Average case analysis of leaf-centric binary tree sources

Louisa Seelbach Benkner
Universität Siegen
Germany
seelbach@eti.uni-siegen.de

Markus Lohrey
Universität Siegen
Germany
lohrey@eti.uni-siegen.de

Abstract

We study the average size of the minimal directed acyclic graph (DAG) with respect to so-called leaf-centric binary tree sources as studied by Zhang, Yang, and Kieffer. A leaf-centric binary tree source induces for every $n \geq 2$ a probability distribution on all binary trees with n leaves. We generalize a result shown by Flajolet, Gourdon, Martinez and Devroye according to which the average size of the minimal DAG of a binary tree that is produced by the binary search tree model is $\Theta(n/\log n)$.

2012 ACM Subject Classification Mathematics of computing → Enumeration

Keywords and phrases Directed acylic graphs, average case analysis, tree compression

Digital Object Identifier 10.4230/LIPIcs.MFCS.2018.16

1 Introduction

One of the most important and widely used compression methods for trees is to represent a tree by its minimal *directed acyclic graph*, shortly referred to as minimal DAG. The minimal DAG of a tree t is obtained by keeping for each subtree s of t only one isomorphic copy of s to which all edges leading to roots of s-copies are redirected. DAGs found applications in numerous areas of computer science; let us mention compiler construction [1, Chapter 6.1 and 8.5], unification [14], XML compression and querying [5, 9], and symbolic model-checking (binary decision diagrams) [4]. Recently, in information theory the average size of the minimal DAG with respect to a probability distribution turned out to be the key in order to obtain tree compressors whose average redundancy converges to zero [10, 16].

In this paper, we consider the problem of deriving asymptotic estimates for the average size of the minimal DAG of a randomly chosen binary tree of size n. So far, this problem has been analyzed mainly for two particular distributions: In [8], Flajolet, Sipala and Steyaert proved that the average size of the minimal DAG with respect to the uniform distribution on all binary trees of size n is asymptotically equal to $c \cdot n/\sqrt{\ln n}$, where c is the constant $2\sqrt{\ln(4/\pi)}$. This result was extended to unranked and node-labelled trees in [3] (with a different constant c). An alternative proof to the result of Flajolet et al. was presented in [15] by Ralaivaosaona and Wagner. For the so-called binary search tree model, Flajolet, Gourdon and Martinez [7] and Devroye [6] proved that the average size of the minimal DAG becomes $\Theta(n/\log n)$. In the binary search tree model, a binary search tree of size n is built by inserting the keys $1, \dots, n$ according to a uniformly chosen random permutation on $1, \dots, n$.

43rd International Symposium on Mathematical Foundations of Computer Science (MFCS 2018).
Editors: Igor Potapov, Paul Spirakis, and James Worrell; Article No. 16; pp. 16:1–16:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A general concept to produce probability distributions on the set of binary trees of size n was introduced by Zhang, Yang, and Kieffer in [16] (see also [11]), where the authors extend the classical notion of an information source on finite sequences to so-called *structured binary tree sources*, or binary tree sources for short. This yields a general framework for studying the average size of a minimal DAG. Let $\mathcal{T}$ denote the set of all binary trees[1] and let $\mathcal{T}_n$ denote the set of binary trees with n leaves. A binary tree source is a tuple $(\mathcal{T}, (\mathcal{T}_n)_{n\in\mathbb{N}}, P)$, in which P is a mapping from the set of binary trees to the unit intervall $[0, 1]$, such that $\sum_{t\in\mathcal{T}_n} P(t) = 1$ for every $n \geq 1$. This is a very general definition that was further restricted by Zhang et al. in order to yield interesting results. More precisely, they considered so-called *leaf-centric binary tree sources*, which are induced by a mapping $\sigma : (\mathbb{N} \setminus \{0\}) \times (\mathbb{N} \setminus \{0\}) \to [0, 1]$ that satisfies $\sum_{i=1}^{n-1} \sigma(i, n-i) = 1$ for every $n \geq 2$. In other words, σ restricted to $S_n := \{(i, n-i) \colon 1 \leq i \leq n-1\}$ is a probability mass function for every $n \geq 2$. To randomly produce a tree with n leaves, one starts with a single root node labelled with n and randomly chooses a pair $(i, n-i)$ according to the distribution σ on S_n. Then, a left (resp., right) child labelled 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. This yields a function P_σ that restricts to a probability mass function on every set $\mathcal{T}_n$ for $n \geq 2$.

The binary search tree model is the leaf-centric binary tree source where σ corresponds to the uniform distribution on S_n for every $n \geq 2$. Moreover, also the uniform distribution on all trees with n leaves can be obtained from a leaf-centric binary tree source by choosing σ suitably, see Section 4. Another well-known leaf-centric binary tree source is the *digitial search tree model* [13], where the distribution on S_n is a binomial distribution.

Let $\mathcal{D}_t$ denote the minimal DAG of a binary tree t and let $|\mathcal{D}_t|$ denote the number of nodes of $\mathcal{D}_t$. The average size of the minimal DAG with respect to a leaf-centric binary tree source $(\mathcal{T}, (\mathcal{T}_n)_{n\in\mathbb{N}}, P_\sigma)$ is the mapping

$$\mathcal{D}_\sigma(n) := \sum_{t\in\mathcal{T}_n} P_\sigma(t)|\mathcal{D}_t|. \tag{1}$$

In this work, we generalize the results of [6, 7] on the average size of the minimal DAG with respect to the binary search tree model in several ways. For this, we consider three classes of leaf-centric binary tree sources, which are defined by the following three properties of the corresponding σ-mappings:

(i) There exists an integer $N \geq 2$ and a monotonically decreasing function $\psi : \mathbb{R} \to (0, 1]$ such that $\psi(n) \geq \frac{2}{n-1}$ and $\sigma^*(i, n-i) \leq \psi(n)$ for every $n \geq N$ and $1 \leq i \leq n-1$. Here, σ^* is defined by $\sigma^*(i, i) = \sigma(i, i)$ and $\sigma^*(i, j) = \sigma(i, j) + \sigma(j, i)$ for $i \neq j$.

(ii) There exists an integer $N \geq 2$ and a constant $0 < \rho < 1$, such that $\sigma(i, n-i) \leq \rho$ for every $n \geq N$ and $1 \leq i \leq n-1$.

(iii) There is a monotonically decreasing function $\phi : \mathbb{N} \to (0, 1]$ and a constant $c \geq 3$ such that for every $n \geq 2$,

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

Property (iii) generalizes the concept of *balanced* binary tree sources from [10, 11]: When randomly constructing a binary tree with respect to a leaf-centric source of type (iii), the probability that the current weight is roughly equally splitted among the two children is

[1] We consider binary trees, where every non-leaf node has a left and a right child, but the whole framework can be easily extended to binary trees, where a node may have only a left or right child.

lower bounded by a function. Therefore, for slowly decreasing functions ϕ, balanced trees are preferred by this model. The binary search tree model satisfies all three conditions (i), (ii) and (iii). As our main results, we obtain for each of these three types of leaf-centric binary tree sources asymptotic bounds for the average size of the minimal DAG:

(a) For leaf-centric sources of type (i), the average size of the minimal DAG is upper bounded by $\mathcal{O}\left(\psi\left(\frac{1}{2}\log_4(n)\right)n\right)$, which is in $o(n)$ if $\psi(x) \in o(1)$.

(b) Using a simple entropy argument based on a result from [11], we show that for every leaf-centric binary tree source of type (ii), the average size of the minimal DAG is lower bounded by $\Omega(n/\log n)$.

(c) For leaf-centric binary tree sources of type (iii), the average size of the minimal DAG is upper bounded by $\mathcal{O}\big(\frac{n}{\phi(n)\log n}\big)$, which is in $o(n)$ if $\phi(n) \in \omega(1/\log n)$.

Both (a) and (c) imply the upper bound $\mathcal{O}(n/\log n)$ for the binary search tree model [7], whereas (b) yields an information-theoretic proof of the lower bound $\Omega(n/\log n)$ from [6].

The upper bounds (a) and (c) can be applied to the problem of universal tree compression [10, 16]. It is shown in [16] that a suitable binary encoding of the DAG yields a tree encoding whose average-case redundancy converges to zero assuming the trees are produced by a leaf-centric tree source for which the average DAG size is $o(n)$. See [16] for precise definitions.

2 Preliminaries

We use the classical Landau notations $\mathcal{O}$, o, Ω and ω. Quite often, we write sums of the form $\sum_{q_0 \le k \le q_1} a_k$ for rational numbers q_0, q_1. With this, we mean the sum $\sum_{k=\lceil q_0 \rceil}^{\lfloor q_1 \rfloor} a_k$. In the following, $\log x$ will always denote the binary logarithm $\log_2 x$ of a positive real number x. With $[0,1]$ we denote the unit interval of reals, and $(0,1] = [0,1] \setminus \{0\}$.

2.1 Trees and DAGs

We define binary trees as terms over the two symbols a (for leaves) and f (for binary nodes). The set $\mathcal{T}$ of *binary trees* is the smallest set of terms in f and a such that (i) $a \in \mathcal{T}$, and (ii) if $t_1, t_2 \in \mathcal{T}$, then $f(t_1, t_2) \in \mathcal{T}$. Thus, if we consider elements in $\mathcal{T}$ as graphs in the usual way, a binary tree is an ordered, rooted tree such that each node has either exactly two or no children. With $\mathcal{T}_n$ we denote the set of binary trees which have exactly n leaves. The *size* of a binary tree t is the number of leaves of t and denoted with $|t|$. A *fringe subtree* of a binary tree t is a subtree which consists of a node of t 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. The *leaf-size* of a node v of t is the size of the subtree $t[v]$. For a binary tree $t \in \mathcal{T}$ and an integer $k \ge 1$, let $N(t,k)$ denote the number of nodes of t of leaf-size greater than k.

For a binary tree $t \in \mathcal{T}$, let $\mathcal{D}_t$ denote its minimal *directed acyclic graph*, often shortly referred to as its minimal DAG. It is obtained by merging nodes u and v if $t[u]$ and $t[v]$ are isomorphic. The size $|\mathcal{D}_t|$ of $\mathcal{D}_t$ is the number of different pairwise non-isomorphic fringe subtrees of t. An example of a binary tree and its minimal DAG can be found in Figure 1.

2.2 Leaf-centric binary tree sources

In this paper we are interested in the average size of minimal DAGs. For this, we need for every $n \ge 1$ a probability distribution on $\mathcal{T}_n$. We restrict here to so-called leaf-centric binary tree sources that were studied in [11, 16]. Let Σ denote the set of all functions

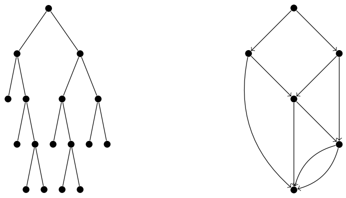

Figure 1 A binary tree (left) and its minimal DAG (right).

$\sigma : (\mathbb{N} \setminus \{0\}) \times (\mathbb{N} \setminus \{0\}) \to [0, 1]$ which satisfy

$$\sum_{i=1}^{n-1} \sigma(i, n-i) = 1$$

for every integer $n \geq 2$. We define $P_\sigma : \mathcal{T} \to [0, 1]$ inductively by $P_\sigma(a) = 1$ and $P_\sigma(f(u, v)) = \sigma(|u|, |v|) \cdot P_\sigma(u) \cdot P_\sigma(v)$. For every $n \geq 1$, P_σ restricts to a probability mass function on $\mathcal{T}_n$. The tuple $(\mathcal{T}, (\mathcal{T}_n)_{n\in\mathbb{N}}, P_\sigma)$ is called a *leaf-centric binary tree source.*

For an element $\sigma \in \Sigma$ define the mapping $\sigma^* : (\mathbb{N} \setminus \{0\}) \times (\mathbb{N} \setminus \{0\}) \to [0, 1]$ by

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

Note that $\sigma^*(i, j) \leq 1$ for all i, j and that $\sum_{k=1}^{\lfloor n/2 \rfloor} \sigma^*(k, n-k) = 1$.

3 Average size of the minimal DAG

Consider $\sigma \in \Sigma$. The *average size of the minimal DAG* with respect to the leaf-centric binary tree source $(\mathcal{T}, (\mathcal{T}_n)_{n\in\mathbb{N}}, P_\sigma)$ is the function $\mathcal{D}_\sigma : \mathbb{N} \to \mathbb{R}$ defined by equation (1). In the following, we present three natural classes of leaf-centric binary tree sources and investigate the average size of the minimal DAG with respect to these leaf-centric binary tree sources. In particular, we present conditions on $\sigma \in \Sigma$ that imply $\mathcal{D}_\sigma(n) \in o(n)$. In order to estimate $\mathcal{D}_\sigma$, we use the so-called cut-point argument that was applied in several papers [6, 15].

For a mapping $\sigma \in \Sigma$ and integers $b \geq 1$ and $n \geq 1$, let $E_{\sigma,b}(n)$ denote the expected value of $N(t, b)$ with respect to the probability mass function P_σ on the set of binary trees $\mathcal{T}_n$:

$$E_{\sigma,b}(n) = \sum_{t\in\mathcal{T}_n} P_\sigma(t) \cdot N(t, b).$$

Clearly, $E_{\sigma,b}(n) = 0$ if $n \leq b$. The following lemma constitutes the crucial argument we need in order to estimate the average size of a minimal DAG.

▶ **Lemma 1.** *Let $\sigma \in \Sigma$ and let $n \geq b \geq 1$. Then $\mathcal{D}_\sigma(n) \leq E_{\sigma,b}(n) + 4^b/3$.*

Proof. Let $t \in \mathcal{T}_n$. The size of the minimal DAG $\mathcal{D}_t$ of t is upper bounded by

(i) the number $N(t, b)$ of nodes of t of leaf-size greater than b plus

(ii) the number of binary trees with at most b leaves.

Recall that the number of binary trees with k leaves is the $(k-1)^{\text{th}}$ Catalan number C_{k-1}, which is bounded by 4^{k-1}. Hence, the number in (ii) is upper bounded by $\sum_{k=1}^{b} 4^{k-1} \leq 4^b/3$. This proves the lemma. ◀

The integer $b \geq 1$ from Lemma 1 is called the cutpoint. In order to apply Lemma 1 to estimate $\mathcal{D}_\sigma$, we first have to obtain estimates for $E_{\sigma,b}(n)$. This will be done inductively: Let $t = f(u,v) \in \mathcal{T}_n$ and let $b < n$. The number of nodes of t of leaf-size greater than b is composed of the number of nodes of the left subtree u of leaf-size greater than b plus the number of nodes of the right subtree v of leaf-size greater than b plus one (for the root), i.e., $N(t,b) = N(u,b) + N(v,b) + 1$. This observation easily yields the following recurrence relation for the expected value $E_{\sigma,b}(n)$:

$$E_{\sigma,b}(n) \;=\; 1 + \sum_{k=b+1}^{n-1} (\sigma(k, n-k) + \sigma(n-k,k)) \cdot E_{\sigma,b}(k).$$

With our definition of σ^*, this is equivalent to

$$E_{\sigma,b}(n) = \; 1 + \sum_{k=b+1}^{n-1} \sigma^*(k, n-k) \cdot E_{\sigma,b}(k) \tag{2}$$

if $b+1 > \frac{n}{2}$ and

$$E_{\sigma,b}(n) = 1 + \sum_{k=b+1}^{n-b-1} \sigma(k,n-k)(E_{\sigma,b}(k) + E_{\sigma,b}(n-k)) + \sum_{k=n-b}^{n-1} \sigma^*(k,n-k)E_{\sigma,b}(k) \tag{3}$$

if $b+1 \leq \frac{n}{2}$.

3.1 Average size of the minimal DAG for bounded σ-functions

First, we consider leaf-centric binary tree sources $(\mathcal{T}, (\mathcal{T}_n)_{n\in\mathbb{N}}, P_\sigma)$, where the function values of σ (or σ^*) are upper bounded by a function. We will prove an upper as well as a lower bound on the average DAG size.

3.1.1 Upper bound on the average DAG size

▶ **Definition 2** (the class Σ_*^ψ)**.** For a monotonically decreasing function $\psi : \mathbb{R} \to (0,1]$ such that $\psi(x) \geq 2/(x-1)$ for all large enough $x > 1$, let $\Sigma_*^\psi \subseteq \Sigma$ denote the set of mappings $\sigma \in \Sigma$ such that $\sigma^*(k, n-k) \leq \psi(n)$ for all large enough $n \geq 2$ and all $1 \leq k \leq n-1$.

The restriction $\psi(x) \geq 2/(x-1)$ is quite natural, at least for odd $x \in \mathbb{N}$, because $\sum_{k=1}^{n-1} \sigma^*(k, n-k) = 2$ if n is odd.

As our first main theorem, we prove an upper bound on $\mathcal{D}_\sigma(n)$ for every $\sigma \in \Sigma_*^\psi$:

▶ **Theorem 3.** *For every $\sigma \in \Sigma_*^\psi$, we have $\mathcal{D}_\sigma(n) \in \mathcal{O}\left(\psi\left(\frac{1}{2}\log_4(n)\right) \cdot n\right)$.*

Note that Theorem 3 only makes a nontrivial statement if ψ converges to zero: if ψ is lower bounded by a nonzero constant then we only obtain the trivial bound $\mathcal{D}_\sigma(n) \in \mathcal{O}(n)$. Moreover, the bound $\mathcal{D}_\sigma(n) \in \mathcal{O}\left(\psi\left(\frac{1}{2}\log_4(n)\right) \cdot n\right)$ also holds if we require that $\sigma(k, n-k) \leq \psi(n)$ for all large enough n and $1 \leq k \leq n-1$, since the latter implies that $\sigma^*(k, n-k) \leq 2\psi(n)$.

Let us fix a monotonically decreasing function $\psi : \mathbb{R} \to (0,1]$ such that $\psi(n) \geq 2/(n-1)$ for all large enough n. Moreover, let $\sigma \in \Sigma_*^\psi$. We can choose a constant N_σ such that $\psi(n) \geq 2/(n-1)$ and $\sigma^*(k, n-k) \leq \psi(n)$ for all $n \geq N_\sigma$ and all $1 \leq k \leq n-1$. In order to prove Theorem 3, we use the cut-point argument from Lemma 1. Thus, we start with an upper bound for $E_{\sigma,b}(n)$. A similar statement for the special case of the binary search tree model was shown by Knuth [12, p. 121].

▶ **Lemma 4.** *For all n, b with $n \geq b+1 > N_\sigma$ we have $E_{\sigma,b}(n) \leq 4n\psi(b) - 2$.*

In the proof of Lemma 4, we make use of the following lemma from linear optimization:

▶ **Lemma 5.** *Let $a_0 \leq a_1 \leq \cdots \leq a_{n-1}$ be a finite sequence of monotonically increasing positive real numbers and let $0 \leq c, \omega \leq 1$ and $l := \lfloor \omega/c \rfloor$. Moreover, let $x_0, \dots, x_{n-1}$ denote real numbers satisfying $0 \leq x_i \leq c$ for every $0 \leq i \leq n-1$ and $\sum_{k=0}^{n-1} x_k = \omega$. Then*

$$\sum_{i=0}^{n-1} a_i x_i \leq c \sum_{i=n-l}^{n-1} a_i + (\omega - lc) a_{n-l-1}. \tag{4}$$

Proof. Since $0 \leq a_0 \leq a_1 \leq \cdots \leq a_{n-1}$ and $0 \leq x_i \leq c$, the sum $\sum_{i=0}^{n-1} a_i x_i$ is maximized if we choose the maximal weight c for the l largest values $a_{n-l} \leq \cdots \leq a_{n-1}$ (i.e., $x_{n-l} = \cdots = x_{n-1} = c$), and put the remaining weight $\omega - lc$ (note that $\omega/c - 1 \leq l \leq \omega/c$, which implies $0 \leq \omega - lc \leq c$) on the $(l-1)$-th largest value a_{n-l-1} (i.e., $x_{n-l-1} = \omega - l \cdot c$). The remaining $x_1, \dots, x_{n-l-2}$ are set to zero. Then $\sum_{i=0}^{n-1} a_i x_i$ becomes the right-hand side of (4). ◀

Proof of Lemma 4. We prove the statement inductively in $n \geq b+1 > N_\sigma$. For the base case, let $n = b+1$. We have $E_{\sigma,b}(b+1) = 1 \leq 4(b+1)\psi(b) - 2$, as $\psi(b) \geq \frac{2}{b-1}$ by assumption.

For the induction step take an $n > b+1 > N_\sigma$ such that $E_{\sigma,b}(k) \leq 4k\psi(b) - 2$ for every $b < k \leq n-1$. By assumption, we have $n - 1 \geq n - \frac{1}{\psi(n)} > \frac{n}{2}$, as $n > N_\sigma$. We distinguish three subcases:

Case 1: $\frac{n}{2} < b+1 < n - \frac{1}{\psi(n)}$. By equation (2) and the induction hypothesis, we have

$$E_{\sigma,b}(n) \leq 1 + \sum_{k=b+1}^{n-1} \sigma^*(k, n-k)\,(4k\psi(b) - 2)\,. \tag{5}$$

Note that $\frac{n}{2} < b+1$ implies that $\sum_{k=b+1}^{n-1} \sigma^*(k, n-k) \leq 1$. Without loss of generality, we can assume that $\sum_{k=b+1}^{n-1} \sigma^*(k, n-k) = 1$: since $4k\psi(b) - 2 > 0$ for every k with $b+1 \leq k \leq n-1$, this makes the right-hand side in (5) only larger. Let $l := \left\lfloor \frac{1}{\psi(n)} \right\rfloor$ and $\delta := \frac{1}{\psi(n)} - l$. Applying Lemma 5 (with $a_k = 4k\psi(b) - 2$, $x_k = \sigma^*(k, n-k)$, $c = \psi(n)$ and $\omega = 1$), we get

$$E_{\sigma,b}(n) \leq 1 + \psi(n) \sum_{k=n-l}^{n-1} (4k\psi(b) - 2) + (1 - l\psi(n))\,(4(n-l-1)\psi(b) - 2)\,. \tag{6}$$

By simplifying the right hand side and using $0 \leq \delta < 1$ and $\psi(n) \leq \psi(b)$, we get

$$\begin{aligned}
E_{\sigma,b}(n) &\leq 4n\psi(b) - 1 - 4l\psi(b) - 4\psi(b) + 2l^2\psi(n)\psi(b) + 2l\psi(n)\psi(b) \\
&= 4n\psi(b) - 1 - \frac{2\psi(b)}{\psi(n)} - 2\psi(b) - 2\delta\psi(n)\psi(b) + 2\delta^2\psi(n)\psi(b) \\
&\leq 4n\psi(b) - 1 - \frac{2\psi(b)}{\psi(n)} - 2\psi(b) \leq 4n\psi(b) - 2.
\end{aligned}$$

Case 2: $b+1 \geq n - \frac{1}{\psi(n)}$. By equation (2) and by the induction hypothesis, we get

$$E_{\sigma,b}(n) \leq 1 + \sum_{k=b+1}^{n-1} \sigma^*(k, n-k)\,(4k\psi(b) - 2)\,.$$

Again, let $l := \left\lfloor \frac{1}{\psi(n)} \right\rfloor$ and $\delta := \frac{1}{\psi(n)} - l$. Since $b+1 \geq n - \frac{1}{\psi(n)}$ by assumption and $b+1 \in \mathbb{N}$ we have $b+1 \geq n-l$. Moreover, $n - \frac{1}{\psi(n)} > \frac{n}{2}$ implies $n - l > \frac{n}{2}$. Since $n-l$ is an integer, we get $n-l-1 \geq \frac{n-1}{2}$. This implies

$$4(n-l-1)\psi(b) - 2 \geq 2(n-1)\psi(b) - 2 \geq 2(n-1)\psi(n) - 2 \geq 2(n-1)\frac{2}{n-1} - 2 > 0$$

and hence also $4k\psi(b) - 2 > 0$ for all $n-l-1 \leq k \leq n-1$. As $\sigma \in \Sigma_*^\psi$, we have $\sigma^*(k, n-k) \leq \psi(n)$ for every $1 \leq k \leq n-1$. We get

$$E_{\sigma,b}(n) \leq 1 + \psi(n) \sum_{k=n-l}^{n-1} (4k\psi(b) - 2)\,.$$

Moreover, we have $1 - \psi(n)l \geq 0$ and $4(n-l-1)\psi(b) - 2 \geq 0$ and thus

$$E_{\sigma,b}(n) \leq 1 + \psi(n) \sum_{k=n-l}^{n-1} (4k\psi(b) - 2) + (1 - \psi(n)l)\,(4(n-l-1)\psi(b) - 2)\,.$$

This is equation (6) from Case 1. The statement follows now as in Case 1.

Case 3: $b+1 \leq \frac{n}{2}$. By equation (3) and the induction hypothesis, we have

$$\begin{aligned} E_{\sigma,b}(n) &= 1 + \sum_{k=b+1}^{n-b-1} \sigma(k, n-k)\,(E_{\sigma,b}(k) + E_{\sigma,b}(n-k)) + \sum_{k=n-b}^{n-1} \sigma^*(k, n-k) E_{\sigma,b}(k) \\ &\leq 1 + (4n\psi(b) - 4) \sum_{k=b+1}^{n-b-1} \sigma(k, n-k) + \sum_{k=n-b}^{n-1} \sigma^*(k, n-k)\,(4k\psi(b) - 2)\,. \end{aligned}$$

We set $\alpha := \sum_{k=b+1}^{n-b-1} \sigma(k, n-k)$. Hence, we have $\sum_{k=n-b}^{n-1} \sigma^*(k, n-k) = 1 - \alpha$. Set $l := \lfloor \frac{1-\alpha}{\psi(n)} \rfloor$. Note that $l \leq \frac{1}{\psi(n)} \leq \frac{n-1}{2}$ and that $4k\psi(b) - 2 \geq 0$ for all $n-b \leq k \leq n-1$ since $n-b > \frac{n}{2}$ and $\psi(b) \geq \psi(n) > \frac{2}{n}$. We distinguish two subcases:

Case 3.1: $b > l$ and thus, $n-b < n-l$. Applying Lemma 5 (with $a_k = 4k\psi(b) - 2$ and $x_k = \sigma^*(k, n-k)$ for $n-b \leq k \leq n-1$ and $c = \psi(n)$, $\omega = 1-\alpha$) yields

$$\begin{aligned} E_{\sigma,b}(n) &\leq 1 + (4n\psi(b) - 4)\,\alpha + \psi(n) \sum_{k=n-l}^{n-1} (4k\psi(b) - 2) \\ &\quad + (1 - \alpha - l\psi(n))\,(4(n-l-1)\psi(b) - 2)\,. \end{aligned} \tag{7}$$

Simplifying the right-hand side yields

$$E_{\sigma,b}(n) \leq 4n\psi(b) - 2\alpha - 1 + 2l\psi(n)\psi(b) + 2l^2\psi(n)\psi(b) - 4(1-\alpha)\psi(b) - 4(1-\alpha)l\psi(b).$$

Setting $\delta := \frac{(1-\alpha)}{\psi(n)} - l$, we get

$$E_{\sigma,b}(n) \leq 4n\psi(b) - 2\alpha - 1 - 2(1-\alpha)\psi(b) - \frac{2\psi(b)(1-\alpha)^2}{\psi(n)} - 2\delta\psi(n)\psi(b) + 2\delta^2\psi(n)\psi(b).$$

As $0 \leq \delta < 1$ and $\psi(n) \leq \psi(b)$, we have

$$\begin{aligned} E_{\sigma,b}(n) &\leq 4n\psi(b) - 2\alpha - 1 - 2(1-\alpha)\psi(b) - \frac{2\psi(b)(1-\alpha)^2}{\psi(n)} \\ &\leq 4n\psi(b) - 2\alpha - 1 - 2(1-\alpha)\psi(b) - 2(1-\alpha)^2. \end{aligned}$$

With $-2\alpha - 2(1-\alpha)^2 \leq -1$ for every value $0 \leq \alpha \leq 1$, the statement follows.

Case 3.2: $b \leq l$ and thus $n-b \geq n-l$. Since $n-l-1 \geq n - \frac{n-1}{2} - 1 = \frac{n-1}{2}$ and $\psi(b) \geq \psi(n) \geq \frac{2}{n-1}$ we have $4(n-l-1)\psi(b) - 2 \geq 0$. Thus, we also have $4k\psi(b) - 2 \geq 0$ for every $n-l \leq k \leq n-1$. Moreover, as $\sigma^*(k, n-k) \leq \psi(n)$, we get

$$E_{\sigma,b}(n) \leq 1 + (4n\psi(b) - 4)\,\alpha + \psi(n) \sum_{k=n-l}^{n-1} (4k\psi(b) - 2)\,.$$

Furthermore, as $1 - \alpha - l\psi(n) \geq 0$, we obtain

$$\begin{aligned} E_{\sigma,b}(n) \ \leq\ & 1 + (4n\psi(b) - 4)\,\alpha + \psi(n) \sum_{k=n-l}^{n-1} (4k\psi(b) - 2) \\ & + (1 - \alpha - l\psi(n))\,(4(n-l-1)\psi(b) - 2)\,. \end{aligned}$$

This is equation (7) from Case 3.1, and we can conclude as in Case 3.1. This finishes the proof of Lemma 4. ◀

With Lemma 4, we are able to prove Theorem 3 using the cut-point argument from Lemma 1:

Proof of Theorem 3. Let $\sigma \in \Sigma_*^\psi$, $n > 4^{2N_\sigma}$ and $N_\sigma \leq b < n$. By Lemma 1 and 4 we have $\mathcal{D}_\sigma(n) \leq E_{\sigma,b}(n) + 4^b/3 \leq 4n \cdot \psi(b) + 4^b/3$. Choose $b := \lceil \log_4(n)/2 \rceil$. As $n > 4^{2N_\sigma}$, this accords with $b \geq N_\sigma$. We obtain $\mathcal{D}_\sigma(n) \leq 4n\cdot\psi\left(\log_4(n)/2\right) + \Theta(\sqrt{n})$. Since $n \cdot \psi\left(\log_4(n)/2\right) \geq \frac{2n}{\log_4(n)/2 - 1}$ grows faster than $\Theta(\sqrt{n})$, this finishes the proof. ◀

In the following examples, we consider the results of Theorem 3 with respect to some concrete functions ψ:

▶ **Example 6.** Let $\sigma_{\text{bst}}(k, n-k) = \frac{1}{n-1}$ for every integer $1 \leq k \leq n-1$ and $n \geq 2$. The leaf-centric binary tree source $(\mathcal{T}, (\mathcal{T}_n)_{n\geq 1}, P_{\sigma_{\text{bst}}})$ corresponds to the well-known binary search tree model. Let $\psi(x) = \frac{2}{x-1}$ for every $x > 1$. We find $\sigma_{\text{bst}} \in \Sigma_*^\psi$. With Theorem 3, we have $\mathcal{D}_{\sigma_{\text{bst}}} \in \mathcal{O}(n/\log n)$, which accords with the results of [6]. ◀

▶ **Example 7.** There are plenty of other ways to choose ψ in Theorem 3. For example $\psi(x) \in \Theta(1/x^\alpha)$ with $0 \leq \alpha \leq 1$ yields $\mathcal{D}_\sigma(n) \in \mathcal{O}(n/\log(n)^\alpha)$ for every $\sigma \in \Sigma_*^\psi$. For $\psi(x) \in \Theta(1/\log x)$ we get $\mathcal{D}_\sigma(n) \in \mathcal{O}(n/\log\log n)$ for every $\sigma \in \Sigma_*^\psi$. ◀

3.1.2 Lower bound on the average DAG size

In this section we prove a lower bound for $\mathcal{D}_\sigma(n)$.

▶ **Definition 8** (the class Σ^ρ)**.** For a constant ρ with $0 < \rho < 1$ let Σ^ρ denote the set of mappings $\sigma \in \Sigma$ such that $\sigma(k, n-k) \leq \rho$ for all large enough n and all $1 \leq k \leq n-1$.

By Theorem 3, we only know $\mathcal{D}_\sigma(n) \in \mathcal{O}(n)$ for $\sigma \in \Sigma^\rho$. In the following theorem, we present a lower bound for $\mathcal{D}_\sigma(n)$ with respect to a mapping $\sigma \in \Sigma^\rho$:

▶ **Theorem 9.** *If $\sigma \in \Sigma^\rho$, then $\mathcal{D}_\sigma(n) \in \Omega(n/\log n)$.*

Let us fix a mapping $\sigma \in \Sigma^\rho$, where $0 < \rho < 1$, and let $N_\sigma \geq 2$ such that $\sigma(k, n-k) \leq \rho$ for all $n \geq N_\sigma$ and all $1 \leq k \leq n-1$. In order to prove Theorem 9, we make use of an information-theoretic argument. We need the following notations: For a mapping $\sigma \in \Sigma$, let X_σ^n denote the random variable taking values in $\mathcal{T}_n$ according to the probability mass function P_σ on $\mathcal{T}_n$. Moreover, let $H(X_\sigma^n)$ denote the Shannon entropy of X_σ^n, i.e.,

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

▶ **Lemma 10.** *If $\sigma \in \Sigma^\rho$, then $H(X_\sigma^n) \geq \log\left(\frac{1}{\rho}\right)\left(\frac{n}{4N_\sigma - 4}\right)$ for every $n \geq N_\sigma$.*

In order to prove Lemma 10, we need a lower bound for $E_{\sigma,b}(n)$:

▶ **Lemma 11.** *For a mapping $\sigma \in \Sigma$ and integers $n > b \geq 1$, we have $E_{\sigma,b}(n) \geq \frac{n}{4b}$.*

Proof. We prove the statement inductively in $n \geq b+1$: For the base case, let $n = b+1$. A binary tree $t \in \mathcal{T}_{b+1}$ has exactly one node of leaf-size greater than b, which is the root of t. Thus, $E_{\sigma,b}(b+1) = 1 \geq \frac{b+1}{4b}$ for every integer $b \geq 1$. For the induction hypothesis, take an integer $n > b+1$ such that $E_{\sigma,b}(k) \geq \frac{k}{4b}$ for every integer $b+1 \leq k \leq n-1$.

In the induction step, we distinguish two cases:

Case 1: $\frac{n}{2} < b+1 \leq n-1$: We thus have $\frac{n}{4b} \leq 1$. By equation (2), we have

$$E_{\sigma,b}(n) = 1 + \sum_{k=b+1}^{n-1} \sigma^*(k, n-k) E_{\sigma,b}(k) \geq 1 \geq \frac{n}{4b}.$$

Case 2: $b+1 \leq \frac{n}{2}$: Let $\alpha := \sum_{k=b+1}^{n-b-1} \sigma(k, n-k)$. From equation (3) and the induction hypothesis we get

$$\begin{aligned} E_{\sigma,b}(n) &= 1 + \sum_{k=b+1}^{n-b-1} \sigma(k,n-k)(E_{\sigma,b}(k) + E_{\sigma,b}(n-k)) + \sum_{k=n-b}^{n-1} \sigma^*(k,n-k) E_{\sigma,b}(k) \\ &\geq 1 + \sum_{k=b+1}^{n-b-1} \sigma(k,n-k)\frac{n}{4b} + \sum_{k=n-b}^{n-1} \sigma^*(k,n-k)\frac{k}{4b} \\ &\geq 1 + \frac{n}{4b}\sum_{k=b+1}^{n-b-1} \sigma(k,n-k) + \frac{n-b}{4b}\sum_{k=n-b}^{n-1} \sigma^*(k,n-k) \\ &= 1 + \alpha\frac{n}{4b} + (1-\alpha)\left(\frac{n-b}{4b}\right) = \frac{n}{4b} + \frac{3}{4} + \frac{\alpha}{4}. \end{aligned}$$

As $0 \leq \alpha \leq 1$, the statement follows. ◀

Proof of Lemma 10. Lemma 10 follows from identity (4) in [11]: Define

$$h_k(\sigma) := \sum_{\substack{i,j \geq 1 \\ i+j=k}} \sigma(i,j) \log\left(\frac{1}{\sigma(i,j)}\right),$$

that is, $h_k(\sigma)$ is the Shannon entropy of the random variable taking values in $\{(i, k-i) \colon 1 \leq i \leq k-1\}$ according to the probility mass function σ. As $\sigma(i,j) \leq \rho$ for $i+j \geq N_\sigma$, we find

$$h_k(\sigma) \geq \log\left(\frac{1}{\rho}\right) \sum_{\substack{i,j \geq 1 \\ i+j=k}} \sigma(i,j) = \log\left(\frac{1}{\rho}\right)$$

for every $k \geq N_\sigma$. Identity (4) in [11] states that $H(X_\sigma^n) = \sum_{j=2}^{n} (E_{\sigma,j-1}(n) - E_{\sigma,j}(n))\, h_j(\sigma)$. With $n \geq N_\sigma$, we obtain

$$\begin{aligned} H(X_\sigma^n) &\geq \sum_{j=N_\sigma}^{n} (E_{\sigma,j-1}(n) - E_{\sigma,j}(n))\, h_j(\sigma) \geq \log\left(\frac{1}{\rho}\right) \sum_{j=N_\sigma}^{n} (E_{\sigma,j-1}(n) - E_{\sigma,j}(n)) \\ &= \log\left(\frac{1}{\rho}\right) (E_{\sigma,N_\sigma-1}(n) - E_{\sigma,n}(n)) = \log\left(\frac{1}{\rho}\right) E_{\sigma,N_\sigma-1}(n). \end{aligned}$$

By Lemma 11, we have $H(X_\sigma^n) \geq \log\left(\frac{1}{\rho}\right)\left(\frac{n}{4N_\sigma-4}\right)$. This proves the statement. ◀

With Lemma 10, we are able to prove Theorem 9:

Proof of Theorem 9. We first show that a binary tree $t \in \mathcal{T}_n$ can be encoded with at most $2m\lceil\log(2n-1)\rceil$ bits, where $m = |\mathcal{D}_t| \leq 2n-1$ (note that t has exactly $2n-1$ nodes). It suffices to encode $\mathcal{D}_t$. W.l.o.g. assume that the nodes of $\mathcal{D}_t$ are the numbers $1, \dots, m$, where m is the unique leaf node of $\mathcal{D}_t$. For $1 \leq k \leq m-1$ let l_k (resp., r_k) be the left (resp., right) child of node k. We encode each number $1, \dots, m$ by a bit string of length exactly $\lceil\log(2n-1)\rceil$. The DAG $\mathcal{D}_t$ can be uniquely encoded by the bit string $l_1 r_1 l_2 r_2 \cdots l_{m-1} r_{m-1}$, which has length $2(m-1)\lceil\log(2n-1)\rceil$.

Let $\sigma \in \Sigma^\rho$. By Lemma 10, we know that $H(X_\sigma^n) \geq \log\left(\frac{1}{\rho}\right)\left(\frac{n}{4N_\sigma-4}\right)$ for every $n \geq N_\sigma$. Shannon's coding theorem implies

$$H(X_\sigma^n) \leq 2\lceil\log(2n-1)\rceil \sum_{t \in \mathcal{T}_n} P_\sigma(t)|\mathcal{D}_t| = 2\lceil\log(2n-1)\rceil \mathcal{D}_\sigma(n).$$

We get $\log(1/\rho)\left(\frac{n}{4N_\sigma-4}\right) \leq 2\lceil\log(2n-1)\rceil \mathcal{D}_\sigma(n)$ for all $n \geq 2$, which concludes the proof. ◀

3.2 Average size of the minimal DAG for weakly balanced tree sources

In this subsection, we present so-called *weakly balanced* binary tree sources, which represent a generalization of balanced binary tree sources introduced in [11] and further analysed in [10]. Let us fix a constant $c \geq 3$ for the rest of this subsection.

▶ **Definition 12** (the class Σ_ϕ)**.** For a monotonically decreasing function $\phi : \mathbb{N} \to (0,1]$ let $\Sigma_\phi \subseteq \Sigma$ denote the set of mappings σ such that for every $n \geq 2$,

$$\sum_{\frac{n}{c} \leq k \leq n - \frac{n}{c}} \sigma(k, n-k) \geq \phi(n).$$

We call a binary tree source $(\mathcal{T}, (\mathcal{T}_n)_{n\geq 1}, P_\sigma)$ with $\sigma \in \Sigma_\phi$ *weakly balanced.* We obtain the following upper bound for $\mathcal{D}_\sigma$ with respect to a weakly balanced tree source:

▶ **Theorem 13.** *For every $\sigma \in \Sigma_\phi$, we have $\mathcal{D}_\sigma(n) \in \mathcal{O}\left(\frac{n}{\phi(n)\log n}\right)$.*

Theorem 13 can be used to reprove the upper bound $\mathcal{D}_{\sigma_{\text{bst}}}(n) \in \mathcal{O}(n/\log n)$ for the binary search tree model from Example 6 (note that $\sum_{n/4 \leq k \leq 3n/4} \frac{1}{n-1} > \frac{1}{2}$). More generally, if $\phi(n) \in \omega(1/\log n)$, then Theorem 13 yields $\mathcal{D}_\sigma(n) \in o(n)$ for every $\sigma \in \Sigma_\phi$.

Analogously to Theorem 3, we show Theorem 13 using the cut-point argument from Lemma 1. The strategy in the proof of the following lemma is similar to Lemma 4.

▶ **Lemma 14.** *For every $\sigma \in \Sigma_\phi$ and all $b \geq 1$, $n \geq b+1$ we have $E_{\sigma,b}(n) \leq \frac{cn}{\phi(n)b} - \frac{1}{\phi(n)}$.*

Proof. We prove the statement inductively in $n \geq b+1$. For the base case, note that a binary tree $t \in \mathcal{T}_{b+1}$ has exactly one node of leaf-size $> b$, which is the root of t. Thus,

$$E_{\sigma,b}(b+1) = 1 \leq \frac{c(b+1)}{\phi(b+1)b} - \frac{1}{\phi(b+1)}.$$

Let us now deal with the induction step. Take an integer $n > b+1$ such that $E_{\sigma,b}(k) \leq \frac{ck}{\phi(k)b} - \frac{1}{\phi(k)}$ for every integer $b+1 \leq k \leq n-1$. We distinguish six cases:

Case 1: $c \geq n$ and thus $c > b$. We have $\frac{n}{c} \leq 1$ and $n-1 \leq n - \frac{n}{c}$. Case 1 splits into two subcases:

Case 1.1: $\frac{n}{2} < b+1 \leq n-1$: By equation (2), the induction hypothesis, and the fact that ϕ is monotonically decreasing, we get

$$\begin{aligned} E_{\sigma,b}(n) &= 1 + \sum_{k=b+1}^{n-1} \sigma^*(k,n-k) \left(\frac{ck}{\phi(k)b} - \frac{1}{\phi(k)}\right) \\ &\leq 1 + \left(\frac{c(n-1)}{\phi(n)b} - \frac{1}{\phi(n)}\right) \sum_{k=b+1}^{n-1} \sigma^*(k,n-k). \end{aligned}$$

As $b+1 > \frac{n}{2}$ and $\sigma \in \Sigma$, we have $\sum_{k=b+1}^{n-1} \sigma^*(k,n-k) \leq 1$ and thus

$$E_{\sigma,b}(n) \leq \frac{cn}{\phi(n)b} - \frac{c}{\phi(n)b} - \frac{1}{\phi(n)} + 1 \leq \frac{cn}{\phi(n)b} - \frac{1}{\phi(n)}.$$

Case 1.2: $b+1 \leq \frac{n}{2}$: Equation (3), the induction hypothesis, and the fact that ϕ is monotonically decreasing yield

$$\begin{aligned} E_{\sigma,b}(n) &= 1 + \sum_{k=b+1}^{n-b-1} \sigma(k,n-k)\left(E_{\sigma,b}(k) + E_{\sigma,b}(n-k)\right) + \sum_{k=n-b}^{n-1} \sigma^*(k,n-k) E_{\sigma,b}(k) \\ &\leq 1 + \sum_{k=b+1}^{n-b-1} \sigma(k,n-k) \left(\frac{cn}{\phi(n)b} - \frac{2}{\phi(n)}\right) + \sum_{k=n-b}^{n-1} \sigma^*(k,n-k)\left(\frac{ck}{\phi(k)b} - \frac{1}{\phi(k)}\right) \\ &\leq 1 + \left(\frac{cn}{\phi(n)b} - \frac{2}{\phi(n)}\right) \sum_{k=b+1}^{n-b-1} \sigma(k,n-k) \\ &\quad + \left(\frac{c(n-1)}{\phi(n)b} - \frac{1}{\phi(n)}\right) \sum_{k=n-b}^{n-1} \sigma^*(k,n-k). \end{aligned}$$

We set $\alpha := \sum_{k=b+1}^{n-b-1} \sigma(k,n-k)$ and get

$$\begin{aligned} E_{\sigma,b}(n) &\leq 1 + \left(\frac{cn}{\phi(n)b} - \frac{2}{\phi(n)}\right)\alpha + \left(\frac{c(n-1)}{\phi(n)b} - \frac{1}{\phi(n)}\right)(1-\alpha) \\ &= \frac{cn}{\phi(n)b} - \frac{c}{\phi(n)b} + 1 - \frac{1}{\phi(n)} + \alpha\left(\frac{c}{\phi(n)b} - \frac{1}{\phi(n)}\right). \end{aligned}$$

As $c > b$ by assumption, the last term is monotonically increasing in α. With $\alpha \leq 1$, we have

$$E_{\sigma,b}(n) \leq \frac{cn}{\phi(n)b} - \frac{2}{\phi(n)} + 1 \leq \frac{cn}{\phi(n)b} - \frac{1}{\phi(n)}.$$

Case 2: $n > c$. We have $\frac{n}{c} > 1$ and $n - \frac{n}{c} < n-1$. Case 2 splits into four subcases:

Case 2.1: $n - \frac{n}{c} < b + 1 \leq n - 1$: This case is very similar to Case 1.1 and left to the reader; see also the long version [2].

Case 2.2: $\frac{n}{2} < b + 1 \leq n - \frac{n}{c}$. Equation (2), the induction hypothesis and the monotonicity of ϕ yield

$$\begin{aligned}
E_{\sigma,b}(n) &= 1 + \sum_{b+1 \leq k \leq n - \frac{n}{c}} \sigma^*(k, n-k) E_{\sigma,b}(k) + \sum_{n - \frac{n}{c} < k \leq n-1} \sigma^*(k, n-k) E_{\sigma,b}(k) \\
&\leq 1 + \left(\frac{(c-1)n}{\phi(n)b} - \frac{1}{\phi(n)} \right) \sum_{b+1 \leq k \leq n - \frac{n}{c}} \sigma^*(k, n-k) \\
&\quad + \left(\frac{c(n-1)}{\phi(n)b} - \frac{1}{\phi(n)} \right) \sum_{n - \frac{n}{c} < k \leq n-1} \sigma^*(k, n-k).
\end{aligned}$$

We set $\alpha := \sum_{n - \frac{n}{c} < k \leq n-1} \sigma^*(k, n-k)$. Since $b+1 > \frac{n}{2}$ we have $\sum_{b+1 \leq k \leq n - \frac{n}{c}} \sigma^*(k, n-k) \leq 1 - \alpha$ and get

$$\begin{aligned}
E_{\sigma,b}(n) &\leq 1 + (1-\alpha) \left(\frac{(c-1)n}{\phi(n)b} - \frac{1}{\phi(n)} \right) + \alpha \left(\frac{c(n-1)}{\phi(n)b} - \frac{1}{\phi(n)} \right) \\
&= \frac{cn}{\phi(n)b} - \frac{n}{\phi(n)b} - \frac{1}{\phi(n)} + 1 + \alpha \frac{(n-c)}{\phi(n)b}.
\end{aligned}$$

As $n > c$ by assumption, the last term is monotonically increasing in α. With $\alpha \leq 1 - \phi(n)$ as $\sigma \in \Sigma_\phi$, we find

$$E_{\sigma,b}(n) \leq \frac{cn}{\phi(n)b} - \frac{1}{\phi(n)} + 1 - \frac{c}{\phi(n)b} + \frac{c}{b} - \frac{n}{b} \leq \frac{cn}{\phi(n)b} - \frac{1}{\phi(n)}.$$

Case 2.3: $\frac{n}{c} \leq b + 1 \leq \frac{n}{2}$. Equation (3), the induction hypothesis, and the monotonicity of ϕ yield

$$\begin{aligned}
E_{\sigma,b}(n) &= 1 + \sum_{k=b+1}^{n-b-1} \sigma(k, n-k)(E_{\sigma,b}(k) + E_{\sigma,b}(n-k)) \\
&\quad + \sum_{n-b \leq k \leq n - \frac{n}{c}} \sigma^*(k, n-k) E_{\sigma,b}(k) + \sum_{n - \frac{n}{c} < k \leq n-1} \sigma^*(k, n-k) E_{\sigma,b}(k) \\
&\leq 1 + \left(\frac{cn}{\phi(n)b} - \frac{2}{\phi(n)} \right) \sum_{k=b+1}^{n-b-1} \sigma(k, n-k) \\
&\quad + \left(\frac{(c-1)n}{\phi(n)b} - \frac{1}{\phi(n)} \right) \sum_{n-b \leq k \leq n - \frac{n}{c}} \sigma^*(k, n-k) \\
&\quad + \left(\frac{c(n-1)}{\phi(n)b} - \frac{1}{\phi(n)} \right) \sum_{n - \frac{n}{c} < k \leq n-1} \sigma^*(k, n-k).
\end{aligned}$$

With $\alpha := \sum_{k=b+1}^{n-b-1} \sigma(k, n-k)$ and $\beta := \sum_{n - \frac{n}{c} < k \leq n-1} \sigma^*(k, n-k)$ one can simplify this to

$$E_{\sigma,b}(n) = \frac{cn}{\phi(n)b} - \frac{n}{\phi(n)b} - \frac{1}{\phi(n)} + 1 + \alpha \left(\frac{n}{\phi(n)b} - \frac{1}{\phi(n)} \right) + \beta \left(\frac{n-c}{\phi(n)b} \right). \tag{8}$$

As $b < n$ and $c < n$ by assumption, the term in the last line is monotonically increasing in α and β. Using this fact, as well as $0 \leq \beta \leq 1 - \phi(n)$ (as $\sigma \in \Sigma_\phi$), $0 \leq \alpha \leq 1$ and $\alpha + \beta \leq 1$, one can show that

$$\alpha \left(\frac{n}{\phi(n)b} - \frac{1}{\phi(n)} \right) + \beta \left(\frac{n-c}{\phi(n)b} \right) \leq \frac{n}{\phi(n)b} - 1$$

(see the long version [2] for details). Plugging this into (8) yields

$$E_{\sigma,b}(n) \leq \frac{cn}{\phi(n)b} - \frac{n}{\phi(n)b} - \frac{1}{\phi(n)} + 1 + \frac{n}{\phi(n)b} - 1 = \frac{cn}{\phi(n)b} - \frac{1}{\phi(n)}.$$

Case 2.4: $b + 1 < \frac{n}{c}$. This case is very similar to Case 1.2 and left to the reader; see also the long version [2]. ◀

Proof of Theorem 13. Let $n \geq 2$ and let $1 \leq b < n$. By Lemma 1 and 14, we have

$$\mathcal{D}_\sigma(n) \leq E_{\sigma,b}(n) + \frac{4^b}{3} \leq \frac{cn}{\phi(n)b} + \frac{4^b}{3}.$$

Choosing $b := \left\lceil \frac{1}{2}\log_4(n) \right\rceil$, the statement follows. ◀

In the following corollary we identify a constant $\nu \in (0,1]$ with the function mapping every $n \in \mathbb{N}$ to ν. The corollary follows immediately from Theorem 13 and 9.

▶ **Corollary 15.** *For all* $0 < \nu, \rho < 1$ *and all* $\sigma \in \Sigma_\nu \cap \Sigma^\rho$ *we have* $\mathcal{D}_\sigma(n) \in \Theta(n/\log n)$.

▶ **Example 16.** In this example, we investigate the binomial random tree model, which was studied in [11] for the case $p = 1/2$, and which is a slight variant of the digital search tree model, see [13]. Let $0 < p < 1$ and define $\sigma_p \in \Sigma$ by

$$\sigma_p(k, n-k) = p^{k-1}(1-p)^{n-k-1}\binom{n-2}{k-1}$$

for every integer $n \geq 2$ and $1 \leq k \leq n-1$. We use the abbreviation $\pi(i) = \sigma_p(i, n-i)$ in the following. By the binomial theorem, we have $\sum_{k=1}^{n-1} \pi(k) = 1$. In the following, we will prove that $\mathcal{D}_{\sigma_p}(n) \in \mathcal{O}(n/\log n)$. We distinguish two cases.

Case 1: $0 < p \leq \frac{1}{2}$. Let $\nu := 1 - \frac{4-4p}{4+p}$. We find $\nu > 0$ for $0 < p \leq \frac{1}{2}$. We claim that with $c := \frac{6}{p}$, we have $\sigma_p \in \Sigma_\nu$. Then Theorem 13 yields $\mathcal{D}_{\sigma_p}(n) \in \mathcal{O}(n/\log n)$.

In order to prove $\sigma_p \in \Sigma_\nu$, we show

$$\sum_{\frac{np}{6} \leq i \leq n - \frac{np}{6}} \sigma_p(i, n-i) = \sum_{\frac{np}{6} \leq i \leq n - \frac{np}{6}} \pi(i) \geq 1 - \frac{4-4p}{4+p}.$$

Without loss of generality, let $n \geq 3$. Let X_p^n denote the random variable taking values in the set $\{1, \dots, n-1\}$ according to the probability mass function π. Thus, $X_p^n = 1 + Y_p^n$, where Y_p^n is binomially distributed with parameters $n-2$ and p. For the expected value and variance of X_p^n we obtain $\mathsf{E}[X_p^n] = p(n-2) + 1$ and $\mathsf{Var}[X_p^n] = p(1-p)(n-2)$. Let $\kappa := p(n-2)/2$ so that $\mathsf{E}[X_p^n] = 2\kappa + 1$ and $\mathsf{Var}[X_p^n] = 2\kappa(1-p)$. By Chebyshev's inequality, we have ($\mathsf{Prob}(A)$ denotes the probability of the event A)

$$\begin{aligned}\mathsf{Prob}\left(\left|X_p^n - \mathsf{E}[X_p^n]\right| < \kappa + 1\right) &\geq 1 - \frac{\mathsf{Var}[X_p^n]}{(\kappa+1)^2} = 1 - \frac{2\kappa(1-p)}{\kappa^2 + 2\kappa + 1} \geq 1 - \frac{2(1-p)}{\kappa + 2} \\ &= 1 - \frac{4(1-p)}{p(n-2)+4} \geq 1 - \frac{4(1-p)}{p+4}\end{aligned}$$

where the last inequality holds due to $n \geq 3$. Moreover, with $\mathsf{E}[X_p^n] = 2\kappa + 1$, we have

$$\mathsf{Prob}\left(\left|X_p^n - \mathsf{E}[X_p^n]\right| < \kappa + 1\right) = \sum_{\kappa < i < 3\kappa + 2} \pi(i).$$

As $n \geq 3$ and $0 < p \leq \frac{1}{2}$, we have $\kappa \geq \frac{pn}{6}$ and $3\kappa + 2 \leq n - \frac{pn}{6}$. Thus, we have

$$\sum_{\frac{pn}{6} \leq i \leq n - \frac{pn}{6}} \pi(i) \geq \sum_{\kappa < i < 3\kappa+2} \pi(i) = \mathsf{Prob}\left(|X_p^n - \mathsf{E}[X_p^n]| < \kappa + 1\right) \geq 1 - \frac{4(1-p)}{p+4}.$$

This finishes the proof of Case 1.

Case 2: $\frac{1}{2} < p < 1$. Define a mapping $\vartheta : \mathcal{T} \to \mathcal{T}$ inductively by $\vartheta(a) = a$ and $\vartheta(f(u,v)) = f(\vartheta(v), \vartheta(u))$. Intuitively, ϑ exchanges the right child node and the left child node of every node of a binary tree t. It is easy to see that $\vartheta : \mathcal{T}_n \to \mathcal{T}_n$ is a bijection for every $n \geq 1$ and that ϑ^2 is the identity mapping. Moreover, t and $\vartheta(t)$ have the same number of different pairwise non-isomorphic subtrees and thus, $|\mathcal{D}_t| = |\mathcal{D}_{\vartheta(t)}|$. We show inductively in $n \geq 1$, that $P_{\sigma_p}(\vartheta(t)) = P_{\sigma_{1-p}}(t)$ for a binary tree $t \in \mathcal{T}_n$: For the base case, let $t = a$. We find $P_{\sigma_p}(\vartheta(a)) = 1 = P_{\sigma_{1-p}}(a)$.

For the induction step, let $t = f(u,v) \in \mathcal{T}_n$. We have

$$\begin{aligned} P_{\sigma_p}(\vartheta(t)) = P_{\sigma_p}(f(\vartheta(v), \vartheta(u))) &= \sigma_p(|\vartheta(v)|, |\vartheta(u)|) P_{\sigma_p}(\vartheta(v)) P_{\sigma_p}(\vartheta(u)) \\ &= \sigma_p(|v|, |u|) P_{\sigma_{1-p}}(u) P_{\sigma_{1-p}}(v), \end{aligned}$$

where the last equality holds by the induction hypothesis. Moreover, with $|u| = n - |v|$ and by definition of σ_p, we find that $\sigma_p(|v|, |u|) = \sigma_{1-p}(|u|, |v|)$. Thus, we have

$$\sigma_p(|v|, |u|) P_{\sigma_{1-p}}(u) P_{\sigma_{1-p}}(v) = \sigma_{1-p}(|u|, |v|) P_{\sigma_{1-p}}(u) P_{\sigma_{1-p}}(v) = P_{\sigma_{1-p}}(t).$$

This finishes the induction. Altogether, and as $\vartheta : \mathcal{T}_n \to \mathcal{T}_n$ is a bijection, we get

$$\mathcal{D}_{\sigma_p}(n) = \sum_{t \in \mathcal{T}_n} P_{\sigma_p}(t) |\mathcal{D}_t| = \sum_{t \in \mathcal{T}_n} P_{\sigma_p}(\vartheta(t)) |\mathcal{D}_{\vartheta(t)}| = \sum_{t \in \mathcal{T}_n} P_{\sigma_{1-p}}(t) |\mathcal{D}_t| = \mathcal{D}_{\sigma_{1-p}}(n).$$

Since $\frac{1}{2} < p < 1$, we have $0 < 1 - p < \frac{1}{2}$. Thus, the result for Case 2 follows from Case 1. ◀

4 Open Problems

Perhaps the most natural probability distribution on the set of binary trees with n leaves is the uniform distribution with $P_\sigma(t) = 1/C_{n-1}$ for every $t \in \mathcal{T}_n$, where C_n denotes the n^{th} Catalan number. The corresponding leaf-centric binary tree source is induced by the mapping $\sigma_{\text{eq}} \in \Sigma$ with $\sigma_{\text{eq}}(k, n-k) = C_{k-1} C_{n-k-1} / C_{n-1}$. In [8], it was shown that $\mathcal{D}_{\sigma_{\text{eq}}}(n) \in \Theta(n/\sqrt{\log n})$. Unfortunately, our main results Theorem 3 and Theorem 13 only yield the trivial bound $\mathcal{D}_{\sigma_{\text{eq}}} \in \mathcal{O}(n)$: An easy computation shows that $\sigma_{\text{eq}} \in \Sigma^\rho$ with $\rho = 1/4$ and $\sigma_{\text{eq}} \in \Sigma_\phi$ with $\phi(n) \in \Theta(1/\sqrt{n})$. An interesting open problem would be to find a nontrivial subset $\Sigma' \subseteq \Sigma$ that contains σ_{eq} and such that $\mathcal{D}_\sigma(n) \in \mathcal{O}(n/\sqrt{\log n})$ for all $\sigma \in \Sigma'$.

Another type of binary tree sources are so-called *depth-centric binary tree sources*, which yield probability distributions on the set of binary trees of a fixed depth; see for example [10, 16]. Depth-centric binary tree sources resemble leaf-centric binary tree sources in many ways. An interesting problem would be to estimate the average size of the minimal DAG with respect to certain classes of depth-centric binary tree sources.

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 Louisa Seelbach Benkner and Markus Lohrey. Average case analysis of leaf-centric binary tree sources. *CoRR*, abs/1804.10396, 2018. URL: `http://arxiv.org/abs/1804.10396`, `arXiv:1804.10396`.

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

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

5 Peter Buneman, Martin Grohe, and Christoph Koch. Path queries on compressed XML. In Johann Christoph Freytag et al., editors, *Proceedings of the 29th Conference on Very Large Data Bases, VLDB 2003*, pages 141–152. Morgan Kaufmann, 2003.

6 Luc Devroye. On the richness of the collection of subtrees in random binary search trees. *Information Processing Letters*, 65(4):195–199, 1998.

7 Philippe Flajolet, Xavier Gourdon, and Conrado Martínez. Patterns in random binary search trees. *Random Structures & Algorithms*, 11(3):223–244, 1997.

8 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.

9 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.

10 Danny Hucke and Markus Lohrey. Universal tree source coding using grammar-based compression. In *Proceedings of the 2017 IEEE International Symposium on Information Theory, ISIT 2017*, pages 1753–1757. IEEE, 2017.

11 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.

12 Donald E. Knuth. *The Art of Computer Programming: Volume 3 – Sorting and Searching.* Addison-Wesley, 1998.

13 Conrado Martínez. *Statistics under the BST model.* Dissertation, Universidad Politécnica de Cataluna, 1991.

14 Mike Paterson and Mark N. Wegman. Linear unification. *Journal of Computer and System Sciences*, 16(2):158–167, 1978.

15 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.

16 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.