A NOTE ON THE SCALING LIMITS OF RANDOM PÓLYA TREES

BERNHARD GITTENBERGER, EMMA YU JIN AND MICHAEL WALLNER*

Abstract. Panagiotou and Stufler (arXiv:1502.07180v2) recently proved one important fact on their way to establish the scaling limits of random Pólya trees: a uniform random Pólya tree of size n consists of a conditioned critical Galton-Watson tree $\mathcal{T}_n$ and many small forests, where with probability tending to one as n tends to infinity, each forest $F_n(v)$ is maximally of size $|F_n(v)| = O(\log n)$. Their proof used the framework of a Boltzmann sampler and deviation inequalities.

In this paper, first we employ a unified framework in analytic combinatorics to prove this fact with additional improvements on the bound of $|F_n(v)|$, namely $|F_n(v)| = \Theta(\log n)$. Second we give a combinatorial interpretation of all weights on the D-forests and C-trees in terms automorphisms associated to a given Pólya tree. Finally, we derive the limit probability that for a random node v the attached forest $F_n(v)$ is of a given size.

1. Introduction and main results

First we recall the asymptotic estimation of the number of Pólya trees with n nodes from the literatures [8, 9, 11]. Second we present Theorem 1 that leads to the proof of the scaling limits of random Pólya trees in [10].

1.1. **Pólya trees.** *Pólya trees* are unlabelled rooted trees considered up to symmetry. For any $\Omega \subseteq \mathbb{N}_0$ such that $0 \in \Omega$ and $\{0,1\} \neq \Omega$, we call a Pólya tree with outdegree set Ω *an* Ω*-Pólya tree* and note that a Pólya tree is an $\mathbb{N}_0$-Pólya tree, that is, the outdegree set is $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. The *size* of a tree is referred to the number of its nodes. We denote by t_n the number of Pólya trees of size n and by $T(z)$ the corresponding ordinary generating function. That is, $t_n = [z^n]T(z)$. From the Pólya enumeration theory [11] or the Burnside's Lemma, the generating function $T(z)$ satisfies

$$T(z) = z \exp\left(\sum_{i=1}^{\infty} \frac{T(z^i)}{i}\right). \tag{1.1}$$

By differentiating both sides of (1.1) with respect to z, one can derive a recurrence relation of t_n (see [8, Chapter 29] and [9]). The first few terms of $T(z)$ are then

$$T(z) = z + z^2 + 2z^3 + 4z^4 + 9z^5 + 20z^6 + 48z^7 + 115z^8 + 286z^9 + 719z^{10} + \cdots. \tag{1.2}$$

Pólya [11] showed that the radius of convergence $z = \rho$ of $T(z)$ satisfies $0 < \rho < 1$ and that $z = \rho$ is the only singularity on the circle of convergence $|z| = \rho$. Subsequently, Otter [9] proved that $T(\rho) = 1$ as well as the asymptotic expansion

$$T(z) = 1 - b\,(\rho - z)^{1/2} + c(\rho - z) + \mathcal{O}\left((\rho - z)^{3/2}\right). \tag{1.3}$$

*Correspondence author email: michael.wallner@tuwien.ac.at; The first author is partially supported by the Austrian Research Fund (FWF), grant SFB F50-03. The second author and the third author are fully supported the Austrian Research Fund (FWF), grant SFB F50-03.

By transfer theorems he derived

$$t_n = \frac{b\sqrt{\rho}}{2\sqrt{\pi}}\frac{\rho^{-n}}{\sqrt{n^3}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{1.4}$$

where $\rho \approx 0.3383219$, $b \approx 2.68112$ and $c = b^2/3 \approx 2.39614$.

We will see that $T(z)$ is connected with the exponential generating function of Cayley trees, which belongs to the class of *simply generated trees*. Simply generated trees have been introduced by Meir and Moon [7] to describe a weighted version of rooted trees. They are defined by the functional equation

$$y(z) = z\Phi(y(z)), \qquad \text{where} \qquad \Phi(z) = \sum_{j\geq 0}\phi_j\, z^j \tag{1.5}$$

is a power series with non-negative coefficients and $y(x) = \sum_{n\geq 1} y_n x^n$ is the generating function of *weighted* simply generated trees. One usually assumes that $\phi_0 > 0$ and $\phi_j > 0$ for some $j \geq 2$ to exclude the trivial cases. In particular *Cayley trees* are simply generated trees which are characterized by $\Phi(z) = \exp(z)$. For this purpose, let

$$C(z) = \sum_{n\geq 0} c_n \frac{z^n}{n!}$$

be the exponential generating function of Cayley trees of size n, which is also a generating function of weighted Cayley trees where every Cayley tree of size n has weight $1/n!$. Then, by construction it satisfies $C(z) = z\exp(C(z))$.

In order to analyze the dominant singularity of $T(z)$, we follow [9, 11], see also [4, Chapter VII.5], and we rewrite (1.2) into

$$T(z) = ze^{T(z)}D(z), \qquad \text{where} \qquad D(z) = \sum_{n\geq 0} d_n z^n = \exp\left(\sum_{i=2}^{\infty}\frac{T(z^i)}{i}\right). \tag{1.6}$$

We observe that $D(z)$ is analytic for $|z| < \sqrt{\rho} < 1$, so $\sqrt{\rho} > \rho$ and it follows that

$$T(z) = C(zD(z)) \tag{1.7}$$

is convergent for $|z| < \rho$. In fact, $T(z) = C(zD(z))$ is a case of super-critical composition schema where $T(z)$ becomes singular before reaching the singularity of $D(z)$. In other words, the dominant singularity of $T(z)$ is determined by the outer function $C(z)$. That is, $T(\rho) = 1$ or equivalently $\rho\, D(\rho) = e^{-1}$ where $z = e^{-1}$ is the unique dominant singularity of $C(z)$.

We set $d_n = [z^n]D(z)$ which counts the number of *weighted D-forest* of size n. These are forests of Pólya trees and the weights on each Pólya tree are assigned according to the function $D(z)$. From (1.2) and (1.6) one gets its first values

$$D(z) = \sum_{n=0}^{\infty} d_n z^n = 1 + \frac{1}{2}z^2 + \frac{1}{3}z^3 + \frac{7}{8}z^4 + \frac{11}{30}z^5 + \frac{281}{144}z^6 + \frac{449}{840}z^7 + \cdots. \tag{1.8}$$

Combinatorially, the composition (1.7) means that a Pólya tree is constructed from a "Cayley" tree where a weighted D-forest is attached to each node. We note that Cayley trees are labeled objects, while this composition (1.7) is on the level of unlabeled objects. Thus, we are not really working with Cayley trees, but with the Cayley function. The difference is that we interpret the exponential generating function of Cayley trees as a *weighted* ordinary generating function. To emphasize this fact we will from now on speak of *C-trees* in this context.

1.2. **Main results.** We will present Theorem 1 in the context of Pólya trees because it shows the main ideas in the proof without the notational and technical complications in the more general random Ω-Pólya trees.

Consider a random Pólya tree of size n, denoted by A_n, which is a Pólya tree that is uniformly selected from such trees with n vertices. We use $\mathcal{T}_n$ to denote the random C-tree that is contained in a random Pólya tree A_n. For every vertex v from $\mathcal{T}_n$, we use $F_n(v)$ to represent the D-forest that is attached to the vertex v in A_n. See Figure 1.1.

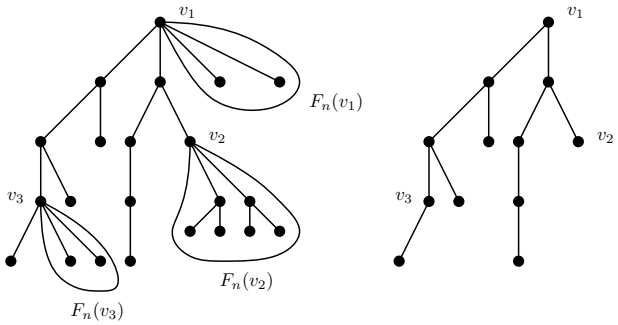

FIGURE 1.1. A random Pólya tree A_n (left) and a C-tree $\mathcal{T}_n$ (right) that is contained in A_n where all D-forests $F_n(v)$, except $F_n(v_1), F_n(v_2), F_n(v_3)$ are empty.

Let L_n be the maximal size of a D-forest contained in A_n, that is, $|F_n(v)| \le L_n$ holds for all $v \in \mathcal{T}_n$. The first theorem is stated as follows:

Theorem 1 (the upper bound is from (5.5) of [10])**.** *For* $0 < s < 1$,

$$(1 - (\log n)^{-s}) \left(\frac{-2 \log n}{\log \rho} \right) \le L_n \le (1 + (\log n)^{-s}) \left(\frac{-2 \log n}{\log \rho} \right) \tag{1.9}$$

holds with high probability $1 - o(1)$.

Our first main result is a new proof of Theorem 1 by applying the unified framework by Gourdon [5]. Our second main result is a combinatorial interpretation of all weights on the D-forests and C-trees in terms automorphisms associated to a given Pólya tree.

Let $c_{n,k}$ denote the total weight of C-trees of size k that are contained in the Pólya trees of size n, let $t_{c,n}(u)$ and $T_c(z,u)$ denote the corresponding generating function and bivariate generating function of $c_{n,k}$, that is,

$$t_{c,n}(u) = \sum_{k=1}^{n} c_{n,k} u^k \quad \text{and} \quad T_c(z,u) = \sum_{n \ge 0} t_{c,n}(u) z^n.$$

where $c_{n,k}$ is not an integer for general n, k. By marking the nodes of all C-trees in Pólya trees, we find a functional equation of the bivariate generating function $T_c(z,u)$, which is

$$T_c(z,u) = zu \exp\left(T_c(z,u)\right) \exp\left(\sum_{i=2}^{\infty} \frac{T(z^i)}{i} \right) = zu \exp\left(T_c(z,u)\right) D(z). \tag{1.10}$$

Our second main result is the following:

Theorem 2. *Let $\mathcal{T}$ be the set of all Pólya trees, and $\mathrm{MSET}^{(\geq 2)}(\mathcal{T})$ be the multiset of Pólya trees where each tree appears at least twice if it appears at all. Combinatorially this is a forest without unique trees. Then their weights d_n defined in (1.8) is equal to*

$$d_n = \sum_{\substack{F\in \mathrm{MSET}^{(\geq 2)}(\mathcal{T})\\ |F|=n}} \frac{|\{\sigma\in \mathrm{Aut}(F)\mid \sigma_1=0\}|}{|\mathrm{Aut}(F)|}$$

where $\mathrm{Aut}(F)$ is given in Definition 1. Furthermore, the polynomial of C-trees in Pólya trees of size n is given by

$$t_{c,n}(u)=\sum_{\substack{T\in\mathcal{T}\\ |T|=n}} t_T(u) \quad \text{where} \quad t_T(u)=\frac{1}{|\mathrm{Aut}(T)|}\sum_{\sigma\in\mathrm{Aut}(T)} u^{\sigma_1}.$$

In particular, for all $T\in\mathcal{T}$, it holds that $t_T'(1)=|\mathcal{P}(T)|$ where $\mathcal{P}(T)$ is the pointing operation on the tree T, that is, $\mathcal{P}(T)$ is the set of T with one single pointed (or colored) node.

Finally, we derive the limit probability that for a random node v the attached forest $F_n(v)$ is of a given size, which is consistent with the Boltzmann sampler from [10]. The precise statement of our third main result is the following:

Theorem 3. *The generating function $T^{[m]}(z,u)$ of Pólya trees, where each vertex is marked by z, and each weighted D-forest of size m is marked by u, is given by*

$$T^{[m]}(z,u)=C\left(uzd_mz^m+z\left(D(z)-d_mz^m\right)\right). \tag{1.11}$$

where $d_m=[z^m]D(z)$ and

$$\mathbb{P}\left[|F_n(v)|=m\right]=\frac{d_m\rho^m}{D(\rho)}\left(1+\mathcal{O}\left(n^{-1}\right)\right). \tag{1.12}$$

1.3. **Paper outline.** The paper is organized as follows. In Section 2 we prove Theorem 1 and discuss the size of the C-tree $\mathcal{T}_n$ in a random Pólya tree A_n. In Section 3 we prove Theorem 2 and 3. In Section 4 we conclude with final remarks.

2. The maximal size of a D-forest

We will use the generating function approach from [5] to analyze the maximal size L_n of D-forests in the random Pólya tree A_n, which provides a new proof of Theorem 1. Following the same approach, we can establish a central limit theorem of random variable $|\mathcal{T}_n|$, which has been done in [13] in the more general random $\mathcal{R}$-enriched trees.

Proof of Theorem 1. In (5.5) of [10], only an upper bound of L_n is given. By directly applying Gourdon's Theorem (Theorem 4 and Corollary 3 of [5]) for the super-critical composition schema, we find that for any positive m,

$$\mathbb{P}[L_n\leq m]=\exp(-\frac{c_1n}{m^{3/2}}\rho^{m/2})(1+\mathcal{O}(\exp(-m\varepsilon))),\qquad \text{where } c_1\sim\frac{b}{2\sqrt{\pi}(1-\sqrt{\rho})(D(\rho)+\rho D'(\rho))}.$$

Moreover, the random maximal size L_n satisfies asymptotically, as $n\to\infty$,

$$\mathbb{E}L_n=-\frac{2\log n}{\log\rho}-\frac{3}{2}\frac{2}{\log\rho}\log\log n+\mathcal{O}(1)\qquad\text{and}\qquad \mathrm{Var}\,L_n=\mathcal{O}(1).$$

By using Chebyshev's inequality, one can prove that L_n is highly concentrated around the mean $\mathbb{E}L_n$. We set $\varepsilon_n = (\log n)^{-s}$ where $0 < s < 1$ and we get

$$\mathbb{P}(|L_n - \mathbb{E}L_n| \geq \varepsilon_n \cdot \mathbb{E}L_n) \leq \frac{\mathbb{V}\mathrm{ar}\, L_n}{\varepsilon_n^2 \cdot (\mathbb{E}L_n)^2} = o(1),$$

which means that (1.9) holds with high probability $1 - o(1)$. □

It was shown in [13] that the size $|\mathcal{T}_n|$ of the C-tree $\mathcal{T}_n$ in A_n satisfies a central limit theorem and $|\mathcal{T}_n| = \Theta(n)$ holds with high probability $1 - o(1)$. The precise statement is the following.

Theorem 4 (Page 38, (3.9), (3.10) of [13] and (5.6) of [10]). *The size of the C-tree $|\mathcal{T}_n|$ in a random Pólya tree A_n of size n satisfies a central limit theorem where the expected value $\mathbb{E}|\mathcal{T}_n|$ and the variance $\mathbb{V}ar|\mathcal{T}_n|$ are asymptotically*

$$\mathbb{E}|\mathcal{T}_n| = \frac{2n}{b^2\rho}(1 + \mathcal{O}(n^{-1})), \quad \text{and} \quad \mathbb{V}ar|\mathcal{T}_n| = \frac{11n}{12b^2\rho}(1 + \mathcal{O}(n^{-1})). \tag{2.1}$$

Furthermore, for any s such that $0 < s < 1/2$, with high probability $1 - o(1)$ it holds that

$$(1 - n^{-s})\frac{2n}{b^2\rho} \leq |\mathcal{T}_n| \leq (1 + n^{-s})\frac{2n}{b^2\rho}. \tag{2.2}$$

The random Pólya trees belong to the class of random $\mathcal{R}$-enriched trees and we refer the readers to [13] for the proof of Theorem 4 in the general setting. Here we provide a proof of Theorem 4 to show the connection between bivariate generating function and normal distribution.

Proof of Theorem 4 (see also [13]). It follows from Theorem 2.23 in [2] that the random variable $|\mathcal{T}_n|$ satisfies a central limit theorem. In the present case, we set $F(z, y, u) = zu\exp(y)D(z)$. It is easy to verify that $F(z, y, u)$ is an analytic function in z, y around 0 such that $F(0, y, u) \equiv 0$, $F(x, 0, u) \not\equiv 0$ and all coefficients $[z^n y^m]F(z, y, 1)$ are real and non-negative. From Theorem 2.23 in [2] we know that $T_c(z, u)$ is the unique solution of the functional identity $y = F(z, y, u)$. Since all coefficients of $F_y(z, y, 1)$ are non-negative and the coefficients of $T(z)$ are positive as well as monotonically increasing, this implies $(\rho, T(\rho), 1)$ is the unique solution of $F_y(z, y, 1) = 1$, which leads to the fact that $T(\rho) = 1$. Moreover, the expected value is

$$\mathbb{E}|\mathcal{T}_n| = \frac{nF_u(z, y, u)}{\rho F_z(z, y, u)} = \frac{[z^n]\partial_u T_c(z, u)|_{u=1}}{[z^n]T(z)} = \left([z^n]\frac{T(z)}{1 - T(z)}\right)([z^n]T(z))^{-1} = \frac{2n}{b^2\rho}(1 + \mathcal{O}(\frac{1}{n})).$$

The asymptotics are directly derived from (1.3). Likewise, we can compute the variance

$$\begin{aligned}\mathbb{V}\mathrm{ar}\,|\mathcal{T}_n| &= \frac{[z^n]\partial_u^2 T(z, u)|_{u=1}}{[z^n]T(z)} + \mathbb{E}\,|\mathcal{T}_n| - (\mathbb{E}\,|\mathcal{T}_n|)^2 \\ &= \frac{[z^n]T(z)(1 - T(z))^{-3}}{[z^n]T(z)} - (\mathbb{E}\,|\mathcal{T}_n|)^2 = \frac{11n}{12b^2\rho}(1 + O(\frac{1}{n})).\end{aligned}$$

Furthermore, $|\mathcal{T}_n|$ is highly concentrated around $\mathbb{E}\,|\mathcal{T}_n|$, which can be proved again by using Chebyshev's inequality. We set $\varepsilon_n = n^{-s}$ where $0 < s < 1/2$ and consequently we get

$$\mathbb{P}(|\mathcal{T}_n| - \mathbb{E}|\mathcal{T}_n|| \geq \varepsilon_n \cdot \mathbb{E}|\mathcal{T}_n|) \leq \frac{\mathbb{V}\mathrm{ar}|\mathcal{T}_n|}{\varepsilon_n^2 \cdot (\mathbb{E}|\mathcal{T}_n|)^2} = \mathcal{O}(n^{2s-1}) = o(1),$$

which yields (2.2). □

As a simple corollary, we also get the total size of all weighted D-forests in A_n. Let $\mathcal{D}_n$ denote the union of all D-forests in a random Pólya tree A_n of size n.

Corollary 5. *The size of weighted D-forests in a random Pólya tree of size n satisfies a central limit theorem where the expected value $\mathbb{E}|\mathcal{D}_n|$ and the variance $\mathbb{V}ar|\mathcal{D}_n|$ are asymptotically*

$$\mathbb{E}|\mathcal{D}_n| = n\left(1-\frac{2}{b^2\rho}\right)(1+\mathcal{O}(n^{-1})), \quad \text{and} \quad \mathbb{V}ar|\mathcal{D}_n| = \frac{11n}{12b^2\rho}(1+\mathcal{O}(n^{-1})). \tag{2.3}$$

Theorem 4 and Corollary 5 tell us that a random Pólya tree A_n consists on average out of $\frac{2}{b^2\rho}\approx 82.2\%$ C-tree $\mathcal{T}_n$ and $1-\frac{2}{b^2\rho}\approx 17.8\%$ weighted D-forests $\mathcal{D}_n$. Furthermore, the average size of a weighted D-forest $F_n(v)$ attached to a random C-tree vertex in A_n is $\frac{b^2\rho}{2}-1\approx 0.216$, which indicates that on average the D-forest $F_n(v)$ is very small, although the maximal size of all D-forests in a random Pólya tree A_n reaches $\Theta(\log n)$.

Remark 1. *We will describe the connection of (1.7) to the Boltzmann sampler in [10]. We know that $F(z,y,1)=z\Phi(y)D(z)$ where $\Phi(x)=\exp(x)$ and $y=T(z)$. By dividing both sides of this equation by $y=T(z)$, one obtains from (1.6) that*

$$1=\frac{zD(z)}{T(z)}\exp(T(z))=\exp(-T(z))\sum_{k\geq 0}\frac{T^k(z)}{k!}$$

which implies that in the Boltzmann sampler $\Gamma T(x)$, the number of offspring contained in the C-tree $\mathcal{T}_n$ is Poisson distributed with parameter $T(x)$. As an immediate result, this random C-tree $\mathcal{T}_n$ contained in the Boltzmann sampler $\Gamma T(\rho)$ is a critical Galton-Watson tree since the expected number of offspring is $F_y(z,y,1)=1$ which holds only when $(z,y)=(\rho,1)$.

3. D-forests and C-trees

In order to get a better understanding of D-forests and C-trees, we need to return to the original proof of Pólya on the number of Pólya trees [11]. The important step is the treatment of tree automorphisms by the cycle index. Let us recall what it means that two graphs are isomorphic.

Definition 1. Two graphs G_1 and G_2 are *isomorphic* if there exists a bijection between the vertex sets of G_1 and G_2,

$$f: V(G_1)\mapsto V(G_2)$$

such that two vertices v and w of G_1 are adjacent if and only if $f(v)$ and $f(w)$ are adjacent in G_2. If $G_1=G_2$ we call the bijection f an *automorphism*. The automorphism group on the graph G_1 is denoted by $\mathrm{Aut}(G_1)$.

For any permutation σ, let σ_i be the number of cycles of length i of σ. We define the *type* of σ, to be the sequence $(\sigma_1,\sigma_2,\dots,\sigma_k)$ if $\sigma\in S_k$. Note that $k=\sum_{i=1}^k i\sigma_i$.

Definition 2 (Cycle index)**.** Let G be a subgroup of the symmetric group S_k. Then, the *cycle index* is

$$Z(G;s_1,s_2,\dots,s_k)=\frac{1}{|G|}\sum_{\sigma\in G}s_1^{\sigma_1}s_2^{\sigma_2}\cdots s_k^{\sigma_k}. \tag{3.1}$$

Now we are ready to prove Theorem 2.

Proof of Theorem 2. From the Pólya enumeration theory [11] or the Burnside's Lemma, the generating function $T(z)$ satisfies the functional equation:

$$\begin{aligned} T(z) &= z\sum_{k\geq 0} Z(S_k; T(z), T(z^2), \dots, T(z^k)) \\ &= z\sum_{k\geq 0}\frac{1}{k!}\sum_{\sigma\in S_k}(T(z))^{\sigma_1}(T(z^2))^{\sigma_2}\cdots(T(z^k))^{\sigma_k}. \end{aligned} \tag{3.2}$$

which, by a simple calculation, can be simplified to (1.1), the starting point of our research. However, this shows that the generating function of D-forests from (1.6) is given by

$$D(z) = \exp\left(\sum_{i=2}^{\infty}\frac{T(z^i)}{i}\right) = \sum_{k\geq 0} Z(S_k; 0, T(z^2), \dots, T(z^k)) = \sum_{k\geq 0}\frac{1}{k!}\sum_{\substack{\sigma\in S_k\\ \sigma_1=0}}(T(z^2))^{\sigma_2}\cdots(T(z^k))^{\sigma_k}.$$

This representation enables us to interpret the weights d_n of D-forests of size n: A D-forest of size n is a multiset of k Pólya trees, where every tree occurs at least twice. Its weight is given by the ratio of fix point free automorphisms over the total number of automorphisms. (Equivalently, it is given by the number of fix point free permutations $\sigma \in S_k$ of these trees rescaled by the total number of orderings $k!$.)

Let $\mathfrak{T}$ be the set of all Pólya trees, and $\mathrm{MSET}^{(\geq 2)}(\mathfrak{T})$ be the multiset of Pólya trees where each tree appears at least twice if it appears at all. Combinatorially this is a forest without unique trees. Then their weights are given by

$$d_n = \sum_{\substack{F\in \mathrm{MSET}^{(\geq 2)}(\mathfrak{T})\\ |F|=n}} \frac{|\{\sigma\in \mathrm{Aut}(F) \mid \sigma_1 = 0\}|}{|\mathrm{Aut}(F)|}.$$

Example 1. The smallest D-forest is of size 2, and it consists of a pair of single nodes. There is just one fix point free automorphism on this forest, thus $d_2 = 1/2$. For $n = 3$ the forest consists of 3 single nodes. The fix point free permutations are the 3 cycles, thus $d_3 = 2/6 = 1/3$. The case $n = 4$ is more interesting. A forest consists either of 4 single nodes, or of 2 identical trees, each consisting of 2 nodes and one edge. In the first case we have 6 4-cycles and 3 pairs of two involutions. In the second case we have 1 involution swapping the two trees. Thus, $d_4 = \frac{6+3}{24} + \frac{1}{2} = \frac{7}{8}$.

These results also yield a natural interpretation of C-trees. We recall that by definition

$$T_c(z, u) = \sum_{n\geq 0} t_{c,n}(u) z^n,$$

where $t_{c,n}(u) = \sum_k c_{n,k}u^k$ is the polynomial marking the C-trees in Pólya trees of size n. From the decomposition (1.7) and (1.10) we get the first few terms

$$\begin{aligned} t_{c,1}(u) &= u, & t_{c,2}(u) &= u^2, \\ t_{c,3}(u) &= \frac{3}{2}u^3 + \frac{1}{2}u, & t_{c,4}(u) &= \frac{8}{3}u^4 + u^2 + \frac{1}{3}u. \end{aligned}$$

Evaluating these polynomials at $u = 1$ obviously returns $t_{c,n}(1) = t_n$ the number of Pólya trees. Their coefficients however are weighted sums depending on the number of C-tree nodes. A C-tree node is a node, that does not belong to a decoration. For a given Pólya tree there are in general several ways to decide what is a C-tree node and what is a D-node. The possible

choices are encoded in the automorphisms of the tree and these are responsible for the above weights.

Let T be a Pólya tree, and $\mathrm{Aut}(T)$ be its automorphism group. For an automorphism $\sigma \in \mathrm{Aut}(T)$ the nodes which are fix points by σ are C-tree nodes. All other nodes are part of D-forests. Summing over all automorphisms and normalizing by the total number gives the C-tree generating polynomial for T.

$$t_T(u) = Z(\mathrm{Aut}(T); u, 1, \dots, 1) = \frac{1}{|\mathrm{Aut}(T)|} \sum_{\sigma \in \mathrm{Aut}(T)} u^{\sigma_1}. \tag{3.3}$$

The polynomial of C-trees in Pólya trees of size n are then given by

$$t_{c,n}(u) = \sum_{\substack{T \in \mathcal{T} \\ |T|=n}} t_T(u).$$

Example 2. For $n = 3$ we have 2 Pólya trees, namely the sequence T_1 and a cherry T_2. Thus, $\mathrm{Aut}(T_1) = \{\mathrm{id}\}$, where $\mathrm{Aut}(T_2) = \{\mathrm{id}, \sigma\}$, where σ swaps the two leaves and the root is unchanged. Thus,

$$t_{T_1}(u) = u^3, \qquad t_{T_2}(u) = \frac{1}{2}(u^3 + u).$$

For $n = 4$ we have 4 Pólya trees shown in Figure 3.1. Their automorphism groups are given by $\mathrm{Aut}(T_1) = \mathrm{Aut}(T_2) = \{\mathrm{id}\}$, $\mathrm{Aut}(T_3) = \{\mathrm{id}, (v_3\, v_4)\} \cong S_2$, and

$$\mathrm{Aut}(T_4) = \{\mathrm{id}, (v_2\, v_3), (v_3\, v_4), (v_2\, v_4), (v_2\, v_3\, v_4), (v_2\, v_4\, v_3)\} \cong S_3.$$

This gives

$$t_{T_1}(u) = t_{T_2}(u) = u^4, \qquad t_{T_3}(u) = \frac{1}{2}(u^4 + u^2), \qquad t_{T_4}(u) = \frac{1}{6}(u^4 + 3u^2 + 2u).$$

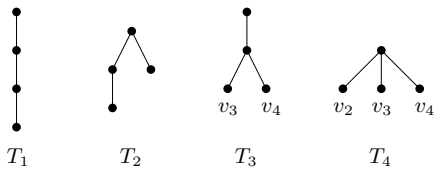

FIGURE 3.1. All Pólya trees of size 4

In the same way as we got the composition scheme in (1.7), we can rewrite $T_c(z, u)$ from (1.10) as $T_c(z, u) = C(uzD(z))$. The expected total weight of all C-trees contained in all Pólya trees of size n is the n-th coefficient of $T_c(z)$, which is,

$$T_c(z) := \frac{\partial}{\partial u} T_c(z, u)\Big|_{u=1} = \frac{T(z)}{1 - T(z)} = z + 2z^2 + 5z^3 + 13z^4 + 35z^5 + 95z^6 + 262z^7 + \cdots. \tag{3.4}$$

Let us explain why these numbers are integers, although the coefficients of $t_{c,n}(u)$ are in general not. We will show an even stronger result. Recall that $\mathcal{P}(T)$ is the pointing operation on the tree T, that is, $\mathcal{P}(T)$ is the set of T with one single pointed (or colored) node.

Lemma 6. *For all $T \in \mathcal{T}$ it holds that $t'_T(1) = |\mathcal{P}(T)|$.*

Proof. From (3.3) we get that $t'_T(1) = \sum_{\sigma \in \mathrm{Aut}(T)} \frac{\sigma_1}{|\mathrm{Aut}(T)|}$ is the expected number of fixed points in a uniformly at random chosen automorphism on T. Let X_T be the associated random variable for the number of fixed points in a random chosen automorphism of T. We will prove $\mathbb{E}X_T = |\mathcal{P}(T)|$ by induction on the size of T.

The most important observation is, that only if the root of a subtree is a fixed point, its children can also be fixed points. Obviously the root of the tree is always a fixed point.

For $|T| = 1$, it obviously holds as $\mathbb{E}(X_T) = 1$ and there is just one tree with a single node and a marker on it. For larger T consider the construction of Pólya trees. A Pólya tree consists of a root and its children, which are a multiset of smaller trees. Thus, the children set is of the form

$$\{T_{1,1}, \dots, T_{1,k_1}, T_{2,1}, \dots, T_{2,k_2}, \dots, T_{r,1}, \dots, T_{r,k_r}\}, \quad \text{with} \quad T_{i,j} \in \mathcal{T}.$$

On the level of children, the possible behaviors of automorphisms are permutations within the same class of trees. In other words, an automorphism may interchange the trees $T_{1,1}, \dots, T_{1,k_1}$ in $k_1!$ many ways, etc. Here the main observation comes into play: only subtrees, of which the root is a fixed point, might also have other fixed points. Thus, the expected number of fixed points are given by the expected number of fixed points in a random permutation of S_{k_i} times the expected number of fixed points in T_{k_i}. By linearity of expectation we get

$$\mathbb{E}(X_T) = \sum_{i=0}^{r} \underbrace{\mathbb{E}(\text{Fixed points in } S_{k_i})}_{=1} \mathbb{E}(X_{T_i}),$$

where $\mathbb{E}(X_{T_i}) = \mathbb{E}(X_{T_{i,j}})$ for all $1 \le j \le k_i$ and $X_{T_0} = 1$ because the root is a fixed point of any automorphism. Since the expected number of fixed points for each permutations is 1, we get on average 1 representative for each class of trees. This is exactly the operation of labeling one tree among each equivalent class. Finally, by induction the claim holds. □

In view of Lemma 6, we complete the proof of Theorem 2. □

As an immediate consequence of Lemma 6, $t'_{c,n}(1)$ counts the number of Pólya trees with n nodes and a single labeled node (see OEIS A000107, [12]). This also explains the construction of non-empty sequences of trees in (3.4): Following the connection of [1, p. 61, 62] one can draw a path from the root to each labeled node. The nodes on the path are the roots of a sequence Pólya trees.

Remark 2. *Note that Lemma 6 also implies that the total number of fixed points in all automorphisms of a tree is a multiple of the number of automorphisms.*

Remark 3. *Lemma 6 can also be proved by considering the cycle-pointed Pólya trees; see subsection 3.2 of [6] for a full description. Let (T, c) be a cycle-pointed structure considered up to symmetry where T is a Pólya tree and c is a cycle of an automorphism $\sigma \in \mathrm{Aut}(T)$. Then, the number of such cycle-pointed structures (T, c) where c has length 1, is exactly the number $t'_T(1)$.*

Let us analyze the D-forests in A_n more carefully. We want to count the number of D-forests that have size m in a random Pólya tree A_n, so we label such D-forest with additional weight u in (1.7). From the bivariate generating function (1.11) we can recover the probability $\mathbb{P}[|F_n(v)| = m]$ to generate a D-forest of size m in the Boltzmann sampler from [10].

Proof of Theorem 3. The first result is a direct consequence of (1.7), where only vertices with weighted D-forests of size m are marked. For the second result we differentiate both

sides of (1.11) and get

$$T_u^{[m]}(z,1) = \frac{T(z)}{1-T(z)}\frac{d_m z^m}{D(z)} = T_c(z)\frac{d_m z^m}{D(z)}.$$

Then, the sought probability is given by

$$\mathbb{P}\left[|F_n(v)| = m\right] = \frac{[z^n]T_u^{[m]}(z,1)}{[z^n]T_c(z)} = \frac{d_m\rho^m}{D(\rho)}\left(1+\mathcal{O}\left(n^{-1}\right)\right).$$

For the last equality we used the fact that $D(z)$ is analytic in a neighborhood of $z=\rho$.

Let $P_n(u)$ be the probability generating function for the size of a weighted D-forest $F_n(v)$ attached to a vertex v of $\mathcal{T}_n$ in a random Pólya tree A_n. From the previous Theorem it follows that

$$P_n(u) = \sum_{m\geq 0}\frac{[z^n]T_u^{[m]}(z,1)}{[z^n]T_c(z)}u^m = \frac{[z^n]T_c(z)\frac{D(zu)}{D(z)}}{[z^n]T_c(z)} = \frac{D(\rho u)}{D(\rho)}\left(1+\mathcal{O}\left(n^{-1}\right)\right).$$

This is exactly (5.2) in [10] when $\Omega=\mathbb{N}_0$. □

Summarizing, we state the asymptotic probabilities that a weighted D-forest $F_n(v)$ in A_n has size equal to or greater than m.

m	0	1	2	3	4	5	6	7
$\mathbb{P}[\lvert F_n(v)\rvert = m] \approx$	0.9197	0.0000	0.0526	0.0119	0.0105	0.0015	0.0027	0.0003
$\mathbb{P}[\lvert F_n(v)\rvert \geq m] \approx$	1.0000	0.0803	0.0803	0.0277	0.0161	0.0060	0.0041	0.0014

TABLE 1. The probability that a weighted D-forest $F_n(v)$ has size equal to or greater than m when $0 \leq m \leq 7$.

4. Conclusion and perspectives

In this paper we provide an alternative proof of the maximal size of D-forests in a random Pólya tree, we interpret all weights on D-forests and C-trees in terms of automorphisms associated to a Pólya tree, and we derive the limiting probability that for a random node v the attached D-forest $F_n(v)$ is of a given size.

Our work can be extended to Ω-Pólya with more technical details. In view of the connection between Boltzmann sampler and generating functions, it comes as no surprise that the 'colored' Boltzmann sampler from [10] is closely related to a bivariate generating function. But certainly the unified frameworks in analyzing the (bivariate) generating functions offer stronger results on the limiting distributions of the size of the C-trees and the maximal size of D-forests.

Now the door is open to studying shape characteristics of D-forests. The C-tree is the Cayley tree within a Polya tree and thus already well-known.

References

[1] F. Bergeron, G. Labelle and P. Leroux, *Combinatorial Species and Tree-Like Structures*, Camb. 1998.
[2] M. Drmota, *Random trees, An Interplay between Combinatorics and Probability*, Springer Verlag (2008).
[3] M. Drmota and B. Gittenberger, *The distribution of nodes of given degree in random trees*, Journal of Graph Theory, 31:227-253, 1999.
[4] P. Flajolet and R. Sedgewick, Analytic Combinatorics, Cambridge University Press, 2009.

[5] X. Gourdon, *Largest component in random combinatorial structures*, Discrete Mathematics, 180, 185-209, 1998.
[6] M. Bodirsky, É. Fusy, M. Kang and S. Vigerske, *Boltzmann samplers, Pólya theory, and cycle pointing*, SIAM J. Comput., 40(3), 721769, 2011.
[7] A. Meir and J.W. Moon, *On the altitude of nodes in random trees*, Canad. J. Math., 30(5) (1978), 997-1015.
[8] A. Nijenhuis and H.S. Wilf, *Combinatorial algorithms*, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, second edition, 1978. For computers and calculators, Computer Science and Applied Mathematics.
[9] R. Otter, *The number of trees*, Ann. of Math., 49(2)(1948), 583-599.
[10] K. Panagiotou and B. Stufler, *Scaling limits of random Pólya trees*, arXiv preprint arXiv:1502.07180v2, 2015.
[11] G. Pólya, *Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen*, Acta Mathematica (68)(1), 145-254, 1937.
[12] N.J.A. Sloane, *The on-line Encyclopedia of integer sequences (OEIS)*.
[13] B. Stufler, *Random enriched trees with applications to random graphs*, arXiv:1504.02006v6, 2015.

Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien, Wiedner Hauptstr. 810, 1040 Vienna, Austria
E-mail address: bgitten@dmg.tuwien.ac.at
E-mail address: yu.jin@tuwien.ac.at
E-mail address: michael.wallner@tuwien.ac.at