On the Collection of Fringe Subtrees in Random Binary Trees

Louisa Seelbach Benkner[1*✉] and Stephan Wagner[2,3][0000−0001−5533−2764]

[1] Department für Elektrotechnik und Informatik, Universität Siegen, Hölderlinstrasse 3, 57076 Siegen, Germany
seelbach@eti.uni-siegen.de

[2] Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
swagner@sun.ac.za

[3] Department of Mathematics, Uppsala Universitet, Box 480, 751 06 Uppsala, Sweden
stephan.wagner@math.uu.se

Abstract. A fringe subtree of a rooted tree is a subtree consisting of one of the nodes and all its descendants. In this paper, we are specifically interested in the number of non-isomorphic trees that appear in the collection of all fringe subtrees of a binary tree. This number is analysed under two different random models: uniformly random binary trees and random binary search trees.

In the case of uniformly random binary trees, we show that the number of non-isomorphic fringe subtrees lies between $c_1 n/\sqrt{\ln n}(1 + o(1))$ and $c_2 n/\sqrt{\ln n}(1 + o(1))$ for two constants $c_1 \approx 1.0591261434$ and $c_2 \approx 1.0761505454$, both in expectation and with high probability, where n denotes the size (number of leaves) of the uniformly random binary tree. A similar result is proven for random binary search trees, but the order of magnitude is $n/\ln n$ in this case.

Our proof technique can also be used to strengthen known results on the number of distinct fringe subtrees (distinct in the sense of ordered trees). This quantity is of the same order of magnitude in both cases, but with slightly different constants in the upper and lower bounds.

Keywords: Uniformly Random Binary Trees · Random Binary Search Trees · Fringe Subtrees · Tree Compression

1 Introduction

A subtree of a rooted tree that consists of a node and all its descendants is called a *fringe subtree*. Fringe subtrees are a natural object of study in the context of random trees, and there are numerous results for various random tree models, see e.g. [3, 9, 11, 13].

* This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 731143 and the DFG research project LO 748/10-1 (QUANT-KOMP).

Fringe subtrees are of particular interest in computer science: One of the most important and widely used lossless compression methods for rooted trees is to represent a tree as a directed acyclic graph, which is obtained by merging nodes that are roots of identical fringe subtrees. This compressed representation of the tree is often shortly referred to as *minimal DAG* and its size (number of nodes) is the number of distinct fringe subtrees occurring in the tree. Compression by minimal DAGs has found numerous applications in various areas of computer science, as for example in compiler construction [2, Chapter 6.1 and 8.5], unification [24], symbolic model checking (binary decision diagrams) [7], information theory [20, 28] and XML compression and querying [8, 19].

In this work, we investigate the number of fringe subtrees in random binary trees, i.e. random trees such that each node has either exactly two or no children. So far, this problem has mainly been studied with respect to ordered fringe subtrees in random ordered binary trees: A *uniformly random ordered binary tree* of size n (with n leaves) is a random tree whose probability distribution is the uniform probability distribution on the set of ordered binary trees of size n. In [18], Flajolet, Sipala and Steyaert proved that the expected number of distinct ordered fringe subtrees in a uniformly random ordered binary tree of size n is asymptotically equal to $c \cdot n/\sqrt{\ln n}$, where c is the constant $2\sqrt{\ln 4/\pi}$. This result of Flajolet et al. was extended to unranked labelled trees in [6] (for a different constant c). Moreover, an alternative proof to the result of Flajolet et al. was presented in [25] in the context of simply-generated families of trees.

Another important type of random trees are so-called *random binary search trees*: A random binary search tree of size n is a binary search tree built by inserting the keys $\{1, \dots, n\}$ according to a uniformly chosen random permutation on $\{1, \dots, n\}$. Random binary search trees naturally arise in theoretical computer science, see e.g. [12]. In [16], Flajolet, Gourdon and Martinez proved that the expected number of distinct ordered fringe subtrees in a random binary search tree of size n is $O(n/\ln n)$. This result was improved in [10] by Devroye, who showed that the asymptotics $\Theta(n/\ln n)$ holds. Moreover, the result of Devroye was generalized from random binary search trees to a broader class of random ordered binary trees in [26], where the problem of estimating the expected number of distinct ordered fringe subtrees in random binary trees was considered in the context of so-called leaf-centric binary tree sources, which were introduced in [22, 28] as a general framework for modeling probability distributions on the set of ordered binary trees of size n.

In this work, we focus on estimating the number of *non-isomorphic* fringe subtrees in random ordered binary trees, where we call two binary trees non-isomorphic if they are distinct as unordered binary trees. This question arises quite naturally for example in the context of XML compression: Here, one distinguishes between so-called document-centric XML, for which the corresponding XML document trees are ordered, and data-centric XML, for which the corresponding XML document trees are unordered. Understanding the interplay between ordered and unordered structures has thus received considerable attention in the context of XML (see, for example, [1, 5, 29]). In particular, in [23], it

was investigated whether tree compression can benefit from unorderedness. For this reason, so-called *unordered minimal DAGs* were considered. An unordered minimal DAG of a binary tree is a directed acyclic graph obtained by merging nodes that are roots of isomorphic fringe subtrees, i.e. of fringe subtrees which are identical as unordered trees. From such an unordered minimal DAG, an unordered representation of the original tree can be uniquely retrieved. The size of this compressed representation is the number of non-isomorphic fringe subtrees occurring in the tree. So far, only some worst-case estimates comparing the size of a minimal DAG to the size of its corresponding unordered minimal DAG are known: Among other things, it was shown in [23] that the size of an unordered minimal DAG of a binary tree can be exponentially smaller than the size of the corresponding (ordered) minimal DAG.

However, no average-case estimates comparing the size of the minimal DAG of a binary tree to the size of the corresponding unordered minimal DAG are known so far. In particular, in [23] it is stated as an open problem to estimate the expected number of non-isomorphic fringe subtrees in a uniformly random ordered binary tree of size n and conjectured that this number asymptotically grows as $\Theta(n/\sqrt{\ln n})$.

In this work, as one of our main theorems, we settle this open conjecture by proving upper and lower bounds of order $n/\sqrt{\ln n}$ for the number of non-isomorphic fringe subtrees which hold both in expectation and with high probability (i.e., with probability tending to 1 as $n \to \infty$). Our approach can also be used to obtain an analogous result for random binary search trees, though the order of magnitude changes to $\Theta(n/\ln n)$. Again, we have upper and lower bounds in expectation and with high probability. Our two main theorems read as follows.

Theorem 1 *Let F_n be the total number of non-isomorphic fringe subtrees in a uniformly random ordered binary tree with n leaves. For two constants $c_1 \approx 1.0591261434$ and $c_2 \approx 1.0761505454$, the following holds:*

(i) $c_1 \dfrac{n}{\sqrt{\ln n}}(1+o(1)) \le \mathbb{E}(F_n) \le c_2 \dfrac{n}{\sqrt{\ln n}}(1+o(1))$,

(ii) $c_1 \dfrac{n}{\sqrt{\ln n}}(1+o(1)) \le F_n \le c_2 \dfrac{n}{\sqrt{\ln n}}(1+o(1))$ *with high probability.*

Theorem 2 *Let G_n be the total number of non-isomorphic fringe subtrees in a random binary search tree with n leaves. For two constants $c_3 \approx 1.5470025923$ and $c_4 \approx 1.8191392203$, the following holds:*

(i) $c_3 \dfrac{n}{\ln n}(1+o(1)) \le \mathbb{E}(G_n) \le c_4 \dfrac{n}{\ln n}(1+o(1))$,

(ii) $c_3 \dfrac{n}{\ln n}(1+o(1)) \le G_n \le c_4 \dfrac{n}{\ln n}(1+o(1))$ *with high probability.*

To prove the above Theorems 1 and 2, we refine techniques from [25]. Our proof technique also applies to the problem of estimating the number of distinct ordered fringe subtrees in uniformly random binary trees or in random binary search trees. In this case, upper and lower bounds for the expected value have

already been proven by other authors. Our new contribution is to show that they also hold with high probability.

Theorem 3 *Let H_n denote the total number of distinct fringe subtrees in a uniformly random ordered binary tree with n leaves. Then, for the constant $c = 2\sqrt{\ln 4/\pi} \approx 1.3285649405$, the following holds:*

(i) $\mathbb{E}(H_n) = c\dfrac{n}{\sqrt{\ln n}}(1 + o(1))$,

(ii) $H_n = c\dfrac{n}{\sqrt{\ln n}}(1 + o(1))$ *with high probability.*

Here, the first part (i) was already shown in [18] and [25], part (ii) is new. Similarly, we are able to strengthen the results of [10] and [26]:

Theorem 4 *Let J_n be the total number of distinct fringe subtrees in a random binary search tree with n leaves. For two constants $c_5 \approx 2.4071298335$ and $c_6 \approx 2.7725887222$, the following holds:*

(i) $c_5\dfrac{n}{\ln n}(1 + o(1)) \leq \mathbb{E}(J_n) \leq c_6\dfrac{n}{\ln n}(1 + o(1))$,

(ii) $c_5\dfrac{n}{\ln n}(1 + o(1)) \leq J_n \leq c_6\dfrac{n}{\ln n}(1 + o(1))$ *with high probability.*

The upper bound in part (i) can already be found in [16] and [10]. Moreover, a lower bound of the form $\mathbb{E}(J_n) \geq \frac{\alpha n}{\ln n}(1 + o(1))$ was already shown in [10] for the constant $\alpha = (\ln 3)/2 \approx 0.5493061443$ and in [26] for the constant $\alpha \approx 0.6017824584$. So our new contributions are part (ii) and the improvement of the lower bound on $\mathbb{E}(J_n)$.

2 Preliminaries

Let $\mathcal{T}$ denote the set of ordered binary trees, i.e. of ordered rooted trees such that each node has either exactly two or no children. We define the *size* $|t|$ of a binary tree $t \in \mathcal{T}$ as the number of leaves of t and by $\mathcal{T}_k$ we denote the set of binary trees of size k for every integer $k \geq 1$. It is well known that $|\mathcal{T}_k| = C_{k-1}$, where C_k denotes the k-th *Catalan number* [17]: We have

$$C_k = \frac{1}{k+1}\binom{2k}{k} \sim \frac{4^k}{\sqrt{\pi}k^{3/2}}(1 + O(1/k)), \tag{1}$$

where the asymptotic growth of the Catalan numbers follows from Stirling's Formula [17]. Analogously, let $\mathcal{U}$ denote the set of unordered binary trees, i.e. of unordered rooted trees such that each node has either exactly two or no children. The *size* $|u|$ of an unordered tree $u \in \mathcal{U}$ is again the number of leaves of u and by $\mathcal{U}_k$ we denote the set of unordered binary trees of size k. We have $|\mathcal{U}_k| = W_k$, where W_k denotes the k-th *Wedderburn-Etherington number*. Their asymptotic growth is

$$W_k \sim A \cdot k^{-3/2} \cdot b^k, \tag{2}$$

for certain positive constants A, b [4,15]. In particular, we have $b \approx 2.4832535362$.

A *fringe subtree* of a binary tree is a subtree consisting of a node and all its descendants. For a binary tree t and a given node $v \in t$, let $t(v)$ denote the fringe subtree of t rooted at v. Two fringe subtrees are called *distinct* if they are distinct as ordered binary trees.

Every tree $t \in \mathcal{T}$ can be considered as an element of $\mathcal{U}$ by simply forgetting the ordering on t's nodes. If two binary trees t_1, t_2 correspond to the same unordered tree $u \in \mathcal{U}$, we call them *isomorphic*: Thus, we obtain a partition of $\mathcal{T}$ into isomorphism classes. If two binary trees $t_1, t_2 \in \mathcal{T}$ belong to the same isomorphism class, we can obtain t_1 from t_2 and vice versa by reordering the children of some of t_1's (respectively, t_2's) inner nodes. An inner node v of an ordered or unordered binary tree t is called a *symmetrical node* if the fringe subtrees rooted at v's children are isomorphic. Let $\mathrm{sym}(t)$ denote the number of symmetrical nodes of t. The cardinality of the automorphism group of t is given by $|\mathrm{Aut}(t)| = 2^{\mathrm{sym}(t)}$. Thus, by the orbit-stabilizer theorem, there are $2^{k-1-\mathrm{sym}(t)}$ many ordered binary trees in the isomorphism class of $t \in \mathcal{T}_k$, and likewise $2^{k-1-\mathrm{sym}(t)}$ many ordered representations of $t \in \mathcal{U}_k$.

We consider two types of probability distributions on the set of ordered binary trees of size n:

(i) The *uniform probability distribution* on $\mathcal{T}_n$, that is, every binary tree of size n is assigned the same probability $\frac{1}{C_{n-1}}$. A random variable taking values in $\mathcal{T}_n$ according to the uniform probability distribution is called a *uniformly random (ordered) binary tree* of size n.

(ii) The probability distribution induced by the so-called *Binary Search Tree Model* (see e.g. [12,16]): The corresponding probability mass function $P_{\mathrm{bst}} : \mathcal{T}_n \to [0,1]$ is given by

$$P_{\mathrm{bst}}(t) = \prod_{\substack{v \in t \\ |t(v)|>1}} \frac{1}{|t(v)|-1}, \tag{3}$$

for every $n \geq 1$. A random variable taking values in $\mathcal{T}_n$ according to this probability mass function is called a *random binary search tree* of size n.

Before we prove our main results, we need two preliminary lemmas:

Lemma 1. *Let a, ε be positive real numbers with $\varepsilon < \frac{1}{3}$. For every positive integer k with $a \ln n \leq k \leq n^{\varepsilon}$, let $\mathcal{S}_k \subset \mathcal{T}_k$ be a set of ordered binary trees with k leaves. We denote the cardinality of $\mathcal{S}_k$ by s_k. Let $X_{n,k}$ denote the (random) number of fringe subtrees with k leaves in a uniformly random ordered binary tree with n leaves that belong to $\mathcal{S}_k$. Moreover, let $Y_{n,\varepsilon}$ denote the (random) number of arbitrary fringe subtrees with more than n^{ε} leaves in a uniformly random ordered binary tree with n leaves. We have*

(1) $\mathbb{E}(X_{n,k}) = s_k 4^{1-k} n\big(1 + O(k/n)\big)$ *for all k with $a \ln n \leq k \leq n^{\varepsilon}$, the O-constant being independent of k,*

(2) $\mathbb{V}(X_{n,k}) = s_k 4^{1-k} n(1 + O(k^{-1/2}))$ *for all* k *with* $a \ln n \leq k \leq n^{\varepsilon}$*, again with an* O*-constant that is independent of* k*,*
(3) $\mathbb{E}(Y_{n,\varepsilon}) = O(n^{1-\varepsilon/2})$ *and*
(4) with high probability, the following statements hold:
  (i) $|\sum_k X_{n,k} - \mathbb{E}(X_{n,k})| \leq \sum_k s_k^{1/2} 2^{-k} n^{1/2+\varepsilon}$*, where the sums are taken over all* k *with* $a \ln n \leq k \leq n^{\varepsilon}$*,*
  (ii) $Y_{n,\varepsilon} \leq n^{1-\varepsilon/3}$*.*

Lemma 2. *Let* a, ε *be positive real numbers with* $\varepsilon < \frac{1}{3}$ *and let* n *and* k *denote positive integers. Moreover, for every* k*, let* $\mathcal{S}_k \subset \mathcal{T}_k$ *be a set of ordered binary trees with* k *leaves and let* p_k *denote the probability that a random binary search tree is contained in* $\mathcal{S}_k$*, that is,* $p_k = \sum P_{\mathrm{bst}}(t)$*, where the sum is taken over all binary trees in* $\mathcal{S}_k$*. Let* $X_{n,k}$ *denote the (random) number of fringe subtrees with* k *leaves in a random binary search tree with* n *leaves that belong to* $\mathcal{S}_k$*. Moreover, let* $Y_{n,\varepsilon}$ *denote the (random) number of arbitrary fringe subtrees with more than* n^{ε} *leaves in a random binary search tree with* n *leaves. We have*

(1) $\mathbb{E}(X_{n,k}) = \frac{2p_k n}{k(k+1)}$ *for* $1 \leq k < n$*,*
(2) $\mathbb{V}(X_{n,k}) = O(p_k n/k^2)$ *for all* k *with* $a \ln n \leq k \leq n^{\varepsilon}$*, where the* O*-constant is independent of* k*,*
(3) $\mathbb{E}(Y_{n,\varepsilon}) = 2n/\lceil n^{\varepsilon} \rceil - 1 = O(n^{1-\varepsilon})$ *and*
(4) with high probability, the following statements hold:
  (i) $|\sum_k X_{n,k} - \mathbb{E}(X_{n,k})| \leq \sum_k p_k^{1/2} k^{-1} n^{1/2+\varepsilon}$*, where the sums are taken over all* k *with* $a \ln n \leq k \leq n^{\varepsilon}$*,*
  (ii) $Y_{n,\varepsilon} \leq n^{1-\varepsilon/2}$*.*

For the proofs of Lemma 1 and Lemma 2, see the long version of the paper [27].

3 Fringe Subtrees in Uniformly Random Binary Trees

3.1 Ordered Fringe Subtrees

We provide the proof of Theorem 3 first, since it is simplest and provides us with a template for the other proofs. Basically, it is a refinement of the proof for the corresponding special case of Theorem 3.1 in [25]. In the following sections, we refine the argument further to prove Theorems 1, 2 and 4. For further details, see the long version of the paper [27].

Proof (Proof of Theorem 3). We prove the statement in two steps: In the first step, we show that the upper bound $H_n \leq cn/\sqrt{\ln n}(1 + o(1))$ holds for $c = 2\sqrt{\ln 4/\pi}$ both in expectation and with high probability. In the second step, we prove the corresponding lower bound.

The upper bound: Let $k_0 = \log_4 n$. The number H_n of distinct fringe subtrees in a uniformly random ordered binary tree with n leaves equals (i) the number of such distinct fringe subtrees of size at most k_0 plus (ii) the number of such distinct fringe subtrees of size greater than k_0. We upper-bound (i) by the number

of all ordered binary trees of size at most k_0 (irrespective of their occurrence as fringe subtrees) and (ii) by the total number of such fringe subtrees occurring in the tree to obtain, using the notation of Lemma 1,

$$H_n \leq \sum_{k \leq k_0} C_{k-1} + \Big(\sum_{k_0 < k \leq n^\varepsilon} X_{n,k} \Big) + Y_{n,\varepsilon}.$$

Here, $\mathcal{S}_k$ is the full set $\mathcal{T}_k$, so that $s_k = C_{k-1}$. The first sum is $O(n/(\ln n)^{3/2})$ by (1). This upper bound holds deterministically. In order to estimate the other two terms, we apply Lemma 1 with $a = \frac{1}{\ln 4}$ and $\varepsilon = \frac{1}{6}$. We thus find that the two terms are bounded from above by $\frac{2\sqrt{\ln 4}}{\sqrt{\pi}} \cdot \frac{n}{\sqrt{\ln n}} + O(n/(\ln n)^{3/2})$, both in expectation and with high probability.

The lower bound: Again, let $k_0 = \log_4 n$ and $\varepsilon = \frac{1}{6}$. In order to lower-bound the number H_n of distinct fringe subtrees in a uniformly random ordered tree with n leaves, we only count distinct fringe subtrees of sizes k with $k_0 < k \leq n^\varepsilon$. To this end, let $X_{n,k}^{(2)}$ denote the number of pairs of identical fringe subtrees of size k in a uniformly random ordered binary tree of size n. Each such pair can be obtained by taking an ordered tree with $n-2k+2$ leaves, picking two leaves, and replacing them by the same ordered binary tree of size k. The total number of such pairs of identical fringe subtrees of size k is thus

$$C_{n-2k+1} \cdot \binom{n-2k+2}{2} \cdot C_{k-1} = \frac{4^{n-k}}{2\pi k^{3/2}} (n-2k+1)^{1/2} (1 + O(1/k)).$$

By dividing by C_{n-1}, i.e. the total number of binary trees of size n, we thus obtain the expected value: $\mathbb{E}(X_{n,k}^{(2)}) = O(4^{-k} n^2 k^{-3/2})$ and consequently $\sum \mathbb{E}(X_{n,k}^{(2)}) = O(n/(\ln n)^{3/2})$, where the sum is taken over all k with $k_0 < k \leq n^\varepsilon$. If a binary tree of size k occurs m times as a fringe subtree in a uniformly random binary tree of size n, it contributes $m - \binom{m}{2}$ to the random variable $X_{n,k} - X_{n,k}^{(2)}$. Since $m - \binom{m}{2} \leq 1$ for all non-negative integers m, we find that $X_{n,k} - X_{n,k}^{(2)}$ is a lower bound on the number of distinct fringe subtrees with k leaves. Hence, we have

$$H_n \geq \sum_{k_0 < k \leq n^\varepsilon} X_{n,k} - \sum_{k_0 < k \leq n^\varepsilon} X_{n,k}^{(2)}.$$

The second sum is $O(n/(\ln n)^{3/2})$ in expectation and thus with high probability as well by the Markov inequality. As the first sum is $\frac{2\sqrt{\ln 4}}{\sqrt{\pi}} \cdot \frac{n}{\sqrt{\ln n}}(1 + o(1))$, both in expectation and with high probability by our estimate from the first part of the proof, the statement of Theorem 3 follows. ∎

As the main idea of the proof is to split the number of distinct fringe subtrees into the number of distinct fringe subtrees of size at most k_0 plus the number of distinct fringe subtrees of size greater than k_0 for some suitably chosen integer k_0, this type of argument is called a *cut-point argument* and the integer k_0 is called the *cut-point* (see [16]). This basic technique is applied in several previous papers to similar problems (see for instance [10], [16], [25], [26]). Moreover, we remark that the statement of Theorem 3 can be easily generalized to simply generated families of trees.

3.2 Unordered Fringe Subtrees

In this subsection, we prove Theorem 1. For this, we refine the cut-point argument we applied in the proof of Theorem 3: In particular, for the lower bound on F_n, we need a result due to Bóna and Flajolet [4] on the number of automorphisms of a uniformly random ordered binary tree. It is stated for random phylogenetic trees in [4], but the two probabilistic models are equivalent.

Theorem 5 ([4], Theorem 2) *Consider a uniformly random ordered binary tree T_k with k leaves, and let $A_k = |\mathrm{Aut}(T_k)|$ be the cardinality of its automorphism group. The logarithm of this random variable satisfies a central limit theorem: For certain positive constants γ and σ_1, we have*

$$\mathbb{P}(A_k \leq 2^{\gamma k+\sigma_1\sqrt{k}x}) \overset{k\to\infty}{\to} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt$$

for every real number x. The numerical value of the constant γ is 0.2710416936.

With Theorem 5, we are able to upper-bound the probability that two fringe subtrees of the same size are isomorphic in our proof of Theorem 1:

Proof (Proof of Theorem 1). We prove the statement in two steps: First, we show that the upper bound on F_n stated in Theorem 1 holds both in expectation and with high probability, then we prove the respective lower bound.

The upper bound: The proof for the upper bound in Theorem 1 exactly matches the first part of the proof of Theorem 3, except that we choose a different cut-point: Let $k_0 = \log_b n$, where $b \approx 2.4832535362$ is the constant in the asymptotic formula (2) for the Wedderburn-Etherington numbers. We then find

$$F_n \leq \sum_{k<k_0} W_k + \Big(\sum_{k_0\leq k\leq n^\epsilon} X_{n,k}\Big) + Y_{n,\epsilon} = \frac{2\sqrt{\ln b}}{\sqrt{\pi}}\cdot\frac{n}{\sqrt{\ln n}} + O(n(\ln n)^{-3/2}),$$

both in expectation and with high probability, where the estimates for $X_{n,k}$ and $Y_{n,\varepsilon}$ follow again from Lemma 1. We have $2\sqrt{\ln b}/\sqrt{\pi} \approx 1.0761505454$.

The lower bound: As a consequence of Theorem 5, the probability that the cardinality of the automorphism group of a uniformly random binary tree T_k of size k satisfies $|\mathrm{Aut}(T_k)| \leq 2^{\gamma k-k^{3/4}}$ tends to 0 as $k\to\infty$. We define $\mathcal{S}_k$ as the set of ordered trees with k leaves that do not satisfy this inequality, so that $s_k = |\mathcal{S}_k| = C_{k-1}(1+o(1))$. Our lower bound is based on counting only fringe subtrees in $\mathcal{S}_k$ for suitable k. The reason for this choice is that we have an upper bound on the number of ordered binary trees in the same isomorphism class for every tree in $\mathcal{S}_k$. Recall that the number of possible ordered representations of an unordered binary tree t with k leaves is given by $2^{k-1}/|\mathrm{Aut}(t)|$ by the orbit-stabiliser theorem. Hence, the number of ordered binary trees in the same isomorphism class as a tree $t\in\mathcal{S}_k$ is bounded above by $2^{k-1-\gamma k+k^{3/4}}$.

Now set $k_1 = \frac{1+\delta}{1+\gamma}\log_2 n$ for some positive constant $\delta < \frac{2}{3}$, and consider only fringe subtrees that belong to $\mathcal{S}_k$, where $k_1 \leq k \leq n^{\delta/2}$. By Lemma 1, the

number of such fringe subtrees in a random ordered binary tree with n leaves is $s_k 4^{1-k} n(1 + O(k/n + s_k^{-1/2} 2^k n^{(\delta-1)/2}))$ both in expectation and with high probability. Since $s_k = C_{k-1}(1+o(1))$, the number of fringe subtrees that belong to $\mathcal{S}_k$ in a random ordered binary tree of size n becomes $\frac{n}{\sqrt{\pi k^3}}(1+o(1))$. We show that most of these trees are the only representatives of their isomorphism classes as fringe subtrees. To this end, we consider all fringe subtrees in $\mathcal{S}_k$ for some k that satisfies $k_1 \le k \le n^{\delta/2}$. Let the sizes of the isomorphism classes of trees in $\mathcal{S}_k$ be $r_1, r_2, \dots, r_\ell$, so that $r_1 + r_2 + \cdots + r_\ell = s_k$. By definition of $\mathcal{S}_k$, we have $r_i \le 2^{k-1-\gamma k+k^{3/4}}$ for every i. Let us condition on the event that their number $X_{n,k}$ is equal to N for some $N \le n$. Each of these N fringe subtrees $S_1, S_2, \dots, S_N$ follows a uniform distribution among the elements of $\mathcal{S}_k$, so the probability of being in an isomorphism class with r_i elements is r_i/s_k. Moreover, the N fringe subtrees are also all independent. Let $X^{(2)}_{n,k}$ be the number of pairs of isomorphic trees among the fringe subtrees with k leaves. We have

$$\mathbb{E}\big(X^{(2)}_{n,k} | X_{n,k} = N\big) = \binom{N}{2} \sum_i \left(\frac{r_i}{s_k}\right)^2 \le \frac{n^2}{2s_k^2} \sum_i r_i^2 \le \frac{n^2}{s_k} 2^{k-2-\gamma k+k^{3/4}}.$$

Since this holds for all N, the law of total expectation yields

$$\mathbb{E}\big(X^{(2)}_{n,k}\big) \le \frac{n^2}{s_k} 2^{k-2-\gamma k+k^{3/4}} = \sqrt{\pi} n^2 k^{3/2} 2^{-k-\gamma k+k^{3/4}}(1+o(1)).$$

Since $k \ge k_1 = \frac{1+\delta}{1+\gamma} \log_2 n$, we find that

$$\mathbb{E}\big(X^{(2)}_{n,k}\big) \le n^2 2^{-(1+\gamma)k+O(k^{3/4})} \le n^{1-\delta} \exp\big(O((\ln n)^{3/4})\big).$$

Thus

$$\sum_{k_1 \le k \le n^{\delta/2}} \mathbb{E}\big(X^{(2)}_{n,k}\big) \le n^{1-\delta/2} \exp\big(O((\ln n)^{3/4})\big) = o(n/\sqrt{\ln n}).$$

As in the previous proof, we see that $X_{n,k} - X^{(2)}_{n,k}$ is a lower bound on the number of non-isomorphic fringe subtrees with k leaves. This gives us

$$F_n \ge \sum_{k_1 \le k \le n^{\delta/2}} X_{n,k} - \sum_{k_1 \le k \le n^{\delta/2}} X^{(2)}_{n,k}.$$

The second sum is negligible since it is $o(n/\sqrt{\ln n})$ in expectation and thus also with high probability by the Markov inequality. For the first sum, a calculation similar to that for the upper bound shows that it is

$$\frac{2\sqrt{(1+\gamma)\ln 2}}{\sqrt{\pi(1+\delta)}} \cdot \frac{n}{\sqrt{\ln n}}(1+o(1)),$$

both in expectation and with high probability. Since δ is arbitrary, we can choose any constant smaller than $\frac{2\sqrt{(1+\gamma)\ln 2}}{\sqrt{\pi}} \approx 1.0591261434$ for c_1. ∎

4 Fringe Subtrees in Random Binary Search Trees

In order to show the respective lower bounds of Theorem 2 and Theorem 4, we need two theorems similar to Theorem 5: The first one shows that the logarithm of the random variable $B_k = P_{\mathrm{bst}}(T_k)^{-1}$, where T_k denotes a random binary search tree of size k, satisfies a central limit theorem and is needed to estimate the probability that two fringe subtrees in a random binary search tree are identical. The second one transfers the statement of Theorem 5 from uniformly random binary trees to random binary search trees and is needed in order to estimate the probability that two fringe subtrees in a random binary search tree are isomorphic. The first of these two central limit theorems is shown in [14]:

Theorem 6 ([14], Theorem 4.1) *Consider a random binary search tree T_k with k leaves, and let $B_k = P_{\mathrm{bst}}(T_k)^{-1}$. The logarithm of this random variable satisfies a central limit theorem: For certain positive constants μ and σ_2, we have*

$$\mathbb{P}\left(B_k \leq 2^{\mu k+\sigma_2\sqrt{k}x}\right) \overset{k\to\infty}{\to} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt$$

for every real number x. The numerical value of the constant μ is

$$\mu = \sum_{k=1}^{\infty} \frac{2\log_2 k}{(k+1)(k+2)} \approx 1.7363771368.$$

The second of these two central limit theorems follows from a general theorem devised by Holmgren and Janson [21]: The proof of Theorem 7 can be found in the long version of the paper [27].

Theorem 7 *Consider a random binary search tree T_k with k leaves, and let $A_k = |\mathrm{Aut}(T_k)|$ be the cardinality of its automorphism group. The logarithm of this random variable satisfies a central limit theorem: for certain positive constants ν and σ_3, we have*

$$\mathbb{P}(A_k \leq 2^{\nu k+\sigma_3\sqrt{k}x}) \overset{k\to\infty}{\to} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt$$

for every real number x. The numerical value of ν is $\nu \approx 0.3795493473$.

For the proofs of Theorems 2 and 4, we refer to the long version of the paper [27]: The techniques used in the proofs are mostly the same as in the proof of Theorem 1. In order to show the corresponding upper bounds, we make use of the cut-point technique presented in the proofs of Theorems 3 and 1, combined with Lemma 2. For the lower bounds, we suitably define, as in the proof of Theorem 1, respective sets $\mathcal{S}_k$ using Theorems 6 and 7. We then lower-bound the number of distinct (non-isomorphic, respectively) fringe subtrees by the number of such fringe subtrees of size k that belong to the respective set $\mathcal{S}_k$. The sets $\mathcal{S}_k$ and the range of k are again chosen in a way that allows us to bound the probability that two fringe subtrees from the set $\mathcal{S}_k$ are identical (isomorphic, respectively).

5 Open Problems

The following natural question arises from our results: Is it possible to determine constants $\alpha_1, \alpha_2, \alpha_3$ with $c_1 \le \alpha_1 \le c_2$, $c_3 \le \alpha_2 \le c_4$ and $c_5 \le \alpha_3 \le c_6$, such that

$$\mathbb{E}(F_n) = \frac{\alpha_1 n}{\sqrt{\log n}}(1+o(1)),\ \mathbb{E}(G_n) = \frac{\alpha_2 n}{\log n}(1+o(1)),\ \mathbb{E}(J_n) = \frac{\alpha_3 n}{\log n}(1+o(1)),$$

respectively, and

$$\frac{F_n}{n/\sqrt{\log n}} \xrightarrow{P} \alpha_1,\ \frac{G_n}{n/\log n} \xrightarrow{P} \alpha_2,\ \text{and}\ \frac{J_n}{n/\log n} \xrightarrow{P} \alpha_3\ ?$$

In order to prove such estimates, it seems essential to gain a better understanding of the random variables $P_{\mathrm{bst}}(T_k)^{-1}$ and $|\mathrm{Aut}(T_k)|$, in particular their distributions further away from the mean values, for random binary search trees or uniformly random ordered binary trees T_k of size k.

References

1. Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Highly expressive query languages for unordered data trees. *Theory of Computing Systems*, 57(4):927–966, 2015.
2. Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley series in computer science / World student series edition. Addison-Wesley, 1986.
3. David Aldous. Asymptotic fringe distributions for general families of random trees. *The Annals of Applied Probability*, 1(2):228–266, 1991.
4. Miklós Bóna and Philippe Flajolet. Isomorphism and symmetries in random phylogenetic trees. *Journal of Applied Probability*, 46(4):1005–1019, 2009.
5. Iovka Boneva, Radu Ciucanu, and Slawek Staworko. Schemas for unordered XML on a DIME. *Theory of Compuing Systems*, 57(2):337–376, 2015.
6. Mireille Bousquet-Mélou, Markus Lohrey, Sebastian Maneth, and Eric Noeth. XML compression via DAGs. *Theory of Computing Systems*, 57(4):1322–1371, 2015.
7. Randal E. Bryant. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys*, 24(3):293–318, 1992.
8. Peter Buneman, Martin Grohe, and Christoph Koch. Path queries on compressed XML. In Johann Christoph Freytag et al., editors, *Proceedings of the 29th Conference on Very Large Data Bases, VLDB 2003*, pages 141–152. Morgan Kaufmann, 2003.
9. Florian Dennert and Rudolf Grübel. On the subtree size profile of binary search trees. *Combinatorics, Probability and Computing*, 19(4):561–578, 2010.
10. Luc Devroye. On the richness of the collection of subtrees in random binary search trees. *Information Processing Letters*, 65(4):195–199, 1998.
11. Luc Devroye and Svante Janson. Protected nodes and fringe subtrees in some random trees. *Electronic Communications in Probability*, 19:1–10, 2014.
12. Michael Drmota. *Random Trees: An Interplay Between Combinatorics and Probability*. Springer Publishing Company, Incorporated, 1st edition, 2009.

13. Qunqiang Feng and Hosam M. Mahmoud. On the variety of shapes on the fringe of a random recursive tree. *Journal of Applied Probability*, 47(1):191–200, 2010.
14. James Allen Fill. On the distribution of binary search trees under the random permutation model. *Random Structures & Algorithms*, 8(1):1–25, 1996.
15. Steven R. Finch and Gian-Carlo Rota. *Mathematical Constants*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2003.
16. Philippe Flajolet, Xavier Gourdon, and Conrado Martínez. Patterns in random binary search trees. *Random Structures & Algorithms*, 11(3):223–244, 1997.
17. Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.
18. Philippe Flajolet, Paolo Sipala, and Jean-Marc Steyaert. Analytic variations on the common subexpression problem. In *Proceedings of the 17th International Colloquium on Automata, Languages and Programming, ICALP 1990*, volume 443 of *Lecture Notes in Computer Science*, pages 220–234. Springer, 1990.
19. Markus Frick, Martin Grohe, and Christoph Koch. Query evaluation on compressed trees (extended abstract). In *Proceedings of the 18th Annual IEEE Symposium on Logic in Computer Science, LICS 2003*, pages 188–197. IEEE Computer Society Press, 2003.
20. Moses Ganardi, Danny Hucke, Markus Lohrey, and Louisa Seelbach Benkner. Universal tree source coding using grammar-based compression. *IEEE Transactions on Information Theory*, 65(10):6399–6413, 2019.
21. Cecilia Holmgren and Svante Janson. Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electronic Journal of Probability*, 20:1–51, 2015.
22. John C. Kieffer, En-Hui Yang, and Wojciech Szpankowski. Structural complexity of random binary trees. In *Proceedings of the 2009 IEEE International Symposium on Information Theory, ISIT 2009*, pages 635–639. IEEE, 2009.
23. Markus Lohrey, Sebastian Maneth, and Carl Philipp Reh. Compression of unordered XML trees. In *20th International Conference on Database Theory, ICDT 2017, March 21-24, 2017, Venice, Italy*, pages 18:1–18:17, 2017.
24. Mike Paterson and Mark N. Wegman. Linear unification. *Journal of Computer and System Sciences*, 16(2):158–167, 1978.
25. Dimbinaina Ralaivaosaona and Stephan G. Wagner. Repeated fringe subtrees in random rooted trees. In *Proceedings of the Twelfth Workshop on Analytic Algorithmics and Combinatorics, ANALCO 2015*, pages 78–88. SIAM, 2015.
26. Louisa Seelbach Benkner and Markus Lohrey. Average case analysis of leaf-centric binary tree sources. In *43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August 27-31, 2018, Liverpool, UK*, pages 16:1–16:15, 2018.
27. Louisa Seelbach Benkner and Stephan Wagner. On the collection of fringe subtrees in random binary trees, 2020. `https://arxiv.org/abs/2003.03323`.
28. Jie Zhang, En-Hui Yang, and John C. Kieffer. A universal grammar-based code for lossless compression of binary trees. *IEEE Transactions on Information Theory*, 60(3):1373–1386, 2014.
29. Sen Zhang, Zhihui Du, and Jason Tsong-Li Wang. New techniques for mining frequent patterns in unordered trees. *IEEE Transactions on Cybernetics*, 45(6):1113–1125, 2015.