting correlation coefficients $\varphi(\alpha,\beta)/\sqrt{\varphi(\alpha,\alpha)\varphi(\beta,\beta)}$; compare Figures 5 and 6.

Note that $\varphi(1,1) = c_2 = 2 - \pi^2/6$. Thus $\rho(I_{n,k}, I_{n,h}) \to 1$ when (i) $k, h \sim \alpha \log n$ where $\alpha \neq 2$ and (ii) $k, h \sim 2\log n$ and $|k - 2\log n|, |h - 2\log n| \to \infty$.

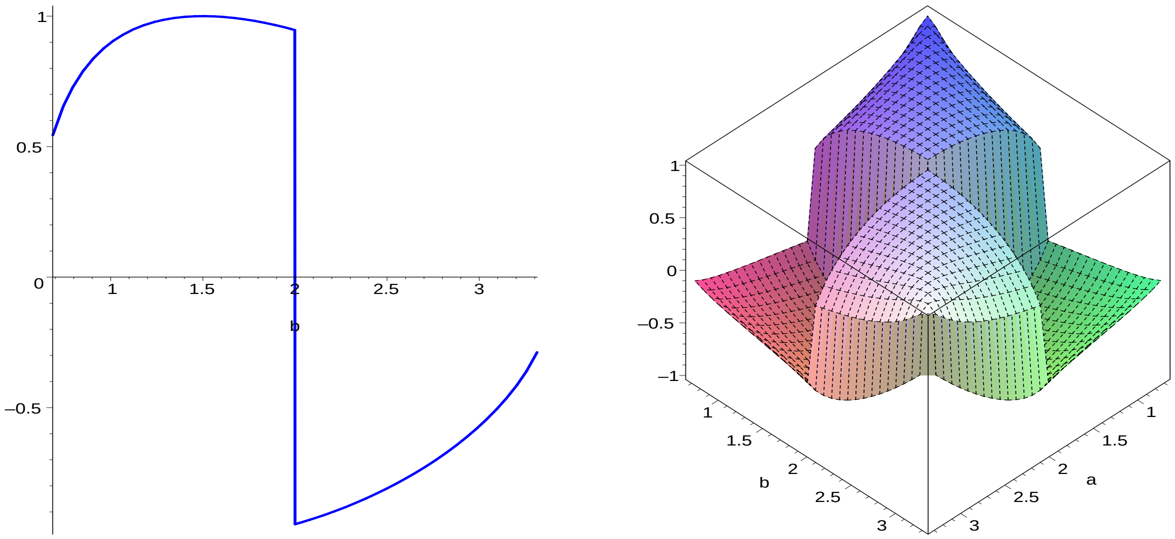

Figure 7: *Asymptotic correlation coefficients: $\alpha = 1.5$ and β varies (left), and a 3-dimensional plot (right) for $\alpha, \beta \in (2-\sqrt{2}, 2+\sqrt{2})$.*

An intuitive interpretation of the sign-change. For internal nodes, the behavior and the corresponding intuitive interpretation of the limiting correlation coefficients are similar to those of $Y_{n,k}$ (of recursive trees). The double sign-change of the limit of $\rho(X_{n,k}, X_{n,h})$ is roughly explained as follows. Observe first that

$$\mathbb{E}(I_{n,k}) \sim \begin{cases} 2^k - \dfrac{\mathbb{E}(X_{n,k})}{1-\alpha_{n,k}}, & \text{if } 1 \le k \le \log n - (\log n)^{2/3-\varepsilon}, \\ 2^k \Phi\left(\dfrac{\log n - k}{\sqrt{\log n}}\right), & \text{if } |k - \log n| \le (\log n)^{2/3-\varepsilon}, \\ \dfrac{\mathbb{E}(X_{n,k})}{\alpha_{n,k}-1}, & \text{if } k \ge \log n + (\log n)^{2/3-\varepsilon}, \end{cases}$$

for any $\varepsilon > 0$, where $\Phi(x)$ is the standard normal distribution function; see FHN. This says roughly that levels up to $(1-\varepsilon)\log n$ are full of internal nodes (since in this range $\mathbb{E}(X_{n,k}) = o(2^k)$) with less room for external nodes; outside this range, the number of internal nodes at each level is asymptotically of the same order as that of external nodes. Thus if $X_{n,k}$ with, say $\alpha \in (1,2)$ has more nodes, then this means that there are also more internal nodes at this and neighboring levels, which implies that there are fewer nodes available at levels $\leq (1-\varepsilon)\log n$ and levels $\geq (2+\varepsilon)\log n$, similar to the interpretation given in Introduction for recursive trees.

5 Conclusions

In this paper we describe the sharp sign-change phenomena in the correlation coefficients of two level sizes in random recursive trees and random BSTs. Such sign-changes are consistent with the bimodality of the variance in the middle range ($k \sim \log n$ for recursive trees and $k \sim 2\log n$ for BSTs).

We conclude this paper with a brief comparison of the different approaches we used for the variance (and covariance) of profiles. In Hwang and Drmota (2004), we introduced two approaches for $\mathbb{V}(X_{n,k})$ and $\mathbb{V}(Y_{n,k})$, one based on explicit integral representations in terms of Bessel functions and the other on explicit expressions in terms of Stirling numbers of the first kind. But extending the two approaches to $\mathbb{V}(I_{n,k})$ is not easy. In FHN, we used an approach based on recurrence and asymptotic transfer, which applies well to all three profiles we discussed in this paper. But getting more terms in the asymptotic expansions by this approach is effortful. The approach we present in this paper is not only more general (applicable to covariance and to more profiles) but also useful in deriving asymptotic expansions if needed. Note that by the L_2-convergence of the normalized profiles (established by, say the contraction method), the leading estimates for the variance or covariance can also be derived by the fixed-point equation of the limit laws. But this approach fails for the ranges when the limit laws are degenerate.

The major open question here is: what is the limit distribution (if it exists) of the width? Is it the same as the limit law of total path length (in both class of random trees considered in this paper)?

Acknowledgement

We thank Luc Devroye for pointing our an error in an earlier version of the paper. Part of the work of the second author was carried out during a visit to Institut für Stochastik und Mathematische Informatik, J. W. Goethe-Universität (Frankfurt). He thanks the Institute for hospitality and support; he is also indebted to Tsung-Hsi Tsai for helpful discussions.

References

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

[2] A. D. Booth and J. T. Colin (1960). On the efficiency of a new method of dictionary construction. *Information and Control* **3** 327–334.

[3] G. G. Brown and B. O. Shubert (1984). On random binary trees. *Mathematics of Operations Research* **9** 43–65.

[4] B. Chauvin, M. Drmota and J. Jabbour-Hattab (2001). The profile of binary search trees. *Annals of Applied Probability* **11** 1042–1062.

[5] B. Chauvin, T. Klein, J.-F. Marckert and A. Rouault (2003). Martingales, embedding and tilting of binary trees. Preprint, 2003.

[6] B. Chauvin and A. Rouault (2004). Connecting Yule process, bisection and binary search tree via martingales. *Journal of the Iranian Statistical Society 3* 88-116.

[7] M. Drmota and H.-K. Hwang (2004). Bimodality and phase transitions in the profile variance of random binary search trees. *SIAM Journal on Discrete Mathematics* accepted for publication.

[8] P. Flajolet and A. M. Odlyzko (1990). Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics* **3** 216–240.

[9] J. Françon (1977). On the analysis of algorithms for trees. *Theoretical Computer Science* **4** 155–169.

[10] M. Fuchs, H.-K. Hwang and R. Neininger (2004). Profiles of random trees: Limit theorems for random recursive trees and binary search trees. Manuscript submitted for publication. Available at algo.stat.sinica.edu.tw.

[11] J. L. Gastwirth (1977). A probability model of a pyramid scheme, *The American Statistician* **31** 79–82.

[12] R. Grossman and R. G. Larson (1989). Hopf-algebraic structure of families of trees. *Journal of Algebra* **126** 184–210.

[13] T. N. Hibbard (1962). Some combinatorial properties of certain trees with applications to searching and sorting. *Journal of the Association for Computing Machinery* **9** 13–28.

[14] H.-K. Hwang (1995). Asymptotic expansions for the Stirling numbers of the first kind. *Journal of Combinatorial Theory, Series A* **71** 343–351.

[15] H.-K. Hwang (1997). Asymptotic estimates for elementary probability distributions. *Studies in Applied Mathematics* **99** 393–417.

[16] D. E. Knuth (1997). *The Art of Computer Programming, Volume III: Sorting and Searching*. Second edition. Addison-Wesley, Reading, MA.

[17] G. Louchard (1987). Exact and asymptotic distributions in digital and binary search trees. *RAIRO Informatique Théorique et Applications* **21** 479–495.

[18] W. C. Lynch (1965). More combinatorial properties of certain trees. *Computer Journal* **7** 299–302.

[19] H. M. Mahmoud (1992). *Evolution of Random Search Trees*. John Wiley & Sons, New York.

[20] A. Meir and J. W. Moon (1974). Cutting down recursive trees. *Mathematical Biosciences* **21** 173–181.

[21] S. L. Mitchell, E. J. Cockayne and S. T. Hedetniemi (1979). Linear algorithms on recursive representations of trees. *Journal of Computer System and Science* **18** 76–85.

[22] J. W. Moon (1974). The distance between nodes in recursive trees. In "*Combinatorics*", edited by T. P. McDonough and V. C. Marron. London Mathematical Society Lecture Notes, Series 13, London, pp. 125–132.

[23] H. S. Na and A. Rapoport (1970). Distribution of nodes of a tree by degree. *Mathematical Biosciences* **6** 313–329.

[24] M. A. Tapia and B. R. Myers (1967). Generation of concave node-weighted trees, *IEEE Transactions on Circuits and Systems* **14** 229–230.

[25] R. van der Hofstad, G. Hooghiemstra and P. van Mieghem (2002). On the covariance of the level sizes in random recursive trees. *Random Structures and Algorithms* **20** 519–539.

[26] P. F. Windley (1960). Trees, forests and rearranging. *Computer Journal* **3** 84–88.