SOME DISTRIBUTIONS ON FINITE ROOTED BINARY TREES

SEAN CLEARY, MAREIKE FISCHER, ROBERT C. GRIFFITHS,
AND RAAZESH SAINUDIIN

ABSTRACT. We introduce some natural families of distributions on rooted binary ranked plane trees with a view toward unifying ideas from various fields, including macroevolution, epidemiology, computational group theory, search algorithms and other fields.

1. INTRODUCTION

We study some families of distributions on $\widehat{\mathcal{T}}_n^{\downarrow}$, the set of *rooted binary ranked plane trees* with n unlabeled terminal nodes. $\widehat{\mathcal{T}}_n^{\downarrow}$ and their equivalence classes represent various binary tree spaces that are encountered under a myriad of names across several mathematical sciences. Our main objective is to introduce existing and novel families of distributions, based on recursively constructive randomized tree-generation algorithms, whereby the trees are grown from the root node by splitting one of the existing leaf nodes according to a probabilistic scheme, to generate distributions on $\widehat{\mathcal{T}}_{1:n}^{\downarrow} := \bigcup_{k=1}^{n} \widehat{\mathcal{T}}_k^{\downarrow}$ and their equivalence classes.

The plan of the paper is as follows. In Section 2 we give a brief introduction to four main classes of finite rooted binary trees and a recursive randomized construction scheme for the finest class of trees that are in bijective correspondence to the permutations. In Section 3 we revisit the class of trees and their recursive construction using other representations. In Section 4 we introduce a nonparametric Markov splitting model on the finest class of ranked plane trees both directly and indirectly through their bijective correspondence with permutations or with dyadic partitions. Specific examples of the Markov splitting model are further characterized by split-exchangeability and plane-invariance in order to readily obtain probabilities over various equivalence classes of trees in phylogenetics, epidemiology and group theory.

2. CLASSES OF FINITE ROOTED BINARY TREES

2.1. Preliminaries. Recall that a *rooted tree*, in the abstract graph-theoretic sense, is a connected acyclic graph with a specific node distinguished as the *root*. The *size* of tree is given by the number of its nodes. A *finite tree* has finitely many nodes. In a rooted tree, the *outdegree* of a node is the number of its descendants. A *leaf* is a node of a tree without any descendants. Non-leaf nodes are also called *internal nodes* and leaf nodes are also called terminal nodes. In a *rooted binary tree*, every internal node has two descendants. All trees in this study are finite, rooted and binary and hence we do not explicitly mention this. Thus, unless stated otherwise,

---

by a tree we mean a finite rooted binary tree. Let $\mathcal{T}_n$ denote the set of all such trees with n leaf nodes.

A tree whose nodes are labeled by distinct elements of a non-empty label set is a *labeled tree*. We can have a *semi-labeled* tree when only a subset of its nodes are labeled by distinct elements. Node labels are assigned by a *labeling function* from a set of nodes to a set of labels. For a formal treatment of semi-labeled trees in a phylogenetic setting see [28, Dfn. 2.1.1–2]. For example, if only the leaf nodes are labeled we get a *leaf-labeled tree* and if only the internal nodes are labeled we get an *internal-labeled tree*. An *internal-ranking* is a labeling function from the $n-1$ internal nodes of a tree to the set of integers in $[n-1] := \{1, 2, \ldots, n-1\}$, which satisfies the following requirements. The root node has label or rank 1, and if v is an interior node which is on the path from an interior node w to a leaf node, then the label or rank of w is less than that of v. A tree together with an internal-ranking gives a *ranked tree*. Let $\mathcal{T}_n^{\downarrow}$ denote the set of all such ranked trees with n leaf nodes. Such trees are also known as *increasing trees* [14].
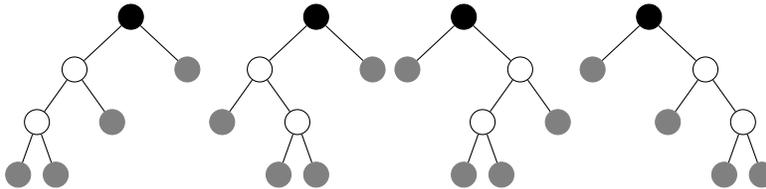
FIGURE 1. Four distinct plane trees that represent the same tree with 4 leaf nodes. In each tree, the root node is solid black, all other internal nodes are white with black boundary and leaf nodes are solid gray.

By superimposing additional structure on graph-theoretic trees we obtain trees that are known as plane, planar, oriented or ordered trees. A *plane tree* is defined as a tree in which subtrees dangling from a common node are ordered between themselves and represented from left to right in order. In addition to labels, nodes may also have *addresses* to encode their planar embedding or orientation. It is convenient to assign addresses to the nodes of a plane tree to encode their planar embedding using strings formed by concatenation of L's and R's in the obvious manner as shown in Figure 2. Note that all four plane trees in Figure 1 are subtrees of the tree in Figure 2 with addressed nodes. The addresses of the nodes of plane trees are clear from the planar drawing and usually not shown as in Figure 1. A finite rooted binary tree that is plane is called a *plane tree* in this work. Figure 1 shows four distinct plane trees that represent the same (non-plane) tree. They are known as *plane binary trees* in enumerative combinatorics [33, Ex. 6.19(d), p. 220], *finite, rooted binary trees* in geometric group theory [22, Ch. 10], or *binary search trees* in computer science [19]. They are less well known in evolutionary biology and may be referred to as *rooted binary unranked oriented tree shapes* by a natural extension of phylogenetic notions in [28, Section 2.4]. Let $\widehat{\mathcal{T}}_n$ denote the set of all such plane trees with n leaf nodes or equivalently with $n-1$ internal nodes.

We obtain a *ranked plane tree* when an internal ranking is used to assign rank labels from $[n-1]$ to the $n-1$ internal (addressed) nodes of a plane tree with n leaf nodes. We emphasize that the integer rank label associated with each node of
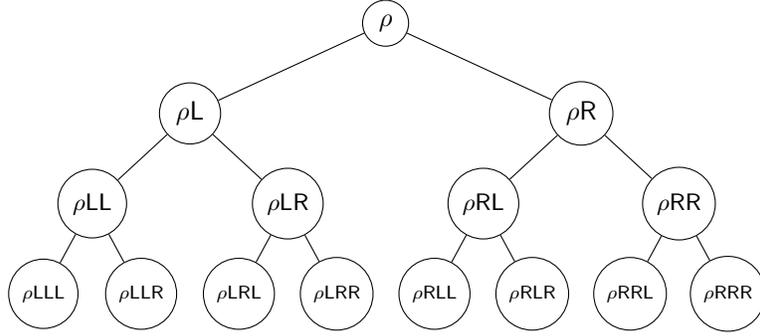
FIGURE 2. Addresses of nodes in plane trees.

a ranked plane tree is in addition to the node's address given by a string of L's and R's. Let $\widehat{\mathcal{T}}_n^{\downarrow}$ denote the set of all ranked plane trees with n leaf nodes.

2.2. **Ranked Plane Trees.** We can see the rank labels or ranks of a ranked plane tree $\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow}$ as the result of a recursive splitting process of Construction 2.1 where at the k-th step an unlabeled leaf node of $\hat{t}_k^{\downarrow} \in \widehat{\mathcal{T}}_k^{\downarrow}$ is split and thus transformed into an internal node with rank k in order to produce a ranked plane tree $\hat{t}_{k+1}^{\downarrow} \in \widehat{\mathcal{T}}_{k+1}^{\downarrow}$ with k internal nodes (that have been split) and $k+1$ unlabeled leaf nodes. Thus, the ranks encode the splitting order and thereby give the entire history of the process that recursively created the tree $\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow}$ after $n-1$ recursive splits starting from the root node in $\widehat{\mathcal{T}}_1^{\downarrow}$.

One can assign a random variable I_k over the current set of k leaf nodes (indexed by their addresses) of a ranked plane tree and choose a random leaf node according to I_k for the next split. This can be used to recursively generate random trees in $\widehat{\mathcal{T}}_{1:n}^{\downarrow} := \bigcup_{k=1}^n \widehat{\mathcal{T}}_k^{\downarrow}$ with up to n leaves starting from the root node on the basis of the random variables $I_{1:n-1} := \{I_1, I_2, \ldots, I_{n-1}\}$ as follows.

Construction 2.1 (Ranked Plane Trees). Consider the following process:

- Initialize:
  - $i \leftarrow 1$, set counter
  - let $\hat{t}_1^{\downarrow} = (V, E) = (\{\rho\}, \{\})$ be a ranked plane tree which consists only of a single root node ρ in V, which we consider an unlabeled leaf node, and no edges in E.
- Randomize: Choose a leaf node ℓ of $\hat{t}_i^{\downarrow}$ according to the random variable I_i that may depend on $\hat{t}_i^{\downarrow}$.
- Split ℓ:
  - label ℓ by rank i,
  - attach two new leaf nodes ℓL and ℓR to the left and the right of ℓ, respectively – i.e. introduce nodes ℓL and ℓR, and edges $(\ell, \ell$L$)$ and $(\ell, \ell$R$)$ to $\hat{t}_i^{\downarrow}$,
  - $i \leftarrow i + 1$, increment counter
- Repeat: Go to the Randomize step if i, the number of leaf nodes in $\hat{t}_i^{\downarrow}$, is less than a given $n \in \mathbb{N}$.

The distribution of the random ranked plane tree $\hat{t}_n^\downarrow$ produced by the recursive splitting process in Construction 2.1 is determined by $I_{0:n-1}$. See Figure 3(A) for the Hasse diagram on $\widehat{\mathcal{T}}_{1:4}^\downarrow$ when $n = 4$. By ignoring planarity but not the ranks in $\widehat{\mathcal{T}}_{1:n}^\downarrow$ we get random ranked trees (Figure 3(B)). Ignoring the ranks (internal node labels) in $\widehat{\mathcal{T}}_{1:n}^\downarrow$ gives random plane trees (Figure 3(C)). Finally by ignoring planarity in addition to the ranks in $\widehat{\mathcal{T}}_{1:n}^\downarrow$ we get random trees (Figure 3(D)).

(A) ranked plane trees

(B) ranked trees

(C) plane trees
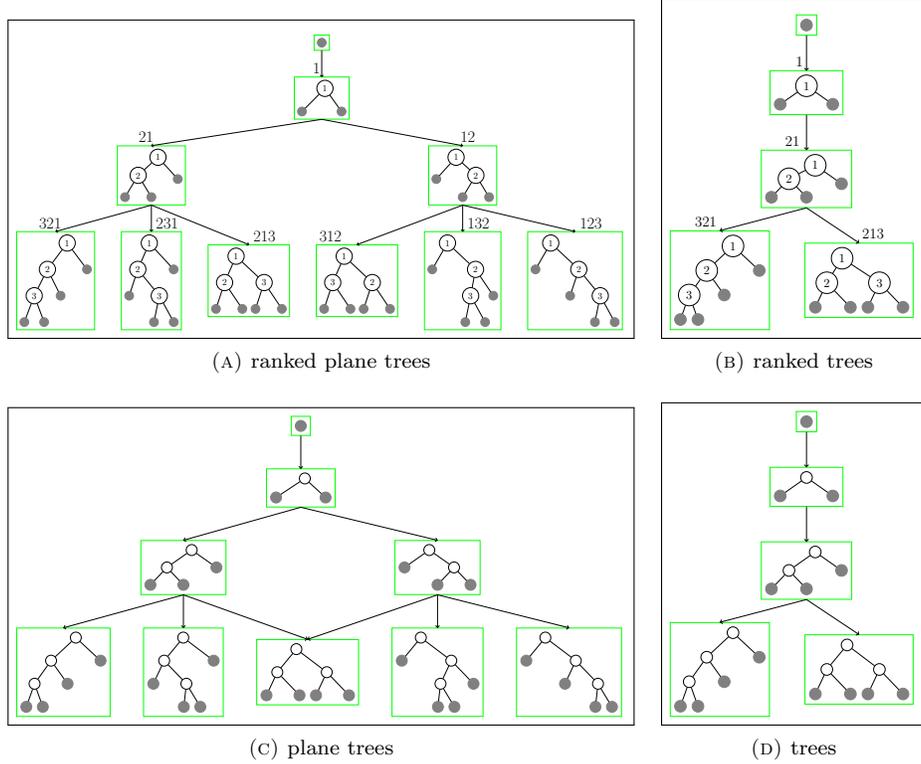
(D) trees

FIGURE 3. Hasse diagram of the recursive splitting process in Construction 2.1 to generate (A) ranked plane trees in $\widehat{\mathcal{T}}_{1:4}^\downarrow$, (B) ranked (non-plane) trees in $\mathcal{T}_{1:4}^\downarrow$, (C) plane trees (unranked) in $\widehat{\mathcal{T}}_{1:4}$, and (D) trees (unranked non-plane) in $\mathcal{T}_{1:4}$ with up to three splits and four leaf nodes. The permutation representations for trees in (A) and (B) are also given.

Clearly, every tree $\hat{t}_n^\downarrow \in \widehat{\mathcal{T}}_n^\downarrow$ with n leaves is obtained by Construction 2.1 in a unique way since at the k-th splitting step we choose exactly one of the available k leaf nodes to split for each $k \in [n-1]$. Thus $|\widehat{\mathcal{T}}_n^\downarrow| = (n-1)!$. There is a simple bijective correspondence between $\widehat{\mathcal{T}}_n^\downarrow$ and the $(n-1)!$ permutations of $[n-1]$ using the *increasing binary tree lifting* (see [14, Ex. 17, p. 132] and the references therein). The bijection, $\widehat{\mathcal{T}}_n^\downarrow \ni \hat{t}_n^\downarrow \leftrightarrow \sigma \in \mathcal{S}_{n-1}$, shown for $n \leq 4$ in Figure 3(A), is given by the following Construction 2.2.

Construction 2.2 (Tree lifting bijection). Consider the following process:

- Write the permutation $\sigma \in \mathcal{S}_{n-1}$ as a word $\sigma = \sigma_1 \sigma_1 \cdots \sigma_{n-1}$.
- If $\min(\sigma)$ is the minimum letter of σ, then σ, as a word, can be decomposed into three terms of the form $\sigma = \sigma_{\mathsf{L}} \cdot \min(\sigma) \cdot \sigma_{\mathsf{R}}$, with $\sigma_{\mathsf{L}}, \sigma_{\mathsf{R}}$ the words to the left and right of $\min(\sigma)$.
- Then $\hat{t}_n^{\downarrow}(\sigma)$, the ranked plane tree corresponding to the given permutation σ is obtained by recursively using this decomposition:
  - The empty tree goes with the empty permutation ε.
  - The root node of the tree $\hat{t}_n^{\downarrow}(\sigma)$ gets rank $\min(\sigma)$ with the left and right subtrees constructed recursively with $\hat{t}_n^{\downarrow}(\sigma_{\mathsf{L}})$ and $\hat{t}_n^{\downarrow}(\sigma_{\mathsf{R}})$, respectively.
- Conversely, you can get σ from a ranked plane tree $\hat{t}_n^{\downarrow}$ by simply reading the ranks at the $n-1$ internal nodes of $\hat{t}_n^{\downarrow}$ in symmetric (in-fix) order.

2.3. **Ranked Trees.** The number of ranked trees with n leaves, $|\mathcal{T}_n^{\downarrow}|$, is given by the *Euler zigzag numbers* [31]. Ranked trees have been studied in evolutionary biology by Tajima [34] as *evolutionary relationships* among n nucleons and have recently been given a coalescent re-formulation [27]. They are called *unlabeled ranked binary dendrograms* in [23] and can be represented by a subset of $\mathcal{S}_{n-1}$ [15]. Recall the *increasing binary tree lifting* that gave a bijection between $\mathcal{S}_{n-1}$ and $\widehat{\mathcal{T}}_n^{\downarrow}$. The idea is to choose a *standard* permutation to represent each $t^{\downarrow}$. We can use a permutation in $\mathcal{S}_{n-1}$ to construct a ranked tree in $\mathcal{T}_n^{\downarrow}$ by modifying Construction 2.2 with a non-planar standard form constraint akin to [23, Section 6].

Thus, to obtain a ranked tree $t^{\downarrow} \in \mathcal{T}_n^{\downarrow}$ that can be drawn in the plane in a unique way from any permutation $\pi \in \mathcal{S}_{n-1}$ just apply the Construction 2.2 with the additional constraints that must be satisfied by the rank labels at the internal nodes:

(1) if only one of the child nodes v of an internal node u is internal, then v is drawn to the left of u;
(2) if both child nodes v and w of an internal node u are internal with $v < w$, then v is drawn on the left of u and w on the right.

Thus, we can use the following non-plane and possibly flipped decomposition of σ

$$\sigma = \sigma_{\overline{\mathsf{L}}} \cdot \min(\sigma) \cdot \sigma_{\overline{\mathsf{R}}} = \begin{cases} \sigma_{\mathsf{L}} \cdot \min(\sigma) \cdot \sigma_{\mathsf{R}} & \text{if } \min(\sigma_{\mathsf{L}}) < \min(\sigma_{\mathsf{R}}) \text{ or } \sigma_{\mathsf{R}} = \varepsilon \\ \sigma_{\mathsf{R}} \cdot \min(\sigma) \cdot \sigma_{\mathsf{L}} & \text{if } \min(\sigma_{\mathsf{L}}) > \min(\sigma_{\mathsf{R}}) \text{ or } \sigma_{\mathsf{L}} = \varepsilon \end{cases}$$

as the only modification in Construction 2.2 such that the root node of the tree $t_n^{\downarrow}(\sigma)$ gets rank $\min(\sigma)$ with the left and right subtrees constructed recursively with $t_n^{\downarrow}(\sigma_{\overline{\mathsf{L}}})$ and $t_n^{\downarrow}(\sigma_{\overline{\mathsf{R}}})$, respectively. After constructing such a ranked tree $t_n^{\downarrow}$ from a permutation $\sigma \in \mathcal{S}_{n-1}$ we can obtain the representative non-plane permutation (corresponding to an equivalence class in $\mathcal{S}_{n-1}$ or equivalently in $\widehat{\mathcal{T}}_n^{\downarrow}$) by just reading the ranks at the internal nodes of $t_n^{\downarrow}$ according to in-fix order as before. Let the set of such representative non-plane permutations be $\mathcal{A}_{n-1}$. Clearly $\mathcal{A}_{n-1} \subset \mathcal{S}_{n-1}$ and the above modification to Construction 2.2 due to [15] gives a bijective correspondence $\mathcal{A}_{n-1} \ni \sigma \leftrightarrow t_n^{\downarrow} \in \mathcal{T}_n^{\downarrow}$. The non-plane permutation representation of ranked trees can be used to enumerate $\mathcal{A}_{n-1}$ which is equinumerous to $\mathcal{T}_n^{\downarrow}$ using the following recursion:

$$\begin{aligned} |\mathcal{A}_n| = |\mathcal{T}_{n+1}^{\downarrow}| = e(n) &= \frac{1}{2} \sum_{k=0}^{n-1} \left(\binom{n-1}{k} e(k) e(n-k-1) \right) \\ e(0) &= e(1) = 1 \end{aligned}$$

The proof is identical to that in [23, p. 196] although ranks are assigned there in decreasing order from the root node. Thus, for

$$n = 1, 2, \ldots, 10, \ |\mathcal{T}_n^\downarrow| = 1, 1, 1, 2, 5, 16, 61, 272, 1385, 7936,$$

respectively [31]. Elements of $\mathcal{T}_n^\downarrow$ and $\mathcal{A}_{n-1}$ for $n \in \{1, 2, 3, 4\}$ are shown in Figure 3(B). We remark in passing that $\mathcal{A}_{n-1}$ is the set of André permutations of the second kind (in reverse) and that the two kinds of alternating permutations of $[n-1]$ and $\mathcal{A}_{n-1}$ or $\widehat{\mathcal{T}}_n^\downarrow$ are in bijective correspondence [10].

Let a *cherry node* be an internal node that has two leaf nodes as its children. If $\beth(t_n^\downarrow)$ be the number of *cherry* nodes of $t_n^\downarrow$ then

$$|\{\hat{t}_n^\downarrow \in \widehat{\mathcal{T}}_n^\downarrow : \hat{t}_n^\downarrow \mapsto t_n^\downarrow\}| = 2^{n-1-\beth(t_n^\downarrow)} \ .$$

Tajima [34] shows that there are $2^{n-1-\beth(t_n^\downarrow)}$ ranked planar trees for a given ranked (non-planar) tree $t_n^\downarrow$. For an intuitive justification of Tajima's result, suppose we want to turn the ranked tree $t_n^\downarrow$ into a ranked planar tree. Then, for each of the $n-1$ internal nodes of $t_n^\downarrow$, there are two choices for the child node that is said to be 'left' except if they are both leaf nodes that carry no ranks (i.e., the internal node is a cherry node). Thus, all internal nodes except the cherry nodes (a total of $n-1-\beth(t_n^\downarrow)$ nodes) give two possible orderings for their child nodes with ranks.

2.4. Plane Trees. Recall the n-th Catalan number [20]:

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!} = \prod_{k=2}^{n} \frac{n+k}{k} \ .$$

The number of plane trees with $n-1$ internal nodes and n leaf nodes is given by C_{n-1}, i.e., $|\widehat{\mathcal{T}}_n| = C_{n-1}$. Recall that the number of ranked plane trees with n leaf nodes is $|\widehat{\mathcal{T}}_n^\downarrow| = (n-1)!$, and this is greater than the number of plane trees, i.e., $|\widehat{\mathcal{T}}_n^\downarrow| > |\widehat{\mathcal{T}}_n|$ for any $n > 2$. Thus, if one ignores the ranks at the internal nodes of ranked plane trees and considers them only as (unranked) plane trees then by the pigeon-hole principle there may be more than one ranked plane tree that corresponds to a plane tree. This combinatorics has to be accounted for when obtaining the distribution on plane trees from that over ranked plane trees. Thus, $\widehat{\mathcal{T}}_n$, the set of *plane trees* with n leaf nodes and without any internal node labels or ranks is an equivalence class of $\widehat{\mathcal{T}}_n^\downarrow$. The next Lemma gives the needed counting argument. We suppress sub-scripting trees by the number of leaves for simplicity.

Lemma 2.3. Let $\hat{t}$ be a plane tree with n leaf nodes and $n-1$ internal nodes, $\check{V}(\hat{t}) := \{v \in V : deg(v) > 1\}$ be the set of internal nodes of $\hat{t}$, $\lfloor \hat{t} \rfloor := |\check{V}(\hat{t})|$ be the number of internal nodes of $\hat{t}$, and $\hat{t}(u)$ be the subtree of $\hat{t}$ with root node u. Then the *Catalan coefficient* [24] of $\hat{t}$, that gives the number of ranked plane trees in $\widehat{\mathcal{T}}_n^\downarrow$ corresponding to the plane tree $\hat{t} \in \widehat{\mathcal{T}}_n$, is:

$$(1) \qquad B(\hat{t}) = \frac{(n-1)!}{\displaystyle\prod_{u \in \check{V}(\hat{t})} \lfloor \hat{t}(u) \rfloor} = \frac{(\# \text{ of internal nodes of } \hat{t})!}{\displaystyle\prod_{u \in \check{V}(\hat{t})} (\# \text{ of internal nodes of } \hat{t}(u))} \ .$$

Proof. Let $L(\hat{t})$ and $R(\hat{t})$ be left and right subtrees of $\hat{t}$. Then the number of distinct binary inter-leavings between the interior (or split) nodes of $L(\hat{t}_n)$ and $R(\hat{t}_n)$ is:

$$
\binom{\lfloor L(\hat{t}) \rfloor + \lfloor R(\hat{t}) \rfloor}{\lfloor L(\hat{t}) \rfloor} = \frac{\left(\lfloor L(\hat{t}) \rfloor + \lfloor R(\hat{t}) \rfloor \right)!}{\lfloor L(\hat{t}) \rfloor! \times \lfloor R(\hat{t}) \rfloor!} = \frac{\lfloor \hat{t} \rfloor \times \left(\lfloor L(\hat{t}) \rfloor + \lfloor R(\hat{t}) \rfloor \right)!}{\lfloor \hat{t} \rfloor \times \lfloor L(\hat{t}) \rfloor! \times \lfloor R(\hat{t}) \rfloor!}
$$

$$
= \frac{\lfloor \hat{t} \rfloor!}{\lfloor \hat{t} \rfloor \times \lfloor L(\hat{t}) \rfloor! \times \lfloor R(\hat{t}_n) \rfloor!} \; .
$$

And the number of distinct binary inter-leavings between the interior nodes of $L(\hat{t})$ and $R(\hat{t})$ as well as their subtrees and their sub-subtrees and so on gives the Catalan coefficient by the following recursion with cancellations:

$$
B(\hat{t}) = \frac{\lfloor \hat{t} \rfloor!}{\lfloor \hat{t} \rfloor \times \lfloor L(\hat{t}) \rfloor! \times \lfloor R(\hat{t}) \rfloor!} \times B(L(\hat{t})) \times B(R(\hat{t}))
$$

$$
= \frac{\lfloor \hat{t} \rfloor!}{\lfloor \hat{t} \rfloor \times \lfloor L(\hat{t}) \rfloor \times \lfloor R(\hat{t}) \rfloor \times \lfloor L(L(\hat{t})) \rfloor \times \lfloor R(L(\hat{t})) \rfloor \times \cdots \times 1}
$$

$$
= \frac{\lfloor \hat{t} \rfloor!}{\prod\limits_{v \in \hat{t}} \lfloor \hat{t}(v) \rfloor} = \frac{(n-1)!}{\prod\limits_{v \in \breve{V}(\hat{t})} \lfloor \hat{t}(v) \rfloor} \; .
$$

$\square$

Remark 2.4. Lemma 2.3 can be proved using poset theoretic ideas as in [32, Ch. 3, Ex. 1.b., p. 312]. The proof given above closely follows that of [9, Cor. 4.1] where $B(\hat{t}_n)$ is called the *shape functional* of the planar tree $\hat{t}_n$ in the context of uniform permutations on binary search trees. Observe that if a rooted binary phylogenetic tree τ with n leaf nodes (but with the leaf labels ignored) is viewed as a plane tree $\hat{t}$ then the Catalan coefficient of $\hat{t}$ is identical to the number of rankings of τ [28, Prop. 2.3.2].

Thus, Lemma 2.3 consolidates [9, Cor. 4.1], [32, Ch. 3, Ex. 1.b., p. 312] and [28, Prop. 2.3.2]. Our nomenclature is motivated by our need of Catalan coefficients to obtain probabilities on tree spaces with up to n leaves akin to how binomial coefficients are needed to obtain probabilities on $\{0, 1, 2 \ldots, n\}$. To fix ideas we consider an example and some visualizations of the Catalan coefficients next.

Example 2.5. We can compute the Catalan coefficient of the perfectly balanced plane tree with $k = 7$ splits and 8 leaves (all with depth 3) using (1), as follows:

$$
B \left(\bigwedge \right) = \frac{7!}{7 \times 3 \times 3 \times 1 \times 1 \times 1 \times 1} = \frac{\overset{2}{\cancel{6}} \times 5 \times 4 \times \cancel{3} \times 2}{\cancel{3} \times \cancel{3}} = 80 \; .
$$

This appears at frequency 1 in the third row of Figure 4. Similarly,

$$
B \left(\bigwedge \right) = \frac{5!}{5 \times 2 \times 2 \times 1 \times 1} = \frac{\cancel{5} \times \cancel{4} \times 3 \times 2}{\cancel{5} \times \cancel{2} \times \cancel{2}} = 6 \; .
$$

Four of the 42 plane trees with five splits have Catalan coefficient of 6 as shown in the third stem plot of the top row of Figure 4.
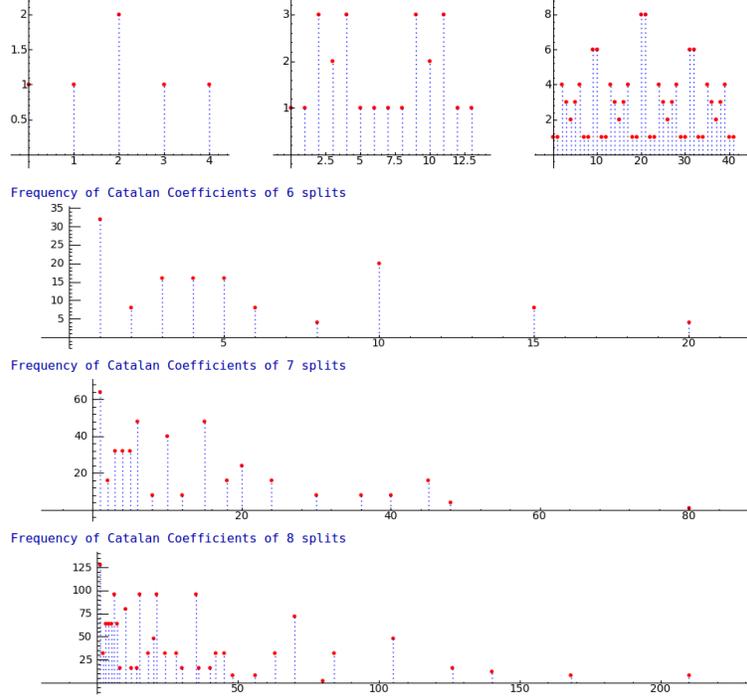
FIGURE 4. Catalan coefficients of plane trees with $3, 4, 5$ splits (top row) and frequency of Catalan coefficients of plane trees with $6, 7, 8$ splits.

The smallest value of the Catalan coefficient is 1 (there is only one path leading to a fully left-only branching unbalanced planar tree, for example). Thus, $\min(B(\hat{t}_n)) = 1$. Moreover, the number of planar trees with $n - 1$ splits that correspond to a non-planar fully unbalanced tree with Catalan coefficient equal to 1 is 2^{n-2}. This is because, out of the $n - 1$ splits in the fully unbalanced tree there are $n - 2$ possibilities for planarity (since there are $n - 2$ non-cherry internal nodes). At the other extreme, the maximum possible value for the Catalan coefficient over trees with n leaves is given by the number of heaps of n elements [30]:

$$1, 1, 1, 2, 3, 8, 20, 80, 210, 896, 3360, 19200, 79200, \ldots .$$

Let $Q(\hat{t}_n) = B(\hat{t}_n)/(n-1)!$ be a measure of the balance of the planar tree. Clearly, more balanced trees will have a higher value of Q compared to less balanced trees. Fill [13] shows that

$$-\frac{\ln Q(\hat{t}_n)}{n} \xrightarrow{P} \sum_{j=2}^{\infty} 4^{-j} C_j \ln(j) \approx 2.03$$

if $\hat{t}_n$ is uniformly distributed on $\widehat{\mathcal{T}}_n$ with $\Pr(\hat{t}_n) = 1/C_{n-1}$, and

$$-\frac{\ln Q(\hat{t}_n)}{n} \xrightarrow{P} 2 \sum_{k=2}^{\infty} \frac{\ln(k)}{(k+1)(k+2)} \approx 1.204$$

if the probability of $\hat{t}_n \in \widehat{\mathcal{T}}_n$ is $Q(\hat{t}_n)$, i.e., induced by the uniform distribution on $\widehat{\mathcal{T}}_n^\downarrow$ with $\Pr(\hat{t}_n^\downarrow) = 1/(n-1)!$. A limiting Normal law is established in [13, Thm. 4.1] for $-\ln(Q(\hat{t}_n))/n$ when $\Pr(\hat{t}_n) = Q(\hat{t}_n)$.

2.5. **Trees.** Trees in this work are finite rooted and binary without node labels as stated in the Introduction. Such trees are called *unlabeled non-ranked binary dendrograms* in [23, Section 5] and as *rooted binary tree shapes* in [28, Section 2.4]. Let $\mathcal{T}_n$ denote the set of such trees with n leaf nodes and $n-1$ internal nodes. The number of such trees as a function of n is given by Wedderburn-Etherington numbers [29]:

$$1, 1, 1, 2, 3, 6, 11, 23, 46, 98, 207, 451, 983, 2179, 4850, 10905, 24631, 56011, \ldots \ .$$

If we generate ranked plane trees in $\widehat{\mathcal{T}}_n^\downarrow$ according to Construction 2.1 and ignore the ranks then we can obtain plane trees in $\widehat{\mathcal{T}}_n$. We can further ignore the planarity of trees in $\widehat{\mathcal{T}}_n$ to obtain trees in $\mathcal{T}_n$. From (1) we know the number of elements in $\widehat{\mathcal{T}}_n^\downarrow$ that map to a given plane tree $\hat{t} \in \widehat{\mathcal{T}}_n$ when we ignore the ranks. Next we find the number of plane trees in $\widehat{\mathcal{T}}_n$ that map to a given tree $t \in \mathcal{T}_n$.

Recall that $\check{V}(t)$ is the set of internal nodes of t and $\lfloor t \rfloor := |\check{V}(t)|$. For each internal node v of $\check{V}_t$, denote by $L(t(v))$ and $R(t(v))$ the left and right subtrees of t below v, i.e., with $v\mathsf{L}$ and $v\mathsf{R}$ as their roots, respectively. Let the set of symmetry nodes of t be

$$S(t) := \{v \in \check{V}(t) : L(t(v)) \text{ and } R(t(v)) \text{ are isomorphic } \} \text{ and } s(t) := |S(t)| \ .$$

Lemma 2.6. Let $t = (V, E) \in \mathcal{T}_n$ with $n \geq 2$. Then, t corresponds to $2^{n-1-s(t)}$ plane trees, i.e.

$$(2) \qquad C(t) := |\{\hat{t} \in \widehat{\mathcal{T}}_n : \hat{t} \mapsto t\}| = 2^{\lfloor t \rfloor - s(t)} = 2^{n-1-s(t)} \ .$$

Proof. The last equality in (2) is merely due to the fact that a tree with n leaves has $n-1$ internal nodes, i.e., if $t \in \mathcal{T}_n$ then $\lfloor t \rfloor = n-1$. We use induction on the number n of leaves of t to prove the main equality in (2).

Initial case $n = 2$: There is only one tree t with two leaves, namely the one consisting of the root node with two attached leaves. Clearly, the root is a symmetry node, so $s(t) = 1$ and since the root is the only inner node of t we get $\lfloor t \rfloor = 1$. Also, there is only one plane tree $\hat{t}$ with two leaves. It consists of the root node ρ and the two leaf nodes with addresses $\rho\mathsf{L}$ and $\rho\mathsf{R}$. So altogether, for the number of plane trees we have: $1 = 2^{\lfloor t \rfloor - s(t)} = 2^{1-1}$. This completes the proof for $n = 2$.

Inductive case $n \to n+1$: Let t have $n+1$ leaves and assume that the lemma is already proven for any natural number up to and including n. Let ρ be the root of t. Let $\rho\mathsf{L}$ and $\rho\mathsf{R}$ denote the children of ρ and let $L(t)$ and $R(t)$ denote the subtrees rooted at $\rho\mathsf{L}$ and $\rho\mathsf{R}$, respectively. Let n_R, n_L denote the number of leaves in $L(t)$ and $R(t)$, respectively. Note that $n_L + n_R = n + 1$. Moreover, we have $\lfloor t \rfloor = \lfloor L(t) \rfloor + \lfloor R(t) \rfloor + 1$ because of the root. As both n_L and n_R are smaller than $n + 1$, we know by the inductive assumption that $L(t)$ corresponds to $2^{\lfloor L(t) \rfloor - s(L(t))}$ plane trees and $R(t)$ to $2^{\lfloor R(t) \rfloor - s(R(t))}$. Now if $L(t)$ and $R(t)$ are isomorphic, ρ is a symmetry node of t. In this case, $s(t) = s(L(t)) + s(R(t)) + 1$. Else, $s(t) = s(L(t)) + s(R(t))$. We consider the two cases next.

Symmetric sub-case: If ρ is a symmetry node, the number of plane trees induced by t is just the number of such trees induced by $L(t)$ times that induced by $R(t)$, because swapping the left and the right subtrees would not lead to any more distinct

plane trees. Therefore, we conclude that the number of plane trees induced by t in this case is

$$
\begin{aligned}
2^{\lfloor L(t) \rfloor - s(L(t))} 2^{\lfloor R(t) \rfloor - s(R(t))} &= 2^{(\lfloor L(t) \rfloor + \lfloor R(t) \rfloor) - (s(L(t)) + s(R(t)))} \\
&= 2^{(\lfloor t \rfloor - 1) - (s(t) - 1)} = 2^{\lfloor t \rfloor - s(t)} \ .
\end{aligned}
$$

Asymmetric sub-case: If ρ is not a symmetry node, the number of plane trees induced by τ is the number of such trees induced by $L(t)$ times that induced by $R(t)$ times 2, because the roles of $L(t)$ and $R(t)$ can be swapped about the asymmetric node ρ to obtain two distinct plane trees. Therefore, the number of plane trees induced by t in this case is

$$
\begin{aligned}
2^{\lfloor L(t) \rfloor - s(L(t))} 2^{\lfloor R(t) \rfloor - s(R(t))} 2 &= 2^{(\lfloor L(t) \rfloor + \lfloor R(t) \rfloor) - (s(L(t)) + s(R(t))) + 1} \\
&= 2^{(\lfloor t \rfloor - 1) - s(t) + 1} = 2^{\lfloor t \rfloor - s(t)} \ .
\end{aligned}
$$

This completes the proof. $\qquad\square$

More than one plane tree in $\widehat{\mathcal{T}}_n$ may map to a given tree $t \in \mathcal{T}_n$ when we ignore planarity. However, $B(\hat{t})$, the Catalan coefficient of any plane tree $\hat{t}$ in $\{\hat{t} \in \widehat{\mathcal{T}}_n : \hat{t} \mapsto t\}$, that maps to any given tree $t \in \mathcal{T}_n$, is identical. This is because $B(\hat{t})$ in (1) only depends on $\prod_{v \in \check{V}(\hat{t})} \lfloor \hat{t}(v) \rfloor$, the product of the number of internal nodes in each subtree with an internal node in $\hat{t}$ as its root, a quantity that is preserved when planarity is ignored. Thus,

$$
B(\hat{t}_n) = B(t_n) = \frac{(n-1)!}{\prod_{v \in \check{V}} \lfloor t_n(v) \rfloor} \ .
$$

This leads to the next lemma.

Lemma 2.7. The number of ranked planar trees that map to a tree $t_n \in \mathcal{T}_n$ is:

$$
(3) \qquad |\{\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow} : \hat{t}_n^{\downarrow} \mapsto t_n\}| = B(t_n) 2^{n-1-s(t_n)} \ .
$$

Lemma 2.8. The number of ranked trees in $\mathcal{T}_n^{\downarrow}$ corresponding to a given tree $t_n \in \mathcal{T}_n$ is:

$$
(4) \qquad |\{t_n^{\downarrow} \in \mathcal{T}_n^{\downarrow} : t_n^{\downarrow} \mapsto t_n\}| = \frac{(n-1)!}{\displaystyle\prod_{u \in \check{V}(\hat{t}_n)} \lfloor \hat{t}_n(u) \rfloor} 2^{\beth(t_n) - s(t_n)} \ .
$$

Proof. Recall that a cherry node is an internal node that has two leaf nodes as its children. If $\beth(t_n^{\downarrow})$ be the number of cherry nodes of $t_n^{\downarrow}$ then

$$
|\{\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow} : \hat{t}_n^{\downarrow} \mapsto t_n^{\downarrow}\}| = 2^{n-1-\beth(t_n^{\downarrow})} \ .
$$

Also, we know the following two facts:

$$
|\{\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow} : \hat{t}_n^{\downarrow} \mapsto \hat{t}_n\}| = B(\hat{t}_n) = \frac{(n-1)!}{\displaystyle\prod_{u \in \check{V}(\hat{t}_n)} \lfloor \hat{t}_n(u) \rfloor}
$$

and

$$
C(t_n) := |\{\hat{t}_n \in \widehat{\mathcal{T}}_n : \hat{t}_n \mapsto t_n\}| = 2^{\lfloor t_n \rfloor - s(t_n)} = 2^{n-1-s(t_n)} \ ,
$$

where $s(t_n)$ is the size of the set of symmetry nodes of $t_n \in \mathcal{T}_n$:

$$
S(t_n) := \{v \in \check{V}(t_n) : L(t_n(v)) \text{ and } R(t_n(v)) \text{ are isomorphic } \} \text{ and } s(t_n) := |S(t_n)| \ .
$$

Due to the invariance of $\lrcorner$, s, B and C to the equivalence classes in $\widehat{\mathcal{T}}_n^{\downarrow}$, $\widehat{\mathcal{T}}_n$, $\mathcal{T}_n^{\downarrow}$ and $\mathcal{T}_n$, we obtain:

$$|\{\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow} : \hat{t}_n^{\downarrow} \mapsto t_n\}| = B(t_n)C(t_n) = \frac{(n-1)!}{\prod\limits_{u \in \check{V}(t_n)} \lfloor t_n(u) \rfloor} 2^{n-1-s(t_n)} \ .$$

Finally, we obtain:

$$|\{\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow} : \hat{t}_n^{\downarrow} \mapsto t_n\}| = |\{\hat{t}_n^{\downarrow} \in \widehat{\mathcal{T}}_n^{\downarrow} : \hat{t}_n^{\downarrow} \mapsto t_n^{\downarrow}\}| \times |\{t_n^{\downarrow} \in \mathcal{T}_n^{\downarrow} : t_n^{\downarrow} \mapsto t_n\}|$$

$$\frac{(n-1)!}{\prod\limits_{u \in \check{V}(t_n)} \lfloor t_n(u) \rfloor} 2^{n-1-s(t_n)} = 2^{n-1-\lrcorner(t_n^{\downarrow})} \times |\{t_n^{\downarrow} \in \mathcal{T}_n^{\downarrow} : t_n^{\downarrow} \mapsto t_n\}| \ .$$

Thus

$$|\{t_n^{\downarrow} \in \mathcal{T}_n^{\downarrow} : t_n^{\downarrow} \mapsto t_n\}| = \frac{(n-1)!}{\prod\limits_{u \in \check{V}(\hat{t}_n)} \lfloor \hat{t}_n(u) \rfloor} 2^{\lrcorner(t_n)-s(t_n)} \ .$$

$\square$

3. Other Tree Constructions

There are a few representations of the state space for the probabilistic construction of trees. The different representations allow different classes of distributions to be defined easily on ranked plane trees.

We can turn any randomized algorithm that generates permutations on $[n-1]$ into one that generates ranked planar trees in $\widehat{\mathcal{T}}_n^{\downarrow}$ by simply going from $\sigma \mapsto \hat{t}^{\downarrow}$ using Construction 2.2. A simple way to generate random permutations is through the *sampling without replacement scheme*, where you start with n balls labelled $1, 2 \ldots, n$ from an urn, picking one by one, uniformly at random, noting its label and setting it outside the urn in a row. Another simple way is through a *Knuth Shuffle*, where you start with any permutation (say, the identity permutation), and then go through the positions 1 through $n-1$, such that for each position i swap the element currently at i with a randomly chosen element from positions $i, i+1, \ldots, n$. Although these randomized algorithms over permutations can be transformed using Construction 2.2 into randomized trees, they are not evolutionary as in Construction 2.1 since the trees are not grown randomly in an incremental manner by splitting one of the existing leaves.

For an evolutionary and incremental construction over permutations, consider a recursive sampling scheme that inserts the i-th ball into one of the i gaps between the $i-1$ balls that have been inserted up to step i. This is equivalent to splitting one of the current leaf nodes of the corresponding ranked plane tree. A natural construction of this idea using trees is described next.

A *binary search tree* is a rooted planar binary tree, whose internal nodes each store a key (say, a real number) and each internal node has left and right subtrees (see for e.g. [19]). The tree additionally satisfies the *binary search tree property*, whereby the key in each node must be greater than all keys stored in the left subtree, and smaller than all keys stored in right subtree. The leaf nodes of the tree contain no key and are usually left unlabelled. We are interested in inserting a new key into the tree and growing the tree as summarized in Construction 3.1.
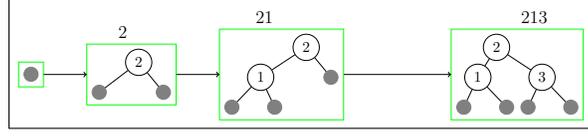
FIGURE 5. Binary search tree grown by inserting $2, 1, 3$.

Construction 3.1 (Inserting Random Permutation into Binary Search Tree). Suppose you are given $\sigma = \sigma_1 \sigma_2 \cdots \sigma_{n-1}$, a random permutation of $[n-1]$. First, insert the key σ_1 into the root node of the binary search tree, a planar tree in $\widehat{\mathcal{T}}_n$. In order to insert the i-th node in the tree, its key σ_i is first compared with that of the root node, i.e., with σ_1. If its key is less than that of the root, it is then compared with that of the root's left child node. If its key is greater than that of the root, it is then compared with that of the root's right child node. This process continues, until the new node to be inserted is compared with a sub-terminal node, and then it is added as this node's left or right child, depending on whether its key is greater than or less than the key of the sub-terminal node, respectively.

For example, the planar tree with three internal nodes and four leaves that is grown by inserting $2, 1, 3$ is shown in Figure 5. The first element 2 is inserted into the root node. The second element 1 is less than 2 at the root, so it is inserted into the left child node to ensure the binary search tree property. Finally, the third element 3 in the sequence is inserted into the right child node of the root since it is greater than 2.

Construction 3.2 (Dyadic Partition Edges). Let the set of dyadic fractions be

$$\mathbb{X} := \left\{ x = (0.b_1 b_2 \ldots b_{n_x})_2 = \sum_{j=1}^{n_x} b_j 2^{-j} : b_j \in \{0,1\}, 1 \le j < n_x, b_{n_x} = 1, n_x < \infty \right\},$$

where $(0.b_1 b_2 \ldots b_{n_x})_2$ is the finite binary expansion of the dyadic fraction $x = m/2^n$ with $m, n \in \mathbb{N}$. Let $\mathrm{mid}\,(a,b) = (b+a)/2$ be the mid-point of a and b. Let $\check{x} := 2^{-n_x}$ be the smallest additive constituent of x due to the terminal binary digit $b_{n_x} = 1$.

Our construction gives a sequence of $\mathbb{X}$-valued random variables $(X_0, X_1, \ldots, X_k)$, such that $X_0 = 0$, $X_1 = 1$ and for $k \ge 2$ we obtain X_k from $(X_0, \ldots, X_{k-1})$ from a randomly chosen index I for bisecting or splitting as follows:

$$X_k \leftarrow \mathrm{mid}\,\left(X_{(I)}, X_{(I+1)} \right),$$

where, $X_{(0:k)} := \left(X_{(0)}, X_{(1)}, \ldots, X_{(i)}, X_{(i+1)}, \ldots, X_{(k)} \right)$ is the order statistics of the sequence $X_{0:k} := (X_0, X_1, \ldots, X_i, X_{i+1}, \ldots, X_k)$. We also refer to $X_{(0:k)}$ as a partition since it represents the following partition of $[0,1]$

$$[X_{(0)}, X_{(1)}) \cup \cdots \cup [X_{(i)}, X_{(i+1)}) \cdots \cup [X_{(k-1)}, X_{(k)}] \ .$$

The general probabilistic *splitting rule* to obtain X_k from the mid-point of a randomly chosen interval $(X_{(I)}, X_{(I+1)})$ of the partition $X_{(0:k-1)}$ generated by $X_{0:k-1}$ is given by the transition probability matrix P with entries:

(5) $P(x_{0:k-1}, i) := \Pr\{I = i \mid X_{0:k-1} = x_{0:k-1}\}, \ i \in \{0, 1, \ldots, k-2\}, k \in 2, 3, \ldots$

Construction 3.3 (Dyadic Partition Depths). We can equivalently represent $X_{(0:k)}$ by the width of the successive intervals partitioning $[0, 1]$ as follows:

$$W(X_{(0:k)}) := W_{1:k} = (W_1, \ldots, W_k) = \big((X_{(1)} - X_{(0)}), \ldots, (X_{(k)} - X_{(k-1)})\big)$$

Note that $\sum_{i=1}^k W_i = 1$ and $0 < W_i \leq 1$ and therefore for each $k \geq 1$ we can think of $W_{1:k} = (W_1, \ldots, W_k)$ as a probability distribution over k outcomes. It is convenient to denote $W_{1:k}$ in terms of integer sequences as follows:

$$Y(X_{(0:k)}) := Y_{1:k} = (Y_1, \ldots, Y_k) = \big(-\lg(X_{(1)} - X_{(0)}), \ldots, -\lg(X_{(k)} - X_{(k-1)})\big)$$

This is called the depth encoding corresponding to the dyadic partition. We can obtain $Y_{1:k}$ from $Y_{1:k-1}$ by choosing I at random according to Equation (5) and replacing it by two consecutive entries that are deeper by 1 as follows:

$$
\begin{aligned}
Y_1 &\leftarrow Y(X_{(0:1)}) = (-\lg(X_{(1)} - X_{(0)})) = (0) \\
Y_{1:k} &\leftarrow (Y_1, Y_2, \ldots, Y_I + 1, Y_I + 1, \ldots, Y_{k-1}) .
\end{aligned}
$$

Figure 6 depicts the three representations of the binary tree generation process under two splits and Figure 7 gives the state transition diagram of the process over dyadic partition depths.

$$X_{0:1} = (0, 1) \qquad X_{0:2} = (0, 1, \tfrac{1}{2}) \qquad X_{0:3} = (0, 1, \tfrac{1}{2}, \tfrac{1}{4})$$

$$X_{(0:1)} = (0, 1) \qquad X_{(0:2)} = (0, \tfrac{1}{2}, 1) \qquad X_{(0:3)} = (0, \tfrac{1}{4}, \tfrac{1}{2}, 1)$$

$$Y_1 = (0) \qquad Y_{1:2} = (1, 1) \qquad Y_{1:3} = (2, 2, 1)$$
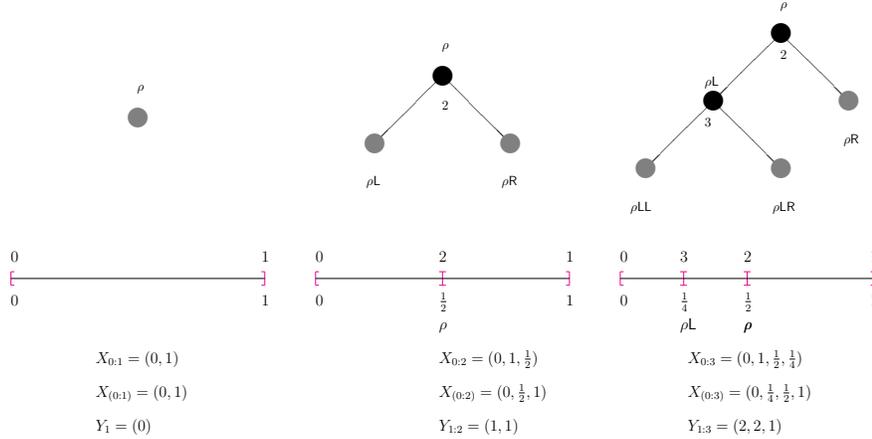
FIGURE 6. A growing ranked plane tree and the corresponding partition of $[0, 1]$.

A specific sequence $x_{0:k}$ represents a sequence of $k-1$ bisections or splits of $[0, 1]$ into k intervals and the order statistics $x_{(0:k)}$ represents the corresponding partition formed by the sequence of splits. Let $\mathcal{X}_k$ denote the set of all dyadic partitions of size k which is in bijective correspondence with $\widehat{\mathcal{T}}_k$, the set of plane trees with k leaves. Thus, $\#\mathcal{X}_k = \#\widehat{\mathcal{T}}_k = C_{k-1}$, the $(k-1)$-th Catalan number. When there are k intervals in $X_{0:k}$ there are k possible splits leading to k choices for $X_{0:k+1}$. Thus, there are $(k-1)!$ distinct sequences for $X_{0:k}$ which is in bijective correspondence with $\widehat{\mathcal{T}}_k^{\downarrow}$, the set of ranked plane trees with k leaves.
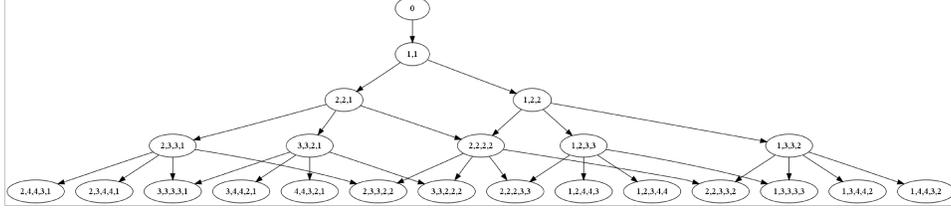
FIGURE 7. State Transition Diagram with 0, 1, 2, 3, and 4 splits.

4. NICE PROPERTIES OF SOME FAMILIAR PROBABILITY MODELS

Model 4.1 (General Splitting). In the most general model, we allow the transition probabilities given in Equation (5) to possibly depend on the entire history of $X_{0:k-1}$. For each level $k \in \{2, 3, \ldots\}$, corresponding to $k-1$ splits, the transition matrix $P(x_{0:k-1}, i)$ has $(k-1)!$ rows, corresponding to the number of distinct possibilities for $x_{0:k-1}$, and $k-1$ columns, corresponding to the number of intervals or leaves in the partition $x_{(0:k-1)}$ associated with each such sequence $x_{0:k-1}$. Since the rows of P must be non-negative and sum to 1, we can think of each row as being a point in the $(k-1)$-simplex:

$$\Delta_{k-1} := \{(p_0, p_1, \ldots, p_{k-2}) : p_i \geq 0, \forall i, \sum_{i=0}^{k-2} p_i = 1\}$$

So each family of such transition probabilities can be thought of as an element of:

$$\{\Delta_{k-1}^{(k-1)!} : k \in \{2, 3, \ldots\}\}$$

that can index the law of a partitioning or tree-building process.

Besides specifying the largest non-parametric family for the tree-building process, Model 4.1 is too general to provide useful insights. We next restrict the construction to satisfy a Markov property on state space $\mathcal{X}$, the set of all dyadic partitions of $[0, 1]$. We want the probability of X_k given the entire history $X_{0:k-1}$ to only depend on the partition $X_{(0:k-1)}$:

$$\Pr\{X_k \mid X_{0:k-1}\} = \Pr\{X_k \mid X_{(0:k-1)}\} .$$

Model 4.2 (Markov Splitting). Here we allow the transition probabilities given in Equation (5) to only depend on the entire history of $X_{0:k-1}$ up to the most recent partition $X_{(0:k-1)}$. We can accomplish this by ensuring that the transition probabilities satisfy:

$$P(x_{0:k-1}, i) = \Pr\{I \mid X_{0:k-1}\} = \Pr\{I \mid X_{(0:k-1)}\} = P(x_{(0:k-1)}, i),$$

where, $X_k = \text{mid}(X_{(I)}, X_{(I+1)})$. For each level $k \in \{2, 3, \ldots\}$, corresponding to $k-1$ splits, the transition matrix $P(x_{0:k-1}, i) = P(x_{(0:k-1)}, i)$ has only C_k rows, corresponding to the number of distinct possibilities for $x_{(0:k-1)}$, and $k-1$ columns, corresponding to the number of intervals or leaves in the partition $x_{(0:k-1)}$. Thus, the construction gives a Markov chain on state space $\mathcal{X}$, the set of all dyadic partitions of $[0, 1]$, that satisfies the following Markov property

$$\Pr\{X_{(0:k)} \mid X_{(0:1)}, X_{(0:2)}, \ldots, X_{(0:k-1)}\} = \Pr\{X_{(0:k)} \mid X_{(0:k-1)}\}$$

So each family of such transition probabilities can be thought of as an element of:

$$\{\Delta_{k-1}^{C_{k-1}} : k \in \{2, 3, \ldots\}\}$$

that can index the law of a partitioning or tree-building process. Finally, instead of the bisection scheme where $X_k = \text{mid}\,(X_{(I)}, X_{(I+1)})$, we can substitute a more general way of splitting the interval $(X_{(I)}, X_{(I+1)})$ into two subintervals. For instance, we can sample a point s from a density g rescaled over $(X_{(I)}, X_{(I+1)})$, such that $\int_{X_{(I)}}^{X_{(I+1)}} g(x)dx = 1$, and use it to split $(X_{(I)}, X_{(I+1)})$ into $(X_{(I)}, s)$ and $(s, X_{(I+1)})$.

Next we present some concrete Markov splitting models that are special cases of Model 4.2.

Model 4.3 (Uniform Splitting). A concrete example of the conditional random variable I is Uniform$\{0, 1, \ldots, k-2\}$ with

$$P(x_{0:k-1}, i) = P(x_{(0:k-1)}, i) = \begin{cases} 1/(k-1) & \text{if } i \in \{0, 1, \ldots, k-2\}, k \in \{2, 3, \ldots\} \\ 0 & \text{otherwise} \end{cases},$$

This corresponds to producing the next split by choosing one of the current intervals or leaves uniformly at random. This model assigns uniform probability $1/(k-1)!$ to every ranked planar tree $\hat{t}_k^\downarrow \in \hat{T}_k^\downarrow$ with $k-1$ splits and k leaves and is equivalent to the speciation model due to Yule [35] in phylogenetics and the random permutation model for binary search trees [9].

Model 4.4 (Statistically Equivalent Block or SEB Splitting). The distribution of I can be given by a probability density function f on $[0, 1]$ such that $\int_0^1 f(u)du = 1$, $f(u) \geq 0$ for every $u \in [0, 1]$ and $f(u) = 0$ for every $u \notin [0, 1]$. Under this model we choose the next leaf for splitting according to:

$$P(x_{0:k-1}, i) = P(x_{(0:k-1)}, i) = \int_{x_{(i)}}^{x_{(i+1)}} f(u)du, \quad i \in \{0, 1, \ldots, k-2\}, k \in \{2, 3, \ldots\}.$$

And having chosen a leaf interval, we split that leaf interval exactly at its midpoint. For the special case of the continuous random variable on $[0, 1]$ with uniform density: $f(u) = 1$ if $u \in [0, 1]$ and $f(u) = 0$ if $u \notin [0, 1]$, we have the Uniform SEB Splitting Model where intervals are bisected according to probabilities given by their widths. This model is indexed by a density f on $[0, 1]$ and produces trees such that its leaf intervals have nearly uniform probability under f since the leaf interval with the most probability under f is immediately bisected. This is related to the statistically equivalent blocks rule as a consistent partitioning strategy in density estimation [16].

Model 4.5 (Depth-proportional Splitting). The distribution of I is obtained by normalizing the depth of each leaf in $x_{(0:k-1)}$ or $y_{1:k}$ as follows:

$$P(x_{0:k-1}, i) = P(x_{(0:k-1)}, i) = \frac{-\lg(x_{(i)} - x_{(i-1)})}{\sum_{i=1}^{k-1} -\lg(x_{(i)} - x_{(i-1)})} = \frac{y_i}{\sum_{i=1}^{k-1} y_i} = P(y_{1:k-1}, i),$$

where, $i \in \{1, \ldots, k-1\}, k \in \{2, 3, \ldots\}$. Thus, deeper nodes have a higher probability of splitting under this model.

Blum and François [2] introduced an evolutionary Beta-splitting model based on ideas of Kirkpatrick and Slatkin [18], and Aldous [1]. This model is further extended to a biparametric Beta-splitting model for diversification in [25] and for

epidemiological transmission in [26]. Under this model, a current interval is chosen for bisection with probability given by its width and the point of bisection is drawn from a Beta$(\alpha + 1, \beta + 1)$ distribution rescaled over the interval being bisected, for some $\alpha > -1$ and $\beta > -1$. This way of splitting a leaf interval is in contrast with earlier models where we always bisected the chosen leaf interval at its mid-point. The following model is from [25].

Model 4.6 (A biparametric Beta-splitting model). Let $(B_1, B_2, \ldots)$ be a sequence of independent and identically distributed (i.i.d.) random variables, with the $\mathcal{B}(\alpha + 1, \beta + 1)$ distribution. Let also $(U_1, U_2, \ldots)$ be a sequence of i.i.d. random variables with the uniform distribution on $[0, 1]$, that is independent of $(B_1, B_2, \ldots)$.

Let $((u_i, b_i))_{i \in \mathbb{N}}$ be a realization of $((U_i, B_i))_{i \in \mathbb{N}}$. The tree construction proceeds incrementally as follows, until the tree created has n leaves. We start with a single root node, labelled by the interval $[0, 1]$.

- Step 1: Split the root into a left leaf labelled by $[0, b_1]$ and a right leaf labelled by $[b_1, 1]$. Change the label of the root to the integer 1.
- Step 2: If $u_2 \in [0, b_1]$, split the left child node of the root into a left leaf and a right leaf respectively labelled by $[0, b_1 b_2]$ and $[b_1 b_2, b_1]$. If $u_2 \in [b_1, 1]$, then instead split the right child node of the root into left and right leaves with respective labels $[b_1, b_1 + (1 - b_1)b_2]$, $[b_1 + (1 - b_1)b_2, 1]$. Label the former leaf that is split during this step by 2.
- Step i: Find the leaf whose interval label $[a, b]$ contains u_i. Change its label to the integer i and split it into a left leaf with label $[a, a + (b - a)b_i]$ and a right leaf with label $[a + (b - a)b_i, b]$.
- Stop at the end of Step $n - 1$.

In words, at each step i the labels of the leaves form a partition of the interval $[0, 1]$. We find the next leaf to be split by checking which interval contains the corresponding u_i and then b_i is used to split the interval of that former leaf, say with length ℓ, into two intervals of lengths $b_i \ell$ and $(1 - b_i)\ell$. The internal node just created is then labelled by i to record the order of the splits. At the end of step i, the tree has $i + 1$ leaves, and so we stop the procedure at step $n - 1$. Figure 8 shows an example of such construction for $n = 4$. The probability of obtaining a ranked plane tree $\hat{t}_n^\downarrow$ under this model by erasing the interval leaf labels is given in [25, Thm. 1] by integrating over all possible splits. An interpretation of this model for transmission trees in terms of the underlying contact network of hosts undergoing an epidemic is given in [26].

4.1. Two properties and their consequences.

Definition 4.7 (split-exchangeable model). If $\Pr\{X_{0:k} = x_{0:k}\}$ or $\Pr\{\hat{T}_k^\downarrow = \hat{t}_k^\downarrow\}$ obtained from a Markov splitting model is identical for every $x_{0:k}$ or $\hat{t}_k^\downarrow$ that has the same partition $x_{(0:k)}$ or the same planar tree $\hat{t}_k$, then the model is said to be *split-exchangeable*. Note that Models 4.3, 4.4 and 4.6 are split-exchangeable while Model 4.5 is not.

Theorem 4.8 (Split-exchangeable planar tree probability). *The probability of reaching a given partition or plane binary tree under a Markov splitting model that satisfies split-exchangeability is:*

$$\Pr\{X_{(0:k)} = x_{(0:k)}\} = B(x_{(0:k)}) \times \Pr\{x_{0:k}\}.$$
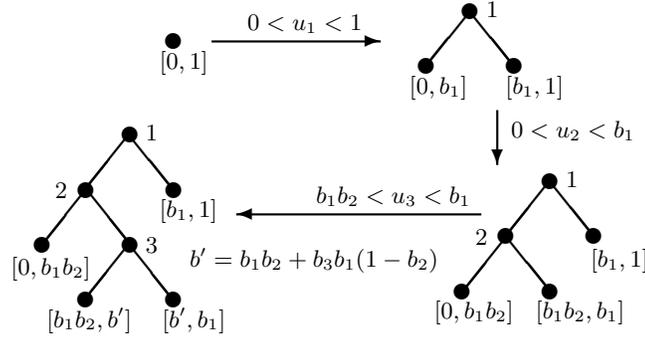
FIGURE 8. An example of a Beta-splitting tree construction for $k = 3$.

or equivalently in tree notation:

$$\Pr\{\hat{T}_k = \hat{t}_k\} = B(\hat{t}_k) \times \Pr\{\hat{t}_k^\downarrow\}$$

where $\hat{t}_k^\downarrow \in \{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto \hat{t}_k\}$ and B is the Catalan coefficient.

Proof. Since the probabilities are split-exchangeable, $\Pr(\hat{t}^\downarrow)$ is identical for each $\hat{t}_k^\downarrow \in \{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto \hat{t}_k\}$.

$$\Pr\{\hat{T}_k = \hat{t}_k\} = \Pr\{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto \hat{t}_k\} = \#\{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto \hat{t}_k\} \times \Pr\{\hat{t}_k^\downarrow\} = B(\hat{t}_k) \times \Pr\{\hat{t}_k^\downarrow\}$$

The last equality is due to (1). $\qquad\square$

Definition 4.9 (plane-invariant model). If $\Pr\{\hat{T}_k = \hat{t}_k\}$ is identical for every $\hat{t}_k$ that has the same tree t_k then the model is said to be *plane-invariant*. For example, Model 4.3 is plane-invariant but Model 4.4 is not.

Theorem 4.10 (Split-exchangeable plane-invariant tree probability). *The probability of a tree under a Markov splitting model that satisfies split-exchangeability and plane-invariance is:*

$$\Pr\{T_k = t_k\} = B(t_k) \times 2^{k-1-s(t_k)} \times \Pr\{\hat{t}_k^\downarrow\}$$

where $\hat{t}_k^\downarrow \in \{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto \hat{t}_k\}$, and $\hat{t}_k \in \{\hat{t}_k : \hat{t}_k \mapsto t_k\}$, B is the Catalan coefficient and $s(t_k)$ is the number of symmetry nodes in t_k.

Proof. Since the probabilities are split-exchangeable, $\Pr(\hat{t}^\downarrow)$ is identical for each $\hat{t}_k^\downarrow \in \{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto \hat{t}_k\}$. And, since the probabilities are plane-invariant, $\Pr(\hat{t})$ is identical for each $\hat{t}_k \in \{\hat{t}_k : \hat{t}_k \mapsto t_k\}$. Therefore,

$$\Pr\{T_k = t_k\} = \Pr\{\hat{t}_k^\downarrow : \hat{t}_k^\downarrow \mapsto t_k\} = B(\hat{t}_k) \times 2^{k-1-s(t_k)} \times \Pr\{\hat{t}_k^\downarrow\}$$

The last equality is due to (3). $\qquad\square$

4.2. **Applications of split-exchangeability and plane-invariance.** Theorems 4.8 and 4.10 can be useful in obtaining probabilities of trees at coarser resolutions from the probabilities at the finer resolution of ranked planar trees if they satisfy split-exchangeability (and plane-invariance). For example, Model 4.3 for Yule trees
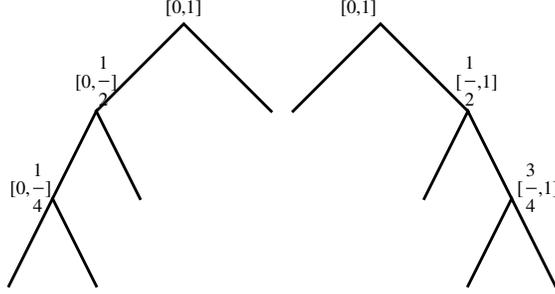
FIGURE 9. A pair of trees giving partitions $(0, \frac{1}{4}, \frac{1}{2}, 1)$ and $(0, \frac{1}{2}, \frac{3}{4}, 1)$ respectively

with $\Pr(\hat{t}_k^\downarrow) = 1/(k-1)!$ for every $\hat{t}_k^\downarrow \in \hat{T}_k^\downarrow$, is split-exchangeable and plane-invariant and therefore by Theorem 4.10,

$$\Pr\{T_k = t_k\} = \frac{(n-1)!}{\prod_{v \in \check{V}} \lfloor t_k(v) \rfloor} \times 2^{k-1-s(t_k)} \times \frac{1}{(k-1)!}.$$

This gives the nonuniform probability of an unranked and nonplanar Yule tree in terms of the product of its subtree splits and symmetry nodes.

The Beta-splitting model is split-exchangeable for any α and β, but only plane-invariant when $\alpha = \beta$. These properties of the Beta-splitting model are used in [25] to readily obtain probabilities of planar trees, nonplanar ranked trees and nonplanar unranked trees from those of their corresponding ranked planar trees.

4.3. **Thompson's group** F. We next describe how probability models on plane binary trees are used to obtain insights in geometric group theory involving Thompson's group. Thompson's group F is a group with a range of unusual properties and a wide range of characterizations. Here, we consider elements of Thompson's group F as piecewise-linear orientation-preserving homeomorphisms of the unit interval $[0, 1]$ to itself, with slopes that are powers of two, and with breakpoint sets that are contained in the dyadic rationals. Such elements can be described by pairs of rooted binary trees of the same size, where the corresponding group element is the piecewise-linear interpolation of the corresponding dyadic partitions described by the trees. The widths of the intervals is exactly the dyadic width described in Representation 3.3. For further background on Thompson's group F, see the introduction by Cannon, Floyd and Parry [4].

4.3.1. *Sampling in F*. Thompson's group F is the simplest known example of a wide range of pathological group-theoretic behavior, with it serving as counterexamples to a wide range of conjectures. Furthermore, there are a number of properties of F which are not known despite a great deal of study over the last 40 years. There have thus been a number of computational experiments by a range of authors [7, 5, 8, 12, 11, 6, 3] designed to yield insight into the group-theoretic properties of F. For those involving sampling, there are three methods that have been used to sample elements at random of increasingly large subsets of F. The first is to sample words chosen at random from the balls of size n with respect to the standard word metric for F (using the generating set $\{x_0, x_1\}$). This has a number
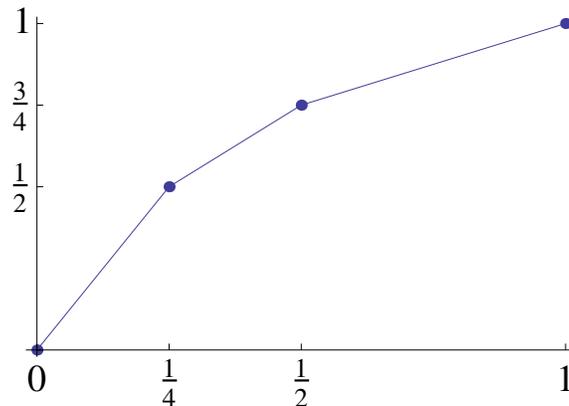
FIGURE 10. Interpolating between the two partitions of the unit interval $\{0, \frac{1}{4}, \frac{1}{2}, 1\}$ and $\{0, \frac{1}{2}, \frac{3}{4}, 1\}$ to make a piecewise-linear homeomorphism

of desired properties but unfortunately the sizes of the metric balls are not known generally, not even asymptotically, and though it is known that the growth rate is exponential, the growth rate is unknown. There are proven upper and lower bounds for the exponential growth rate [17] and compelling computational evidence [11] that the growth rate is extremely close to the upper bound but not knowing the growth properties prevents the analysis for sampling to understand the asymptotic behavior from being feasible. The second method of sampling [6, 5, 8] analyzes sampling in F by choosing tree pairs of size n uniformly at random, performing the appropriate reductions, and considering the resulting group elements. Here, the trees are selected uniformly at random from all trees of size n. The third way [7], selects trees via a bifurcation process modeled on the Yule distribution [35] for rooted binary trees.

4.3.2. *Distortion of sampling in F.* The Catalan coefficients described here exactly describe the sampling bias between the two methods of tree pair generation for unreduced tree pair diagrams. That is, a given pair $(\hat{s}_n, \hat{t}_n) \in \widehat{\mathcal{T}}_n^2$ has weight 1 with respect to the uniform distribution on trees, and has chance of selection $1/C_n^2$, where C_n is the n-th Catalan number. The same tree pair has weight $B(\hat{s}_n)B(\hat{t}_n)$ where $B(\hat{t}_n)$ is the number of ways that tree $\hat{t}_n$ can arise via a bifurcation process, giving a chance of selection of $B(\hat{s}_n)B(\hat{t}_n)/(n!)^2$. We note that a number of authors have analyzed different properties of these two tree distributions in the unordered case– see, for example, McKenzie and Steel [21] where the distribution of the number of sibling pairs (or "cherries") is analyzed.

Acknowledgments

References

[1] David J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.

[2] Michael G. B. Blum and Olivier François. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, 2006.

[3] José Burillo, Sean Cleary, and Bert Wiest. Computational explorations in Thompson's group *F*. In *Geometric group theory*, Trends Math., pages 21–35. Birkhäuser, Basel, 2007.

[4] J. W. Cannon, W. J. Floyd, and W. R. Parry. Introductory notes on Richard Thompson's groups. *Enseign. Math. (2)*, 42(3-4):215–256, 1996.

[5] Timothy Chu and Sean Cleary. Expected conflicts in pairs of rooted binary trees. *Involve*, 6(3):323–332, 2013.

[6] Sean Cleary, Murray Elder, Andrew Rechnitzer, and Jennifer Taback. Random subgroups of Thompson's group *F*. *Groups Geom. Dyn.*, 4(1):91–126, 2010.

[7] Sean Cleary, John Passaro, and Yasser Toruno. Average reductions in yule-generated binary trees. "to appear".

[8] Sean Cleary, Andrew Rechnitzer, and Thomas Wong. Common edges in rooted trees and polygonal triangulations. *Electron. J. Combin.*, 20(1):Paper 39, 22, 2013.

[9] Robert P. Dobrow and James Allen Fill. On the markov chain for the move-to-root rule for binary search trees. *The Annals of Applied Probability*, 5(1):1–19, 02 1995.

[10] Robert Donaghey. Alternating permutations and binary increasing trees. *Journal of Combinatorial Theory, Series A*, 18(2):141 – 148, 1975.

[11] Murray Elder, Éric Fusy, and Andrew Rechnitzer. Counting elements and geodesics in Thompson's group *F*. *J. Algebra*, 324(1):102–121, 2010.

[12] Murray Elder, Andrew Rechnitzer, and Thomas Wong. On the cogrowth of Thompson's group *F*. *Groups Complex. Cryptol.*, 4(2):301–320, 2012.

[13] James Allen Fill. On the distribution of binary search trees under the random permutation model. *Random Structures & Algorithms*, 8(1):1–25, 1996.

[14] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, New York, NY, USA, 1 edition, 2009.

[15] D. Foata and M.P. Schützenberger. Nombres d'euler et permutations alternantes (unabriged version, 71 pages). Department of mathematics, University of Florida, Gainesville, available from http://www.mat.univie.ac.at/slc/, 1971.

[16] M. P. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *The Annals of Mathematical Statistics*, 41(4):pp. 1344–1346, 1970.

[17] V. S. Guba. On the properties of the Cayley graph of Richard Thompson's group *F*. *Internat. J. Algebra Comput.*, 14(5-6):677–702, 2004. International Conference on Semigroups and Groups in honor of the 65th birthday of Prof. John Rhodes.

[18] Mark Kirkpatrick and Montgomery Slatkin. Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. *Evolution*, 47(4):1171–1181, 1993.

[19] Hosam M. Mahmoud. *Evolution of random search trees*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1992.

[20] G. McGarvey and B. Cloitre. Sequence A000108, The On-line Encyclopedia of Integer Sequences. published electronically, Feb 2005.

[21] Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. *Math. Biosci.*, 164(1):81–92, 2000.

[22] J. Meier. *Groups, graphs and trees: an introduction to the geometry of infinite groups*. London Mathematical Society student texts. Cambridge University Press, Cambridge, 2008.

[23] Fionn Murtagh. Counting dendrograms: a survey. *Discrete Applied Mathematics*, 7(2):191–199, 1984.

[24] R. Sainudiin. Sequence A185155, The On-line Encyclopedia of Integer Sequences. published electronically, Feb 2012.

[25] R Sainudiin and A Veber. A beta-splitting model for evolutionary trees. *UCDMS Research Report 2015/3*, pages 1–20, 2015.

[26] R Sainudiin and D Welch. The transmission process: A combinatorial stochastic process on binary trees over the contact network of hosts in an epidemic. *UCDMS Research Report 2015/4*, pages 1–31, 2015.

[27] Raazesh Sainudiin, Tanja Stadler, and Amandine Véber. Finding the best resolution for the kingmantajima coalescent: theory and applications. *Journal of Mathematical Biology*, pages 1–41, 2014.

[28] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[29] N.J.A. Sloane. Sequence A001190, The On-line Encyclopedia of Integer Sequences. published electronically, May 1995.

[30] N.J.A. Sloane. Sequence A056971, The On-line Encyclopedia of Integer Sequences. published electronically, Nov 2007.

[31] N.J.A. Sloane. Sequence A000111, The On-line Encyclopedia of Integer Sequences. published electronically, May 2013.

[32] Richard P. Stanley. *Enumerative combinatorics. Vol. 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997. With a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original.

[33] Richard P. Stanley. *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.

[34] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

[35] G. U. Yule. A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. *Philos. Trans. Roy. Soc. London Ser. B*, 213:21–87, 1924.

Department of Mathematics, The City College of New York, NY 10031, USA
E-mail address: `scleary@ccny.cuny.edu`

Ernst-Moritz-Arndt University of Greifswald, Department for Mathematics and Computer Science, Walther-Rathenau-Str. 47, 17487 Greifswald, Germany
E-mail address: `email@mareikefischer.de`

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom
E-mail address: `griff@stats.ox.ac.uk`

School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
E-mail address: `raazesh.sainudiin@gmail.com`