

AVERAGE CASE ANALYSIS OF LEAF-CENTRIC BINARY TREE SOURCES

LOUISA SEELBACH BENKNER, MARKUS LOHREY AND STEPHAN WAGNER

ABSTRACT. We study the average number of distinct fringe subtrees in random trees generated by leaf-centric binary tree sources as introduced by Zhang, Yang and Kieffer in [28]. A leaf-centric binary tree source induces for every $n \geq 2$ a probability distribution on the set of binary trees with n leaves. We generalize a result by Flajolet, Gourdon and Martinez [12] and Devroye [8], according to which the average number of distinct fringe subtrees in a random binary search tree of size n is in $\Theta(n/\log n)$, as well as a result by Flajolet, Sipala and Steyaert [14], according to which the number of distinct fringe subtrees in a uniformly random binary tree of size n is in $\Theta(n/\sqrt{\log n})$.

1. INTRODUCTION

A *fringe subtree* of a rooted tree is a subtree which consists of a node and all its descendants. Fringe subtrees are a natural object of study in the context of random trees, and there is a great variety of results for various random tree models, see e.g. [2, 7, 9, 11].

Fringe subtrees are of particular interest in computer science. A widely used lossless compression method for rooted trees is to represent a tree as a directed acyclic graph (DAG), which is obtained by merging nodes that are roots of identical fringe subtrees (see Figure 1 for an example). 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 applications in many areas of computer science, as for example in compiler construction [1, Chapter 6.1 and 8.5], unification [22], symbolic model checking (binary decision diagrams) [5], information theory [16, 28] and XML compression and querying [6, 15].

In this work, we consider the problem of deriving asymptotic estimates for the expected number of distinct fringe subtrees in random binary trees. Binary trees will be rooted ordered trees where every node is either a leaf or has exactly two children (a left child and a right child). We define the size of a binary tree as the number of leaves.¹ So far, the average number of distinct fringe subtrees has mainly been studied with respect to two probability distributions: uniformly random binary trees, and random binary search trees. A uniformly random ordered binary tree of size n is a random tree whose corresponding probability distribution is the uniform probability distribution on the set of ordered binary trees of size n . In [14], Flajolet, Sipala and Steyaert proved that the expected number of distinct 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 $c = 2\sqrt{\ln 4/\pi}$. This result of Flajolet et al. was extended to unranked labeled trees in [4] (for a different constant c).

A short version of this paper appeared in the Proceedings of MFCS 2018 [25]. The first and the second author have been supported by the DFG research project LO 748/10-1 (QUANT-KOMP). The third author is supported by the Knut and Alice Wallenberg Foundation, grant KAW 2017.0112.

¹This has the technical advantage that for every n there is a binary tree of size n .

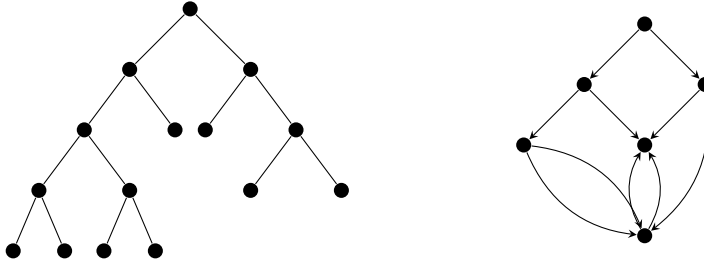


FIGURE 1. A binary tree (left) and its corresponding minimal DAG (right).

Moreover, an alternative proof to the result of Flajolet et al. was presented in [23] in the context of simply generated families of trees.

Another important type of random trees are 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. [10]. In [12], Flajolet, Gourdon and Martinez proved that the expected number of distinct fringe subtrees in a random binary search tree of size n is in $\mathcal{O}(n/\log n)$. This result was improved in [8] by Devroye, who showed that the asymptotics $\Theta(n/\log n)$ holds, and reproved in a recent paper by Bodini et al. [3].

The goal of this work is to generalize the results from [8, 12, 14] by proving upper and lower bounds for the average number of distinct fringe subtrees that hold for large classes of distributions. A very general concept to model probability distributions on the set of binary trees of size n was presented by Kieffer et al. in [17] (see also [28]), where the authors introduce *leaf-centric binary tree sources* as an extension of the classical notion of an information source on finite sequences of binary trees. A leaf-centric binary tree source is induced by a mapping $\sigma : \mathbb{N} \times \mathbb{N} \rightarrow [0, 1]$ that satisfies $\sum_{i=1}^{n-1} \sigma(i, n-i) = 1$ for every $n \geq 2$. In other words, σ restricted to the set of ordered pairs $S_n := \{(i, n-i) \mid 1 \leq i \leq n-1\}$ is a probability mass function for every $n \geq 2$. A leaf-centric binary tree source randomly generates a binary tree of size n as follows: we start with a single root node labeled with n and randomly choose a pair $(i, n-i)$ according to the distribution σ on S_n . Then, a left (resp. right) child labeled with i (resp. $n-i$) is attached to the root, and the process is repeated with these two nodes. The process stops at nodes with label 1 (and in the end, the node labels are omitted). In particular, such a mapping σ induces a function P_σ which restricts to a probability mass function on the set of binary trees of size n for every $n \geq 1$.

The binary search tree model is the leaf-centric binary tree source where the mapping σ corresponds to the uniform probability distribution on the set S_n for every $n \geq 2$ (Example 2.1). Moreover, the uniform probability distribution on the set of binary trees of size n can be modeled as a leaf-centric binary tree source as well (Example 2.2). Another well-known leaf-centric tree source is the binomial random tree model [17] (see Example 2.3), where the mapping σ corresponds to a binomial distribution on S_n .

Let $F_{n,\sigma}$ denote the (random) number of distinct (ordered) fringe subtrees occurring in a random tree of size n generated by the leaf-centric binary tree source corresponding to the mapping σ . In this work, we investigate the average $\mathbb{E}(F_{n,\sigma})$ under certain conditions on the mapping σ , that is, we assume that the mapping σ satisfies certain properties and then derive upper and lower bounds on $\mathbb{E}(F_{n,\sigma})$ with respect to these conditions. In particular, we consider four classes of leaf-centric

binary tree sources, listed below, for which we will be able to derive asymptotic upper or lower bounds on $\mathbb{E}(F_{n,\sigma})$. Here and in the rest of the paper, increasing/decreasing functions are not necessarily strictly monotone: if f is increasing, then $f(x) \geq f(y)$ whenever $x > y$, but not necessarily $f(x) > f(y)$ (and analogously for decreasing functions).

- (i) A leaf-centric tree source with mapping σ is called *ψ -upper-bounded* for a decreasing function $\psi : \mathbb{R} \rightarrow (0, 1]$ if there is a constant N such that for all $1 \leq i \leq n - 1$ and $n \geq N$:

$$\sigma(i, n - i) \leq \psi(n).$$

- (ii) A leaf-centric tree source with mapping σ is called *ϕ -weakly-balanced* for a decreasing function $\phi : \mathbb{N} \rightarrow (0, 1]$ if there is a constant $\gamma \in (0, \frac{1}{2})$ and an integer N such that for all $n \geq N$:

$$\sum_{\gamma n \leq i \leq (1-\gamma)n} \sigma(i, n - i) \geq \phi(n).$$

- (iii) A leaf-centric tree source with mapping σ is *ϑ -strongly-balanced* for a decreasing function $\vartheta : \mathbb{R} \rightarrow (0, 1]$ if there is a constant $\gamma \in (0, \frac{1}{2})$, an integer N and a constant C such that for all $n \geq N$ and $C \leq r \leq \lceil \gamma n \rceil$:

$$\sum_{r \leq i \leq n-r} \sigma(i, n - i) \geq \vartheta(r).$$

- (iv) A leaf-centric tree source with mapping σ is called *ξ -unbalanced* for a decreasing function ξ if there is a constant $\gamma \in (0, \frac{1}{2})$, an integer N and a constant C such that for all $n \geq N$ and $C \leq r \leq \lceil \gamma n \rceil$:

$$\sum_{r \leq i \leq n-r} \sigma(i, n - i) \leq \xi(r).$$

Property (i) quantifies how close σ is to the binary search tree model. The latter is ψ -upper-bounded for $\psi(n) = 1/(n - 1)$, which is the smallest function ψ for which (i) makes sense (since the sum over all $\sigma(i, n - i)$ for $1 \leq i \leq n - 1$ has to be one). Property (ii) generalizes the concept of balanced binary tree sources from [16, 28]: When randomly constructing a binary tree with respect to a leaf-centric tree source of type (ii), the probability that the current weight is roughly equally split among the two children is bounded below by the function ϕ . Therefore, for slowly decreasing functions ϕ , balanced trees are preferred by this model. Property (iii) is a stronger constraint than property (ii): Every leaf-centric tree source that is ϑ -strongly-balanced for a function ϑ is also ϕ -weakly-balanced for the function ϕ defined by $\phi(n) = \vartheta(\lceil \gamma n \rceil)$.

As our main results, we obtain the following asymptotic bounds on the expected number of distinct fringe subtrees for these classes of leaf-centric binary tree sources. Let σ be a leaf-centric tree source.

- (a) If σ is ψ -upper-bounded, then $\mathbb{E}(F_{n,\sigma}) \in \mathcal{O}(n\psi(\log n))$ (Theorem 3.3, where as slightly more general result is shown).
- (b) If ψ -upper-bounded and there are constants $N \in \mathbb{N}$, $\rho < 1$ such that $\psi(x) < \rho$ for all $x \geq N$, then $\mathbb{E}(F_{n,\sigma}) \in \Omega(\frac{n}{\log n})$ (Theorem 4.1).
- (c) If σ is ϕ -weakly-balanced, then $\mathbb{E}(F_{n,\sigma}) \in \mathcal{O}(\frac{n}{\phi(n)\log n})$ (Theorem 3.8).
- (d) If σ is ϑ -strongly-balanced, then $\mathbb{E}(F_{n,\sigma}) \in \mathcal{O}(\frac{n}{\vartheta(\log n)\log n})$ (Theorem 3.13).
- (e) If σ is ξ -unbalanced and some additional technical conditions are satisfied, then $\mathbb{E}(F_{n,\sigma}) \in \Omega(\frac{n}{\xi(\log n)\log n})$ (Theorem 4.10).

The precise statements and asymptotic bounds including leading constants for the main term are given in the respective theorems. These results generalize the results

from [8, 12] on the binary search tree model in the following sense: from each of the results (a), (c) and (d) (respectively, (b) and (e)), we obtain the upper bound $\mathcal{O}(n/\log n)$ (respectively, the lower bound $\Omega(n/\log n)$) on the expected number of distinct fringe subtrees occurring in a random binary search tree of size n . Furthermore, results (d) and (e) combined yield that the expected number of distinct fringe subtrees in a uniformly random binary tree of size n is in $\Theta(n/\sqrt{\log n})$, as shown in [14, 23]. Finally, results (b) and (c) imply the asymptotic bound $\Theta(n/\log n)$ on the expected number of distinct fringe subtrees in a random binary tree of size n generated according to the binomial random tree model.

The general message of our results could be stated as follows: if a leaf-centric tree source produces balanced trees with high (resp., small) probability (which is formalized by the above classes (ii), (iii) and (iv)), then the average number of distinct fringe subtrees gets smaller (resp., higher).

A short version of this work appeared in the Proceedings of MFCS 2018 [25]: The results (a), (b) and (c) already appeared in the conference version [25] (however, among others, several proofs needed for the respective results have been simplified, and leading constants in front of the main terms in the asymptotic bounds have been improved in this version). The results (d) and (e) are new contributions of this version.

Related work. Recently, in [27], the problem of estimating the number of distinct fringe subtrees in random trees was investigated in a more general setting, i.e., under a generalized interpretation of “distinctness”, which allows for many different interpretations of what “distinct” fringe subtrees are. While in the aforementioned results from [8, 12, 14, 23], two fringe subtrees were considered as “distinct” if they are distinct as ordered trees, the generalized interpretation of distinctness from [27] allows for example to count the number of distinct *unordered* fringe subtrees in ordered random trees of size n , i.e., two fringe subtrees are considered as “distinct” if their corresponding unordered trees are distinct. Amongst others, as special cases of more general theorems, it is shown in [27] that the number of distinct unordered fringe subtrees in a uniformly random binary tree of size n is in $\Theta(n/\sqrt{\log n})$ and that the number of distinct unordered fringe subtrees in a random binary search tree of size n is in $\Theta(n/\log n)$, where both asymptotic results hold not only in expectation, but also with high probability, i.e., with probability tending to 1 as $n \rightarrow \infty$ (for the precise upper and lower bounds including leading constants, see [26, 27]). Moreover, the results from [14, 23] (for the uniform distribution) and [8, 12] (for the binary search tree model) were generalized in [27] in the sense that these results hold not only in expectation, but also with high probability.

Leaf-centric binary tree sources have been used in information theory in order to obtain universal tree coders [16, 17, 21, 28]. In particular, it has been shown in [28] that for every leaf-centric tree source σ with $\mathbb{E}(F_{n,\sigma}) \in o(n)$, the average-case redundancy of a certain tree encoder converges to zero; see [28] for precise definitions. Our upper bounds from Theorems 3.3, 3.8 and 3.13 yield large classes of leaf-centric tree sources with this property.

2. PRELIMINARIES

We use the classical Landau notations \mathcal{O} , o , Ω and ω . In the following, $\log x$ will denote the binary logarithm $\log_2 x$ of a positive real number x . Let \mathbb{N} denote the natural numbers without zero.

2.1. Binary Trees. We write \mathcal{T} for the set of (*full*) *binary trees*, i.e., of rooted ordered trees such that each node has either exactly two or zero children. The *size* of a binary tree t is the number of leaves of t and denoted by $|t|$. We use \mathcal{T}_n to

denote the set of binary trees which have exactly n leaves. It is well known that $|\mathcal{T}_n| = C_{n-1}$, where C_n denotes the n -th *Catalan number* [13]. We have

$$(1) \quad C_n = \frac{1}{n+1} \binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n^{3/2}}} (1 + \mathcal{O}(1/n)),$$

where the asymptotic growth follows from Stirling's formula (see e.g. [13]).

A *fringe subtree* of a tree is a subtree that consists of a node and all its descendants. For a node v of a binary tree $t \in \mathcal{T}$, let $t[v]$ denote the fringe subtree of t which is rooted at v . Furthermore, for an inner node v of t , let $t_\ell[v]$ (resp. $t_r[v]$) denote the fringe subtree rooted at v 's left (resp. right) child node. If v is the root node of t , we write t_ℓ and t_r for $t_\ell[v]$ and $t_r[v]$ and call t_ℓ (resp. t_r) the *left subtree* (resp. *right subtree*) of t .

2.2. Leaf-centric binary tree sources. Let \mathcal{L} denote the set of all functions $\sigma: \mathbb{N} \times \mathbb{N} \rightarrow [0, 1]$ that satisfy

$$(2) \quad \sum_{\substack{i, j \geq 1 \\ i+j=k}} \sigma(i, j) = 1$$

for every integer $k \geq 2$. A mapping $\sigma \in \mathcal{L}$ induces a probability mass function $P_\sigma: \mathcal{T}_n \rightarrow [0, 1]$ for every $n \geq 1$ in the following way: define $P_\sigma: \mathcal{T} \rightarrow [0, 1]$ inductively by

$$(3) \quad P_\sigma(t) = \begin{cases} 1 & \text{if } |t| = 1, \\ \sigma(|t_\ell|, |t_r|) \cdot P_\sigma(t_\ell) \cdot P_\sigma(t_r) & \text{otherwise.} \end{cases}$$

A tuple $(\mathcal{T}, (\mathcal{T}_n)_{n \in \mathbb{N}}, P_\sigma)$ with $\sigma \in \mathcal{L}$ is called a *leaf-centric tree source* [28, 17]. Furthermore, the following notation will be useful: for a function $\sigma \in \mathcal{L}$, we define $\sigma^*: \mathbb{N} \times \mathbb{N} \rightarrow [0, 1]$ as

$$(4) \quad \sigma^*(i, j) = \begin{cases} \sigma(i, j) + \sigma(j, i) & \text{if } i \neq j, \\ \sigma(i, j) & \text{if } i = j. \end{cases}$$

Note that $\sigma^*(i, j) \leq 1$ for all $i, j \in \mathbb{N}$ and that $\sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) = 1$ (where the sum goes over all integers in the interval $[1, n/2]$).

Intuitively, a leaf-centric binary tree source randomly generates a binary tree of size n as follows: we start at the root node and determine the sizes of the left and of the right subtree, where the probability that the left subtree is of size i for $i \in \{1, \dots, n-1\}$ (and consequently, the right subtree is of size $n-i$) is given by $\sigma(i, n-i)$. This process then recursively continues in the left and right subtree, i.e., the leaf-centric binary tree source then randomly generates a binary tree of size i as the left subtree and a binary tree of size $n-i$ as the right subtree. In particular, for $t \in \mathcal{T}$, we have

$$P_\sigma(t) = \prod_{v \text{ inner node of } t} \sigma(|t_\ell[v]|, |t_r[v]|).$$

Several well-known random tree models can be characterized as leaf-centric tree sources (see Figure 2 for an example of two binary trees together with the probabilities assigned to them by the leaf-centric tree sources from the following examples).

Example 2.1. The *binary search tree model* [10, 12] corresponds to the leaf-centric binary tree source defined by

$$\sigma_{bst}(i, n-i) = \frac{1}{n-1}$$

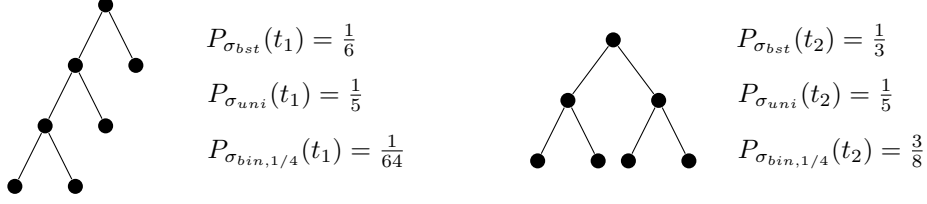


FIGURE 2. Two binary trees t_1 (left) and t_2 (right), with the probabilities assigned to them by the binary search tree model (Example 2.1), the uniform model (Example 2.2) and the binomial random tree model (Example 2.3) with $p = 1/4$.

for every $n \geq 2$ and $1 \leq i \leq n - 1$. This distribution over binary trees arises if a binary search tree of size n is built by inserting the keys $1, \dots, n$ according to a uniformly chosen random permutation on $1, \dots, n$ [10].

Example 2.2. Perhaps the most natural probability distribution on the set of binary trees \mathcal{T}_n is the uniform probability distribution, i.e., every binary tree $t \in \mathcal{T}_n$ is assigned the same probability $1/C_{n-1}$. This probability distribution corresponds to the leaf-centric binary tree source with

$$\sigma_{uni}(i, n - i) = \frac{C_{i-1}C_{n-i-1}}{C_{n-1}}$$

for every $n \geq 2$ and $1 \leq i \leq n - 1$ [17].

Example 2.3. Fix a constant $p \in (0, 1)$ and define

$$(5) \quad \sigma_{bin,p}(i, n - i) = p^{i-1}(1-p)^{n-i-1} \binom{n-2}{i-1}$$

for every $n \geq 2$ and $1 \leq i \leq n - 1$. This leaf-centric binary tree source corresponds to the *binomial random tree model*, which was studied in [17] for the case $p = 1/2$, and which is a slight variant of the *digital search tree model*, [10, 20].

In this paper, we are interested in the expected number of *distinct fringe subtrees* in random binary trees (see Figure 3 for an example of a binary tree and its distinct fringe subtrees). Formally, the number of distinct fringe subtrees occurring in a binary tree t can be defined as follows: We define an equivalence relation \sim on the set of nodes of t by $v \sim v'$ if and only if the two fringe subtrees $t[v]$ and $t[v']$ are identical as ordered binary trees. Let $[v]$ denote the equivalence class of node v with respect to this equivalence relation. Then the number of distinct fringe subtrees occurring in t equals the number of equivalence classes $|\{[v] \mid v \text{ node of } t\}|$. With $T_{n,\sigma}$ we denote a random tree of size n drawn from the set \mathcal{T}_n according to the probability mass function $P_\sigma: \mathcal{T}_n \rightarrow [0, 1]$ from (3), and with $F_{n,\sigma}$ we denote the (random) number of distinct fringe subtrees occurring in the random tree $T_{n,\sigma}$.

In the following, we investigate $F_{n,\sigma}$ under certain conditions on the mapping σ . That is, we will assume that the mapping σ satisfies certain properties and then derive upper and lower bounds on $\mathbb{E}(F_{n,\sigma})$.

In order to obtain upper bounds on $\mathbb{E}(F_{n,\sigma})$, we make use of a technique called *cut-point argument*, which was used before in several works for similar investigations (see for example [8, 12, 23, 26, 27]). The main idea of this cut-point technique is the following: Fix an integer $k > 0$. Let $t \in \mathcal{T}$ be a binary tree. The number of distinct fringe subtrees occurring in t equals

- (i) the number of distinct fringe subtrees of size larger than k occurring in t plus
- (ii) the number of distinct fringe subtrees of size at most k occurring in t .

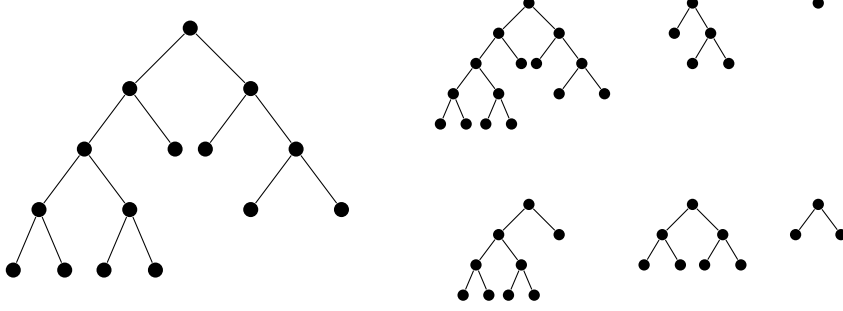


FIGURE 3. A binary tree (left) and its six distinct fringe subtrees (right).

The number (i) of distinct fringe subtrees of size larger than k is then bounded above by the number of *all fringe subtrees* in t of size larger than k ,² whereas the number (ii) of distinct fringe subtrees of size at most k is bounded above by the number of *all binary trees* of size at most k (irrespective of their occurrence in t). Note that this upper bound on number (ii) is a deterministic quantity; it is a sum of Catalan numbers. Let $Y_{n,k,\sigma}$ denote the (random) number of (all) fringe subtrees of size larger than k occurring in the random tree $T_{n,\sigma}$ (if σ is clear from the context, we shortly write $Y_{n,k}$ for $Y_{n,k,\sigma}$). In other words, $Y_{n,k,\sigma}$ is the following random variable:

$$(6) \quad Y_{n,k,\sigma} = |\{v \text{ node of } T_{n,\sigma} \mid |T_{n,\sigma}[v]| > k\}|.$$

Note that $Y_{n,k,\sigma} = 0$ for all $\sigma \in \mathcal{L}$ if $n \leq k$. Note that the definition of $Y_{n,k,\sigma}$ also makes sense if k is not an integer. Using the cut-point argument, we obtain the following upper bound on $\mathbb{E}(F_{n,\sigma})$.

Lemma 2.4. *Let $\sigma \in \mathcal{L}$. The number $F_{n,\sigma}$ of distinct fringe subtrees occurring in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(F_{n,\sigma}) \leq \mathbb{E}(Y_{n,k,\sigma}) + \sum_{i=0}^{k-1} C_i$$

with $Y_{n,k,\sigma}$ as defined in (6).

Thus, in Section 3, we will focus on obtaining upper bounds on $\mathbb{E}(Y_{n,k,\sigma})$ with respect to certain conditions on σ . The integer k is called the *cut-point*; it has to be chosen in a suitable way in order to obtain suitable upper bounds on $\mathbb{E}(F_{n,\sigma})$. First, we observe that $\mathbb{E}(Y_{n,k,\sigma})$ satisfies the following recurrence relation.

Lemma 2.5. *Let $\sigma \in \mathcal{L}$, and let $n > k \geq 0$. Then the number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$(7) \quad \mathbb{E}(Y_{n,k}) = 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})).$$

Moreover, if $k+1 > n/2$ we have

$$(8) \quad \mathbb{E}(Y_{n,k}) = 1 + \sum_{i=k+1}^{n-1} \sigma^*(i, n-i) \mathbb{E}(Y_{i,k}),$$

²For instance, the tree in Figure 3 (left) has 7 fringe subtrees of size larger than one, but only 5 different fringe subtrees of size larger than one.

and if $k + 1 \leq n/2$ we have

$$(9) \quad \mathbb{E}(Y_{n,k}) = 1 + \sum_{i=k+1}^{n-k-1} \sigma(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})) + \sum_{i=n-k}^{n-1} \sigma^*(i, n-i) \mathbb{E}(Y_{i,k}).$$

Proof. Let $t \in \mathcal{T}_n$ be a binary tree. Then the number of fringe subtrees of size larger than k occurring in t equals the number of fringe subtrees of size larger than k occurring in its left subtree t_ℓ plus the number of fringe subtrees of size larger than k occurring in the right subtree t_r plus one (for the tree itself, i.e., the fringe subtree rooted in the root node of t). As the left and right subtree of a random tree $T_{n,\sigma}$ conditioned on their sizes i and $n-i$ for some integer $1 \leq i \leq n-1$ are again independent random trees $T_{i,\sigma}$ and $T_{n-i,\sigma}$, and as the probability that the left and right subtree are of sizes i and $n-i$ is given by $\sigma(i, n-i)$, we find

$$\mathbb{E}(Y_{n,k}) = 1 + \sum_{i=1}^{n-1} \sigma(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})).$$

From this and the definition of σ^* (see (4)) we obtain (7). With $\mathbb{E}(Y_{i,k}) = 0$ for $i \leq k$ we can write (7) as

$$\mathbb{E}(Y_{n,k}) = 1 + \sum_{k+1 \leq i \leq n/2} \sigma^*(i, n-i) \mathbb{E}(Y_{i,k}) + \sum_{1 \leq i \leq \min\{n-(k+1), n/2\}} \sigma^*(i, n-i) \mathbb{E}(Y_{n-i,k}).$$

By considering the case $k+1 > n/2$ and $k+1 \leq n/2$, we finally obtain (8) and (9). \square

Several times we have to bound expressions of the form $\sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \cdot \min\{\alpha i, \beta\}$ for constants $\alpha, \beta > 0$. It turns out to be convenient to use the (cumulative) distribution function $D_{\sigma,n} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ corresponding to σ^* , which is defined as follows:

$$(10) \quad D_{\sigma,n}(x) = \sum_{i \leq x} \sigma^*(i, n-i).$$

The sum goes over all integers i with $1 \leq i \leq \min\{n/2, x\}$. If σ is clear from the context, we also write D_n for $D_{\sigma,n}$. In the following lemma we use a Riemann–Stieltjes integral; see e.g. [24, Chapter 6].

Lemma 2.6. *Let $\sigma \in \mathcal{L}$ and $\alpha, \beta > 0$. Then we have*

$$\sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \cdot \min\{\alpha i, \beta\} = \beta - \int_0^{\beta/\alpha} \alpha D_n(x) dx.$$

Proof. Using integration by parts for Riemann–Stieltjes integrals (see e.g. [24, p. 141]), we obtain

$$\begin{aligned} \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \cdot \min\{\alpha i, \beta\} &= \int_0^\infty \min\{\alpha x, \beta\} dD_n(x) \\ &= \int_0^{\beta/\alpha} \alpha x dD_n(x) + \int_{\beta/\alpha}^\infty \beta dD_n(x) \\ &= \left[\alpha x D_n(x) \right]_0^{\beta/\alpha} - \int_0^{\beta/\alpha} \alpha D_n(x) dx + \left[\beta D_n(x) \right]_{\beta/\alpha}^\infty \\ &= \beta - \int_0^{\beta/\alpha} \alpha D_n(x) dx. \end{aligned}$$

This proves the lemma. \square

3. UPPER BOUNDS

In this section of the paper, we focus on upper bounds on the expected number $\mathbb{E}(F_{n,\sigma})$ of distinct fringe subtrees occurring in a random tree $T_{n,\sigma}$. In particular, we present three classes of leaf-centric tree sources, for which we will be able to derive asymptotic upper bounds on $\mathbb{E}(F_{n,\sigma})$.

3.1. Upper-bounded sources. The first natural class of leaf-centric tree sources we consider is the following class, where the mapping σ (resp. σ^*) is bounded from above by a function ψ :

Definition 3.1 (ψ -upper-bounded sources). For a decreasing function $\psi: \mathbb{R} \rightarrow (0, 1]$, let $\mathcal{L}_{up}(\psi) \subseteq \mathcal{L}$ denote the set of mappings $\sigma \in \mathcal{L}$ for which there is an integer N_σ such that

$$\sigma(i, n-i) \leq \psi(n)$$

for all $n \geq N_\sigma$ and all integers $1 \leq i \leq n-1$. In the same way, let $\mathcal{L}_{up}^*(\psi) \subseteq \mathcal{L}$ denote the set of mappings $\sigma \in \mathcal{L}$ for which there is an integer N_σ such that

$$\sigma^*(i, n-i) \leq \psi(n)$$

for all $n \geq N_\sigma$ and all integers $1 \leq i \leq n-1$.

Note that by definition of σ^* (see (4)), we have $\mathcal{L}_{up}^*(\psi) \subseteq \mathcal{L}_{up}(\psi) \subseteq \mathcal{L}_{up}^*(2\psi)$. Furthermore, note that without loss of generality, we can assume that $\psi(n) \geq 1/(n-1)$ for every integer $n \geq 2$, as the values $\sigma(i, n-i)$ have to add up to 1 for $1 \leq i \leq n-1$ by condition (2). In the same way, if we consider the class $\mathcal{L}_{up}^*(\psi)$, we can assume that $\psi(n) \geq 2/n$ for every integer $n \geq 2$. The class of ψ -upper-bounded functions can be generalized as follows:

Definition 3.2 (ψ -weakly-upper-bounded sources). For a decreasing function $\psi: \mathbb{R} \rightarrow (0, 1]$, let $\mathcal{L}_{wup}(\psi) \subseteq \mathcal{L}$ denote the set of mappings $\sigma \in \mathcal{L}$ for which there is an integer N_σ such that

$$\sum_{1 \leq i \leq k} \sigma^*(i, n-i) \leq k\psi(n)$$

for all $n \geq N_\sigma$ and all integers $1 \leq k \leq n/2$.

Note that we have $\mathcal{L}_{up}^*(\psi) \subseteq \mathcal{L}_{wup}(\psi)$ and hence also $\mathcal{L}_{up}(\psi) \subseteq \mathcal{L}_{wup}(2\psi)$. For ψ -weakly-upper-bounded sources we can assume again that $\psi(n) \geq 2/n$ for every integer $n \geq 2$, as the values $\sigma^*(i, n-i)$ have to add up to 1 for $1 \leq i \leq n/2$ by condition (2). As our first main theorem, we obtain the following upper bound on $\mathbb{E}(F_{n,\sigma})$ for mappings $\sigma \in \mathcal{L}_{wup}(\psi)$:

Theorem 3.3. *Let $\psi: \mathbb{R} \rightarrow (0, 1]$ be a decreasing function, and let $\sigma \in \mathcal{L}_{wup}(\psi)$. The number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(F_{n,\sigma}) \leq 2n\psi(\log_4 n) + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right).$$

As $\mathcal{L}_{up}(\psi) \subseteq \mathcal{L}_{up}^*(2\psi)$ and $\mathcal{L}_{up}^*(\psi) \subseteq \mathcal{L}_{wup}(\psi)$, we immediately obtain an asymptotic upper bound on $\mathbb{E}(F_{n,\sigma})$ for $\sigma \in \mathcal{L}_{up}(\psi)$ and $\sigma \in \mathcal{L}_{up}^*(\psi)$ from Theorem 3.3 as well.

In order to prove Theorem 3.3, we make use of the cut-point technique as described in Section 2 (see Lemma 2.4). For this, we start with an upper bound on the expectation $\mathbb{E}(Y_{n,k,\sigma})$ of fringe subtrees of size larger than k in a random tree $T_{n,\sigma}$, where $\sigma \in \mathcal{L}_{wup}(\psi)$ for some mapping ψ .

Lemma 3.4. *Let $\psi : \mathbb{R} \rightarrow (0, 1]$ be a decreasing function, and let $\sigma \in \mathcal{L}_{wup}(\psi)$ with integer N_σ . Then the (random) number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(Y_{n,k}) \leq 2n\psi(k) - 2$$

for all n, k with $n > k \geq N_\sigma$.

Proof. As $\sigma \in \mathcal{L}_{wup}(\psi)$, we assume that $1 \geq \psi(n) \geq 2/n$ for every integer $n \geq N_\sigma$. For the distribution function $D_n = D_{\sigma,n}$ in (10), we have

$$(11) \quad D_n(x) \leq x\psi(n)$$

for all $x \geq 0$. We prove the statement by induction on $n > k \geq N_\sigma$. For the base case, let $n = k + 1$. As there is exactly one fringe subtree of size larger than k in a binary tree of size $k + 1$ (the fringe subtree rooted at the root node), we have

$$\mathbb{E}(Y_{k+1,k}) = 1 \leq 2(k+1)\psi(k) - 2,$$

as $\psi(k) \geq 2/k$ by assumption. For the induction step, take an integer $n > k + 1 > N_\sigma$, so that $\mathbb{E}(Y_{i,k}) \leq 2i\psi(k) - 2$ for every $k < i \leq n - 1$. Lemma 2.5 yields

$$\mathbb{E}(Y_{n,k}) = 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})).$$

Note that $\mathbb{E}(Y_{i,k}) \leq \max\{2i\psi(k) - 2, 0\}$ for all $1 \leq i \leq n/2$: if $i > k$, $\mathbb{E}(Y_{i,k}) \leq 2i\psi(k) - 2$ holds by the induction hypothesis, and otherwise, if $i \leq k$, we trivially have $\mathbb{E}(Y_{i,k}) = 0$. Furthermore, we have $\mathbb{E}(Y_{n-i,k}) \leq 2(n-i)\psi(k) - 2$ for all $1 \leq i \leq n/2$: if $n-i > k$, this holds by the induction hypothesis, and if $n-i \leq k$, we find $\mathbb{E}(Y_{n-i,k}) = 0 < \frac{2n}{k} - 2 \leq n\psi(k) - 2 \leq 2(n-i)\psi(k) - 2$. We can combine these to

$$\begin{aligned} \mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k}) &\leq \max\{2i\psi(k) - 2, 0\} + 2(n-i)\psi(k) - 2 \\ &= 2n\psi(k) - 2 - \min\{2i\psi(k), 2\}. \end{aligned}$$

Thus, we obtain

$$\begin{aligned} \mathbb{E}(Y_{n,k}) &\leq 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) (2n\psi(k) - 2 - \min\{2i\psi(k), 2\}) \\ &= 2n\psi(k) - 1 - \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \min\{2i\psi(k), 2\} \\ &\stackrel{\text{Lem. 2.6}}{=} 2n\psi(k) - 3 + \int_0^{1/\psi(k)} 2\psi(k)D_n(x) \, dx \\ &\stackrel{(11)}{\leq} 2n\psi(k) - 3 + \int_0^{1/\psi(k)} 2\psi(k)\psi(n)x \, dx \\ &= 2n\psi(k) - 3 + \left[\psi(k)\psi(n)x^2 \right]_0^{1/\psi(k)} \\ &= 2n\psi(k) - 3 + \frac{\psi(n)}{\psi(k)} \\ &\leq 2n\psi(k) - 2, \end{aligned}$$

where we use the fact that ψ is decreasing and $n > k$ for the last inequality. \square

We are now able to prove Theorem 3.3:

Proof of Theorem 3.3. Let $\sigma \in \mathcal{L}_{wup}(\psi)$ with integer N_σ . Let $n > 4^{N_\sigma}$, and set $k := \lceil \log_4 n \rceil > N_\sigma$. We use the cut-point argument from Section 2 (see Lemma 2.4) and find that the expected number $\mathbb{E}(F_{n,\sigma})$ of distinct fringe subtrees in the random

tree $T_{n,\sigma}$ is bounded above by the expected number $\mathbb{E}(Y_{n,k})$ of fringe subtrees of size larger than k in $T_{n,\sigma}$, plus the number of all binary trees of size at most k :

$$\mathbb{E}(F_{n,\sigma}) \leq \mathbb{E}(Y_{n,k}) + \sum_{i=0}^{k-1} C_i.$$

By Lemma 3.4, we have

$$\mathbb{E}(F_{n,\sigma}) \leq 2n\psi(k) + \sum_{i=0}^{k-1} C_i.$$

With the asymptotic growth of the Catalan numbers (1), we obtain

$$\begin{aligned} \mathbb{E}(F_{n,\sigma}) &\leq 2n\psi(k) + \mathcal{O}\left(\frac{4^k}{k^{3/2}}\right) \\ &\leq 2n\psi(\log_4 n) + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right), \end{aligned}$$

as $k = \lceil \log_4 n \rceil$. This yields the asymptotic upper bound from Theorem 3.3. \square

Using Theorem 3.3, we obtain an upper bound on the expected number of distinct fringe subtrees in random binary search trees:

Example 3.5. For the binary search tree model from Example 2.1, we find that $\sigma_{bst} \in \mathcal{L}_{up}^*(\psi_{bst})$, where $\psi_{bst}: \mathbb{R} \rightarrow (0, 1]$ is given by $\psi_{bst}(x) = 2/(x-1)$. Theorem 3.3 gives us the following upper bound on the expected number of distinct fringe subtrees in a random binary search tree:

$$\mathbb{E}(F_{n,\sigma_{bst}}) \leq \frac{8n}{\log n} (1 + o(1)).$$

In [26], it is shown that

$$(12) \quad \frac{c_1 n}{\log n} (1 + o(1)) \leq \mathbb{E}(F_{n,\sigma_{bst}}) \leq \frac{4n}{\log n} (1 + o(1)),$$

where $c_1 \approx 3.48$ (see also [8, 12]). Thus, except for the leading constant, the upper bound from Theorem 3.3 yields the correct asymptotic growth of $\mathbb{E}(F_{n,\sigma_{bst}})$.

For the binomial random tree model, we are also able to obtain a non-trivial upper bound on $\mathbb{E}(F_{n,\sigma_{bin,p}})$ from Theorem 3.3:

Example 3.6. For the binomial random tree model from Example 2.3, a short computation shows that for a fixed integer n , the maximal value of $\sigma_{bin,p}(i, n-i)$ with $1 \leq i \leq n-1$ is attained at $i_{\max}(n) = \lfloor p(n-1) \rfloor + 1$. In particular, we have

$$\sigma_{bin,p}(i_{\max}(n), n - i_{\max}(n)) \in \Theta(1/\sqrt{n}).$$

Thus, we find that $\sigma_{bin,p} \in \mathcal{L}_{up}^*(\psi_{bin,p})$ for a function $\psi_{bin,p}$ with $\psi_{bin,p}(x) \in \Theta(1/\sqrt{x})$. By Theorem 3.3, we obtain the bound $\mathbb{E}(F_{n,\sigma_{bin,p}}) \in \mathcal{O}(n/\sqrt{\log n})$. However, in the following sections, we will be able to prove the stronger upper bound $\mathbb{E}(F_{n,\sigma_{bin,p}}) \in \mathcal{O}(n/\log n)$ (see Example 3.11) and also a corresponding lower bound $\mathbb{E}(F_{n,\sigma_{bin,p}}) \in \Omega(n/\log n)$ (see Example 4.7).

There are plenty of other ways to choose the mapping ψ in order to obtain non-trivial upper bounds on $\mathbb{E}(F_{n,\sigma})$ for $\sigma \in \mathcal{L}_{up}^*(\psi)$: For example, $\psi(x) \in \Theta(x^{-\alpha})$ for a constant $0 < \alpha \leq 1$ yields $\mathbb{E}(F_{n,\sigma}) \in \mathcal{O}(n/(\log n)^\alpha)$, and $\psi(x) \in \Theta(1/\log x)$ yields $\mathbb{E}(F_{n,\sigma}) \in \mathcal{O}(n/\log \log n)$. Note that Theorem 3.3 only makes a non-trivial statement if $\psi(n) < 1/2$ for $n \geq N_\sigma$. For the uniform probability distribution (see Example 2.2), Theorem 3.3 only yields an upper bound of the form $\mathbb{E}(F_{n,\sigma_{uni}}) \in \mathcal{O}(n)$, as we have $\sigma_{uni}(1, n-1) > 1/4$ for every $n \geq 2$.

3.2. Weakly-balanced sources. In this subsection, we investigate another class of leaf-centric binary tree sources. We consider *weakly-balanced* binary tree sources, which represent a generalization of balanced binary tree sources introduced in [28] and further analyzed in [16].

Definition 3.7 (ϕ -weakly-balanced sources). For a decreasing function $\phi : \mathbb{N} \rightarrow (0, 1]$ and $\gamma \in (0, \frac{1}{2})$, let $\mathcal{L}_{wbal}(\phi, \gamma) \subseteq \mathcal{L}$ denote the set of mappings $\sigma \in \mathcal{L}$ for which there is an integer N_σ such that

$$\sum_{\gamma n \leq i \leq (1-\gamma)n} \sigma(i, n-i) \geq \phi(n)$$

for all $n \geq N_\sigma$.

For the class of weakly-balanced leaf-centric tree sources, we obtain the following main result:

Theorem 3.8. *Let $\phi : \mathbb{N} \rightarrow (0, 1]$ be a decreasing function, let $\gamma \in (0, \frac{1}{2})$, and let $\sigma \in \mathcal{L}_{wbal}(\phi, \gamma)$. The number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(F_{n,\sigma}) \leq \frac{n}{\gamma\phi(n)\log_4 n} + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right).$$

In order to get a nontrivial statement from Theorem 3.8, we should have $\phi(n) \in \Omega(1/\log n)$. Again, in order to prove Theorem 3.8, we make use of the cut-point argument from Section 2 (see Lemma 2.4). We start with the following lemma:

Lemma 3.9. *Let $\phi : \mathbb{N} \rightarrow (0, 1]$ be a decreasing function, let $\gamma \in (0, \frac{1}{2})$, and let $\sigma \in \mathcal{L}_{wbal}(\phi, \gamma)$ with integer N_σ . Then for $k \geq N_\sigma$ and $n \geq k+1$, the (random) number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(Y_{n,k}) \leq \frac{n}{\gamma\phi(n)k} - \frac{1}{\phi(n)}.$$

Proof. We prove the statement by induction on $n \geq k+1$. For the base case, let $n = k+1$. As there is exactly one fringe subtree of size larger than k in a binary tree of size $k+1$ (the fringe subtree rooted at the root node), we have

$$\mathbb{E}(Y_{k+1,k}) = 1 \leq \frac{(\frac{1}{\gamma} - 1)(k+1)}{\phi(k+1)k} \leq \frac{k+1}{\gamma\phi(k+1)k} - \frac{1}{\phi(k+1)},$$

as $\frac{1}{\gamma} > 2$ and $\phi(k+1) \leq 1$ by assumption. For the induction step, take an integer $n > k+1$, so that

$$(13) \quad \mathbb{E}(Y_{i,k}) \leq \frac{i}{\gamma\phi(i)k} - \frac{1}{\phi(i)}$$

for every integer $k+1 \leq i \leq n-1$. As $\gamma k < \gamma n$ and $\sigma \in \mathcal{L}_{wbal}(\phi, \gamma)$, we have

$$(14) \quad D_n(x) \leq 1 - \phi(n)$$

for every $0 \leq x \leq \gamma k$. Lemma 2.5 yields

$$\mathbb{E}(Y_{n,k}) = 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})).$$

First, we observe that for all $1 \leq i \leq n/2$, we have

$$\mathbb{E}(Y_{i,k}) \leq \max\left\{\frac{i}{\gamma\phi(n)k} - \frac{1}{\phi(n)}, 0\right\}.$$

To see this, note that (i) $\mathbb{E}(Y_{i,k}) = 0$ for $i \leq k$, and (ii) (13) holds for $i > k$ and ϕ is decreasing. Furthermore, for all $1 \leq i \leq n/2$ we have

$$\mathbb{E}(Y_{n-i,k}) \leq \frac{n-i}{\gamma\phi(n)k} - \frac{1}{\phi(n)}.$$

If $n-i > k$, this follows from (13) and the fact that ϕ is decreasing. On the other hand, if $n-i \leq k$, then we have

$$\mathbb{E}(Y_{n-i,k}) = 0 < \frac{n-2\gamma k}{2\gamma\phi(n)k} = \frac{n}{2\gamma\phi(n)k} - \frac{1}{\phi(n)} \leq \frac{n-i}{\gamma\phi(n)k} - \frac{1}{\phi(n)},$$

as $\gamma < 1/2$ by assumption. Thus, we obtain

$$\begin{aligned} \mathbb{E}(Y_{n,k}) &\leq 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \left(\max \left\{ \frac{i}{\gamma\phi(n)k} - \frac{1}{\phi(n)}, 0 \right\} + \frac{n-i}{\gamma\phi(n)k} - \frac{1}{\phi(n)} \right) \\ &= 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \left(\frac{n}{\gamma\phi(n)k} - \frac{1}{\phi(n)} - \min \left\{ \frac{i}{\gamma\phi(n)k}, \frac{1}{\phi(n)} \right\} \right) \\ &= 1 + \frac{n}{\gamma\phi(n)k} - \frac{1}{\phi(n)} - \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \min \left\{ \frac{i}{\gamma\phi(n)k}, \frac{1}{\phi(n)} \right\} \\ &\stackrel{\text{Lem. 2.6}}{=} 1 + \frac{n}{\gamma\phi(n)k} - \frac{2}{\phi(n)} + \int_0^{\gamma k} \frac{D_n(x)}{\gamma\phi(n)k} dx \\ &\stackrel{(14)}{\leq} 1 + \frac{n}{\gamma\phi(n)k} - \frac{2}{\phi(n)} + \int_0^{\gamma k} \frac{1-\phi(n)}{\gamma\phi(n)k} dx \\ &= \frac{n}{\gamma\phi(n)k} - \frac{1}{\phi(n)}, \end{aligned}$$

which concludes the proof. \square

With Lemma 3.9, we are now able to prove Theorem 3.8.

Proof of Theorem 3.8. Let $\sigma \in \mathcal{L}_{wbal}(\psi)$ with integer N_σ . Let $n > 4^{N_\sigma}$, and set $k := \lceil \log_4 n \rceil > N_\sigma$. Again, we use the cut-point technique from Section 2 (Lemma 2.4). The expected number $\mathbb{E}(F_{n,\sigma})$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ is bounded from above by $\mathbb{E}(Y_{n,k})$ plus the number of all binary trees of size at most k :

$$\mathbb{E}(F_{n,\sigma}) \leq \mathbb{E}(Y_{n,k}) + \sum_{i=0}^{k-1} C_i.$$

With Lemma 3.9, and with the asymptotic growth of the Catalan numbers, we find

$$\begin{aligned} \mathbb{E}(F_{n,\sigma}) &\leq \frac{n}{\gamma\phi(n)k} + \sum_{i=0}^{k-1} C_i \\ &\leq \frac{n}{\gamma\phi(n)\log_4 n} + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right). \end{aligned}$$

This finishes the proof. \square

The application of Theorem 3.8 yields the following upper bound on the expected number of distinct fringe subtrees in random binary search trees:

Example 3.10. For the binary search tree model from Example 2.1, we find that $\sigma_{bst} \in \mathcal{L}_{wbal}(\phi_{bst}, \gamma_{bst})$ with $\phi_{bst}(n) = 1/2$ for every $n \geq 2$ and $\gamma_{bst} = 1/4$. Hence, by Theorem 3.8, we have

$$\mathbb{E}(F_{n,\sigma_{bst}}) \leq \frac{16n}{\log n} + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right).$$

The asymptotic growth coincides with the upper bound on $\mathbb{E}(F_{n,\sigma_{bst}})$ from Example 3.5 obtained by Theorem 3.3, except for the leading constant. Recall the upper and lower bound (12) for the binary search tree model from [26]. So except for the leading constant, the upper bound from Theorem 3.8 yields the correct asymptotic growth of $\mathbb{E}(F_{n,\sigma_{bst}})$ as well. Note that it is possible to choose $\phi_{bst}(n) = 1 - 2\gamma$ for any $\gamma < 1/2$. For $\gamma = 1/4$, we obtain the best leading constant in the upper bound.

Example 3.11. For the binomial random tree model from Example 2.3, we obtain the following result from Theorem 3.8. Let S_p^n be the random variable taking values in the set $\{1, \dots, n-1\}$ according to the probability mass function $\sigma_{bin,p}$ from (5). Then $S_p^n = Z_p^n + 1$, where Z_p^n is a binomially distributed random variable with parameters $n-2$ and p . For the expected value μ of S_p^n , we thus obtain $\mu = \mathbb{E}(S_p^n) = p(n-2) + 1$. Chernoff's bound implies

$$\mathbb{P}(|\mu - S_p^n| < \mu^{3/4}) \geq 1 - 2e^{-\frac{\sqrt{\mu}}{3}}.$$

Moreover, with $i_1 := \mu - \mu^{3/4}$ and $i_2 := \mu + \mu^{3/4}$ we have

$$\mathbb{P}(|\mu - S_p^n| < \mu^{3/4}) = \sum_{i_1 < i < i_2} \sigma_{bin,p}(i, n-i).$$

Next, let $\gamma < \min\{p, 1-p\} \leq 1/2$. There is an integer N_σ (depending on p and γ) such that

$$\gamma n \leq \mu - \mu^{3/4} \leq \mu + \mu^{3/4} \leq (1-\gamma)n$$

for all $n \geq N_\sigma$. All in all, we thus have

$$\sum_{\gamma n \leq i \leq n-\gamma n} \sigma_{bin,p}(i, n-i) \geq \sum_{i_1 < i < i_2} \sigma_{bin,p}(i, n-i) \geq 1 - 2e^{-\frac{\sqrt{p(n-2)+1}}{3}}$$

for $n \geq N_\sigma$. Choose any constant ε with $0 < \varepsilon < 1$, and let $\phi(n) = 1 - \varepsilon$ (i.e., ϕ is a constant function). Then, if n is large enough, we have

$$\sum_{\gamma n \leq i \leq n-\gamma n} \sigma_{bin,p}(i, n-i) \geq 1 - \varepsilon,$$

i.e., $\sigma_{bin,p} \in \mathcal{L}_{wbal}(\phi, \gamma)$ for N_σ large enough. From Theorem 3.8, we obtain

$$\mathbb{E}(F_{n,\sigma_{bin,p}}) \leq c_p \cdot \frac{n}{\log n} + \mathcal{O}\left(\frac{n}{(\log n)^{3/2}}\right),$$

where we can choose any constant $c_p > 2/p$ (if $p \leq 1/2$) respectively $c_p > 2/(1-p)$ (if $p > 1/2$), as $\varepsilon > 0$ is arbitrary. In particular, Theorem 3.8 yields a more precise upper bound on $\mathbb{E}(F_{n,\sigma_{bin,p}})$ than Theorem 3.3 (see Example 3.6). In Example 4.7, we will show a corresponding lower bound of the form $\mathbb{E}(F_{n,\sigma_{bin,p}}) \in \Omega(n/\log n)$.

It remains to remark that for the uniform probability distribution (from Example 2.2), Theorem 3.8 only yields a trivial upper bound on $\mathbb{E}(F_{n,\sigma_{uni}})$. On the one hand, we would need $\phi(n) \in \omega(1/\log n)$ in order to obtain a non-trivial upper bound from Theorem 3.8, but on the other hand, for every constant $0 < \gamma < 1/2$, we find that if $\sigma_{uni} \in \mathcal{L}_{wbal}(\phi, \gamma)$, then $\phi(n) \in \mathcal{O}(1/\sqrt{n})$. In the next section (Section 3.3), we present a class of leaf-centric binary tree sources which contains the uniform distribution σ_{uni} from Example 2.2, and for which we will be able to derive a non-trivial asymptotic upper bound on $\mathbb{E}(F_{n,\sigma})$.

3.3. Strongly-balanced sources. In this section, we focus on another class of leaf-centric binary tree sources, which represents a refinement of the class of weakly-balanced leaf-centric tree sources from Section 3.2:

Definition 3.12 (ϑ -strongly-balanced sources). For a decreasing function $\vartheta : \mathbb{R} \rightarrow (0, 1]$ and a constant $\gamma \in (0, \frac{1}{2})$, let $\mathcal{L}_{sbal}(\vartheta, \gamma) \subseteq \mathcal{L}$ denote the set of mappings σ for which there is an integer N_σ and a constant $c_\sigma \geq 1$ such that for every $n \geq N_\sigma$ and every integer r with $c_\sigma \leq r \leq \lceil \gamma n \rceil$, the following inequality holds:

$$\sum_{r \leq i \leq n-r} \sigma(i, n-i) \geq \vartheta(r).$$

Let $\sigma \in \mathcal{L}_{sbal}(\vartheta, \gamma)$, and define $\phi : \mathbb{N} \rightarrow (0, 1]$ by $\phi(n) = \vartheta(\lceil \gamma n \rceil)$. Then we have $\sigma \in \mathcal{L}_{wbal}(\phi, \gamma)$, i.e., every strongly-balanced leaf-centric tree source is also weakly-balanced. In particular, Theorem 3.8 thus holds for strongly balanced tree sources as well. However, we are able to prove a stronger asymptotic upper bound on the expected number $\mathbb{E}(F_{n,\sigma})$ of distinct fringe subtrees in a random tree $T_{n,\sigma}$ if σ corresponds to a strongly-balanced tree source:

Theorem 3.13. *Let $\vartheta : \mathbb{R} \rightarrow (0, 1]$ be a decreasing function, let $\gamma \in (0, \frac{1}{2})$, and let $\sigma \in \mathcal{L}_{sbal}(\vartheta, \gamma)$. The number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(F_{n,\sigma}) \leq \frac{n}{\gamma \vartheta(\gamma \log_4 n) \log_4 n} (1 + o(1)).$$

In order to prove Theorem 3.13, we start again with an upper bound on the expected number $\mathbb{E}(Y_{n,k})$ of fringe subtrees of size larger than k in a random tree $T_{n,\sigma}$, where $\sigma \in \mathcal{L}_{sbal}(\vartheta, \gamma)$. The following lemma and its proof are in fact quite similar to Lemma 3.9 and its proof:

Lemma 3.14. *Let $\vartheta : \mathbb{R} \rightarrow (0, 1]$ be a decreasing function, let $\gamma \in (0, \frac{1}{2})$, and let $\sigma \in \mathcal{L}_{sbal}(\vartheta, \gamma)$ with integer N_σ and constant c_σ . Then for $k \geq \max\{N_\sigma, c_\sigma/\gamma\}$ and $n \geq k+1$, the (random) number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(Y_{n,k}) \leq \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}.$$

Proof. We prove the statement by induction on $n \geq k+1$. For the base case, let $n = k+1$. A binary tree t of size $k+1$ has exactly one fringe subtree of size larger than k , thus

$$\mathbb{E}(Y_{k+1,k}) = 1 \leq \frac{(\frac{1}{\gamma} - 1)(k+1)}{\vartheta(\lceil \gamma(k+1) \rceil) k} \leq \frac{k+1}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)},$$

as $\frac{1}{\gamma} > 2$ and $\vartheta(\lceil \gamma(k+1) \rceil) \leq 1$ by assumption. Let us now deal with the induction step. Take an integer $n > k+1$, so that

$$(15) \quad \mathbb{E}(Y_{i,k}) \leq \frac{i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}$$

for every integer $k+1 \leq i \leq n-1$ by the induction hypothesis. As $c_\sigma \leq \gamma k < \lceil \gamma(k+1) \rceil \leq \lceil \gamma n \rceil$ and $\sigma \in \mathcal{L}_{sbal}(\vartheta, \gamma)$, we have

$$(16) \quad D_n(x) \leq 1 - \vartheta(\lceil \gamma(k+1) \rceil)$$

for every $0 \leq x \leq \gamma k$ (take $r = \lceil \gamma(k+1) \rceil$ in Definition 3.12). By Lemma 2.5, we have

$$\mathbb{E}(Y_{n,k}) = 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})).$$

First, note that for all $1 \leq i \leq n/2$ we have

$$\mathbb{E}(Y_{i,k}) \leq \max \left\{ \frac{i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}, 0 \right\}.$$

This follows from (15) for $i > k$, and from the fact that $\mathbb{E}(Y_{i,k}) = 0$ for $i \leq k$. Furthermore, for all $1 \leq i \leq n/2$ we have

$$\mathbb{E}(Y_{n-i,k}) \leq \frac{n-i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}.$$

This follows again from (15) if $n-i > k$. For $n-i \leq k$ note that

$$\begin{aligned} \mathbb{E}(Y_{n-i,k}) &= 0 < \frac{n-2\gamma k}{2\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} = \frac{n}{2\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)} \\ &\leq \frac{n-i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}, \end{aligned}$$

since $i \leq n/2$ and $\gamma < 1/2$ by assumption. Thus, we obtain

$$\begin{aligned} \mathbb{E}(Y_{n,k}) &\leq 1 + \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \left(\max \left\{ \frac{i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}, 0 \right\} \right. \\ &\quad \left. + \frac{n-i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)} \right) \\ &= 1 + \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)} \\ &\quad - \sum_{1 \leq i \leq n/2} \sigma^*(i, n-i) \min \left\{ \frac{i}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k}, \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)} \right\} \\ &\stackrel{\text{Lem. 2.6}}{=} 1 + \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{2}{\vartheta(\lceil \gamma(k+1) \rceil)} + \int_0^{\gamma k} \frac{D_n(x)}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} dx \\ &\stackrel{(16)}{\leq} 1 + \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{2}{\vartheta(\lceil \gamma(k+1) \rceil)} + \int_0^{\gamma k} \frac{1 - \vartheta(\lceil \gamma(k+1) \rceil)}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} dx \\ &= \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}, \end{aligned}$$

which concludes the proof. \square

We are now able to prove Theorem 3.13.

Proof of Theorem 3.13. Let $m = \max\{N_\sigma, c_\sigma/\gamma\}$. By Lemma 3.14, the expected number $\mathbb{E}(Y_{n,k})$ of fringe subtrees of size larger than k in $T_{n,\sigma}$ satisfies

$$\mathbb{E}(Y_{n,k}) \leq \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} - \frac{1}{\vartheta(\lceil \gamma(k+1) \rceil)}$$

for all $n > k \geq m$. Set $k := \lfloor \log_4 n - 1/\gamma \rfloor - 1$ (this implies $\vartheta(\lceil \gamma(k+1) \rceil) \geq \vartheta(\gamma \log_4 n)$), and let n be sufficiently large, so that $k \geq m$. By the cut-point argument from Section 2 (Lemma 2.4), we find that

$$\mathbb{E}(F_{n,\sigma}) \leq \mathbb{E}(Y_{n,k}) + \sum_{i=0}^{k-1} C_i.$$

With the asymptotic growth of the Catalan numbers (1), and with Lemma 3.14, we find

$$\begin{aligned} \mathbb{E}(F_{n,\sigma}) &\leq \frac{n}{\gamma \vartheta(\lceil \gamma(k+1) \rceil) k} + \mathcal{O}\left(\frac{4^k}{k^{3/2}}\right) \\ &\leq \frac{n}{\gamma \vartheta(\gamma \log_4 n) \log_4 n} (1 + o(1)). \end{aligned}$$

This finishes the proof. \square

With Theorem 3.13, we obtain a non-trivial asymptotic upper bound on the expected number of distinct fringe subtrees in a uniformly random binary tree of size n :

Example 3.15. Consider the uniform distribution corresponding to σ_{uni} from Example 2.2. Let $0 < \gamma < 1/2$, and let $\varepsilon > 0$. From the asymptotic bound of the Catalan numbers (1), we find that there is an integer N_ε (depending on ε) such that

$$\frac{(1-\varepsilon)4^n}{\sqrt{\pi}(n+1)^{3/2}} \leq C_n \leq \frac{(1+\varepsilon)4^n}{\sqrt{\pi}(n+1)^{3/2}}$$

for all $n \geq N_\varepsilon$. Let $N_\varepsilon \leq r \leq \lceil \gamma n \rceil$, then

$$\begin{aligned} \sum_{r \leq i \leq n-r} \sigma_{uni}(i, n-i) &= \sum_{r \leq i \leq n-r} \frac{C_{i-1}C_{n-i-1}}{C_{n-1}} \\ &\geq \frac{(1-\varepsilon)^2}{4(1+\varepsilon)\sqrt{\pi}} \sum_{r \leq i \leq n-r} \frac{n^{3/2}}{i^{3/2}(n-i)^{3/2}} \\ &\geq \frac{(1-\varepsilon)^2 n^{3/2}}{2(1+\varepsilon)\sqrt{\pi}} \sum_{r \leq i < \frac{n}{2}} i^{-3/2}(n-i)^{-3/2} \\ &\geq \frac{(1-\varepsilon)^2 n^{3/2}}{2(1+\varepsilon)\sqrt{\pi}} \int_r^{\frac{n}{2}} x^{-3/2}(n-x)^{-3/2} dx \\ &= \frac{(1-\varepsilon)^2}{(1+\varepsilon)\sqrt{\pi}} \frac{n-2r}{\sqrt{n(n-r)r}}. \end{aligned}$$

If N_ε is chosen large enough, then we also have

$$\frac{n-2r}{\sqrt{n(n-r)}} \leq 1-2\gamma$$

for all r, n with $n \geq N_\varepsilon$ and $r \leq \lceil n\gamma \rceil$. So set $\delta := \frac{(1+\varepsilon)}{(1-\varepsilon)^2}$, and define

$$\vartheta_{uni}(r) := \frac{(1-2\gamma)}{\delta\sqrt{\pi r}}.$$

Then $\sigma_{uni} \in \mathcal{L}_{sbal}(\vartheta_{uni}, \gamma)$. By Theorem 3.13, we thus obtain

$$\mathbb{E}(F_{n, \sigma_{uni}}) \leq \frac{\delta\sqrt{2\pi}}{\sqrt{\gamma}(1-2\gamma)} \cdot \frac{n}{\sqrt{\log n}}(1+o(1)).$$

As $0 < \gamma < 1/2$ is arbitrary, we can choose the optimal value $\gamma = 1/6$ to obtain

$$\mathbb{E}(F_{n, \sigma_{uni}}) \leq 3\delta\sqrt{3\pi} \cdot \frac{n}{\sqrt{\log n}}(1+o(1)),$$

where $\delta > 1$, as $\varepsilon > 0$. In [14, 23, 26], it is shown that

$$\mathbb{E}(F_{n, \sigma_{uni}}) = \frac{2\sqrt{2}}{\sqrt{\pi}} \cdot \frac{n}{\sqrt{\log n}}(1+o(1)).$$

Thus, except for the leading constant, the upper bound from Theorem 3.13 yields the correct asymptotic growth of $\mathbb{E}(F_{n, \sigma_{uni}})$.

Example 3.16. For the binary search tree model σ_{bst} from Example 2.1, we also obtain an upper bound on $\mathbb{E}(F_{n, \sigma_{bst}})$ from Theorem 3.13. Let $\gamma_{bst} = 1/4$, let $n \geq 2$, and let $1 \leq r \leq \lceil \gamma n \rceil$. Then

$$\sum_{r \leq i \leq n-r} \sigma_{bst}(i, n-i) = \frac{n-2r+1}{n-1} \geq \frac{1}{2}.$$

Set $\vartheta_{bst}(r) := 1/2$. By Theorem 3.13, we find

$$\mathbb{E}(F_{n,\sigma_{uni}}) \leq \frac{16n}{\log n}(1 + o(1)).$$

This coincides with the asymptotic upper bound on $\mathbb{E}(F_{n,\sigma_{uni}})$ from Example 3.10 as well as Example 3.5 (except for the leading constant).

4. LOWER BOUNDS

So far, we have only considered upper bounds on the expected number of distinct fringe subtrees in random trees $T_{n,\sigma}$. In this section, we focus on lower bounds instead. We present some classes of leaf-centric binary tree sources for which we will be able to show asymptotic lower bounds on the number $\mathbb{E}(F_{n,\sigma})$ of distinct fringe subtrees in the corresponding random trees $T_{n,\sigma}$.

4.1. Upper-bounded sources. First, we consider a specific subclass of the class of upper-bounded leaf-centric tree sources from Definition 3.1. Let $\rho \in (0, 1)$ be a constant. With $\mathcal{L}_{up}(\rho) \subseteq \mathcal{L}$, we denote the set of mappings $\sigma \in \mathcal{L}$ that are eventually bounded above by the constant ρ : that is, a mapping $\sigma \in \mathcal{L}$ belongs to $\mathcal{L}_{up}(\rho)$ if there is an integer N_σ such that

$$\sigma(i, n - i) \leq \rho$$

for all $n \geq N_\sigma$ and all integers $1 \leq i \leq n - 1$. In the following theorem, we obtain a lower bound on $\mathbb{E}(F_{n,\sigma})$ for $\sigma \in \mathcal{L}_{up}(\rho)$.

Theorem 4.1. *Let $\rho \in (0, 1)$, and let $\sigma \in \mathcal{L}_{up}(\rho)$ with integer N_σ . Then the number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(F_{n,\sigma}) \geq \log\left(\frac{1}{\rho}\right) \frac{n}{2(N_\sigma - 1) \log n} (1 + o(1)).$$

In order to prove Theorem 4.1, we make use of an information-theoretic argument. Let $H(T_{n,\sigma})$ denote the Shannon entropy of the random variable $T_{n,\sigma}$, i.e.,

$$H(T_{n,\sigma}) = \sum_{t \in \mathcal{T}_n} P_\sigma(t) \log(1/P_\sigma(t)).$$

We obtain the following general result.

Theorem 4.2. *Let $\sigma \in \mathcal{L}$. Then the number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(F_{n,\sigma}) \geq \frac{H(T_{n,\sigma})}{2 \log n} (1 + o(1)).$$

Proof. Let $F(t)$ denote the number of distinct fringe subtrees occurring in the binary tree $t \in \mathcal{T}$. Recall that $t \in \mathcal{T}_n$ has exactly n leaves and $n - 1$ internal nodes. This implies that $F(t) \leq n$ for $t \in \mathcal{T}_n$. We first show that we need at most $2F(t)(\lceil \log n \rceil + 1)$ many bits in order to encode t . Recall that $F(t)$ equals the size of the minimal DAG of t , i.e., the number of nodes in the directed acyclic graph obtained from t by merging identical fringe subtrees of t . As we can uniquely retrieve t from its minimal DAG (see e.g. [18]), it suffices to show that we can encode the minimal DAG of t with at most $2F(t)(\lceil \log n \rceil + 1)$ many bits. Without loss of generality, assume that the nodes of the minimal DAG of t are enumerated as $1, 2, \dots, F(t)$, where $F(t)$ is the unique leaf node of the minimal DAG of t . For $1 \leq i \leq F(t) - 1$, let ℓ_i (resp. r_i) denote the left (resp. right) child of node k . We encode each number $1, \dots, F(t)$ by a bit string of length exactly $\lceil \log n \rceil + 1$. The minimal DAG of t can be uniquely encoded as the bit string $\ell_1 r_1 \cdots \ell_{F(t)-1} r_{F(t)-1}$,

which has length $2(F(t) - 1)(\lceil \log n \rceil + 1)$. Let $\sigma \in \mathcal{L}$. Shannon's coding theorem implies

$$H(T_{n,\sigma}) \leq 2(\lceil \log n \rceil + 1) \sum_{t \in \mathcal{T}_n} P_\sigma(t) F(t) = 2(\lceil \log n \rceil + 1) \mathbb{E}(F_{n,\sigma}).$$

This finishes the proof. \square

In order to show Theorem 4.1, we make use of Theorem 4.2 together with a lower bound on the entropy $H(T_{n,\sigma})$ for $\sigma \in \mathcal{L}_{up}(\rho)$. We will show the following lemma.

Lemma 4.3. *Let $\rho \in (0, 1)$, and let $\sigma \in \mathcal{L}_{up}(\rho)$ with integer $N_\sigma \geq 2$. Then*

$$H(T_{n,\sigma}) \geq \log\left(\frac{1}{\rho}\right) \left(\frac{n}{N_\sigma - 1} - 1\right)$$

for every $n \geq N_\sigma$.

In order to prove Lemma 4.3, we need the following result:

Lemma 4.4. *Let $k \geq 1$. Every binary tree of size n contains at least $\frac{n}{k} - 1$ fringe subtrees of size larger than k .*

Proof. Let t be a binary tree of size n , and let s_1, s_2, \dots, s_r be the maximal fringe subtrees of size at most k (i.e., the fringe subtrees that have at most k leaves and are not contained in any other fringe subtree with that property). Every leaf of t is contained in one of these: since a leaf is itself a fringe subtree of size at most k (namely size 1), it must be contained in a maximal fringe subtree with that property. Moreover, s_1, s_2, \dots, s_r must be disjoint, since for any two fringe subtrees of the same tree, either one is contained in the other, or they are disjoint. Hence we have $n = |t| = |s_1| + |s_2| + \dots + |s_r| \leq kr$, which implies $r \geq \frac{n}{k}$. Lastly, note that the roots of s_1, s_2, \dots, s_r and all internal nodes of t that are not contained in these r fringe subtrees form a binary tree of size r . This tree has $r - 1$ internal nodes, and all of them are roots of fringe subtrees of t whose size is at least $k + 1$. \square

In particular, we obtain the following corollary from Lemma 4.4:

Corollary 4.5. *Let $\sigma \in \mathcal{L}$, and let $n > k$. The random number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(Y_{n,k}) \geq \frac{n}{k} - 1.$$

We are now able to prove Lemma 4.3:

Proof of Lemma 4.3. Lemma 4.3 follows from identity (4) in [17]: Define

$$h_k(\sigma) = \sum_{i=1}^{k-1} \sigma(i, k-i) \log\left(\frac{1}{\sigma(i, k-i)}\right).$$

Thus, $h_k(\sigma)$ is the Shannon entropy of the random variable corresponding to the probability mass function $\sigma : \{(i, k-i) \mid 1 \leq i \leq k-1\} \rightarrow [0, 1]$. As $\sigma(i, k-i) \leq \rho$ for $k \geq N_\sigma$, we find

$$h_k(\sigma) \geq \log(1/\rho) \sum_{i=1}^{k-1} \sigma(i, k-i) = \log(1/\rho)$$

for every $k \geq N_\sigma$. Identity (4) in [17] states that

$$H(T_{n,\sigma}) = \sum_{k=2}^n (\mathbb{E}(Y_{n,k-1}) - \mathbb{E}(Y_{n,k})) h_k(\sigma),$$

where $Y_{n,k}$ again denotes the random number of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$. For every $n \geq N_\sigma \geq 2$, we obtain

$$\begin{aligned} H(T_{n,\sigma}) &\geq \sum_{k=N_\sigma}^n (\mathbb{E}(Y_{n,k-1}) - \mathbb{E}(Y_{n,k})) h_k(\sigma) \\ &\geq \log\left(\frac{1}{\rho}\right) \sum_{k=N_\sigma}^n (\mathbb{E}(Y_{n,k-1}) - \mathbb{E}(Y_{n,k})) \\ &= \log\left(\frac{1}{\rho}\right) (\mathbb{E}(Y_{n,N_\sigma-1}) - \mathbb{E}(Y_{n,n})) \\ &= \log\left(\frac{1}{\rho}\right) \mathbb{E}(Y_{n,N_\sigma-1}). \end{aligned}$$

By Corollary 4.5, this implies

$$H(T_{n,\sigma}) \geq \log\left(\frac{1}{\rho}\right) \left(\frac{n}{N_\sigma-1} - 1\right).$$

This proves the statement. \square

Theorem 4.1 now follows immediately from Theorem 4.2 and Lemma 4.3.

Example 4.6. Let us apply Theorem 4.2 to the binary search tree model from Example 2.1. In [17], it is shown that

$$H(T_{n,\sigma_{bst}}) \sim 1.7363771368 n (1 + o(1)).$$

Hence, we obtain from Theorem 4.2 that

$$\mathbb{E}(F_{n,\sigma_{bst}}) \geq \frac{0.8681 n}{\log n} (1 + o(1)).$$

In Example 3.5, we have shown the upper bound

$$\mathbb{E}(F_{n,\sigma_{bst}}) \leq \frac{8n}{\log n} (1 + o(1))$$

(see also Examples 3.10 and 3.16). Altogether, we find that $\mathbb{E}(F_{n,\sigma}) \in \Theta(n/\log n)$ (as shown in [8, 26]).

Example 4.7. For the binomial random tree model (Example 2.3), a simple induction over n (using $\binom{n-2}{i-1} = \binom{n-3}{i-1} + \binom{n-3}{i-2}$ for $2 \leq i \leq n-2$ in the induction step) shows that for $n \geq 3$, we have

$$\sigma_{bin,p}(i, n-i) \leq \max\{p, 1-p\}$$

for all $1 \leq i \leq n-1$. With $N_\sigma = 3$ and $\rho = \max\{p, 1-p\}$, we thus obtain from Theorem 4.1 that

$$(17) \quad \mathbb{E}(F_{n,\sigma_{bin,p}}) \geq \log\left(\frac{1}{\rho}\right) \frac{n}{4 \log n} (1 + o(1))$$

if $p > 1/2$ and

$$(18) \quad \mathbb{E}(F_{n,\sigma_{bin,p}}) \geq \log\left(\frac{1}{1-p}\right) \frac{n}{4 \log n} (1 + o(1))$$

if $p \leq 1/2$. If we consider the proof of Lemma 4.3 again, we can slightly improve the leading constant in the lower bounds (17) and (18). With identity (4) in [17], $h_2(\sigma) = 0$, and Corollary 4.5, we get

$$H(T_{n,\sigma}) = \sum_{k=2}^n (\mathbb{E}(Y_{n,k-1}) - \mathbb{E}(Y_{n,k})) h_k(\sigma)$$

$$\begin{aligned}
 &\geq \min_{3 \leq i \leq n} h_i(\sigma) \cdot \sum_{k=3}^n (\mathbb{E}(Y_{n,k-1}) - \mathbb{E}(Y_{n,k})) \\
 &= \min_{3 \leq i \leq n} h_i(\sigma) \cdot \mathbb{E}(Y_{n,2}) \\
 &\geq \min_{3 \leq i \leq n} h_i(\sigma) \cdot \left(\frac{n}{2} - 1\right).
 \end{aligned}$$

Together with Theorem 4.2, we thus obtain

$$\mathbb{E}(F_{n,\sigma}) \geq \min_{3 \leq i \leq n} h_i(\sigma) \cdot \frac{n}{4 \log n} (1 + o(1)).$$

For the binomial random tree model we have $h_k(\sigma_{bin,p}) \leq h_{k+1}(\sigma_{bin,p})$: if $\text{Ber}(p)$ is a Bernoulli random variable that takes the value 1 (resp., 0) with probability p (resp., $1 - p$), then the binomial random variable $\text{Bin}(k, p)$ is the sum of k independent copies of $\text{Ber}(p)$. Hence, we have $\text{Bin}(k + 1, p) = \text{Bin}(k, p) + \text{Ber}(p)$. Moreover, for independent random variables X and Y the Shannon entropy satisfies $H(X + Y) \geq H(X)$; see e.g. [19]. Hence, we get

$$\min_{3 \leq i \leq n} h_i(\sigma_{bin,p}) = h_3(\sigma_{bin,p}) = p \log \left(\frac{1}{p}\right) + (1 - p) \log \left(\frac{1}{1 - p}\right).$$

This yields

$$\mathbb{E}(F_{n,\sigma_{bin,p}}) \geq \left(p \log \left(\frac{1}{p}\right) + (1 - p) \log \left(\frac{1}{1 - p}\right)\right) \cdot \frac{n}{4 \log n} (1 + o(1)).$$

Together with the upper bound (Example 3.11), we get $\mathbb{E}(F_{n,\sigma_{bin,p}}) \in \Theta(n/\log n)$.

Example 4.8. For the uniform distribution from Example 2.2, we have (see [17])

$$H(T_{n,\sigma_{uni}}) \sim 2n(1 + o(1)).$$

Theorem 4.2 thus yields a lower bound

$$\mathbb{E}(F_{n,\sigma_{uni}}) \geq \frac{n}{\log n} (1 + o(1)).$$

However, in [14, 23], it is shown that

$$\mathbb{E}(F_{n,\sigma_{uni}}) = \frac{2\sqrt{2}}{\sqrt{\pi}} \cdot \frac{n}{\sqrt{\log n}} (1 + o(1)).$$

In the next subsection, we will prove a general result that implies a lower bound of the form $\mathbb{E}(F_{n,\sigma_{uni}}) \in \Omega(n/\sqrt{\log n})$ (see Example 4.14).

4.2. Unbalanced sources. For upper-bounded sources, Theorem 4.1 can only yield lower bounds of the form $\Omega(n/\log n)$ for $\mathbb{E}(F_{n,\sigma})$. In this subsection, we present another class of leaf-centric binary tree sources, for which we will be able to derive stronger lower bounds on $\mathbb{E}(F_{n,\sigma})$, such as the lower bound $\Omega(n/\sqrt{\log n})$ for the uniform distribution. We start with the following definition:

Definition 4.9 (ξ -unbalanced sources). Let $\xi: \mathbb{R} \rightarrow (0, 1]$ be a decreasing function, and let $\gamma \in (0, \frac{1}{2})$. With $\mathcal{L}_{unbal}(\xi, \gamma) \subseteq \mathcal{L}$, we denote the set of mappings σ for which there are integers N'_σ and c_σ such that

$$\sum_{r \leq i \leq n-r} \sigma(i, n-i) \leq \xi(r)$$

for every $n \geq N'_\sigma$ and every $c_\sigma \leq r \leq \lceil \gamma n \rceil$.

Moreover, let $\rho \in (0, 1)$ be a constant. As in Section 4.1, let again $\mathcal{L}_{up}(\rho) \subseteq \mathcal{L}$ denote the set of mappings $\sigma \in \mathcal{L}$ that are upper-bounded with respect to the

constant ρ . In other words, a mapping $\sigma \in \mathcal{L}$ is contained in $\mathcal{L}_{up}(\rho)$ if there is an integer N_σ such that

$$\sigma(i, n-i) \leq \rho$$

for all $n \geq N_\sigma$ and $1 \leq i \leq n-1$.

Our main result in this subsection is the following theorem:

Theorem 4.10. *Let ξ be a decreasing function, let $\gamma \in (0, \frac{1}{2})$ and $\rho \in (0, 1)$. Moreover, let $\sigma \in \mathcal{L}_{up}(\rho)$ with integer N_σ , and let $\sigma \in \mathcal{L}_{unbal}(\xi, \gamma)$ with integer N'_σ and constant c_σ . If*

- (a) $\xi(x) \in \omega(1/x)$,
- (b) *there is a constant $c_\xi > 0$ such that $1/c_\xi + \gamma \leq 1$ and*

$$\sum_{i=c_\sigma}^k \xi(i) \leq c_\xi(k+1)\xi(k+1)$$

for every $k \geq c_\sigma$, and

- (c) *there is an integer $\ell_\sigma \geq 2$ such that the number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(Y_{n,(\log n)^{\ell_\sigma}}) \in o(n/\log n),$$

then the number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies

$$\mathbb{E}(F_{n,\sigma}) \geq \frac{\gamma n \log(1/\rho)}{c_\xi N_\sigma \xi\left(\frac{N_\sigma \log n}{\log(1/\rho)}\right) \log n} (1 + o(1)).$$

Note that every leaf-centric binary tree source is unbalanced with respect to the constant function ξ with $\xi(x) = 1$ for every $x \in \mathbb{R}$. This yields the following corollary of Theorem 4.10, which is a weaker version of Theorem 4.1:

Corollary 4.11. *Let $\rho \in (0, 1)$, and let $\sigma \in \mathcal{L}_{up}(\rho)$ with integer N_σ . If there is an integer $\ell_\sigma \geq 2$ such that the number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies*

$$\mathbb{E}(Y_{n,(\log n)^{\ell_\sigma}}) \in o(n/\log n),$$

then the number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ satisfies

$$\mathbb{E}(F_{n,\sigma}) \geq \frac{\gamma n \log(1/\rho)}{2N_\sigma \log n} (1 + o(1)),$$

for every $\gamma \in (0, \frac{1}{2})$.

Proof. Let $\sigma \in \mathcal{L}$ satisfy the requirements of Corollary 4.11. We show that σ then satisfies all further requirements of Theorem 4.10 as well.

Clearly, σ is ξ_1 -unbalanced for every constant $\gamma \in (0, 1/2)$, where ξ_1 is the constant function defined by $\xi_1(x) = 1$ for every $x \in \mathbb{R}$. Hence, we have $\sigma \in \mathcal{L}_{unbal}(\xi_1, \gamma)$ with $N'_\sigma = 1$ and $c_\sigma = 1$. We now find that condition (a) of Theorem 4.10 is clearly satisfied, as is condition (b), where the constant c_ξ can be chosen for example as $c_\xi = 2$. Finally, condition (c) of Theorem 4.10 is satisfied by assumption. Thus, we obtain the lower bound

$$\mathbb{E}(F_{n,\sigma}) \geq \frac{\gamma n \log(1/\rho)}{2N_\sigma \log n} (1 + o(1))$$

from Theorem 4.10, where $\gamma \in (0, 1/2)$ is arbitrary. \square

Note that Corollary 4.11 resembles Theorem 4.1 in many ways, except that we have the additional requirement (c) from Theorem 4.10 and the additional constant $\gamma \in (0, 1/2)$ in the lower bound. The technique to prove Theorem 4.10 however will be quite different from the technique used for Theorem 4.1.

In order to prove Theorem 4.10, we make use of a refinement of the cut-point technique from Section 2 (Lemma 2.4), as was applied to a similar problem in [23]. More precisely, we refine the cut-point technique with an inclusion-exclusion-principle-like argument in order to obtain a lower bound on the expected number of distinct fringe subtrees. We first need a lower bound on the expected number $\mathbb{E}(Y_{n,k})$ of fringe subtrees of size larger than k in a random tree $T_{n,\sigma}$, where $\sigma \in \mathcal{L}_{\text{unbal}}(\xi, \gamma)$:

Lemma 4.12. *Let $\xi: \mathbb{R} \rightarrow (0, 1]$ be a decreasing function, let $\gamma \in (0, \frac{1}{2})$, and let $\sigma \in \mathcal{L}_{\text{unbal}}(\xi, \gamma)$ with integers N'_σ and c_σ . If*

- (a) $\xi(x) \in \omega(1/x)$, and
- (b) *there is a constant $c_\xi > 0$ such that $1/c_\xi + \gamma \leq 1$ and*

$$\sum_{i=c_\sigma}^k \xi(i) \leq c_\xi(k+1)\xi(k+1)$$

for every $k \geq c_\sigma$,

then there is an integer m such that for $k \geq m$ and $n \geq k+1$, the random number $Y_{n,k}$ of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$ satisfies

$$\mathbb{E}(Y_{n,k}) \geq \frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)}.$$

In order to prove Lemma 4.12, the following lemma will be helpful:

Lemma 4.13 (summation by parts [24, Thm. 3.41]). *Let $m \in \mathbb{N}$ and $a_1, \dots, a_m, b_1, \dots, b_m$ be real numbers. Then*

$$\sum_{i=1}^m a_i b_i = \left(\sum_{i=1}^m a_i \right) b_m + \sum_{i=1}^{m-1} \left(\sum_{\ell=1}^i a_\ell \right) (b_i - b_{i+1}).$$

We are now able to prove Lemma 4.12.

Proof of Lemma 4.12. As $\xi(x) \in \omega(1/x)$ by condition (a), there is an integer N_ξ such that

$$(19) \quad \frac{c_\sigma - 1}{\xi(k+1)} \leq k$$

for all $k \geq N_\xi$. We choose the integer m in such a way that $m \geq \max\{N_\xi, N'_\sigma, c_\sigma\}$ and $\lceil \gamma n \rceil \leq n/2$ for $n > m$. We prove the statement using induction for $n \geq k+1 > m$. For the base case, let $n = k+1$. A binary tree $t \in \mathcal{T}_{k+1}$ has exactly one fringe subtree of size larger than k , and we have

$$\mathbb{E}(Y_{k+1,k}) = 1 > \frac{\gamma}{c_\xi\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)}.$$

For the induction step, take an integer $n > k+1$, so that

$$\mathbb{E}(Y_{i,k}) \geq \frac{\gamma i}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)}$$

for every integer $k+1 \leq i \leq n-1$ by the induction hypothesis. We distinguish two cases.

Case 1: $\lceil \gamma n \rceil < k + 1 \leq n - 1$. We thus have

$$\frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} < \frac{1}{c_\xi\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} = 0.$$

Furthermore, as $n > k$, we find $\mathbb{E}(Y_{n,k}) \geq 1$. Thus, the statement follows in this case.

Case 2: $k + 1 \leq \lceil \gamma n \rceil$. As $\lceil \gamma n \rceil \leq n/2$ for $n > m$, we find by Lemma 2.5:

$$\mathbb{E}(Y_{n,k}) = 1 + \sum_{i=k+1}^{n-k-1} \sigma(i, n-i) (\mathbb{E}(Y_{i,k}) + \mathbb{E}(Y_{n-i,k})) + \sum_{i=n-k}^{n-1} \sigma^*(i, n-i) \mathbb{E}(Y_{i,k}).$$

By the induction hypothesis, we have

$$\begin{aligned} \mathbb{E}(Y_{n,k}) &\geq 1 + \sum_{i=k+1}^{n-k-1} \sigma(i, n-i) \left(\frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{2}{c_\xi\xi(k+1)} \right) \\ &\quad + \sum_{i=n-k}^{n-1} \sigma^*(i, n-i) \left(\frac{\gamma i}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} \right). \end{aligned}$$

We introduce the abbreviation

$$S(i) = \frac{\gamma i}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)}.$$

Clearly, $S(i)$ is increasing in i . Moreover, set

$$\sum_{i=k+1}^{n-k-1} \sigma(i, n-i) =: \alpha_1, \quad \sum_{i=n-k}^{n-c_\sigma} \sigma^*(i, n-i) =: \alpha_2 \quad \text{and} \quad \sum_{i=n-c_\sigma+1}^{n-1} \sigma^*(i, n-i) =: \alpha_3.$$

We obtain

$$\begin{aligned} (20) \quad \mathbb{E}(Y_{n,k}) &\geq 1 + \alpha_1 \left(\frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{2}{c_\xi\xi(k+1)} \right) + \alpha_3 S(n - c_\sigma + 1) \\ &\quad + \sum_{i=n-k}^{n-c_\sigma} \sigma^*(i, n-i) S(i). \end{aligned}$$

As $\sigma \in \mathcal{L}_{unbal}(\xi, \gamma)$ and additionally $N'_\sigma < n$ and $c_\sigma < k + 1 \leq \lceil \gamma n \rceil$, we have $0 \leq \alpha_1 \leq \xi(k+1)$, $1 - \xi(c_\sigma) \leq \alpha_3 \leq 1$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Furthermore, we have

$$\frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{2}{c_\xi\xi(k+1)} \leq S(i)$$

for every integer i with $n-k \leq i \leq n-c_\sigma+1$. Hence, we minimize the right-hand side of (20) subject to the condition $\alpha_1 + \alpha_2 + \alpha_3 = 1$ if we assign the maximal possible weight $\alpha_1 = \xi(k+1)$ to the smallest term $\gamma n / (c_\xi(k+1)\xi(k+1)) - 2 / (c_\xi\xi(k+1))$ and the minimal possible weight $\alpha_3 = 1 - \xi(c_\sigma)$ to the largest term $S(n - c_\sigma + 1)$. It remains to bound the sum

$$\sum_{i=n-k}^{n-c_\sigma} \sigma^*(i, n-i) S(i)$$

from below subject to the condition $\alpha_2 = \xi(c_\sigma) - \xi(k+1)$. First note that $\sigma \in \mathcal{L}_{unbal}(\xi, \gamma)$, $N'_\sigma < n$ and $\alpha_3 = 1 - \xi(c_\sigma)$ imply

$$(21) \quad \sum_{\ell=c_\sigma}^i \sigma^*(\ell, n-\ell) = \sum_{\ell=1}^i \sigma^*(\ell, n-\ell) - \alpha_3 \geq \xi(c_\sigma) - \xi(i+1),$$

for every $c_\sigma \leq i \leq k$ (note that $c_\sigma \leq i + 1 \leq \lceil \gamma n \rceil$; therefore we can take $r = i + 1$ in Definition 4.9). Using summation by parts (Lemma 4.13), we obtain

$$\begin{aligned}
& \sum_{i=n-k}^{n-c_\sigma} \sigma^*(i, n-i) S(i) \\
&= \sum_{i=c_\sigma}^k \sigma^*(i, n-i) S(n-i) \\
&= \left(\sum_{i=c_\sigma}^k \sigma^*(i, n-i) \right) S(n-k) + \sum_{i=c_\sigma}^{k-1} \left(\sum_{\ell=c_\sigma}^i \sigma^*(\ell, n-\ell) \right) (S(n-i) - S(n-i-1)) \\
&\stackrel{(21)}{\geq} (\xi(c_\sigma) - \xi(k+1)) S(n-k) + \sum_{i=c_\sigma}^{k-1} (\xi(c_\sigma) - \xi(i+1)) (S(n-i) - S(n-i-1)) \\
&= -\xi(k+1) S(n-k) + \xi(c_\sigma) S(n-c_\sigma) - \sum_{i=c_\sigma}^{k-1} \xi(i+1) (S(n-i) - S(n-i-1)).
\end{aligned}$$

Note that we have

$$S(n-i) - S(n-i-1) = \frac{\gamma}{c_\xi(k+1)\xi(k+1)}.$$

Altogether, we obtain from (20) that

$$\begin{aligned}
\mathbb{E}(Y_{n,k}) &\geq 1 + \xi(k+1) \left(\frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{2}{c_\xi\xi(k+1)} \right) \\
&\quad + \frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{\gamma(c_\sigma-1)}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} \\
&\quad - \xi(k+1) \left(\frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{\gamma k}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} \right) \\
&\quad - \sum_{i=c_\sigma}^k \frac{\xi(i)\gamma}{c_\xi(k+1)\xi(k+1)} \\
&= 1 - \frac{1}{c_\xi} + \frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{\gamma(c_\sigma-1)}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} \\
&\quad + \frac{\gamma k}{c_\xi(k+1)} - \sum_{i=c_\sigma}^k \frac{\xi(i)\gamma}{c_\xi(k+1)\xi(k+1)}.
\end{aligned}$$

Since we have

$$\sum_{i=c_\sigma}^k \xi(i) \leq c_\xi(k+1)\xi(k+1)$$

by the assumption from condition (b), we get

$$\begin{aligned}
\mathbb{E}(Y_{n,k}) &\geq 1 - \frac{1}{c_\xi} + \frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{\gamma(c_\sigma-1)}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} \\
&\quad + \frac{\gamma k}{c_\xi(k+1)} - \gamma.
\end{aligned}$$

Furthermore, from (19) we obtain

$$-\frac{\gamma(c_\sigma-1)}{c_\xi(k+1)\xi(k+1)} + \frac{\gamma k}{c_\xi(k+1)} \geq 0$$

and hence

$$\mathbb{E}(Y_{n,k}) \geq 1 - \frac{1}{c_\xi} + \frac{\gamma n}{c_\xi(k+1)\xi(k+1)} - \frac{1}{c_\xi\xi(k+1)} - \gamma.$$

Finally, as $1/c_\xi + \gamma \leq 1$ by assumption (by condition (b)), the statement follows. \square

We are now able to prove Theorem 4.10. The main idea of the proof is based on techniques from [23, 26].

Proof of Theorem 4.10. Let δ_n for $n \in \mathbb{N}$ be defined by $\delta_n = 2\ell_\sigma \log \log n / \log n$, where $\ell_\sigma \geq 2$ is the constant from condition (c) in Theorem 4.10. Moreover, let

$$k_1 = \left\lceil N_\sigma \left(1 + \frac{(1 + \delta_n) \log n}{\log(1/\rho)} \right) \right\rceil.$$

Furthermore, assume that n is sufficiently large, so that the lower bound on $\mathbb{E}(Y_{n,k})$ from Lemma 4.12 applies for $k \geq k_1 - 1$, and that $k_1 < (\log n)^{\ell_\sigma}$. The random number $F_{n,\sigma}$ of distinct fringe subtrees in the random tree $T_{n,\sigma}$ is bounded below by the number of distinct fringe subtrees in $T_{n,\sigma}$ of sizes k for $k_1 \leq k \leq n^{\delta_n/2}$, where $n^{\delta_n/2} = (\log n)^{\ell_\sigma}$ by definition of δ_n . Let $X_{n,k}$ denote the (random) number of fringe subtrees of size exactly k occurring in $T_{n,\sigma}$, and let $X_{n,k}^{(2)}$ denote the (random) number of unordered pairs of identical fringe subtrees of size exactly k occurring in a random tree $T_{n,\sigma}$. Using the inclusion-exclusion principle, we find that $\mathbb{E}(F_{n,\sigma})$ can be bounded from below as follows:

$$(22) \quad \mathbb{E}(F_{n,\sigma}) \geq \sum_{k_1 \leq k \leq n^{\delta_n/2}} \mathbb{E}(X_{n,k}) - \sum_{k_1 \leq k \leq n^{\delta_n/2}} \mathbb{E}(X_{n,k}^{(2)}).$$

We first bound the second sum on the right-hand side of (22). By Lemma 4.4, we find that the number of fringe subtrees of size at least $N_\sigma + 1$ in a binary tree t of size $k > N_\sigma$ is at least $k/N_\sigma - 1$. As $\sigma \in \mathcal{L}_{up}(\rho)$ with integer N_σ , we find

$$P_\sigma(t) = \prod_{v \text{ inner node of } t} \sigma(|t_\ell[v]|, |t_r[v]|) \leq \prod_{\substack{v \text{ node of } t \\ |t[v]| > N_\sigma}} \sigma(|t_\ell[v]|, |t_r[v]|) \leq \rho^{k/N_\sigma - 1}$$

for every binary tree $t \in \mathcal{T}_k$. Let us condition on the event that $X_{n,k} = N$ for some natural number N . These N fringe subtrees are independent random trees of size k , and the probability that such a fringe subtree equals a given binary tree $t \in \mathcal{T}_k$ is given by $P_\sigma(t)$. Thus, we have

$$\mathbb{E}(X_{n,k}^{(2)} \mid X_{n,k} = N) = \binom{N}{2} \sum_{t \in \mathcal{T}_k} P_\sigma(t)^2 \leq n^2 \rho^{k/N_\sigma - 1} \sum_{t \in \mathcal{T}_k} P_\sigma(t) = n^2 \rho^{k/N_\sigma - 1}.$$

Since this upper bound on $\mathbb{E}(X_{n,k}^{(2)} \mid X_{n,k} = N)$ holds independently of N , we obtain with the law of total expectation

$$\mathbb{E}(X_{n,k}^{(2)}) = \sum_{N=0}^n \mathbb{E}(X_{n,k}^{(2)} \mid X_{n,k} = N) \mathbb{P}(X_{n,k} = N) \leq n^2 \rho^{k/N_\sigma - 1}.$$

Since $k_1 \geq N_\sigma \left(1 + \frac{(1 + \delta_n) \log n}{\log(1/\rho)} \right)$, we obtain for all $k \geq k_1$:

$$\mathbb{E}(X_{n,k}^{(2)}) \leq n^{1 - \delta_n}.$$

We thus have

$$\sum_{k_1 \leq k \leq n^{\delta_n/2}} \mathbb{E}(X_{n,k}^{(2)}) \leq n^{1 - \delta_n/2} = \frac{n}{(\log n)^{\ell_\sigma}} \in o\left(\frac{n}{\log n}\right).$$

It remains to bound the first sum on the right-hand side of (22): If $Y_{n,k}$ again denotes the random number of fringe subtrees of size larger than k in the random tree $T_{n,\sigma}$, we have

$$\sum_{k_1 \leq k \leq n^{\delta_n/2}} \mathbb{E}(X_{n,k}) = \mathbb{E}(Y_{n,k_1-1}) - \mathbb{E}(Y_{n,n^{\delta_n/2}}).$$

By condition (c) of Theorem 4.10, we have $\mathbb{E}(Y_{n,n^{\delta_n/2}}) \in o(n/\log n)$. Moreover, by Lemma 4.12, we find that

$$\mathbb{E}(Y_{n,k_1-1}) \geq \frac{\gamma n}{c_\xi k_1 \xi(k_1)} - \frac{1}{c_\xi \xi(k_1)} \in \Omega\left(\frac{n}{\log n}\right),$$

since $\xi(x) \leq 1$. By monotonicity of ξ , we have $\xi(k_1) \leq \xi\left(\frac{N_\sigma \log n}{\log(1/\rho)}\right)$, so we obtain from (22) that

$$\mathbb{E}(F_{n,\sigma}) \geq \frac{\gamma \log(1/\rho)n}{c_\xi N_\sigma \xi\left(\frac{N_\sigma \log n}{\log(1/\rho)}\right) \log n} (1 + o(1)).$$

This finishes the proof. \square

Example 4.14. In this example, we use Theorem 4.10 in order to prove a lower bound on the expected number of distinct fringe subtrees in a random tree $T_{n,\sigma_{uni}}$, i.e., in a random tree of size n drawn from the set \mathcal{T}_n according to the uniform probability distribution σ_{uni} from Example 2.2. We have to show that all requirements of Theorem 4.10 are satisfied. Let us write σ for σ_{uni} in the following. Recall that

$$\sigma(i, n-i) = \frac{C_{i-1}C_{n-i-1}}{C_{n-1}}.$$

(i) $\sigma \in \mathcal{L}_{up}(\rho)$ for a constant $\rho \in (0, 1)$. A short computation shows that for a fixed integer n , the maximal value of $\sigma(i, n-i)$ with $1 \leq i \leq n-1$ is attained at $i_{max}(n) = 1$ and $i'_{max}(n) = n-1$. In particular, we have

$$\sigma(i, n-i) \leq \sigma(1, n-1) \leq \frac{n}{4n-6}$$

for all $1 \leq i \leq n-1$. If we fix $N_\sigma \in \mathbb{N}$, we thus find that $\sigma \in \mathcal{L}_{up}(\rho)$ with integer N_σ , where $\rho = N_\sigma/(4N_\sigma - 6)$. For the sake of simplicity, let $N_\sigma = 3$. Then $\sigma \in \mathcal{L}_{up}(\rho)$ with $\rho = 1/2$.

(ii) $\sigma \in \mathcal{L}_{unbal}(\xi, \gamma)$ for a decreasing function ξ and a constant γ . Let $0 < \gamma < 1/2$, and let $\varepsilon_0 > 0$ and $\delta_0 > 1$. From the asymptotic formula for the Catalan numbers (1), we find that there is an integer N'_σ (depending on ε_0 and δ_0) such that both

$$\frac{(1-\varepsilon_0)4^n}{\sqrt{\pi}n^{3/2}} \leq C_n \leq \frac{(1+\varepsilon_0)4^n}{\sqrt{\pi}n^{3/2}}$$

and

$$\frac{n^{5/2}}{(n-2)^2 \sqrt{(r-2)(n-\gamma n-1)}} \leq \frac{\delta_0}{\sqrt{(1-\gamma)r}}$$

for all $n, r \geq N'_\sigma$. Set $c_\sigma = N'_\sigma$, and suppose that $c_\sigma \leq r \leq \lceil \gamma n \rceil$. Then

$$\begin{aligned} \sum_{r \leq i \leq n-r} \sigma(i, n-i) &= \sum_{r \leq i \leq n-r} \frac{C_{i-1}C_{n-i-1}}{C_{n-1}} \\ &\leq \frac{(1+\varepsilon_0)^2 (n-1)^{3/2}}{(1-\varepsilon_0)4\sqrt{\pi}} \sum_{r \leq i \leq n-r} (i-1)^{-3/2} (n-i-1)^{-3/2} \\ &\leq \frac{(1+\varepsilon_0)^2 n^{3/2}}{(1-\varepsilon_0)2\sqrt{\pi}} \int_{r-1}^{n/2} (x-1)^{-3/2} (n-x-1)^{-3/2} dx \end{aligned}$$

$$\begin{aligned}
&= \frac{(1 + \varepsilon_0)^2 n^{3/2}}{(1 - \varepsilon_0)\sqrt{\pi}} \frac{n - 2r + 2}{(n - 2)^2 \sqrt{(r - 2)(n - r)}} \\
&\leq \frac{(1 + \varepsilon_0)^2}{(1 - \varepsilon_0)\sqrt{\pi}} \frac{n^{5/2}}{(n - 2)^2 \sqrt{(r - 2)(n - \gamma n - 1)}} \\
&\leq \frac{(1 + \varepsilon_0)^2}{(1 - \varepsilon_0)\sqrt{\pi}} \frac{\delta_0}{\sqrt{(1 - \gamma)r}}.
\end{aligned}$$

Set

$$\delta_1 := \frac{(1 + \varepsilon_0)^2 \delta_0}{1 - \varepsilon_0} > 1 \quad \text{and} \quad \xi(r) := \frac{\delta_1}{\sqrt{\pi(1 - \gamma)r}}.$$

Then $\sigma \in \mathcal{L}_{unbal}(\xi, \gamma)$, where the constant $\delta_1 > 1$ can take values arbitrarily close to 1 for suitable choices of ε_0 and δ_0 .

(iii) Condition (a) of Theorem 4.10 is clearly satisfied, i.e., $\xi_{uni}(x) \in \omega(1/x)$.

(iv) Condition (b) is satisfied as well: We have

$$\begin{aligned}
\sum_{i=N'_\sigma}^k \xi(i) &= \frac{\delta_1}{\sqrt{\pi(1 - \gamma)}} \sum_{i=N'_\sigma}^k i^{-1/2} \leq \frac{\delta_1}{\sqrt{\pi(1 - \gamma)}} \int_{N'_\sigma - 1}^k x^{-1/2} dx \\
&\leq \frac{2\delta_1 \sqrt{k+1}}{\sqrt{\pi(1 - \gamma)}} = 2(k+1)\xi(k+1).
\end{aligned}$$

Thus, condition (b) holds with $c_{\xi_{uni}} = 2$ (as $\gamma < 1/2$, we also have $1/c_{\xi_{uni}} + \gamma \leq 1$).

(v) Finally, condition (c) of Theorem 4.10 holds as well: From Example 3.15, we know that $\sigma \in \mathcal{L}_{sbal}(\vartheta, \gamma)$ with $\vartheta(x) \in \Theta(x^{-1/2})$ for $0 < \gamma < 1/2$. If we set $\ell_\sigma = 4$, then Lemma 3.14 yields

$$\mathbb{E}(Y_{n,(\log n)^4}) \in o\left(\frac{n}{\log n}\right).$$

We can therefore apply Theorem 4.10. The expected number of distinct fringe subtrees in a random tree of size n drawn according to the uniform probability distribution satisfies

$$\begin{aligned}
\mathbb{E}(F_{n,\sigma_{uni}}) &\geq \frac{\gamma n \log(1/\rho_{uni})}{c_\xi N_\sigma \xi\left(\frac{N_\sigma \log n}{\log(1/\rho)}\right) \log n} (1 + o(1)) \\
&= \frac{\gamma \sqrt{(1 - \gamma)\pi \log(1/\rho)}}{\delta_1 c_\xi \sqrt{N_\sigma}} \cdot \frac{n}{\sqrt{\log n}} (1 + o(1)) \\
&= \frac{\gamma \sqrt{(1 - \gamma)\pi}}{2\sqrt{3}\delta_1} \cdot \frac{n}{\sqrt{\log n}} (1 + o(1)).
\end{aligned}$$

As $0 < \gamma < 1/2$ is arbitrary, we can choose γ arbitrarily close to the optimal value $1/2$. We then obtain for $\sigma = \sigma_{uni}$

$$\mathbb{E}(F_{n,\sigma_{uni}}) \geq \frac{\sqrt{\pi}}{4\sqrt{6}\delta_1} \cdot \frac{n}{\sqrt{\log n}} (1 + o(1))$$

for any constant $\delta_1 > 1$. Recall that in Example 3.15, we have already shown that

$$\mathbb{E}(F_{n,\sigma_{uni}}) \leq 8\delta\sqrt{2\pi} \cdot \frac{n}{\sqrt{\log n}} (1 + o(1)),$$

for any constant $\delta > 1$. In particular, we thus have $\mathbb{E}(F_{n,\sigma_{uni}}) \in \Theta(n/\sqrt{\log n})$. We remark again that in [14, 23, 26], it is shown that in fact

$$\mathbb{E}(F_{n,\sigma_{uni}}) = \frac{2\sqrt{2}}{\sqrt{\pi}} \cdot \frac{n}{\sqrt{\log(n)}}(1 + o(1)).$$

It remains to remark that Theorem 4.10 can be applied to a wide variety of leaf-centric binary tree sources, including the binary search tree model from Example 2.1 and the binomial random tree model from Example 2.3: For these two leaf-centric binary tree sources, Corollary 4.11 again yields asymptotic lower bounds of the form $\mathbb{E}(F_{n,\sigma_{bst}}) \in \Omega(n/\log n)$ and $\mathbb{E}(F_{n,\sigma_{bin,p}}) \in \Omega(n/\log n)$, respectively.

5. CONCLUSION AND OPEN PROBLEMS

We have proposed several classes of leaf-centric binary tree sources and derived upper and lower bounds on the number of distinct fringe subtrees occurring in a random tree generated by a leaf-centric binary tree source from the respective class.

Another type of binary tree sources are *depth-centric binary tree sources* [16, 28], which yield probability distributions on the set of binary trees of a fixed depth and resemble leaf-centric tree sources in many ways. Furthermore, leaf-centric binary tree sources have been generalized to random tree models for plane trees called *fixed-size ordinal tree sources* in [21].

An interesting problem would be to estimate the number of distinct fringe subtrees with respect to classes of depth-centric or fixed-size ordinal tree sources. Furthermore, another open problem is to consider the question of estimating the number of distinct fringe subtrees in a leaf-centric binary tree source under a generalized interpretation of “distinctness”, as was done in [27] for simply generated families of trees and families of increasing trees.

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] David Aldous. Asymptotic fringe distributions for general families of random trees. *The Annals of Applied Probability*, 1(2):228–266, 1991.
- [3] Olivier Bodini, Antoine Genitrini, Bernhard Gittenberger, Isabella Larcher, and Mehdi Naima. Compaction for two models of logarithmic-depth trees: Analysis and experiments. *CoRR*, abs/2005.12997, 2020.
- [4] Mireille Bousquet-Mélou, Markus Lohrey, Sebastian Maneth, and Eric Noeth. XML compression via DAGs. *Theory of Computing Systems*, 57(4):1322–1371, 2015.
- [5] Randal E. Bryant. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys*, 24(3):293–318, 1992.
- [6] 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.
- [7] Florian Dennert and Rudolf Grübel. On the subtree size profile of binary search trees. *Combinatorics, Probability and Computing*, 19(4):561–578, 2010.
- [8] Luc Devroye. On the richness of the collection of subtrees in random binary search trees. *Information Processing Letters*, 65(4):195–199, 1998.
- [9] Luc Devroye and Svante Janson. Protected nodes and fringe subtrees in some random trees. *Electronic Communications in Probability*, 19:1–10, 2014.
- [10] Michael Drmota. *Random Trees: An Interplay Between Combinatorics and Probability*. Springer Publishing Company, Incorporated, 1st edition, 2009.
- [11] 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.
- [12] Philippe Flajolet, Xavier Gourdon, and Conrado Martínez. Patterns in random binary search trees. *Random Structures & Algorithms*, 11(3):223–244, 1997.
- [13] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

- [14] 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.
- [15] 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.
- [16] 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.
- [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.
- [18] 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.
- [19] Mokshay Madiman. On the entropy of sums. In *Proceedings of the 2008 IEEE Information Theory Workshop, ITW 2008*, pages 303–307. IEEE, 2008.
- [20] Conrado Martínez. *Statistics under the BST model*. PhD thesis, University Barcelona, 1992.
- [21] J. Ian Munro, Patrick K. Nicholson, Louisa Seelbach Benkner, and Sebastian Wild. Hyper-succinct trees - new universal tree source codes for optimal compressed tree data structures and range minima. In Petra Mutzel, Rasmus Pagh, and Grzegorz Herman, editors, *29th Annual European Symposium on Algorithms, ESA 2021, September 6-8, 2021, Lisbon, Portugal (Virtual Conference)*, volume 204 of *LIPICs*, pages 70:1–70:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.
- [22] Mike Paterson and Mark N. Wegman. Linear unification. *Journal of Computer and System Sciences*, 16(2):158–167, 1978.
- [23] 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.
- [24] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953.
- [25] 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.
- [26] 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.
- [27] Louisa Seelbach Benkner and Stephan G. Wagner. Distinct fringe subtrees in random trees. *CoRR*, abs/2105.04231, 2021.
- [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.