Additive tree functionals with small toll functions and subtrees of random trees

Stephan Wagner[†]

Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

Many parameters of trees are additive in the sense that they can be computed recursively from the sum of the branches plus a certain toll function. For instance, such parameters occur very frequently in the analysis of divide-and-conquer algorithms. Here we are interested in the situation that the toll function is small (the average over all trees of a given size n decreases exponentially with n). We prove a general central limit theorem for random labelled trees and apply it to a number of examples. The main motivation is the study of the number of subtrees in a random labelled tree, but it also applies to classical instances such as the number of leaves.

Keywords: additive tree functional, small toll function, number of subtrees, size of subtrees, random trees

1 Introduction

A parameter $F(T)$ defined for rooted trees T is called an additive tree functional if it satisfies the recursion

$$F(T) = \sum_{i=1}^{k} F(T_i) + t(T), \tag{1}$$

where $T_1, \dots, T_k$ are the branches of the tree and $t(T)$ is a so-called toll function, which often only depends on the size of T. The recursion remains true for the tree $T = \bullet$ of order 1 if we assume without loss of generality that $F(\bullet) = t(\bullet)$. Examples include:

- The number of leaves, corresponding to the toll function

$$t(T) = \begin{cases} 1 & |T| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

- The number of vertices with a certain fixed outdegree k, in which case one can simply take

$$t(T) = \begin{cases} 1 & \text{if the root of } T \text{ has outdegree } k, \\ 0 & \text{otherwise.} \end{cases}$$

[†] This material is based upon work supported financially by the National Research Foundation of South Africa under grant number 70560.

- The internal path length, i.e., the sum of the distances from the root to all vertices, which can be obtained from the toll function $t(T) = |T| - 1$.

- The log-product of the subtree sizes [10], also called the "shape functional" [4], corresponding to $t(T) = \log |T|$.

Such functionals also arise frequently in the study of divide-and-conquer algorithms, e.g. quicksort [6]. Central limit theorems for different classes of functionals of this type have been studied for various families of random trees. See for instance [5, Chapter IX.7] or [2, Chapter 3] for examples including the number of leaves and general patterns in rooted trees, which are asymptotically normally distributed for a wide class of rooted trees. The internal path length has been shown to follow an Airy distribution by Takács [13, 14]. It also occurs prominently in Janson's analysis of the Wiener index [8] (the sum of all distances between pairs of vertices), which can itself be regarded as an additive tree functional. Fill and Kapur [4] determine limiting distributions for Catalan trees (binary trees enumerated by the Catalan numbers) for the special toll functions $t(T) = |T|^\alpha$ and $t(T) = \log |T|$ (the aforementioned shape functional). In the analysis of these toll functions, Hadamard products occur naturally, which is also an important theme in the paper of Fill, Flajolet and Kapur [3], where a very general approach leading to asymptotic formulas for the mean of additive tree functionals is provided. In [4], Fill and Kapur also mention conditions that (at least heuristically) guarantee a Gaussian central limit theorem, but a formal proof still seems to be missing.

The original motivation for the present work was the study of the number of subtrees in random trees. Let T be a random labelled tree, and let $s(T)$ denote the number of its subtrees (connected induced subgraphs of T, excluding the empty graph). The mean of this parameter was first studied by Meir and Moon [9] for simply generated families of trees (which include, amongst others, labelled trees, d-ary trees and plane trees). In the case of random labelled trees, it is asymptotically equal to $(e/(e-1))^{3/2}e^{n/e}$, and it is not much harder to determine that the variance is of asymptotic order K^n for a constant K whose numerical value is $2.15483 > e^{2/e}$, so the variance grows faster than the mean squared (K is a solution of the equation $T(T(1/(eK)))^2 = T(1/(eK))/e$, where $T(x)$ is the exponential generating function for rooted labelled trees). In view of this growth, one cannot expect the "usual" renormalisation (subtract the mean and divide by the standard deviation) to yield a limiting distribution. However, one can hope for such a limit distribution if one takes $\log s(T)$ instead. As it turns out, this parameter is essentially an additive tree functional.

It will be technically convenient to consider the number $s_1(T)$ of subtrees that include the root of a rooted tree T instead of $s(T)$ itself. We will, however, be able to infer the limiting distribution of $\log s(T)$. Note that $s_1(T)$ satisfies the recursion

$$s_1(T) = \prod_{i=1}^{k} (1 + s_1(T_i))$$

if $T_1, \ldots, T_k$ are the branches (each subtree induces either a subtree containing the root or the empty set on each of the branches). It follows that

$$\log(1 + s_1(T)) = \sum_{i=1}^{k} \log(1 + s_1(T_i)) + \log(1 + s_1(T)^{-1}), \tag{2}$$

hence $\log(1+s_1(T))$ is an additive tree functional (although the toll function involves s_1 itself). However, a-priori estimates show that the toll function is very small, and that the average of the toll function over all rooted labelled trees of a given size n decreases exponentially with n. In the following section, we prove a general theorem that an additive tree functional whose toll function satisfies this property will always follow a Gaussian distribution in the limit. In Sections 3 and 4, we apply this theorem to both old and new examples, in particular the number and the average size of subtrees of labelled trees.

2 The general central limit theorem

As indicated in the introduction, we restrict ourselves to random labelled trees, which we can assume to be rooted. Our goal is to prove that certain additive functionals asymptotically follow a Gaussian distribution. To be precise, we assume that the toll function is bounded and that it satisfies

$$n^{-(n-1)} \sum_{|T|=n} |t(T)| = O(c^n) \tag{3}$$

for some positive constant $c < 1$, where the sum is taken over all rooted labelled trees of order n (recall that their number is n^{n-1}). We will see that this is indeed the case for various examples. The theorem that is proved in this section reads as follows:

Theorem 1 *Suppose that the toll function $t(T)$ satisfies* (3). *Then the mean of $F(T_n)$ is asymptotically*

$$\mathbb{E}(F(T_n)) = \mu n + O(1).$$

The centred moments are asymptotically

$$\mathbb{E}((F(T_n) - \mu n)^r) = \begin{cases} (r-1)!!\sigma^r n^{r/2} + O\left(n^{r/2-1}\right) & r \text{ even,} \\ O\left(n^{(r-1)/2}\right) & r \text{ odd.} \end{cases}$$

Here the constants μ and σ are given by

$$\mu = \sum_T t(T) \frac{e^{-|T|}}{|T|!} \quad \text{and} \quad \sigma^2 = \sum_T t(T)(2F(T) - t(T)) \frac{e^{-|T|}}{|T|!} - 2\mu \sum_T t(T) \frac{e^{-|T|}}{(|T|-1)!}.$$

The sums are taken over all rooted labelled trees. If $\sigma \neq 0$, then the distribution of the additive functional $F(T)$ induced by $t(T)$ converges, upon renormalisation, to a Gaussian distribution.

Remark 1 *The theorem remains true if the conditions are satisfied by $t(T) + C$ for some constant C. This is because the function $F(T)$ only changes to $F(T) + C|T|$ when $t(T)$ is replaced by $t(T) + C$, so the difference is deterministic.*

While proofs of central limit theorems, especially in the Gaussian case, are often based on studying the analytic behaviour of bivariate generating functions, this does not seem to be feasible in our setting: an extra variable is needed to obtain a functional equation (see (6) below), and it is not clear how analytical information on the multivariate generating function can be extracted from it. Hence we appeal to the classical method of “moment pumping” in the proof that follows. An alternative might be to consider rooted labelled trees as a conditioned Galton-Watson process and apply probabilistic tools.

Proof of Theorem 1: The exponential generating function $T(x)$ is well-known to satisfy the functional equation

$$T(x) = x\exp(T(x)). \tag{4}$$

For the sake of convenience, we introduce a multivariate generating function that suits our purpose. Let $G(x, y, z)$ be defined by

$$G(x,y,z) = \sum_T \frac{x^{|T|}}{|T|!} y^{-t(T)} z^{F(T)}, \tag{5}$$

where the sum is taken over all rooted labelled trees. Clearly, $G(x, 1, 1) = T(x)$. The recursive relation (1) for $F(T)$ now translates to a simple functional equation: any rooted labelled tree T is constructed from an unordered collection of k rooted labelled trees $T_1, T_2, \ldots, T_k$ ($k = 0$ if $T = \bullet$), and the labels of T can be assigned in $|T| \cdot \binom{|T_1|+\cdots+|T_k|}{|T_1|,\ldots,|T_k|}$ ways. Then since

$$F(T) - t(T) = F(T_1) + \cdots + F(T_k),$$

we have

$$\begin{aligned}
G(x,z,z) &= \sum_T \frac{x^{|T|}}{|T|!} z^{-t(T)} z^{F(T)} \\
&= \sum_{k\geq 0} \frac{1}{k!} \sum_{T_1,\ldots,T_k} (1 + |T_1| + \cdots + |T_k|) \binom{|T_1| + \cdots + |T_k|}{|T_1|, \ldots, |T_k|} \frac{x^{1+|T_1|+\cdots+|T_k|} z^{F(T_1)+\cdots+F(T_k)}}{(1 + |T_1| + \cdots + |T_k|)!} \\
&= x \sum_{k\geq 0} \frac{1}{k!} \sum_{T_1,\ldots,T_k} \frac{x^{|T_1|}}{|T_1|!} z^{F(T_1)} \cdots \frac{x^{|T_k|}}{|T_k|!} z^{F(T_k)} \\
&= x \sum_{k\geq 0} \frac{1}{k!} \left(\sum_T \frac{x^{|T|}}{|T|!} z^{F(T)} \right)^k \\
&= x \exp(G(x,1,z)).
\end{aligned} \tag{6}$$

In order to obtain the moments of $F(T)$, we need the partial derivatives with respect to z. Define

$$G_r(x) = \left(\frac{\partial}{\partial z}\right)^r G(x,1,1) = \sum_T \frac{x^{|T|}}{|T|!} F(T)^{\underline{r}}.$$

Here the derivative is taken with respect to the variable z in our definition (5). Faà di Bruno's formula, applied to (6), yields

$$\sum_{k=0}^r \binom{r}{k} \frac{\partial^r}{\partial y^k \partial z^{r-k}} G(x,1,1) = x \exp(G(x,1,1)) \cdot r! \sum_{\mathbf{m} \vdash r} \prod_{j=1}^r \frac{(G_j(x)/j!)^{m_i}}{m_i!},$$

where the sum is over all partitions $\mathbf{m} = (m_1, m_2, \ldots, m_r)$ of r (i.e., $m_1 + 2m_2 + \ldots + rm_r = r$ has to be satisfied). Let $A_r(x)$ denote the left hand side without the term $G_r(x)$ corresponding to $k = 0$. Note that $x\exp(G(x, 1, 1)) = G(x, 1, 1) = T(x)$. Summing the identity above over all $r \geq 1$, we easily get

$$\sum_{r\geq 1} \frac{G_r(x) + A_r(x)}{r!} v^r = T(x) \left(\exp\Big(\sum_{r\geq 1} \frac{G_r(x)}{r!} v^r \Big) - 1 \right).$$

Write

$$\mathcal{G}(v,x) = \sum_{r\ge 1} \frac{G_r(x)}{r!} v^r \qquad \text{and} \qquad \mathcal{A}(v,x) = \sum_{r\ge 1} \frac{A_r(x)}{r!} v^r,$$

so that

$$\mathcal{G}(v,x) + \mathcal{A}(v,x) = T(x)\left(\exp(\mathcal{G}(v,x)) - 1\right), \tag{7}$$

considered as a formal power series in v. Now define the differential operator Φ by $\Phi Q(x) = xQ'(x)$ for any function $Q(x)$. It satisfies the product rule as well as

$$\Phi Q(R(x)) = Q'(R(x))\Phi(R(x)),$$

hence we can apply Faà di Bruno's formula to it as well to obtain from (4) that

$$\Phi^r T(x) = \sum_{\ell=0}^{r} \binom{r}{\ell} \Phi^{r-\ell}(x) \Phi^{\ell} \exp(T(x)) = x \exp(T(x)) \sum_{\ell=0}^{r} \binom{r}{\ell} \cdot \ell! \sum_{\mathbf{m}\vdash \ell} \prod_{j=1}^{\ell} \frac{(\Phi^j T(x)/j!)^{m_j}}{m_j!},$$

from which one deduces

$$\sum_{r\ge 1} \frac{\Phi^r T(x)}{r!} v^r = T(x) \left(\exp\left(v + \sum_{r\ge 1} \frac{\Phi^r T(x)}{r!} v^r \right) - 1 \right)$$

or, with

$$\mathcal{H}(v,x) = v + \sum_{r\ge 1} \frac{\Phi^r T(x)}{r!} v^r,$$

the identity

$$\mathcal{H}(v,x) - v = T(x)\left(\exp(\mathcal{H}(v,x)) - 1\right).$$

Comparing with (7), we infer that

$$\mathcal{G}(v,x) = \mathcal{H}(-\mathcal{A}(v,x), x).$$

Now we can simply extract $G_r(x)$ as the coefficient of v^r. For a partition $\mathbf{m} \vdash r$, we use the abbreviation $|\mathbf{m}| = m_1 + m_2 + \ldots + m_r$ for the length. Then we obtain

$$\begin{aligned} G_r(x) &= r![v^r] \left(-\mathcal{A}(v,x) + \sum_{\ell\ge 1} \frac{\Phi^{\ell} T(x)}{\ell!} (-\mathcal{A}(v,x))^{\ell} \right) \\ &= -A_r(x) + r! \sum_{\mathbf{m}\vdash r} (-1)^{|\mathbf{m}|} \Phi^{|\mathbf{m}|} T(x) \prod_{j=1}^{r} \frac{A_j(x)^{m_j}}{m_j! j!^{m_j}}. \end{aligned} \tag{8}$$

Now we are ready to apply singularity analysis to derive the asymptotic behaviour of the moments, which is where our condition on the toll function comes into play. Let M be an upper bound on $|t(T)|$. Then it follows by a simple induction that $|F(T)| \le M \cdot |T|$ for any tree T. Now note that

$$\frac{\partial^r}{\partial y^k \partial z^{r-k}} G(x,1,1) = \sum_{T} (-t(T))^{\underline{k}} F(T)^{\underline{r-k}} \frac{x^{|T|}}{|T|!}.$$

Then the absolute value of the coefficient of x^n is

$$\frac{1}{n!}\left|\sum_{|T|=n}(-t(T))^{\underline{k}}F(T)^{\underline{r-k}}\right| \leq \frac{(M+r)^{k-1}(Mn+r)^{r-k}}{n!}\sum_{|T|=n}|t(T)| \ll \frac{n^{r-k}c^n n^{n-1}}{n!} \ll n^r(ce)^n$$

for any $k > 0$, which shows that the radius of convergence of $A_r(x)$ is strictly greater than $1/e$, so that it is analytic in a disk around zero of radius greater than $1/e$. On the other hand it is well known that $T(x)$ has a square-root singularity at $x = 1/e$ with expansion

$$T(x) = \sum_{n\geq 1}\frac{n^{n-1}}{n!}x^n = 1 - \sqrt{2(1-ex)} + \frac{2}{3}(1-ex) + \cdots .$$

It further follows that

$$\Phi^r T(x) = 2^{1/2-r}(2r-3)!!(1-ex)^{1/2-r} - \frac{2}{3}[r=1] + O\left((1-ex)^{3/2-r}\right).$$

The mean of our parameter F is given by

$$\mathbb{E}(F(T_n)) = \frac{n!}{n^{n-1}}[x^n]G_1(x).$$

We see from (8) that

$$G_1(x) = -A_1(x)(1+\Phi T(x)).$$

A_1 is analytic in a disk of radius greater than $1/e$ and in particular at $1/e$ itself, so the dominant singularity of $G_1(x)$ is still $1/e$, and the expansion around $1/e$ starts with

$$G_1(x) = -\frac{A_1(1/e)}{\sqrt{2}}(1-ex)^{-1/2} + C + O((1-ex)^{1/2}).$$

Applying the standard theorems of singularity analysis [5, Chapter VI], we obtain

$$\mathbb{E}(F(T_n)) = \frac{n!}{n^{n-1}}[x^n]G_1(x) = \frac{n!}{n^{n-1}}\cdot\frac{-A_1(1/e)}{\sqrt{2\pi n}}e^n(1+O(1/n)) = -A_1(1/e)\cdot n + O(1),$$

as it was claimed. Let us now consider the higher centred moments. Set $\mu = -A_1(1/e)$. We have

$$\mathbb{E}\left((F(T_n)-\mu n)^r\right) = \sum_{\ell=0}^{r}\binom{r}{\ell}(-\mu n)^{r-\ell}\mathbb{E}(F^\ell(T_n)),$$

and since the generating function of the ℓth moment is, with $\left\{{\ell \atop k}\right\}$ denoting a Stirling number of the second kind,

$$\sum_{k=1}^{\ell}\left\{{\ell \atop k}\right\}G_k(x),$$

and multiplication by n amounts to an application of the operator Φ on the level of generating functions, we have to consider

$$K_r(x) = \sum_{\ell=0}^{r} \binom{r}{\ell} (-\mu)^{r-\ell} \sum_{k=0}^{\ell} \left\{ {\ell \atop k} \right\} \Phi^{r-\ell} G_k(x),$$

where $G_0(x)$ is interpreted as $T(x)$. Plug in the representation (8) to obtain

$$K_r(x) = \sum_{\ell=0}^{r} \binom{r}{\ell} (-\mu)^{r-\ell} \sum_{k=0}^{\ell} \left\{ {\ell \atop k} \right\} \Bigg(- \Phi^{r-\ell} A_k(x) \\ + k! \sum_{\mathbf{m} \vdash k} (-1)^{|\mathbf{m}|} \sum_{h=0}^{r-\ell} \binom{r-\ell}{h} \Phi^{r-\ell-h+|\mathbf{m}|} T(x) \Phi^h \bigg(\prod_{j=1}^{k} \frac{A_j(x)^{m_j}}{m_j! j!^{m_j}} \bigg) \Bigg).$$

The term $\Phi^{r-\ell} A_k(x)$ is irrelevant and only yields an exponentially small error term in view of the aforementioned properties of A_k. We now collect the coefficient of $\Phi^s T(x)$ for fixed s, which is

$$(-1)^s \sum_{h=0}^{r} \sum_{\ell=0}^{r-h} (-1)^h \binom{r}{h} \binom{r-h}{\ell} \mu^{r-\ell} \sum_{k=0}^{\ell} k! \left\{ {\ell \atop k} \right\} \sum_{\substack{\mathbf{m} \vdash k \\ |\mathbf{m}| = s+\ell+h-r}} \Phi^h \bigg(\prod_{j=1}^{k} \frac{A_j(x)^{m_j}}{m_j! j!^{m_j}} \bigg).$$

We first simplify the resulting expression for fixed h; since Φ^h is a linear operator, it can be taken out of the sum. Now use the abbreviation $q = r - s - h$ for simplicity. Notice that $\mathbf{m} \vdash k$ and $|\mathbf{m}| = \ell - q$ is only possible if $\ell - q \leq k \leq \ell$. If we fix $m_2, m_3, \ldots$, then

$$m_1 = \ell - q - \sum_{j>1} m_j \qquad \text{and} \qquad k = \ell - q + \sum_{j>1} (j-1) m_j,$$

hence $0 \leq \sum_{j>1} (j-1) m_j = k + q - \ell \leq q$. Now set $p = \ell - k$ and rewrite the sum above to obtain

$$(-1)^{s+h} \binom{r}{h} \sum_{p=0}^{q} \sum_{\substack{m_2, m_3, \ldots \\ m_2 + 2m_3 + \cdots = q-p}} \prod_{j=2}^{q} \frac{A_j(x)^{m_j}}{m_j! j!^{m_j}} \sum_{\ell=0}^{r-h} \binom{r-h}{\ell} \mu^{r-\ell} (\ell-p)! \left\{ {\ell \atop \ell-p} \right\} \frac{A_1(x)^{\ell-q-\sum_{j>1} m_j}}{(\ell - q - \sum_{j>1} m_j)!}$$

for any fixed h. Finally observe that

$$\frac{(\ell-p)!}{(\ell - q - \sum_{j>1} m_j)!}$$

is a polynomial in ℓ of degree

$$q - p + \sum_{j>1} m_j \leq q - p + \sum_{j>1} (j-1) m_j = 2(q-p), \tag{9}$$

and the Stirling number $\left\{ {\ell \atop \ell-p} \right\}$ can be expressed as a polynomial of degree $2p$. But any sum of the form

$$\sum_{\ell=0}^{r-h} \binom{r-h}{\ell} a^{r-\ell} P(\ell) b^{\ell},$$

where $P(\ell)$ is a polynomial of degree d, is a linear combination of the derivatives (with respect to a) up to order d of $(a+b)^{r-h}$, hence it has a factor $(a+b)^{r-h-d}$. So the part corresponding to a fixed h has a factor $(A_1(x)+\mu)^{r-h-2q} = (A_1(x)+\mu)^{2s+h-r}$. The operator Φ^h only reduces this to a factor of at least $(A_1(x)+\mu)^{2s-r}$. We conclude that the coefficient of $\Phi^s T(x)$ in $K_r(x)$, which we were interested in, is a polynomial in $A_1(x), A_2(x), \ldots$ and their derivatives with a factor $(A_1(x)+\mu)^{2s-r}$ if $2s \geq r$.

What does this mean for the dominant singularity $1/e$ of $K_r(x)$? $(A_1(x)+\mu)^{2s-r}$ has a zero of order $2s-r$ at $1/e$, while $\Phi^s T(x)$ has a singularity of the form $(1-ex)^{1/2-s}$. Hence the order of the singularity of any term with $s \geq (r+1)/2$ is $1/2 - s + 2s - r = 1/2 + s - r \geq 1 - r/2$, which means that the contribution to the rth centred moment that we are interested in is only $O(n^{(r-1)/2})$. Terms involving $\Phi^s T(x)$ with $s \leq (r-1)/2$ also contribute only $O(n^{(r-1)/2})$. It follows that the centred mean is $O(n^{(r-1)/2})$ if r is odd. If r is even, then the main contribution comes from the term $s = r/2$ if r is even, with an error term $O(n^{r/2-1})$ from the rest.

Now we have to revisit our argument that gave us the factor $(A_1(x)+\mu)^{2s-r}$ to determine the actual contribution of the term $s = r/2$. We get an even higher power unless equality holds in (9), which only happens if $m_3 = m_4 = \ldots = 0$ and m_2 is the only nonzero component (apart from m_1) in our partition $\mathbf{m}$. Hence the main term in the asymptotic behaviour comes from partitions of the form $(m_1, m_2, 0, 0, \ldots)$, so that $k = m_1 + 2m_2$ and $s + \ell + h - r = \ell + h - r/2 = m_1 + m_2$. Such partitions only exist for $h \leq r/2$. We obtain

$$(-1)^{r/2}\sum_{h=0}^{r/2}(-1)^h\binom{r}{h}\Phi^h\Bigg(\sum_{m_2=0}^{r/2}\sum_{m_1=0}^{r/2-m_2}\binom{r-h}{m_1+m_2+r/2-h}(m_1+2m_2)!\left\{{m_1+m_2+r/2-h \atop m_1+2m_2}\right\} \mu^{r/2-m_1-m_2+h}\frac{A_1(x)^{m_1}A_2(x)^{m_2}}{m_1!m_2!2^{m_2}}\Bigg),$$

whose value at the singularity $1/e$ needs to be evaluated. Now we first simplify the inner double sum, making use of the fact that $\sum_{n\geq 0}\sum_{k\geq 0}\left\{{n \atop k}\right\}\frac{k!u^kv^n}{n!} = 1/(1-u(e^v-1))$:

$$\begin{aligned}
&\sum_{m_2=0}^{r/2}\sum_{m_1=0}^{r/2-m_2}\frac{(r-h)!A_1(x)^{m_1}A_2(x)^{m_2}\mu^{r/2-m_1-m_2+h}}{(r/2-m_1-m_2)!m_1!m_2!2^{m_2}}[u^{m_1+2m_2}v^{r/2+m_1+m_2-h}]\frac{1}{1-u(e^v-1)}\\
&= [u^0v^{-h}]\frac{(r-h)!\mu^h}{(r/2)!(1-u(e^v-1))}\sum_{m_2=0}^{r/2}\binom{r/2}{m_2}\left(\frac{\mu}{v}\right)^{r/2-m_2}\left(\frac{A_2(x)}{2u^2v^2}\right)^{m_2}\sum_{m_1=0}^{r/2-m_2}\binom{r/2-m_2}{m_1}\left(\frac{A_1(x)}{\mu uv}\right)^{m_1}\\
&= [u^0v^{-h}]\frac{(r-h)!\mu^h}{(r/2)!(1-u(e^v-1))}\sum_{m_2=0}^{r/2}\binom{r/2}{m_2}\left(\frac{\mu}{v}+\frac{A_1(x)}{uv^2}\right)^{r/2-m_2}\left(\frac{A_2(x)}{2u^2v^2}\right)^{m_2}\\
&= [u^0v^{-h}]\frac{(r-h)!\mu^h}{(r/2)!(1-u(e^v-1))}\left(\frac{\mu}{v}+\frac{A_1(x)}{uv^2}+\frac{A_2(x)}{2u^2v^2}\right)^{r/2}\\
&= \frac{(r-h)!\mu^h}{2^{r/2}(r/2)!}[u^rv^{r-h}]\frac{(A_2(x)+2uA_1(x)+2\mu u^2v)^{r/2}}{1-u(e^v-1)}\\
&= -\frac{(r-h)!\mu^h}{2^{r/2}(r/2)!}[v^{r-h}]\operatorname*{Res}_{u=1/(e^v-1)}\frac{(A_2(x)+2uA_1(x)+2\mu u^2v)^{r/2}}{u^{r+1}(1-u(e^v-1))}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(r-h)!\mu^h}{2^{r/2}(r/2)!}[v^{r-h}](e^v-1)^r\left(A_2(x)+\frac{2A_1(x)}{e^v-1}+\frac{2\mu v}{(e^v-1)^2}\right)^{r/2}\\
&= \frac{(r-h)!\mu^h}{2^{r/2}(r/2)!}[v^{r-h}]\left(A_2(x)(e^v-1)^2+2A_1(x)(e^v-1)+2\mu v\right)^{r/2}.
\end{aligned}
$$

Now note that

$$
\left(A_2(x)(e^v-1)^2+2A_1(x)(e^v-1)+2\mu v\right)^{r/2}=\sum_{i=0}^{r/2}\binom{r/2}{i}2^i(A_1(x)+\mu)^iv^{r-i}\left(A_1(x)+A_2(x)+O(v)\right)^{r/2-i}.
$$

For $i<h$, the coefficient of v^{r-h} is 0, for $i>h$ it has a factor $(A_1(x)+\mu)^{h+1}$, so even after application of Φ^h a factor $A_1(x)+\mu$ remains, which is 0 at $x=1/e$. So only the term $i=h$ contributes, and we are finally left with

$$
\begin{aligned}
&(-1)^{r/2}\sum_{h=0}^{r/2}(-1)^h\binom{r}{h}\frac{(r-h)!(2\mu)^h}{2^{r/2}(r/2)!}\binom{r/2}{h}\Phi^h\left((A_1(x)+\mu)^h(A_1(x)+A_2(x))^{r/2-h}\right)\\
&\qquad=(-1)^{r/2}(r-1)!!\sum_{h=0}^{r/2}\binom{r/2}{h}\frac{(-2\mu)^h}{h!}\Phi^h\left((A_1(x)+\mu)^h(A_1(x)+A_2(x))^{r/2-h}\right).
\end{aligned}
$$

The value at $x=1/e$ is

$$
\begin{aligned}
&(-1)^{r/2}(r-1)!!\sum_{h=0}^{r/2}\binom{r/2}{h}\frac{(-2\mu)^h}{h!}\cdot h!(\Phi A_1(1/e))^h(A_1(1/e)+A_2(1/e))^{r/2-h}\\
&\qquad=(r-1)!!\left(-A_1(1/e)-A_2(1/e)+2\mu\Phi A_1(1/e)\right)^{r/2}.
\end{aligned}
$$

Hence with

$$
\sigma^2=-A_1(1/e)-A_2(1/e)+2\mu\Phi A_1(1/e)=\sum_T t(T)(2F(T)-t(T))\frac{e^{-|T|}}{|T|!}-2\mu\sum_T t(T)\frac{e^{-|T|}}{(|T|-1)!},
$$

it follows that

$$
\begin{aligned}
K_r(x)&=(r-1)!!\sigma^r\Phi^{r/2}T(x)+O((1-ex)^{3/2-r/2})\\
&=(r-1)!!\sigma^r2^{1/2-r/2}(r-3)!!(1-ex)^{1/2-r/2}+O((1-ex)^{3/2-r/2})
\end{aligned}
$$

around the singularity, hence by singularity analysis

$$
\begin{aligned}
\mathbb{E}((F(T_n)-\mu n)^r)&=\frac{[x^n]K_r(x)}{[x^n]T(x)}=\frac{(r-1)!!\sigma^r2^{1/2-r/2}(r-3)!!n^{r/2-3/2}e^n}{\Gamma(r/2-1/2)e^n/\sqrt{2\pi n^3}}+O(n^{r/2-1})\\
&=(r-1)!!\sigma^rn^{r/2}+O(n^{r/2-1})
\end{aligned}
$$

for even r, while

$$\mathbb{E}((F(T_n) - \mu n)^r) = O(n^{(r-1)/2})$$

for odd r. It follows that the moments of the renormalised random variable $(F(T_n) - \mu n)/(\sigma\sqrt{n})$ converge to those of a standard normal distribution, and since this distribution is characterised by its moments, we infer the central limit theorem, thus completing our proof. □

3 Simple examples

The simplest example to which our theorem applies is probably the number of leaves, for which we easily obtain $\mu = 1/e$ and $\sigma^2 = (e-2)/e^2$, since $t(T) \neq 0$ only when $|T| = 1$. Hence we rederived the well-known result that the distribution of the number of leaves in a random labelled tree of order n is asymptotically normal with mean $\sim n/e$ and variance $\sim (e-2)n/e^2$ (see [5, Example IX.25]).

More generally, one can consider the number of occurrences of a fixed rooted tree H of order m. We take H to be unlabelled (since this is somewhat more natural in our setting) and let L be the number of its labellings (equal to $m!/|\operatorname{Aut}(H)|$). The tree functional $F_H(T)$ that counts the number of occurrences of H satisfies

$$F_H(T) = \sum_{i=1}^{k} F_H(T_i) + t_H(T)$$

if $T_1, T_2, \ldots, T_k$ are the branches of T, where

$$t_H(T) = \begin{cases} 1 & \text{if } T \text{ is isomorphic to } H \text{ as a rooted tree,} \\ 0 & \text{otherwise.} \end{cases}$$

In this more general setting, the mean and variance are asymptotically

$$\frac{L}{m!e^m} \cdot n \qquad \text{and} \qquad \left(\frac{L}{m!e^m} - \frac{2L^2}{e^{2m}m!(m-1)!} \right) \cdot n$$

respectively. Even more generally, one can count the occurrences of any finite set of trees or a sufficiently "thin" subset of trees. For instance, suppose we are interested in the number of vertices all of whose children are leaves. This amounts to counting the number of stars, rooted at their centres, occurring as rooted subtrees. Hence the distribution is Gaussian with mean $\sim \mu n$ and variance $\sim \sigma^2 n$, where

$$\mu = \sum_{n=2}^{\infty} \frac{ne^{-n}}{n!} = e^{-1}\left(e^{1/e} - 1\right) \approx 0.163584$$

and

$$\sigma^2 = \sum_{n=2}^{\infty} \frac{ne^{-n}}{n!} - 2\mu \sum_{n=2}^{\infty} \frac{n^2 e^{-n}}{(n-1)!} = e^{-3}\left(e^{1/e} - 1\right)\left(e^2 + 2e - 2(e+1)e^{1/e}\right) \approx 0.0460984.$$

4 Number and size of subtrees

This final section is concerned with two parameters defined in terms of subtrees of trees, which were the original motivation for the investigations in this paper. The calculations are somewhat more involved in these cases. We first consider the number $s(T)$ of subtrees (connected induced subgraphs). As mentioned in the introduction, it is technically convenient to deal with the number of subtrees containing the root only. This parameter satisfies, as it was also mentioned in the introduction, the recursion

$$s_1(T) = \prod_{i=1}^{k}(1 + s_1(T_i)).$$

We first need the following lemma:

Lemma 2 *The probability that a random rooted labelled tree of order n has less than $2^{n/4}$ subtrees containing the root is $O(\alpha^n)$, where $\alpha = 2\sqrt{e-1}/e < 1$.*

Proof: It is well known that the number of vertices of a given degree in a random tree follows a central limit theorem [2], and that the tails are exponentially small. For completeness, we give the precise estimate needed for our lemma. The bivariate generating function $F(x,u)$ for rooted labelled trees in which u marks the number of vertices of outdegree 1 satisfies $F(x,u) = x\exp(F(x,u)) + x(u-1)F(x,u)$, and thus $F(x,u) = T(x/(1+x-ux))$, which has its square root singularity at $1/(e+u-1)$. Now singularity analysis shows that

$$\frac{[x^n]F(x,u)}{n^{n-1}/n!} = O\left((1 + (u-1)/e)^n\right)$$

for any u. If X_n denotes the random variable "number of vertices of outdegree 1", then by the Markov inequality

$$\mathbb{P}(X_n \geq n/2) = \mathbb{P}\left(u^{X_n} \geq u^{n/2}\right) \leq \frac{\mathbb{E}(u^{X_n})}{u^{n/2}} = \frac{[x^n]F(x,u)}{u^{n/2}n^{n-1}/n!}.$$

Now we set $u = e - 1$ to obtain that $\mathbb{P}(X_n \geq n/2) = O(\alpha^n)$, with α as in the statement of the lemma. Now consider a rooted tree with at most $n/2$ vertices of outdegree 1. If we contract them, we obtain a rooted tree T' of order at least $n/2$ without vertices of outdegree 1, and every subtree of T' also corresponds to a unique subtree in T. It is a straightforward induction that a tree T' without vertices of outdegree 1 has at least $2^{(|T'|+1)/2} - 1$ subtrees containing the root: this is trivial for $|T'| = 1$, and if the branches of T' are $T_1, \ldots, T_k$ ($k \geq 2$), then by the induction hypothesis

$$s_1(T) = \prod_{i=1}^{k}(1 + s_1(T_i)) \geq \prod_{i=1}^{k} 2^{(|T_i|+1)/2} = 2^{(|T'|+(k-1))/2} \geq 2^{(|T'|+1)/2},$$

completing the proof of the lemma. □

Now we see that our condition (3) is satisfied for the toll function $t(T) = \log(1 + 1/s_1(T))$ (see (2)), since it is clearly bounded (between 0 and $\log 2$), and since $t(T) = O(2^{-n/4})$ for all but $O(\alpha^n n^{n-1})$ rooted labelled trees. Hence the distribution of $\log(1 + s_1(T))$ (and thus clearly also $\log s_1(T)$) is asymptotically Gaussian. It remains to show that the difference between $\log s(T)$ and $\log s_1(T)$ is actually small. For this purpose, let $s_0(T) = s(T) - s_1(T)$ denote the number of subtrees of T not containing the root. Then we can determine $s_0(T)$ recursively from the branches as follows:

$$s_0(T) = \sum_{i=1}^{k} (s_0(T_i) + s_1(T_i)).$$

Now we prove by induction that $s_0(T) \leq |T| s_1(T)$. This is clearly true for $|T| = 1$, and for the induction step note that

$$\frac{s_0(T)}{s_1(T)} = \frac{\sum_{i=1}^{k} (s_0(T_i) + s_1(T_i))}{\prod_{i=1}^{k} (1 + s_1(T_i))} \leq \frac{\sum_{i=1}^{k} (1 + |T_i|) s_1(T_i)}{\sum_{i=1}^{k} s_1(T_i)} \leq 1 + \max_i |T_i| \leq |T|,$$

completing the induction. Hence we have $\log s(T) = \log s_1(T) + O(\log |T|)$, and a central limit theorem for $\log s(T)$ follows, since the renormalised random variables only differ by $o(1)$. Summing up, we have the following result:

Theorem 3 *The logarithm $\log s(T_n)$ of the number of subtrees of a random labelled tree T_n of order n is asymptotically normally distributed, with mean and variance asymptotically equal to μn and $\sigma^2 n$ respectively, where the numerical values of μ and σ^2 are $\mu \approx 0.35$ (slightly less than $1/e$) and $\sigma^2 \approx 0.04$.*

Remark 2 *Obtaining good numerical approximations (here and later) seems to be a difficult task, as the sums in Theorem 1 all contain the toll function, which has to be evaluated for all "small" trees to get an approximation. This is only feasible for up to about 15-20 vertices.*

Finally, we take a look at the parameter "average size of a subtree" in a tree T, which was first treated in the graph-theoretical literature by Jamison [7]. Meir and Moon [9] remark that the average size of all subtrees in all labelled trees of order n is asymptotically $(1 - 1/e)n$, and Baron and Drmota [1] and Moon [12] give formulas for the total number of subtrees of a given order in all labelled trees of order n. However, the underlying probabilistic model in the aforementioned result of Meir and Moon is perhaps not quite optimal, as it assumes that all subtrees in all trees are given the same probability measure, regardless of the number of subtrees in the underlying tree. It seems more intuitive to first pick a random tree and then study the average size, averaged only over the subtrees of the tree that was chosen. Here we briefly show how to apply Theorem 1 to prove that the average size of a subtree in a random labelled tree T satisfies a central limit theorem. It is again advantageous to consider subtrees containing the root of T first. Let $t_1(T)$ be the total number of vertices of all such subtrees. Then

$$t_1(T) = s_1(T) + \prod_{i=1}^{k} (1 + s_1(T_i)) \sum_{j=1}^{k} \frac{t_1(T_j)}{1 + s_1(T_j)}.$$

The term $s_1(T)$ takes account of the root that occurs in all $s_1(T)$ subtrees, and

$$\prod_{i=1}^{k} (1 + s_1(T_i)) \cdot \frac{t_1(T_j)}{1 + s_1(T_j)} = t_1(T_j) \prod_{i \neq j} (1 + s_1(T_i))$$

is the total number of vertices in T_j in all subtrees that contain the root. One obtains easily that

$$\frac{t_1(T)}{1+s_1(T)} = 1 + \sum_{j=1}^{k} \frac{t_1(T_j)}{1+s_1(T_j)} - \frac{t_1(T)}{s_1(T)(1+s_1(T))}.$$

Hence $t_1(T)/(1+s_1(T))$ is an additive tree functional with toll function $t(T) = 1 - t_1(T)/(s_1(T)(1+s_1(T)))$. Since $t_1(T)/s_1(T)$ is at most $|T|$, while $s_1(T)$ grows exponentially for almost all trees (as proven in Lemma 2), the toll function (minus 1, compare Remark 1) satisfies our conditions (it is also clearly bounded, since $s_1(T) \geq |T|$ holds trivially). Hence Theorem 1 applies to the average size $t_1(T)/s_1(T)$ of subtrees containing the root (the difference between $t_1(T)/s_1(T)$ and $t_1(T)/(1+s_1(T))$ clearly does not matter), with $\mu \approx 0.65$ (note that this is slightly more than $1 - 1/e$) and $\sigma^2 \approx 0.04$.

To show that this remains true if all subtrees are counted, not just those that contain the root, one can e.g. assume that the vertex labelled 1 is the centroid, and that its neighbour in the largest branch has label 2 (a labelled tree almost surely has a unique centroid and largest centroid branch). Then both components of the tree that remain when the edge between vertices 1 and 2 is removed almost surely have linear size (much more is known on centroid branches, see [11]). In view of Lemma 2 and the recursions for s_0 and s_1, it follows that (excluding an exceptional set of trees whose proportion goes exponentially to 0) the proportion of subtrees not containing both vertices 1 and 2 is exponentially small, and one can apply the result on subtrees containing the root to the two aforementioned components.

5 Final remarks

It is quite likely that the approach shown in this paper can be carried over to unlabelled trees (the arising functional equations typically only differ from those for labelled trees in some asymptotically irrelevant terms) and to the wider class of simply generated families of trees (e.g., plane trees, binary trees). The main difficulty lies in the fact that the application of Faà di Bruno's formula in the proof of Theorem 1 becomes more complicated for these families, which affects the asymptotic analysis that follows. It is probably also possible to relax the conditions on the toll function a little further, so that the decay no longer has to be exponential.

References

[1] G. Baron and M. Drmota. Distribution properties of induced subgraphs of trees. *Ars Combin.*, 35:193–213, 1993.

[2] M. Drmota. *Random trees*. SpringerWienNewYork, Vienna, 2009.

[3] J. A. Fill, P. Flajolet, and N. Kapur. Singularity analysis, Hadamard products, and tree recurrences. *J. Comput. Appl. Math.*, 174(2):271–313, 2005.

[4] J. A. Fill and N. Kapur. Limiting distributions for additive functionals on Catalan trees. *Theoret. Comput. Sci.*, 326(1-3):69–102, 2004.

[5] P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[6] H.-K. Hwang and R. Neininger. Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM J. Comput.*, 31(6):1687–1722 (electronic), 2002.

[7] R. E. Jamison. On the average number of nodes in a subtree of a tree. *J. Combin. Theory Ser. B*, 35(3):207–223, 1983.

[8] S. Janson. The Wiener index of simply generated random trees. *Random Structures Algorithms*, 22(4):337–358, 2003.

[9] A. Meir and J. W. Moon. On subtrees of certain families of rooted trees. *Ars Combin.*, 16(B):305–318, 1983.

[10] A. Meir and J. W. Moon. On the log-product of the subtree-sizes of random trees. *Random Structures Algorithms*, 12(2):197–212, 1998.

[11] A. Meir and J. W. Moon. On centroid branches of trees from certain families. *Discrete Math.*, 250(1-3):153–170, 2002.

[12] J. W. Moon. On the number of induced subgraphs of trees. *Discrete Math.*, 167/168:487–496, 1997. 15th British Combinatorial Conference (Stirling, 1995).

[13] L. Takács. Conditional limit theorems for branching processes. *J. Appl. Math. Stochastic Anal.*, 4(4):263–292, 1991.

[14] L. Takács. On the total heights of random rooted trees. *J. Appl. Probab.*, 29(3):543–556, 1992.