On Symmetries of Non-Plane Trees in a Non-Uniform Model*

Jacek Cichoń* Abram Magner† Wojciech Szpankowski‡ Krzysztof Turowski§

October 31, 2016

Abstract

Binary trees come in two varieties: plane trees, often simply called binary trees, and non-plane trees, in which the order of subtrees does not matter. Non-plane trees find many applications; for example in modeling epidemics, in studying phylogenetic trees, and as models in data compression. While binary trees have been studied very extensively, non-plane trees still pose some challenges. Moreover, in most analyses a uniform probabilistic model is assumed; that is, a tree is selected uniformly from among all trees. Such a model limits significantly applications of the analysis. In this paper we study by analytic techniques non-plane trees in a non-uniform model. In our model, we grow the tree on n leaves by selecting randomly a leaf and appending two children to it. We can show that this is equivalent to an alternative model, also used to analyze the average-case performance of binary search trees, that is more easily amenable to study by recurrences and generating functions. Here, one of the most important questions is the number of symmetries in such trees (i.e., the number of internal nodes with two isomorphic subtrees), or the sizes of such symmetric subtrees. We first present a functional-differential equation characterizing tree symmetries, and then analyze it. In this conference paper we focus on the expected number of symmetries, the size of symmetric subtrees, and the tree entropy.

Keywords: Binary trees, non-plane trees, symmetry, Hadamard product, entropy, functional-differential equations.

1 Introduction

Binary trees come in two flavors: plane-oriented trees and non-plane or Otter trees. A plane tree, that we shall call just a binary tree, is defined as a rooted tree in which two subtrees of the same parent node are ordered between themselves and represented from left to right; that is, they have a distinguished embedding in the plane. A non-plane tree, also called unordered tree or an Otter tree, is just a rooted tree in the graph theoretic sense, so that there is no order between subtrees.

Non-plane trees are harder to analyze and less attention in literature has been devoted to these kind of trees. Nevertheless, non-plane trees find many important applications from data compression to biology. For example, in a phylogenetic tree describing n species there are n leaves representing extant species and $n-1$ internal nodes. By design there is no specific order between the two children of a binary node.

In combinatorics both trees have been studied [5, 11] but only in a uniform probability model in which trees are selected uniformly from a collection of all distinct trees. However, such a model is often too simple to capture nuances of various applications. Therefore in this paper we study non-plane (and the corresponding plane) trees under a non-uniform model that we describe below.

There are two equivalent models that are useful in applications such as data compression [9]. Let $\mathcal{T}$ be the set of all binary rooted plane trees having finitely many vertices and, for each positive integer n, let $\mathcal{T}_n$ be the subset of $\mathcal{T}$ consisting of all trees with exactly n leaves. Similarly, let $\mathcal{S}$ and $\mathcal{S}_n$ be the set of all binary

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

*Jacek Cichoń is with Department of Computer Science Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland (email: jacek.cichon@pwr.edu.pl)

†Abram Magner is with Coordinated Science Lab at Univ. of Illinois at Urbana-Champaign, Urbana, IL, USA, (email: anmagner@illinois.edu).

‡Krzysztof Turowski is with Dept. of Algorithms and System Modeling, at Gdańsk University of Technology, Gdansk, Poland, (email: krzturow@pg.gda.pl).

§W. Szpankowski is with the Department of Computer Science, Purdue University, IN 47907, USA (e-mail: spa@cs.purdue.edu)

rooted non-plane trees with finitely many vertices and exactly n leaves, respectively. In the first model to grow a tree built on n leaves we pick up randomly a leaf v, and then append two children to it, say v^L and v^R. Our second model (also known as the binary search tree model), is ubiquitous in the computer science literature, arising for example in the context of binary search trees formed by inserting a random permutation of $[n-1]$ into a binary search tree [1, 6]. Under this model we generate a random tree T_n as follows: at first, t is equal to the unique tree in $\mathcal{T}_1$ and we associate a number n with its single vertex. Then, in each recursive step, let $v_1, v_2, \ldots, v_k$ be the leaves of t, and let integers $n_1, n_2, \ldots, n_k$ be the values assigned to these leaves, respectively. For each leaf v_i with value $n_i > 1$, randomly select integer s_i from the set $\{1, \ldots, n_i - 1\}$ with probability $\frac{1}{n_i-1}$ (independently of all other such leaves), and then grow two edges from v_i with left edge terminating at a leaf of the extended tree with value s_i and right edge terminating at a leaf of the extended tree with value $n_i - s_i$. The extended tree is the result of the current recursive step. Clearly, the recursion terminates with a binary tree having exactly n leaves, in which each leaf has assigned value 1; this tree is T_n. In [9, 13] it was proved that both models are probabilistically equivalent.

In data compression and other applications symmetries of plane and non-plane trees are of interest. In particular, for compression one needs to know the number of internal nodes that have two isomorphic subtrees, the size of such isomorphic trees, and the entropy of trees. In this conference paper we set the stage to analyze these quantities for non-plane (and the corresponding plane) trees. We derive a functional-differential equation for the bivariate generating function from which we compute the average number of nodes with two isomorphic subtrees as well as the entropy of non-plane trees (see also the recent paper [9]).

There are a few scattered results for non-plane trees in a uniform model [2, 5, 11]. The non-plane trees in the binary search tree-like model were studied by analytic tools in [1, 6] in a different context, and by a different method in [9]. In this work we use analytic combinatorics to initiate systematic studies of non-plane trees.

2 Main Results

Let $\mathcal{S}$ be the set of all binary rooted non-plane trees having finitely many vertices. Such trees are often called Otter trees. Let $\mathcal{S}_n$ be the subset of $\mathcal{S}$ consisting of all trees with exactly n leaves. We also denote by $\mathcal{T}$ the set of all binary rooted plane trees having finitely many vertices. Let $\mathcal{T}_n$ be the subset of $\mathcal{T}$ consisting of all trees with exactly n leaves. For any $s \in \mathcal{S}$ and $t \in \mathcal{T}$ let $t \sim s$ mean that the plane tree t is isomorphic to the non-plane tree s. Furthermore, we define $[s] = \{t \in \mathcal{T} : t \sim s\}$ as a collection of all plane trees t having the same non-plane tree s.

In this paper we denote by $|t|$ the number of leaves of a tree t. The probability of a tree $t \in \mathcal{T}_n$ in our non-uniform model is [9]

$$P(T_n = t) = \prod_{v \in t^o} (\Delta(v) - 1)^{-1} ,$$

where t^o is the set of internal nodes of t and $\Delta(v)$ is the number of leaves of a tree rooted at v.[1] These probabilities satisfies the following recurrence

$$P(T_n = t) = \frac{1}{n-1} P(T_{\Delta(t_1)} = t_1) P(T_{\Delta(t_2)} = t_2) ,$$

where t_1, t_2 are the two subtrees of tree t whose roots are the two children of the root of t. The recursion starts with the unique tree t_1 from T_1, when we have $P(T_1 = t_1) = 1$. It should be noted that this recursion resembles a recurrence for binary search trees which was studied in [1, 6].

For any $t_1, t_2 \in T_n$ such that $t_1 \sim s$ and $t_2 \sim s$ it holds that $P(T_n = t_1) = P(T_n = t_2)$. By definition, s corresponds to $|[s]|$ isomorphic plane trees, so for any $t \in [s]$ it holds that

$$P(S_n = s) = |[s]| \cdot P(T_n = t), \qquad t \in [s]. \tag{2.1}$$

In other words, we put

$$P(S_n = s) = \sum_{t \sim s} P(T_n = t) .$$

Furthermore,

$$P(T_n = t | S_n = s) = \frac{1}{|[s]|}. \tag{2.2}$$

Let sym(t) be the the number of non-leaf (internal) nodes v of tree t such that the two subtrees stemming

[1] We use T_n to denote a random variable representing a random tree in $\mathcal{T}_n$; however, we often abuse notation and write T_n for both the set of trees and the random variable.

from v are isomorphic. Observe that $|\mathcal{S}_1| = |\mathcal{S}_2| = |\mathcal{S}_3| = 1$, and that $|\mathcal{S}_4| = 2$. If $t \in \mathcal{S}_1$, then clearly $\mathrm{sym}(t) = 0$. If $t \in \mathcal{S}_2$ or $t \in \mathcal{S}_3$ then $\mathrm{sym}(t) = 1$. Notice that (see [2]) if t_1 and t_2 are the two subtrees of a tree t whose roots are the two children of the root of t, then

$$\mathrm{sym}(t) = \begin{cases} \mathrm{sym}(t_1) + \mathrm{sym}(t_2) + 1 & \text{if } t_1 = t_2 \\ \mathrm{sym}(t_1) + \mathrm{sym}(t_2) & \text{if } t_1 \neq t_2. \end{cases}$$

Observe also that

$$\mathrm{sym}(s) = \mathrm{sym}(t) \quad \forall_t \ t \in [s].$$

Furthermore, and more interestingly

$$|[s]| = 2^{n-1-\mathrm{sym}(s)}. \tag{2.3}$$

Indeed, we can form a new tree by rotating both subtrees of a plane tree at every internal node that is *not* symmetric, that is, for those nodes whose subtrees are not isomorphic.

Our goal is to evaluate the average $\mathbb{E}[\mathrm{sym}(S_n)]$ for non-plane trees. We shall accomplish it through generating function approach.

We first derive a differential equation for the following bivariate generating function

$$\begin{aligned} F(u,z) &= \sum_{t \in \mathcal{T}} P(T = t) u^{\mathrm{sym}(t)} z^{|t|} \\ &= \sum_{n=1}^{\infty} \sum_{t \in \mathcal{T}_n} P(T_n = t) u^{\mathrm{sym}(t)} z^{|t|} \ . \end{aligned}$$

Define also

$$\begin{aligned} &B(u,z) \\ &= \sum_{t \in \mathcal{T}} P^2(T = t) u^{\mathrm{sym}(t)} z^{|t|-1} \sum_{t \in \mathcal{T}} p^2(t) u^{\mathrm{sym}(t)} z^{|t|-1} \end{aligned}$$

where we write $p(t) = P(T = t)$ to simplify our notation.

In the next section we prove the following lemma characterizing $F(u,z)$.

Lemma 2.1. *Let $f(u,z) = \frac{F(u,z)}{z}$. Then it satisfies the following Riccati differential equation*

$$\frac{\partial f(u,z)}{\partial z} = f(u,z)^2 + (u-1)B(u^2, z^2). \tag{2.4}$$

Furthermore after the substitution

$$f(u,z) = -\frac{\frac{\partial g(u,z)}{\partial z}}{g(u,z)}$$

equation (2.4) becomes

$$\frac{\partial^2 g(u,z)}{\partial^2 z} + (u-1)B(u^2, z^2) g(u,z) = 0 \tag{2.5}$$

which is a second order linear equation assuming $B(u,z)$ is known.

Remark (i) We should point out differences between our non-uniform model and uniform model. For the uniform model, following Bóna and Flajolet [2] the bivariate generating function

$$\tilde{F}(u,z) = \sum_{s \in \mathcal{S}} u^{\mathrm{sym}(s)} z^{|s|}$$

satisfies the following functional equation

$$\tilde{F}(u,z) = z + \frac{1}{2}\tilde{F}(u,z)^2 + (u - \frac{1}{2})\tilde{F}(u^2, z^2).$$

(ii) Furthermore, we observe that (2.4) could be viewed as a functional-differential equation. Indeed, let us introduce a special Hadamard product of two generating functions $A(u,z)$, $B(u,z)$ defined by

$$\begin{aligned} A(u,z) &= \sum_{t \in \mathcal{T}} c_A(t) u^{r(t)} z^{s(t)}, \\ B(u,z) &= \sum_{t \in \mathcal{T}} c_B(t) u^{r(t)} z^{s(t)}, \end{aligned}$$

where $r(t)$ and $s(t)$ map combinatorial objects $t \in \mathcal{T}$ to the non-negative integers. We then define a Hadamard product $A(u,z) \boxdot B(u,z)$ by

$$A(u,z) \boxdot B(u,z) = \sum_{t \in \mathcal{T}} c_A(t) c_B(t) u^{r(t)} z^{s(t)}.$$

We contrast this with the standard bivariate Hadamard product, defined by

$$\begin{aligned} &A(u,z) \odot B(u,z) = \\ &\sum_{n,m=0}^{\infty} \left(\sum_{t \,:\, r(t)=m, s(t)=n} c_A(t) \right) \\ &\qquad \left(\sum_{t \,:\, r(t)=m, s(t)=n} c_B(t) \right) u^m z^n. \end{aligned}$$

Note that $A(u,z) \boxdot B(u,z)$ is in general not equal to $A(u,z) \odot B(u,z)$, unless $r(t)$ and $s(t)$ uniquely determine t. With this definition in mind we can rewrite (2.4) as

$$\frac{\partial}{\partial z} f_z(u,z) = f(u,z)^2 + (u-1)(f(u^2, z^2) \boxdot f(u^2, z^2)) \tag{2.6}$$

which is a functional-differential equation. □

Our first goal is to understand probabilistic behavior of $\mathrm{sym}(S_n) = \mathrm{sym}(T_n)$. In this preliminary conference paper we focus on the average $\mathbb{E}[\mathrm{sym}(S_n)]$. We start with a definition

$$\mathcal{E}(z) = \sum_{n=1}^{\infty} \mathbb{E}[\mathrm{sym}(S_n)] z^{n-1}.$$

Using Lemma 2.1 in the next section we show that it satisfies the following ODE

$$\mathcal{E}'(z) = \frac{2\mathcal{E}(z)}{z(1-z)} + B(z^2) \tag{2.7}$$

with $\mathcal{E}(0) = 0$. Here we define

$$\begin{aligned} B(z) &= \sum_{t \in \mathcal{T}} p^2(t) z^{|t|-1} \\ &= \sum_{n=1}^{\infty} z^{n-1} \sum_{t_n \in \mathcal{T}_n} p^2(t_n) \\ &= \sum_{n=1}^{\infty} b_n z^{n-1} \end{aligned}$$

where $b_n = [z^{n-1}]B(z) = \sum_{t_n \in \mathcal{T}_n} p^2(t_n)$ for $n \geq 1$. It is easy to compute b_n for a few small values of n, namely (2.8)

$$b_1 = b_2 = 1, \quad b_3 = \frac{1}{2}, \quad b_4 = \frac{2}{9}, \quad b_5 = \frac{13}{144}, \quad b_6 = \frac{7}{200}.$$

Actually, to compute precisely b_n for all values of n we need a better approach. Define $C(z) = zB(z)$ and notice that

$$b_n = [z^n]C(z) = [z^{n-1}]B(z).$$

We will derive a differential equation for $C(z)$, from which a recurrence for b_n will follow. It is easy to notice that $C(z)$ satisfies the following

$$C(z) = z + \sum_{u,v \in \mathcal{T}} \frac{1}{(|v|+|u|-1)^2} p^2(u) p^2(v) z^{|u|+|v|}.$$

Furthermore,

$$B'(z) = \sum_{u,v \in \mathcal{T}} \frac{1}{(|v|+|u|-1)} p^2(u) p^2(v) z^{|u|+|v|-2}$$

and

$$C^2(z) = \sum_{u,v \in \mathcal{T}} p(u) p(v) z^{|u|+|v|}.$$

Combining all of these, we arrive at the following differential equation

$$C(z) - zC'(z) + z^2 C''(z) = C^2(z). \tag{2.9}$$

By standard tools we can extract from the above a recurrence for $b_n = [z^n]C(z)$. Indeed, for $n \geq 2$ we have

$$b_n = \frac{1}{(n-1)^2} \sum_{j=1}^{n-1} b_j b_{n-j} \tag{2.10}$$

with $b_1 = 1$. In passing we should point out that asymptotic growth of b_n satisfying the above recurrence can be found in [3]. We come back to this issue later in this paper.

Remark. We can write (2.7) as a functional-differential equation using a different special Hadamard product defined as follows. For generating functions

$$A(z) = \sum_{t \in \mathcal{T}} c_A(t) z^{s(t)}, \qquad B(z) = \sum_{t \in \mathcal{T}} c_B(t) z^{s(t)},$$

we define $A(z) \boxtimes B(z)$ by

$$A(z) \boxtimes B(z) = \sum_{t \in \mathcal{T}} c_A(t) c_B(t) z^{s(t)}.$$

Define now

$$A(z) = \sum_{t \in \mathcal{T}} p(t)^2 z^{|t|-1}.$$

Then

$$\begin{aligned} (A(z) \boxtimes A(z))\Big|_{z^2} &= \sum_{t \in \mathcal{T}} p(t)^2 z^{2(|t|-1)} \\ &= \left[(f(1,z) \boxdot f(1,z))\Big|_{(u^2,z^2)} \right], \end{aligned}$$

and (2.7) becomes

$$\mathcal{E}'(z) = \frac{2\mathcal{E}(z)}{(1-z)} + (A(z^2) \boxtimes A(z^2)),$$

which is functional-differential equation. □

In the next section we solve the ODE (2.7) leading to our second main result.

Theorem 2.1. *Consider a binary plane tree and its corresponding non-plane tree. The expected number of* *internal nodes with two isomorphic subtrees is*

(2.11)

$$\begin{aligned}&\mathbb{E}[\mathrm{sym}(S_n)]\\ &= n\sum_{\ell=1}^{\lfloor (n+1)/2\rfloor} \frac{b_\ell}{(2\ell-1)\ell(2\ell+1)} + (-1)^{n+1} b_{\lfloor (n+1)/2\rfloor}\\ &\approx n\dot{(}0.3725 \pm 10^{-4})\end{aligned}$$

where, we recall, $b_\ell = \sum_{t_\ell} p^2(t_\ell)$.

Remark. Observe that we can use just computed $\mathbb{E}[\mathrm{sym}(S_n)]$ to evaluate certain compression algorithms on non-plane (and plane) trees. Indeed, for every internal node with two isomorphic subtrees we can replace the second identical subtree with a pointer to the first subtrees. We need about $n\dot{(}0.3725 \pm 10^{-4})$ bits to accomplish it. On the other hand, if we do this, we can save some storage on the replaced subtree. How much? Let us, as a preliminary assessment, compute the total size, $\mathrm{size(S_n)}$, of the saved isomorphic subtrees. Similar computations as above lead to

$$\mathbb{E}[\mathrm{size}(S_n)] = n\sum_{k=1}^{\lfloor (n+1)/2\rfloor} \frac{b_k}{(2k-1)(2k+1)} \approx n\cdot 0.4190.$$

This result can also be recovered from [6]. □

Finally we deal with the entropy of the non-plane tree $H(S_n)$ and its rate $h(s) = \lim_{n\to\infty} H(S_n)/n$. Observe that $H(S_n) = H(T_n) - H(T_n|S_n)$. In [8, 9] it was proved that

$$H(T_n) = \log_2(n-1) + 2n\sum_{k=2}^{n-1} \frac{\log_2(k-1)}{k(k+1)}. \tag{2.12}$$

Remark. We know that (see also [8])

$$2\sum_{k=2}^{n} \frac{\log_2(k-1)}{k(k+1)} \approx 1.736$$

for large n (we took $n = 10^6$). The above series converges slowly at the rate $O(\log n/n)$. Therefore one would prefer a more explicit formula for large n. We approximate the sum using the Euler-Maclaurin formula [14] by the integral

$$\begin{aligned}&\sum_{k=1}^{n-2} \frac{\log k}{(k+1)(k+2)} \sim \int_1^{n-2} \frac{\log(x)}{(x+1)(x+2)}dx\\ &= -\mathrm{Li}_2\left(1-\frac{n}{2}\right)\\ &+ \mathrm{Li}_2(2-n) + \log\left(2-\frac{2}{n}\right)\log(n-2)\\ &+ \mathrm{Li}_2\left(-\frac{1}{2}\right) + \frac{\pi^2}{12}\end{aligned}$$

where

$$\mathrm{Li}_2(x) = \int_1^x \frac{\log t}{1-t}dt$$

is the *dilogarithmic integral.* For large x the following holds [7]

$$\mathrm{Li}_2(x) = -\frac{1}{2}\log^2 x - \frac{\pi^2}{6} + O(\log x/x).$$

In fact, to get a better approximation of the sum we need two extra terms in the Euler-Maclaurin formula which leads to the following approximation

$$\begin{aligned}&\sum_{k=1}^{n-2} \frac{\log k}{(k+1)(k+2)}\\ &\approx \mathrm{Li}_2(3/2) + \frac{1}{12}\pi^2 + \frac{1}{2}\log^2(2) - \frac{1}{72} + \frac{23}{12960}\\ &= 0.868\ldots\end{aligned}$$

which matches the first three digits of the sum. □

To complete our analysis we now evaluate $H(T_n|S_n)$. From (2.3) we conclude

$$\begin{aligned}&H(T_n|S_n)\\ &= -\sum_{t\in\mathcal{T}_n, s\in\mathcal{S}_n} P(T_n = t, S_n = s)\log P(T_n = t|S_n = s)\\ &= \sum_{s\in\mathcal{S}_n} P(S_n = s)\log|[s]|\\ &= \sum_{t\in\mathcal{T}_n} P(T_n = t)(n-1-\mathrm{sym}(t))\\ &= n-1-\mathbb{E}[\mathrm{sym}(S_n)] = n-1-\mathbb{E}[\mathrm{sym}(T_n)]\end{aligned}$$

Thus from Theorem 2.1 we derive our third main result (see also [9]).

Theorem 2.2. *The entropy rate* $h(s) = \lim_{n\to\infty} H(S_n)/n$ *of the non-plane trees is*

$$h(s) = h(t) - h(t|s) \approx 1.109\ldots$$

where

$$\begin{aligned} h(t|s) &= 1 - \sum_{k=1}^{\infty} \frac{b_k}{(2k-1)k(2k+1)}, \\ h(t) &= 2\sum_{k=1}^{\infty} \frac{\log_2(k)}{(k+1)(k+2)} \end{aligned}$$

with $b_n = \sum_{t_n \in \mathcal{T}_n} p^2(t_n)$.

Remark 4. The entropy $h(s)$ is related to the Rényi entropy $h_1(t)$ of order 1 of non-plane trees. We define $h_1(t)$ as follows, if it exists:
(2.13)

$$h_1(t) = \lim_{n\to\infty} \frac{-\log \mathbb{E}[p(T_n)]}{n} = \lim_{n\to\infty} \frac{-\log \sum_{t_n\in\mathcal{T}_n} p^2(t_n)}{n}.$$

From the above we conclude that for large n

$$b_n = \sum_{t_n\in\mathcal{T}_n} p^2(t_n) \sim \exp(-nh_1(t)).$$

But appealing to (2.9)-(2.10) and applying the method discussed in [3] we observe that

$$b_n = \rho^n \left(6n - \frac{22}{5} + O(n^{-5})\right)$$

where $\rho = 0.3183843834378459\ldots$. Thus $h_1(t) = -\log(\rho)$. □

3 Analysis and Proofs

In this section we prove our main results.

3.1 Proof of Lemma 2.1

Observe that

$$\begin{aligned} F(u,z) &= \sum_{n=1}^{\infty}\sum_{t\in\mathcal{T}_n} P(T_n = t)u^{\mathrm{sym}(t)}z^{|t|} \\ &= z + \sum_{n\geq 2}\sum_{t\in\mathcal{T}_n} P(T_n = t)u^{\mathrm{sym}(t)}z^{|t|} \ . \end{aligned}$$

Recall that we write $p(t) = P(T = t)$. Since every tree $t \in \mathcal{T}_n$ for $n \geq 2$ can be divided into two trees, we have

$$\begin{aligned} &F(u,z) = z + \\ &\sum_{s,t\in\mathcal{T}} \frac{1}{|s|+|t|-1} p(s)p(t)u^{\mathrm{sym}(s)+\mathrm{sym}(t)+[|s=t|]} \cdot z^{|s|+|t|} \\ &= z + \sum_{s,t\in\mathcal{T}} \frac{1}{|s|+|t|-1} p(s)p(t)u^{\mathrm{sym}(s)+\mathrm{sym}(t)} \cdot z^{|s|+|t|} \\ &\quad - \sum_{t\in\mathcal{T}} \frac{1}{2|t|-1} p^2(t)u^{2\mathrm{sym}(t)} \cdot z^{2|t|} \\ &\quad + \sum_{t\in\mathcal{T}} \frac{1}{2|t|-1} p^2(t)u^{2\mathrm{sym}(t)+1} \cdot z^{2|t|} \\ &= z + \sum_{s,t\in\mathcal{T}} \frac{1}{|s|+|t|-1} p(s)p(t)u^{\mathrm{sym}(s)+\mathrm{sym}(t)} \cdot z^{|s|+|t|} \\ &\quad + (u-1)\sum_{t\in\mathcal{T}} \frac{1}{2|t|-1} p^2(t)u^{2\mathrm{sym}(t)} \cdot z^{2|t|} \ . \end{aligned}$$

Notice that, from the original definition of $F(u,z)$,

$$F(u,z)^2 = \sum_{s,t\in\mathcal{T}} p(s)p(t)u^{\mathrm{sym}(s)+\mathrm{sym}(t)}z^{|s|+|t|} \ ;$$

therefore,

$$\begin{aligned} &z\int_0^z \frac{F(u,w)^2}{w^2}dw \\ &= \sum_{s,t\in\mathcal{T}} \frac{1}{|s|+|t|-1} p(s)p(t)u^{\mathrm{sym}(s)+\mathrm{sym}(t)} \cdot z^{|s|+|t|}. \end{aligned}$$

Recall that

$$B(u,z) = \sum_{t\in\mathcal{T}} p^2(t)u^{\mathrm{sym}(t)}z^{|t|-1}.$$

Notice that

$$z\int_0^z B(u^2,w^2)dw = \sum_{t\in\mathcal{T}} \frac{1}{2|t|-1} p^2(t)u^{2\mathrm{sym}(t)} \cdot z^{2|t|} \ .$$

Hence

$$\begin{aligned} &F(u,z) \\ &= z + z\int_0^z \frac{F(u,w)^2}{w^2}dw + (u-1)z\int_0^z B(u^2,w^2)dw. \end{aligned}$$

Let $f(u,z) = \frac{F(u,z)}{z}$. From the last equation we get

$$\frac{\partial f(u,z)}{\partial z} = f(u,z)^2 + (u-1)B(u^2,z^2).$$

This proves Lemma 2.1. □

3.2 Average Symmetry: Proof of Theorem 2.1

We now compute $\mathbb{E}[\mathrm{sym}(T_n)] = \mathbb{E}[\mathrm{sym}(S_n)]$ using generating function

$$\phi_n(u) = [z^n]F(u,z) = [z^{n-1}]f(u,z)$$

defined above. We observe that

$$\mathbb{E}[\mathrm{sym}(T_n)] = \left.\frac{d\phi_n(u)}{du}\right|_{u=1}.$$

Moreover, we have that taking the coefficient (actually an integral) commutes with taking the derivative with respect to u:

$$\frac{d[z^{n-1}]f(u,z)}{du} = [z^{n-1}]f_u(u,z).$$

Thus, taking the derivative with respect to u in (2.4) gives (using (2.6))

$$\begin{aligned}(3.14)\qquad & f_{z,u}(u,z)\\ &= 2f(u,z)f_u(u,z)\\ &+ (u-1)\frac{\partial}{\partial u}\left[\left.(f(u,z)\boxdot f(u,z))\right|_{(u^2,z^2)}\right]\\ &+ \left[\left.(f(u,z)\boxdot f(u,z))\right|_{(u^2,z^2)}\right],\end{aligned}$$

and setting $u=1$ results in

$$\begin{aligned}f_{z,u}(1,z) &= 2f(1,z)f_u(1,z)\\ &+ \left[\left.(f(1,z)\boxdot f(1,z))\right|_{(u^2,z^2)}\right].\end{aligned}$$

Now, we use the fact that $\phi_n(1) = 1$ for $n \geq 1$ implies

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

We also recall the definition of the almost-OGF of $\mathbb{E}[\mathrm{sym}(T_n)]$:

$$\mathcal{E}(z) = \sum_{n=1}^{\infty} \mathbb{E}[\mathrm{sym}(T_n)]z^{n-1}$$

and note that

$$f_u(1,z) = \sum_{n=1} \mathbb{E}[\mathrm{sym}(T_n)]z^{n-1} = \mathcal{E}(z).$$

With the notation as in the Remark above Theorem 2.1 we find

(3.15)

$$\mathcal{E}'(z) = \frac{2\mathcal{E}(z)}{(1-z)} + \left.(A(z)\boxtimes A(z))\right|_{z^2} = \frac{2\mathcal{E}(z)}{z(1-z)} + B(z^2),$$

with $\mathcal{E}(0) = 0$, as needed.

We now solve (3.15). It is easy to see that

$$\mathcal{E}(z) = \frac{1}{(1-z)^2}\left(\int_0^z B(x^2)(1-x)^2dx + C\right).$$

But $\mathcal{E}(0) = 0$ implies $C = 0$. Thus

$$\begin{aligned}\mathbb{E}[\mathrm{sym}(T_n)] &= [z^{n-1}]\mathcal{E}(z)\\ &= [z^{n-1}]\frac{1}{(1-z)^2}\int_0^z B(x^2)(1-x)^2dx.\end{aligned}$$

We first observe, after some algebra, that

$$\begin{aligned}&\frac{1}{(1-z)^2}\int_0^z B(x^2)(1-x)^2dx\\ &= z + \sum_{n=1}^{\infty}\left(\frac{b_n+b_{n+1}}{2n+1}z^{2n+1} - \frac{b_n}{n}z^{2n}\right).\end{aligned}$$

Define

$$(3.16)\qquad c_k = \begin{cases} 0 & k = 0,\\ 1 & k = 1,\\ -\frac{b_\ell}{\ell} & k = 2\ell,\ \ell \geq 1,\\ \frac{b_\ell+b_{\ell+1}}{2\ell+1} & k = 2\ell+1,\ \ell \geq 1.\end{cases}$$

Then

$$\mathbb{E}[\mathrm{sym}(T_n)] = [z^{n-1}]\frac{1}{(1-z)^2}\sum_{k=0}^{\infty} c_k z^k = \sum_{k=0}^{n} c_k(n-k).$$

But it is easy to see that

$$\begin{aligned}\sum_{k=0}^{n} kc_k &= 1 + \sum_{\ell=1}^{\lfloor n/2\rfloor}(b_{\ell+1} - b_\ell)\\ &= 1 - b_1 + b_{\lfloor n/2\rfloor+1} = (-1)^{n+1}b_{\lfloor (n+1)/2\rfloor}.\end{aligned}$$

This leads to the final formula

$$(3.17)\qquad \mathbb{E}[\mathrm{sym}(T_n)] = n\sum_{k=1}^{n} c_k + (-1)^{n+1}b_{\lfloor (n+1)/2\rfloor}.$$

We can further simply this expression by computing $\sum_k c_k$. Some algebra is needed to show that

$$\begin{aligned}&\sum_k c_k\\ &= 1 - \frac{2}{3}b_1 +\\ &\sum_{\ell=2}^{\lfloor (n+1)/2\rfloor - 1} b_\ell\left(\frac{1}{2\ell-1} - \frac{1}{\ell} + \frac{1}{2\ell+1}\right)\\ &+ b_{\lfloor (n+1)/2\rfloor}\left(\frac{I(n \leq 2\ell+1)}{2\ell-1} + \frac{I(n \leq 2\ell+1)}{\ell}\right).\end{aligned}$$

The last term is $O(1/n)$, because $b_{\lfloor (n+1)/2 \rfloor} = O(1)$ and the indicators ensure that the only nonzero contributions of this term happen when $\ell \geq (n-1)/2$, so that $\ell = \Omega(n)$ and there are only $O(1)$ such terms. We thus get the following formula:

$$\sum_k c_k = \frac{1}{3} + \sum_{\ell=2}^{\lfloor (n+1)/2 \rfloor - 1} \frac{b_\ell}{(2\ell-1)\ell(2\ell+1)} + O(n^{-1}) \tag{3.18}$$

$$= \sum_{\ell=1}^{\lfloor (n+1)/2 \rfloor - 1} \frac{b_\ell}{(2\ell-1)\ell(2\ell+1)} + O(n^{-1}). \tag{3.19}$$

In summary, we have

$$\mathbb{E}[\mathrm{sym}(T_n)] = n \sum_{\ell=1}^{\lfloor (n+1)/2 \rfloor} \frac{b_\ell}{(2\ell-1)\ell(2\ell+1)} + O(1) \tag{3.20}$$

$$+ (-1)^{n+1} b_{\lfloor (n+1)/2 \rfloor} := nh(t|s) + O(1), \tag{3.21}$$

where, we recall, $b_\ell = \sum_{t_\ell} p^2(t_\ell)$. This proves Theorem 2.1.

References

[1] R. Baeza-Yates, R. Casas, J. Diaz, C. Martinez, On the average size of the intersection of binary trees, *SIAM J. Computing*, 21, 24-32, 1992

[2] M. Bona, P. Flajolet, Isomorphism and symmetries in random phylogenetic trees, *Journal of Applied Probability* 46(4):1005-1019.

[3] H-H Chern, M. Fernández-Camacho, H-K. Hwang, and C. Martinez, Psi-series method for equality of random trees and quadratic convolution recurrences, *Random Structures & Algorithms*, 44, 67-108, 2012.

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

[5] M. Drmota, *Random Trees: An Interplay between Combinatorics and Probability.* Springer Publishing Company, Inc., 2009.

[6] P. Flajolet, X. Gourdon, and C. Martinez, Patterns in random binary search trees, *Random Structures & Algorithms*, 11, 223-244, 1997

[7] M. Hassani, Approximation of the Dilogarithm Function, *J. Inequalities in Pure and Applied Mathematics*, 8, 1-7, 2007.

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

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

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

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

[12] M. Steel, A. McKenzie, Distributions of cherries for two models of trees. *Mathematical Biosciences*, 2000, 164:81-92.

[13] M. Steel, A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 2001, 170(1):91-112.

[14] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences.* John Wiley & Sons, Inc., New York, NY, 2001.

[15] 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.