Compaction for two models of logarithmic-depth trees: Analysis and Experiments

Olivier Bodini, Antoine Genitrini, Bernhard Gittenberger, Isabella Larcher, Mehdi Naima

COMPACTION FOR TWO MODELS OF LOGARITHMIC-DEPTH TREES: ANALYSIS AND EXPERIMENTS

OLIVIER BODINI, ANTOINE GENITRINI, BERNHARD GITTENBERGER, ISABELLA LARCHER, AND MEHDI NAIMA

Abstract. We are interested in the quantitative analysis of the compaction ratio for two classical families of trees: recursive trees and plane binary increasing trees. These families are typical representatives of tree models with a small depth. Once a tree of size n is compacted by keeping only one occurrence of all fringe subtrees appearing in the tree the resulting graph contains only $O(n/\ln n)$ nodes. This result must be compared to classical results of compaction in the families of simply generated trees, where the analogous result states that the compacted structure is of size of order $n/\sqrt{\ln n}$. The result about the plane binary increasing trees has already been proved, but we propose a new and generic approach to get the result. Finally, an experimental study is presented, based on a prototype implementation of compacted binary search trees that are modeled by plane binary increasing trees.

Keywords: Analytic Combinatorics; Tree compaction; Common subexpression recognition; Increasing trees; Binary search trees

1. Introduction

Tree-shape data structures are omnipresent in computer science. The syntax structure of a program is a tree, symbolic expressions in computer algebra systems have a tree structure. Syntax trees arise in the context of parsing, XML data structures are also built on trees. However in order to reduce redundancy in the storage, usually an algorithmic step called the *common subexpression recognition* is run to identify identical fringe subtrees (*i.e.* a node and all its descendants) so that only one occurrence is stored and all other are replaced by pointers to the first one. Thus the trees are then replaced by directed acyclic graphs. In the context of tree compaction several studies attempt to quantitatively analyze the process of compaction. We mention here in particular two important research lines about compaction properties.

The first line occurs in the context of information theory and data compression studies. There researchers are interested in designing compression algorithms for advanced data structures. One of the main parameters of interest is the entropy of the data structure: it represents an optimal lower bound on the average number of bits required to

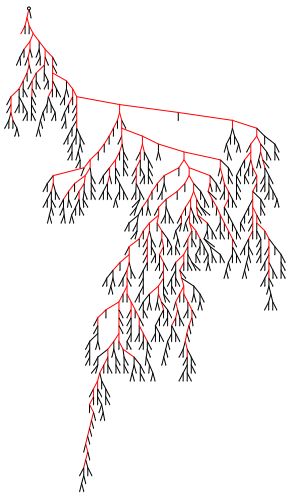

Figure 1. A uniformly sampled plane binary tree with 500 internal nodes: Black fringe subtrees are removed by the compaction process. The red part (that is what remains of the tree after pruning the black fringe subtrees) is of size 250.

Date: September 10, 2021.

This work was partially supported by the anr project Metaconc ANR-15-CE40-0014, by the phc # 39454SF, by the ÖAD grant FR04/2018 and by the Austrian Science Foundation (FWF), grant SFB F50-03.

represent the data structure: see for example [10] for an introduction to the subject. For trees, the entropy of some models of plane trees[1] has been studied in particular in [26, 9, 20].

An analysis of a model of non-plane binary trees has been presented in [9]. The authors focus on the number of symmetry nodes (internal nodes having two isomorphic subtrees as children) and its relation with Rényi entropy. In all investigations of that kind, the probability distribution used for the tree model is central. The aforementioned work [9] is focusing on a growing tree model that is also seen as the classical binary search tree distribution model. Likewise, it can be rephrased as the binary increasing tree model we will deal with in Section 3, as it was already pointed out in [4]. We are, however, interested in different aspects of these trees (see below for more details).

The second line of research has been started by the seminal paper of Flajolet *et al* [19]. In this paper the authors consider the compaction ratio of classical binary trees compared with their corresponding compacted structures. They prove, starting from a large binary tree of size n (containing n nodes) and then compacting it, then the average size of the compacted result is $\alpha n/\sqrt{\ln n}$ with a computable constant α. In the end of the paper the authors finally state that their analysis is fully adapted to all families of simply generated trees as defined by Meir and Moon in their fundamental paper [28] and thus for all kinds of tree structures we mentioned above as examples, we get the same kind of ratio for the compaction. In Figure 1 we have represented a uniformly sampled binary tree with 500 internal nodes. If we compact it then all the fringe subtrees in black are removed and only the red structure is kept with addition of several pointers (that are not represented in the figure). The remaining red tree is of size 250. We recall that in the context of simply generated trees of size n, the typical depth is of order $\sqrt{n}$ (this is the case for the binary trees). Bousquet-Mélou *et al.* [7] present the complete proof for the compaction quantitative analysis of simply generated tree families and apply it experimentally on XML-trees. Finally, in [32] the authors are interested in the number of fringe subtrees with at least r occurrences in a random simply generated tree. This approach is an extension of the previous results where it was dealt with subtrees appearing at least once (thus for $r = 1$).

But there are also several other kinds of tree structures that cannot be modeled through the concept of simply generated trees. In particular, we have in mind all structures used for searching, and thus usually with a small depth of order $\ln n$ for a whole structure of size n. The classical binary search trees (BST), red-black trees or AVL trees belong to these families. But we can also point out priority heaps like binary or binomial heaps. The reader can refer for example to Knuth's book [24] for details about all these structures. In this context, all nodes contain different labels (or information) and thus the compaction process as described before has no effect (no two subtrees are identical due to the labeling). But, if we remove the labels from the nodes, then a tree structure remains whose typical depth is of order $\ln n$ for n nodes. Hence we can compact the tree structure. In Figure 2 we have depicted a binary search tree structure of size 500. Once the

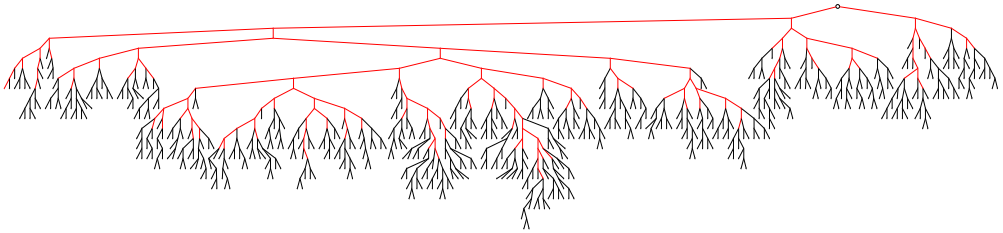

FIGURE 2. A uniformly sampled (plane) binary search tree structure with 500 internal nodes: Black fringe subtrees are removed by the compaction. The red part is only of size 172.

[1]Plane trees are such that the descendants of a node are ordered contrary to non-plane trees where the descendants are seen as a set instead of a sequence of subtrees.

structure is compacted, it remains a tree with 172 nodes (represented in red).

Our study focuses on the number of non-isomorphic subtrees in a tree and this corresponds also to the size of the compacted tree (also called minimal DAG representation in [33]). This parameter is different from the study of symmetry nodes mentioned above (see [9]), since there symmetries happen if an internal node has two isomorphic children (a local symmetry) whereas the number of non-isomorphic subtrees of a tree is capturing a global symmetry. Using the results in [9] to design and analyze a data compression algorithm leads to constant compression rate on average, as was already shown in [16]. In our case, we gain on average at least a logarithmic factor.

For both investigations, a Riccati-like functional-differential equation must be analyzed. But, not only the equations in [9] and in Section 3 are different, but the global nature of the parameter of our interest is reflected by the need of uniform asymptotics, which required a delicate singularity analysis.

In this paper, we analyze the underlying *unlabeled tree structure* of a plane and a non-plane model of increasingly labeled trees, namely increasing binary trees and recursive trees. For these two models of trees picking a tree uniformly at random and erasing the labels from it gives an unlabeled plane binary tree or an unlabeled non-plane general tree (also called Pólya tree). However, for each model the probability distribution of the resulting unlabeled tree is non-uniform. The distribution on plane binary trees we use is the same as the one of [9, 26]. Even if the analyzed parameters are not the same, for all such studies the mathematical tools are based on differential equation analyses due to the underlying distribution on trees.

Finally, another way to reach the non-uniform distribution is as a very simple natural evolution process. First let us mention the plane binary tree model: start with a single node, and at each step select randomly one of the leaves (external node) and replace with a binary node. While for Pólya trees, start with a node and at each step select randomly one of the nodes and append a new leaf to it.

We are interested in the analysis of the compaction ratio, relating the tree size and its minimal DAG size as in [33] for two families of trees that are not simply generated trees. The first family consists of recursive trees (Section 2). The family has been introduced by Moon [30] and further studied by Meir and Moon in the 70s [28]. Their motivation was to present a tree model for the spread of epidemics. The second tree family we are interested in is the class of plane binary increasing tree (Section 3). It corresponds to the tree model for binary search trees. Both families have been extensively studied in the last two decades with probabilistic methods [27, 12, 8, 14] as well as with combinatorial ones [4, 25, 31].

For recursive trees and binary increasing trees, informally speaking we prove that, asymptotically, if a tree of size n is compacted, then the resulting structure has on average size $\mathcal{O}(n/\ln n)$, with a lower bound of $\Omega(\sqrt{n})$.

In the context of binary increasing trees the result has already been derived. The upper bound $\mathcal{O}(n/\ln n)$ was proved in [16] as a specific result in the context of patterns in random binary search trees. The proof is based on some bivariate generating function analysis in the Analytic Combinatorics context. The stronger Θ-result has then been proved in [11] based on a preliminary result in [15]. These papers are based on probability theory rather than Analytic Combinatorics. But recently other authors [2, 3] presented new proofs based on Analytic Combinatorics. We, however, decided to briefly present a further proof based on Analytic Combinatorics, as it is generic in the following sense: the same approach is valid for recursive trees as well as for binary increasing trees. Especially in order to derive our results, we analyze a perturbation of the differential equation defining the tree models, observing that analogous functions related to the increasing labeling of the tree structure are central in both tree models. And under the assumption that a certain experimentally supported conjecture is true, almost the same proof can be used to improve the lower bound and get a Θ-result for both classes.

We thus remark that such a kind of trees are compacted in a more efficient way (in the sense of the number of remaining nodes) than simply generated trees. Finally, we end the paper (Section 4) with a section dedicated to the compaction of binary search trees (BST) in practice, in order to exhibit the way we can compact the tree structure, but by keeping some extra information we lose no information (about the labeling of the initial BST). An experimental study is provided by using

a prototype in *python* for our new data structure, the *compacted* BST. The experiments are very encouraging for the development of such new compacted search tree structures.

So, as a synthesis, Section 2 is dedicated to the compaction analysis of recursive trees. Then Section 3 contains the key elements to derive the same result for binary increasing trees and finally, Section 4 presents an experimental approach to verify the latter result in the context of data structures.

Remarks. We note that for all figures we present, we use a postorder traversal of the tree representation in order to compact them. However, whatever traversal is chosen, the quantitative results are always identical.

Recall that the size of the compacted tree also equals the number of distinct unlabeled fringe subtrees appearing in the original tree.

2. Recursive trees

The class of recursive trees has been studied by Meir and Moon [28]. These trees are models in several contexts as e.g. for the study of epidemic spreads, and thus many quantitative study have focused on this family. Some details are presented either in [13] or in [18]. Using the classical operators from Analytic Combinatorics, recursive trees can be specified by the so-called boxed product, or Greene operator,

$$\mathcal{T} = \mathcal{Z} \,^{\square}\star \textsc{Set}(\mathcal{T}),$$

meaning that the structure of a recursive tree (in the class $\mathcal{T}$) is defined as a root $\mathcal{Z}$ attached to a set of recursive trees (the set may be empty, then $\mathcal{Z}$ is a leaf) and such that the whole structure is canonically labeled (1, 2, ..., up to the size). The box in the boxed product indicates that the lowest label goes into the left component (the atom in this case). The atoms $\mathcal{Z}$ in the structure are therefore labeled increasingly on each path from the root of the tree to any leaf. See [18, Section II.6.3] for details about the constraint labeling operators. The class of recursive trees is also presented in [13, Section 1.3].

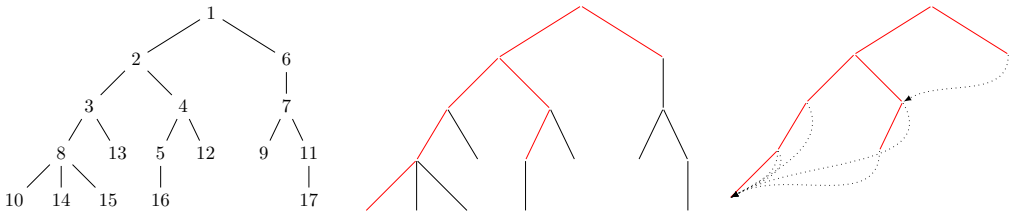

Figure 3. Example of a recursive tree of size 17

On the left side of Figure 3 we have represented a recursive tree of size 17. The children of a node are put in lexicographic order of their root labels. We remark that the unlabeled structures underlying the fringe subtrees rooted at 4 and 7 have the same unlabeled non-plane structure. And obviously the leaves are also identical. So, in the middle of the figure we represent with black edges the fringe subtrees whose unlabeled non-plane structure has already been seen through a postorder traversal of the leftmost tree. Finally, on the right side of the figure we replace the multiple occurrences of a subtree by pointers to the first occurrence.

In Figure 4 we have represented a recursive tree structure containing 5,000 nodes on the left side. It has been uniformly sampled among all trees with the same size. The original root of the tree is represented using a small circle ∘. On the right side we have depicted the nodes that are kept after the compaction of the latter tree. Only 663 nodes remain.

We define the exponential generating function $T(z) = \sum_{n\geq 1} T_n \frac{z^n}{n!}$, where T_n corresponds to the number of trees containing n nodes *i.e.* of *size* n. Using the now classical *symbolic method* from Analytic Combinatorics, from the latter unambiguous specification we deduce the following

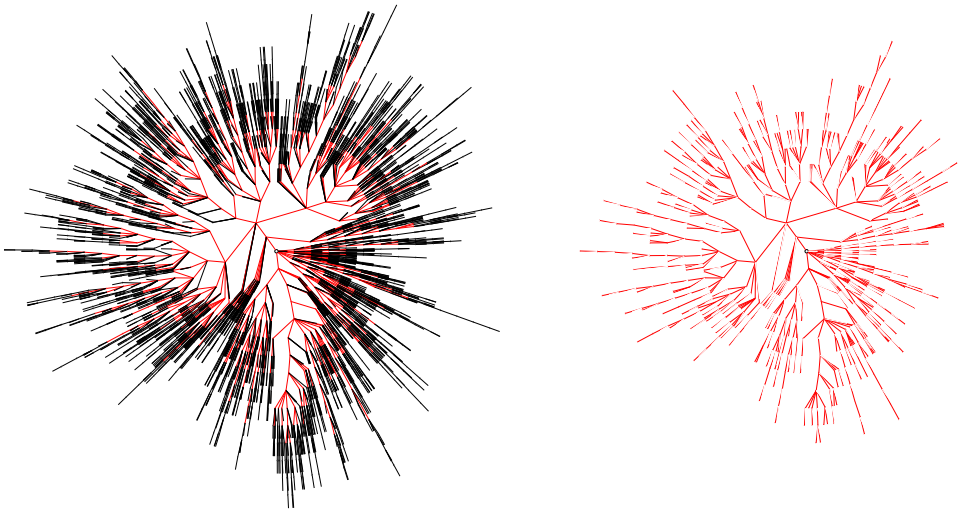

FIGURE 4. (left) A uniformly sampled non-plane recursive tree of size 5,000: Black fringe subtrees are removed by the compaction. (right) The red part is of size 663.

functional equation satisfied by $T(z)$:

$$T(z) = \int_0^z \exp(T(v))\,\mathrm{d}v.$$

The unique power series solution satisfying $T(0) = 0$ is

$$T(z) = \ln \frac{1}{1-z},$$

whose dominant singularity is $\rho = 1$. Finally, we get the value $T_n = (n-1)!$.

Let $\mathcal{T}_n$ be the class of recursive trees of size n; the size of a tree τ is defined as the number of its nodes and is denoted by $|\tau|$. Let X_n be the size of the compacted tree corresponding to a random recursive tree τ of size n. In other words, X_n is the number of distinct fringe subtree shapes in τ. We define $\mathcal{P}$ as the set of Pólya trees, *i.e.*, non-plane unlabeled trees such that the degrees of their nodes are arbitrary. This class of trees is presented in detail in Drmota's book [13, Section 1.2.5]. and it corresponds to the possible shapes of the recursive trees, once the increasing labeling has been removed. We denote by $\mathcal{P}_{\leq n}$ the set of all Pólya trees with size at most n. Then we have

$$\mathbb{E}\,(X_n) = \sum_{t\in\mathcal{P}_{\leq n}} \mathbb{P}(t \text{ occurs as subtree of } \tau) = \sum_{t\in\mathcal{P}_{\leq n}} 1 - \mathbb{P}(t \text{ does not occur as subtree of } \tau). \quad (1)$$

Recall that the tree t corresponds to a tree shape, it is unlabeled, while τ is a recursive tree and therefore is increasingly labeled.

Now, for a given Pólya tree $t \in \mathcal{P}$ let us consider a perturbed combinatorial class $\mathcal{S}_t$ that contains all recursive trees except for those that contain a t-shape as a (fringe) subtree. The corresponding exponential generating function satisfies the differential equation

$$S_t'(z) = \exp(S_t(z)) - P_t'(z), \quad (2)$$

where $P_t(z) = \ell(t)\frac{z^{|t|}}{|t|!}$, with $\ell(t)$ denoting the number of ways to increasingly label the tree shape t.

So, using (1) we obtain

$$\begin{aligned}\mathbb{E}\,(X_n) &= \sum_{t\in\mathcal{P}_{\le n}} (1-\mathbb{P}(t \text{ does not occur as shape of a fringe subtree of } \tau)) \\ &= \sum_{t\in\mathcal{P}_{\le n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right). \end{aligned} \tag{3}$$

Therefore, the problem is now essentially reduced to the analysis of the asymptotic behavior of $[z^n]S_t(z)$.

Solving (2) we obtain the exponential generating function

$$S_t(z) = \ln\left(\frac{1}{1-\int_0^z \exp(-P_t(v))\,\mathrm{d}v}\right) - P_t(z). \tag{4}$$

Thus, for the dominant singularity $\tilde{\rho}_t$ of $S_t(z)$, the following equation must hold:

$$\int_0^{\tilde{\rho}_t} \exp(-P_t(v))\,\mathrm{d}v = 1. \tag{5}$$

As $\exp(-P_t(v)) < 1$ for positive v, the dominant singularity $\tilde{\rho}_t$ is greater than 1. Recall that ρ denotes the dominant singularity of $T(z)$, thus $\rho = 1$ and therefore we write $\tilde{\rho}_t = \rho(1+\epsilon_t) = 1+\epsilon_t$ with suitable $\epsilon_t > 0$.

Notations. Before we proceed, let us introduce some frequently used notations: For the size and the weight of a Pólya tree t we use

$$k := |t| \qquad \text{and} \qquad w(t) := \frac{\ell(t)}{|t|!},$$

respectively. Moreover, let

$$G(z) := \int_0^z e^{-P_t(v)}\,\mathrm{d}v = \int_0^z e^{-w(t)v^k}\,\mathrm{d}v,$$

if $z \ge 0$ and its complex continuation if z is not a nonnegative real number. With this notation (5) reads as $G(1+\epsilon_t) = 1$. By expanding the integrand, we obtain

$$G(z) = \sum_{\ell\ge 0} (-w(t))^\ell \frac{z^{\ell k+1}}{(\ell k+1)\cdot \ell!},$$

which shows that $G(z)$ is an entire function.

How to proceed. Taking a random recursive tree of size n, we are interested in the asymptotic behavior of the size of the compacted tree issued from the compaction of the recursive one. In order to obtain bounds for this compacted size we proceed as follows: First, in Lemma 1, we compute a upper bound for $\tilde{\rho}_t$.

Then, in Proposition 1, we provide uniform asymptotics for the n-th coefficient of the generating function $S_t(z)$ when n tends to infinity, thereby showing that the error term is sufficiently small for what is needed later on.

The average size of a compacted tree corresponding to a random recursive tree is expressed as a sum over the forbidden trees. Thereby, the two cases where the size k of the forbidden tree t is smaller or larger than $\log n$ are treated in a different way: Upper bounds for the size of the compacted tree are derived in Proposition 2 (small trees) and Proposition 3 (large trees). Finally, Proposition 4, gives a (crude) lower bound for the size of the compacted tree.

Lemma 1. *Let $S_t(z)$ be the generating function of the perturbed combinatorial class (*cf. *Equation (2)) of recursive trees that do not contain a subtree of shape t and $\tilde{\rho}_t$ be the dominant singularity of $S_t(z)$ (*cf. *Equation (5)). Furthermore, let $k = |t|$ and $w(t) = \ell(t)/k!$ where $\ell(t)$ denotes the number of possible increasing labelings of the Pólya tree t. Then*

$$\tilde{\rho}_t = 1+\epsilon_t < 1 + \frac{2w(t)}{k}.$$

Proof. First observe that the number of increasing labelings of the Pólya tree t is bounded by $(k-1)!$, which gives the very crude bound $w(t) \leq 1/k$ which is valid for any tree t.

Next, as $\tilde{\rho}_t$ satisfies $G(1+\epsilon_t) = 1$, it suffices to show the inequality $G\left(1+\frac{2w(t)}{k}\right) > G(1+\epsilon_t)$. We show the equivalent inequality $G\left(1+\frac{2w(t)}{k}\right) - G(1) > G(1+\epsilon_t) - G(1)$: If $k=2$, then t is a path of length one and therefore $w(t) = 1/2$. This gives explicitly $\int_1^{3/2} e^{-v^2/2}\,\mathrm{d}v > 1/6$ which is easily verified.
If $k \geq 3$, then we have the lower bound

$$\begin{aligned} G\left(1+\frac{2w(t)}{k}\right) - G(1) &\geq \frac{2w(t)}{k}\exp\left(-w(t)\left(1+\frac{2w(t)}{k}\right)^k\right) \\ &\geq \frac{2w(t)}{k}\exp\left(-w(t)\left(1+\frac{2}{k^2}\right)^k\right), \end{aligned}$$

because $w(t) \leq 1/k$. Then for $k \geq 3$ we have $\left(1+\frac{2}{k^2}\right)^k < 2$ and again, since $w(t) \leq 1/k$, we obtain $2e^{-2/3} > 1$ and thus

$$G\left(1+\frac{2w(t)}{k}\right) - G(1) \geq \frac{w(t)}{k}\cdot 2e^{-2w(t)} > \frac{w(t)}{k+1}.$$

On the other hand, we have

$$G(1+\epsilon_t) - G(1) = 1 - \int_0^1 e^{-w(t)v^k}\,\mathrm{d}v \leq 1 - \int_0^1 (1-w(t)v^k)\,\mathrm{d}v = \frac{w(t)}{k+1},$$

which implies the assertion. □

With a similar reasoning as in the above proof a lower bound for $\tilde{\rho}_t$ can be shown:

Corollary 1. *With the notations of Lemma 1 we have the following estimate:*

$$\tilde{\rho}_t > 1 + \frac{w(t)}{k+1}.$$

Corollary 2. *With the notations of Lemma 1 we have the following asymptotic relation:*

$$\tilde{\rho}_t = 1 + \epsilon_t \sim 1 + \frac{w(t)}{k}, \text{ as } k \to \infty.$$

Proof. Write $G(z)$ as $G(z) = z + R(z)$ with

$$R(z) = \sum_{\ell \geq 1} (-w(t))^\ell \frac{z^{\ell k+1}}{(\ell k+1)\cdot \ell!} \tag{6}$$

As $\tilde{\rho}_t = 1+\epsilon_t$ is the smallest positive solution of $G(z) = 1$, it is the smallest positive zero of $z - 1 + R(z)$. From Lemma 1 we know that $\epsilon_t = \mathcal{O}\left(1/k^2\right)$ and thus $\tilde{\rho}_t^k \sim 1$, as k tends to infinity, and $R(\tilde{\rho}_t) = w(t)\tilde{\rho}_t^{k+1}/(k+1) + \mathcal{O}\left(1/k^3\right)$. This implies

$$\epsilon_t \sim \frac{w(t)}{k+1}\tilde{\rho}_t^{k+1} \sim \frac{w(t)}{k},$$

as desired. □

Remark. In the paper [21], which is related to pattern exclusion in recursive trees, the same result about the singularity $\tilde{\rho}_t$ is proved. Using more terms of the expansion of $G(z)$, it is possible to derive a more accurate asymptotic expression for ϵ_t (in principle up to arbitrary order). As an example, we state

$$\tilde{\rho}_t = 1 + \frac{w(t)}{k+1} + \frac{w(t)^2(3k+1)}{(k+1)(4k+2)} + \frac{w(t)^3(29k^3+32k^2+10k+1)}{6(k+1)^2(2k+1)(3k+1)} + \mathcal{O}\left(\frac{w(t)^4}{k}\right).$$

Note that in the sequel we will have to evaluate the coefficient $[z^n]S_t(z)$ for n tending to infinity and $|t|$ tending to infinity with n as well. Thus a standard transfer lemma in the sense of Flajolet and Odlyzko [17] is not sufficient. We need a tight and uniform error term. In order to find this, we need to know where the second dominant singularity is, or rather where we can be sure that there will not be any singularity. The next lemma provides information about an eventually large enough singularity-free region.

Lemma 2. *Let $S_t(z)$ be the generating function of the perturbed class of recursive trees defined in (4). Then $S_t(z)$ has no singularity in the domain*

$$\tilde{\rho}_t < |z| < 1 + \frac{\ln(1/w(t)) + \ln\ln\ln(1/w(t))}{k}.$$

Proof. Recall that by (4) we have

$$S_t(z) = \ln\left(\frac{1}{1-G(z)}\right) - P_t(z).$$

Since $G(z)$ is an entire function, the singularities of $S_t(z)$ are exactly the zeros of $G(z)-1$. Therefore, consider z_0 such that $G(z_0)=1$ and write $G(z) = z + R(z)$ with $R(z)$ as in (6). Then the chosen number z_0 must satisfy the inequality

$$|R(z_0)| \le \frac{|z_0|}{k+1}\sum_{\ell\ge 1}\frac{|w(t)|^\ell |z_0|^{k\ell+1}}{\ell!} < \frac{1}{k}(e^{|w(t)||z_0|^k} - 1). \tag{7}$$

The first step is to show that $G(z)-1$ does not have any zeros (except $\tilde{\rho}_t$) in a sufficiently large domain, *i.e.* that either $z_0 = \tilde{\rho}_t$ or $|z_0|$ is large. We have to approach this in two steps.

Case 1: Assume first that $|z_0| \le 1 + \frac{\alpha\ln(1/w(t))}{k}$ for some $\alpha < 1$. As the dominant singularity of $S_t(z)$ is $\tilde{\rho}_t$ and $\tilde{\rho}_t > 1$, we must have $|z_0| > 1$. Thus, the upper bound on $|z_0|$ implies $|z_0|^k \le \exp(\alpha\ln(1/w(t))) = (1/w(t))^\alpha = o(1/w(t))$ and by (6) we obtain then

$$1 - z_0 = R(z_0) \sim -\frac{w(t)}{k}z_0^k = o\left(\frac{1}{k}\right). \tag{8}$$

This implies further that $z_0^k \sim 1$, hence z_0 is asymptotically equal to a k-th root of unity. But then $z_0 = \tilde{\rho}_t$, because the distance between the other k-th roots of unity and 1 is greater than $1/k$, which contradicts (8).

Case 2: Now, assume that $|z_0| = 1+\eta$ with $\alpha\ln(1/w(t))/k < \eta \le (\ln(1/w(t))+\ln\ln\ln(1/w(t))-\delta)/k$ for some arbitrary but small $\delta > 0$. Then $w(t)|z_0|^k \le \ln\ln(1/w(t))e^{-\delta}$ and so by (7) we have then

$$|R(z_0)| \le \frac{\left(\ln\frac{1}{w(t)}\right)^{e^{-\delta}} - 1}{k}. \tag{9}$$

But we assumed $|z_0 - 1| > \alpha\ln(1/w(t))/k$ and so $R(z_0)$ would be too small to compensate the value of $z_0 - 1$. Indeed, we observe that in this region

$$|G(z)-1| > \frac{1}{k} + \frac{\alpha\ln\frac{1}{w(t)} - \left(\ln\frac{1}{w(t)}\right)^{e^{-\delta}}}{k} \ge \frac{\gamma\ln\frac{1}{w(t)}}{k} \tag{10}$$

holds, where γ is a suitable positive constant.

Summarizing what we have so far, we infer that either $z_0 = \tilde{\rho}_t$ or

$$|z_0| \ge 1 + \frac{\ln(1/w(t)) + \ln\ln\ln(1/w(t))}{k},$$

as claimed. □

Now we are able to derive a uniform asymptotic expression for the coefficients of $S_t(z)$ with a sufficiently small error term.

Proposition 1. *Let $S_t(z)$ be the generating function of the perturbed class of recursive trees defined in (4) and fix a constant $L > 2$. Then, uniformly for $D \le |t| \le n$ with D independent of n and sufficiently large, the following asymptotic relations hold, depending of the magnitude of $w(t)$:*

- *If* $\ln\frac{1}{w(t)} = o\left(\sqrt{k}\right)$*, then the coefficients of* $S_t(z)$ *behave asymptotically as follows:*

$$[z^n]S_t(z) = \frac{\tilde{\rho}_t^{-n}}{n}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{k}}\left(\frac{kcw(t)}{\ln\ln\frac{1}{w(t)}}\right)^{n/k}\right)\right),\ \text{as } n\to\infty,$$

  where c *is an arbitrary constant satisfying* $c > 1$*.*
- *If* $\ln\frac{1}{w(t)} = \Omega\left(\sqrt{k}\right)$ *and* $\ln\frac{1}{w(t)} \le Lk$*, then*

$$[z^n]S_t(z) = \frac{\tilde{\rho}_t^{-n}}{n}\left(1 + \mathcal{O}\left(\exp\left(\frac{n}{k}\cdot\frac{\ln(L+1)}{L}\ln(kw(t))\right)\right)\right),\ \text{as } n\to\infty.$$

- *If* $\ln\frac{1}{w(t)} > Lk$*, then*

$$[z^n]S_t(z) = \frac{\tilde{\rho}_t^{-n}}{n}\left(1 + \mathcal{O}\left(\ln(k)\exp\left(-n\left(\ln\left(\ln\frac{1}{w(t)} - \ln k\right) - \ln k\right)\right)\right)\right),\ \text{as } n\to\infty.$$

Proof. Notice that $G'(\tilde{\rho}_t) = \exp\left(-w(t)\tilde{\rho}_t^k\right) \neq 0$ and therefore $\tilde{\rho}_t$ is a simple zero of $G(z)-1$. Thus $G(z)-1 = (z-\tilde{\rho}_t)\tilde{G}(z)$ where $\tilde{G}(z)$ is analytic in the considered domain and does not have any zeros there. Thus,

$$S_t(z) = \ln\left(\frac{1}{1-G(z)}\right) - P_t(z) = -\ln\left(1-\frac{z}{\tilde{\rho}_t}\right) - \ln(\tilde{\rho}_t\tilde{G}(z)) - P_t(z),$$

where, apart from the first summand, there are no singularities in $|z| < 1 + \frac{\ln(1/w(t)) + \ln\ln\ln(1/w(t))}{k}$ (see Lemma 2). Expanding the logarithm gives

$$[z^n]S_t(z) = \frac{\tilde{\rho}_t^{-n}}{n}\left(1 + \mathcal{O}\left(n\tilde{\rho}_t^n[z^n]\ln\tilde{G}(z)\right)\right)$$

and we want to estimate $[z^n]\ln\tilde{G}(z)$ using Cauchy's estimate. Therefore we use the integration contour

$$\Gamma := \left\{z:\ |z| = 1 + \frac{\ln\frac{1}{w(t)} + \ln\ln\ln\frac{1}{w(t)} - \delta}{k}\right\}$$

for some small $\delta > 0$, which we split into a part Γ_1 where $|z-1| \le 5\ln(1/w(t))/k$ and its complement Γ_2.

As we want to estimate the logarithm of $\tilde{G}(z) = (G(z)-1)/(z-\tilde{\rho}_t)$, we need an upper and a lower bound for $\tilde{G}(z)$.

First of all, note that on the whole integration contour certain useful inequalities hold, provided that k is sufficiently large:

$$|z-\tilde{\rho}_t| \ge |z-1| - |1-\tilde{\rho}_t| \ge |z-1| - \frac{2w(t)}{k} \ge |z-1|\left(1 - \frac{2w(t)}{\ln\frac{1}{w(t)}}\right) \ge \frac{|z-1|}{2},$$

$$|z-\tilde{\rho}_t| \le |z-1| + |\tilde{\rho}_t - 1| \le |z-1| + \frac{2w(t)}{k} \le |z-1|\left(1 + \frac{2w(t)}{\ln\frac{1}{w(t)}}\right) \le 2|z-1|,$$

which is true, because $|1-\tilde{\rho}_t| < 2w(t)/k$ due to Lemma 1 and $\ln(1/w(t))/k \le |z-1|$. For $z\in\Gamma_1$ the upper bound can be slightly improved: Indeed, we even have $|z-\tilde{\rho}_t| \le |z-1|$. Moreover, recall the inequality

$$|R(z)| \le \frac{\alpha\ln\frac{1}{w(t)} - 1}{k},$$

which follows from (9). On Γ_1 we also have

$$\frac{\ln(1/w(t))}{2k} \le |z-\tilde{\rho}_t| \le \frac{5\ln(1/w(t))}{k}. \tag{11}$$

From all these inequalities we infer a universal upper bound (for all $z\in\Gamma_1\cup\Gamma_2$) for $\tilde{G}(z)$:

$$\left|\frac{G(z)-1}{z-\tilde{\rho}_t}\right| \le \frac{|z-1|}{|z-\tilde{\rho}_t|} + \frac{|R(z)|}{|z-\tilde{\rho}_t|} \le 2 + \frac{2(e-1)}{\ln(1/w(t))} \le 3.$$

Here we used that the first inequality in (11) actually holds on the whole integration contour. Using (10) and the second inequality in (11) we get for $z \in \Gamma_1$ the lower bound

$$\left|\frac{G(z)-1}{z-\tilde{\rho}_t}\right| \geq \frac{\gamma}{5} > \frac{1}{5}.$$

These two bounds and the fact that the length of the curve Γ_1 is less than $10\ln(1/w(t))/k$ imply

$$\left|[z^n]\ln\tilde{G}(z)\right| \leq \left(1+\frac{\ln\frac{1}{w(t)}+\ln\ln\ln\frac{1}{w(t)}-\delta}{k}\right)^{-n}\frac{10\ln\left(\frac{1}{w(t)}\right)\ln 5}{k}+\frac{1}{2\pi}\int_{\Gamma_2}\frac{|\ln\tilde{G}(z)|}{|z|^{n+1}}|\,\mathrm{d}z|.$$

Turning to Γ_2, we obtain the lower bound

$$\left|\frac{G(z)-1}{z-\tilde{\rho}_t}\right| \geq \frac{|z-1|}{|z-\tilde{\rho}_t|}-\frac{|R(z)|}{|z-\tilde{\rho}_t|} \geq \frac{1}{2}-\frac{\alpha\ln\frac{1}{w(t)}-1}{k}\frac{k}{5\ln\frac{1}{w(t)}} \geq \frac{1}{10},$$

and so $|\ln\tilde{G}(z)|$ is bounded on Γ_2.

Finally, let $M := \max(\ln(10), 10\ln(1/w(t))\ln(10)/k)$. Altogether the above estimates show that for sufficiently large k we have

$$\begin{aligned}
n\tilde{\rho}_t^n|[z^n]\ln\tilde{G}(z)| &\leq n\tilde{\rho}_t^n\left(1+\frac{\ln\frac{1}{w(t)}+\ln\ln\ln\frac{1}{w(t)}-\delta}{k}\right)^{-n}\left(\frac{10\ln\left(\frac{1}{w(t)}\right)\ln 5}{k}+\ln(10)M\right)\\
&\leq n\left(1+\frac{\ln\frac{1}{w(t)}+\ln\ln\ln\frac{1}{w(t)}-2\delta}{k}\right)^{-n}\left(\frac{10\ln\left(\frac{1}{w(t)}\right)\ln 5}{k}+\ln(10)M\right)\\
&= \mathcal{O}\left(n\left(1+\frac{\ln\frac{1}{w(t)}-\ln k+\ln\ln\ln\frac{1}{w(t)}-2\delta}{k}+\frac{\ln n}{n}\right)^{-k\cdot\frac{n}{k}}\cdot\frac{\ln\frac{1}{w(t)}}{k}\right) \qquad (12)\\
&= \mathcal{O}\left(\frac{\ln\frac{1}{w(t)}}{k}\left(\frac{w(t)ke^{2\delta}}{\ln\ln\frac{1}{w(t)}}\right)^{n/k}\right),
\end{aligned}$$

where the last step is only true in the case where $\ln(1/w(t)) = o\left(\sqrt{k}\right)$ and yields the desired result after all.

In all the other cases, only the last step is different. Indeed, going back to (12), we can estimate $\ln\ln\ln\frac{1}{w(t)}-2\delta > 0$ and thus

$$n\tilde{\rho}_t^n|[z^n]\ln\tilde{G}(z)| = \mathcal{O}\left(\frac{\ln\frac{1}{w(t)}}{k}(1+X)^{-n}\right),$$

with $X = (\ln(1/w(t))-\ln k)/k$.

If $\ln(1/w(t)) = \Omega\left(\sqrt{k}\right)$, but $\ln(1/w(t)) \leq Lk$, we write $(1+X)^{-n} = \exp(-n\ln(1+X))$ and get the final result by using $\ln(1+X) \geq X\ln(L+1)/L$, which is true for $0 \leq X \leq L$. The prefactor $\ln(1/w(t))/k$ is bounded by L in the considered case.

And finally, if $\ln(1/w(t)) > Lk$ (and so $X > L$), then simply use $\ln(1+X) > \ln X$. This yields

$$(1+X)^{-n} \leq \exp\left(-n\left(\ln\left(\ln\frac{1}{w(t)}-\ln k\right)-\ln k\right)\right).$$

As $w(t) \leq (k-1)!$, we get $\ln(1/w(t))/k = \mathcal{O}(\ln k)$ and the proof is complete. □

The uniform error term in Proposition 1 allows us to derive a simple upper bound for k not too large. It turns out that the bound in Corollary 1 is actually good enough to cover the error term from Proposition 1.

Corollary 3. *If k is sufficiently large, then*

$$[z^n]S_t(z) \leq \frac{1}{n}\left(1+\frac{w(t)}{k+1}\right)^{-n},$$

as n tends to infinity and $k = \mathcal{O}(\sqrt{n})$.

Proof. We know from Proposition 1 that $[z^n]S_t(z) = \tilde{\rho}_t^{-n} n^{-1}(1+r_n)$ with $r_n = o(1)$. Thus we must show that

$$r_n \leq \tilde{\rho}_t^n \left(1+\frac{w(t)}{k+1}\right)^{-n} - 1 = \left(1+\mathcal{O}\left(w(t)^2\right)k\right)^n - 1.$$

As r_n tends to 0, the inequality is trivial if $nw(t)^2/k$ does not tend to 0, as in this case the right-hand side grows exponentially. Otherwise we are left with having to show the estimate $r_n = \mathcal{O}\left(nw(t)^2/k\right)$. Let us compare $nw(t)^2/k$ with the exponential part of the error term given by Proposition 1. In the case where $w(t)$ is large ($\ln\frac{1}{w(t)} = o(\sqrt{k}\,)$) this gives

$$\begin{aligned}
&\left(\frac{kcw(t)}{\ln\ln\frac{1}{w(t)}}\right)^{n/k} \frac{k}{nw(t)^2} \\
&\quad = \exp\left(\left(-\frac{n}{k}+2\right)\ln\frac{1}{w(t)} + \left(\frac{n}{k}+1\right)\ln k - \frac{n}{k}\ln\ln\ln\frac{1}{w(t)} + \frac{n}{k}\ln c - \ln n\right) \\
&\quad \leq \exp\left(3\ln k - \frac{n}{k}\ln\ln\ln k + \frac{n}{k}\ln c - \ln n\right),
\end{aligned}$$

where the inequality holds because of $\ln\frac{1}{w(t)} \geq \ln k$. As our assumptions imply $n/k \to \infty$ and so the dominant term in the exponent, $-\frac{n}{k}\ln\ln\ln k$, is negative, we obtain $r_n = o(nw(t)^2/k)$ as desired.

In the case where $w(t)$ has intermediate size, the difference of the logarithms of the exponential term in the error and of $nw(t)^2/k$ is equal to

$$\left(-\frac{n}{k}\frac{\ln(L+1)}{L}+2\right)\ln\frac{1}{w(t)} + \left(\frac{n}{k}\frac{\ln(L+1)}{L}+1\right)\ln k - \ln n$$

which is negative if $k = \mathcal{O}(\sqrt{n})$.

Finally, if $w(t)$ is small, then the difference of the logarithms equals

$$\begin{aligned}
&-n\left(\ln\left(\ln\frac{1}{w(t)} - \ln k\right) - \ln k\right) + 2\ln\frac{1}{w(t)} + \ln k - \ln n \\
&\quad \leq -n(\ln((L-1)k) - \ln k) + 2\ln\frac{1}{w(t)} + \ln k - \ln n \\
&\quad \leq -n\ln(L-1) + 2\ln\frac{1}{w(t)} + \ln k - \ln n
\end{aligned}$$

which is again negative if $k = \mathcal{O}(\sqrt{n})$. □

Within this section many logarithms that occur are with respect to the base $\frac{1}{\sigma} \approx 2.9955765$, where $\sigma \approx 0.3383218$ denotes the dominant singularity of the generating function of Pólya trees (*cf.* [18, Section VII.5]). We thus use the notation $\log_{\frac{1}{\sigma}}$ for the logarithm with respect to base $\frac{1}{\sigma}$.

Now we decompose the sum (3) into

$$\mathbb{E}(X_n) = \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log_{\frac{1}{\sigma}} n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) + \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log_{\frac{1}{\sigma}} n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right), \tag{13}$$

and investigate the two sums individually, starting with the first one, whose summands are probabilities and thus bounded by 1.

Proposition 2. *The first sum in (13) behaves asymptotically as*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log_{\frac{1}{\sigma}} n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \underset{n\to\infty}{=} \mathcal{O}\left(\frac{n}{\sqrt{(\ln n)^3}}\right).$$

Proof. Remember that we have set $k := |t|$. Furthermore, we denote by $P(z)$ the generating function of Pólya trees and by σ its dominant singularity. Then

$$\begin{aligned}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log_{\frac{1}{\sigma}} n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) &\leq \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log_{\frac{1}{\sigma}} n}} 1 = \sum_{k<\log_{\frac{1}{\sigma}} n} [z^k]P(z)\\ &\sim \frac{1}{1-\sigma}[z^{\lfloor \log_{\frac{1}{\sigma}} n\rfloor}]P(z) = \mathcal{O}\left(\frac{\sigma^{-\lfloor \log_{\frac{1}{\sigma}} n\rfloor}}{\sqrt{(\log_{\frac{1}{\sigma}} n)^3}}\right).\end{aligned}$$

Since $\log_{\frac{1}{\sigma}} n$ has the base $1/\sigma$, we estimate $\sigma^{-\lfloor \log_{\frac{1}{\sigma}} n\rfloor} \leq n$, which completes the proof. □

In order to analyze the second sum from (13) we rely on counting arguments, which were presented in [21, Remark 4.2]. For the sake of self-containedness we restate the counting arguments here.

Proposition 3. *The second sum in* (13) *behaves asymptotically as*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log_{\frac{1}{\sigma}} n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \mathcal{O}\left(\frac{n}{\log_{\frac{1}{\sigma}} n}\right).$$

Proof. Remember that we have set $k := |t|$ and k tends to infinity in this proof. We are interested in $1-\frac{[z^n]S_t(z)}{[z^n]T(z)}$, the probability that a tree of size n contains a fringe subtree of shape t.

We start with a counting argument, allowing multiple counting, to construct a tree of size n having a fringe subtree of shape t. Let ν denote the root label of t in the tree of size n. If several occurrences of t do appear, we consider one of them.

First suppose $k < n$. Then choose a tree of size $n-k$ to which t will be attached. Recall that the number of possible choices for that tree equals $(n-k-1)!$. The number of ways to choose the labels of t is $\binom{n-\nu}{k-1}$, as ν is the smallest label in t and $|t| = k$. Once the labels for t have been chosen, there are $\ell(t)$ possibilities to distribute them over the vertices of t in order to obtain a proper labeling. The initially chosen (and already labeled) tree of size $n-k$ gets the remaining labels (that have not been chosen for t), which replace the original label in an order-preserving way. Finally, there are $\nu-1$ possible parent nodes to which t can be attached.

Putting all this together, we get the number of all recursive trees of size n having t as a fringe subtree, but each counted as many times as there are occurrences of t. This is clearly an upper bound. We obtain

$$\begin{aligned}1-\frac{[z^n]S_t(z)}{[z^n]T(z)} &\leq \frac{(n-k-1)!}{(n-1)!}\sum_{\nu=2}^{n-k+1}(\nu-1)\binom{n-\nu}{k-1}\ell(t) \qquad (14)\\ &= \frac{\ell(t)}{(k-1)!}\frac{n}{k(k+1)} = \frac{nw(t)}{k(k+1)}.\end{aligned}$$

Now let $k = n$. This means that we are interested in the probability that a recursive tree has shape t. In this case,

$$1-\frac{[z^n]S_t(z)}{[z^n]T(z)} = \frac{\ell(t)}{(n-1)!} = nw(t).$$

Now we apply this to the sum we want to estimate. Recall that $\sum_{t\in\mathcal{P}_k} w(t) = 1/k$. We get

$$\begin{aligned}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log_{\frac{1}{\sigma}} n}}\left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) &\leq n\sum_{t\in\mathcal{P}_n} w(t) + \sum_{k\geq\log_{\frac{1}{\sigma}} n}\frac{n}{k+1}\sum_{t\in\mathcal{P}_k} w(t)\\ &= 1+\sum_{k\geq\log_{\frac{1}{\sigma}} n}\frac{n}{k(k+1)}\\ &= 1+\sum_{k\geq\log_{\frac{1}{\sigma}} n} n\left(\frac{1}{k}-\frac{1}{k+1}\right) = \Theta\left(\frac{n}{\log_{\frac{1}{\sigma}} n}\right)\end{aligned}$$ □

Theorem 1. *Let X_n be the size of the compacted tree corresponding to a random recursive tree τ of size n. Then*

$$\mathbb{E}\left(X_n\right) \underset{n\to\infty}{=} \mathcal{O}\left(\frac{n}{\ln n}\right).$$

Proof. The result follows directly by combining the previous propositions. □

Finally, we prove a lower bound for the average size of the compacted tree based on a random recursive tree of size n.

Proposition 4. *Let $\mathcal{P}_{\leq n}$ denote the class of Pólya trees of size at most n. Then*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ \log_{\frac{1}{\sigma}} n\leq k\leq\sqrt{n}}}\left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \underset{n\to\infty}{=} \Omega\left(\sqrt{n}\right).$$

Proof. For the sake of simplified reading we will use the abbreviation $\sum_t := \sum_{t\in\mathcal{P}_k}$ in this proof.

First, we use Corollary 3 and the inequality $(1+x)^{-n} \leq \exp\left(-nx+\frac{nx^2}{2}\right)$ in order to estimate

$$\begin{aligned}A_n := \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor}\sum_t\left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) &\geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor}\sum_t\left(1-\left(1+\frac{w(t)}{k+1}\right)^{-n}\right)\\ &\geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor}\sum_t\left(1-\exp\left(-\frac{nw(t)}{k+1}+\frac{nw(t)^2}{(k+1)^2}\right)\right). \quad (15)\end{aligned}$$

Since $x\mapsto 1-\exp\left(-nx+\frac{nx^2}{2}\right)$, $0\leq x\leq 2$, is a concave nonnegative function with a zero in the origin and $x = w(t)/(k+1)$ certainly falls in this range for all t, we can estimate the inner sum in (15), which yields

$$A_n \geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor}\left(1-\exp\left(-n\sum_t\frac{w(t)}{k+1}+n\left(\sum_t\frac{w(t)}{k+1}\right)^2\right)\right)$$

As $\sum_t w(t)\leq 1/k$, we get

$$\begin{aligned}A_n &\geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor}\left(1-\exp\left(-\frac{n}{(k+1)^2}+\mathcal{O}\left(\frac{n}{k^4}\right)\right)\right)\\ &\underset{n\to\infty}{\sim} \int_{\log_{\frac{1}{\sigma}} n}^{\sqrt{n}}\left(1-\exp\left(-\frac{n}{x^2}+\mathcal{O}\left(\frac{n}{x^4}\right)\right)\right)\mathrm{d}x\\ &= \sqrt{n}\int_{n^{-1/2}\log_{\frac{1}{\sigma}} n}^{1}\left(1-\exp\left(-\frac{1}{y^2}+\mathcal{O}\left(\frac{1}{ny^4}\right)\right)\right)\mathrm{d}y.\end{aligned}$$

Since the integral is convergent this gives a lower bound that is $\Theta(\sqrt{n})$. □

We strongly believe that the upper bound presented in Theorem 1 is in fact the actual order of magnitude. Unfortunately, we cannot prove this. It seems that a finer knowledge on the distribution of the values of $w(t)$ is necessary.

Conjecture 1. *If $k \geq \log_{\frac{1}{\sigma}} n$, then $\sum_{t\in\mathcal{P}_k} w(t)^2 = \mathcal{O}\left(1/n\right)$.*

It is not easy to carry out experiments to support or disprove this conjecture. But for small value of n this works and they seem to confirm the conjecture. If it is true, then our conjecture on the order of magnitude $\mathbb{E}(X_n)$ is true as well.

Theorem 2. *If Conjecture 1 is true, then*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ \log_{\frac{1}{\sigma}} n\leq k\leq\sqrt{n}}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \underset{n\to\infty}{=} \Omega\left(\frac{n}{\ln n}\right).$$

Consequently, then $\mathbb{E}(X_n) = \Theta(n/\ln n)$.

Proof. Let us again use the notation $\sum_t := \sum_{t\in\mathcal{P}_k}$. Then by Corollary 3 we have

$$\sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor} A_n := \sum_t \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor} \sum_t \left(1-\left(1+\frac{w(t)}{k+1}\right)^{-n}\right).$$

The function $f(x) = 1-(1+x)^{-n}$ is concave, monotonically increasing for $x \geq 0$ and nonnegative there. Moreover, $f(0) = 0$. Thus $f(x) \geq xf'(x)$, since the slope at some $x_0 > 0$ is flatter than the slope at 0 and so the line $x \mapsto xf'(x_0)$ stays below the graph of f at least until $x = x_0$. This implies

$$A_n \geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor} \sum_t \frac{nw(t)}{k+1}\left(1+\frac{w(t)}{k+1}\right)^{-n-1} = \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor} \frac{n}{k(k+1)} \sum_t kw(t)\left(1+\frac{w(t)}{k+1}\right)^{-n-1}.$$

Now observe that $\sum_t kw(t) = 1$ and that $g(x) = 1/(1+x)^{n+1}$ is a convex function. Thus the last sum is a convex linear combination of values of $g(x)$ and so Jensen's inequality gives

$$A_n \geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor} \frac{n}{k(k+1)}\left(1+\sum_t \frac{k}{k+1}w(t)^2\right)^{-n-1}.$$

Under our assumption that Conjecture 1 is true, this can be further transformed into

$$A_n \geq \sum_{k=\lfloor\log_{\frac{1}{\sigma}} n\rfloor}^{\lfloor\sqrt{n}\rfloor} \frac{n}{k(k+1)}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)^{-n-1} = \Theta\left(\frac{n}{\ln n}\right). \qquad \square$$

3. Plane increasing binary trees

As already mentioned in the introduction, the main result of this section related to the size of the compaction of a random binary increasing tree (or a random binary search tree) has already been proved. But here we want to show that the methodology of the previous section is applicable to other classes of increasing trees as well. Thus we aim at presenting a new proof of this known result based on the same approach as the one we used for random recursive trees. Thus, many proofs will only be sketched.

Plane binary increasing trees have a classical specification in the context of Analytic Combinatorics, once again by using the Greene operator, or boxed product, allowing to define increasing labeling constraint for decomposable objects. Thus the specification of this class $\mathcal{T}$ is

$$\mathcal{T} = \mathcal{Z}^{\,\square} \star (1+\mathcal{T})^2. \tag{16}$$

This specification defines a tree to be rooted with an atom $\mathcal{Z}$ associated to a pair of elements that are either the empty element (representing no subtree) or a subtree itself from the class $\mathcal{T}$. Once

again the operator $\cdot \,^{\square}_{\star}\, \cdot$ ensures the fact that the smallest available label must be used for the atom $\mathcal{Z}$.

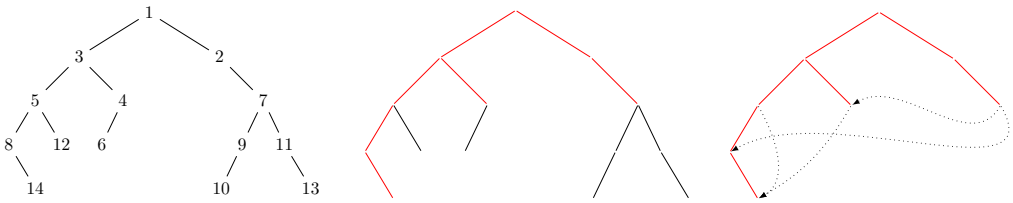

FIGURE 5. Example of a plane increasing binary tree of size 14

On the left side of Figure 5 we present an example of a plane increasing binary tree. Note that the internal nodes have a left child or a right child or both children. In particular, the unlabeled subtree rooted at 8 is the same as the one rooted at 11, but they are not the same as the one rooted either at 4 or at 9. The two other structures in the right of the figure are the compacted version of the plane increasing binary tree. In [13, Section 1.3.3] Drmota exhibits the link between plane increasing binary trees and binary search trees.

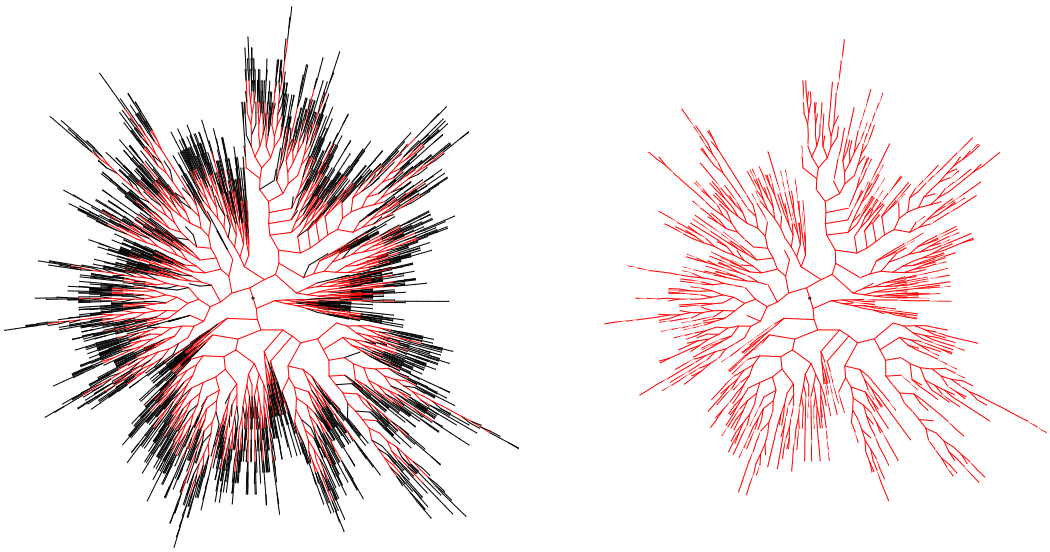

FIGURE 6. (left) A uniformly sampled (plane) increasing binary tree of size 5,000: Black fringe subtrees are removed by the compaction. (right) The red part is of size 1,361.

In Figure 6 we have represented on the left side a plane increasing binary tree structure containing $5,000$ nodes. It has been uniformly sampled among all trees with the same size. The original root of the tree is represented using a small circle $\circ$. On the right side, we have depicted the nodes that are kept after the compaction of the latter tree. Only $1,361$ nodes remain.

By using the symbolic method [18], the latter specification (16) translates as

$$T(z) = \int_0^z (1 + T(v))^2 \, \mathrm{d}v,$$

in terms of $T(z)$ the exponential generating function for $\mathcal{T}$. We can also rewrite it as a differential equation

$$T'(z) = (1 + T(z))^2, \qquad \text{with } T(0) = 0.$$

The equation can be solved such that

$$T(z) = \frac{z}{1-z},$$

with the dominant singularity $\rho = 1$.

The exponential generating function $S_t(z)$ of the perturbed class of plane increasing binary trees that do not contain the tree shape t (where t is a non-labeled binary tree) as a fringe subtree, fulfills the equation

$$S_t'(z) = (1 + S_t(z))^2 - P_t'(z) \qquad \text{with } S_t(0) = 0 \tag{17}$$

where $P_t(z) = \frac{\ell(t) z^{|t|}}{|t|!}$ and $\ell(t)$ denotes the number of ways to increasingly label the plane binary tree t. The quantity $\ell(t)$ is also called the hook length of t and it is well known that $\ell(t)$ equals $|t|!$ divided by the product of the sizes of all fringe subtrees of t (*cf.* e.g. [24, p.67] or [6]). We first start with a lemma establishing an upper bound for the normalized hook length.

Lemma 3. *Let t be a binary tree of size k. By defining the weight of the tree t as $w(t) := \frac{\ell(t)}{k!}$, where $\ell(t)$ denotes the hook length of t, we have*

$$w(t) \leq \frac{1}{2^{k-2}}.$$

Key ideas of the proof. Transforming the hook length formula into a recursive relation we prove

$$w(t) = \begin{cases} \frac{1}{k} w(t') & \text{if the root of } t \text{ has one child } t' \\ \frac{1}{k} w(t_\ell) w(t_r) & \text{if the root of } t \text{ has the two children } t_\ell \text{ and } t_r. \end{cases}$$

Set $w_n := \max_{t \in \mathcal{P}_n} w(t)$. Then compute the first values w_1 up to w_7, which confirms the claim, and then proceed by induction for $k \geq 8$. □

Finally, note that the term by term inverse of the sequence $(w_n)_{n \geq 0}$ corresponds to the sequence stored as `OEIS A132862`[2].

By the same combinatorial argument as in the previous section we know that $S_t(z)$ has a unique dominant singularity $\tilde{\rho}_t$, which is greater than the dominant singularity $\rho = 1$ of $T(z)$. Thus, we set again $\tilde{\rho}_t = \rho(1+\epsilon_t) = 1+\epsilon_t$. Since (17) is a Riccati differential equation (*cf.* [23] for a background on Riccati equations), we use the ansatz $S_t(z) = \frac{-u'(z)}{u(z)}$ to get the transformed equation

$$u''(z) - 2u'(z) + (1 - w(t) k z^{k-1}) u(z) = 0, \tag{18}$$

where we use the same abbreviations as in the previous section, namely $k := |t|$ and $w(t) := \frac{\ell(t)}{k!}$. Note that the condition $S_t(0) = 0$ implies $u'(0) = 0$ and $u(0) \neq 0$.

The singularities of a function $u(z)$ solving a linear differential equation (with polynomial coefficients) are given by the singularities of the coefficient of the highest derivative, *i.e.*, in our case the coefficient of $u''(z)$, which is 1. The reader can refer to Miller [29] for details. Thus, we can conclude that $u(z)$ is an entire function. As a direct consequence we know that the singularities of $S_t(z)$ are given by the zeros of $u(z)$ (that are not zeros of $u'(z)$) and are therefore poles. More precisely the dominant singularity $\tilde{\rho}_t$ must be a simple pole for $S_t(z)$, since for $u(z) = (\tilde{\rho}_t - z)^l v(z)$, (such that ρ is not a zero of $v(z)$), it follows that $u'(z) = -(\tilde{\rho}_t - z)^{l-1} v(z) + (\tilde{\rho}_t - z)^l v'(z)$. Thus

$$S_t(z) = \frac{l}{\tilde{\rho}_t - z} - \frac{v'(z)}{v(z)},$$

which implies

$$S_t(z) \underset{z \to \tilde{\rho}_t}{\sim} \frac{l/\tilde{\rho}_t}{1 - z/\tilde{\rho}_t}.$$

[2]This corresponds to the reference of this sequence in Sloane's Online Encyclopedia of: Integer Sequences `www.oeis.org`.

Taking the derivative we get $S_t'(z) \sim \frac{1}{\tilde{\rho}_t^2}\frac{l}{(1-z/\tilde{\rho}_t)^2}$. Plugging in the asymptotic expressions for S_t and S_t' in the original differential equation (17) we get

$$\frac{1}{\tilde{\rho}_t^2}\frac{l}{\left(1-\frac{z}{\tilde{\rho}_t}\right)^2} \underset{z\to\tilde{\rho}_t}{\sim} \left(1+\frac{l/\tilde{\rho}_t}{1-\frac{z}{\tilde{\rho}_t}}\right)^2,$$

since the monomial P_t is analytic in $\tilde{\rho}_t$. Comparing the main coefficients yields $l=1$, and thus $\tilde{\rho}_t$ is a simple zero of the function $u(z)$ and

$$S_t(z) \underset{z\to\tilde{\rho}_t}{\sim} \frac{1}{\tilde{\rho}_t - z}.$$

How to proceed. As in the previous section, we have a singularity $\tilde{\rho}_t = 1+\epsilon_t$ with $\epsilon_t > 0$ depending on t, or k. In order to get results on the average size of the compacted tree of a random increasing binary tree we proceed similarly to the recursive tree case. Lemma 5 gives an asymptotic expression for $\tilde{\rho}_t$ that quantifies its dependence on t, when the size k of the "forbidden" tree tends to infinity.

As a next step, Lemma 6 shows that $S_t(z)$ has a unique dominant singularity $\tilde{\rho}_t$ on the circle of convergence, which is used in Proposition 5 to obtain the asymptotic behavior of the coefficients of the generating function $S_t(z)$.

Again, the average size of a compacted tree can be represented as a sum over the forbidden trees, where we distinguish between the two cases whether the size of the trees is smaller or larger than $\ln n$ in order to get an upper bound (see Propositions 6 and 7).

We start from the equation $u''(z) - 2u'(z) + (1-w(t)kz^{k-1})u(z) = 0$ with the initial conditions $u(0)=\gamma$, and $u'(0)=0$. The value γ can be chosen arbitrarily, as $S_t(z) = \gamma u'(z)/(\gamma u(z))$, and thus, γ cancels. For simplification reasons in the following we choose $u(0)=-1$ together with the initial condition $u'(0)=0$.

Lemma 4. *The function $u(z)$ defined by the differential equation* (18) *and the initial conditions $u(0)=-1$ and $u'(0)=0$ satisfies*

$$u(z) = ze^z \sum_{m\geq 0}\left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!(m+\alpha)_m} z^{(k+1)m} - e^z \sum_{m\geq 0}\left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!(m-\alpha)_m} z^{(k+1)m},$$

where $(x)_m$ denotes the falling factorials $(x)_m = x(x-1)\cdots(x-m+1)$ and $\alpha = 1/(k+1)$.

Recall (for details refer to the book of Bender and Orszag [1]) that the solutions $y(z)$ of the ordinary differential equation

$$z^2y''(z) + zy'(z) + (z^2-\alpha^2)y(z) = 0,$$

with α not being an integer are linear combinations of the Bessel functions $J_\alpha(z)$ and $Y_\alpha(z)$. Some modifications on (18) let us exhibit the combination of Bessel functions that yields the result of Lemma 4.

Proof (sketch). Substituting $y(z) := u(z)\cdot\exp(-z)/\sqrt{z}$ and then $x := \left(\frac{k+1}{2\sqrt{-w(t)k}}z\right)^{2/(k+1)}$ transforms the differential equation for $u(z)$ into

$$\beta^2 y''(\beta) + \beta y'(\beta) + \left(\beta^2 - \frac{1}{(k+1)^2}\right) y(\beta) = 0,$$

with $\beta = \frac{2\sqrt{-w(t)k}}{k+1} t^{(k+1)/2}$. We recognize the Bessel equation and thus $y(\beta)$ is a linear combination of the Bessel functions $J_\alpha(\beta)$ and $Y_\alpha(\beta)$.

Due to the relationship between the function $u(z), y(\beta)$ and the Bessel functions, we deduce $u(z)$ is a linear combination of the functions $f(z)$ and $\bar{f}(z)$ where

$$f(z) = \sqrt{z}\exp(z)J_\alpha\left(2\tilde{\beta}z^{\frac{1}{2\alpha}}\right) \quad \text{and} \quad \bar{f}(z) = \sqrt{z}\exp(z)J_{-\alpha}\left(2\tilde{\beta}z^{\frac{1}{2\alpha}}\right),$$

with $\tilde{\beta} := \frac{\sqrt{-w(t)k}}{k+1}$ and $\alpha := \frac{1}{k+1}$.

By means of the initial conditions $u(0) = -1$ and $u'(0) = 0$ the coefficients of the linear combination can be computed. Finally, using the well-known power series expansions for $J_\alpha(x)$ and $J_{-\alpha}(x)$ as well as the formula $\frac{\Gamma(1+\alpha)}{\Gamma(m+1+\alpha)} = \frac{1}{(m+\alpha)_m}$, with $(x)_m$ being the falling factorials $(x)_m = x(x-1)\cdots(x-m+1)$, the previously obtained sum of power series eventually simplifies to the expression in the assertion. □

We are now ready to analyze the dominant singularity of $S_t(z)$.

Lemma 5. *Let $S_t(z)$ be the generating function of the perturbed combinatorial class of plane increasing binary trees that do not contain the shape t as a subtree (of size k). With $\tilde{\rho}_t$ denoting the dominant singularity of $S_t(z)$, we get*

$$\tilde{\rho}_t = 1 + \epsilon_t \underset{k\to\infty}{\sim} 1 + \frac{2w(t)}{k^2},$$

where $w(t) = \frac{\ell(t)}{k!}$ and $\ell(t)$ denotes the hook length of t.

Proof. For combinatorial reasons we deduced that the equation $u(z) = 0$ must have a solution $\tilde{\rho}_t > 1$ and no smaller positive solution. When k tends to infinity we expect that $\tilde{\rho}_t = 1+\epsilon_t$ tends to 1, *i.e.* ϵ_t tends to 0.

First observe that $u(0) = -1$ and

$$u\left(1 + \frac{1}{k^2}\right) = \frac{1}{k^2} + \mathcal{O}\left(\frac{w(t)}{k}\right) > 0,$$

as $w(t)$ decays exponentially due to Lemma 3. Thus $\epsilon_t = \mathcal{O}\left(1/k^2\right)$ and plugging $z = 1+\epsilon_t$ into $u(z) = 0$ gives then

$$\epsilon_t + (1+\epsilon_t)^{k+1}\frac{w(t)k}{(k+1)^2}\left(\frac{1+\epsilon_t}{1+\alpha} - \frac{1}{1-\alpha}\right) = \mathcal{O}\left(\frac{w(t)^2}{k^2}\right).$$

This implies $\epsilon_t - 2w(t)/k^2 = \mathcal{O}\left(w(t)^2/k^2\right)$ and hence $\epsilon_t \sim 2w(t)/k^2$, which finishes the proof. □

So, Lemma 5 ensures that for $|t| = k$ tending to infinity the generating function $S_t(z)$ has a dominant singularity at $\tilde{\rho}_t \sim 1 + 2w(t)/k^2$. Now we show that in a sufficiently large disk there is no other singularity for $S_t(z)$.

Lemma 6. *Let $\tilde{\rho}_t$ be the dominant singularity of $S_t(z)$. Then, for all $\delta > 0$ the following assertion holds: If k is sufficiently large, then the generating function $S_t(z)$ does not have any singularity in the domain $\tilde{\rho}_t < |z| < 1 + \frac{(1-\delta)\ln(1/w(t))+\ln k}{k}$.*

Proof (sketch). First note that the singularities of $S_t(z)$ are exactly the zeros of $u(z)$. Define $\tilde{u}(z) := u(z)\exp(-z)$ and note that $u(z)$ and $\tilde{u}(z)$ have the same zeros. By Lemma 4 we can write $\tilde{u}(z) = zF(z) - G(z)$ with

$$F(z) = \sum_{m\geq 0}\left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!}\frac{1}{(m+\alpha)_m} z^{(k+1)m}, \quad \text{and}$$

$$G(z) = \sum_{m\geq 0}\left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!}\frac{1}{(m-\alpha)_m} z^{(k+1)m},$$

with $\alpha := 1/(k+1)$. Therefore we get $|F(z) - G(z)| = \mathcal{O}\left(w(t)|z|^{k+1}/k^2\right)$. Now, let us rewrite $\tilde{u}(z)$ as

$$\tilde{u}(z) = (z-1)F(z) + F(z) - G(z), \tag{19}$$

set $|z| = 1+\eta$ and perform a distinction of two cases:

Case 1: $\eta = \mathcal{O}(1/k)$. This implies $|z|^{k+1} = \Theta(1)$ for k tending to infinity. Thus $F(z) \sim 1$, $G(z) \sim 1$, and then $F(z)-G(z) \to 0$. In view of this, (19) and $\tilde{u}(z) = 0$ imply $z-1 \sim G(z)-F(z) = \mathcal{O}\left(w(t)/k^2\right)$ and thus $|z-1| = \mathcal{O}(\tilde{\rho}_t - 1) = o(1/k)$.

But for zeros z_0 of $\tilde{u}(z)$ with $|z_0| = 1 + o\left(1/k\right)$ we know $z_0 - 1 \sim (2w(t)/k^2) \cdot z_0^k \sim 2w(t)/k^2$, so $z_0^k \sim 1$. Hence $z_0 \sim \sqrt[k]{1} = \cos\left(\frac{2\pi}{k}\right) + i \sin\left(\frac{2\pi}{k}\right)$ which contradicts $z_0 - 1 \sim 2w(t)/k^2$. Thus, the function $\tilde{u}(z)$ has no zeros in the domain $\tilde{\rho}_t < |z| \leq 1 + \mathcal{O}\left(1/k\right)$.

Case 2: $\eta = C_k/k$, with $C_k \leq (1-\delta)\ln\frac{1}{w(t)} + \ln k$. In this case we have $|z|^{k+1} \leq e^{C_k} = \mathcal{O}\left(k/w(t)^{1-\delta}\right)$, and thus $|F(z) - G(z)| = o(1/k)$ and $F \sim 1 + \mathcal{O}\left(w(t)^{\delta}\right)$. Using again (19) yields

$$\tilde{u}(z) = z - 1 + o(|z-1|w^{\delta}) + o(1/k) \sim z - 1. \tag{20}$$

Since $|z| = 1 + \eta$ we have $|z-1| \geq C_k/k > 1/k$ and thus $\tilde{u}(z)$ cannot be zero in $\tilde{\rho}_t < |z| < 1 + ((1-\delta)\ln\frac{1}{w(t)} + \ln k)/k$. □

Now we are interested in the ratio $[z^n]S_t(z)/[z^n]T(z)$, which corresponds to the probability that a random plane binary tree of size n does not contain the binary tree shape t as a fringe subtree.

Proposition 5. *Let $T(z)$ be the generating function of plane binary increasing trees and $S_t(z)$ the generating function of the perturbed class that has the dominant singularity $\tilde{\rho}_t$. Fix a constant $L > 2$. Then, uniformly for $D \leq |t| \leq n$ with D independent of n and sufficiently large, the following asymptotic relations hold, depending on the magnitude of $w(t)$:*

- *If $\ln\frac{1}{w(t)} \leq Lk$, then*

$$[z^n]S_t(z) = \tilde{\rho}_t^{-n-1}\left(1 + \mathcal{O}\left(\exp\left(-\frac{n}{k}\cdot\frac{\ln(L+1)}{L}\ln\frac{1}{w(t)}\right)\right)\right),\ \text{as } n \to \infty.$$

- *If $\ln\frac{1}{w(t)} > Lk$, then*

$$[z^n]S_t(z) = \tilde{\rho}_t^{-n-1}\left(1 + \mathcal{O}\left(\ln(k)\exp\left(-n\left(\ln\left((1-\delta)\ln\frac{1}{w(t)}\right) - \ln k\right)\right)\right)\right),\ \text{as } n \to \infty,$$

with arbitrary $\delta > 0$.

Remark. In contrast to Proposition 1 there are only two cases. The reason is that we know from Lemma 3 that $\ln\frac{1}{w(t)}$ cannot be too small. In fact, we have $(k-1)\ln 2 \leq \ln\frac{1}{w(t)} \leq k\ln k$.

Proof (sketch). First, let us remember that $\tilde{\rho}_t$ is a unique zero of the function $u(z)$. Thus, we can write

$$u(z) = \left(1 - \frac{z}{\tilde{\rho}_t}\right)v(z), \tag{21}$$

with $v(\tilde{\rho}_t) \neq 0$ and by Lemma 6 we additionally know that $v(z) \neq 0$ in $\tilde{\rho}_t < |z| < 1 + \frac{(1-\delta)\ln(1/w(t))+\ln k}{k}$, provided that k is sufficiently large. This implies

$$S_t(z) = \frac{1}{\tilde{\rho}_t - z} - \frac{v'(z)}{v(z)}.$$

And thus,

$$[z^n]S_t(z) = \tilde{\rho}_t^{-n-1} - [z^n]\frac{v'(z)}{v(z)} = \tilde{\rho}_t^{-n-1} - (n+1)[z^{n+1}]\ln v(z). \tag{22}$$

Now, we estimate the second summand in (22). First we use a Cauchy integral to write

$$n\ [z^n]\ln v(z) = \frac{n}{2\pi i}\int_{\mathcal{C}}\frac{\ln v(t)}{t^{n+1}}\,\mathrm{d}t, \tag{23}$$

where the curve $\mathcal{C}$ is described by $|t| = 1 + \frac{(1-\delta)\ln(1/w(t))+\ln k}{k}$ with some $\delta > 0$. The absolute value of the logarithm of $v(z)$ is given by $|\ln v(z)| = \left|\ln\left(|v(z)|e^{i\arg v(z)}\right)\right| = |\ln|v(z)| + i\arg(v(z))|$. Furthermore, by (21) we have $|v(z)| = |u(z)|/\left|1 - z/\tilde{\rho}_t\right|$, which can be estimated along $\mathcal{C}$ via

$$\frac{|u(z)|}{2 + \ln k} \leq |v(z)| \leq \frac{k|u(z)|}{(1-\delta)\ln(1/w(t))}.$$

Now, we have to estimate $|u(z)|$. Using the expansion in Lemma 4 and estimating, we find that there is a $\mu > 0$ such that

$$|u(z)| \le e^{|z|} \sum_{m\ge 0} \left(\frac{w(t)}{k}\right)^m \frac{1}{m!} \left| \frac{z}{(m+\alpha)_m} - \frac{1}{(m-\alpha)_m} \right| |z|^{(k+1)m}$$
$$\le e^{|z|} \sum_{m\ge 0} w^{\delta m} \frac{2+\mu}{m!(m-\alpha)_m} = \mathcal{O}\left(k\right).$$

To get a lower bound, observe that $|u(z)| \ge e^{-|z|}|\tilde{u}(z)| \ge |\tilde{u}(z)|/(ke)$ and by (20) we have $\tilde{u}(z) \sim z-1$. This yields $|u(z)| \ge (1-\delta)\ln(1/w(t))/(k^2 e) = \Omega(\ln(k)/k^2)$.

Putting all together, we can estimate the integral (23) by

$$n|[z^n]\ln v(z)| = \mathcal{O}\left(n\ln k \left(1 + \frac{(1-\delta)\ln(1/w(t)) + \ln k}{k}\right)^{-n}\right)$$
$$= \mathcal{O}\left(n\tilde{\rho}_t^{-n}\ln k \left(1 + \frac{(1-\delta)\ln(1/w(t)) + \ln k - \delta}{k}\right)^{-n}\right)$$
$$= \mathcal{O}\left(n\tilde{\rho}_t^{-n}\ln k \left(1 + \frac{(1-\delta)\ln(1/w(t))}{k} + \frac{\ln n}{n}\right)^{-n}\right).$$

Finally, proceed as at the end of the proof of Proposition 1 to complete the proof. □

Now, we split the sum of interest, *i.e.* $\sum_{t\in\mathcal{B}} \mathbb{P}\left[t \text{ occurs at subtree of } \tau\right]$, where τ denotes a plane increasing binary tree of size n and $\mathcal{B}$ denotes the class of (unlabeled) plane binary trees, analogously as we did in the previous section for recursive trees.

Remark. Now our underlying class of tree shapes is the class of plane binary trees and no more the class of Pólya trees. Since the dominant singularity of the generating function of binary trees is $1/4$, we use henceforth $\log_4 n$, the logarithm with respect to base 4.

$$\mathbb{E}\left(X_n\right) = \sum_{\substack{t\in\mathcal{B}_{\le n}\\ k<\log_4 n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) + \sum_{\substack{t\in\mathcal{B}_{\le n}\\ k\ge\log_4 n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right). \tag{24}$$

In order to estimate the first sum, we proceed analogously to Proposition 2.

Proposition 6. *Let $B(z)$ be the generating function associated to $\mathcal{B}$, of (unlabeled) binary trees, whose dominant singularity is $1/4$. Then asymptotically when n tends to infinity we have*

$$\sum_{\substack{t\in\mathcal{B}_{\le n}\\ k<\log_4 n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \underset{n\to\infty}{=} \mathcal{O}\left(\frac{n}{\sqrt{(\ln n)^3}}\right).$$

Proof. A crude estimate gives

$$\sum_{\substack{t\in\mathcal{B}_{\le n}\\ k<\log_4 n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \le \sum_{\substack{t\in\mathcal{B}_{\le n}\\ k<\log_4 n}} 1 = \sum_{k<\log_4 n} [z^k]B(z) \underset{n\to\infty}{\sim} \frac{1}{1-\frac{1}{4}}[z^{\lfloor\log_4 n\rfloor}]B(z)$$
$$\underset{n\to\infty}{=} \mathcal{O}\left(\frac{\left(\frac{1}{4}\right)^{-\lfloor\log_4 n\rfloor}}{\sqrt{(\log_4 n)^3}}\right).$$

This is already sufficient, since $\left(\frac{1}{4}\right)^{-\lfloor\log_4 n\rfloor} \le n$, which completes the proof. □

Estimating the second sum in (24) is based on some counting arguments, analogously to the proof of Proposition 3 in the previous section. However, due to the fewer grafting possibilities for the tree shape t a straight-forward analog of the proof of Proposition 3 yields a too crude upper bound. Thus a finer analysis is needed.

Proposition 7. *Let $\mathcal{B}_{\leq n}$ denote the class of binary trees of size at most n. Then*

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq \log_4 n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \underset{n\to\infty}{=} \mathcal{O}\left(\frac{n}{\ln n}\right).$$

Proof. Using a similar counting approach as the one proposed in Proposition 3, we obtain for $k=|t|<n$

$$1-\frac{[z^n]S_t(z)}{[z^n]T(z)} \leq \frac{1}{n!}\sum_{\nu=2}^{n-k+1}\binom{n-\nu}{k-1}\ell(t)(n-k)!(\nu-1).$$

Let t be the shape that appears in a tree of size n and ν be the root label of this occurrence of t. Then there at most $\nu-1$ possibilities to attach t to a tree of size $n-k$, because the node to which t is attached must have a label smaller than ν and a free place, as we consider incomplete binary trees here. Moreover, there are $\binom{n-\nu}{k-1}$ ways to choose the labels for t, $\ell(t)$ ways to make a proper labeling on t with the chosen labels, and $(n-k)!$ trees of size $n-k$ to which t will be attached. As we want an upper bound, we do not care for multiple counting.

After simplification we obtain $\ell(t)/k!\cdot(n-k)/(k+1)$, but this is too large to get the analog of Proposition 3. The problem here comes from the fact there are usually much fewer possibilities to graft t, in particular if $|t|$ is small. To get a better upper bound, we rely on [5, Theorem 5], where it is proved that the number of binary increasing trees of size $i+j$ having exactly k subtrees attached to the head of the tree (that is the minimal subtree that contains all nodes labeled with the smallest i labels) is

$$\binom{i+1}{k}\binom{j-1}{k-1}\, i!\; j!.$$

Here we are interested in trees of size $n-k$ containing the first $\nu-1$ labels and having exactly r available possibilities to graft the tree t, thus $\nu-r$ trees are already attached to the head. According to the above formula the number of such trees is

$$\binom{\nu}{\nu-r}\binom{n-k-\nu}{\nu-r-1}\,(\nu-1)!\,(n-k-\nu+1)!.$$

If $\nu<n-k+1$ this value is correct for r ranging from 0 to $\nu-1$. Otherwise if $\nu=n-k+1$ then r can also be equal to ν and then the number of possibilities to attach t to all heads of size $n-k$ is $(n-k+1)!$. So we obtain the better upper bound

$$\begin{aligned}1-\frac{[z^n]S_t(z)}{[z^n]T(z)} \leq & \frac{1}{n!}\sum_{\nu=2}^{n-k+1}\binom{n-\nu}{k-1}\ell(t)\sum_{r=1}^{\nu-1} r\binom{\nu}{\nu-r}\binom{n-k-\nu}{\nu-r-1}(\nu-1)!(n-k-\nu+1)!\\ &+\ell(t)\frac{(n-k+1)!}{n!}.\end{aligned}$$

Following [22, p. 169] this simplifies to

$$\sum_{r=1}^{\nu-1} r\binom{\nu}{\nu-r}\binom{n-k-\nu}{\nu-r-1} = \nu\binom{n-k-1}{\nu-2}.$$

Using this result we deduce

$$\begin{aligned}1-\frac{[z^n]S_t(z)}{[z^n]T(z)} &\leq \frac{\ell(t)}{n!}\sum_{\nu=2}^{n-k+1}\frac{\nu!(n-\nu)!(n-k-1)!}{(k-1)!(\nu-2)!(n-k-\nu+1)!}+\ell(t)\frac{(n-k+1)!}{n!}\\ &=\ell(t)\frac{(n-k-1)!}{n!}\sum_{\nu=2}^{n-k+1}\nu(\nu-1)\binom{n-\nu}{k-1}+\ell(t)\frac{(n-k+1)!}{n!}.\end{aligned}$$

Again using [22, p. 169], we further simplify and get

$$\sum_{\nu=2}^{n-k+1}\nu(\nu-1)\binom{n-\nu}{k-1} = 2\sum_{\nu=2}^{n-k+1}\binom{\nu}{2}\binom{n-\nu}{k-1} = 2\binom{n+1}{k+2}.$$

We thus conclude

$$1 - \frac{[z^n]S_t(z)}{[z^n]T(z)} \leq \frac{\ell(t)}{k!}\left(\frac{2(n+1)}{(k+2)(k+1)} + \frac{k}{n}\frac{k-1}{n-1}\cdots\frac{1}{n-k+1}\right)$$
$$\leq \frac{\ell(t)}{k!}\left(\frac{2(n+1)}{(k+2)(k+1)} + \frac{1}{n-k+1}\right).$$

Furthermore, for $|t| = n$, we have

$$1 - \frac{[z^n]S_t(z)}{[z^n]T(z)} = \frac{\ell(t)}{n!}.$$

Finally, we finish the proof like in Proposition 3 and get the stated result. □

Theorem 3. *Let X_n be the size of the compacted tree corresponding to a random binary increasing tree of size n. Then*

$$\mathbb{E}(X_n) \underset{n\to\infty}{=} \mathcal{O}\left(\frac{n}{\ln n}\right).$$

Proof. The result follows directly by combining the previous propositions. □

Recall that this result has already been shown in [16, 11], even with Θ instead of big-$\mathcal{O}$. Other proofs were presented as well, see [2, 3].

To get a crude lower bound for the number of non-isomorphic subtree shapes in a random increasing binary tree, we may proceed as in the case of recursive trees. Indeed, the uniform asymptotics given in Proposition 5 enable us to derive a lower bound for for the dominant singularity: $\tilde{\rho}_t > 1 + w(t)/k^2$ (*cf.* Corollary 3). With this bound we can perform all the steps of the proof of Proposition 4 and get the lower bound $\Omega(\sqrt{n})$.

Likewise, if the analog of Conjecture 1 for plane binary increasing trees were true, then we would be able to redo the proof of Theorem 2 to obtain the lower bound $\Omega(n/\ln n)$. Unfortunately, even the better knowledge on $w(t)$ given by Lemma 3 is not sufficient to show the analog of Conjecture 1.

4. A compressed data structure

The probability model induced by plane increasing binary trees is the classical permutation model of *binary search trees* (or BST). Thus the typical shape of a uniformly sampled plane increasing binary tree consisting of n internal nodes corresponds to the typical shape of a binary search tree built using a uniform random permutation of n elements. See Drmota [13, Section 1.3.3] for details about the latter correspondence. Thus the tree structure of a typical BST has the properties we have found out in the previous section. In particular, by removing the information stored in the nodes the typical compaction of the tree gives a compacted structure consisting of $\mathcal{O}(n/\ln n)$ nodes (on average).

Throughout this section, we aim at designing a new data structure based on the tree structure induced by the compaction of a BST associated to some extra information in the nodes and the edges in order to keep all the information (the integer values) from the original BST. And of course we must be able to retrieve information efficiently, as in BSTs. Our approach is supported with a python prototype and the experiments are obtained through this implementation.

The BST built for example on the permutation $(4, 8, 6, 2, 9, 1, 3, 7, 5)$ is represented with the classical tree structure in the left-hand side of Figure 7. This example will be used as an illustration throughout the whole section. In order to compress the tree structure, first the node labels must be removed, as presented before. Thus by using a compaction through a postorder traversal of the tree, the example becomes the tree structure presented in the right-hand side Figure 7. By adding the values stored in the original BST we get the tree of Figure 8. When a substructure has been removed through the compaction process, then in addition to the red pointer, the list of the labels, obtained through a *preorder traversal* of the substructure is stored. The latter, associated to the size of the substructures, depicted with the circled blue values, allows efficient searching. Let us present an example. We would like to know if 7 is stored in the structure. 7 is larger than 4, thus from the root we take the right edge to reach 8. The value we are looking for is smaller than 8. We take the left black pointer, and take also in consideration the list $L := [6, 5, 7]$. We

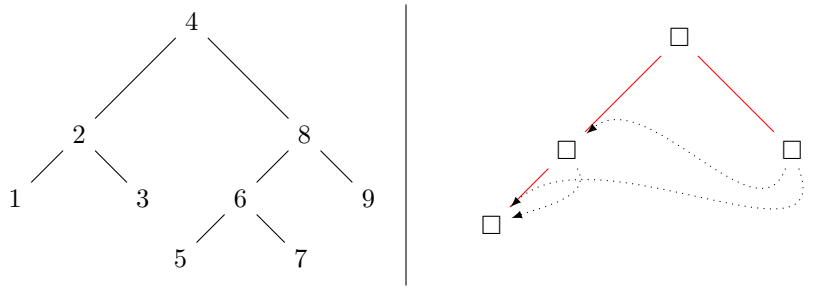

FIGURE 7. (left) A BST built e.g. on $(4, 8, 6, 2, 9, 1, 3, 7, 5)$; (right) The compacted tree structure associated to the BST

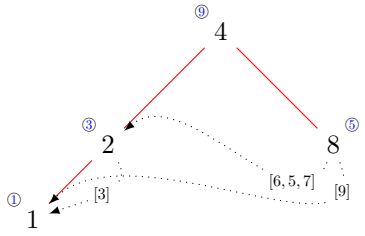

FIGURE 8. Labeled compacted structure associated to the original BST

define an index $i = 0$ corresponding to the actual index in the list we are interested in. Using the pointer, we reach 2 that corresponds in fact to $L[0] = 6$. Since 7 is larger than 6, we must follow the right child of 2, thus the new index is $i := i + 2$ (the list stores the values obtained through the preorder traversal), the constant 2 is the size of the left subtree attached to 2 plus 1 for the node labeled by 2. Now $L[2] = 7$, we have reached the value we were interested in.

Proposition 8. *In the compacted* BST *containing n values, the search complexity is the same as in the* BST *with respect to the number of value comparisons. There may be an extra-cost corresponding to the number of additions (related to the index) to traverse a list. The number of additions is at most equal to the number of comparisons to search for the value.*

Proof. The number of value comparisons is exactly the same in the compacted structure as in the original BST. In fact, we just share the identical unlabeled tree structure, thus the number of comparisons does not change. For the same reason, if we must search inside a list associated to a black pointer, then, for each comparison there is one addition to shift inside the list. □

In the following Figure 9 we have represented two experiments through our python prototype. On the left-hand side we are interested in the compaction ratio between the compressed data structure and the original BST. Here we are interested in the whole size needed in memory. In particular the size of the integer values is counted but further the data structure size itself is important. It is this latter that is in fact compressed: in the BST many pointers are needed to reach the nodes of the tree. Many pointers and nodes are replaced in the compressed data by lists of integers that need much less memory in practice. In the figure, in the abscissa we represent the number of integers stored in the data structures; and in the ordinate, we compute the ratio between the size in memory of the compressed data structure in front of the size of its corresponding BST. Each dot corresponds to one sample, and the green curve is the average value among all samples. The experiments are starting with 250 integer values up to 20,000 with steps every 250 values, and for each size we have used 30 uniformly sampled BSTs. We observe that even for small BSTs, the compression ratio is very interesting, smaller than 0.5. Further we remark that the green curve looks like the theoretical result: it is very close to a function $x \mapsto \alpha / \ln x$ for a given α.

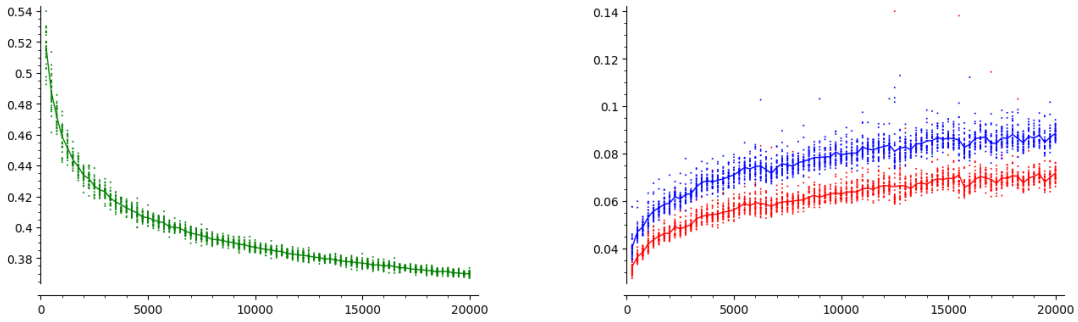

FIGURE 9. (left) Experimental compression ratio; (right) Experimental search time comparison

On the right-hand side of Figure 9, for the same set of BSTs and associated compacted structures, we search for 1,000 randomly sampled values present in the two structures. Each red dot is the average time, in milliseconds, (among the 1,000 searches) for finding the value inside the BST, and the blue point is the analogous time for the search in the compressed structure. For both complexity measures (number of comparisons or of arithmetic additions) the average complexity stays of the same order $O(\ln n)$ as for the original BST, as we see it in the figure. By computing the ratio of the blue values and the red values, the mean seems oscillating around 1.25 for the whole range of sampled structures.

Let us conclude this section with the following remark. The point of view we have chosen is to build first the BST and then, once the insertion and deletion process is done, we convert the BST into a compressed data structure that is used only for searching. We could develop a prototype data structure that manages insertion in deletion but the efficiency would probably be much less than the one of BST, because of the substructure recognition problem.

5. Conclusion

We showed for two exemplary families of increasing trees that the size of the compacted tree is smaller than for simply generated trees. This was done for recursive trees and plane increasing binary trees. Though the result for the latter family was already known (and even with better lower bound), we presented a new proof here and showed that our approach might work for more classes of increasing trees.

More precisely, we proved that the compacted tree belonging to a random recursive or increasing binary tree of size n is on average of size $\Omega(\sqrt{n})$ and $\mathcal{O}(n/\ln n)$. Numerical simulations on recursive trees suggest that this upper bound is already sharp, *i.e.*, that the size of the compacted tree is $\Theta(n/\ln)$. For the binary case that was already shown with other methods.

However, in order to prove this conjecture, one has to find the distribution of the weights $w(t)$, which turns out to be a very challenging task, especially in case of non-plane trees due to the appearance of automorphisms. However, we could formulate a simple to state condition under which we can prove the sharpness of the lower bound. Thus, obtaining the (maximum) number of labelings of non-plane trees of a given size is still work in progress, with the aim to improve the lower bounds such that we can show the Θ-result. Furthermore, we conjecture that on average the compacted tree is of size $\Theta\left(\frac{n}{\ln n}\right)$ for all classes of increasing trees.

We explain the choice of the two classes of increasing trees, that were investigated within this paper. The reason to choose recursive trees and increasing binary trees was that for these two classes our computer algebra system is able to solve the differential equation defining $S_t(z)$, although in case of increasing binary trees the solution is already more complicated and involves some Bessel functions. On the other hand, this makes it easier, as we could then deal with explicit expansions. However, in case of the third prominent class of increasing trees, PORTs (plane oriented

recursive trees), we did not get any explicit solution for the analogous of $S_t(z)$; thus this case is still an open question.

As a final note, remember the way we have compacted the BSTs in the last section. Using a pointer to describe the erased fringe subtree and the list of the labels in a specific traversal (labels that must be kept in the compacted tree), we are able to search in the compacted structure efficiently. But more generally, the way we have compacted the tree can be used for all possible tree structures. In the original paper [19] by Flajolet *et al.*, the authors compact only identical fringe subtrees in simply generated trees. We focus on the tree structure and its compaction as well, but the probability model on the tree shapes is a different one, induced by the labeling. Moreover, we use a different additional information management in order to cope with labels and could there extend the compaction to labeled tree models. It is desirable to study other natural labeled tree classes and the resulting compaction ratio.

Acknowledgments

The authors thank the anonymous referees for pointing out several references, but also for their comments and suggested improvements. In particular, we express our gratitude to one of the referees who pointed out a subtle error and several smaller ones. All these persistent remarks have greatly increased the quality of the paper.

References

[1] C. Bender and S. Orszag. *Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory*, volume 1. Springer, 1999.

[2] L. Seelbach Benkner and M. Lohrey. Average case analysis of leaf-centric binary tree sources. In *43rd International Symposium on Mathematical Foundations of Computer Science, MFCS*, volume 117 of *LIPIcs*, pages 16:1–16:15, 2018.

[3] L. Seelbach Benkner and S. G. Wagner. On the collection of fringe subtrees in random binary trees. In *Theoretical Informatics - 14th Latin American Symposium, (LATIN)*, volume 12118 of *Lecture Notes in Computer Science*, pages 546–558. Springer, 2020.

[4] F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In *CAAP '92 (Rennes, 1992)*, volume 581 of *Lecture Notes in Comput. Sci.*, pages 24–48. Springer, Berlin, 1992.

[5] O. Bodini and A. Genitrini. Cuts in increasing trees. In *12th SIAM Meeting on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 66–77, San Diego, USA, January 2015.

[6] O. Bodini, A. Genitrini, and F. Peschanski. A Quantitative Study of Pure Parallel Processes. *Electronic Journal of Combinatorics*, 23(1):P1.11, 39 pages, (electronic), 2016.

[7] M. Bousquet-Mélou, M. Lohrey, S. Maneth, and E. Noeth. XML compression via directed acyclic graphs. *Theory of Computing Systems*, 57(4):1322–1371, 2015.

[8] N. Broutin, L. Devroye, E. McLeish, and M. de la Salle. The height of increasing trees. *Random Struct. Algorithms*, 32(4):494–518, 2008.

[9] J. Cichoń, A. Magner, W. Szpankowski, and K. Turowski. *On Symmetries of Non-Plane Trees in a Non-Uniform Model*, pages 156–163. 2017.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd edition*. Wiley-Interscience, 2006.

[11] L. Devroye. On the richness of the collection of subtrees in random binary search trees. *Inf. Process. Lett.*, 65(4):195–199, 1998.

[12] M. Drmota. An analytic approach to the height of binary search trees II. *J. ACM*, 50(3):333–374, 2003.

[13] M. Drmota. *Random Trees*. Springer, Vienna-New York, 2009.

[14] M. Drmota, A. Iksanov, M. Moehle, and U. Roesler. A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Struct. Algorithms*, 34(3):319–336, 2009.

[15] J. Fill. On the distribution of binary search trees under the random permutation model. *Random Struct. Algorithms*, 8:1–25, 1996.

[16] P. Flajolet, X. Gourdon, and C. Martínez. Patterns in random binary search trees. *Random Structures Algorithms*, 11(3):223–244, 1997.

[17] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on discrete mathematics*, 3(2):216–240, 1990.

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

[19] P. Flajolet, P. Sipala, and J.-M. Steyaert. Analytic variations on the common subexpression problem. In *Automata, languages and programming (Coventry, 1990)*, volume 443 of *Lecture Notes in Comput. Sci.*, pages 220–234. Springer, New York, 1990.

[20] Z. Gołębiewski, A. Magner, and W. Szpankowski. Entropy and optimal compression of some general plane trees. *ACM Transactions on Algorithms (TALG)*, 15(1):1–23, 2018.

[21] M. Gopaladesikan, S. Wagner, and M. D. Ward. On the asymptotic probability of forbidden motifs on the fringe of recursive trees. *Experimental Mathematics*, 25(3):237–245, 2016.
[22] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., 2nd edition, 1994.
[23] E. L. Ince. *Ordinary Differential Equations*. Dover Publications, New York, 1944.
[24] D. E. Knuth. *The Art of Computer Programming, volume 3: (2nd ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.
[25] M. Kuba and A. Panholzer. On the degree distribution of the nodes in increasing trees. *J. Comb. Theory, Ser. A*, 114(4):597–618, 2007.
[26] A. Magner, K. Turowski, and W. Szpankowski. Lossless compression of binary trees with correlated vertex names. *IEEE Transactions on Information Theory*, 64(9):6070–6080, 2018.
[27] H. M. Mahmoud and R. T. Smythe. A Survey of Recursive Trees. *Theo. Probability and Mathematical Statistics*, 51:1–37, 1995.
[28] A. Meir and J. W. Moon. On the altitude of nodes in random trees. *Canadian Journal of Mathematics*, 30(5):997–1015, 1978.
[29] P. D. Miller. *Applied Asymptotic Analysis*. Graduate studies in mathematics. American Mathematical Society, 2006.
[30] J. Moon. The distance between nodes in recursive trees. In *London Math. Soc. Lecture Note Ser.*, volume 13, pages 125–132, 1974.
[31] A. Panholzer and H. Prodinger. Level of nodes in increasing trees revisited. *Random Struct. Algorithms*, 31(2):203–226, 2007.
[32] D. Ralaivaosaona and S. Wagner. Repeated fringe subtrees in random rooted trees. In *2015 Proceedings of the Twelfth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 78–88. SIAM, Philadelphia, PA, 2015.
[33] J. Zhang, E.-H. Yang, and J. C. Kieffer. Redundancy analysis in lossless compression of a binary tree via its minimal dag representation. In *2013 IEEE International Symposium on Information Theory*, pages 1914–1918. IEEE, 2013.

Olivier Bodini and Mehdi Naima. Université Sorbonne Paris Nord, Laboratoire d'Informatique de Paris Nord, CNRS, UMR 7030, F-93430, Villetaneuse, France.
Email address: {Olivier.Bodini, Mehdi.Naima}@lipn.univ-paris13.fr

Antoine Genitrini. Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6 -LIP6- UMR 7606, F-75005 Paris, France.
Email address: Antoine.Genitrini@lip6.fr

Bernhard Gittenberger and I. Larcher. Department of Discrete Mathematics and Geometry, Technische Universität Wien, Wiedner Hauptstrasse 8-10/104, 1040 Wien, Austria.
Email address: {Gittenberger, Larcher}@dmg.tuwien.ac.at