Entropy of Some General Plane Trees

Zbigniew Gołębiewski
Wroclaw University of Science and Technology
Wrocław, Poland
Email: zbigniew.golebiewski@pwr.edu.pl

Abram Magner
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
Email: anmagner@illinois.edu

Wojciech Szpankowski
Purdue University
West Lafayette, IN, USA
Email: spa@cs.purdue.edu

***Abstract*—We continue developing the information theory of advanced data structures. In our previous work, we introduced structural entropy of unlabeled graphs and designed lossless compression algorithms for *binary* trees (with structure-correlated vertex names). In this paper, we consider d-ary trees ($d \geq 2$) and trees with unrestricted degree for which we compute the entropy (the first step to design optimal compression algorithms). It turns out that extending from binary trees to general trees is mathematically quite challenging and leads to new recurrences that find ample applications in the information theory of structures.**

I. Introduction

Advances in sensing, communication, and storage technologies have created a state of the art in which our ability to collect data from richly instrumented environments has far outpaced our ability to process, understand, and analyze it in a (provably) rigorous manner. A significant component of this complexity arises from the *multimodal* and *heterogeneous* nature of data. This poses significant challenges for theoretical characterization of limits of information storage and transmission and methods that achieve these limits. While ad hoc approaches are often currently deployed, critical issues regarding their performance, robustness, and scalability remain. These challenges have motivated our recent research program [3], [4], [8] and others [1], [7], [10]. It provides the basis for our effort in developing a comprehensive theory of information for multimodal and structured data, that is, multitype and context dependent structures.

As a start to understand advanced data structures in an information-theoretic setting, we focused on graphs [3] and trees with vertex names [8]. In [3] the entropy and an optimal compression algorithm (up to two leading terms of the entropy) for Erdős-Rényi graph structures were presented. Furthermore, in [9] an automata approach was used to design an optimal graph compression scheme. For binary plane-oriented trees, rigorous information-theoretic results were obtained in [7], complemented by a universal grammar-based lossless coding scheme [10].

In our recent work [8] (see also [4]) we study *binary* trees (with structure-correlated vertex names) and design an optimal compression scheme based on arithmetic encoding. In this paper, we extend our study on entropy of advanced data structures to d-ary trees (i.e., trees with degree $d \geq 2$) and general trees without any restriction on the degree. It turns out that moving from binary trees to d-ary (general) trees is mathematically quite challenging. First of all, in [8] we proved for binary trees an equivalence between two models: the *binary search tree model* and a model in which leaves are selected randomly to expand the tree by adding two additional nodes (new leaves). This equivalence allowed us to analyze the entropy of such trees by solving a relatively simple recurrence, namely

$$x_n = a_n + \frac{2}{n}\sum_{i=1}^{n-1} x_i$$

for some given a_n (e.g., for the entropy, $a_n = \log n$), where n denotes the number of internal nodes. However, for d-ary trees T_n on n internal nodes the entropy $H(T_n)$ satisfies

$$H(T_n) = H(\text{root}) + d\sum_{k=0}^{n-1} H(T_k)\, p_{n,k}$$

where $H(\text{root})$ is the entropy of the split probability at the root, and $p_{n,k}$ is the probability of one specified subtree being of size k. For the m-ary search tree model discussed in Section II, this recurrence can be handled by results from [2], [6]. In a more interesting d-ary tree model, we randomly select a leaf and add exactly d leaves to it. We studied this model previously in the special case of $d = 2$, but the analysis is more complicated when $d > 2$.

After some tedious algebra, we prove in Section III that the new type of recurrence we need to solve to find the entropy is of the following form (see Lemma 2):

$$x_n = a_n + \frac{\alpha}{n}\frac{n!}{\Gamma(n+\alpha-1)}\sum_{k=0}^{n-1}\frac{\Gamma(k+\alpha-1)}{k!}x_k \qquad (1)$$

where $\alpha = d/(d-1)$, a_n is given sequence, and Γ is the Euler gamma function. The situation is even more involved when we consider general trees in Section III-C where no restrictions on degrees are imposed.

We present our main results in Section III. We first provide in Corollary 1 the entropy rate for m-ary search trees. Then we consider d-ary recursive (also called *increasing*) trees and in Theorem 1 give our expression for the entropy of such trees. We extend it to general recursive trees in Theorem 2.

II. Models

In this section we describe the concepts of unlabeled plane trees with and without restrictions on the nodes' out-degrees. This will allow us to introduce three models of tree generation.

A. Unlabeled m-Ary Search Tree Generation

Search trees are plane trees built from a set of n distinct keys taken from some totally ordered set, for instance a random permutation of the numbers $\{1, 2, \ldots, n\}$. An m-ary search tree is a tree in which each node has at most m children; moreover, each node stores up to $m-1$ keys. We define the size of a search tree as the number of keys n. The construction of m-ary search tree can be described as follows [5]. If $n = 0$ the tree is empty. If $1 \leq n \leq m-1$ the tree consists of a root only, with all keys stored in the root. If $n \geq m$ we select $m-1$ keys that are called pivots. The pivots are stored in the root. The $m-1$ pivots split the set of remaining $n-m+1$ keys into m sublists $I_1, \ldots, I_m$: if the pivots are $p_1 < p_2 < \cdots < p_{m-1}$, then $I_1 := (p_i : p_i < p_1), I_2 := (p_i : p_1 < p_i < p_2), \ldots, I_m := (p_i : p_{m-1} < p_i)$. We then recursively construct a search tree for each of the sets I_i of keys. In order to obtain an unlabeled search tree of size n we remove the keys from a search tree.

B. Unlabeled d-ary Recursive Plane Tree Generation

We consider the following generation model of an unlabeled d-ary recursive plane tree. Suppose that the process starts with an empty tree, that is with just an external node (leaf). The first step in the growth process is to replace this external node by an internal one with d successors that are external nodes (see Figure 1). Then with probability $\frac{1}{d}$ each, one of these d external nodes is selected and again replaced by an internal node with d successors. At the end, we remove the labels (which describe the history of tree evolution) from internal nodes.

C. Unlabeled General Recursive Plane Trees Generation

We consider the following generation model of unlabeled plane trees. Suppose that the process starts with the root node carrying a label 1. Then we add a node with label 2 to the root. The next step is to attach a node with label 3. However, there are three possibilities: either to add it to the root (as a left or right sibling of 2) or to the node with label 2. Similarly one proceeds further. Now if a node already has out-degree k (where the descendants are ordered), then there are $k+1$ possible ways to add new node (this time we do not distinguish between external and internal nodes). Hence, if a plane tree already has $j-1$ nodes then there are precisely $2j-3$ possibilities to attach the jth node (see Figure 2). More precisely, the probability of choosing a node of out-degree k equals $(k+1)/(2j-3)$. At the end, we remove the labels from internal nodes of a tree.

III. Main Results

In this section we present our main results. In particular, we briefly address the entropy of m-ary search trees. Then we present our derivation of the the entropy of unlabeled d-ary recursive trees and general trees.

We should point out that in all our models, the probability of a tree generation is non-uniform, and root subtrees are conditionally independent given their respective sizes. Indeed, let T_n be a random variable representing a tree $\mathtt{t}_n$ on n internal nodes. Assume now that at the root we split $\mathtt{t}_n$ into d subtrees of size $k_1, \ldots, k_d$, respectively, where $k_1 + \cdots + k_d = n-1$. Then the probability $\mathbb{P}(T_n = \mathtt{t}_n)$ of generating tree $\mathtt{t}_n$ in all our models satisfies $\mathbb{P}(T_n = \mathtt{t}_n) = \mathbb{P}(k_1, \ldots, k_d) \prod_{i=1}^{d} \mathbb{P}(T_{k_i} = \mathtt{t}_{k_i})$ where $\mathbb{P}(k_1, \ldots, k_d)$ is the probability of a split at the root of n internal nodes into subtrees $t_{k_1}, \ldots, t_{k_d}$, respectively. This split probability is different for m-ary search trees, d-ary trees, and general trees, as we shall see in this section.

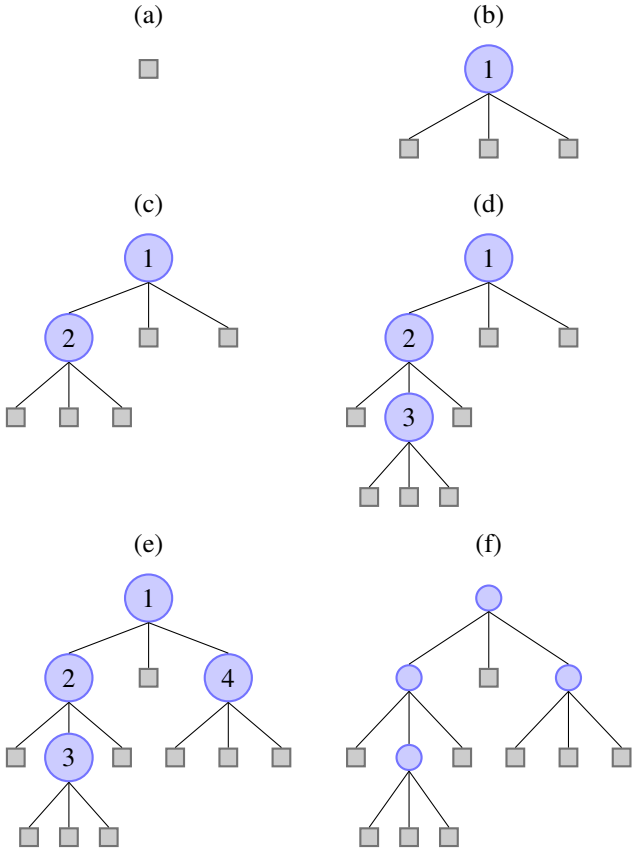

Fig. 1: Labeled and unlabeled 3-ary trees of size 4.

Throughout we shall use the following notation. Let $\mathbf{k}^{(n)} = (k_1, \ldots, k_n)$ denote an n-dimensional vector of non-negative integers and $\|\mathbf{k}^{(n)}\| = k_1 + \ldots + k_m$ be its L^1 norm. Let $(k, \mathbf{k}^{(n-1)}) = (k, k_2 \ldots, k_n)$ denote a n-dimensional vector with the first coordinate equal to k. We often write $\mathbf{k}$ instead of $\mathbf{k}^{(n)}$.

A. The Entropy of the Unlabeled m-ary Search Trees

Let U_n denote a random unlabeled m-ary search tree with n keys, generated according to the process described earlier. We write $\mathtt{u}_n$ for an arbitrary fixed m-ary (unlabeled) search tree with n keys.

We describe the splitting of keys at the root of the search tree by the random vector $\mathbf{Y}_n^{(m)} = (Y_{n,1}, \ldots, Y_{n,m})$, where $Y_{n,j} = |I_j|$ is the number of keys that go into the jth subtree of the root. If $n \geq m-1$ we have $Y_{n,1} + \cdots Y_{n,m} = n-m+1$ and $\mathbb{P}\left(\mathbf{Y}_n^{(m)} = \mathbf{k}^{(m)}\right) = 1/\binom{n}{m-1}$.

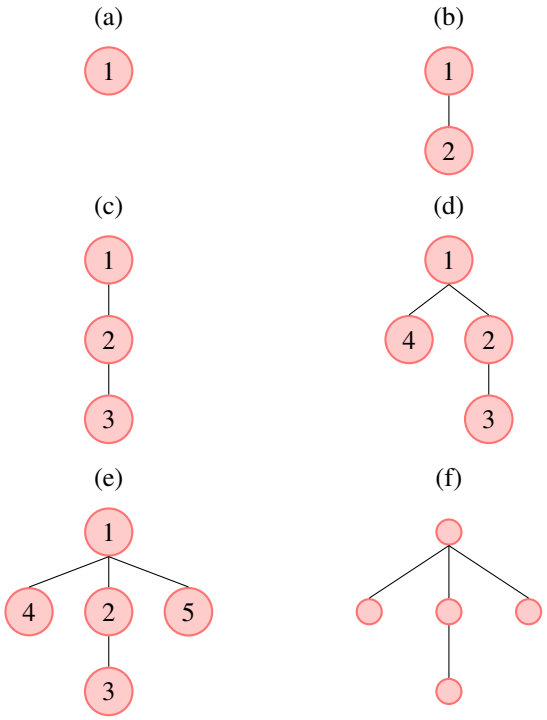

Fig. 2: Labeled and unlabeled general trees of size 5.

Suppose that the tree $\mathtt{u}_n$ has subtrees $\mathtt{u}_{k_1}, \ldots, \mathtt{u}_{k_m}$ of sizes $k_1, \ldots, k_m$, then

$$\mathbb{P}\left(U_n = \mathtt{u}_n\right) = \mathbb{P}\left(\mathbf{Y}_n^{(m)} = \mathbf{k}^{(m)}\right) \prod_{j=1}^{m} \mathbb{P}\left(U_{k_j} = \mathtt{u}_{k_j}\right).$$

If $n = 0$ we have an empty tree, and $H\left(U_0\right) = 0$. Moreover, if $1 \leq n \leq m-1$, all keys are stored in one node, and $H\left(U_n\right) = 0$. For $n > m-1$, we have

$$H\left(U_n\right) = H\left(\mathbf{Y}_n^{(m)}, U_{Y_{n,1}}, \ldots, U_{Y_{n,m}}\right) = H\left(\mathbf{Y}_n^{(m)}\right) +$$

$$\sum_{\|\mathbf{k}\|=n-m+1} H\left(U_{Y_{n,1}}, \ldots, U_{Y_{n,m}} | \mathbf{Y}_n^{(m)} = \mathbf{k}^{(m)}\right) \mathbb{P}\left(\mathbf{Y}_n^{(m)} = \mathbf{k}^{(m)}\right)$$

leading to

$$H\left(U_n\right) = H\left(\mathbf{Y}_n^{(m)}\right) + m \sum_{k=0}^{n-m+1} H\left(U_k\right) \mathbb{P}\left(Y_{n,1} = k\right).$$

For $n \geq m-1$ and $1 \leq j \leq m$, the random variables $Y_{n,j}$ are identically distributed, and for $0 \leq k \leq n-1$, $\mathbb{P}\left(Y_{n,1} = k\right) = \frac{\binom{n-k-1}{m-2}}{\binom{n}{m-1}}$ (see [5]). Hence $H\left(\mathbf{Y}_n^{(m)}\right) = \log\binom{n}{m-1}$ leading to the following recurrence

$$H\left(U_n\right) = \log\binom{n}{m-1} + \frac{m}{\binom{n}{m-1}} \sum_{k=0}^{n-m+1} \binom{n-k-1}{m-2} H\left(U_k\right).$$

The asymptotics of a recurrence like this one have been studied before; see Theorem 2.4 in [6].

Hence, the entropy of the m-ary search tree becomes $H(U_n) = c_m n + o\left(n\right)$ where

$$c_m = \frac{1}{\mathcal{H}_m - 1} \sum_{k \geq 0} \frac{\log\binom{k}{m-1}}{(k+1)(k+2)},$$

where $\mathcal{H}_m$ is the mth harmonic number.

Corollary 1. *The entropy rate $h_m = \lim_{n\to\infty} H\left(U_n\right)/n$ of the unlabeled m-ary trees, is given by*

$$h_m = \frac{1}{\mathcal{H}_m - 1} \sum_{k \geq 0} \frac{\log\binom{k}{m-1}}{(k+1)(k+2)}. \tag{2}$$

B. The Entropy of the Unlabeled d-ary Plane Recursive Trees

Let $\mathcal{F}_n$ be the set of d-ary plane trees with n internal nodes, and let $\mathcal{G}_n$ be the set of d-ary plan recursive trees with exactly n internal nodes. By $g_n = |\mathcal{G}_n|$ we denote the number of d-ary plane recursive trees with n internal nodes. From [5] we know that for $d = 2$ we have $g_n = n!$. Moreover, for $d > 2$ we have

$$g_n = (-1)^n (d-1)^n \frac{\Gamma(2 - \frac{d}{d-1})}{\Gamma(2 - \frac{d}{d-1} - n)}. \tag{3}$$

Let $\mathcal{G}_{\mathtt{f}_n}$ denote the subset of $\mathcal{G}_n$ of trees that have the same structure as the unlabeled tree $\mathtt{f}_n \in \mathcal{F}_n$; that is, $\mathcal{G}_{\mathtt{f}_n}$ is the set of labeled representatives of $\mathtt{f}_n$. Moreover, let $g_{\mathtt{f}_n} = |\mathcal{G}_{\mathtt{f}_n}|$ be the number of d-ary plane recursive trees that have the same structure as a tree $\mathtt{f}_n$. Observe that $\mathbb{P}\left(F_n = \mathtt{f}_n\right) = \frac{g_{\mathtt{f}_n}}{g_n}$. Suppose that the tree $\mathtt{f}_n$ has subtrees $\mathtt{f}_{k_1}, \ldots, \mathtt{f}_{k_d}$ of sizes $k_1, \ldots, k_d$. Then

$$\begin{aligned}\mathbb{P}\left(F_n = \mathtt{f}_n\right) &= \frac{1}{g_n}\binom{n-1}{k_1, \ldots, k_d} \prod_{j=1}^{d} g_{\mathtt{f}_{k_j}} \\ &= \binom{n-1}{k_1, \ldots, k_d} \frac{g_{k_1} \cdots g_{k_d}}{g_n} \prod_{j=1}^{d} \mathbb{P}\left(F_{k_j} = \mathtt{f}_{k_j}\right).\end{aligned} \tag{4}$$

Observe that $\binom{n-1}{k_1,\ldots,k_d}\frac{g_{k_1}\cdots g_{k_d}}{g_n}$ is the probability that the subtrees of the root are of sizes $k_1, \ldots, k_d$. Let us define a random vector $\mathbf{V}_n^{(d)} : \mathcal{G}_n \to \{0, \ldots, n-1\}^d$ whose jth component $V_{n,j}$ denotes the size of the jth subtree. For $n \geq 1$ we have $V_{n,1} + \ldots + V_{n,d} = n-1$ and

$$\mathbb{P}\left(\mathbf{V}_n^{(d)} = \mathbf{k}^{(d)}\right) = \binom{n-1}{k_1, \ldots, k_d} \frac{g_{k_1} \cdots g_{k_d}}{g_n}. \tag{5}$$

The entropy of unlabeled d-ary plane recursive trees of size n is defined as $H\left(F_n\right) = -\sum_{\mathtt{f}_n \in \mathcal{F}_n} \mathbb{P}\left(F_n = \mathtt{f}_n\right) \log\left(\mathbb{P}\left(F_n = \mathtt{f}_n\right)\right)$. If $n = 0$ we have an empty tree, and $H\left(F_0\right) = 0$. If $n = 1$, we have one fixed tree and $H\left(F_1\right) = 0$. By (4) for $n > 1$ there is a bijection between a tree F_n and a tuple $(\mathbf{V}_n^{(d)}, F_{V_1}, \ldots, F_{V_d})$. Therefore, for $n > 1$, we have

$$H\left(F_n\right) = H\left(\mathbf{V}_n^{(d)}, F_{V_{n,1}}, \ldots, F_{V_{n,d}}\right) = H\left(\mathbf{V}_n^{(d)}\right) +$$

$$\sum_{\|\mathbf{k}\|=n-1} H\left(F_{k_1}, \ldots, F_{k_d}\right) \mathbb{P}\left(\mathbf{V}_n^{(d)} = \mathbf{k}^{(d)}\right).$$

Since subtrees $F_{k_1},\ldots,F_{k_d}$ are conditionally independent given their sizes, we have

$$H\left(F_n\right)=H\left(\mathbf{V}_n^{(d)}\right)+ d\sum_{k=0}^{n-1}H\left(F_k\right)\sum_{\|\mathbf{k}^{(d-1)}\|=n-1-k}\mathbb{P}\left(\mathbf{V}_n^{(d)}=\left(k,\mathbf{k}^{(d-1)}\right)\right).$$

For $k=0,\ldots,n-1$, let $p_{n,k}$ be the probability that one specified subtree in a d-ary recursive tree is of size k, that is, $p_{n,k}=\sum_{\|\mathbf{k}^{(d-1)}\|=n-1-k}\mathbb{P}\left(\mathbf{V}_n^{(d)}=\left(k,\mathbf{k}^{(d-1)}\right)\right)$. Then we have the following recurrence

$$H\left(F_n\right)=H\left(\mathbf{V}_n^{(d)}\right)+d\sum_{k=0}^{n-1}H\left(F_k\right)p_{n,k}. \tag{6}$$

We can prove the following lemma.

Lemma 1. *For $k=0,\ldots,n-1$ and $d>1$, $\alpha=\frac{d}{d-1}$, then*

$$p_{n,k}=\frac{(\alpha-1)}{n}\frac{n!\Gamma(k+\alpha-1)}{k!\Gamma(n+\alpha-1)}.$$

In the lemma below we propose a general solution for recurrences of the form (6).

Lemma 2. *For constant α,x_0 and x_1, the recurrence*

$$x_n=a_n+\frac{\alpha}{n}\frac{n!}{\Gamma(n+\alpha-1)}\sum_{k=0}^{n-1}\frac{\Gamma(k+\alpha-1)}{k!}x_k,\qquad n\geq 2 \tag{7}$$

has the following solution for $n\geq 2$:

$$x_n=a_n+\alpha(n+\alpha-1)\sum_{k=0}^{n-1}\frac{a_k}{(k+\alpha-1)(k+\alpha)} +\frac{n+\alpha-1}{\alpha+1}\left(x_1+\frac{x_0}{\alpha-1}\right).$$

Proof. Multiply both sides of the recurrence by the normalizing factor $\frac{\Gamma(n+\alpha-1)}{n!}$. Define also $\hat{x}_n=\frac{x_n\Gamma(n+\alpha-1)}{n!}$ and $\hat{a}_n=\frac{a_n\Gamma(n+\alpha-1)}{n!}$. Then

$$\hat{x}_n=\hat{a}_n+\frac{\alpha}{n}\sum_{k=2}^{n-1}\hat{x}_k. \tag{8}$$

To solve the recurrence (8) we compute $n\hat{x}_n-(n-1)\hat{x}_{n-1}$. This leads us to

$$\hat{x}_n=\hat{a}_n-\left(1-\frac{1}{n}\right)\hat{a}_{n-1}+\left(1+\frac{\alpha-1}{n}\right)\hat{x}_{n-1},$$

which holds for $n\geq 3$. Then after iterating the above we arrive at

$$\hat{x}_n=\hat{x}_2\prod_{j=3}^{n}\left(1+\frac{\alpha-1}{j}\right)+ \sum_{k=3}^{n}\left(\hat{a}_k-\left(1-\frac{1}{k}\right)\hat{a}_{k-1}\right)\prod_{j=k+1}^{n}\left(1+\frac{\alpha-1}{j}\right). \tag{9}$$

The product $\prod_{j=k+1}^{n}\left(1+\frac{\alpha-1}{j}\right)=\frac{k!\Gamma(n+\alpha)}{n!\Gamma(k+\alpha)}$, and after some standard calculations we obtain

$$\hat{x}_n=\hat{a}_n+(\hat{x}_2-\hat{a}_2)\frac{2\Gamma(n+\alpha)}{\Gamma(\alpha+2)n!}+ \frac{\Gamma(n+\alpha)}{n!}\sum_{k=2}^{n-1}\hat{a}_k\frac{k!}{\Gamma(k+\alpha)}\frac{\alpha}{k+\alpha}.$$

Going back from $\hat{x}_n$ and $\hat{a}_n$ to x_n,a_n, respectively, we obtain

$$x_n=a_n+\alpha(n+\alpha-1)\sum_{k=2}^{n-1}\frac{a_k}{(k+\alpha-1)(k+\alpha)} +(x_2-a_2)\frac{n+\alpha-1}{\alpha+1}.$$

But $x_2-a_2=x_1+\frac{x_0}{\alpha-1}$ which completes the proof. □

This leads us to out first main result.

Theorem 1. *The entropy of an unlabeled d-ary plane tree is*

$$H\left(F_n\right)=H\left(\mathbf{V}_n^{(d)}\right)+\alpha(n+\alpha-1)\sum_{k=0}^{n-1}\frac{H\left(\mathbf{V}_k^{(d)}\right)}{(k+\alpha-1)(k+\alpha)}, \tag{10}$$

where $\alpha=\frac{d}{d-1}$ and

$$H\left(\mathbf{V}_n^{(d)}\right)=-\sum_{\|\mathbf{k}\|=n-1}\mathbb{P}\left(\mathbf{V}_n^{(d)}=\mathbf{k}^{(d)}\right)\log\mathbb{P}\left(\mathbf{V}_n^{(d)}=\mathbf{k}^{(d)}\right).$$

Furthermore, the entropy rate $h_d=\lim_{n\to\infty}H\left(F_n\right)/n$ is

$$h_d=\alpha\sum_{k=0}^{\infty}\frac{H\left(\mathbf{V}_k\right)}{(k+\alpha-1)(k+\alpha)}. \tag{11}$$

C. The Entropy of the Unlabeled General Plane Trees

Let $\mathcal{T}_n$ be the set all unlabeled trees with n internal nodes (of any degree), while $\mathcal{T}_n^{(d)}$ be the subset of $\mathcal{T}_n$ consisting of all trees that contain exactly n nodes and have root degree equal to d. Finally, by $\mathcal{R}_n$ we denote a set of labeled plane trees oriented recursive trees with exactly n nodes.

Let $r_n=|\mathcal{R}_n|$. From [5] we know that there are

$$r_n=(2n-3)!!=\frac{n!}{n2^{n-1}}\binom{2n-2}{n-1}$$

different labeled plane oriented recursive trees of size n.

As in the case of the d-ary plane recursive trees, let $\mathcal{R}_{\mathtt{t}_n}$ denote the subset of trees in $\mathcal{R}_n$ that have the same structure as a given unlabeled tree $\mathtt{t}_n\in\mathcal{T}_n$ (i.e., $\mathcal{R}_{\mathtt{t}_n}$ is the set of labeled *representatives* of $\mathtt{t}_n$); moreover, let $r_{\mathtt{t}_n}=|\mathcal{R}_{\mathtt{t}_n}|$ be the number of such trees. Observe that $\mathbb{P}\left(T_n=\mathtt{t}_n\right)=\frac{r_{\mathtt{t}_n}}{r_n}$.

Let D_n denote the random variable representing the number of subtrees of the root. Observe that $\mathbb{P}\left(D_n=d\right)=\frac{r_n^{(d)}}{r_n}$, where $r_n^{(d)}=|\mathcal{R}_n^{(d)}|$ is the number of plane recursive trees with root degree equal to d. Suppose that the tree $\mathtt{t}_n$ has d subtrees $\mathtt{t}_{k_1},\ldots,\mathtt{t}_{k_d}$ of sizes $k_1,\ldots,k_d$. Then

$$\mathbb{P}\left(T_n=\mathtt{t}_n\right)=\mathbb{P}\left(D_n=d\right)\mathbb{P}\left(T_n=\mathtt{t}_n|D_n=d\right) =\binom{n-1}{k_1,\ldots,k_d}\frac{r_{k_1}\cdots r_{k_d}}{r_n}\prod_{j=1}^{d}\mathbb{P}\left(T_{k_j}=\mathtt{t}_{k_j}\right). \tag{12}$$

Observe that $\binom{n-1}{k_1,\ldots,k_d}\frac{r_{k_1}\cdots r_{k_d}}{r_n}$ is the probability that the root of a plane recursive tree of size n has degree equal to d and the root's subtrees are of sizes $k_1,\ldots,k_d$. Let $\mathbf{W}_n^{(d)}:\mathcal{R}_n^{(d)}\to\{1,\ldots,n-d\}^d$, where its jth component $W_{n,j}$ denotes the size of the jth subtree when the root is of degree d. For $n\geq 1$ we have $W_{n,1}+\ldots+W_{n,d}=n-1$ and

$$\mathbb{P}\left(D_n=d\right)\mathbb{P}\left(\mathbf{W}_n^{(D_n)}=\mathbf{k}^{(D_n)}|D_n=d\right)=\binom{n-1}{k_1,\ldots,k_d}\frac{r_{k_1}\cdots r_{k_d}}{r_n}. \tag{13}$$

For the entropy of unlabeled plane recursive trees of size n, using the conditional independence of $T_{k_1},\ldots,T_{k_d}$, we have

$$H\left(T_n\right)=\sum_{d=1}^{n-1}H\left(\mathbf{W}_n^{(d)}|D_n=d\right)\mathbb{P}\left(D_n=d\right)+\sum_{d=1}^{n-1}\mathbb{P}\left(D_n=d\right)d\sum_{k=1}^{n-d}H\left(T_k\right)\sum_{\|\mathbf{k}^{(d-1)}\|=n-1-k}\mathbb{P}\left(\mathbf{W}_n^{(d)}=\left(k,\mathbf{k}^{(d-1)}\right)\right).$$

For $k=1,\ldots,n-1$, let $q_{n,k}=q_{n,k}^{(d)}$ be defined as the probability that the root of a plane recursive tree has degree d and that one specified root subtree is of size k. Then

$$q_{n,k}^{(d)}=\mathbb{P}\left(D_n=d\right)\sum_{\|\mathbf{k}^{(d-1)}\|=n-1-k}\mathbb{P}\left(\mathbf{W}_n^{(d)}=\left(k,\mathbf{k}^{(d-1)}\right)\right) \tag{14}$$

leading to

$$H\left(T_n\right)=\sum_{d=1}^{n-1}H\left(\mathbf{W}_n^{(d)}|D_n=d\right)\mathbb{P}\left(D_n=d\right)+\sum_{d=1}^{n-1}d\sum_{k=1}^{n-d}H\left(T_k\right)q_{n,k}^{(d)}. \tag{15}$$

We need an expression for the probability $q_{n,k}^{(d)}$ which we present in the next lemma.

Lemma 3. *For $k=1,\ldots,n-1$ we have $q_{n,n-1}^{(1)}=\frac{1}{2n-3}$ and if $k\neq n-1:\ q_{n,k}^{(1)}=0$, while for $d>1$:*

$$q_{n,k}^{(d)}=2^d\frac{d-1}{k(n-1-k)}\frac{\binom{2k-2}{k-1}\binom{2(n-1-k)-d}{n-2-k}}{\binom{2n-2}{n-1}}.$$

The recurrence (15) is another recurrence that we need to analyze. Its general solution is presented next. The proof can be found in the journal version of the paper.

Lemma 4. *For constant y_1 and y_2, the recurrence*

$$y_n=b_n+\sum_{d=1}^{n-1}d\sum_{k=1}^{n-d}q_{n,k}^{(d)}\cdot y_k,\qquad n>2 \tag{16}$$

has the following solution for $n>2$:

$$y_n=\frac{2(2n-1)}{3}b_1+b_n+\frac{1}{2}\left(n-\frac{1}{2}\right)\sum_{j=2}^{n-1}\frac{b_j}{\left(j-\frac{1}{2}\right)\left(j+\frac{1}{2}\right)}.$$

This leads us to our second main result.

Theorem 2. *The entropy of an unlabeled general plane tree is*

$$H\left(T_n\right)=\sum_{d=1}^{n-1}H\left(\mathbf{W}_n^{(d)}|D_n=d\right)\mathbb{P}\left(D_n=d\right) \tag{17}$$

$$+\frac{1}{2}\left(n-\frac{1}{2}\right)\sum_{j=2}^{n-1}\frac{\sum_{d=1}^{j-1}H\left(\mathbf{W}_j^{(d)}|D_j=d\right)\mathbb{P}\left(D_j=d\right)}{\left(j-\frac{1}{2}\right)\left(j+\frac{1}{2}\right)}, \tag{18}$$

where

$$H\left(\mathbf{W}_n^{(d)}|D_n=d\right)=-\sum_{\|\mathbf{k}\|=n-1}\mathbb{P}\left(\mathbf{W}_n^{(d)}=\mathbf{k}^{(d)}|D_n=d\right)\cdot\log\mathbb{P}\left(\mathbf{W}_n^{(d)}=\mathbf{k}^{(d)}|D_n=d\right).$$

Furthermore, the entropy rate $h_t=\lim_{n\to\infty}H\left(T_n\right)/n$ is

$$h_t=\frac{1}{2}\sum_{j=2}^{\infty}\frac{\sum_{d=1}^{j-1}H\left(\mathbf{W}_k^{(d)}|D_n=d\right)\mathbb{P}\left(D_n=d\right)}{\left(j-\frac{1}{2}\right)\left(j+\frac{1}{2}\right)}. \tag{19}$$

Finally, we did not address how to optimally compress these trees, but it is not hard to see that a direct generalization of the arithmetic encoding proposed in [8] can be used. For example, in the d-ary tree cases, we traverse the tree in depth-first order from left to right, taking advantage of the fact that conditioned on the size k of the leftmost root subtree, the rest of the tree is a random $d-1$-ary tree of size $n-k$.

ACKNOWLEDGMENT

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, by NSF Grant CCF-1524312, and NIH Grant 1U01CA198941-01. Z. Gołębiewski was in addition supported by Polish NCN grant 2013/09/B/ST6/02258.

REFERENCES

[1] D. Aldous and N. Ross, Entropy of Some Models of Sparse Random Graphs With Vertex-Names. *Probability in the Engineering and Informational Sciences*, 2014, 28:145-168.

[2] Hua-Huai Chern and Hsien-Kuei Hwang. Phase changes in random m-ary search trees and generalized quicksort. *Random Struct. Algorithms*, 19(3-4):316–358, 2001.

[3] Y. Choi, W. Szpankowski: Compression of Graphical Structures: Fundamental Limits, Algorithms, and Experiments. *IEEE Transactions on Information Theory*, 2012, 58(2):620-638.

[4] J. Cichon, A. Magner, W. Szpankowski, K. Turowski, On symmetries of non-plane trees in a non-uniform model, *ANALCO*, Barcelona, 2017.

[5] M. Drmota. *Random Trees, An Interplay between Combinatorics and Probability*. Springer-Verlag Wien, 2009.

[6] J. Fill and N. Kapur. Transfer theorems and asymptotic distributional results for m-ary search trees. *Random Structures & Algorithms*, 26(4):359–391, 2005.

[7] J. C. Kieffer, E.-H. Yang, W. Szpankowski, Structural complexity of random binary trees. *ISIT 2009* , pp. 635-639.

[8] A. Magner, W. Szpankowski, K. Turowski, Lossless Compression of Binary Trees with Correlated Vertex Names, *ISIT'16*, Barcelona, 2016.

[9] M. Mohri, M. Riley, A. T. Suresh, Automata and graph compression. *ISIT 2015*, pp. 2989-2993.

[10] J. Zhang, E.-H. Yang, J. C. Kieffer, A Universal Grammar-Based Code for Lossless Compression of Binary Trees. *IEEE Transactions on Information Theory*, 2014, 60(3):1373-1386.