Repeated fringe subtrees in random rooted trees*

Dimbinaina Ralaivaosaona and Stephan Wagner†

Abstract

A fringe subtree of a rooted tree is a subtree that consists of a node and all its descendants. In this paper, we are particularly interested in the number of fringe subtrees that occur repeatedly in a random rooted tree. Specifically, we show that the average number of fringe subtrees that occur at least r times is of asymptotic order $n/(\log n)^{3/2}$ for every $r \geq 2$ (with small periodic fluctuations in the main term) if a tree is taken uniformly at random from a simply generated family. Moreover, we also prove a strong concentration result for a related parameter: the size of the smallest tree that does not occur as a fringe subtree is with high probability equal to one of at most two different values. The main proof ingredients are singularity analysis, bootstrapping and the first and second moment methods.

1 Introduction

Subtrees of a rooted tree that consist of a node and all its descendants are known as *fringe subtrees*. Fringe subtrees are a classical object of study in the context of random trees, and there is a wealth of results for various random tree models, starting with the fundamental work of Aldous [1]; see [5, 6, 9, 13–15] for recent results on the distribution of fringe subtrees. Fringe subtrees and subtree patterns play an important role, for instance, in the study of phylogenetic trees in mathematical biology [2].

It is well-known that the number of occurrences of a fixed rooted tree as a fringe subtree of a uniformly random simply generated tree (among other random tree models) follows a Gaussian distribution with linear mean and variance, see e.g. [15, Corollary 1.8], and compare also Section 3.3 in [7].

In their recent work [3, 4], Christou et al. describe a linear-time algorithm for computing all "subtree repeats" in rooted ordered trees (see also the even more recent article [12], which deals with unordered trees) whose nodes may also have labels (taken from some finite alphabet). By a subtree repeat, they mean a collection of several (at least two) fringe subtrees that are isomorphic as ordered trees. This question parallels the problem of computing repetitions in strings and is related to the common subexpression problem that arises e.g. in compiler design. For further applications in, among others, the implementations of functional programming languages, automatic theorem proving, and computational biology, let us refer to the references given in [4].

Several natural questions arise in this context: how many subtree repeats is the algorithm going to find on average, i.e., how many different fringe subtrees occur more than once in a random tree? Here, we will always assume that a tree is chosen uniformly at random from some given family of trees (usually a simply generated family, see the discussion in Section 2). More generally, how many fringe subtrees occur at least r times, where r is some fixed positive integer? For example, the tree in Figure 1 has four distinct fringe subtrees, two of which (the trivial one and the two-node tree) are repeated.

For $r = 1$ (i.e., the total number of distinct subtrees), this question was answered by Flajolet, Sipala and Steyaert [11]: their work was specifically motivated by the aforementioned common subexpression problem, which is based on the fact that a tree can be written in a more compact form as a directed acyclic graph by identifying nodes that are roots of identical fringe subtrees. The average

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

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

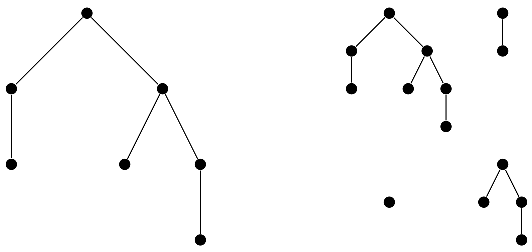

Figure 1: A tree (left) and its four distinct fringe subtrees (right).

number of distinct fringe subtrees is a measure of how much is saved in this way.

Flajolet, Sipala and Steyaert proved that, under very general assumptions, the number of distinct fringe subtrees is of order $n/\sqrt{\log n}$. For plane (rooted ordered) trees, their result reads as follows:

Theorem 1.1. (cf. [11]) *The average number of distinct fringe subtrees in a plane tree of order n taken uniformly at random is*

$$\sqrt{\frac{\log 4}{\pi}} \cdot \frac{n}{\sqrt{\log n}} + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right).$$

Their approach is based on generating functions; for later use, let us explain how such a generating function is obtained for plane trees. Recall that the generating function $R(x)$ for plane trees satisfies the functional equation

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

and that the number of such trees of order n is the Catalan number $\frac{1}{n}\binom{2n-2}{n-1}$. If we only want to count trees that do not contain a fixed tree S as a fringe subtree, then the functional equation simply becomes

$$(1) \qquad R_S(x) = \frac{x}{1 - R_S(x)} - x^{|S|}.$$

Denoting the size (number of nodes) of S by $k = |S|$, we find that the solution is

$$R_S(x) = \frac{1}{2}\left(1 - x^k - \sqrt{1 - 4x + 2x^k + x^{2k}}\right).$$

Any fixed plane tree S occurs as a fringe subtree in all trees that are not counted by R_S; it follows that the generating function for the total number of distinct fringe subtrees is

$$\sum_S (R(x) - R_S(x)) = \sum_{k=1}^{\infty} \frac{1}{k}\binom{2k-2}{k-1} \Big(\frac{1 - \sqrt{1-4x}}{2} - \frac{1 - x^k - \sqrt{1 - 4x + 2x^k + x^{2k}}}{2}\Big).$$

Its expansion starts $x + 2x^2 + 5x^3 + 15x^4 + 48x^5 + 162x^6 + \cdots$. Our proof of Theorem 1.1 will not make use of this generating function (we will need (1) for other purposes, though); it will rather be based on distinguishing different cases for the number of nodes of a fringe subtree. As a side result of this proof, we find that the greatest contribution to the number of distinct fringe subtrees comes from subtrees whose number of nodes is of order $\log n$ (cf. Remark 3.1). More importantly, our proof serves as a template for our treatment of the number of repeated fringe subtrees. For this parameter, we obtain a similar asymptotic formula, but surprisingly some periodic fluctuations occur as well:

Theorem 1.2. *The average number of trees that occur more than once as fringe subtree in a random plane tree of order n is*

$$\psi(\log_4 n) \cdot \frac{n}{(\log n)^{3/2}} + \mathcal{O}\left(\frac{n}{(\log n)^{5/2}}\right),$$

where ψ is a continuous periodic function with period 1 that is bounded above and below by positive constants.

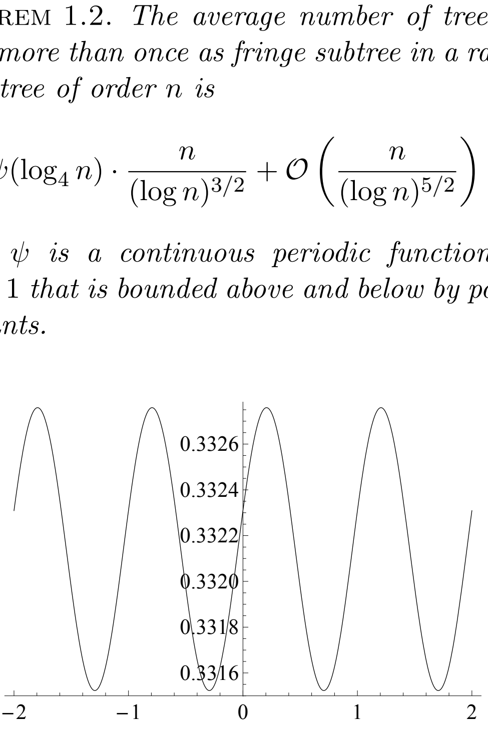

Figure 2: The periodic function ψ in Theorem 1.2.

The periodic function ψ is smooth (infinitely differentiable) and has small amplitude; see the remark on the generalized version (Theorem 5.1). Figure 2 shows a plot of the function.

As it was mentioned before, the number of occurrences of a specific fringe subtree satisfies a central limit theorem with mean and variance of linear order. Small trees thus occur almost surely; this raises another related question: what is the typical size of the smallest rooted ordered tree that does not occur as a fringe subtree of a randomly chosen rooted ordered tree T_n of order n? For this parameter, we will prove the following statement, which shows that there is a very sharp threshold:

THEOREM 1.3. *Let M_n be the (random) size of the smallest plane tree that does not occur as a fringe subtree of a random plane tree of order n. With high probability (i.e., probability tending to* 1*), M_n is either equal to $m_n = \lceil \log_4 \frac{2n}{\log n} \rceil$ or $m_n + 1$.*

We will even make this statement a little more precise, thereby also showing that typically M_n is concentrated at m_n only.

The paper is structured as follows: in the following section, we formulate the problem in more generality in the natural context of simply generated trees. Section 3 deals with the number of distinct fringe subtrees, where we re-prove the theorem of Flajolet, Sipala and Steyaert. In Section 4, we show the concentration result for the smallest missing subtree. A mix of the techniques required to prove these two results is needed to obtain Theorem 1.2, whose proof is sketched in Section 5. Remarks on labeled trees and Pólya trees in Section 6 conclude the paper.

2 Simply generated families of trees

Simply generated families of trees were first introduced by Meir and Moon [16], and there is a range of literature on their properties – see [7] and the references therein. They are a natural common generalization of plane trees, binary trees, labeled trees, and others, and they are also essentially equivalent to Galton-Watson trees. Therefore, they will provide the natural setup for our purposes. Recall that a simply generated family of trees is characterized by a sequence $\phi_0, \phi_1, \phi_2, \dots$ of nonnegative weights. Typically, $\phi_0 = 1$, and we will assume this (without actual loss of generality) in the following. Every plane tree T is then equipped with a weight given by

$$w(T) = \prod_{k \geq 0} \phi_k^{N_k(T)},$$

where $N_k(T)$ is the number of nodes in T whose outdegree is k. The weight generating function

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

where $|T|$ denotes the number of nodes of T again and the sum goes over all plane trees, then satisfies the crucial functional equation

$$Y(x) = x\Phi(Y(x))$$

with $\Phi(t) = \sum_{j \geq 0} \phi_j t^j$. Special cases that will be important to us include the following:

- Plane trees themselves correspond to the sequence of weights given by $\phi_j = 1$ for all j,
- d-ary trees are obtained for $\phi_j = \binom{d}{j}$,
- unary-binary trees are obtained by setting $\phi_0 = \phi_1 = \phi_2 = 1$ and $\phi_j = 0$ for $j > 2$,
- labeled trees are obtained for $\phi_j = \frac{1}{j!}$. In this case, the function Y is actually an exponential generating function. It turns out that this makes a difference for our particular problem.

The main asymptotic result regarding the enumeration of simply generated families of trees reads as follows:

THEOREM 2.1. (CF. [7, THEOREM 3.6]) *Let R be the radius of convergence of $\Phi(t) = \sum_{j \geq 0} \phi_j t^j$, 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 $\phi_j > 0$. The following asymptotic formula for the coefficients of $Y(x)$ holds, with $\rho = \tau/\Phi(\tau) = \Phi'(\tau)^{-1}$:*

$$t_n = [x^n]Y(x) = d\sqrt{\frac{\Phi(\tau)}{2\pi\Phi''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}} \left(1 + \mathcal{O}(n^{-1})\right)$$

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

We will now study the problem of determining the average number of distinct fringe subtrees of a random tree in the context of a simply generated family of trees. Note that it may depend on the specific family what is considered to be distinct. For example, two trees that are identical as plane trees may be considered distinct when regarded as binary trees. Figure 3 gives an example of the four different binary trees associated to one plane tree, which accordingly has a weight of 4 when the weights $\phi_j = \binom{2}{j}$ corresponding to the family of binary trees are used.

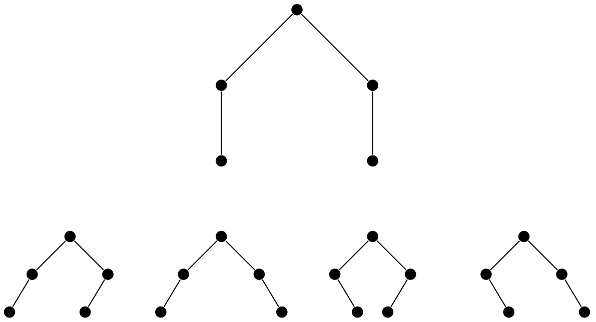

Figure 3: A plane tree and the four binary trees associated to it.

In the following, we make the assumption that all weights are integers and every possible fringe subtree simply corresponds to a weight of 1 unit, so the total number of trees of order n is just the coefficient of x^n in $Y(x)$, or the total weight of all plane trees of order n. This also implies that $\rho < 1$ in the setting of Theorem 2.1. Our assumption is natural not only for the simply generated family of plane trees, but also for instance for d-ary trees or unary-binary trees. Labeled trees are different, however, and as we will see, the results also differ slightly from other simply generated families. We will briefly deal with labeled trees (as well as with Pólya trees) in Section 6.

As a generalization of the functional equation (1), we now obtain the following equation that is satisfied by the generating function Y_k for trees which do not contain a specified fringe subtree of order k:

$$Y_k(x) = x\Phi(Y_k(x)) - x^k.$$

The slight change in the functional equation results in a shift of the dominating singularity, which we will make precise by means of bootstrapping later on.

3 The average number of distinct fringe subtrees

We are now prepared to present our alternative proof of the result of Flajolet, Sipala and Steyaert [11], which gives a general version of Theorem 1.1, in the setting of simply generated families of trees.

Theorem 3.1. *Suppose that a random tree T_n of order $n \equiv 1 \mod d$ is drawn from a simply generated family of trees satisfying the conditions of Theorem 2.1, and write $b = \rho^{-1}$ and $c = \sqrt{\Phi(\tau)/(2\pi\Phi''(\tau))}$, so that the coefficients of Y_n satisfy $t_n \sim cdn^{-3/2}b^n$. Then the average number of distinct fringe subtrees of T_n is asymptotically given by*

$$\frac{2c}{\tau} \cdot \frac{n}{\sqrt{\log_b n}} + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right).$$

Proof. The proof proceeds by distinguishing two different cases for the sizes of the fringe subtrees. In the following, we assume for simplicity that $d = 1$ in Theorem 2.1.

Small trees. Let us start with fringe subtrees whose order is at most $k_1 = \log_b n$. Their total contribution is clearly at most

$$\sum_{k \le k_1} t_k = \mathcal{O}\left(k_1^{-3/2} b^{k_1}\right) = \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right)$$

by Theorem 2.1 and can therefore be neglected.

Large trees. Now we consider fringe subtrees whose order is greater than $k_1 = \log_b n$. We first count the total number of occurrences of such fringe subtrees. This means some overcounting (as the same fringe subtree can occur multiple times), which we will correct for later on.

To this end, we note that every occurrence of a fringe subtree of order k can be generated by taking a tree of order $n - k + 1$ and replacing one of the leaves with a fringe subtree of order k. Let ℓ_r be the average number of leaves of a random tree of order r from our specific family of trees. It is well known (see [7, Theorem 3.13]) that $\ell_r = Lr + \mathcal{O}(1)$ as $r \to \infty$, where the constant $L = \rho/\tau = 1/(b\tau)$ (in the notation of Theorem 2.1) depends on the family of trees. The total number of occurrences of fringe subtrees of order $k > k_1$ is now equal to

$$\sum_{k_1 < k \leq n} t_{n-k+1} \ell_{n-k+1} t_k. \tag{2}$$

Once again by Theorem 2.1, the sum over all $k \geq n/2$ is

$$\begin{aligned}
&\mathcal{O}\left(\sum_{n/2 \leq k \leq n} b^{n-k+1}(n-k+1)^{-1/2} \cdot b^k k^{-3/2}\right) \\
&\quad = \mathcal{O}\left(b^n n^{-3/2} \sum_{n/2 \leq k \leq n} (n-k+1)^{-1/2}\right) \\
&\quad = \mathcal{O}\left(t_n \sqrt{n}\right),
\end{aligned}$$

so the contribution to the average number of distinct fringe subtrees is $\mathcal{O}(\sqrt{n})$ and thus again negligible. The remaining sum is

$$\begin{aligned}
&\sum_{k_1 < k < n/2} c b^{n-k+1} \cdot L(n-k+1)^{-1/2} \cdot c b^k k^{-3/2} \\
&\quad \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right) \\
&\quad = bcLn^{3/2} t_n \sum_{k_1 < k < n/2} (n-k+1)^{-1/2} k^{-3/2} \\
&\qquad \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right) \\
&\quad = bcLnt_n \sum_{k > k_1} k^{-3/2} \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right) \\
&\quad = \frac{2c}{\tau} n t_n k_1^{-1/2} \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right),
\end{aligned}$$

which means that the contribution to the average number of distinct fringe subtrees is

$$\frac{2cn}{\tau\sqrt{k_1}} \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right). \tag{3}$$

This already provides us with the main term in Theorem 3.1. We only have to prove that we did not overcount too much.

We will therefore also count the total number of occurrences of a pair of two identical fringe subtrees. If a certain fringe subtree occurs m times, then this gives us an overcount of $m-1$, which is less than the number of pairs (namely $\binom{m}{2}$) that can be formed from them. Thus the total number of pairs of identical fringe subtrees provides an upper bound for the overcount. We remark that this also follows from the inclusion-exclusion principle.

The number of pairs of identical fringe subtrees can be determined in a similar fashion: each such pair is obtained from a tree of order $n-2k+2$ by replacing two leaves with the same fringe subtree of order k. Note that k has to be less than $n/2$ and that there are clearly less than $(n-2k+2)^2$ choices for the two leaves, so the number (total weight) of pairs of identical fringe subtrees of order $k > k_1$ is at most

$$\sum_{k_1 < k < n/2} (n-2k+2)^2 t_{n-2k+2} t_k. \tag{4}$$

Again by Theorem 2.1, this is

$$\begin{aligned}
&\mathcal{O}\left(\sum_{k_1 < k < n/2} b^{n-2k+2}(n-2k+2)^{1/2} b^k k^{-3/2}\right) \\
&\quad = \mathcal{O}\left(\sqrt{n} b^n \sum_{k > k_1} b^{-k} k^{-3/2}\right) \\
&\quad = \mathcal{O}\left(n^2 t_n b^{-k_1} k_1^{-3/2}\right) = \mathcal{O}\left(\frac{nt_n}{(\log n)^{3/2}}\right).
\end{aligned}$$

Thus the error made by overcounting in (3) is also negligible, which concludes our proof.

REMARK 3.1. *The proof shows that the main contribution comes from fringe subtrees of order $A \log_b n$ with $A \geq 1$. Analyzing the contributions a bit more carefully, one also finds that the contributions to the average, regarded as a function of the order k, follow a bimodal distribution, with a peak at $k \approx \log_b n$ and a second peak at $k = n$.*

REMARK 3.2. *It is tempting to conjecture that the number of distinct fringe subtrees asymptotically follows a Gaussian distribution. However, at this stage we do not even have a conjecture for the* *asymptotic behavior of the variance in view of the rather nontrivial dependencies between the possible fringe subtrees.*

4 The smallest missing subtree

In this section we settle the question regarding the smallest missing subtree by proving that there is a sharp threshold k^* depending on n (and the specific family of trees) such that with high probability, all trees of order $k \leq k^*$ occur as fringe subtrees of a random tree T_n of order n, while for every $k \geq k^* + \epsilon$, there is a tree of order k that is not present as a fringe subtree.

If the weight ϕ_1 associated with outdegree 1 is not zero (so that the root can have degree 1), then trees of order k contain all trees of lower order as subtrees automatically as well. Generally, under the assumption that d (the gcd of all outdegrees with nonzero weights) in Theorem 2.1 is 1, there exists a positive integer γ (depending on the specific family of trees) such that trees with positive weight exist for all orders greater than γ. This means that every rooted tree S of order $m \leq k - \gamma$ is contained as a fringe subtree in a tree of order k, obtained by taking any tree of order $k - m + 1$ and replacing any leaf by a copy of S. Thus if all trees of order k are present as fringe subtrees in a tree, then all trees of order at most $k - \gamma$ are automatically present as well. This argument can be modified to apply to the case $d > 1$ as well.

The proof of our result is based on the first and second moment method, paired with a careful analysis of the generating function Y_k introduced in Section 2 for trees that do not contain a specific fringe subtree of order k. Recall that the generating function Y that counts all trees has a dominant square root singularity of the form

$$Y(x) = \tau - a(1 - x/\rho)^{1/2} + \mathcal{O}(|x - \rho|),$$

where τ and ρ satisfy the system of equations

$$\tau = \rho\Phi(\tau), \qquad 1 = \rho\Phi'(\tau),$$

and $a = \sqrt{2\Phi(\tau)/\Phi''(\tau)}$. The asymptotic expansion of Y leads to Theorem 2.1 by means of singularity analysis (see [10, Section VII.4] or [7, Section 3.1.4]). Let us assume again, for simplicity, that $d = 1$.

The functional equation is now perturbed slightly. For $h = 1$ or $h = 2$ and an integer $k > 1$, consider the generating function defined by the functional equation

$$Y_{h,k}(x) = x\Phi(Y_{h,k}(x)) - hx^k.$$

$Y_{1,k}$ is the generating function for trees in which one specific fringe subtree of order k does not occur, $Y_{2,k}$ analogously for trees with two given fringe subtrees missing.

In order to apply singularity analysis and thus obtain the asymptotic behavior of the coefficients of $Y_{h,k}$, we need to identify the dominant singularity, which is done by means of bootstrapping. This method is classical (see [17] and the references therein). It is a priori clear that $Y_{h,k}$ must have a dominant positive singularity $\rho_{h,k} \geq \rho$, and this singularity must occur at a solution to the system of equations

$$y = x\Phi(y) - hx^k, \qquad x\Phi'(y) = 1,$$

cf. [10, Theorem VII.3] or [7, Theorem 2.19]. Note that the right hand side of the functional equation does not have nonnegative coefficients only as is usually required, but the argument given in [7] applies nonetheless. Thus let y be implicitly defined by $x\Phi'(y) = 1$, so that $y'(x) = -\Phi'(y)/(x\Phi''(y))$, and set $F(x) = y + hx^k - x\Phi(y)$. Then

$$F'(x) = hkx^{k-1} - \Phi(y),$$

which is negative on the interval $[\rho, \rho + \epsilon]$ (where $\epsilon > 0$ is chosen in such a way that $\rho + \epsilon < 1$) for sufficiently large k since $\Phi(y)$ is bounded below by 1. Moreover, $F(\rho) = h\rho^k$ (since $y = \tau$ if $x = \rho$) and

$$F\left(\rho + \frac{2h\rho}{\tau}\rho^k\right) = -h\rho^k + \mathcal{O}(k\rho^{2k}),$$

so for sufficiently large k there must be a unique solution $(x, y) = (\rho_{h,k}, \tau_{h,k})$ to the system of equations above that satisfies

$$\rho \leq \rho_{h,k} \leq \rho + \frac{2h\rho}{\tau}\rho^k.$$

Refining this argument further, we easily obtain the asymptotic expansions

$$\rho_{h,k} = \rho + \frac{h\rho}{\tau}\rho^k + \mathcal{O}(k\rho^{2k}), \tag{5}$$

$$\tau_{h,k} = \tau - \frac{h}{\rho\tau\Phi''(\tau)}\rho^k + \mathcal{O}(k\rho^{2k}). \tag{6}$$

Around the singularity, the generating function $Y_{h,k}$ has an asymptotic expansion

$$Y_{h,k}(x) = \tau_{h,k} - a_{h,k}(1 - x/\rho_{h,k})^{1/2} + \mathcal{O}(|x - \rho_{h,k}|)$$

with

$$a_{h,k} = \sqrt{\frac{2(\Phi(\tau_{h,k}) - hk\rho_{h,k}^{k-1})}{\Phi''(\tau_{h,k})}} = a + \mathcal{O}(k\rho^k),$$

so we can deduce the following asymptotic formula for the coefficients:

$$t_{h,k,n} = [x^n]Y_{h,k}(x) = \frac{a_{h,k}}{2\sqrt{\pi}} \cdot \frac{\rho_{h,k}^{-n}}{n^{3/2}} \left(1 + \mathcal{O}(n^{-1})\right),$$

and this holds uniformly in k by a simple compactness argument. It follows that the proportion of trees of order n that do not contain h given fringe subtrees of order k is

$$\begin{aligned} \frac{t_{h,k,n}}{t_n} &= \frac{[x^n]Y_{h,k}(x)}{[x^n]Y(x)} \\ &= \left(\frac{\rho}{\rho_{h,k}}\right)^n \left(1 + \mathcal{O}(k\rho^k + n^{-1})\right). \end{aligned} \tag{7}$$

Let us now denote by $X(n,k)$ the number of missing fringe subtrees of order k in a random tree of order n, and set

$$k^* = \log_b \frac{n}{\tau \log n}, \tag{8}$$

where $b = \rho^{-1}$ again.

LEMMA 4.1. *With probability tending to 1, every tree of order at most $\lfloor k^* \rfloor$ appears as a fringe subtree of a random tree of order n.*

Proof. As mentioned at the beginning of this section, all trees of order $\leq k^* - \gamma$ are contained in a tree of order $\lfloor k^* \rfloor$ for some fixed γ, so we focus on trees whose order k lies in the interval $[k^* - \gamma, k^*]$. Let us first show that for such a k, $\mathbb{E}(X(n,k))$, i.e., the expected number of missing fringe subtrees of order k, tends to zero as $n \to \infty$.

By definition, the expected value of $X(n,k)$ is

$$\mathbb{E}(X(n,k)) = t_k \cdot \frac{t_{1,k,n}}{t_n},$$

and from the asymptotic formula for t_k in Theorem 2.1, combined with (7), we obtain the estimate

$$\begin{aligned} &\mathbb{E}(X(n,k)) \\ &\quad = \frac{c\rho^{-k}}{k^{3/2}} \left(\frac{\rho}{\rho_{1,k}}\right)^n \left(1 + \mathcal{O}(k\rho^k + n^{-1} + k^{-1})\right). \end{aligned}$$

In view of (5), this becomes

$$\mathbb{E}(X(n,k)) = \frac{c\rho^{-k}}{k^{3/2}} \exp\left(-\frac{n}{\tau}\rho^k\right) \left(1 + \mathcal{O}\Big(\frac{1}{\log n}\Big)\right).$$

Since $k \leq k^*$ and $\rho^{-k^*} = b^{k^*} = \frac{n}{\tau \log n}$ by our choice of k^*, this gives us

$$\mathbb{E}(X(n,k)) = \mathcal{O}\left(\rho^{-k^*} \exp\left(-\frac{n}{\tau}\rho^{k^*}\right)\right) = \mathcal{O}\Big(\frac{1}{\log n}\Big).$$

Since γ is a fixed constant, this also implies

$$\sum_{k^*-\gamma \leq k \leq k^*} \mathbb{E}(X(n,k)) = \mathcal{O}\Big(\frac{1}{\log n}\Big).$$

Now a simple application of the Markov inequality gives us

$$\mathbb{P}\Big(\sum_{k^*-\gamma \leq k \leq k^*} X(n,k) > 0\Big) = \mathcal{O}\Big(\frac{1}{\log n}\Big).$$

If any tree of order $\leq k^*$ is missing as a fringe subtree, then some fringe subtree whose order lies between $k^* - \gamma$ and k^* must be missing (by the argument stated at the beginning of the section). Thus

$$\mathbb{P}\Big(\sum_{k \leq k^*} X(n,k) > 0\Big) = \mathbb{P}\Big(\sum_{k^*-\gamma \leq k \leq k^*} X(n,k) > 0\Big),$$

and this probability goes to 0 as $n \to \infty$. This completes the proof.

Lemma 4.2. *Let $\epsilon > 0$ be a fixed constant. There exists a constant $\delta > 0$ (depending only on ϵ and the specific family of trees) such that*

$$\mathbb{E}(X(n,k)) \geq n^{\delta}$$

for $k \geq k^* + \epsilon$ and sufficiently large n. Moreover,

$$\frac{\mathbb{E}(X(n,k)^2)}{\mathbb{E}(X(n,k))^2} = 1 + \mathcal{O}(n^{-\delta}) \tag{9}$$

for $k \geq k^* + \epsilon$ as $n \to \infty$. Therefore, with probability tending to 1, a random tree of order n does not contain all trees of order k as fringe subtrees whenever $k \geq k^* + \epsilon$.

Proof. As in the proof of the previous lemma, we obtain

$$\mathbb{E}(X(n,k)) = \frac{c\rho^{-k}}{k^{3/2}} \exp\left(-\frac{n}{\tau}\rho^k\right)\left(1 + \mathcal{O}\Big(\frac{1}{\log n}\Big)\right).$$

Since we now have $k \geq k^* + \epsilon$, we get

$$\mathbb{E}(X(n,k)) \geq \frac{An^{1-\rho^{\epsilon}}}{(\log n)^{5/2}}$$

for some constant $A > 0$ and all sufficiently large n. We can therefore choose any $\delta < 1 - \rho^{\epsilon}$ in our first statement.

Next we estimate the second moment and thus the variance of $X(n,k)$. The generating function for the number of trees of order n that do not contain two given distinct subtrees of order k as fringe subtrees is $Y_{2,k}(x)$, and the number of such trees is $t_{2,k,n}$. Therefore, we have

$$\mathbb{E}(X(n,k)^2) = t_k(t_k - 1) \cdot \frac{t_{2,k,n}}{t_n} + \mathbb{E}(X(n,k)).$$

Now we use the asymptotic formula for t_k in Theorem 2.1 and the asymptotic formula (7) again to obtain

$$\frac{\mathbb{E}(X(n,k)^2)}{\mathbb{E}(X(n,k))^2} = \left(\frac{\rho_{1,k}^2}{\rho\rho_{2,k}}\right)^n \Big(1 + \mathcal{O}(t_k^{-1} + k\rho^k + n^{-1})\Big) + \frac{1}{\mathbb{E}(X(n,k))}.$$

By (5), we have

$$\begin{aligned}\frac{\rho_{1,k}^2}{\rho\rho_{2,k}} &= \frac{\Big(1 + \frac{1}{\tau}\rho^k + \mathcal{O}(k\rho^{2k})\Big)^2}{\Big(1 + \frac{2}{\tau}\rho^k + \mathcal{O}(k\rho^{2k})\Big)} \\ &= 1 + \mathcal{O}(k\rho^{2k}),\end{aligned}$$

so putting everything together we end up with

$$\begin{aligned}&\frac{\mathbb{E}(X(n,k)^2)}{\mathbb{E}(X(n,k))^2} \\ &= 1 + \mathcal{O}(t_k^{-1} + k\rho^k + n^{-1} + nk\rho^{2k}) + \frac{1}{\mathbb{E}(X(n,k))} \\ &= 1 + \mathcal{O}(n^{-\delta}),\end{aligned}$$

where the last step follows from our assumptions on k. This completes the proof of (9).

It follows from our estimates that the variance $\mathbb{V}(X(n,k))$ satisfies

$$\frac{\mathbb{V}(X(n,k))}{\mathbb{E}(X(n,k))^2} \to 0,$$

i.e., the random variable $X(n,k)$ is concentrated around its mean. A standard application of the Chebyshev inequality yields

$$\mathbb{P}\Big(X(n,k) > 0\Big) = 1 + \mathcal{O}(n^{-\kappa})$$

for some positive constant κ, which is what we wanted to prove.

Now we can combine the two lemmas to conclude with the main theorem of this section (the generalization to arbitrary values of d is straightforward):

Theorem 4.1. *Fix some $\epsilon > 0$, and suppose that a random tree T_n of order $n \equiv 1 \mod d$ is drawn from a simply generated family of trees satisfying the conditions of Theorem 2.1. Let m_n be the smallest integer greater than $k^* = \log_b \frac{n}{\tau \log n}$ that is congruent to 1 modulo d. With high probability, the size of the smallest tree that is missing as a fringe subtree in T_n is m_n if $m_n \geq k^* + \epsilon$, and either m_n or $m_n + 1$ otherwise.*

Remark 4.1. *In the generic situation that k^* is not too close to an integer, we have concentration at a single value.*

5 The number of repeated fringe subtrees

In this section we deal with the number of fringe subtrees that occur more than once. The approach will be similar to the proof of Theorem 3.1, however the main contribution will come from a much smaller range of subtree orders, and we will have to carry out a careful analysis based on the bootstrapping results from the previous section. We encounter two phenomena that are perhaps surprising: first of all, we obtain periodic fluctuations (albeit tiny) in the main term of the asymptotics, and secondly the basic asymptotic behavior remains the same even if we count fringe subtrees that are repeated at least r times, where $r \geq 2$ can be any fixed integer. The main result, whose proof we will only sketch, reads as follows:

THEOREM 5.1. *Suppose that a random tree T_n of order $n \equiv 1 \mod d$ is drawn from a simply generated family of trees satisfying the conditions of Theorem 2.1, and write $b = \rho^{-1}$. Then the average number of trees that occur at least r times as a fringe subtree of T_n is asymptotically given by*

$$\psi_r(\log_b n) \cdot \frac{n}{(\log n)^{3/2}} + \mathcal{O}\left(\frac{n}{(\log n)^{5/2}}\right),$$

where ψ_r is a smooth periodic function with period 1 that depends on r and the specific family of simply generated trees. Moreover, ψ_r is bounded above and below by positive constants.

Proof. (Sketch for $r = 2$). As in the proof of Theorem 3.1, we find that the number of trees of order $k \leq \log_b(n/\log n)$ is $\mathcal{O}(n/(\log n)^{5/2})$, so these trees are negligible. Likewise, the number of pairs of identical fringe subtrees whose order is at least $\log_b(n \log n)$ can be bounded in the same way as (4) to yield another $\mathcal{O}(n/(\log n)^{5/2})$.

So we set $k_1 = \log_b(n/\log n)$ and $k_2 = \log_b(n \log n)$ and focus on trees of order k with $k_1 < k < k_2$. We estimate the probability that the random tree T_n contains a given tree S of order k more than once as a fringe subtree. To this end, we define a bivariate generating function $V_k(x, u)$, in which the second variable u marks the number of occurrences of S as a fringe subtree. This generating function satisfies the functional equation

$$V_k(x, u) = x\Phi(V_k(x, u)) + (u - 1)x^k.$$

Clearly, $V_k(x, 0)$ is simply the function Y_k defined in the previous section, and $\frac{\partial V_k}{\partial u}(x, 0) = [u^1]V_k(x, u)$ is the generating function for the number of trees that contain S exactly once as a fringe subtree. It is not difficult to obtain the identity

$$\frac{\partial V_k}{\partial u}(x, 0) = x^k \cdot \frac{x\frac{\partial V_k}{\partial x}(x, 0)}{V_k(x, 0) - (k - 1)x^k}$$

by differentiating with respect to u and x and solving the resulting system of equations. Applying singularity analysis to $V_k(x, 0)$ (as in the previous section) and $\frac{\partial V_k}{\partial u}(x, 0)$, we obtain

$$[x^n]V_k(x, 0) = t_n \left(\frac{\rho}{\rho_{1,k}}\right)^n \left(1 + \mathcal{O}(k\rho^k + n^{-1})\right)$$

as well as

$$[x^n]\frac{\partial V_k}{\partial u}(x, 0) = \frac{n\rho^k t_n}{\tau} \left(\frac{\rho}{\rho_{1,k}}\right)^n \left(1 + \mathcal{O}(k\rho^k + kn^{-1})\right).$$

So by (5), the probability that the fixed subtree S occurs more than once as a fringe subtree in T_n is

$$1 - \left(1 + \frac{n\rho^k}{\tau}\right) \exp\left(-\frac{n\rho^k}{\tau}\right) + \mathcal{O}\left(\frac{\log^2 n}{n}\right)$$

for k in the relevant range. This means that the main contribution to the average number of repeated fringe subtrees is given by the sum

$$\sum_{k_1 < k < k_2} t_k \left(1 - \left(1 + \frac{n\rho^k}{\tau}\right) \exp\left(-\frac{n}{\tau}\rho^k\right)\right).$$

Here, one simply approximates t_k by $ck^{-3/2}\rho^{-k}$ according to Theorem 2.1 and extends the sum to the full range of integers at the expense of further small error terms to end up with

$$\sum_{k=-\infty}^{\infty} \frac{1}{\rho^k n} \left(1 - \left(1 + \frac{n\rho^k}{\tau}\right) \exp\left(-\frac{n\rho^k}{\tau}\right)\right) \cdot \frac{cn}{(\log_b n)^{3/2}} + \mathcal{O}\left(\frac{n}{(\log n)^{5/2}}\right).$$

Is is also not difficult to verify that the infinite sum indeed converges. Setting $\lambda = \log_b n$ (and recalling that $b = \rho^{-1}$), it can be written as

$$(10) \qquad \sum_{k=-\infty}^{\infty} b^{k-\lambda} \left(1 - \left(1 + \frac{b^{\lambda-k}}{\tau}\right) e^{-\frac{b^{\lambda-k}}{\tau}}\right),$$

which shows that it is a continuous, periodic function in λ with period 1. This completes the proof of Theorem 5.1 in the case $r = 2$, and for larger values of r the only change are additional terms inside the brackets in the infinite sum.

Remark 5.1. *The Fourier series of the periodic function in* (10) *is easily found to be*

$$\frac{1}{\tau \log b} + \sum_{k \neq 0} \frac{2k\pi i}{(\log b)^2} \tau^{-1-\frac{2k\pi i}{\log b}} \bigg(-1 - \frac{2k\pi i}{\log b} \bigg) e^{2k\pi i \lambda}.$$

The coefficients are, with the exception of the constant term, very small (for plane trees, the coefficients corresponding to $k = \pm 1$ already have an absolute value of only approximately 0.001342*) and also rapidly decreasing, so the periodic function in our theorem has small amplitude. Since the Γ-function decays exponentially as the imaginary part goes to $\pm\infty$, the periodic function is also smooth.*

6 Some remarks on Pólya trees and labeled trees.

Let us finally say a few words about Pólya trees and labeled trees. It is particularly natural to consider analogous questions for Pólya trees, i.e., rooted unordered trees (where the order of branches does not matter). In this setting, *distinct* simply means *nonisomorphic* (as rooted trees). Pólya trees are technically not simply generated [8], but it is well known that their structure is very similar to simply generated trees. The generating function for Pólya trees satisfies the functional equation

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

and Pólya trees also satisfy the asymptotic formula of Theorem 2.1 with suitable constants:

$$[x^n]R(x) = c \cdot \frac{b^n}{n^{3/2}} \left(1 + \mathcal{O}(n^{-1})\right), \tag{11}$$

where $c \approx 0.439924$ and $b \approx 2.955765$ (see [7, Theorem 3.8]). The treatment of Pólya trees is almost identical to that of simply generated trees, and Theorem 3.1, Theorem 4.1 and Theorem 5.1 hold in exactly the same way (with $\tau = 1$). The only small subtlety lies in the fact that (2) is no longer entirely correct, since attaching a fringe subtree to two distinct leaves of the same tree may result in two trees that are considered identical as Pólya trees. This can easily be overcome e.g. by using a bivariate generating function as in the previous section.

Labeled trees, on the other hand, are a bit different, since the weight of a single tree is no longer 1, but rather $1/k!$, where k is the order of the tree. Note also that two fringe subtrees are considered as distinct in this context if the labels are in a different relative order. This changes the argument only slightly, though, the main difference being the threshold between "small" and "large" trees, which is now the solution to the equation $k^{k+1/2} = n$. Asymptotically, this solution is $k \sim (\log n)/(\log \log n)$. We obtain the following results in analogy to Theorem 3.1 and Theorem 5.1:

Theorem 6.1. *The average number of distinct fringe subtrees of a random rooted labeled tree T_n of order n is asymptotically given by*

$$\sqrt{\frac{2}{\pi}} \cdot \frac{n\sqrt{\log\log n}}{\sqrt{\log n}} \left(1 + \mathcal{O}\left(\frac{\log\log\log n}{\log\log n}\right)\right).$$

Theorem 6.2. *The average number of trees that occur at least r times as a fringe subtree of a random rooted labeled tree T_n is asymptotically given by*

$$\eta_r(n) \cdot \frac{n(\log\log n)^{3/2}}{(\log n)^{3/2}} \left(1 + \mathcal{O}\left(\frac{\log\log\log n}{\log\log n}\right)\right),$$

where η_r is a function that is bounded above and below by positive constants.

Note that the function η_r is no longer a periodic function of $\log_b n$. Likewise, the threshold in Theorem 4.1 changes somewhat when labeled trees are considered:

Theorem 6.3. *Fix some $\epsilon > 0$, and suppose that a random rooted labeled tree T_n of order n is drawn. Let k^* be such that $e^{k^*}k^*! = \frac{n}{\log n}$, and let m_n be the smallest integer greater than k^*. With high probability, the size of the smallest tree that is missing as a fringe subtree in T_n is m_n if $m_n \geq k^* + \epsilon$, and either m_n or $m_n + 1$ otherwise.*

7 Conclusion

We have seen that the parameters studied in this paper show some rather unusual asymptotic behavior. Several possible extensions of our results are conceivable. In the setting of the paper [4] that was mentioned in the introduction as a starting point to this work, the nodes are equipped with labels from a given finite alphabet, and trees only count as equal if the labels match as well. This was also considered in [11], with probabilities assigned to the labels. Our main results generalize in a straightforward way to this setting as well.

More interestingly, one could study how the situation changes as a different probabilistic model is used, e.g. random recursive or plane-oriented recursive trees, random tries or binary search trees (see [7, Sections 6 and 7]). As it was mentioned earlier, it is also natural to ask for the limiting distribution of the number of distinct fringe subtrees (or even just the asymptotic behavior of the variance), but this appears to be a very challenging task.

Acknowledgment

We are highly indebted to an anonymous referee who pointed out reference [11] to us.

References

[1] David Aldous, *Asymptotic fringe distributions for general families of random trees*, Ann. Appl. Probab. **1** (1991), no. 2, 228–266.

[2] Huilan Chang and Michael Fuchs, *Limit theorems for patterns in phylogenetic trees*, J. Math. Biol. **60** (2010), no. 4, 481–512.

[3] Michalis Christou, Maxime Crochemore, Tomáš Flouri, Costas S. Iliopoulos, Jan Janoušek, Bořivoj Melichar, and Solon P. Pissis, *Computing all subtree repeats in ordered ranked trees*, String processing and information retrieval, Lecture Notes in Comput. Sci., vol. 7024, Springer, Heidelberg, 2011, pp. 338–343.

[4] ———, *Computing all subtree repeats in ordered trees*, Information Processing Letters **112** (2012), no. 24, 958–962.

[5] Florian Dennert and Rudolf Grübel, *On the subtree size profile of binary search trees*, Combin. Probab. Comput. **19** (2010), no. 4, 561–578.

[6] Luc Devroye and Svante Janson, *Protected nodes and fringe subtrees in some random trees*, Electron. Commun. Probab. **19** (2014), no. 6, 10.

[7] Michael Drmota, *Random trees*, SpringerWienNewYork, Vienna, 2009.

[8] Michael Drmota and Bernhard Gittenberger, *The shape of unlabeled rooted random trees*, European J. Combin. **31** (2010), no. 8, 2028–2063.

[9] Qunqiang Feng and Hosam M. Mahmoud, *On the variety of shapes on the fringe of a random recursive tree*, J. Appl. Probab. **47** (2010), no. 1, 191–200.

[10] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[11] Philippe Flajolet, Paolo Sipala, and Jean-Marc Steyaert, *Analytic variations on the common subexpression problem*, Automata, languages and programming (Coventry, 1990), Lecture Notes in Comput. Sci., vol. 443, Springer, New York, 1990, pp. 220–234.

[12] Tomáš Flouri, Kassian Kobert, Solon P. Pissis, and Alexandros Stamatakis, *An optimal algorithm for computing all subtree repeats in trees*, Combinatorial Algorithms, Lecture Notes in Comput. Sci., vol. 8288, Springer, Heidelberg, 2013, pp. 269–282.

[13] Michael Fuchs, *Limit theorems for subtree size profiles of increasing trees*, Combin. Probab. Comput. **21** (2012), no. 3, 412–441.

[14] Cecilia Holmgren and Svante Janson, *Limit laws for functions of fringe trees for binary search trees and recursive trees*, preprint.

[15] Svante Janson, *Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton–Watson trees*, Random Structures Algorithms, to appear.

[16] A. Meir and J. W. Moon, *On the altitude of nodes in random trees*, Canad. J. Math. **30** (1978), no. 5, 997–1015.

[17] Helmut Prodinger and Stephan Wagner, *Bootstrapping and double-exponential limit laws*, preprint.