On the Number of Distinct Fringe Subtrees in Binary Search Trees

Stephan Wagner
Institute of Discrete Mathematics, TU Graz, Austria
Department of Mathematics, Uppsala University, Sweden

Abstract

A fringe subtree of a rooted tree is a subtree that consists of a vertex and all its descendants. The number of distinct fringe subtrees in random trees has been studied by several authors, notably because of its connection to tree compaction algorithms. Here, we obtain a very precise result for binary search trees: it is shown that the number of distinct fringe subtrees in a binary search tree with n leaves is asymptotically equal to $\frac{c_1 n}{\log n}$ for a constant $c_1 \approx 2.4071298335$, both in expectation and with high probability. This was previously shown to be a lower bound, our main contribution is to prove a matching upper bound. The method is quite general and can also be applied to similar problems for other tree models.

2012 ACM Subject Classification Mathematics of computing → Enumeration; Theory of computation → Randomness, geometry and discrete structures; Theory of computation → Data compression

Keywords and phrases Fringe subtrees, binary search trees, tree compression, minimal DAG, asymptotics

Digital Object Identifier 10.4230/LIPIcs.AofA.2024.13

Funding Supported by the Swedish research council (VR), grant 2022-04030.

1 Introduction

A *fringe subtree* of a rooted tree is a subtree that consists of a vertex and all its descendants, see for instance Figure 1. Fringe subtrees of random trees have been studied quite thoroughly under different models of randomness. Typical results include limit theorems for the number of fringe subtrees of a given size or shape (we will use those as an auxiliary tool in this paper as well), see for example [12, 14]. Fringe subtrees are intrinsically related to *additive functionals* of rooted trees [14–16, 19, 24], which can in fact be seen as linear combinations of fringe subtree counts. There are general limit theorems for additive functionals under different assumptions, and many relevant quantities associated with trees can be expressed as additive functionals.

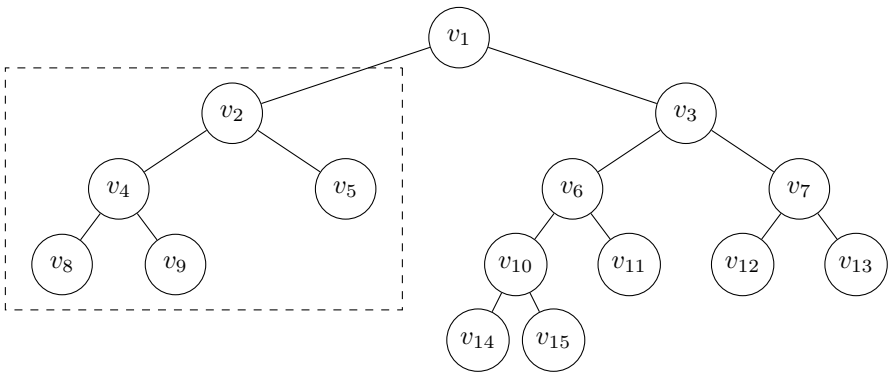

Figure 1 A binary tree. The fringe subtree rooted at v_2 is indicated by the dashed rectangle.

It is clear that an n-vertex tree has n fringe subtrees, one corresponding to each of its vertices. Usually, some of these will be identical/isomorphic as rooted trees, so the number of *distinct fringe subtrees* is generally smaller. In most of this paper, fringe subtrees will be considered identical if they are the same as plane trees (i.e., the order of the children of a vertex matters). The vertex labels are ignored. Otherwise, we regard them as distinct. There are however also other possible notions of distinctness that will be mentioned briefly in the final section.

The number of distinct fringe subtrees is connected to *tree compression*: in a fundamental algorithm to compress trees, vertices whose associated fringe subtrees have the same shape are merged to form what is called the *minimal directed acyclic graph (DAG)*. The precise shape of the tree can be recovered from the minimal DAG. Consider the tree in Figure 1 for a simple example: note that the fringe subtrees rooted at v_2 and v_6 are identical, so they are merged. For the same reason, v_4, v_7, v_{10} are merged as their fringe subtrees are identical in shape. Figure 2 shows the minimal DAG associated with the tree in Figure 1. Observe that the number of vertices of the minimal DAG is precisely the number of distinct fringe subtrees.

There are various applications of this compression technique by means of minimal DAGs. Let us mention XML compression and querying [5, 11], symbolic model checking [4] and compiler construction [1, Chapter 6.1 and 8.5] as notable examples. It is therefore of natural interest in computer science to analyse the extent to which the number of vertices is reduced by compressing a tree to its minimal DAG.

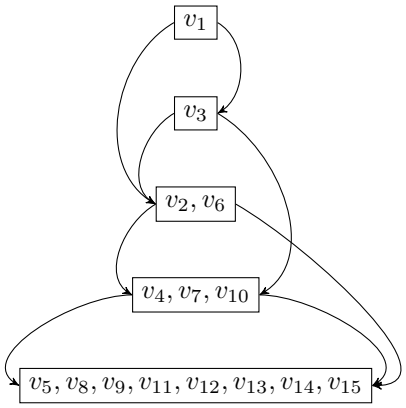

Figure 2 The minimal DAG associated with the tree in Figure 1. The vertices of the original tree that are compressed to a single vertex are listed.

For simply generated trees, it was shown by Flajolet, Sipala and Steyaert [10] that the expected size of the minimal DAG is of order $\frac{n}{\sqrt{\log n}}$. For instance, the average number of vertices in the minimal DAG associated with a uniformly random binary tree (a tree in which every internal vertex has precisely two children) with n leaves is asymptotically equal to $\frac{2n}{\sqrt{\pi \log_4 n}}$. It was also proven (see [22]) that this does not only hold in expectation, but also with high probability: in other words, with probability tending to 1, the size of the minimal DAG lies in an interval of the form $[(1-o(1))\frac{2n}{\sqrt{\pi \log_4 n}}, (1+o(1))\frac{2n}{\sqrt{\pi \log_4 n}}]$. The result of Flajolet, Sipala and Steyaert was further extended to Σ-labelled unranked trees in [3]. Moreover, an extension to the number of fringe subtrees that occur more than once

or generally at least a fixed number of times was considered in [20]. Interestingly, periodic fluctuations start to occur in the asymptotics: the average number of trees that occur at least r times ($r \geq 2$) as a fringe subtree is asymptotically

$$\psi_r(\log n)\frac{n}{(\log n)^{3/2}} + O\Big(\frac{n}{(\log n)^{5/2}}\Big) \tag{1}$$

for a positive periodic function ψ_r (see [20, Theorem 5.1] for the precise statement).

In this paper, we will be concerned with the model of random *binary search trees*. We consider binary trees where all internal vertices have two children: a left child and a right child. In the following, the *size* of a binary tree will always be the number of leaves; the number of internal vertices is always one less. In the probabilistic model that we study, a binary search tree is built from a random permutation of the numbers $1, 2, \dots, n$. These numbers are stored in the internal vertices of the tree in such a way that all numbers less than the root label are in the left branch, while all numbers greater than the root label are in the right branch. See Figure 3 for an example.

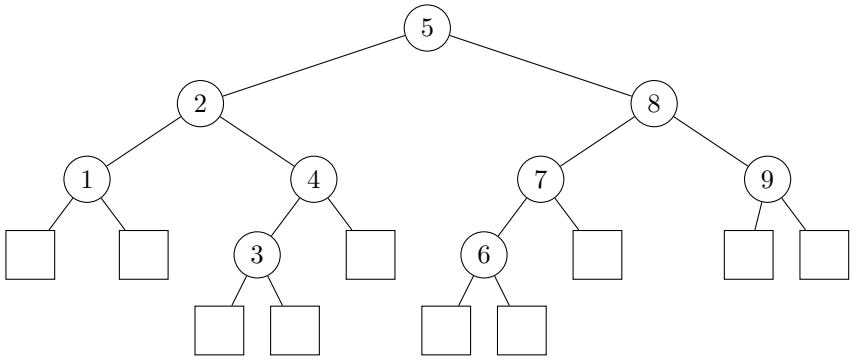

Figure 3 The binary search tree resulting from the permutation $(5, 2, 8, 4, 1, 7, 9, 3, 6)$. Internal vertices are indicated by circles, leaves by squares.

It is well known that this model is also essentially equivalent to that of *binary increasing trees* (binary trees with vertex labels that are increasing from the root to the leaves), see [7, Section 1.4.1]. For this and other types of increasing trees, the typical number of distinct fringe subtrees is of the order $\frac{n}{\log n}$ rather than $\frac{n}{\sqrt{\log n}}$. The main reason for this difference is the fact that the number of fringe subtrees with k vertices in an n-vertex tree is on average $\frac{n}{k^{3/2}}$ (asymptotically, up to a constant factor) for simply generated trees and $\frac{n}{k^2}$ for increasing trees.

The first result on binary search trees is due to Flajolet, Gourdon, and Martínez [9]. Letting F_n be the number of distinct fringe subtrees in a random binary search tree of size n, they proved that

$$\mathbb{E}(F_n) \leq \frac{(4\log 2)n}{\log n} + O\Big(\frac{n\log\log n}{(\log n)^2}\Big).$$

Devroye [6] provided a lower bound of the same order of magnitude (and also reproved the upper bound of Flajolet, Gourdon, and Martínez), showing that

$$\mathbb{E}(F_n) \geq \frac{(\log 3)n}{2\log n}(1 + o(1)).$$

The constant in the lower bound (i.e., $\frac{\log 3}{2} \approx 0.5493061443$) was improved to 0.6017824584 by Seelbach Benkner and Lohrey [21]. Seelbach Benkner and the present author [22] presented a general approach to proving results of this form. Specifically, it is shown in [22] that the number of distinct fringe subtrees is of order $\frac{n}{\sqrt{\log n}}$, both in expectation and with high probability, for simply generated trees/conditioned Bienaymé–Galton–Watson trees under various notions of what "distinct" means (e.g., distinct as plane trees, nonisomorphic as rooted trees). An analogous result for increasing trees holds with an order of magnitude of $\frac{n}{\log n}$ rather than $\frac{n}{\sqrt{\log n}}$. As a special case of the general approach, one obtains the following bounds with $c_1 \approx 2.4071298335$ and $c_2 = 4\log 2 \approx 2.7725887222$:

$$\frac{c_1 n}{\log n}(1+o(1)) \leq \mathbb{E}(F_n) \leq \frac{c_2 n}{\log n}(1+o(1)),$$

which further improves the lower bound (the upper bound is identical with that of Flajolet, Sipala and Steyaert). These inequalities hold not only for the expected value, but also with high probability. Even though upper and lower bound are of the same order of magnitude and the constants c_1 and c_2 in the upper and lower bound are fairly close to each other, it is clear that there is still a gap. The aim of this paper is to close the gap and show that the constant $c_1 = 4\sum_{k\geq 1} \frac{\log k}{(k+1)(k+2)}$ in the lower bound is in fact best possible. We will specifically prove the following theorem.

▶ **Theorem 1.** *For the constant* $c_1 = 4\sum_{k\geq 1} \frac{\log k}{(k+1)(k+2)}$*, we have*

$$\mathbb{E}(F_n) \sim \frac{c_1 n}{\log n}$$

as $n \to \infty$*. Moreover, we also have convergence in probability:*

$$\frac{F_n}{n/\log n} \xrightarrow{p} c_1.$$

The approach taken in [22, 23] leading to the lower bound will be briefly described in Section 3. The proof of the upper bound that is required for Theorem 1 will be presented afterwards in Section 4. Before that, we require some technical results on fringe subtrees in binary search trees as well as an important invariant that is called the *shape functional*. These auxiliary results will be outlined in the following section. The paper concludes with a brief discussion and an outlook to other problems to which the same method applies.

2 Preliminaries

Let us first fix some notation. We let $\mathfrak{B}_n$ be the set of binary trees of size n (for instance, Figure 4 shows the set $\mathfrak{B}_4$), and let $\mathcal{T}_n$ be a random binary tree of size n constructed according to the random binary search tree model. In this section, we gather some results on the distribution of different random variables associated with $\mathcal{T}_n$.

2.1 The binary search tree distribution and the shape functional

We first need some auxiliary results related to the probability distribution of the shape of binary search trees. Let T be a binary tree of size n, and let N_v be the number of *internal vertices* in the fringe subtree rooted at v. We say that a binary search tree has *shape* T if the binary tree obtained by ignoring all labels is T. It is well known that the probability that the shape of a random binary search tree of size n is exactly T can be expressed as

$$p(T) = \prod_v \frac{1}{N_v},$$

the product being over all internal vertices, see for example Fill [8]. The quantity

$$(n-1)!\prod_v \frac{1}{N_v}$$

is also the number of ways to label the internal vertices with labels $1, 2, \dots, n-1$ in an increasing fashion, i.e., in such a way that each vertex other than the root has a greater label than its parent [18, Section 5.1.4, Exercise 20].

Consider for example Figure 4: there are five possible shapes for binary search trees of size 4, occurring respectively with probability $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$ and $\frac{1}{3}$. The negative logarithm of $p(T)$, which can be expressed as

$$-\log p(T) = \sum_v \log N_v,$$

is called the *shape functional* of T [8] – to be more precise, it is the shape functional of the tree formed by the internal vertices.

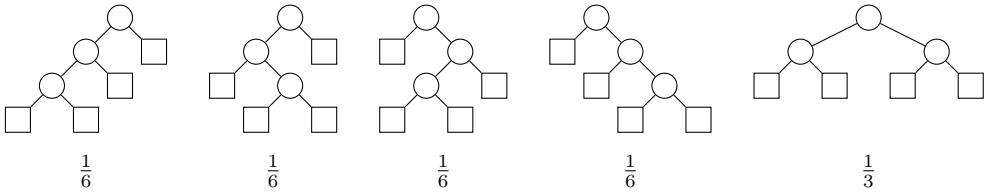

Figure 4 The five different binary trees with four leaves and their respective probabilities.

The distribution of the shape functional in random binary search trees was first studied by Fill in [8]. One can also obtain the following central limit theorem from an application of a general theorem on additive functionals due to Holmgren and Janson [14].

▶ **Lemma 2.** *Let the random variable L_n be defined by $L_n = -\log p(\mathcal{T}_n)$. We have*

$$\mathbb{E}(L_n) = \mu n + O(\log n),$$

where $\mu = \sum_{k=1}^{\infty} \frac{2\log k}{(k+1)(k+2)}$. Moreover, $\mathbb{V}(L_n) = \sigma^2 n + O(1)$ for a constant $\sigma^2 > 0$, and the centred and normalised random variable $\frac{L_n - \mu n}{\sigma\sqrt{n}}$ converges in distribution to a standard normal distribution.

For our purposes, the asymptotic formulas for mean and variance will already be sufficient, since all we actually need is that the random variable L_n is concentrated around its mean.

2.2 The total number of fringe subtrees of a given shape or size

The second key ingredient concerns fringe subtrees that belong to a specific set. As mentioned earlier, there are many results on the number of fringe subtrees of a specific shape or size. The following lemma, which is specifically geared towards our needs, was proven (in greater generality) in [22], see also [23, Lemma 2].

▶ **Lemma 3.** *Let a, ε be two fixed positive real numbers with $\varepsilon < \frac{1}{2}$. For every positive integer k, let $\mathfrak{S}_k \subseteq \mathfrak{B}_k$ be a set whose elements are binary trees of size k. Let $p_k = \sum_{B\in\mathfrak{S}_k} p(B)$ be the probability that a random binary search tree $\mathcal{T}_k$ of size k has a shape that belongs to $\mathfrak{S}_k$.*

Now let $Z_{n,k}$ denote the (random) number of fringe subtrees of size k in a random binary search tree $\mathcal{T}_n$ of size n whose shape belongs to $\mathfrak{S}_k$. Moreover, let $Y_{n,\varepsilon}$ denote the total number of arbitrary fringe subtrees of size greater than n^{ε}. Then

(a) $\mathbb{E}(Z_{n,k}) = \frac{2np_k}{k(k+1)}$ *for all* $k < n$,
(b) $\mathbb{V}(Z_{n,k}) = O(p_k n/k^2)$ *for all* k *with* $a \log n \le k \le n^\varepsilon$,
(c) $\mathbb{E}(Y_{n,\varepsilon}) = O(n^{1-\varepsilon})$, *and*
(d) *with high probability, the following statements hold simultaneously:*
  (i) $|Z_{n,k} - \mathbb{E}(Z_{n,k})| \le p_k^{1/2} k^{-1} n^{1/2+\varepsilon}$ *for all* k *with* $a \log n \le k \le n^\varepsilon$,
  (ii) $Y_{n,\varepsilon} \le n^{1-\varepsilon/2}$.

Equipped with this and the previous lemma, we now have the necessary tools to prove both a lower bound and an upper bound that will ultimately yield Theorem 1. The interval from $a \log n$ to n^ε in Lemma 3 is such that it covers the asymptotically relevant range in the proof of Theorem 1, where we split into several parts according to the fringe subtree size.

3 The lower bound

In this section, we give a brief account of the proof of the lower bound, see [22, 23], slightly adapted to our specific situation to provide more explicit error terms than in those papers.

The key idea to bound the number of distinct fringe subtrees from below is to only consider trees that are relatively "large". Specifically, we set $k_0 := \frac{1}{\mu}(\log n + (\log n)^{3/4})$, with μ as defined in Lemma 2, and only count fringe subtrees whose size is at least k_0, while all smaller fringe subtrees are ignored. It is clear that this will give us a lower bound on the total number of distinct fringe subtrees. It turns out that for this particular choice of k_0, most fringe subtrees of size $k \ge k_0$ only occur once in the tree.

In the setting of Lemma 3, let us choose $\mathfrak{S}_k$ to be the subset of $\mathfrak{B}_k$ consisting of those trees B for which $p(B) \le \exp(-\mu k + k^{2/3})$, or equivalently $-\log p(B) \ge \mu k - k^{2/3}$. We can apply Lemma 2 to show that this condition is satisfied with high probability for random binary trees. Indeed, the Chebyshev inequality yields

$$\mathbb{P}(L_k \le \mu k - k^{2/3}) \le \frac{\mathbb{V}(L_k)}{(\mathbb{E}(L_k) - \mu k + k^{2/3})^2},$$

which by Lemma 2 becomes

$$\mathbb{P}(L_k \le \mu k - k^{2/3}) = O(k^{-1/3}).$$

Thus we can conclude that p_k in Lemma 3 is $1 - O(k^{-1/3})$ for our specific choice of $\mathfrak{S}_k$.

So the expected contribution of trees in $\mathfrak{S}_k$ for $k \ge k_0$ to the total fringe subtree count is, by part (a) of Lemma 3,

$$\begin{aligned}\sum_{k\ge k_0} \mathbb{E}(Z_{n,k}) &= \sum_{k\ge k_0} \frac{2np_k}{k(k+1)} = 2n \sum_{k\ge k_0} k^{-2}\big(1 - O(k^{-1/3})\big) \\ &= \frac{2n}{k_0}\big(1 - O(k_0^{-1/3})\big) = \frac{2\mu n}{\log n}\big(1 - O\big((\log n)^{-1/4}\big)\big). \end{aligned} \tag{2}$$

Moreover, part (d.i) of Lemma 3 guarantees that this is also valid not just in expectation, but also with high probability.

Now we show that there are very few duplicates (identical fringe subtrees) among these. For every $k \ge k_0$, let $Z_{n,k}^{(2)}$ be the number of pairs of identical fringe subtrees in a random binary search tree of size n whose shape is in $\mathfrak{S}_k$. We condition on the total number of fringe subtrees of size k, which we denote by $X_{n,k}$. Since every fringe subtree follows, conditioned on its size k, the probability distribution of a random binary search tree $\mathcal{T}_k$, we have

$$\mathbb{E}(Z_{n,k}^{(2)} \mid X_{n,k} = N) = \binom{N}{2} \sum_{B\in\mathfrak{S}_k} p(B)^2.$$

By the definition of $\mathfrak{S}_k$, this gives us

$$\mathbb{E}(Z_{n,k}^{(2)} \mid X_{n,k} = N) \leq \binom{N}{2} e^{-\mu k + k^{2/3}} \sum_{B \in \mathfrak{S}_k} p(B) \leq \binom{N}{2} e^{-\mu k + k^{2/3}}.$$

Clearly, $X_{n,k} \leq n$, so the law of total expectation gives us

$$\mathbb{E}(Z_{n,k}^{(2)}) \leq \binom{n}{2} e^{-\mu k + k^{2/3}}.$$

Summing over all $k \geq k_0$, we finally find that

$$\sum_{k \geq k_0} \mathbb{E}(Z_{n,k}^{(2)}) \leq \binom{n}{2} \sum_{k \geq k_0} e^{-\mu k + k^{2/3}} = O\Big(n^2 e^{-\mu k_0 + k_0^{2/3}}\Big) = O\Big(n e^{-(\log n)^{3/4} + O((\log n)^{2/3})}\Big).$$

This shows that $\sum_{k \geq k_0} Z_{n,k}^{(2)}$ is (in expectation) negligible compared to $\sum_{k \geq k_0} Z_{n,k}$ (see (2)). By a standard application of the Markov inequality, this also applies with high probability.

Note that $Z_{n,k} - Z_{n,k}^{(2)}$ is a lower bound on the number of distinct fringe subtrees whose shape is in $\mathfrak{S}_k$: a shape that occurs r times contributes $r - \binom{r}{2} = \frac{r(3-r)}{2} \leq 1$ to this quantity. Moreover, the number of distinct fringe subtrees whose shape belongs to $\mathfrak{S}_k$ for some $k \geq k_0$ clearly provides a lower bound on the overall number of distinct fringe subtrees F_n, so we can conclude that

$$F_n \geq \sum_{k \geq k_0} \big(Z_{n,k} - Z_{n,k}^{(2)}\big) = \frac{2\mu n}{\log n}\big(1 - O\big((\log n)^{-1/4}\big)\big),$$

both in expectation and with high probability.

4 The upper bound

Let us now move on to the upper bound. We can express the number of distinct fringe subtrees as a sum of indicators. For every binary tree B, let $I_n(B)$ be the indicator random variable for the event that a random binary search tree of size n has a fringe subtree whose shape is B. With this definition, it is clear that

$$F_n = \sum_{k \geq 1} \sum_{B \in \mathfrak{B}_k} I_n(B).$$

The key to proving the upper bound that yields Theorem 1 is to split this sum into several parts and analyse their contributions. Specifically, the three regions are defined as follows:

- Small: $k \leq k_1 := \frac{1}{2} \log_4 n$;
- Medium: $k_1 < k \leq k_2 := \frac{1}{\mu}(\log n - (\log n)^{3/4})$, with μ as defined in Lemma 2;
- Large: $k_2 < k$.

This cutting technique is also the main idea behind many of the previously mentioned results on the quantity F_n. The novel contribution of this paper lies mainly in the middle region and its precise analysis.

4.1 Bounding the contribution of small fringe subtrees

This part is the easiest: clearly the contribution of trees whose size is at most $k_1 = \frac{1}{2} \log_4 n$ to the random variable F_n is no greater than the total number of distinct binary trees whose size is at most k_1. Since the number of possible trees for every given size k is a Catalan number (thus $|\mathfrak{B}_k| = \frac{1}{k}\binom{2k-2}{k-1} = O(4^k)$), we immediately obtain the (deterministic) upper bound

$$\sum_{k \leq k_1} \sum_{B \in \mathfrak{B}_k} I_n(B) \leq \sum_{k \leq k_1} |\mathfrak{B}_k| = O(4^{k_1}) = O(\sqrt{n}),$$

which renders all these trees negligible.

4.2 Bounding the contribution of medium-sized fringe subtrees

In the medium region, we have to perform a more careful analysis, separating trees not only by their size but also the value of their shape functional. We will split into trees with "large" shape functional and thus (relatively) low probability to occur as a fringe subtree, and trees with "small" shape functional, which have a comparatively high probability to occur. For the former, we show that the expected total number of occurrences is too low to make a significant contribution, while for the latter we prove that there are not enough distinct trees with sufficiently small shape functional to contribute to the main term of the asymptotics.

Let us now make this precise. For an integer k in the range $k_1 < k \le k_2$, let us define a partition of $\mathfrak{B}_k$ into two subsets (depending on n) as follows:

- $\mathfrak{B}_k^1$ contains all trees $B \in \mathfrak{B}_k$ with the property that $p(B) \le \frac{k^3}{n}$,
- $\mathfrak{B}_k^2$ contains all remaining trees in $\mathfrak{B}_k$.

Lemma 2 can be used to show that it is unlikely for the shape of a random binary search tree $\mathcal{T}_k$ to be in $\mathfrak{B}_k^1$: the inequality $p(\mathcal{T}_k) \le \frac{k^3}{n}$ can be rewritten as $e^{-L_k} \le \frac{k^3}{n}$, or $L_k \ge \log n - 3\log k$. This time, the Chebyshev inequality yields

$$\mathbb{P}(L_k \ge \log n - 3\log k) \le \frac{\mathbb{V}(L_k)}{(\log n - 3\log k - \mathbb{E}(L_k))^2}.$$

For $k \le k_2$, we have $\log n - 3\log k - \mathbb{E}(L_k) = \log n - \mu k + O(\log\log n) \ge (\log n)^{3/4} + O(\log\log n)$, thus (by Lemma 2)

$$\mathbb{P}(L_k \ge \log n - 3\log k) = O\Big(\frac{k}{(\log n)^{3/2}}\Big).$$

So if we set $\mathfrak{S}_k = \mathfrak{B}_k^1$ in Lemma 3, then it follows that

$$p_k = \sum_{B\in\mathfrak{B}_k^1} p(B) = O\Big(\frac{k}{(\log n)^{3/2}}\Big).$$

Consequently, by part (d.i) of Lemma 3, we have, with high probability,

$$\begin{aligned}\sum_{k_1<k\le k_2}\sum_{B\in\mathfrak{B}_k^1} I_n(B) &\le \sum_{k_1<k\le k_2} Z_{n,k}\\ &\le \sum_{k_1<k\le k_2}\Big(\frac{2np_k}{k(k+1)} + \frac{p_k^{1/2}n^{1/2+\varepsilon}}{k}\Big)\\ &= O\Big(\frac{n}{(\log n)^{3/2}}\Big).\end{aligned}$$

Observe that this also holds in expectation (even without the term $\frac{p_k^{1/2}n^{1/2+\varepsilon}}{k}$) by part (a) of Lemma 3.

For the remaining part, we prove that there are comparatively few trees in the set $\mathfrak{B}_k^2$ as compared to $\mathfrak{B}_k^1$, even though the majority of the probability mass lies with $\mathfrak{B}_k^2$. Specifically, we bound the contribution as follows: for every $B \in \mathfrak{B}_k^2$, we have $p(B) \ge \frac{k^3}{n}$ by definition and thus

$$\sum_{B\in\mathfrak{B}_k^2} I_n(B) \le \sum_{B\in\mathfrak{B}_k^2} 1 \le \sum_{B\in\mathfrak{B}_k^2} \frac{np(B)}{k^3}.$$

Now by definition of $p(B)$, we have $\sum_{B\in\mathfrak{B}_k^2} p(B) \le \sum_{B\in\mathfrak{B}_k} p(B) = 1$, thus

$$\sum_{B\in\mathfrak{B}_k^2} I_n(B) \le \frac{n}{k^3}.$$

This inequality even holds deterministically. Finally, summing over all k in our range yields

$$\sum_{k_1<k\le k_2} \sum_{B\in\mathfrak{B}_k^2} I_n(B) \le \sum_{k_1<k\le k_2} \frac{n}{k^3} = O\Big(\frac{n}{(\log n)^2}\Big).$$

Both this and the previous error bound that we found for $\mathfrak{B}_k^1$ are negligible compared to the term of order $\frac{n}{\log n}$ that we will obtain in the final case.

4.3 Bounding the contribution of large fringe subtrees

Finally, we look at large fringe subtrees whose size is greater than $k_2 = \frac{1}{\mu}(\log n - (\log n)^{3/4})$. Here, we apply Lemma 3 with $\mathfrak{S}_k = \mathfrak{B}_k$ for all $k > k_2$ to show that the total number of such subtrees (regardless of whether they are distinct or not) is equal to

$$\sum_{k_2<n\le n^\varepsilon} Z_{n,k} + Y_{n,\varepsilon} = \sum_{k_2<n\le n^\varepsilon} \frac{2n}{k(k+1)} + O(n^{1-\varepsilon/2}) = \frac{2n}{k_2}(1+o(1)) = \frac{2\mu n}{\log n}(1+o(1)),$$

both in expectation and with high probability. This term dominates the contribution of the two other cases, so we end up with

$$F_n = \sum_B I_n(B) \le \frac{c_1 n}{\log n}(1+o(1)), \tag{3}$$

both in expectation and with high probability. Since the matching lower bound was already provided (see Section 3), this completes the proof of Theorem 1. ◀

5 Discussion and outlook

As the proof shows, the dominant contribution to the number of distinct fringe subtrees comes from those fringe subtrees that are "large" – specifically, whose size is at least approximately $\frac{1}{\mu}\log n$. The significance of this value is as follows: above this threshold, a typical binary search tree B of size k satisfies $p(B) = o(1/n)$; as a result, the number of duplicates among the fringe subtrees of size k in $\mathcal{T}_n$ becomes insignificant, and the contribution to the number of distinct fringe subtrees is essentially just the number of fringe subtrees. Below the threshold of $\frac{1}{\mu}\log n$, it is precisely the opposite: we have $p(B) = \omega(1/n)$ (i.e., $np(B) \to \infty$) for a typical binary search tree B of size k, which ultimately leads to a negligible contribution.

Further examples of the same kind are presented in [22]: in all these examples, there are upper and lower bounds of the same order of magnitude, namely $\frac{n}{\sqrt{\log n}}$ or $\frac{n}{\log n}$. However, in most of them the constants in the bounds do not quite match.

The same technique as presented in this paper can be applied to other examples of this kind to determine the precise asymptotic behaviour of many similar quantities. To this end, one needs sufficient information on the behaviour of the analogue of the quantity $p(B)$ – specifically, a result of the same type as Lemma 2 is required.

Let us give one concrete example: the number of nonisomorphic fringe subtrees in recursive trees was studied recently by Bodini, Genitrini, Gittenberger, Larcher and Naima [2]. For this quantity, the analogue of $p(B)$ is the probability that a recursive tree of a given size

is isomorphic to a fixed unlabelled tree. The general central limit theorem for additive functionals of recursive trees due to Holmgren and Janson [14] can be applied to show that the analogue of Lemma 2 does indeed hold. At the end of the procedure, we have the following result:

▶ **Theorem 4.** *The number of nonisomorphic fringe subtrees in a random recursive tree with n vertices is asymptotically equal to $\frac{c_3 n}{\log n}$, where the constant c_3 is approximately equal to* 0.9136401430*, both in expectation and with high probability.*

The constant c_3 already appears in the lower bound in [22, Theorem 16]. The numerical computation of this constant is discussed there as well. This and further examples will be considered in the full version of this paper in a broader context.

Let us finally mention an interesting connection to the concept of *entropy* for random tree models (compare [13, 17]): recall that the constant μ in our main theorem stems from the mean of the quantity L_n (the shape functional of a random binary search tree) as given in Lemma 2. Note that we have

$$\mathbb{E}(L_n) = \mathbb{E}(-\log p(\mathcal{T}_n)) = -\sum_{B \in \mathfrak{B}_n} p(B) \log p(B),$$

which can be interpreted as the entropy of the random variable $\mathcal{T}_n$. Thus the growth constant for the number of distinct fringe subtrees is directly connected to the growth constant for the entropy. A similar interpretation is possible in other examples, such as Theorem 4.

References

1 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.

2 Olivier Bodini, Antoine Genitrini, Bernhard Gittenberger, Isabella Larcher, and Mehdi Naima. Compaction for two models of logarithmic-depth trees: analysis and experiments. *Random Structures Algorithms*, 61(1):31–61, 2022. doi:10.1002/rsa.21056.

3 Mireille Bousquet-Mélou, Markus Lohrey, Sebastian Maneth, and Eric Noeth. XML compression via DAGs. *Theory of Computing Systems*, 57(4):1322–1371, 2015. doi:10.1145/2448496.2448506.

4 Randal E. Bryant. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys*, 24(3):293–318, 1992. doi:10.1145/136035.136043.

5 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. doi:10.1016/B978-012722442-8/50021-5.

6 Luc Devroye. On the richness of the collection of subtrees in random binary search trees. *Information Processing Letters*, 65(4):195–199, 1998. doi:10.1016/S0020-0190(97)00206-8.

7 Michael Drmota. *Random Trees: An Interplay Between Combinatorics and Probability*. Springer, 1st edition, 2009.

8 James Allen Fill. On the distribution of binary search trees under the random permutation model. *Random Structures & Algorithms*, 8(1):1–25, 1996. doi:10.1002/(SICI)1098-2418(199601)8:1<1::AID-RSA1>3.0.CO;2-1.

9 Philippe Flajolet, Xavier Gourdon, and Conrado Martínez. Patterns in random binary search trees. *Random Structures & Algorithms*, 11(3):223–244, 1997. doi:10.1002/(SICI)1098-2418(199710)11:3<223::AID-RSA2>3.0.CO;2-2.

10 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. doi:10.1007/BFb0032034.

11 Markus Frick, Martin Grohe, and Christoph Koch. Query evaluation on compressed trees. In *Proceedings of the 18th Annual IEEE Symposium on Logic in Computer Science, LICS 2003*, pages 188–197. IEEE Computer Society Press, 2003. doi:10.1109/LICS.2003.1210058.

12 Michael Fuchs. Limit theorems for subtree size profiles of increasing trees. *Combinatorics, Probability and Computing*, 21(3):412–441, 2012. doi:10.1017/S096354831100071X.

13 Zbigniew Gołębiewski, Abram Magner, and Wojciech Szpankowski. Entropy and optimal compression of some general plane trees. *ACM Trans. Algorithms*, 15(1):Art. 3, 23, 2019. doi:10.1145/3275444.

14 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. doi:10.1214/EJP.v20-3627.

15 Cecilia Holmgren, Svante Janson, and Matas Šileikis. Multivariate normal limit laws for the numbers of fringe subtrees in m-ary search trees and preferential attachment trees. *Electron. J. Combin.*, 24(2):Paper No. 2.51, 49 pp., 2017. doi:10.37236/6374.

16 Svante Janson. Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton-Watson trees. *Random Struct. Algorithms*, 48(1):57–101, 2016. doi:10.1002/rsa.20568.

17 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. doi:10.1109/ISIT.2009.5205704.

18 Donald E. Knuth. *The art of computer programming. Vol. 3*. Addison-Wesley, Reading, MA, 1998.

19 Dimbinaina Ralaivaosaona and Stephan Wagner. A central limit theorem for additive functionals of increasing trees. *Combin. Probab. Comput.*, 28(4):618–637, 2019. doi:10.1017/s0963548318000585.

20 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. doi:10.1137/1.9781611973761.7.

21 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. doi:10.4230/LIPIcs.MFCS.2018.16.

22 Louisa Seelbach Benkner and Stephan Wagner. Distinct fringe subtrees in random trees. *Algorithmica*, 84(12):3686–3728, 2022. doi:10.1007/s00453-022-01013-y.

23 Louisa Seelbach Benkner and Stephan G. Wagner. On the collection of fringe subtrees in random binary trees. In Yoshiharu Kohayakawa and Flávio Keidi Miyazawa, editors, *LATIN 2020: Theoretical Informatics - 14th Latin American Symposium, São Paulo, Brazil, January 5-8, 2021, Proceedings*, volume 12118 of *Lecture Notes in Computer Science*, pages 546–558. Springer, 2020. doi:10.1007/978-3-030-61792-9_43.

24 Stephan Wagner. Central limit theorems for additive tree parameters with small toll functions. *Combin. Probab. Comput.*, 24(1):329–353, 2015. doi:10.1017/S0963548314000443.