Combinatorics, Probability and Computing (2015) 24, 329–353. © Cambridge University Press 2014
doi:10.1017/S0963548314000443

Central Limit Theorems for Additive Tree Parameters with Small Toll Functions

STEPHAN WAGNER[†]

Department of Mathematical Sciences, Stellenbosch University, 7602 Stellenbosch, South Africa
(e-mail: swagner@sun.ac.za)

Received 13 March 2013; revised 22 September 2013; first published online 8 September 2014

Dedicated to the memory of Philippe Flajolet

We call a tree parameter additive if it can be determined recursively as the sum of the parameter values of all branches, plus a certain toll function. In this paper, we prove central limit theorems for very general toll functions, provided that they are bounded and small on average. Simply generated families of trees are considered as well as Pólya trees, recursive trees and binary search trees, and the results are illustrated by several examples of parameters for which we prove normal or log-normal limit laws.

2010 *Mathematics subject classification*: Primary 60C05
Secondary 05A16, 05C05, 05C80, 60F05

1. Introduction

By an *additive tree parameter*, we mean a parameter F which satisfies a recursion of the form

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

where $T_1, T_2, \dots, T_k$ are the branches of the rooted tree T and f is a so-called *toll function.* It is consistent with this recursion to set $f(\bullet) = F(\bullet)$, where $\bullet$ denotes a tree consisting of only a single vertex. *A priori*, every tree parameter is additive if no further assumptions on the toll function are made. In the literature, the following special cases have been treated

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

most extensively:

- the toll function only depends on the order (number of vertices) of T,
- the toll function only depends on the root degree of T.

See also the following examples.

- The *number of leaves* [4, 5, 8, 14, 33] is an additive parameter with toll function

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

- More generally, the *number of vertices of outdegree* k [8, 22, 33], with toll function

$$f(T) = \begin{cases} 1 & \text{if the root degree of } T \text{ is } k, \\ 0 & \text{otherwise.} \end{cases}$$

- The *number of full subtrees* (*i.e.*, subtrees consisting of a vertex and all its successors) *of size* k [1, 4, 5, 10, 14, 16]. Here, the toll function is

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

- The *path length* [6, 25, 34, 35], with toll function

$$f(T) = |T| - 1.$$

- The *log-product of the subtree sizes*, also called the *shape parameter* [11, 13, 28], whose toll function is

$$f(T) = \log |T|.$$

- The *number of subtrees* [2, 23, 27, 31], *i.e.*, all connected induced subgraphs. To turn it into an additive parameter, we have to take the logarithm, and the toll function is somewhat more complicated than in the previous examples.

Note how we distinguish between *subtrees* (all subgraphs that are themselves trees) and *full subtrees* (subtrees consisting of a vertex and all its descendants). The number of subtrees will be one of our toy examples throughout the paper to demonstrate our results, and more examples will follow later in the text as well. In this paper, we are considering rather general toll functions, but we make an assumption on the average growth of f that allows us to prove central limit theorems for the associated tree parameters F under various different random tree models: it is assumed that the average of $|f|$, taken over all trees of order n, goes to 0 at an exponential rate. As we will see, this assumption is satisfied for many natural examples.

There are many existing results about additive parameters of various kinds. However, as mentioned before, the toll functions are often assumed to depend either only on the tree order or only on the root degree. We allow very general toll functions here, but the higher degree of generality has a price in that we need to make rather strong assumptions on their size – we will have to assume exponential decay on average.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

In the conference article [37], the author proved a central limit theorem for additive parameters under the same assumptions as in the present paper, but only for the class of labelled trees. This approach, which makes use of the method of moments, seems too complicated to be applied to more general families of trees, so a different method is used here. We will consider the following families of (random) trees.

Simply generated families of trees. Simply generated families of trees – introduced by Meir and Moon [26] more than 30 years ago – are of central interest in this paper. A simply generated family $\mathcal{F}$ of trees is defined by a sequence $\phi_0, \phi_1, \dots$ of non-negative weights, with the additional assumption that $\phi_0 > 0$ (typically, $\phi_0 = 1$) and $\phi_j > 0$ for some $j > 1$. Let T be a rooted ordered tree (*i.e.*, the order of the children of a vertex matters). We write $D_j(T)$ for the number of vertices whose outdegree is j, and we define the *weight* of T by

$$w(T) = \prod_{j \geqslant 0} \phi_j^{D_j(T)}.$$

The weights define a natural probability distribution on trees of given order, where the probability of any tree is proportional to its weight. Amongst others, random plane trees ($\phi_j = 1$ for all j), random rooted labelled trees ($\phi_j = 1/j!$), random d-ary trees ($\phi_0 = \phi_d = 1$ and $\phi_j = 0$ otherwise), random pruned d-ary trees ($\phi_j = \binom{d}{j}$) and random unary–binary trees ($\phi_0 = \phi_1 = \phi_2 = 1$, $\phi_j = 0$ otherwise) can be generated in this way.

It is well known that the (weight) generating function of a simply generated family of trees, namely

$$T(x) = \sum_T w(T) x^{|T|},$$

where the sum is taken over all rooted ordered trees, satisfies the functional equation

$$T(x) = x\Phi(T(x)),$$

with

$$\Phi(t) = \sum_{j=0}^{\infty} \phi_j t^j.$$

Under certain technical conditions, the asymptotic behaviour of the coefficients of $T(x)$ follows directly from this functional equation. The following theorem, due to Meir and Moon, is classical.

Theorem 1.1 ([7, Theorem 3.6]). *Let R be the radius of convergence of*

$$\Phi(t) = \sum_{j=0}^{\infty} \phi_j t^j,$$

and suppose that there exists some $\tau \in (0, R)$ with $\tau\Phi'(\tau) = \Phi(\tau)$. Finally, let d be the gcd of all indices j with $\phi_j > 0$. Then $T(x)$ has a dominant square-root singularity at $\rho = \tau/\Phi(\tau) =$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

$\Phi'(\tau)^{-1}$ whose asymptotic expansion starts

$$T(x) = \tau - \sqrt{\frac{2\Phi(\tau)}{\Phi''(\tau)}} \cdot \sqrt{1 - x/\rho} + O(|1 - x/\rho|). \tag{1.1}$$

The following asymptotic formula for the coefficients of $T(x)$ holds:

$$t_n = [x^n]T(x) = d\sqrt{\frac{\Phi(\tau)}{2\pi\Phi''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}(1 + O(n^{-1}))$$

if $n \equiv 1 \bmod d$ (and $t_n = 0$ otherwise).

Pólya trees and related families. Pólya trees are rooted *unordered* (and unlabelled) trees, *i.e.*, the order of the branches is irrelevant. One can also regard them as isomorphism classes of plane trees. It is a classical result (see [17, eq. (3.1.4)] or [7, Theorem 3.8]) that their generating function $R(x)$ satisfies

$$R(x) = x \exp\left(\sum_{k=1}^{\infty} \frac{1}{k} R(x^k)\right),$$

which can be seen easily from the fact that a Pólya tree consists of the root and an unordered collection of branches that are themselves Pólya trees.

Just like the generating function of simply generated trees, $R(x)$ has a square-root singularity at $\rho \approx 0.33832185$, from which we obtain the asymptotic number of Pólya trees of order n:

$$r_n \approx 0.43992401 \cdot n^{-3/2} \cdot 2.95576528^n.$$

Pólya trees are also an important step in the enumeration of unrooted unlabelled ('free') trees that is due to Otter [32]. Their generating function is given by

$$T(x) = R(x) - \frac{1}{2}(R(x)^2 - R(x^2)),$$

and it follows that the number of unlabelled trees of order n is asymptotically

$$t_n \approx 0.53494961 \cdot n^{-5/2} \cdot 2.95576528^n.$$

It is well known that Pólya trees behave like simply generated trees in many ways, although they do not technically belong to the class of simply generated trees (see [9]). As it turns out, this also applies to our problems. We also study the related family of homeomorphically irreducible trees, which do not have vertices of degree 2, motivated by a graph-theoretical question due to Vince and Wang [36].

Recursive trees. Recursive trees [7, Section 1.3] can be generated by a simple probabilistic model. At each step, vertex n is (randomly) attached to one of the previous $n - 1$ vertices. This means that there are precisely $(n - 1)!$ recursive trees of order n. Recursive trees belong to the wider class of *increasing trees* [3]: starting at the root, the labels along any path are increasing. Instead of a functional equation, the exponential generating function for increasing trees satisfies a differential equation of the form $T'(x) = \Phi(T(x))$.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

For recursive trees, it is given by

$$T'(x) = \exp(T(x)),$$

with explicit solution $T(x) = -\log(1-x)$.

Binary search trees. A binary search tree of order n is obtained from a (random) permutation π of $\{1,2,\dots,n\}$: $\pi(1)$ becomes the root, and $\pi(2), \pi(3), \dots$ are inserted step by step in such a way that labels that are smaller than the root label are stored in the left subtree, while labels greater than the root label are stored in the right subtree.

Random binary search trees are equivalent to random binary increasing trees (trees with increasing labels such that each vertex has either no child, or one left child, or one right child, or two children) under the uniform model, and they can also be regarded as an analytic model for the famous Quicksort algorithm. Our final section deals with additive parameters of binary search trees. Their exponential generating function also satisfies a differential equation, namely

$$T'(x) = (1 + T(x))^2,$$

whose explicit solution is $T(x) = x/(1-x)$.

Before we start with our analysis, let us review some results on additive parameters from the literature. Parameters that only depend on vertex (out-)degrees are classical; see [7, Sections 3.2 and 6.2.4] for a thorough treatment. Their distribution is Gaussian under very general conditions. In particular, the number of vertices of a certain degree d is normally distributed in all of the aforementioned cases (simply generated trees, Pólya trees, recursive trees, binary search trees). The same is true for the number of occurrences of certain patterns in random simply generated trees and Pólya trees [7, Section 3.3]. However, not all additive parameters follow a Gaussian distribution in the limit: the typical counterexample is the internal path length, which was shown by Takács to follow an Airy distribution [34, 35] for simply generated families of trees. This is also important in the analysis of the Wiener index (the sum of all distances between pairs of vertices) [21] that can also be seen as an additive parameter.

The distributional behaviour of additive tree parameters depends very much on the growth of the associated toll functions. This becomes particularly clear in the paper of Fill and Kapur [13] in which the special toll functions $f(T) = |T|^\alpha$ and $f(T) = \log|T|$ are studied for pruned binary trees (Catalan trees). Fill, Flajolet and Kapur [12] show how to make use of Hadamard products to find the mean behaviour of additive parameters if the toll function only depends on the tree order.

Hwang and Neininger [19] study the phase transitions that occur as the toll function varies. They consider binary search trees (in disguise, by looking at the Quicksort recursion) with toll functions that only depend on the order, but may otherwise be random.

2. Simply generated families of trees

We start with our analysis of simply generated trees. The aim of this section is the proof of the following theorem.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

Theorem 2.1. *Consider a simply generated family $\mathcal{F}$ of trees that satisfies the conditions of Theorem 1.1, and assume that the toll function f is bounded and satisfies*

$$\frac{\sum_{|T|=n} w(T)|f(T)|}{\sum_{|T|=n} w(T)} = O(c^n)$$

for a constant $c \in (0,1)$.

Let T_n denote a random tree of order n in $\mathcal{F}$. The mean $\mu_n = \mathbb{E}(F(T_n))$ of the parameter F is asymptotically

$$\mu_n = \mu n + O(1),$$

where the constant μ is given by

$$\mu = \frac{1}{\tau} \sum_T w(T) f(T) \rho^{|T|}.$$

The variance $\sigma_n^2 = \mathbb{V}(F(T_n))$ of F is asymptotically

$$\sigma_n^2 = \sigma^2 n + O(1),$$

where

$$\sigma^2 = \mu^2\left(1 - \frac{\Phi(\tau)}{\tau^2\Phi''(\tau)}\right) + \frac{1}{\tau}\sum_T w(T)f(T)(2F(T)-f(T))\rho^{|T|} - \frac{2\mu}{\tau}\sum_T w(T)f(T)|T|\rho^{|T|}.$$

Moreover, if $\sigma^2 \neq 0$, the renormalized random variable

$$\frac{F(T_n)-\mu_n}{\sigma_n}$$

converges weakly to a standard normal distribution.

Remark 1. There are instances of tree parameters where σ^2 is indeed zero, *e.g.*, the number of leaves in binary trees.

Remark 2. This and all results for other tree classes remain true if the toll function $f(T)$ is replaced by $f(T)+C$ for some constant C, for the simple reason that this only changes $F(T)$ by a deterministic quantity, namely $C|T|$.

Our proof is based on an analysis of the bivariate generating function

$$Y(x,u) = \sum_T w(T) x^{|T|} u^{F(T)},$$

which satisfies

$$\begin{aligned} x\Phi(Y(x,u)) &= x\sum_{j\geqslant 0}\phi_j \sum_{T_1,T_2,\dots,T_j}\prod_{i=1}^{j} w(T_i)x^{|T_i|}u^{F(T_i)} \\ &= \sum_T w(T)x^{|T|}u^{F(T)-f(T)}. \end{aligned} \tag{2.1}$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

The right-hand side of this equation is 'almost' equal to $Y(x,u)$, except for the additional term $-f(T)$ in the exponent. Of course, for $u=1$, the equation reduces to

$$Y(x,1) = x\Phi(Y(x,1)).$$

Since the toll function f can be rather arbitrary, the right-hand side cannot usually be expressed algebraically in terms of $Y(x,u)$ and elementary functions. However, our assumption on the toll function allows us to obtain analytic information as we will see later. We start with a brief discussion of the moments, although this is technically not necessary since the general theorem that we are going to apply actually covers mean and variance as well. Throughout the proof we assume, without too much loss of generality, that the parameter d in Theorem 1.1 is 1.

2.1. Moments

In order to determine the asymptotic behaviour of the moments (with precise error terms, as stated in the theorem), we need to consider the partial derivatives with respect to u. Differentiating (2.1) with respect to u and setting $u=1$, we get

$$x\Phi'(Y(x,1))Y_u(x,1) = \sum_T w(T)(F(T)-f(T))x^{|T|} = Y_u(x,1) - \sum_T w(T)f(T)x^{|T|}.$$

On the other hand, differentiating with respect to x yields

$$\Phi(Y(x,1)) + x\Phi'(Y(x,1))Y_x(x,1) = Y_x(x,1).$$

Comparing the two, we find that

$$Y_u(x,1) = \frac{xY_x(x,1)}{Y(x,1)} \sum_T w(T)f(T)x^{|T|}.$$

The same can be done with the second derivative: differentiating (2.1) with respect to u twice and setting $u=1$ yields

$$\begin{aligned} x\Phi''(Y(x,1))Y_u(x,1)^2 + x\Phi'(Y(x,1))Y_{uu}(x,1) \\ = \sum_T w(T)(F(T)-f(T))(F(T)-f(T)-1)x^{|T|} \\ = Y_{uu}(x,1) - \sum_T w(T)f(T)(2F(T)-f(T)-1)x^{|T|}. \end{aligned}$$

Set

$$H_1(x) = \sum_T w(T)f(T)x^{|T|} \quad \text{and} \quad H_2(x) = \sum_T w(T)f(T)(2F(T)-f(T)-1)x^{|T|}.$$

By the assumptions on the toll function, H_1 has a larger radius of convergence than $Y(x,1)$, and so it is analytic at the dominating singularity ρ of $Y(x,1)$. The same is true for $H_2(x)$, since $|F(T)| = O(|T|)$ by the boundedness of f. Now the asymptotic expansion of $Y(x,1)$ around the dominating singularity ($Y(x,1) = T(x)$ in the notation of (1.1)) gives us the asymptotic behaviour of $Y_u(x,1)$ as well: since $H_1(x)$ is analytic at ρ, we have

$$Y_u(x,1) \sim \frac{H_1(\rho)}{\tau} xY_x(x,1) \sim \frac{1}{2\tau}\sqrt{\frac{2\Phi(\tau)}{\Phi''(\tau)}} H_1(\rho)\left(1-\frac{x}{\rho}\right)^{-1/2}.$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

It now follows by means of singularity analysis [15, Chapter VI] that the mean is asymptotically equal to $\tau^{-1}H_1(\rho)n + O(1)$. The variance can be treated similarly.

2.2. Limiting distribution

To prove convergence to a Gaussian limit distribution, we make use of the following general result (see [7, Theorem 2.23]).

Lemma 2.2. *Suppose that*

$$F(x,y,u) = \sum_{n,m=0}^{\infty} F_{n,m}(u)x^n y^m$$

is an analytic function in x, y and u around $(0,0,1)$ such that $F(0,y,u) \equiv 0$, $F(x,0,u) \not\equiv 0$ and all coefficients $F_{n,m}(1)$ of $F(x,y,1)$ are real and non-negative. Moreover, let $y = y(x,u)$ be the unique solution of the functional equation

$$y = F(x,y,u)$$

with $y(0,u) = 0$. Assume further that there exist positive solutions $x = x_0$ and $y = y_0$ of the system of equations

$$\begin{aligned} y &= F(x,y,1), \\ 1 &= F_y(x,y,1), \end{aligned}$$

with $F_x(x_0,y_0,1) \neq 0$ and $F_{yy}(x_0,y_0,1) \neq 0$. Let the sequence of random variables $X_1, X_2, \dots$ be defined by their probability generating functions

$$\mathbb{E}(u^{X_n}) = \frac{[x^n]y(x,u)}{[x^n]y(x,1)}.$$

Then there are constants $\mu \geqslant 0$ and $\sigma^2 \geqslant 0$ such that

$$\mathbb{E}(X_n) = \mu n + O(1) \quad \textit{and} \quad \mathbb{V}(X_n) = \sigma^2 n + O(1).$$

Here, $\mu = F_u(x_0,y_0,1)/(x_0F_x(x_0,y_0,1))$, and σ^2 can also be represented in terms of partial derivatives of F at $(x_0,y_0,1)$. Moreover, if $\sigma^2 \neq 0$, then X_n (suitably renormalized) converges weakly to a normal distribution:

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{d} N(0,1).$$

Going back to the functional equation (2.1), let us rewrite it as

$$x\Phi(Y(x,u)) = \sum_T w(T)x^{|T|}u^{F(T)-f(T)} = Y(x,u) - \sum_T w(T)x^{|T|}u^{F(T)}(1-u^{-f(T)}).$$

By the boundedness of f, we have

$$|1-u^{-f(T)}| = O(|f(T)|)$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

Table 1. Values of μ and σ^2 for the distribution of the number of leaves in different families of trees

Tree family	μ	σ^2
Labelled trees, $\Phi(t) = e^t$	$\frac{1}{e}$	$\frac{e-2}{e^2}$
Plane trees, $\Phi(t) = (1-t)^{-1}$	$\frac{1}{2}$	$\frac{1}{8}$
Pruned d-ary trees, $\Phi(t) = (1+t)^d$	$\left(\frac{d-1}{d}\right)^d$	$\left(\frac{d-1}{d}\right)^d - \frac{2d-1}{d-1}\left(\frac{d-1}{d}\right)^{2d}$
Unary–binary trees, $\Phi(t) = 1+t+t^2$	$\frac{1}{3}$	$\frac{1}{18}$

if u is restricted to a fixed disk around 1. It follows (once again by the exponential decay of the average of $|f(T)|$) that the function

$$\sum_T w(T)x^{|T|}u^{F(T)}(1-u^{-f(T)})$$

is analytic in x in a circle of radius $R(\delta) > \rho$ around the origin if $|u-1| < \delta$ for a suitable $\delta > 0$, which makes Lemma 2.2 applicable. The function $F(x, y, u)$ is given by

$$F(x, y, u) = x\Phi(y) + \sum_T w(T)x^{|T|}u^{F(T)}(1-u^{-f(T)})$$

in this specific case, and all technical conditions are easily verified.

2.3. Examples

Let us now consider a couple of examples that show how our main theorem is applied. As a first example, we study the case of the number of leaves (and generalizations thereof), which is quite simple and also well known. The other examples are new and do not seem to be covered by other results in the literature.

2.3.1. The number of leaves and generalizations. The number of leaves is perhaps the simplest and most classical example of an additive tree parameter. Recall that the corresponding toll function is given by

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

Obviously, all conditions are satisfied, which means that we obtain the well-known central limit theorem for the number of leaves (see [7, Theorem 3.13]) as a special case of Theorem 2.1. Explicit values of mean and variance are given in Table 1.

This example can be generalized in many different ways, for instance to the number of full subtrees of order k or at most k (equivalently, the number of vertices with exactly $k-1$ or at most $k-1$ descendants). The associated toll functions are

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

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

and

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

respectively. Even more generally, if $\mathcal{S}$ is a set of trees (sufficiently 'sparse'; otherwise, no conditions on $\mathcal{S}$ need to be made), then the toll function

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

gives rise to a tree parameter F that counts the number of all full subtrees in $\mathcal{S}$. For instance, if $\mathcal{S}$ is the set of all stars (rooted at their centres), then F counts the number of vertices whose children are all leaves. For plane trees, the constants μ and σ^2 associated with this parameter are

$$\mu = 2\sum_{n=2}^{\infty} 4^{-n} = \frac{1}{6}$$

and

$$\sigma^2 = \frac{11}{216}.$$

For labelled trees, the values

$$\mu = e^{-1}(e^{1/e} - 1) \quad \text{and} \quad \sigma^2 = e^{-3}(e^{1/e} - 1)(e^2 + 2e - 2(e+1)e^{1/e})$$

have already been determined in [37]. Generalizing further, we can for example consider the number of vertices in plane trees for which the longest path to a leaf has length k (equivalently, the full subtree S rooted at the vertex has height $h(S) = k$). It turns out that their number satisfies a central limit theorem with

$$\mu_k = 2\sum_{T:h(T)=k} 4^{-|T|} = \frac{1}{(k+1)(k+2)}$$

and

$$\sigma_k^2 = \frac{2k^2 + 6k + 3}{6(k+1)^2(k+2)^2}.$$

2.3.2. The number of antichains and subtrees. The number of antichains in plane trees was studied, amongst other parameters, in a paper by Klazar [23, 24]: we are considering a rooted tree as the Hasse diagram of a poset in this context, and we are counting all possible vertex subsets that form an antichain. Klazar proved that the average number of antichains in a random plane tree of order n is asymptotically equal to

$$\frac{4}{\sqrt{15}} \cdot \left(\frac{25}{16}\right)^n.$$

Our goal here is to show that it asymptotically follows a log-normal distribution, even for arbitrary simply generated families of trees.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

Let $a(T)$ denote the number of (non-empty) antichains of a rooted tree T whose branches are $T_1, T_2, \ldots, T_k$. An antichain either consists only of the root or of a union of arbitrary (possibly empty) antichains in all the branches, hence we have

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

Note here that the product counts all possible combinations of antichains in the branches, including the empty set. However, the antichain consisting only of the root makes up for it. We can rewrite the recursion as

$$\log(a(T) + 1) = \sum_{i=1}^{k} \log(1 + a(T_i)) + \log(1 + a(T)^{-1}).$$

Hence $\log(a(T) + 1)$ can be seen as an additive parameter with toll function

$$f(T) = \log(1 + a(T)^{-1}).$$

This may seem useless, since the toll function itself depends on the parameter a. However, *a priori* estimates are sufficient to show that the toll function does indeed satisfy our conditions. It is clearly bounded, since $a(T) \geqslant 1$ for all trees T. Moreover, since $a(T)$ is exponentially large for most trees, $f(T)$ is small on average. To see why this is true, note that every set of leaves is an antichain. Hence if $\ell(T)$ is the number of leaves, we have $a(T) \geqslant 2^{\ell(T)}$ and thus

$$0 < f(T) = \log(1 + a(T)^{-1}) < 2^{-\ell(T)}.$$

We have already considered the number of leaves in the previous example and found that it is linear in $|T|$ on average, hence it follows from this inequality that $f(T)$ is exponentially small on average. Let us discuss the details in the special case of pruned binary trees ($\Phi(t) = (1 + t)^2$); other simply generated families can be treated in the same way. The bivariate generating function, where u marks the number of leaves, is

$$\sum_{T} x^{|T|} u^{\ell(T)} = \frac{1 - 2x - \sqrt{(1 - 2x)^2 - 4x^2 u}}{2x}.$$

For our purposes, we need the special case $u = 1/2$:

$$\sum_{|T|=n} |f(T)| \leqslant \sum_{|T|=n} 2^{-\ell(T)} = [x^n] \frac{1 - 2x - \sqrt{1 - 4x + 2x^2}}{2x}.$$

The dominating singularity is $1 - 1/\sqrt{2}$. Singularity analysis yields

$$[x^n] \frac{1 - 2x - \sqrt{1 - 4x + 2x^2}}{2x} \sim \frac{\sqrt{1 + \sqrt{2}}}{2\sqrt{\pi} n^{3/2}} \cdot (2 + \sqrt{2})^n.$$

Since the number of pruned binary trees of order n is

$$\frac{1}{n+1} \binom{2n}{n} \sim \pi^{-1/2} n^{-3/2} 4^n,$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

we obtain

$$\frac{\sum_{|T|=n} f(T)}{\sum_{|T|=n} 1} = O\left(\left(\frac{2+\sqrt{2}}{4}\right)^n\right),$$

which proves that the conditions of Theorem 2.1 are satisfied. The difference between $\log a(T)$ and $\log(a(T)+1)$ is of course small, so it follows that the number of antichains is asymptotically log-normally distributed for pruned binary trees (and other simply generated families as well by the same argument). The numerical values of μ and σ^2 for pruned binary trees are

$$\mu \approx 0.272, \quad \sigma^2 \approx 0.034.$$

Remark 3. It should be mentioned how the numerical values are computed. The series for the mean and the variance both converge at an exponential rate, so their values can be determined quite accurately from only a few terms. We compute the values of $a(T)$ for trees of small order (up to about 15 to 20 vertices) explicitly and ignore all other trees in the expressions for mean and variance (or replace them by upper and lower bounds). We can make the estimates for the toll function effective to bound the resulting error. However, it is quite difficult to compute the constants with higher accuracy than just a few digits, since explicitly calculating $a(T)$ for all trees up to a certain order is only feasible if this order is small. All numerical values here and in the following are given to the highest accuracy that we were able to obtain by means of this method.

The number of subtrees is closely related to the number of antichains. Indeed, there is a trivial bijection between antichains and subtrees that contain the root: for any subtree of a rooted tree that contains the root, the leaves form an antichain (we only count the root as a leaf in this context if it is the only vertex of the subtree), and this can easily be reversed.

The enumeration of subtrees in specific families of simply generated trees has been studied quite extensively by Meir and Moon [27], Baron and Drmota [2] and Moon [31]. There are even some nice exact counting formulas in this context, but no results on the distribution so far. It turns out that the distributions of the number of antichains and the number of subtrees are (upon taking the logarithm) essentially identical. If $s(T)$ denotes the number of subtrees, then clearly $a(T) \leqslant s(T)$ by the bijection between antichains and subtrees containing the root described above. On the other hand, we also have $s(T) \leqslant |T|a(T)$, as can be seen by another simple argument. Any subtree of T is uniquely characterized by its root (the closest vertex to the root of T) and its leaves (again counting the root only as a leaf if it is the only vertex). The set of leaves can be any antichain, and there are at most $|T|$ choices for the subtree root, so the inequality follows immediately.

We conclude that $\log s(T) = \log a(T) + O(\log |T|)$, which means that the central limit theorem carries over to the number of subtrees, with the same constants: $\mu \approx 0.272$ and $\sigma^2 \approx 0.034$ for pruned binary trees. For labelled trees, the values $\mu \approx 0.35$ and $\sigma^2 \approx 0.04$ have already been determined in [37].

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

2.3.3. The number of maximal antichains. This example is also taken from Klazar's paper [23]; an antichain is maximal if it is not a proper subset of another antichain. Let $m(T)$ denote the number of maximal antichains of a tree T. The parameter m can be computed recursively from the branches $T_1, T_2, \dots, T_k$ as follows:

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

since a maximal antichain is either a collection of maximal antichains in all the branches or consists of the root only. As in the previous example, we can rewrite this as

$$\log m(T) = \sum_{i=1}^{k} \log m(T_i) - \log(1 - m(T)^{-1}),$$

which means that $\log m(T)$ is an additive parameter with toll function

$$f(T) = \begin{cases} 0 & T = \bullet, \\ -\log(1 - m(T)^{-1}) & \text{otherwise.} \end{cases}$$

Theorem 2.1 shows that the limiting distribution is again a log-normal law; the technical conditions can be verified in the same way as in the previous section (in place of the number of leaves, one can use the number of full subtrees of order 2 in the argument). The constants μ and σ^2 in this example are (for pruned binary trees)

$$\mu \approx 0.175, \quad \sigma^2 \approx 0.03.$$

2.3.4. The shape guessing game. The purpose of this (perhaps not very serious) example is to show that our theorems can deal with toll functions that are defined in a rather complicated way. Let a plane tree T be given. A player plays the following guessing game: she tries to predict the precise shape of T. If she succeeds, her score is the size of the tree. If not, she is given the root degree and the sizes of the branches, and repeats the guessing game for each of the branches. She receives the sum of the branch scores as her total score. It is clear that some trees (*e.g.*, a star) are much more easily 'guessable' in this way than others (*e.g.*, a path). What can be said about the distribution of the expected score?

Let S denote the expected total score, and let s be the associated toll function. Moreover, we write t_n for the total number of plane trees of order n. For a tree T of order n with branches $T_1, T_2, \dots, T_k$, we have

$$s(T) = \frac{1}{t_n}\left(n - \sum_{i=1}^{k} S(T_i)\right),$$

since the probability of guessing correctly is $1/t_n$, in which case the score is n rather than the otherwise expected $\sum_{i=1}^{k} S(T_i)$. Clearly, $0 < s(T) < n/t_n$, and since t_n grows exponentially, the conditions of Theorem 2.1 are satisfied. We find that the expected score associated with a tree is asymptotically normally distributed with mean $\sim 0.6698n$ and variance $\sim 0.1193n$.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

3. Pólya trees and similar families

3.1. Pólya trees

Our technique also applies to the family of Pólya trees. Once again, we define a bivariate generating function

$$Y(x,u) = \sum_T w(T) x^{|T|} u^{F(T)},$$

where the sum is over all Pólya trees T. The analogue of the functional equation (2.1) is now

$$x \exp\left(\sum_{k=1}^{\infty} \frac{1}{k} Y(x^k, u^k)\right) = \sum_T x^{|T|} u^{F(T)-f(T)}. \tag{3.1}$$

The key observation in the analysis of Pólya trees is the fact that only the first term in the infinite series matters asymptotically, since the rest of the sum has a larger radius of convergence. This allows us to prove an analogue of Theorem 2.1 for Pólya trees.

Theorem 3.1. *Let t_n denote the number of Pólya trees of order n. Assume that the toll function f is bounded and satisfies*

$$\frac{\sum_{|T|=n} |f(T)|}{t_n} = O(c^n)$$

for a constant $c \in (0,1)$, the sum being over all Pólya trees of order n.

Let T_n denote a random Pólya tree of order n. The mean $\mu_n = \mathbb{E}(F(T_n))$ of the parameter F is asymptotically

$$\mu_n = \mu n + O(1),$$

where the constant μ is given by

$$\mu = \frac{\sum_T \left(f(T)\rho^{|T|} + F(T) \dfrac{\rho^{2|T|}}{1-\rho^{|T|}} \right)}{1 + \sum_{n=1}^{\infty} n t_n \dfrac{\rho^{2n}}{1-\rho^n}}.$$

The variance $\sigma_n^2 = \mathbb{V}(F(T_n))$ of F is asymptotically

$$\sigma_n^2 = \sigma^2 n + O(1)$$

for a constant $\sigma^2 \geqslant 0$. Moreover, if $\sigma^2 \neq 0$, the renormalized random variable

$$\frac{F(T_n) - \mu_n}{\sigma_n}$$

converges weakly to a standard normal distribution.

Remark 4. The most remarkable difference from Theorem 2.1 is perhaps the formula for the constant μ. An explicit expression can be given for the variance as well, but it is rather long and complicated.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

Proof. One can solve the equation (3.1) ‘explicitly’ for $Y(x,u)$: setting

$$U(x,u)=\sum_{k=2}^{\infty}\frac{1}{k}Y(x^k,u^k)$$

and

$$H(x,u)=\sum_{T}x^{|T|}u^{F(T)}(1-u^{-f(T)}),$$

we have

$$Y(x,u)=R(xe^{H(x,u)+U(x,u)})+H(x,u)=-W(-xe^{H(x,u)+U(x,u)})+H(x,u), \tag{3.2}$$

where

$$R(x)=\sum_{n=1}^{\infty}\frac{n^{n-1}}{n!}x^n$$

is the exponential generating function for rooted labelled trees, and $W(x)$ is the closely related Lambert W-function. They satisfy the functional equations $R(x)=x\exp(R(x))$ and $x=W(x)\exp(W(x))$ respectively. It is well known that $R(x)$ has a square-root singularity at $x=1/e$ with $R(1/e)=1$, which carries over to $Y(x,u)$ – in particular, $Y(x,1)$ has a singularity at $\rho\approx 0.33832185$ with $Y(\rho,1)=1$.

The treatment of mean and variance is analogous to simply generating trees: differentiating (3.1) with respect to u and plugging in $u=1$, we obtain

$$Y(x,1)\Bigg(Y_u(x,1)+\sum_{k\geqslant 2}Y_u(x^k,1)\Bigg)=Y_u(x,1)-\sum_{T}f(T)x^{|T|},$$

and thus

$$Y_u(x,1)=\frac{\sum_T f(T)x^{|T|}+Y(x,1)\sum_{k\geqslant 2}Y_u(x^k,1)}{1-Y(x,1)}.$$

Likewise,

$$xY_x(x,1)=\frac{Y(x,1)\big(1+\sum_{k\geqslant 2}x^kY_x(x^k,1)\big)}{1-Y(x,1)}$$

and thus

$$Y_u(x,1)=xY_x(x,1)\cdot\frac{\sum_T f(T)x^{|T|}+Y(x,1)\sum_{k\geqslant 2}Y_u(x^k,1)}{Y(x,1)\big(1+\sum_{k\geqslant 2}x^kY_x(x^k,1)\big)}.$$

The value of the fraction at $x=\rho$ is exactly the constant μ as stated in the theorem (as can be seen by some elementary manipulations, also making use of the fact that $Y(\rho,1)=1$). The asymptotic formula for the mean can now be obtained by another application of singularity analysis.

The variance is again treated in a similar fashion, and the limiting distribution follows either from the explicit representation (3.2) by direct singularity analysis and Hwang’s quasi-power theorem (see [18] or [15, Theorem IX.8]), or by means of Lemma 2.2 again. □

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

3.1.1. Antichains and subtrees in Pólya trees. Let us now apply our general result to the example of the number of antichains and subtrees in Pólya trees. The technical conditions on the toll function are satisfied for the same reason, and the step from antichains/subtrees containing the root to all subtrees can be performed in the same way. The result is essentially the same: the logarithm of the number of antichains (or the number of subtrees that contain the root) asymptotically follows a Gaussian distribution, and the mean and variance are of linear order with constants that differ only slightly from labelled trees: $\mu \approx 0.38$, $\sigma^2 \approx 0.04$. The same is true for the number of all subtrees (that do not necessarily contain the root) for the same reason as for simply generated trees (see Section 2.3.2). It can also be carried over easily to unrooted (free) trees.

3.2. Homeomorphically irreducible trees and a question of Vince and Wang

The same techniques apply to other similar classes of trees as well. As an example, let us consider *homeomorphically irreducible* trees, *i.e.*, trees that do not have vertices of degree 2, motivated by a question posed by Vince and Wang [36]. They were considering the average subtree order of trees (the average subtree order, first studied by Jamison [20], is the arithmetic mean of the orders of all subtrees), and asked for the average of this parameter over all homeomorphically irreducible trees (which play a special role in this context). They mention that Meir and Moon proved the average to be $1-e^{-1} \approx 0.6321$ for labelled trees. However, this is only true under a rather unintuitive probabilistic model that involves taking all subtrees of all labelled trees of given order and drawing one of them randomly. It seems much more natural to determine the mean subtree order of each tree first and to take the average of all the mean subtree orders, which makes a small yet noticeable difference.

The aim of this section is to provide an answer to the question of Vince and Wang. As already outlined in [37], the average subtree order is also an additive parameter that satisfies our conditions.

The enumeration of homeomorphically irreducible trees is nicely described in [17, Section 3.3]. First of all, we have to consider Pólya trees with the property that no vertex has outdegree 1. The functional equation only changes slightly:

$$x \exp\left(\sum_{k=1}^{\infty} \frac{1}{k} Y(x^k, u^k)\right) - xY(x,u) = \sum_T x^{|T|} u^{F(T)-f(T)},$$

so Lemma 2.2 is still applicable. Now let us verify that the average subtree order does indeed satisfy the required conditions. For a rooted tree T with branches $T_1, T_2, \ldots, T_k$, we write $a(T)$ for the number of subtrees that contain the root (which, as we know, is equal to the number of antichains) and $b(T)$ for the sum of their orders. Then

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

as in Section 2.3.2, and

$$b(T) = a(T) + \sum_{i=1}^{k} b(T_i) \prod_{j \neq i} (1 + a(T_j)).$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

The first term takes the root into account, and the ith term in the sum counts the total number of vertices in branch T_i in all subtrees of T that contain the root. This simplifies to

$$\frac{b(T)}{a(T)} = 1 + \sum_{i=1}^{n} \frac{b(T_i)}{1 + a(T_i)}.$$

Thus the fraction $b(T)/(a(T)+1)$ is an additive parameter with toll function

$$f(T) = 1 - \frac{b(T)}{a(T)} + \frac{b(T)}{1 + a(T)} = 1 - \frac{b(T)}{a(T)(1 + a(T))}.$$

Clearly, $b(T)/a(T) \leqslant |T|$. Recall also that $a(T) \geqslant 2^{\ell(T)}$, where $\ell(T)$ is the number of leaves. For trees without vertices of outdegree 1, the number of leaves is at least $(|T|+1)/2$, and therefore $f(T)-1$ is exponentially small. By Remark 2, this means that our auxiliary function $b(T)/(a(T)+1)$ satisfies the necessary conditions and thus a central limit theorem.

The difference between $b(T)/a(T)$ and $b(T)/(a(T)+1)$ is exponentially small and thus irrelevant for the distribution. However, so far we have only taken into account subtrees that contain the root. To complete the argument, we have to make use of an idea that was outlined in [37]: a random unrooted homeomorphically irreducible tree has almost surely a unique centroid, and this centroid has at least three branches of linear order (see [29,30] on distributional properties of centroids and their branches; homeomorphically irreducible trees are not specifically mentioned there, but the results are the same). Because of that, only a proportion of the subtrees that is exponentially small does not contain the centroid, so that all others can be ignored. Conditioned on the size of the centroid branches, each of them can be regarded as a random rooted (Pólya) tree without vertices of outdegree 1, and the branches are independent of each other. Thus the mean subtree order in the whole tree is essentially (up to a small error term) the sum of the mean subtree orders in the centroid branches, and these summands are all independent. Since the convolution of independent Gaussian random variables is still Gaussian, this completes the argument.

Summing up: the mean subtree order in a random homeomorphically irreducible tree is asymptotically normally distributed with mean and variance of linear order. The answer to the question of Vince and Wang is that the average of the mean subtree order over all homeomorphically irreducible tree is asymptotically $\sim \mu n$, where $\mu \approx 0.625$. This should be compared to ordinary Pólya trees (or also unrooted trees), where the mean is slightly higher ($\mu \approx 0.648$).

4. Recursive trees and binary search trees

A result analogous to Theorem 2.1 also holds for recursive trees. It is somewhat more complicated to extend it to more general families of increasing trees [3], since the singularity type of the generating function strongly depends on the family of trees considered. We only treat the case of recursive trees and binary search trees in more detail, two important special cases that have been studied extensively. Let us begin with a theorem for recursive trees.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

Theorem 4.1. *Assume that the toll function f is bounded and satisfies*

$$\frac{\sum_{|T|=n} |f(T)|}{(n-1)!} = O(c^n)$$

for a constant $c \in (0,1)$, the sum being over all recursive trees T of order n.

Let T_n denote a random recursive tree of order n. The mean $\mu_n = \mathbb{E}(F(T_n))$ of the parameter F is asymptotically

$$\mu_n = \mu n + O(\alpha^n)$$

for any $\alpha \in (c,1)$, where the constant μ is given by

$$\mu = \sum_T \frac{f(T)}{(|T|+1)!}.$$

The variance $\sigma_n^2 = \mathbb{V}(F(T_n))$ of F is asymptotically

$$\sigma_n^2 = \sigma^2 n + O(\alpha^n),$$

again for any $\alpha \in (c,1)$. Here, the constant σ^2 is given by

$$\sigma^2 = \sum_T \frac{f(T)(2F(T)-f(T))}{(|T|+1)!} + \sum_{T_1} \sum_{T_2} \frac{f(T_1)f(T_2)}{(|T_1|+1)!(|T_2|+1)!} \left(\frac{|T_1||T_2|}{|T_1|+|T_2|+1} - |T_1| - |T_2| \right),$$

all sums being over all recursive trees. Moreover, if $\sigma^2 \neq 0$, then the renormalized random variable

$$\frac{F(T_n) - \mu_n}{\sigma_n}$$

converges weakly to a standard normal distribution.

Remark 5. The strong, exponentially small error terms in Theorem 4.1 and later in Theorem 4.2 are quite remarkable: they stem from the fact that the generating functions have pole singularities rather than square-root singularities.

Once again, we make use of a bivariate generating function, which is now necessarily an exponential generating function

$$Y(x,u) = \sum_T \frac{1}{|T|!} x^{|T|} u^{F(T)},$$

the sum being over all recursive trees. The analogue of (2.1), which is obtained in the same way, reads

$$\exp(Y(x,u)) = \sum_T \frac{1}{(|T|-1)!} x^{|T|-1} u^{F(T)-f(T)}. \tag{4.1}$$

For $u = 1$, this becomes the familiar differential equation

$$Y_x(x,1) = \exp(Y(x,1)).$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

4.1. Moments

Differentiating (4.1) with respect to u and plugging in $u = 1$, we obtain

$$\begin{aligned}\exp(Y(x,1))Y_u(x,1) &= \sum_T \frac{1}{(|T|-1)!}(F(T)-f(T))x^{|T|-1} \\ &= Y_{ux}(x,1) - \sum_T \frac{f(T)}{(|T|-1)!}x^{|T|-1},\end{aligned}$$

which is a linear differential equation in $Y_u(x,1)$. Now recall that the exponential generating function for recursive trees is $Y(x,1) = -\log(1-x)$. Then we are left with

$$Y_{ux}(x,1) = \frac{Y_u(x,1)}{1-x} + \sum_T \frac{f(T)}{(|T|-1)!}x^{|T|-1},$$

and the solution to this differential equation is given by

$$\begin{aligned}Y_u(x,1) &= \frac{1}{1-x}\int_0^x (1-v)\sum_T \frac{f(T)}{(|T|-1)!}v^{|T|-1}\,dv \\ &= \frac{1}{1-x}\sum_T \frac{f(T)}{(|T|+1)!}(|T|+1-|T|x)x^{|T|} \\ &= \frac{1}{1-x}\sum_T \frac{f(T)}{(|T|+1)!}x^{|T|} + \sum_T \frac{f(T)|T|}{(|T|+1)!}x^{|T|}.\end{aligned}$$

In the same way, we obtain

$$\begin{aligned}Y_{uu}(x,1) &= \frac{1}{1-x}\int_0^x \left(Y_u(v,1)^2 + (1-v)\sum_T \frac{f(T)}{(|T|-1)!}(2F(T)-f(T)-1)v^{|T|-1}\right)dv \\ &= \frac{x}{(1-x)^2}\left(\sum_T \frac{f(T)}{(|T|+1)!}x^{|T|}\right)^2 + \frac{1}{1-x}\int_0^x \left(\left(\sum_T \frac{f(T)|T|}{(|T|+1)!}v^{|T|}\right)^2\right. \\ &\quad \left. + (1-v)\sum_T \frac{f(T)}{(|T|-1)!}(2F(T)-f(T)-1)v^{|T|-1}\right)dv.\end{aligned}$$

In view of our assumptions on the toll function, the radius of convergence of the series

$$\sum_T \frac{f(T)|T|}{(|T|+1)!}x^{|T|}$$

is at least $c^{-1} > 1$, so it represents an analytic function in the circle around 0 of radius c^{-1}, in particular at 1. The same holds for all the series over all recursive trees in the two formulas above. Singularity analysis (in the meromorphic setting, thus with strong error term; see [15, Theorem IV.10]) now shows that

$$[x^n]Y_u(x,1) = \sum_T \frac{f(T)}{(|T|+1)!} + O(\alpha^n)$$

for every $\alpha \in (c,1)$, so we also have

$$\mu_n = \mu n + O(\alpha^n)$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

for every $\alpha \in (c, 1)$, where

$$\mu = \sum_T \frac{f(T)}{(|T|+1)!}.$$

Applying singularity analysis to $Y_{uu}(x,1)$ as well, we obtain an asymptotic formula for the variance:

$$\sigma_n^2 = \sigma^2 n + O(\alpha^n)$$

for every $\alpha \in (c, 1)$, where

$$\sigma^2 = \sum_T \frac{f(T)(2F(T)-f(T))}{(|T|+1)!} + \sum_{T_1}\sum_{T_2} \frac{f(T_1)f(T_2)}{(|T_1|+1)!(|T_2|+1)!}\left(\frac{|T_1||T_2|}{|T_1|+|T_2|+1} - |T_1| - |T_2|\right).$$

4.2. Limiting distribution

In order to prove the convergence to a limiting distribution, we split the right-hand side of (4.1) in very much the same way as we did in the case of simply generated families of trees: we have

$$\exp(Y(x,u)) = Y_x(x,u) - \sum_T \frac{1}{(|T|-1)!} x^{|T|-1} u^{F(T)}(1-u^{-f(T)}).$$

Now define

$$H(x,u) = \sum_T \frac{1}{|T|!} x^{|T|} u^{F(T)}(1-u^{-f(T)}),$$

so that

$$\frac{\partial}{\partial x}(Y(x,u) - H(x,u)) = \exp(Y(x,u)).$$

$H(x,u)$ is analytic (as a function of x) in a larger region than $Y(x,u)$ if u lies in a suitable neighbourhood of 1, by the same arguments that we used for simply generated families of trees. If we substitute $U(x,u) = Y(x,u) - H(x,u)$, we are left with

$$U_x(x,u) = \exp(U(x,u) + H(x,u))$$

and $U(0,u) = 0$. The solution to this differential equation is given by

$$U(x,u) = -\log\left(1 - \int_0^x \exp(H(v,u))\,dv\right).$$

Obviously, $H(v,1) = 0$, so we get $U(x,1) = Y(x,1) = -\log(1-x)$, as it should be. Note that the type of the singularity is still logarithmic, and the dominating singularity is located at the point $\rho = \rho(u)$, for which

$$\int_0^\rho \exp(H(v,u))\,dv = 1.$$

Note also that

$$\frac{d}{dx}\int_0^x \exp(H(v,u))\,dv = \exp(H(x,u)) \neq 0,$$

hence if u is restricted to a suitable neighbourhood around 1, then ρ is unique and analytic as a function of u. The asymptotic expansion of U around the dominating singularity ρ is found as follows:

$$\begin{aligned} U(x,u) &= -\log\left(1-\int_0^x \exp(H(v,u))\,dv\right) \\ &= -\log\left(\int_x^\rho \exp(H(v,u))\,dv\right) \\ &= -\log\bigl(\exp(H(\rho,u))(\rho-x)+O(|\rho-x|^2)\bigr) \\ &= -\log(\rho-x)-H(\rho,u)+O(|x-\rho|). \end{aligned}$$

Using singularity analysis and the quasi-power theorem once again, we obtain the desired central limit theorem.

4.2.1. Subtrees of recursive trees. As an application, let us again consider the number of subtrees as an example. The technical conditions are satisfied for the same reason as in Section 2.3.2. Once again, we find that the distribution of the number of subtrees is asymptotically log-normal. Plugging into the formulas for mean and variance, we arrive at the numerical values $\mu \approx 0.4505$ and $\sigma^2 \approx 0.017$, which shows that recursive trees tend to have more subtrees than labelled trees, binary trees or Pólya trees (it is well known that they are flatter and wider and thus more star-like than Galton–Watson trees; stars have the largest number of subtrees among all trees of given order) and that the distribution is more concentrated.

4.3. Binary search trees

Let us finally study binary search trees (which are also equivalent to binary increasing trees). Again, a completely analogous theorem holds.

Theorem 4.2. *Assume that the toll function f is bounded and satisfies*

$$\frac{\sum_{|T|=n}|f(T)|}{n!} = O(c^n)$$

for a constant $c \in (0,1)$, the sum being over all binary search trees T of order n.

Let T_n denote a random binary search tree of order n. The mean $\mu_n = \mathbb{E}(F(T_n))$ of the parameter F is asymptotically

$$\mu_n = \mu(n+1) + O(\alpha^n)$$

for any $\alpha \in (c,1)$, where the constant μ is given by

$$\mu = \sum_T \frac{2f(T)}{(|T|+2)!}.$$

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

The variance $\sigma_n^2 = \mathbb{V}(F(T_n))$ *of* F *is asymptotically*

$$\sigma_n^2 = \sigma^2(n+1) + O(\alpha^n),$$

again for any $\alpha \in (c,1)$. *Here, the constant* σ^2 *is given by*

$$\sigma^2 = \sum_T \frac{2f(T)(2F(T)-f(T))}{(|T|+2)!} - \mu^2 + \sum_{T_1}\sum_{T_2} \frac{4f(T_1)f(T_2)}{(|T_1|+2)!(|T_2|+2)!} \\ \times \left(\frac{|T_1||T_2|}{|T_1|+|T_2|+1} - |T_1| - |T_2| + \frac{|T_1||T_2|}{(|T_1|+|T_2|+2)(|T_1|+|T_2|+3)} \right. \\ \left. + \frac{|T_1|^2|T_2|^2}{(|T_1|+|T_2|+1)(|T_1|+|T_2|+2)(|T_1|+|T_2|+3)} \right),$$

all sums being over all binary search trees. Moreover, if $\sigma^2 \neq 0$, *then the renormalized random variable*

$$\frac{F(T_n) - \mu_n}{\sigma_n}$$

converges weakly to a standard normal distribution.

Proof. For binary search trees, the approach is similar to recursive trees, albeit slightly different. The treatment of mean and variance is fully analogous, so we focus on the limiting distribution. Once again, we consider the bivariate generating function $Y(x,u)$, which satisfies the equation

$$(Y(x,u)+1)^2 = \sum_T \frac{1}{(|T|-1)!} x^{|T|-1} u^{F(T)-f(T)}.$$

It is convenient to work with $U(x,u) = Y(x,u)+1$, which satisfies the differential equation

$$U(x,u)^2 = U_x(x,u) + \sum_T \frac{1}{(|T|-1)!} x^{|T|-1} u^{F(T)}(u^{-f(T)} - 1).$$

Let us denote the sum on the right-hand side by $H(x,u)$; once again, if u is restricted to a suitable neighbourhood of 1, this function is analytic (as a function of x) in a larger region than $U(x,u)$. The differential equation is of Riccati type, and we solve it by means of the substitution

$$U(x,u) = -\frac{V_x(x,u)}{V(x,u)}.$$

The differential equation then simplifies to

$$\frac{V_{xx}(x,u)}{V(x,u)} = H(x,u)$$

or

$$V_{xx}(x,u) = H(x,u)V(x,u). \tag{4.2}$$

Recall that by our conditions on the toll function, the coefficients $h_n = h_n(u) = [x^n]H(x,u)$ satisfy $h_n = O(\alpha^n)$ for some $\alpha < 1$ uniformly in u if u is restricted to a suitable neighbourhood of 1. Without loss of generality, we can choose $v_0 = [x^0]V(x,u) = 1$, and since

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

$U(0,u) = 1$, we must have $v_1 = [x^1]V(x,u) = -1$. The remaining coefficients $v_2, v_3, \ldots$ of $V(x,u)$ can be calculated recursively from (4.2):

$$v_j = \frac{1}{j(j-1)} \sum_{i=0}^{j-2} h_i v_{j-2-i}.$$

An easy induction shows that $v_j = O(\alpha^n)$, uniformly in u. Thus we can write

$$Y(x,u) = -1 - \frac{V_x(x,u)}{V(x,u)},$$

where $V(x,u)$ is analytic in a circle of radius $\alpha^{-1} > 1$ around 0, uniformly in u, and $V(x,1) = 1 - x$. Hence the dominating singularity of $Y(x,u)$ is a simple pole at a point $\rho = \rho(u)$, which is the solution to

$$V(x,u) = 0.$$

Note in particular that $\rho(1) = 1$. Once again, we can apply singularity analysis and the quasi-power theorem now to conclude the proof of the central limit theorem. □

4.3.1. How often does Quicksort encounter a sorted list?. As an example of an additive tree parameter for binary search trees, we consider a question that can also be interpreted in terms of the Quicksort algorithm. Binary increasing trees can serve as a model for Quicksort in the following way: the root stands for the pivot, the left and right branches stand for the two sublists to which Quicksort is applied recursively.

Given a random permutation to which Quicksort is applied, we are interested in the number of times that Quicksort is called in the process with a perfectly sorted list as argument. Let us assume that the pivot is always the first element of the list. Then we can simply associate the binary search tree of order n in which each vertex has only a right child (we denote this tree by R_n) to a sorted list.

The toll function associated with our problem is

$$f(T) = \begin{cases} 1 & T = R_n \text{ for some } n, \\ 0 & \text{otherwise,} \end{cases}$$

and it clearly satisfies our condition. Thus we obtain a central limit theorem for our parameter, and the constants are easily calculated as

$$\mu = 2e - 5 \approx 0.43656366 \quad \text{and} \quad \sigma^2 = \frac{55 - 12e - 3e^2}{2} \approx 0.10672488.$$

5. Conclusion

We have seen that a central limit theorem holds for rather general additive parameters in various classes of trees, provided that the associated toll function is (at least on average) very small as the size of the trees goes to infinity. The assumption of exponential decay that we made throughout this paper is still rather strong and can probably be replaced by a much weaker condition. It is not clear, however, where exactly to draw the line between

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

Gaussian (as in all our examples) and non-Gaussian (as for example for the path length of a tree) limiting distributions.

Such results partly exist. Fill and Kapur [12] state sufficient conditions for asymptotic normality in the binary case if the toll function only depends on the order of the tree. However, they mention that their arguments are somewhat heuristic. Hwang and Neininger [19] prove very strong results for the Quicksort recurrence (which is essentially equivalent to the binary search tree model), and they even allow the toll function to be random, but it may only depend on the size of the tree itself, not the whole tree. The fact that in our setting the toll function can depend on the whole tree means that dependences might be amplified, which could result in a shift in the boundary between Gaussian and non-Gaussian behaviour (and conceivably a region where both are possible). It would definitely seem worthwhile to pursue this question further.

References

[1] Aldous, D. (1991) Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.* **1** 228–266.

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

[3] Bergeron, F., Flajolet, P. and Salvy, B. (1992) Varieties of increasing trees. In *CAAP '92: Rennes 1992*, Vol. 581 of *Lecture Notes in Computer Science*, Springer, pp. 24–48.

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

[5] Devroye, L. (2002/2003) Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.* **32** 152–171.

[6] Dobrow, R. P. and Fill, J. A. (1999) Total path length for random recursive trees. *Combin. Probab. Comput.* **8** 317–333.

[7] Drmota, M. (2009) *Random Trees*, Springer.

[8] Drmota, M. and Gittenberger, B. (1999) The distribution of nodes of given degree in random trees. *J. Graph Theory* **31** 227–253.

[9] Drmota, M. and Gittenberger, B. (2010) The shape of unlabeled rooted random trees. *Europ. J. Combin.* **31** 2028–2063.

[10] Feng, Q., Mahmoud, H. M. and Panholzer, A. (2008) Phase changes in subtree varieties in random recursive and binary search trees. *SIAM J. Discrete Math.* **22** 160–184.

[11] Fill, J. A. (1996) On the distribution of binary search trees under the random permutation model. *Random Struct. Alg.* **8** 1–25.

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

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

[14] Flajolet, P., Gourdon, X. and Martínez, C. (1997) Patterns in random binary search trees. *Random Struct. Alg.* **11** 223–244.

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

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

[17] Harary, F. and Palmer, E. M. (1973) *Graphical Enumeration*, Academic Press.

[18] Hwang, H.-K. (1998) On convergence rates in the central limit theorems for combinatorial structures. *Europ. J. Combin.* **19** 329–343.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press

[19] 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.

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

[21] Janson, S. (2003) The Wiener index of simply generated random trees. *Random Struct. Alg.* **22** 337–358.

[22] Janson, S. (2005) Asymptotic degree distribution in random recursive trees. *Random Struct. Alg.* **26** 69–83.

[23] Klazar, M. (1997) Twelve countings with rooted plane trees. *Europ. J. Combin.* **18** 195–210.

[24] Klazar, M. (1997) Addendum: 'Twelve countings with rooted plane trees'. *Europ. J. Combin.* **18** 739–740.

[25] Mahmoud, H. M. (1991) Limiting distributions for path lengths in recursive trees. *Probab. Engrg Inform. Sci.* **5** 53–59.

[26] Meir, A. and Moon, J. W. (1978) On the altitude of nodes in random trees. *Canad. J. Math.* **30** 997–1015.

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

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

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

[30] Moon, J. W. (1985) On the expected distance from the centroid of a tree. *Ars Combin.* **20** (A) 263–276.

[31] Moon, J. W. (1997) On the number of induced subgraphs of trees. *Discrete Math.* **167/168** 487–496.

[32] Otter, R. (1948) The number of trees. *Ann. of Math.* (2) **49** 583–599.

[33] Robinson, R. W. and Schwenk, A. J. (1975) The distribution of degrees in a large random tree. *Discrete Math.* **12** 359–372.

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

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

[36] Vince, A. and Wang, H. (2010) The average order of a subtree of a tree. *J. Combin. Theory Ser. B* **100** 161–170.

[37] Wagner, S. (2012) Additive tree functionals with small toll functions and subtrees of random trees. In *23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms: AofA'12*, DMTCS Proc. AQ, pp. 67–80.

https://doi.org/10.1017/S0963548314000443 Published online by Cambridge University Press