Combinatorics, Probability and Computing (2019), **28**, pp. 618–637
doi:10.1017/S0963548318000585

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

A central limit theorem for additive functionals of increasing trees[†]

Dimbinaina Ralaivaosaona[§] and Stephan Wagner[*,¶]

Department of Mathematical Sciences, Stellenbosch University, 7602 Stellenbosch, South Africa
[*]Corresponding author. Email: swagner@sun.ac.za

(Received 14 September 2016; revised 5 November 2018; first published online 8 March 2019)

Abstract

A tree functional is called additive if it satisfies a recursion of the form $F(T)=\sum_{j=1}^{k} F(B_j)+f(T)$, where $B_1,\dots,B_k$ are the branches of the tree T and $f(T)$ is a toll function. We prove a general central limit theorem for additive functionals of d-ary increasing trees under suitable assumptions on the toll function. The same method also applies to generalized plane-oriented increasing trees (GPORTs). One of our main applications is a log-normal law that we prove for the size of the automorphism group of d-ary increasing trees, but other examples (old and new) are covered as well.

2010 MSC Codes: Primary 60C05; Secondary 05C05, 60F05

1. Introduction

In this paper, we are interested in functionals of rooted trees that satisfy an *additive* relation, *i.e.* a recursion of the form

$$F(T)=\sum_{j=1}^{k} F(B_j)+f(T), \tag{1.1}$$

where $B_1,\dots,B_k$ are the branches of the tree T and $f(T)$ is a so-called toll function, which often only depends on specific features of the tree such as the size or the root degree, but can in principle be arbitrary. The trees in our context will be labelled; it is assumed that the toll function only depends on the relative order of the labels, not the labels themselves, so that it is also well-defined if the labels are not necessarily $1,2,\dots,n$. It is consistent with (1.1) to assume that we have the identity $F(\odot)=f(\odot)$ for the tree $T=\odot$ consisting only of a single labelled vertex. Important examples include the following.

- The number of leaves, which corresponds to the toll function $f(T)$ that is equal to 1 if $|T|=1$ and 0 otherwise.

[†] An extended abstract of this paper was presented at the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, Kraków, 4–8 July 2016: see [15].

[§] Supported by the Division for Research Development (DRD) of Stellenbosch University.

[¶] This material is based upon work supported by the National Research Foundation under grant number 96236.

- The number of vertices of outdegree k, in which case one can simply take $f(T)=1$ if the root of T has outdegree k, and 0 otherwise.
- 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 $f(T)=|T|-1$.
- The log-product of the subtree sizes [13], also called the 'shape functional' [6], corresponding to $f(T)=\log|T|$.
- The logarithm of the size of the automorphism group: here, it is not difficult to see that the toll function is $f(T)=\log(R(T))$, where $R(T)$ is the size of the symmetry group of the collection of root branches.

Such functionals also arise frequently in the study of divide-and-conquer algorithms, *e.g.* quicksort [11]. An alternative viewpoint is based on the notion of *fringe subtrees*: a fringe subtree of a tree is a subtree induced by a vertex and all its descendants. If we let $\mathcal{F}(T)$ denote the collection of all fringe subtrees of a tree T, then it is easy to verify that

$$F(T)=\sum_{S\in\mathcal{F}(T)} f(S). \tag{1.2}$$

In particular, the number of occurrences of a specific tree as a fringe subtree is an additive functional (corresponding to the case that the toll function f is an indicator function), and every additive functional can be obtained as a linear combination of such special functionals.

There are several recent papers providing central limit theorems for rather general additive tree functionals [6, 3, 5, 9, 12, 17]. Specifically, Holmgren and Janson [9] proved such a central limit theorem for binary increasing trees (which are also equivalent to binary search trees) and recursive trees. Both are instances of so-called *increasing trees*: labelled trees with the additional property that the labels increase along any path starting at the root.

Varieties of increasing trees were studied systematically in [1] (see also [4, Section 1.3.3]). The exponential generating function $Y(x)$ associated with a variety of increasing trees satisfies a differential equation of the characteristic shape

$$Y'(x)=\Phi(Y(x)), \qquad Y(0)=0 \tag{1.3}$$

for some function $\Phi(t)$. Varieties of increasing trees for which a uniformly random tree with a given number of vertices can also be generated by a growth process have been of particular interest. There are three such types [14].

- The variety of recursive trees is perhaps the most basic instance: these are simply labelled rooted unordered trees ('unordered' meaning that the order of branches does not matter) with the aforementioned property that the labels increase along paths starting at the root. Uniformly random recursive trees can be obtained by the following growth process: starting from a single vertex (the root, carrying label 1), the vertex labelled n is attached in the nth step to one of the previous vertices, chosen uniformly at random. As mentioned earlier, the order of children attached to a vertex does not matter. To obtain a canonical representation, one can *e.g.* always make the newly added vertex the rightmost child.
- Plane-oriented recursive trees (PORTs) differ from recursive trees in only one aspect: trees are regarded as embedded in the plane, the order of branches is taken into account. The growth process to generate uniformly random PORTs follows a 'preferential attachment' rule: it is essentially the same as for recursive trees, but the probability that the vertex labelled n is attached to a specific vertex v is proportional to 1 plus the current outdegree of v. This reflects the fact that a new vertex can be attached to a vertex v of outdegree k in $k+1$ different places (to the left of a current child or at the end).

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

Generalized plane oriented recursive trees (GPORTs) are obtained by introducing an additional parameter: for some positive real number α, we let the probability that the vertex labelled n is attached to a specific vertex v be proportional to α plus the current outdegree of v (so PORTs correspond to the special case $\alpha = 1$).

- Finally, we have the variety of d-ary increasing trees, which will be the focus of this paper: here, every vertex has d possible places to which a child can be attached (for example, in the binary case, there are left and right children). In the construction of uniform d-ary increasing trees by a growth process, we simply attach the vertex labelled n to one of the $(d-1)(n-1)+1$ places available in total, once again selected uniformly at random. Therefore, the probability that the new vertex is attached to an existing vertex v is proportional to d minus the current outdegree of v (in particular, if v already has d children, no further vertices can be attached to it).

All these examples can be seen as weighted PORTs: for every non-negative integer j, let w_j be a weight associated with outdegree j, and let $N_j(T)$ be the number of vertices whose outdegree is j. The weight $w(T)$ of a tree T is now defined as follows:

$$w(T) = \prod_{j\geqslant 0} c_j^{N_j(T)}.$$

The (weighted) generating function $Y(x)$ associated with such a choice of weight is easily seen to satisfy the differential equation (1.3) with $\Phi(t) = \sum_{j=0}^{\infty} c_j t^j$. Based on the weights, one can select random PORTs of a given size, the probability of a certain tree being proportional to the weight. To obtain uniformly random PORTs, the weights c_j can simply be chosen to be 1 for all values of j. For the other classes, the weights need to be chosen as follows.

- For recursive trees, take
$$c_j = \frac{1}{j!}$$
to factor out the different ways of ordering the branches. In this example, the function Φ is the exponential function. It is easy to see that the generating function is $Y(x) = -\log(1-x)$, which is consistent with the observation that there are $(n-1)!$ recursive trees with n vertices.
- For generalized PORTs, take
$$c_j = \binom{\alpha+j-1}{j} = \frac{\alpha(\alpha+1)\cdots(\alpha+j-1)}{j!}.$$
Note that $c_j = 1$ for $\alpha = 1$. The factors $\alpha, \alpha+1, \dots, \alpha+j-1$ reflect the fact that the ith child of a vertex is attached with probability proportional to $\alpha+i-1$ by definition. As in the previous case, the denominator factors out the different ways of ordering the branches in a PORT. The function Φ in (1.3) is now given by $\Phi(t) = (1-t)^{-\alpha}$. It follows that $Y(x) = 1 - (1-(\alpha+1)x)^{1/(1+\alpha)}$, and the total weight of all trees with n vertices is $\prod_{j=1}^{n-1}((\alpha+1)j-1)$. In particular, the generating function for ordinary PORTs is $Y(x) = 1-\sqrt{1-2x}$, and the number of PORTs with n vertices is $(2n-3)!! = (2n-3)\cdot(2n-5)\cdots 3\cdot 1$.
- For d-ary increasing trees, one sets
$$c_j = \binom{d}{j}$$
to take the d possible positions of a child into account. Here, $\Phi(t) = (1+t)^d$ and $Y(x) = (1-(d-1)x)^{-1/(d-1)} - 1$. The total number of d-ary increasing trees with n vertices is $\prod_{j=1}^{n-1}((d-1)j+1)$.

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

In the following, we state and prove a central limit theorem for additive tree functionals of uniformly random d-ary increasing trees under certain technical conditions on the toll function. As mentioned earlier, binary increasing trees (as well as recursive trees) have already been covered in [9]. The approach in [9] is based on representations of binary increasing trees and recursive trees that are not available for other classes of increasing trees. On the other hand, the generating function method of [17] made use of the fact that certain (separable and Riccati-type) differential equations can be solved explicitly. Since the differential equations arising for d-ary increasing trees and GPORTs are no longer of a form that is explicitly solvable, we use a different approach based on moments, as in a paper of Fuchs [8] on the number of fringe subtrees of given size (which is also an additive functional). Although we only discuss the case of d-ary increasing trees in detail, our method also applies to GPORTs, for which we only state the corresponding result in the following section.

In their very recent paper [10], Holmgren, Janson and Šileikis approached the same question from a different angle: using Pólya urns, they proved that the number of fringe subtrees isomorphic to a given rooted tree is asymptotically normally distributed. They also proved joint normality for different fringe subtrees. In view of the representation (1.2), this implies asymptotic normality of all additive functionals whose toll functions have finite support. Holmgren, Janson and Šileikis also obtained the same result for m-ary search trees ($m < 26$).

2. The general central limit theorem

Let us now formulate our main result. In the following, d is fixed, and $\mathcal{T}_n$ always denotes a random d-ary increasing tree of order n (except for Theorem 2.2). We assume that the toll function $f(T)$ satisfies the following conditions:

(C1) $f(T)$ is bounded,

(C2) $\displaystyle\sum_{k\geqslant 1} \frac{\mathbb{E}|f(\mathcal{T}_k)|}{k} < \infty$ and $\mathbb{E}|f(\mathcal{T}_n)| \to 0$ as $n \to \infty$.

Under these assumptions, our central limit theorem for additive functionals reads as follows.

Theorem 2.1. *Let $\mathcal{T}_n$ be a uniformly random d-ary increasing tree with n vertices. If the toll function $f(T)$ satisfies* (C1) *and* (C2)*, then there exist constants μ and σ such that the mean and variance of $F(\mathcal{T}_n)$ are asymptotically*

$$\mathbb{E}(F(\mathcal{T}_n)) = \mu n + \frac{\mu}{d-1} + o(1), \qquad \mathrm{Var}(F(\mathcal{T}_n)) = \sigma^2 n + o(n).$$

The constants μ and σ can be represented as

$$\mu = (d-1)\sum_T f(T) \prod_{j=1}^{|T|} \frac{1}{(d-1)j+d} = d(d-1)\sum_{n=1}^{\infty} \frac{\mathbb{E}(f(\mathcal{T}_n))}{((d-1)n+1)((d-1)n+d)} \tag{2.1}$$

and

$$\sigma^2 = -\frac{\mu^2}{d-1} - (d-1)\sum_T \frac{f(T)^2 - 2f(T)(F(T)-\mu|T|)}{\prod_{j=1}^{|T|}((d-1)j+d)}$$
$$+ d\sum_{T_1}\sum_{T_2} \frac{(d-1)^{1-|T_1|-|T_2|} f(T_1)f(T_2)}{(|T_1|-1)!(|T_2|-1)!} \int_0^1 \phi_{|T_1|}(x)\phi_{|T_2|}(x)\,dx,$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

where

$$\phi_k(x) = (1-x)^{-1} \int_x^1 (1-w)^{d/(d-1)} w^{k-1}\, dw.$$

The sums are taken over all d-ary increasing trees. If $\sigma \neq 0$, then the renormalized random variable $(F(\mathcal{T}_n) - \mu n)/\sqrt{\sigma^2 n}$ converges weakly to a standard normal distribution.

Remark. (1) We remark that the result remains true if conditions (C1) and (C2) hold for a shifted version $f(T) + c$ (c any constant) of the toll function rather than the toll function itself, since this changes $F(T)$ only by the deterministic quantity $c|T|$.

(2) As the proof shows, one can also replace condition (C1) with a slightly weaker condition, namely that $F(T) = O(|T|)$.

(3) The only degenerate example ($\sigma = 0$) we know of is the case that $f(T)$ and $F(T)$ are identically zero, and we suspect that there are no other degenerate examples that satisfy our technical conditions (compare the analogous situation for Galton–Watson trees [12], where this question is also open). Of course, since every conceivable tree functional becomes additive for a suitable choice of toll function, it is easy to construct degenerate examples violating (C1) or (C2).

As mentioned earlier, the method used in proving Theorem 2.1 also applies to GPORTs. Without going into detail, let us just state the corresponding theorem.

Theorem 2.2. *Let $\mathcal{T}_n$ be a random GPORT (with fixed parameter α) with n vertices. If the toll function $f(T)$ satisfies* (C1) *and* (C2), *then there exist constants μ and σ such that the mean and variance of $F(\mathcal{T}_n)$ are asymptotically*

$$\mathbb{E}(F(\mathcal{T}_n)) = \mu n - \frac{\mu}{\alpha+1} + o(1), \qquad \mathrm{Var}(F(\mathcal{T}_n)) = \sigma^2 n + o(n).$$

The constants μ and σ can be represented as

$$\mu = (\alpha+1) \sum_T w(T) f(T) \prod_{j=1}^{|T|} \frac{1}{(\alpha+1)j + \alpha}$$

and

$$\sigma^2 = \frac{\mu^2}{\alpha+1} - (\alpha+1) \sum_T w(T) \frac{f(T)^2 - 2f(T)(F(T) - \mu|T|)}{\prod_{j=1}^{|T|} ((\alpha+1)j + \alpha)}$$
$$+ \alpha \sum_{T_1} \sum_{T_2} w(T_1) w(T_2) \frac{(\alpha+1)^{1-|T_1|-|T_2|} f(T_1) f(T_2)}{(|T_1|-1)!(|T_2|-1)!} \int_0^1 \varphi_{|T_1|}(x) \varphi_{|T_2|}(x)\, dx,$$

where

$$\varphi_k(x) = (1-x)^{-1} \int_x^1 (1-w)^{\alpha/(\alpha+1)} w^{k-1}\, dw.$$

The sums are taken over all PORTs. If $\sigma \neq 0$, then the renormalized random variable $(F(\mathcal{T}_n) - \mu n)/\sqrt{\sigma^2 n}$ converges weakly to a standard normal distribution.

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

3. Preliminaries

Recall that the exponential generating function $Y(x)$ of d-ary increasing trees satisfies the differential equation

$$Y'(x) = \Phi(Y(x)), \qquad Y(0) = 0, \tag{3.1}$$

where $\Phi(t) = (1+t)^d$. The explicit solution is given by $Y(x) = (1-(d-1)x)^{-1/(d-1)} - 1$, and the total number of d-ary increasing trees with n vertices is

$$Y_n = n! \cdot [x^n]Y(x) = \prod_{j=1}^{n-1} ((d-1)j+1).$$

Let us first define a multivariate generating function that also incorporates the tree functional F and its toll function f. Specifically, we set

$$Y(x, a, b) = \sum_T \frac{x^{|T|}}{|T|!} e^{aF(T)-bf(T)}.$$

In particular, we have

$$\frac{\partial}{\partial x} Y(x, a, a) = \sum_T \frac{x^{|T|-1}}{(|T|-1)!} e^{a(F(T)-f(T))}.$$

The term $(\partial/\partial x)Y(x, a, a)$ corresponds to a bivariate generating function for the ordered forests obtained from cutting off the root of d-ary trees. Such a forest is of the form $(T_1, T_2, \ldots, T_d)$, where each T_i is either an empty tree or a d-ary increasing tree (according to its own set of labels). Moreover, in view of the recursion (1.1), if $T_1, T_2, \ldots, T_d$ are the root branches of T, then

$$e^{a(F(T)-f(T))} = e^{a\sum_i F(T_i)}.$$

Hence, (3.1) becomes

$$\frac{\partial}{\partial x} Y(x, a, a) = \Phi(Y(x, a, 0)), \qquad Y(0, a, b) = 0.$$

Let us now set

$$Z(x, a, b) = 1 + Y(xe^{-a\mu}, a, b) = 1 + \sum_T \frac{x^{|T|}}{|T|!} e^{aF(T)-a\mu|T|-bf(T)},$$

where μ will be determined later, so that

$$\frac{\partial}{\partial x} Z(x, a, a) = e^{-a\mu}\Phi(Y(xe^{-a\mu}, a, 0)) = e^{-a\mu}\Phi(Z(x, a, 0) - 1) = e^{-a\mu} Z(x, a, 0)^d.$$

Note that

$$M_n(a) = \frac{[x^n]Z(x, a, 0)}{[x^n]Z(x, 0, 0)} = \frac{n![x^n]Z(x, a, 0)}{Y_n} \tag{3.2}$$

is the moment generating function for the random variable $F(\mathcal{T}_n) - \mu|\mathcal{T}_n| = F(\mathcal{T}_n) - \mu n$ when a random d-ary increasing tree $\mathcal{T}_n$ with n vertices is generated. Its derivatives with respect to a, evaluated at 0, yield the moments.

Let the rth derivative of Z with respect to a be denoted by $Z^{(r)}(x, a, b)$. Our first goal is to determine a differential equation for the function $Z^{(r)}(x, 0, 0)$. To this end, we need some further notation regarding integer partitions: we represent partitions of a non-negative integer r as sequences $\ell = (\ell_1, \ell_2, \ldots)$, where ℓ_j denotes the multiplicity of j. Thus ℓ is a partition of r if $\sum_j j\ell_j = r$. In this notation, there is only one partition of 0, namely the sequence $(0, 0, \ldots)$.

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

The set of all partitions of r is denoted by $\mathcal{P}(r)$, and we write $|\ell| = \ell_1 + \ell_2 + \cdots$ for the total number of parts in the partition ℓ.

Lemma 3.1. *The function $Z^{(r)}(x,0,0)$ satisfies the differential equation*

$$\frac{\partial}{\partial x}(Z(x,0,0)^{-d} Z^{(r)}(x,0,0)) = -Z(x,0,0)^{-d} H_r(x) + \sum_{s=0}^{r} \binom{r}{s} (-\mu)^{r-s} s! \sum_{\substack{\ell \in \mathcal{P}(s) \\ \ell_r \neq 1}} \frac{d!}{(d-|\ell|)!} \prod_{j \geqslant 1} \frac{1}{\ell_j! j!^{\ell_j}} \left(\frac{Z^{(j)}(x,0,0)}{Z(x,0,0)} \right)^{\ell_j}, \tag{3.3}$$

where

$$H_r(x) = \sum_{s=1}^{r} \binom{r}{s} \sum_{T} \frac{x^{|T|-1}}{(|T|-1)!} (F(T) - \mu|T|)^{r-s} (-f(T))^s.$$

Proof. If we differentiate the identity

$$\frac{\partial}{\partial x} Z(x,a,a) = e^{-a\mu} Z(x,a,0)^d$$

r times with respect to a, we obtain (making use of Faà di Bruno's formula)

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial a}\right)^r Z(x,a,a) = \sum_{s=0}^{r} \binom{r}{s} (-\mu)^{r-s} e^{-a\mu} s! \sum_{\ell \in \mathcal{P}(s)} \frac{d!}{(d-|\ell|)!} Z(x,a,0)^{d-|\ell|} \prod_{j \geqslant 1} \frac{Z^{(j)}(x,a,0)^{\ell_j}}{\ell_j! j!^{\ell_j}}.$$

For $a = 0$, the left side of the equation becomes

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{\partial}{\partial a}\right)^r Z(x,a,a)\bigg|_{a=0} &= \sum_{T} \frac{x^{|T|-1}}{(|T|-1)!} (F(T) - \mu|T| - f(T))^r \\ &= \sum_{s=0}^{r} \binom{r}{s} \sum_{T} \frac{x^{|T|-1}}{(|T|-1)!} (F(T) - \mu|T|)^{r-s} (-f(T))^s. \end{aligned}$$

Separating the term $s = 0$ from the rest, we can write

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial a}\right)^r Z(x,a,a)\bigg|_{a=0} = H_r(x) + \frac{\partial}{\partial x} Z^{(r)}(x,0,0)$$

for every r, where

$$H_r(x) = \sum_{s=1}^{r} \binom{r}{s} \sum_{T} \frac{x^{|T|-1}}{(|T|-1)!} (F(T) - \mu|T|)^{r-s} (-f(T))^s.$$

It follows that

$$H_r(x) + \frac{\partial}{\partial x} Z^{(r)}(x,0,0) = \sum_{s=0}^{r} \binom{r}{s} (-\mu)^{r-s} s! \sum_{\ell \in \mathcal{P}(s)} \frac{d!}{(d-|\ell|)!} Z(x,0,0)^{d-|\ell|} \prod_{j \geqslant 1} \frac{Z^{(j)}(x,0,0)^{\ell_j}}{\ell_j! j!^{\ell_j}}.$$

Dividing by $Z(x,0,0)^d$ gives us

$$\begin{aligned} &Z(x,0,0)^{-d} H_r(x) + Z(x,0,0)^{-d} \frac{\partial}{\partial x} Z^{(r)}(x,0,0) \\ &\quad = \sum_{s=0}^{r} \binom{r}{s} (-\mu)^{r-s} s! \sum_{\ell \in \mathcal{P}(s)} \frac{d!}{(d-|\ell|)!} \prod_{j \geqslant 1} \frac{1}{\ell_j! j!^{\ell_j}} \left(\frac{Z^{(j)}(x,0,0)}{Z(x,0,0)} \right)^{\ell_j}. \end{aligned}$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

One of the terms on the right side of the equation (corresponding to $s=r$ and $\ell_1=\ell_2=\cdots=\ell_{r-1}=0$, $\ell_r=1$) is $dZ^{(r)}(x,0,0)/Z(x,0,0)$. We take this term out of the sum to obtain

$$Z(x,0,0)^{-d}\frac{\partial}{\partial x}Z^{(r)}(x,0,0)-dZ(x,0,0)^{-1}Z^{(r)}(x,0,0)$$
$$=-Z(x,0,0)^{-d}H_r(x)+\sum_{s=0}^{r}\binom{r}{s}(-\mu)^{r-s}s!\sum_{\substack{\ell\in\mathcal{P}(s)\\ \ell_r\neq 1}}\frac{d!}{(d-|\ell|)!}\prod_{j\geqslant 1}\frac{1}{\ell_j!j!^{\ell_j}}\left(\frac{Z^{(j)}(x,0,0)}{Z(x,0,0)}\right)^{\ell_j}.$$

Now recall that

$$Z(x,0,0)=1+Y(x,0,0)=1+Y(x)=(1-(d-1)x)^{-1/(d-1)},\tag{3.4}$$

from which it follows easily that $(\partial/\partial x)Z(x,0,0)^{-d}=-dZ(x,0,0)^{-1}$. Thus we can rewrite it as

$$\frac{\partial}{\partial x}(Z(x,0,0)^{-d}Z^{(r)}(x,0,0))$$
$$=-Z(x,0,0)^{-d}H_r(x)+\sum_{s=0}^{r}\binom{r}{s}(-\mu)^{r-s}s!\sum_{\substack{\ell\in\mathcal{P}(s)\\ \ell_r\neq 1}}\frac{d!}{(d-|\ell|)!}\prod_{j\geqslant 1}\frac{1}{\ell_j!j!^{\ell_j}}\left(\frac{Z^{(j)}(x,0,0)}{Z(x,0,0)}\right)^{\ell_j},$$

which is what we wanted to prove. □

Note that at this stage $H_r(x)$ is only considered as a formal power series, and convergence is not taken into account. We first analyse this differential equation in the special cases $r=1$ and $r=2$ corresponding to mean and variance before we move on to the central limit theorem.

4. Mean and variance

Let us now determine mean and variance of $F(\mathcal{T}_n)$. Since the values of the toll function $f(T)$ for $|T|>n$ will not affect the distribution of $F(\mathcal{T}_n)$, we can assume in this section that $f(T)=0$ for $|T|>n$. This means in particular that the functions $H_r(x)$ also depend on n, so we write $H_r^{(n)}(x)$ to emphasize this dependence. Using the formula (3.2) for the moment generating function of $F(\mathcal{T}_n)-\mu n$, we obtain

$$\mathbb{E}(F(\mathcal{T}_n)-\mu n)=\frac{n!}{Y_n}[x^n]Z^{(1)}(x,0,0).\tag{4.1}$$

For $r=1$, equation (3.3) becomes

$$\frac{\partial}{\partial x}(Z(x,0,0)^{-d}Z^{(1)}(x,0,0))=-Z(x,0,0)^{-d}H_1^{(n)}(x)-\mu,$$

so

$$Z^{(1)}(x,0,0)=Z(x,0,0)^{d}\int_0^x(-Z(w,0,0)^{-d}H_1^{(n)}(w)-\mu)\,dw,$$

where

$$H_1^{(n)}(x)=-\sum_{|T|\leqslant n}\frac{x^{|T|-1}}{(|T|-1)!}f(T).\tag{4.2}$$

If we choose $\mu=\mu^{(n)}$ in such a way that

$$\mu^{(n)}=-(d-1)\int_0^{1/(d-1)}Z(w,0,0)^{-d}H_1^{(n)}(w)\,dw,$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

then we can write (making use of the explicit formula for $Z(x,0,0)$ in (3.4))

$$\begin{aligned} Z^{(1)}(x,0,0) &= Z(x,0,0)^d\left(\frac{\mu^{(n)}}{d-1}+\int_x^{1/(d-1)} Z(w,0,0)^{-d}H_1^{(n)}(w)\,dw-\mu^{(n)}x\right)\\ &= \frac{\mu^{(n)}}{d-1}Z(x,0,0)+R(x), \end{aligned}$$

where

$$R(x)=Z(x,0,0)^d\int_x^{1/(d-1)} Z(w,0,0)^{-d}H_1^{(n)}(w)\,dw. \tag{4.3}$$

Now (4.1) gives us

$$\mathbb{E}(F(\mathcal{T}_n))=\mu^{(n)}n+\frac{n!}{Y_n}[x^n]Z^{(1)}(x,0,0)=\mu^{(n)}\left(n+\frac{1}{d-1}\right)+\frac{n!}{Y_n}[x^n]R(x),$$

so it suffices to determine the contribution from $R(x)$. As we will see in the next lemma, $R(x)$ is a polynomial of degree n whose coefficients can be computed explicitly.

Lemma 4.1. *If* $\beta\in\mathbb{R}\setminus\{-1,-2,-3,\dots\}$,

$$P(x)=\sum_{k=0}^{n-1}a_kx^k \quad \textit{and} \quad Q(x)=(1-x)^{-\beta}\int_x^1(1-w)^\beta P(w)\,dw,$$

then $Q(x)$ is a polynomial of degree n with

$$\begin{aligned}[x^m]Q(x) &= -\frac{a_{m-1}}{m+\beta}+\sum_{k=m}^{n-1}\binom{\beta+m-1}{m}\cdot\frac{\Gamma(\beta+1)k!a_k}{\Gamma(\beta+k+2)}\\ &= O\left(\frac{|a_{m-1}|}{m}+m^{\beta-1}\sum_{k=m}^{n-1}k^{-\beta-1}|a_k|\right).\end{aligned}$$

Proof. To see that $Q(x)$ is indeed a polynomial, one can perform the substitution $w=1-u$ and expand all terms of $P(1-u)$ by means of the binomial theorem; see the calculation below. For $k\geqslant m$, we write

$$(1-x)^{-\beta}\int_x^1(1-w)^\beta w^k\,dw=(1-x)^{-\beta}\int_0^1(1-w)^\beta w^k\,dw-(1-x)^{-\beta}\int_0^x(1-w)^\beta w^k\,dw.$$

The second term does not contribute to $[x^m]Q(x)$ as it is $O(x^{m+1})$, so the only contribution comes from the first term, and this contribution is

$$\left([x^m](1-x)^{-\beta}\right)\cdot\int_0^1(1-w)^\beta w^k\,dw=\binom{\beta+m-1}{m}\cdot\frac{\Gamma(\beta+1)k!}{\Gamma(\beta+k+2)}.$$

On the other hand, if $k<m$, then

$$\begin{aligned}(1-x)^{-\beta}\int_x^1(1-w)^\beta w^k\,dw &= (1-x)^{-\beta}\int_0^{1-x}u^\beta(1-u)^k\,du\\ &= \sum_{j=0}^k\frac{(-1)^j}{j+\beta+1}\binom{k}{j}(1-x)^{j+1}\end{aligned}$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

is a polynomial of degree $k+1$, so the only term that contributes to $[x^m]Q(x)$ comes from $k=m-1$. Putting things together, we have

$$[x^m]Q(x) = -\frac{a_{m-1}}{m+\beta} + \sum_{k=m}^{n-1} \binom{\beta+m-1}{m} \cdot \frac{\Gamma(\beta+1)k!a_k}{\Gamma(\beta+k+2)},$$

which completes the proof of the lemma.

In particular, Lemma 4.1 gives us an expression for $[x^n]R(x)$, since

$$\begin{aligned}[x^n]R(x) &= [x^n](1-(d-1)x)^{-d/(d-1)} \int_x^{1/(d-1)} (1-(d-1)w)^{d/(d-1)} H_1^{(n)}(w)\, dw \\ &= (d-1)^{n-1}[x^n](1-x)^{-d/(d-1)} \int_x^1 (1-u)^{d/(d-1)} H_1^{(n)}\left(\frac{u}{d-1}\right) du.\end{aligned}$$

Evaluating the integral in the expression for $\mu^{(n)}$ explicitly gives us

$$\begin{aligned}\mu^{(n)} &= d(d-1) \sum_{m\leqslant n} \frac{1}{((d-1)m+1)((d-1)m+d)Y_m} \sum_{|T|=m} f(T) \\ &= d(d-1) \sum_{m\leqslant n} \frac{\mathbb{E}(f(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)}.\end{aligned}$$

Putting everything together, we arrive at an explicit formula for the mean:

$$\begin{aligned}\mathbb{E}(F(\mathcal{T}_n)) &= \mu^{(n)}n + \frac{\mu^{(n)}}{d-1} + \frac{n![x^n]R(x)}{Y_n} \\ &= (d(d-1)n+d) \sum_{m\leqslant n} \frac{\mathbb{E}(f(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)} + \frac{n}{(n+d/(d-1))Y_n} \sum_{|T|=n} f(T) \\ &= (d(d-1)n+d) \sum_{m<n} \frac{\mathbb{E}(f(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)} + \mathbb{E}(f(\mathcal{T}_n)).\end{aligned}$$

Completing the series, we have

$$\begin{aligned}&\sum_{m<n} \frac{\mathbb{E}(f(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)} \\ &\qquad = \sum_{m=0}^{\infty} \frac{\mathbb{E}(f(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)} + O\left(\max_{m\geqslant n} |\mathbb{E}(f(\mathcal{T}_m))| \sum_{m=n}^{\infty} \frac{1}{m^2}\right).\end{aligned}$$

Condition (C2) guarantees that the infinite series converges, and the error term is a $o(n^{-1})$. Thus, we arrive exactly at the desired asymptotic formula for the mean in Theorem 2.1. The variance is treated in a similar fashion. Let us return to the differential equation that is satisfied by $Z^{(2)}(x,0,0)$: setting $r=2$ in (3.3) gives us

$$\begin{aligned}&\frac{\partial}{\partial x}(Z(x,0,0)^{-d}Z^{(2)}(x,0,0)) \qquad (4.4)\\ &\quad = -Z(x,0,0)^{-d}H_2^{(n)}(x) + \left(\mu^{(n)}\right)^2 - 2\mu^{(n)}d\frac{Z^{(1)}(x,0,0)}{Z(x,0,0)} + d(d-1)\left(\frac{Z^{(1)}(x,0,0)}{Z(x,0,0)}\right)^2,\end{aligned}$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

where

$$H_2^{(n)}(x) = \sum_{|T|\leqslant n} \frac{x^{|T|-1}}{(|T|-1)!} f(T)^2 - 2\sum_{|T|\leqslant n} \frac{x^{|T|-1}}{(|T|-1)!} f(T)(F(T)-\mu^{(n)}|T|).$$

Let $S(x)$ denote the right side of (4.4) and let

$$c^{(n)} = \int_0^{1/(d-1)} S(x)\,dx \tag{4.5}$$

(note that $S(x)$ depends on n), so that

$$Z^{(2)}(x,0,0) = c^{(n)} Z(x,0,0)^d - Z(x,0,0)^d \int_x^{1/(d-1)} S(w)\,dw. \tag{4.6}$$

We expect the main contribution to the variance to come from the first term on the right side. This is indeed the case, as we can see in the following lemma.

Lemma 4.2. *We have*

$$\mathrm{Var}(F(\mathcal{T}_n)) = (d-1)c^{(n)} n + o(n), \quad \textit{as } n \to \infty.$$

Proof. Since $\mathbb{E}(F(\mathcal{T}_n)) = \mu^{(n)} n + O(1)$, we have

$$\mathrm{Var}(F(\mathcal{T}_n)) = \mathbb{E}\Big(\big(F(\mathcal{T}_n) - \mu^{(n)} n\big)^2\Big) + O(1) = \frac{[x^n]Z^{(2)}(x,0,0)}{[x^n]Z(x,0,0)} + O(1).$$

Now we return to the representation (4.6). We write $S(x)$ in the following way:

$$S(x) = -\frac{\big(\mu^{(n)}\big)^2}{d-1} - Z(x,0,0)^{-d} H_2^{(n)}(x) + d(d-1)\frac{R(x)^2}{Z(x,0,0)^2}, \tag{4.7}$$

where $R(x)$ is defined in (4.3). We now consider the contribution from each of the terms on the right side of (4.7) to the variance. The contributions from the first two terms are not too difficult to estimate.

- Since we have

  $$-Z(x,0,0)^d \int_x^{1/(d-1)} \frac{\big(\mu^{(n)}\big)^2}{d-1}\,dw = -\frac{\big(\mu^{(n)}\big)^2}{(d-1)^2} Z(x,0,0),$$

  its contribution to the variance is

  $$-\frac{\big(\mu^{(n)}\big)^2}{(d-1)^2} = O(1).$$

- By repeating the same argument we used for the mean, we find that the contribution from $-Z(x,0,0)^{-d}H_2^{(n)}(x)$ to the variance is

  $$O(\mathbb{E}|f(\mathcal{T}_n)^2 - 2f(\mathcal{T}_n)(F(\mathcal{T}_n) - \mu^{(n)}|\mathcal{T}_n|)|).$$

  Since the toll function is assumed to be bounded, $F(T) = O(|T|)$, and it follows from condition (C2) that the contribution is in fact $o(n)$.

It remains to estimate the contribution from the third term of (4.7). We first estimate the coefficients of $R(x)$. We have

$$[x^m]H_1^{(n)}(x) = -\frac{Y_{m+1}}{m!}\mathbb{E}(f(\mathcal{T}_{m+1})) = O((d-1)^m m^{1/(d-1)}\mathbb{E}|f(\mathcal{T}_{m+1})|).$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

From Lemma 4.1 and the definition of $R(x)$, we obtain

$$\begin{aligned}[x^m]R(x) &= O\left((d-1)^m\left(m^{(2-d)/(d-1)}\mathbb{E}|f(\mathcal{T}_m)| + m^{1/(d-1)}\sum_{j=m}^{n-1}\frac{\mathbb{E}|f(\mathcal{T}_{j+1})|}{j^2}\right)\right)\\ &= O\Big((d-1)^m m^{(2-d)/(d-1)}\Big).\end{aligned}$$

Note here that

$$\sum_{j=m}^{n-1}\frac{\mathbb{E}|f(\mathcal{T}_{j+1})|}{j^2} = O\big(m^{-1}\big)$$

by condition (C2). We thus have

$$[x^m]R(x)^2 = O\Big((d-1)^m\sum_{k+l=m}(kl)^{(2-d)/(d-1)}\Big) = O\Big((d-1)^m m^{(3-d)/(d-1)}\Big).$$

Again by Lemma 4.1, we also have

$$\begin{aligned}&[x^m]Z(x,0,0)^2\int_x^{1/(d-1)} Z(w,0,0)^{-2}R(w)^2\,dw\\ &\quad= O\Big((d-1)^m\Big(m^{(4-2d)/(d-1)} + m^{(3-d)/(d-1)}\sum_{j=m}^{2n-1}j^{-2}\Big)\Big)\\ &\quad= O\Big((d-1)^m m^{(4-2d)/(d-1)}\Big).\end{aligned}$$

Since the coefficients of $(1-(d-1)x)^{(d-3)/(d-1)}$ ($\log(1-2x)$ if $d=3$) are all negative (except for the constant coefficient) and of order $(d-1)^m m^{(4-2d)/(d-1)}$, we get

$$[x^m]Z(x,0,0)^2\int_x^{1/(d-1)} Z(w,0,0)^{-2}R(w)^2\,dw = \begin{cases} O\Big(\big|[x^m](1-(d-1)x)^{(d-3)/(d-1)}\big|\Big) & d\neq 3,\\ O\left(\big|[x^m]\log(1-2x)\big|\right) & d=3.\end{cases}$$

Multiplying the generating function by $Z(x,0,0)^{d-2} = (1-(d-1)x)^{(2-d)/(d-1)}$ yields

$$[x^n]Z(x,0,0)^d\int_x^{1/(d-1)} Z(w,0,0)^{-2}R(w)^2\,dw = \begin{cases} O(|[x^n](1-(d-1)x)^{-1/(d-1)}|) & d\neq 3,\\ O(|[x^n](1-2x)^{-1/2}\log(1-2x)|) & d=3.\end{cases}$$

Therefore,

$$\frac{[x^n]Z(x,0,0)^d\int_x^{1/(d-1)} Z(w,0,0)^{-2}R(w)^2\,dw}{[x^n]Z(x,0,0)} = \begin{cases} O(1) & d\neq 3,\\ O(\log n) & d=3.\end{cases}$$

Putting all the contributions together, we obtain

$$\mathrm{Var}(F(\mathcal{T}_n)) = c^{(n)}\frac{[x^n]Z(x,0,0)^d}{[x^n]Z(x,0,0)} + o(n) = (d-1)c^{(n)}n + o(n),$$

which is exactly the desired estimate. □

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

The constant $c^{(n)}$ still depends on n, so we write it explicitly in order to determine its behaviour as $n \to \infty$. We have

$$c^{(n)} = -\frac{\left(\mu^{(n)}\right)^2}{(d-1)^2} - \sum_{|T| \leqslant n} \frac{f(T)^2 - 2f(T)(F(T) - \mu^{(n)}|T|)}{\prod_{j=1}^{|T|}((d-1)j+d)}$$
$$+ d \sum_{|T_1| \leqslant n} \sum_{|T_2| \leqslant n} \frac{(d-1)^{-|T_1|-|T_2|} f(T_1) f(T_2)}{(|T_1|-1)!(|T_2|-1)!} \int_0^1 \phi_{|T_1|}(x) \phi_{|T_2|}(x)\, dx,$$

where

$$\phi_k(x) = (1-x)^{-1} \int_x^1 (1-w)^{d/(d-1)} w^{k-1}\, dw.$$

We have already seen that $\mu^{(n)}$ converges to a constant, so let us look at the second term of $c^{(n)}$, which we can split as follows:

$$\sum_{|T| \leqslant n} \frac{f(T)^2 - 2f(T)\,(F(T) - \mu^{(n)}|T|)}{\prod_{j=1}^{|T|}((d-1)j+d)}$$
$$= \sum_{|T| \leqslant n} \frac{f(T)^2 - 2f(T)F(T)}{\prod_{j=1}^{|T|}((d-1)j+d)} + 2\mu^{(n)} \sum_{|T| \leqslant n} \frac{f(T)|T|}{\prod_{j=1}^{|T|}((d-1)j+d)}$$
$$= \sum_{m \leqslant n} \frac{d\mathbb{E}(f(\mathcal{T}_m)^2 - 2f(\mathcal{T}_m)F(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)} + 2\mu^{(n)} \sum_{m \leqslant n} \frac{dm\mathbb{E}(f(\mathcal{T}_m))}{((d-1)m+1)((d-1)m+d)}.$$

Since $|f(T)|$ is assumed to be bounded, we have $F(T) = O(|T|)$, so both sums converge in view of condition (C2).

For the double summation, note first that

$$\phi_m\left(1 - \frac{y}{m}\right) = \frac{m}{y} \int_{1-y/m}^1 (1-w)^{d/(d-1)} w^{m-1}\, dw$$
$$= m^{-d/(d-1)} \int_0^y \frac{t^{d/(d-1)}(1-t/m)^{m-1}}{y}\, dt.$$

The integral is easily seen to be bounded by a constant for all $y \in (0, m]$, so $\phi_m(x) = O\left(m^{-d/(d-1)}\right)$ uniformly for $x \in [0, 1)$. Thus

$$\int_0^1 \phi_{|T_1|}(x)\phi_{|T_2|}(x)\, dx = O\left(|T_1|^{-d/(d-1)}|T_2|^{-d/(d-1)}\right),$$

which means that to prove the convergence of the double sum in the expression for $c^{(n)}$, it suffices to prove convergence of

$$\sum_{|T| \leqslant n} \frac{(d-1)^{-|T|}|T|^{-d/(d-1)}|f(T)|}{(|T|-1)!} = \sum_{m=1}^n (d-1)^{-m} m^{1-d/(d-1)} \frac{Y_m}{m!} \mathbb{E}|f(\mathcal{T}_m)|$$

as $n \to \infty$. Since

$$\frac{Y_m}{m!} = O\left((d-1)^m m^{1/(d-1)-1}\right),$$

this convergence follows from condition (C2). Therefore,

$$\sigma^2 = \lim_{n \to \infty} (d-1)c^{(n)}$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

exists, and

$$\mathrm{Var}(F(\mathcal{T}_n)) = \sigma^2 n + o(n), \qquad \text{as } n \to \infty.$$

5. The central limit theorem

We first consider the case that $f(T)$ has finite support. Conditions (C1) and (C2) are then automatically satisfied, hence the results in the previous section for the mean and variance are valid in this case as well. For the central limit theorem, we also need higher moments, for which we have the following statement.

Lemma 5.1. *If the toll function f has finite support, that is, there exists a constant K such that $f(T) = 0$ whenever $|T| > K$, then the centred moments of the functional F are asymptotically given by*

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

Here, μ and σ are as in Theorem 2.1. Consequently, if $\sigma \neq 0$, then the renormalized random variable

$$\frac{F(\mathcal{T}_n) - \mu n}{\sqrt{\sigma^2 n}}$$

converges weakly to a standard normal distribution.

Proof. Let us return to the general identity (3.3) stated in Lemma 3.1. Since $f(T)$ has finite support, $H_r(x)$ is entire for every $r \geqslant 1$, and μ is given by a finite sum. In particular, $\mu^{(n)} = \mu$ and $c^{(n)} = c$ for large enough n in the notation of the previous section. We prove by induction on p that there exist constants c_p such that

$$Z^{(2p)}(x,0,0) = c_p(1-(d-1)x)^{-1/(d-1)-p} + O(|1-(d-1)x|^{-1/(d-1)-p+1})$$

and

$$Z^{(2p+1)}(x,0,0) = O(|1-(d-1)x|^{-1/(d-1)-p}).$$

This has already been established for $p = 0$, and for $p = 1$, we already know that the first statement holds. Now consider any $r \geqslant 3$ in (3.3). By the induction hypothesis, the product

$$\prod_{j \geqslant 1} \left(\frac{Z^{(j)}(x,0,0)}{Z(x,0,0)} \right)^{\ell_j}$$

is $O(|1-(d-1)x|^{-r/2})$ for every partition ℓ, and the singularity order is exactly $r/2$ if and only if ℓ is a partition of r consisting only of even parts (*i.e.* $\ell_j = 0$ if j is odd). Otherwise, one can improve the bound to $O(|1-(d-1)x|^{-(r-1)/2})$ for odd r (with equality if the partition contains exactly one odd part) and to $O(|1-(d-1)x|^{-r/2+1})$ for even r. After integration and multiplication by $Z(x,0,0)^d$, we end up with the desired result. Moreover, for $p \geqslant 2$, we find that the coefficient c_p can be determined by the recursion

$$c_p(p-1)(d-1) = (2p)! \sum_{\substack{\ell \in \mathcal{P}(p) \\ \ell_p \neq 1}} \frac{d!}{(d-|\ell|)!} \prod_{j \geqslant 1} \frac{c_j^{\ell_j}}{\ell_j!(2j)!^{\ell_j}}.$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

This can be rewritten as

$$\begin{aligned}
c_p(dp-p+1) &= (2p)! \sum_{\ell \in \mathcal{P}(p)} \frac{d!}{(d-|\ell|)!} \prod_{j\geqslant 1} \frac{c_j^{\ell_j}}{\ell_j!(2j)!^{\ell_j}} \\
&= (2p)! \sum_{m=1}^{d} \frac{d!}{(d-m)!}[u^m v^p] \prod_{j\geqslant 1} \left(\sum_{h\geqslant 0} \frac{1}{h!} \left(\frac{uv^j c_j}{(2j)!} \right)^h \right) \\
&= (2p)! \sum_{m=1}^{d} \frac{d!}{(d-m)!}[u^m v^p] \exp \left(\sum_{j\geqslant 1} \frac{uv^j c_j}{(2j)!} \right) \\
&= (2p)![v^p] \left(1 + \sum_{j\geqslant 1} \frac{v^j c_j}{(2j)!} \right)^d .
\end{aligned}$$

Thus if we set

$$C(v) = 1 + \sum_{j\geqslant 1} \frac{v^j c_j}{(2j)!},$$

we have the differential equation

$$(d-1)vC'(v) + C(v) = C(v)^d, \quad C(0) = 1.$$

The general solution of this equation is given by $C(v) = (1-Kv)^{-1/(d-1)}$ (where K is a constant), so

$$c_p = (2p)![v^p](1-Kv)^{-1/(d-1)} = (2p)!K^p \binom{p-1+1/(d-1)}{p},$$

which we can also express in terms of $\sigma^2 = c_1(d-1)$:

$$c_p = (2p)!\sigma^{2p}2^{-p}\binom{p-1+1/(d-1)}{p}.$$

In conclusion, we have

$$Z^{(2p)}(x,0,0) = c_p(1-(d-1)x)^{-1/(d-1)-p} + O(|1-(d-1)x|^{-1/(d-1)-p+1}),$$

so singularity analysis [7, Chapter VI] gives us

$$[x^n]Z^{(2p)}(x,0,0) = \frac{c_p}{\Gamma(p+1/(d-1))} n^{p+1/(d-1)-1}(d-1)^n + O(n^{p+1/(d-1)-2}(d-1)^n)$$

and consequently

$$\begin{aligned}
\mathbb{E}((F(\mathcal{T}_n)-\mu n)^{2p}) &= \frac{[x^n]Z^{(2p)}(x,0,0)}{[x^n]Z(x,0,0)} \\
&= \frac{c_p\Gamma(1/(d-1))}{\Gamma(p+1/(d-1))} n^p + O(n^{p-1}) \\
&= (2p-1)!!\sigma^{2p}n^p + O(n^{p-1}).
\end{aligned}$$

Moreover,

$$\mathbb{E}((F(\mathcal{T}_n)-\mu n)^{2p+1}) = O(n^p),$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

which completes the proof of the asymptotic formulas for the moments. If $\sigma^2 \neq 0$, it follows immediately that the moments of the normalized random variable $(F(\mathcal{T}_n) - \mu n)/\sqrt{\sigma^2 n}$ converge to the moments of a standard normal distribution, which proves the central limit theorem. □

To deal with toll functions that are not finitely supported, we employ a trick that was already used in [9, 12]: we approximate them by truncated versions to which we can apply Lemma 5.1. This approach is based on the following simple yet general lemma (whose proof is also given for completeness).

Lemma 5.2. *If $(X_n)_{n\geqslant 1}$ and $(W_{m,n})_{m,n\geqslant 1}$ are sequences of centred random variables such that*

- *$W_{m,n} \xrightarrow{d}_n W_m$, and $W_m \xrightarrow{d}_m W$, where W has a continuous distribution function,*
- *$\mathrm{Var}(X_n - W_{m,n}) \to_n \gamma_m^2$ and $\gamma_m \to_m 0$,*

then $X_n \xrightarrow{d}_n W$.

Proof. For a random variable X, let $F_X(x)$ denote the distribution function of X. Let x and $\varepsilon > 0$ be fixed real numbers and let m be a fixed positive integer. We have

$$\begin{aligned}\mathbb{P}(X_n \leqslant x) &\leqslant \mathbb{P}(X_n \leqslant x \,\wedge\, X_n - W_{m,n} > -\varepsilon) + \mathbb{P}(X_n - W_{m,n} \leqslant -\varepsilon)\\ &\leqslant \mathbb{P}(W_{m,n} \leqslant x + \varepsilon) + \mathbb{P}(X_n - W_{m,n} \leqslant -\varepsilon).\end{aligned}$$

Similarly,

$$\mathbb{P}(W_{m,n} \leqslant x - \varepsilon) \leqslant \mathbb{P}(X_n \leqslant x \,\vee\, X_n - W_{m,n} \geqslant \varepsilon) \leqslant \mathbb{P}(X_n \leqslant x) + \mathbb{P}(X_n - W_{m,n} \geqslant \varepsilon).$$

Hence

$$\mathbb{P}(W_{m,n} \leqslant x - \varepsilon) - \mathbb{P}(X_n - W_{m,n} \geqslant \varepsilon) \leqslant \mathbb{P}(X_n \leqslant x) \leqslant \mathbb{P}(W_{m,n} \leqslant x + \varepsilon) + \mathbb{P}(X_n - W_{m,n} \leqslant -\varepsilon).$$

Chebyshev's inequality yields

$$\mathbb{P}(W_{m,n} \leqslant x - \varepsilon) - \frac{\mathrm{Var}(X_n - W_{m,n})}{\varepsilon^2} \leqslant \mathbb{P}(X_n \leqslant x) \leqslant \mathbb{P}(W_{m,n} \leqslant x + \varepsilon) + \frac{\mathrm{Var}(X_n - W_{m,n})}{\varepsilon^2}.$$

Taking the limit as $n \to \infty$ and using our assumptions on the convergence of $W_{m,n}$, we find that for every m and $\varepsilon > 0$, we have

$$F_{W_m}(x - \varepsilon) - \frac{\gamma_m^2}{\varepsilon^2} \leqslant \lim_{n\to\infty} \mathbb{P}(X_n \leqslant x) \leqslant F_{W_m}(x + \varepsilon) + \frac{\gamma_m^2}{\varepsilon^2}.$$

By our assumptions on γ_m and W_m, the bounds can simultaneously be made arbitrarily close to $F_W(x)$, which completes the proof of the lemma.

We return to additive functionals and assume that the toll function $f(T)$ satisfies conditions (C1) and (C2). For every positive integer m, consider the truncated toll function f_m and the corresponding function F:

$$f_m(T) = \begin{cases} f(T) & |T| \leqslant m \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad F_m(T) = \sum_{S\in\mathcal{F}(T)} f_m(S) = \sum_{S\in\mathcal{F}(T),|S|\leqslant m} f(S).$$

From Section 4, we know that the mean and variance of $F_m(T)$ have the asymptotic estimates

$$\mathbb{E}(F_m(T)) = \mu_m n + \frac{\mu_m}{d-1} + o(1) \quad \text{and} \quad \mathrm{Var}(F_m(T)) = \sigma_m^2 n + o(n)$$

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

as $n \to \infty$. Furthermore, for each m, if $\sigma_m^2 \neq 0$ then $F_m(T)$ satisfies the central limit theorem, and $\mu_m \to \mu$ and $\sigma_m^2 \to \sigma^2$ as $m \to \infty$. On the other hand, the functional $F(T) - F_m(T)$ is also additive with toll function $f(T) - f_m(T)$. The conditions (C1) and (C2) are both satisfied by the latter toll function, so from the asymptotic formula for the variance we know that

$$\gamma_m^2 = \lim_{n\to\infty} \frac{\mathrm{Var}(F(\mathcal{T}_n) - F_m(\mathcal{T}_n))}{n} \to_m 0$$

under the conditions on the toll function f. Hence, Lemma 5.2 applies to the sequences

$$W_{m,n} = \frac{F_m(\mathcal{T}_n) - \mathbb{E}(F_m(\mathcal{T}_n))}{\sqrt{n}} \quad \text{and} \quad X_n = \frac{F(\mathcal{T}_n) - \mathbb{E}(F(\mathcal{T}_n))}{\sqrt{n}},$$

which proves Theorem 2.1 for arbitrary toll functions f that satisfy (C1) and (C2).

6. Some applications

We conclude this paper with three applications of our general results.

6.1 Fringe subtrees of given size and occurrences of specific fringe subtrees

The simplest example of a toll function is perhaps the indicator function of a specific tree S:

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

The associated additive functional is simply the number of occurrences of S on the fringe of a random tree: by an occurrence of S, we mean a fringe subtree that is isomorphic to S (including the relative order of the labels).

In this case, we obtain a central limit theorem with mean and variance only depending on the size of S: if S has k vertices, then

$$\mu = \frac{d-1}{\prod_{j=1}^{k} ((d-1)j + d)}$$

and

$$\sigma^2 = -\mu^2\left(2k + \frac{1}{d-1}\right) + \mu + \frac{d(d-1)^{1-2k}}{(k-1)!^2} \int_0^1 \phi_k(x)^2 \, dx.$$

A closely related functional is the number of fringe subtrees of some given size k (equivalently, the number of vertices with exactly $k-1$ descendants). In particular, the special case $k = 1$ corresponds to the number of leaves. Here, the toll function is given by

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

and we obtain a central limit theorem with

$$\mu = \frac{d(d-1)}{((d-1)k + d)((d-1)k + 1)}$$

and

$$\sigma^2 = -\mu^2\left(2k + \frac{1}{d-1}\right) + \mu + \frac{d(d-1)^{1-2k}\Upsilon_k^2}{(k-1)!^2} \int_0^1 \phi_k(x)^2 \, dx.$$

This was already shown by Fuchs [8], who also considered the case that k is not fixed but rather tends to infinity with the size of the tree as well.

https://doi.org/10.1017/S0963548318000585 Published online by Cambridge University Press

6.2 The number of subtrees

The number of subtrees is a somewhat more complicated example: for Galton–Watson trees, binary increasing trees and recursive trees, it was already studied in [17]. Here, we count all subtrees, that is, all induced subgraphs that are again trees, not just those on the fringe. It is useful to study an auxiliary quantity first, namely the number of subtrees containing the root: we write $s(T)$ for this number. It is not difficult to see that (for arbitrary trees T with branches $B_1, B_2, \ldots, B_k$)

$$s(T) = \prod_{j=1}^{k} (1 + s(B_j)),$$

since each subtree induces either the empty set or a subtree containing the root in each of the branches. Taking the logarithm gives us

$$\log(1 + s(T)) = \sum_{j=1}^{k} \log(1 + s(B_j)) + \log(1 + s(T)^{-1}),$$

so $F(T) = \log(1 + s(T))$ is an additive functional with toll function $f(T) = \log(1 + s(T)^{-1})$ (*i.e.* $F(T)$ and $f(T)$ satisfy the general definition (1.1)). Simple *a priori* estimates show that the technical conditions of our general central limit theorem are satisfied: this is because $s(T) \geqslant |T|$ (since every path from the root to a vertex is also a subtree), which implies that $f(T) = O(|T|^{-1})$ for all T (even deterministically, not just on average). Thus our main result applies to the functional $s(T)$. As shown in [17], the difference between $F(T) = \log(1 + s(T))$ and the logarithm of the total number of subtrees (not necessarily containing the root) is $O(\log |T|)$, so the central limit theorem remains correct for the total number of subtrees.

6.3 The size of the automorphism group

An important motivating example for this paper is the size of the automorphism group. Bóna and Flajolet [2], motivated by questions in phylogenetics, proved that the logarithm of the size of the automorphism group of uniformly random binary trees is asymptotically normally distributed (they proved this limit law for the number of nodes for which the two branches are isomorphic, which is equivalent). Here, we obtain an analogous statement for d-ary increasing trees. We remark that binary increasing trees are also essentially equivalent to the Yule–Harding model (as opposed to the uniform model) of phylogenetics [16, Section 2.5].

As mentioned in the Introduction, the relevant toll function is $f(T) = \log(R(T))$, where $R(T)$ is the size of the symmetry group of the collection of root branches. This simplifies considerably in the case of binary trees, where we only have two branches B_1 and B_2. In this case, it follows that

$$f(T) = \begin{cases} \log 2 & \text{if } B_1 \text{ and } B_2 \text{ are isomorphic,} \\ 0 & \text{otherwise.} \end{cases}$$

As one would expect, it is very unlikely for large trees that the two branches are actually isomorphic, which is why the technical condition on the toll function is satisfied. In fact, one can show that $\mathbb{E}|f(\mathcal{T}_n)|$ decays exponentially for binary increasing trees. We find that the number of automorphisms of a random binary increasing tree asymptotically follows a log-normal law, which parallels the aforementioned result of Bóna and Flajolet.

The same holds more generally for d-ary trees, although the expected value of the toll function does not decay as quickly: in this case, the probability that two branches are isomorphic only decreases at a rate of $O\big(|T|^{-2/(d-1)}\big)$, which however is still sufficient. Let us prove this fact: the toll function $f(T) = \log(R(T))$ is clearly bounded by $\log d!$, since the symmetry group $R(T)$ is a subgroup of the symmetric group on d elements, so (C1) is satisfied. It remains to prove (C2).

Table 1. Values of μ and σ^2 for selected examples (number of leaves, subtrees and automorphisms) in three cases: binary ($d = 2$) and ternary ($d = 3$) increasing trees and PORTs ($\alpha = 1$). Numerical values are given to the highest accuracy we were able to obtain with our numerical calculations. We remark that the technical conditions of our general theorem are unfortunately not satisfied for the number of automorphisms in PORTs, since its growth can be faster than exponential (*e.g.* for stars).

Example	μ	σ^2
number of leaves, binary	1/3	2/45
number of leaves, ternary	2/5	3/50
number of leaves, PORTs	2/3	1/9
(log of) number of subtrees, binary	0.3507	0.008
(log of) number of subtrees, ternary	0.3926	0.011
(log of) number of subtrees, PORTs	0.5380	0.024
(log of) number of automorphisms, binary	0.0320	0.017
(log of) number of automorphisms, ternary	0.054	0.03

To this end, we determine an upper bound for the probability that $f(\mathcal{T}_n) \neq 0$. This can only happen if two (or more) branches are isomorphic, so in particular two of the branches need to have the same (non-zero) size. There are $\binom{d}{2}$ possible choices for the two branches. If their sizes are k, then there are Y_k^2 possibilities for these branches, and the other branches have to contain $n-2k-1$ vertices (but are otherwise arbitrary). It follows that the number of d-ary increasing trees with at least two isomorphic branches, as well as the number of d-ary increasing trees with at least two branches of equal size, is bounded above by

$$\binom{d}{2}(n-1)!\sum_{k\geqslant 1}\left(\frac{Y_k}{k!}\right)^2[x^{n-2k-1}]Y(x)^{d-2} = \binom{d}{2}(n-1)![x^{n-1}]\sum_{k\geqslant 1}\left(\frac{Y_k}{k!}\right)^2 x^{2k}Y(x)^{d-2}.$$

The series $\sum_{k\geqslant 1}(Y_k/k!)^2x^{2k}$ can be regarded as a Hadamard product of $Y(x^2)$ with itself. By standard closure properties ([5]; see also [7, Section VI.10.2]), it is therefore amenable to singularity analysis. Its dominant singularity at $1/(d-1)$ is of the type

$$(1-(d-1)x)^{(d-3)/(d-1)}$$

if $d \neq 3$, and $-\log(1-2x)$ if $d = 3$. It follows that the singularity of the product with $Y(x)^{d-2}$ is of the form $(1-x)^{-1}$ if $d = 2$, $-(1-2x)^{-1/2}\log(1-2x)$ if $d = 3$ and $(1-(d-1)x)^{-(d-2)/(d-1)}$ for $d > 3$. Applying singularity analysis, we find that

$$\frac{\binom{d}{2}(n-1)![x^{n-1}]\sum_{k\geqslant 1}(Y_k/k!)^2x^{2k}Y(x)^{d-2}}{Y_n} = \begin{cases} O(n^{-1}) & d = 2, \\ O\big(n^{-1}\log n\big) & d = 3, \\ O\big(n^{-2/(d-1)}\big) & d > 3. \end{cases}$$

Since this is an upper bound on the probability that $f(\mathcal{T}_n) \neq 0$, we also obtain

$$\mathbb{E}|f(\mathcal{T}_n)| = \mathbb{E}(f(\mathcal{T}_n)) = \begin{cases} O(n^{-1}) & d = 2, \\ O(n^{-1}\log n) & d = 3, \\ O\Big(n^{-2/(d-1)}\Big) & d > 3, \end{cases}$$

which is sufficient to imply conditions (C1) and (C2). The bounds can be improved somewhat by also taking into account that the probability that two d-ary increasing trees of size k are isomorphic decreases with k.

Conflict of interest. None.

References

[1] Bergeron, F., Flajolet, P. and Salvy, B. (1992) Varieties of increasing trees. In *Colloquium on Trees in Algebra and Programming (CAAP '92)*, Vol. 581 of Lecture Notes in Computer Science, Springer, pp. 24–48.

[2] Bóna, M. and Flajolet, P. (2009) Isomorphism and symmetries in random phylogenetic trees. *J. Appl. Probab.* **46** 1005–1019.

[3] Devroye, L. (2002/03) Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.* **32** 152–171.

[4] Drmota, M. (2009) *Random Trees: An Interplay Between Combinatorics and Probability*, Springer.

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

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

[7] Flajolet, P. and Sedgewick, R. (2009) *Analytic Combinatorics*, Cambridge University Press.

[8] Fuchs, M. (2012) Limit theorems for subtree size profiles of increasing trees. *Combin. Probab. Comput.* **21** 412–441.

[9] Holmgren, C. and Janson, S. (2015) Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.* **20** #4.

[10] Holmgren, C., Janson, S. and Šileikis, M. (2017) Multivariate normal limit laws for the numbers of fringe subtrees in m-ary search trees and preferential attachment trees. *Electron. J. Combin.* **24** P2.51.

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

[12] Janson, S. (2016) Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton–Watson trees. *Random Struct. Alg.* **48** 57–101.

[13] Meir, A. and Moon, J. W. (1998) On the log-product of the subtree-sizes of random trees. *Random Struct. Alg.* **12** 197–212.

[14] Panholzer, A. and Prodinger, H. (2007) Level of nodes in increasing trees revisited. *Random Struct. Alg.* **31** 203–226.

[15] Ralaivaosaona, D. and Wagner, S. (2016) Additive functionals of d-ary increasing trees. In *Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms.* arXiv:1605.03918

[16] Semple, C. and Steel, M. (2003) *Phylogenetics*, Oxford University Press.

[17] Wagner, S. (2015) Central limit theorems for additive tree parameters with small toll functions. *Combin. Probab. Comput.* **24** 329–353.

Cite this article: Ralaivaosaon D and Wagner S (2019). A Central Limit Theorem for Additive Functionals of Increasing Trees. *Combinatorics, Probability and Computing* **28**, 618–637. https://doi.org/10.1017/S0963548318000585