Shape Parameters of Evolutionary Trees in Theoretical Computer Science

Michael Fuchs
Department of Mathematical Sciences
National Chengchi University
Taipei 116
Taiwan

April 29, 2024

Abstract

Shape parameters, e.g., balance indices, of evolutionary trees have been extensively studied under the Yule model in phylogenetics. Independently, many of the same parameters have also been studied for random binary search trees in computer science, where they measure the running time of algorithms. In fact, under these two models, these parameters have the same distribution, resulting in many identical discoveries. In this survey, we explain these connections and introduce some of the tools which have been used in computer science to derive stochastic results for shape parameters.

1 Introduction

The *Yule model* (or, more precisely, *Yule-Harding-Kingman model*) is one of the simplest and most basic random models for *evolutionary trees*. It produces *random* evolutionary trees whose properties have been extensively studied via *shape parameters*. These parameters take only the shape of the evolutionary tree into account and thus can be investigated for any binary tree model, for instance for *random binary search trees*. In fact, many of the shape parameters for evolutionary trees under the Yule model have also independently been studied in computer science for the latter model, where they share the same distribution. The goal of this survey is to explain these connections, survey some of the tools which have been used in computer science, and show some of the results which have been proved with these tools (and which in turn also hold for the corresponding parameters of evolutionary trees).

We give a short outline of the survey. In the next section, we recall the definitions of evolutionary trees and the Yule model. In Section 3, we define what we mean by a shape parameter of an evolutionary tree and show that the distribution of such a parameter under the Yule model coincides with the distribution under the random binary search tree model from computer science (which will be defined in this section as well). Then, in Section 4, we define additive shape parameters and explain two important methods from computer science for deriving distributional results for such parameters. In fact, these parameters have also been considered for evolutionary trees and we explain the connections. Finally, in Section 5, we explain a generalization of the Yule model which also includes other random tree models from combinatorics and computer science.

2 Evolutionary Trees and the YHK model

Throughout this survey, we consider rooted binary trees, i.e., trees with a node of degree 2 distinguished as root and all other nodes either of degree 3 (or outdegree 2 if edges are directed away from the root) or

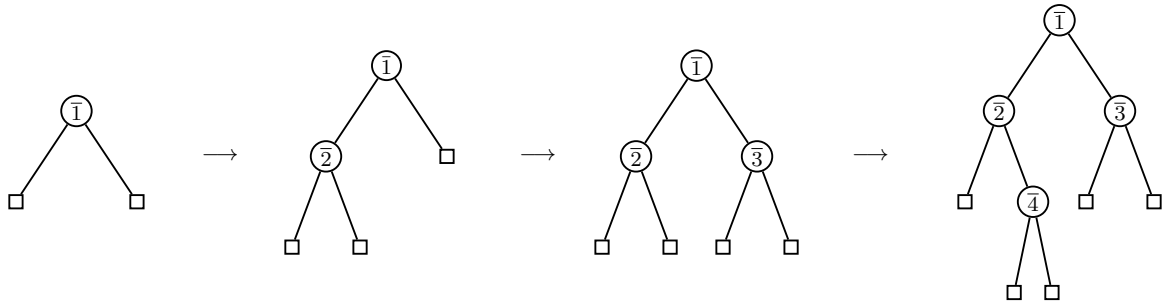


Figure 1: A ranked plane tree constructed in steps (i) & (ii) of the Yule model with probability $1/24$.

1; the former two are called *internal nodes* and the latter *leaves*. Also, we call a tree *plane* if the children (the tips of the outgoing edges) of any internal node have a left-right order and *non-plane* otherwise.

Using these notions, we can now define *evolutionary trees* (or *phylogenetic trees*).

Definition 1 (Evolutionary Tree)**.** An *evolutionary tree*, denoted by t, is a non-plane leaf-labeled tree with the leaves labeled bijectively with elements from the set $\{1, \dots, n\}$, where n is the number of leaves; we write $|t| = n$ and call n the *size* of t. A *tree shape*, denoted by τ, is an evolutionary tree with the labels of the leaves discarded, i.e., τ is merely a non-plane tree.

Evolutionary trees are basic objects in phylogenetics; see [22, 23]. One of the simplest random models for them arose from a seminal paper of Yule [25] and is consequently called the *Yule model*. However, note that the model was not explicitly defined in [25] but rather in a paper of Harding [13] (in which the paper of Yule was not even cited). We use Harding's definition.

Definition 2 (Yule Model)**.** A *random evolutionary tree* of size n under the Yule model is generated by the following steps:

(i) Start with the (unique) plane tree with two leaves (which is called a *cherry*);

(ii) Recursively create a sequence of plane trees as follows: Pick uniformly at random a leaf and replace it by a cherry; do this until n leaves are obtained.

(iii) Label the n leaves in depth-first order by a permutation which is picked uniformly at random from all permutations of the set $\{1, \dots, n\}$;

(iv) Forget the order of the children of all internal nodes.

Remark 1.

- If the roots of the cherries in steps (i) & (ii) are labeled by $\overline{1}, \overline{2}, \dots$ in the order in which they are created, we obtain after the first two steps all plane trees of size n whose $n-1$ internal nodes are labeled so that the labels along paths from the root to a leaf are increasing; see Figure 1. These trees are called *ranked plane trees*. Clearly, there are exactly $(n-1)!$ such trees of size n.
- Likewise, after the third step, we obtain all *ranked plane leaf-labeled trees*; their number is given by $n!(n-1)!$. Note that the first three steps generate all these trees uniformly at random.

Every outcome of an instance of the Yule model has probability $1/(n!(n-1)!)$; see item (ii) of the above remark. However, note that the same evolutionary tree t may result from many different outcomes. More precisely, we have the following (well-known) result for the probability of t.

Theorem 1. *The probability of an evolutionary tree t of size n under the Yule model is given by*

$$\mathbb{P}(t) = \frac{2^{n-1}}{n!} \prod_{v \in I(t)} \frac{1}{|t_v| - 1}, \tag{1}$$

where $I(t)$ denotes the set of internal nodes of t and t_v is the tree rooted at v which consists of v and all the descendants of v.

Proof. We count the number of instances which produce t in the random process underlying the Yule model (by following the random process backwards). First, there are 2^{n-1} ways of embedding t such that the resulting tree is a plane leaf-labeled tree. Next, we need to count the number of ways of ranking these trees. Let $\tilde{t}$ denote such an embedding. Then, the number of rankings of $\tilde{t}$, denoted by $\mathrm{rank}(\tilde{t})$, is recursively given by:

$$\mathrm{rank}(\tilde{t}) = \binom{n-2}{|\tilde{t}^\ell|-1}\mathrm{rank}(\tilde{t}^\ell)\mathrm{rank}(\tilde{t}^r), \tag{2}$$

where $\tilde{t}^\ell$ and $\tilde{t}^r$ denote the subtrees rooted at the left and right child of the root, respectively. This recurrence is explained as follows: any ranking is obtained by distributing the $n-2$ labels of internal nodes minus the label of the root to the subtrees $\tilde{t}^\ell$ and $\tilde{t}^r$ and then ranking these two subtrees. By iterating (2),

$$\mathrm{rank}(\tilde{t}) = (n-2)! \prod_{v\in I(t)\setminus\{r\}} \frac{1}{|\tilde{t}_v|-1} = (n-1)! \prod_{v\in I(t)} \frac{1}{|t_v|-1}, \tag{3}$$

where r denotes the root of $\tilde{t}$ and the last step follows by including the root in the product and noting that $|\tilde{t}_v| = |t_v|$. The claimed result follows from this since $\mathbb{P}(t) = 2^{n-1}\mathrm{rank}(\tilde{t})/(n!(n-1)!)$. ∎

Note that (3) is also the number of possible rankings of an evolutionary tree t, i.e., the number of *ranked non-plane leaf-labeled* trees which give t if the ranking is discarded. Moreover, the total number of these trees of size n, by item (ii) of Remark 1 upon forgetting the order of the children of internal nodes, is given by $n!(n-1)!/2^{n-1}$. Thus, by picking one such tree uniformly at random and discarding the ranking, an evolutionary tree t is again obtained with probability given by (1). Thus, this is another way of defining the Yule model. In fact, the uniform generation of ranked non-plane leaf-labeled trees of size n can be done as follows: start with n items labeled by $\{1,\dots,n-1\}$ and recursively take two items and replace them by one which is labeled by $\{\overline{1},\dots,\overline{n-1}\}$ where we start with $\overline{n-1}$ until $\overline{1}$ is reached. This random process is the *Kingman's coalescent* from population genetics; see [16].

Remark 2. Due to Harding's and Kingman's contributions to the Yule model, the Yule model is sometimes also (more precisely) called *Yule-Harding-Kingman model* (or *YHK model* for short). However, in this paper, we will only refer to it as Yule model.

3 Shape Parameters and Random Binary Search Trees

We start with the definition of a *shape parameter*.

Definition 3 (Shape Parameter)**.** A *shape parameter*, denoted by X_n, is a mapping which assigns every tree shape of size n a real number.

Clearly, any shape parameter can be extended to a class of trees of size n where trees are in addition ranked and/or plane and/or leaf-labeled by assigning the same value to all trees of size n with the same tree shape. With a slight abuse of notation, we use X_n to denote the same shape parameter for all these tree classes.

Many shape parameters have been defined and studied in phylogenetics, e.g., shape parameters which are measures of the balance (or imbalance) of trees; see [10]. We introduce two of them, where the second strictly speaking does not satisfy the definition of an *(im)balance index* from [10] but is nevertheless often included in the list of (im)balance indices.

Example 1. (i) The *Sackin index*, denoted by S_n, of a tree shape τ of size n is defined as the sum of the root-to-leaf distances over all leaves.

(ii) The *cherry index*, denoted by C_n, of a tree shape τ of size n is defined as the number of cherries of τ, i.e., number of internal nodes with both children being leaves.

Another shape parameter, the number of *ancestral configurations*, was investigated in [8, 9], where it was shown that its distribution under the Yule model coincides with its distribution for ranked plane trees which are picked uniformly at random. This is, in fact, a general phenomenon which holds for any shape parameter.

Theorem 2. *The distribution of a shape parameter X_n under the Yule model is the same as the distribution under the uniform model on ranked plane trees of size n.*

Proof. Let t be a ranked non-plane tree of size n. The claim follows by showing that

$$\frac{\mathrm{ord}(t)}{(n-1)!} = \frac{\mathrm{lab(t)}}{n!(n-1)!2^{1-n}},$$

where $\mathrm{ord}(t)$ and $\mathrm{lab}(t)$ are the number of different ranked plane trees and ranked non-plane leaf-labeled trees which correspond to t (after forgetting the ordering of internal nodes resp. discarding leaf labels). Note that the denominators are the number of ranked plane trees (see item (i) of Remark 1) and the number of ranked non-plane leaf-labeled trees (see paragraph after Theorem 1) of size n.

Equivalently, we have to show that

$$\mathrm{ord}(t) = \frac{2^{n-1}}{n!}\mathrm{lab}(t)$$

which follows by taking a ranked non-plane leaf-labeled tree, choosing an order of the children of every internal node, and then discarding the leaf labels. ∎

We can now make the promised connection to computer science. Therefore, we recall the definition of a *(random) binary search tree*, which is a fundamental data structure in computer science; see, e.g., [17, 18] and Figure 2, (a) for an example.

Definition 4 (Binary Search Tree)**.** A *binary search tree* is a plane tree built from a permutation of $\{\overline{1}, \ldots, \overline{n-1}\}$ as follows: the first element goes to the root; all other elements are distributed to the left and right subtree of the root according to whether they are smaller and larger than the root; the subtrees are built recursively using the same rules; finally, an unlabeled left and/or right leaf is added to a node if the left and/or right subtree of the node is empty.

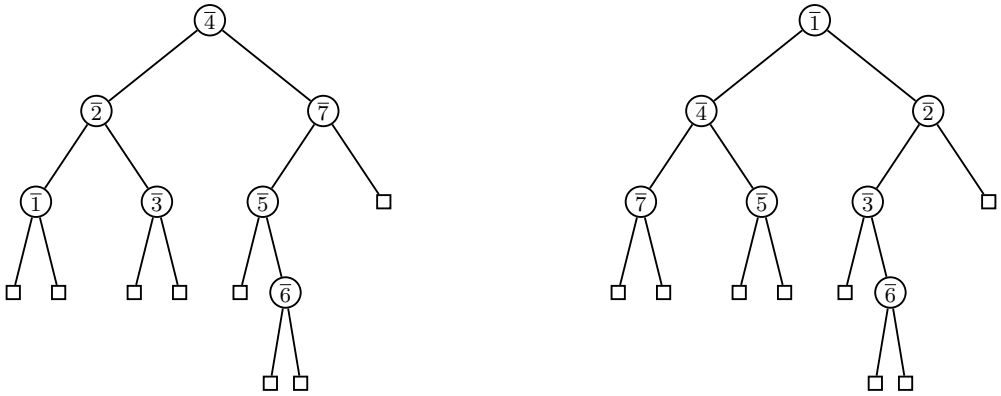
Moreover, a *random binary search tree* is a binary search tree which is built from a random permutation of $\{\overline{1}, \ldots, \overline{n-1}\}$.

Shape parameters of binary search trees have been extensively studied as they are measures of the complexity of algorithms performed on binary search trees. For instance, the Sackin index from Example 1 is related to the *unsuccessful search* in a binary search tree, i.e., a search starting from the root which ends at a leaf; see [17, 18]. Also, since binary search trees are closely related to the *quicksort algorithms* from computer science, the Sackin index is also closely related to an important complexity measure of quicksort, namely, the *number of key comparisons*; see [12]. The number of cherries and more complicated *patterns* have been studied for binary search trees as well; see, e.g., [11].

An important observation is that shape parameters under the Yule model have the same distribution as shape parameters for random binary search trees.

Theorem 3. *The distribution of a shape parameter X_n under the Yule model is the same as the distribution under the random binary search tree model.*

Proof. The claim follows by constructing a map that maps a permutation of the set $\{\overline{1}, \ldots, \overline{n-1}\}$ bijectively to a ranked plane tree of size n that has the same tree shape as the binary search tree built from the permutation. Such a map is obtained by using the recursive procedure for building the binary search tree but storing, instead of the elements of the permutation, the positions of the elements in the permutation into the internal nodes of the tree; see Figure 2, (b). ∎



(a) (b)

Figure 2: (a) A binary search tree built from the permutation $\overline{4}\,\overline{7}\,\overline{5}\,\overline{2}\,\overline{3}\,\overline{6}\,\overline{1}$; (b) The ranked plane tree which is the image of the permutation from (a) under the map in the proof of Theorem 3.

4 Analysis of Additive Shape Parameters

In the sequel, we restrict ourselves to an important "subclass" of shape parameters, namely, *additive shape parameters*.

Definition 5 (Additive Shape Parameter)**.** A shape parameter $X(\tau)$, where τ is a tree shape, is called an *additive shape parameter* if $X(\tau)$ can be recursively computed from the two subtrees of the roots, denoted by τ^ℓ and τ^r, respectively, as follows:

$$X(\tau) = X(\tau^\ell) + X(\tau^r) + T(\tau), \tag{4}$$

where $T(\tau)$ is a function from tree shapes into real numbers which is called the *toll function*.

Example 2. (i) The Sackin index is an additive shape parameter with $T(\tau) = |\tau|$;

(ii) The cherry index is an additive shape parameter with $T(\tau) = 1$ if $|\tau| = 2$ and 0 otherwise.

Note that with the above definition, actually any shape parameter is additive (by a suitable choice of the toll function). Thus, further restrictions are necessary in order to be able to prove (meaningful) general results, i.e., results which hold for a whole class of additive shape parameters; see, e.g., [14, 15, 24]. As in these papers, we are interested in stochastic results for $X(t)$ when the tree t is random (and t is not necessarily just a tree shape).

In the sequel, we denote by X_n the random variable obtained by considering t to be a random binary search tree of size n (or, equivalently, an evolutionary tree of size n under the Yule model). Moreover, we assume that T_n (the random variable corresponding to the toll function) is deterministic (as is the case, e.g., for the Sackin index and the cherry index; see the example above). Then, (4) translates into a distributional recurrence for X_n. Before stating it, we need the following lemma which shows that the left subtree of a random binary search tree of size n has a uniform distribution. (Note that the corresponding property for the Yule model is well-known; see, e.g., Lemma 3.1 in [23].)

Lemma 4. *Let I_n denote the size of the left subtree of the root in a random binary search tree of size n. Then,*

$$\mathbb{P}(I_n = j) = \frac{1}{n-1}, \qquad (1 \le j \le n-1).$$

Proof. Recall that a random binary search tree of size n is built from a random permutation of the set $\{\overline{1}, \ldots, \overline{n-1}\}$. The left subtree of the root has size j if and only if the first element of this random permutation is $\overline{j}$ which clearly happens with probability $1/(n-1)$ as claimed. ∎

Proposition 5. *Let X_n be an additive shape parameter of a random binary search tree of size n. Then,*

$$X_n \stackrel{d}{=} X_{I_n} + X^*_{n-I_n} + T_n, \qquad (n \geq 2), \tag{5}$$

where I_n is as in the last lemma, $X^*_n \stackrel{d}{=} X_n$, and $(X_n)_{n=1}^\infty$, $(X^*_n)_{n=1}^\infty$, and $(I_n)_{n=1}^\infty$ are independent.

This distributional recurrence is the starting point of many studies on additive shape parameters; see [14, 15, 24]. Here, we are going to explain two powerful and (very) general methods for deriving moments and limit laws, namely, the *moment-transfer approach* and the *contraction method*. (Both methods also work if T_n is random and satisfies suitable conditions.) More precisely, we explain in detail the former and then briefly comment on the latter.

First, note from (5) that all (centered or non-centered) moments of X_n satisfy a recurrence of the form (for details, see the proof of Lemma 6 and Proposition 8 below):

$$a_n = \frac{2}{n-1}\sum_{j=1}^{n-1} a_j + b_n, \qquad (n \geq 2), \tag{6}$$

where (for the sake of simplicity) $a_1 = 0$. This recurrence is readily solved:

$$a_n = 2n\sum_{j=2}^{n-1} \frac{b_j}{j(j+1)} + b_n, \qquad (n \geq 2). \tag{7}$$

From this, e.g., for the cherry index C_n, we immediately obtain the following well-known result; see, e.g., [10] and references therein.

Lemma 6. *We have,*

$$\mathbb{E}(C_n) = \frac{n}{3}, \quad (n \geq 3) \qquad \textit{and} \qquad \mathbb{V}(C_n) = \frac{2n}{45}, \quad (n \geq 5).$$

Proof. Note that the mean satisfies (6) with $b_n = T_n$ where T_n is given in item (ii) of Example 2. Thus, from (7), for $n \geq 3$,

$$\mathbb{E}(C_n) = 2n\frac{b_2}{6} = \frac{2n}{3}$$

as claimed. For the variance, a little bit more computation is necessary. We leave the details to the reader. (The sequence b_n for the variance is given in the proof of Proposition 8 below.) ∎

This can be extended to higher moments but it is now better to aim for asymptotic results. For this, we need a so-called *asymptotic transfer result* for (6); see, Lemma 2 in [14]. (We only give a simplified version of the latter result.)

Lemma 7 (Asymptotic Transfer)**.** *(i) If $b_n = \mathcal{O}(n^{1-\epsilon})$, where $\epsilon > 0$ is arbitrarily small, then, $a_n \sim cn$ with*

$$c = 2\sum_{j\geq 2} \frac{b_j}{j(j+1)}.$$

(ii) If $b_n \sim cn^\alpha$, where $\alpha > 1$, then $a_n \sim (\alpha+1)cn^\alpha/(\alpha-1)$.

Proof. (i) This follows immediately from (7) by extending the range of summation to infinity. (Note that the resulting series then converges by the assumption.)

(ii) By plugging the assumption into (7), we have

$$a_n \sim 2cn\sum_{j=2}^{n-1} j^{\alpha-2} + cn^\alpha \sim 2cn^\alpha \int_0^1 x^{\alpha-2}\mathrm{d}x + cn^\alpha = \left(\frac{2}{\alpha-1}+1\right)cn^\alpha,$$

where in the second asymptotic equivalence, we have approximated the sum by an integral. The claim follows from this. ∎

This result can now be used together with induction to obtain the first-order asymptotics of all central moments of C_n which satisfy (6) where b_n is a function of moments of smaller order; see (8) and (9) below. (This method has been nicknamed *moment pumping* in computer science.)

Proposition 8. *For all $m \geq 0$,*

$$\mathbb{E}(C_n - \mathbb{E}(C_n))^m \sim g_m \left(\frac{2n}{45}\right)^{m/2},$$

where g_m is the m-th moment of the standard normal distribution $N(0,1)$.

Remark 3. We use the (convenient but slightly unusual) convention that $f(n) \sim cg(n)$ with $c = 0$ means that $f(n) = o(g(n))$.

Proof. We prove the claim by induction on m. Note that the result holds trivially for $m = 0$ and $m = 1$ (with the convention from Remark 3), and also for $m = 2$ because of Lemma 6.

Assume now that it holds for all $m' < m$. In order to prove it for $\phi_n^{[m]} := \mathbb{E}(C_n - \mathbb{E}(C_n))^m$, first, by a straightforward computation (which starts from (5)):

$$\phi_n^{[m]} = \frac{2}{n-1} \sum_{j=1}^{n-1} \phi_j^{[m]} + \psi_n^{[m]}, \tag{8}$$

where

$$\psi_n^{[m]} = \frac{1}{n-1} \sum_{j=1}^{n-1} \sum_{\substack{a+b+c=m \\ a,b<m}} \binom{m}{a,b,c} \phi_j^{[a]} \phi_{n-j}^{[b]} \Delta(j,n)^c \tag{9}$$

with

$$\Delta(j,n) = T_n + \mathbb{E}(C_j) + \mathbb{E}(C_{n-j}) - \mathbb{E}(C_n)$$

and T_n is given in item (ii) of Example 2. Note that from Lemma 6, we have that $\Delta(j,n) = \mathcal{O}(1)$ which holds uniformly in j and n.

Next, by using the induction hypothesis,

$$\begin{aligned}
\psi_n^{[m]} &= \frac{1}{n-1} \sum_{j=1}^{n-1} \sum_{\substack{a+b+c=m \\ a,b<m}} \binom{m}{a,b,c} \phi_j^{[a]} \phi_{n-j}^{[b]} \Delta(j,n)^c \\
&\sim \frac{1}{n-1} \sum_{j=1}^{n-1} \sum_{\substack{a+b=m \\ a,b<m}} \binom{m}{a,b} g_a \left(\frac{2j}{45}\right)^{a/2} g_b \left(\frac{2(n-j)}{45}\right)^{b/2} \\
&\sim \left(\frac{2n}{45}\right)^{m/2} \sum_{\substack{a+b=m \\ a,b<m}} \binom{m}{a,b} g_a g_b \frac{1}{n-1} \sum_{j=1}^{n-1} \left(\frac{j}{n}\right)^{a/2} \left(1 - \frac{j}{n}\right)^{b/2} \\
&\sim \left(\frac{2n}{45}\right)^{m/2} \sum_{\substack{a+b=m \\ a,b<m}} \binom{m}{a,b} g_a g_b \int_0^1 x^{a/2} (1-x)^{b/2} \mathrm{d}x,
\end{aligned}$$

where the first asymptotics equivalence follows since the terms with $c > 0$ do not contribute to the main term (here, we have used that $\Delta(j,n) = \mathcal{O}(1)$) and the last asymptotics equivalence follows by approximating the sum by an integral.

A straightforward computation reveals that

$$\sum_{\substack{a+b=m\\ a,b<m}} \binom{m}{a,b} g_a g_b \int_0^1 x^{a/2}(1-x)^{b/2}\mathrm{d}x = \frac{m-1}{m+1} g_m.$$

Thus,

$$\psi_n^{[m]} \sim \frac{m-1}{m+1} g_m \left(\frac{2n}{45}\right)^{m/2}$$

from which the claim follows by item (ii) of Lemma 7. ∎

Consequently, from the method of moments in probability theory (see Chapter 30 in [4]), we obtain the central limit theorem

$$\frac{C_n - n/3}{\sqrt{2n/45}} \xrightarrow{d} N(0,1)$$

which for evolutionary trees was first proved in [19] (with a completely different method).

Remark 4. In fact, much more is known, namely, any *pattern* is asymptotically normal; see [14] and [11] where this was proved with the *singularity perturbation method*. Also, this can be extended to patterns whose size is even allowed to moderately grow with n; see [6].

The moment-transfer method can also be applied to the Sackin index S_n. Here, we first obtain from (7) the following result for the mean and variance; see, e.g., [10].

Lemma 9. *We have,*

$$\mathbb{E}(S_n) = 2n(H_n - 1) \sim 2n\log n \qquad \textit{and} \qquad \mathbb{V}(S_n) = 7n^2 - 4n^2 H_n^{(2)} - 2nH_n - n,$$

where $H_n = \sum_{j=1}^n (1/j)$ *and* $H_n^{(2)} = \sum_{j=1}^n (1/j^2)$ *are the first and second Harmonic numbers.*

Also, again from the asymptotic transfer result and moment pumping, we obtain for the central moments; see [14].

Proposition 10. *For* $m \geq 0$,

$$\mathbb{E}(S_n - \mathbb{E}(S_n))^m \sim c_m n^m,$$

where c_m *is recursively given by* $c_0 = 1$, $c_1 = 0$, *and*

$$c_m = \frac{m+1}{m-1} \sum_{\substack{a+b+c=m\\ a,b<m}} \binom{m}{a,b,c} c_a c_b \int_0^1 x^a (1-x)^b \Lambda(x)^c \mathrm{d}x, \qquad (m \geq 2) \tag{10}$$

with $\Lambda(x) = 2x\log x + 2(1-x)\log(1-x) + 1$.

In order to obtain a limit distribution result from this, we need to show that there exists a random variable which is uniquely determined by the moment sequence c_m. For this, we use well-known results from probability theory (see Chapter 30 in [4]), e.g., one sufficient condition that such a random variable exists is that the power series

$$\sum_{m=0}^{\infty} c_m \frac{x^m}{m!}$$

has a non-negative radius of convergence. This follows from an estimate for c_m of the form

$$|c_m| \leq A^m m!$$

for a suitable constant $A > 0$ which is proved by induction and (10). Thus, there exists an S that is uniquely determined by its moments $c_m := \mathbb{E}(S^m)$. Consequently, from the method of moments, we have

$$\frac{S_n - 2n\log n}{n} \xrightarrow{d} S.$$

The moment-transfer approach can be applied to a great number of additive shape parameters of random binary search trees and thus evolutionary trees under the Yule model. To give one more example, we consider the *total cophenetic index* Φ_n which is defined as the sum of distances from the root to the lowest common ancestor over all pairs of (different) leaves. This index, a balance index according to the definition in [10], was introduced in [20]. For evolutionary trees under the Yule model, the moment-transfer approach was used in [7] to derive the following limit distribution result.

Theorem 11. *For the cophenetic index Φ_n of evolutionary trees of size n under the Yule model,*

$$\frac{\Phi_n}{n^2} \xrightarrow{d} \Phi,$$

where Φ is uniquely determined by its moment sequence. More precisely, we have $\mathbb{E}(\Phi^m) = d_m$ with $d_0 = 1$ and

$$d_m = \frac{1}{(2m)!(2m-1)} \sum_{\substack{a+b+c=m \\ a,b<m}} \binom{m}{a,b,c} d_a d_b \sum_{j=0}^{c} \binom{c}{j} \frac{(2a+2j)!(2b+2c-2j)!}{2^c}, \qquad (m \geq 2).$$

Remark 5. The other method mentioned at the beginning of this section, namely the *contraction method*, also starts from (5). This method is based on Banach's fixed-point theorem and a careful choice of a distance; readily applicable black-box results are available (see [21]). However, the method sometimes requires the knowledge of the asymptotics of mean and/or variance and then can only be used in connection with tools from the moment-transfer approach; see [14]. On the other hand, if the latter is not the case, then the contraction method also does yield asymptotic expansions of mean and/or variance.

5 A Generalization of the Yule Model

We give a reformulation of the definition of the Yule model. For this purpose, we first show that the random plane tree created in the first two steps of the Yule process satisfies the property from Lemma 4 for random binary search trees.

Lemma 12. *The size of the left subtree of the random plane tree of size n generated in steps (i) & (ii) of the Yule process is uniformly distributed on $\{1, \dots, n-1\}$.*

Proof. We work with ranked plane trees; see item (i) of Remark 1. The left subtree has size j if $j-1$ out of the $n-2$ ranks (for the internal nodes excluding the root) go to the left subtree. Thus, the probability that the left subtree has size j with $1 \leq j \leq n-1$ equals

$$\binom{n-2}{j-1} \frac{(j-1)!(n-1-j)!}{(n-1)!} = \frac{1}{n-1}.$$

This proves the claim. ∎

Instead of using steps (i) & (ii) of the Yule process, we can generate a random plane tree of size n also as follows (see Figure 3):

(i') Pick n dots uniformly at random in the unit interval;

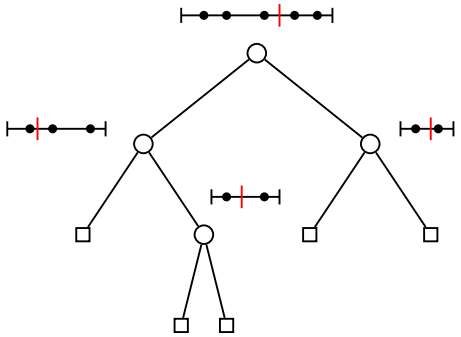


Figure 3: A plane tree constructed by step (i') & (ii'); the interval with dots are next to the node and the red vertical line indicates where the interval is cut.

(ii') Split the unit interval into two (non-empty) subintervals by cutting at a point chosen uniformly at random (if one of the subintervals is empty repeat this step); the dots in the two subintervals are the leaves of the left and right subtree of the root; recursively repeat this step with the two (re-scaled) subintervals until each subinterval just contains one dot.

This in fact, gives the same random model on plane trees.

Lemma 13. *The size of the left subtree of the random plane tree of size n generated by steps (i') & (ii') is uniformly distributed on $\{1,\dots,n-1\}$.*

Proof. The probability that the left subinterval after splitting the unit interval from step (i') contains j dots equals

$$\int_0^1 \binom{n}{j} x^j (1-x)^{n-j} \mathrm{d}x = \binom{n}{j} \beta(j+1, n+1-j) = \binom{n}{j} \frac{j!(n-j)!}{(n+1)!} = \frac{1}{n+1},$$

where $\beta(a,b)$ denotes the β-function. However, j is not allowed to be 0 or n. Thus, the probability that the left subtree has size j with $1 \le j \le n-1$ equals

$$\frac{1/(n+1)}{1-2/(n+1)} = \frac{1}{n-1}$$

which proves the claim. ∎

Consequently, we can replace steps (i) & (ii) in Defintion 2 by steps (i') & (ii') above. The advantage of this is that this modified definition can easily be generalized by using in step (ii') a probability distribution for the cutting which is different from the uniform distribution. In [1], it was suggested to use a β-distribution, i.e., a continuous distribution with density:

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)^2} x^\beta (1-x)^\beta, \qquad x \in [0,1], \qquad (\beta > -1),$$

where $\Gamma(x)$ denotes the Γ-function. The resulting random model on evolutionary trees of size n is called the *β-splitting model*.

For this choice of the distribution, the size of the left subtree of the random plane tree generated after the first two steps has probability:

$$p_{n,j} = \frac{1}{c(\beta)} \frac{\Gamma(\beta+j+1)\Gamma(\beta+n-j+1)}{j!(n-j)!}, \qquad (1 \le j \le n-1),$$

where $c(\beta)$ is a suitable normalization constant (so that $\sum_{j=1}^{n-1} p_{n,j} = 1$). Note that this expression makes also sense for $-2 < \beta \leq -1$ and we can thus extend the range of β to $\beta > -2$.

We conclude by pointing out some important choices of β for which we recover other models from computer science and phylogenetics.

(i) $\beta = \infty$: This gives the random trie model from computer science; see Chapter 5 in [18];

(ii) $\beta = 0$: This is the Yule model;

(iii) $\beta = -1$: The resulting model apparently provides a good fit to many real-world trees; see [5]. We have

$$p_{n,j} = \frac{1}{H_{n-1}} \cdot \frac{1}{j(n-j)}, \qquad (1 \leq j \leq n-1)$$

and additive shape parameters still satisfy (5) but with $P(I_n = j) = p_{n,j}$. However, extending the tools from Chapter 4 has so far turned out to be largely elusive. Nevertheless, there has recently been some progress on this case; see [2, 3].

(iv) $\beta = -3/2$: Here, we have

$$p_{n,j} = \frac{C_{j-1}C_{n-j-1}}{C_{n-1}}, \qquad (1 \leq j \leq n-1),$$

where C_n is the n-th Catalan number, which counts plane trees of size $n+1$. Thus, in this case, we obtain the *uniform model* (or *PDA model*) of evolutionary trees which is another fundamental model in phylogenetics; see [23]. (This model has also been studied in combinatorics where it is usually called the *Catalan model*; see [12].)

Acknowledgement

The author acknowledges partial support by the National Science and Technology Council (NSTC), Taiwan under grant NSTC-111-2115-M-004-002-MY2.

References

[1] D. Aldous (1996). Probability distributions on cladograms, *Random discrete structures (Minneapolis, MN, 1993)*, **76**, 1–18.

[2] D. Aldous. The critical beta-splitting random tree model II: Overview and open problems, arXiv:2303.02529.

[3] D. Aldous and B. Pittel. The critical beta-splitting random tree: Heights and related results, arXiv:2302.05066.

[4] P. Billingsley. *Probability and Measure*, 3rd edition, John Wiley & Sons, New York, 1995.

[5] M. G. B. Blum and O. Francqis (2006). Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance, *Syst. Biol.*, **55**, 685–691.

[6] H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

[7] L.-A. Chen. *Probabilistic Analysis of the Total Cophenetic Index in Phylogenetic Trees*, master thesis, National Chiao Tung Univesity (NCTU), 2015.

[8] F. Disanto, M. Fuchs, C.-Y. Huang, A. R. Paningbatan, N. A. Rosenberg (2024). The distributions under two species-tree models of the total number of ancestral configurations for matching gene trees and species trees, *Adv. in Appl. Math.*, **152**, 102594.

[9] F. Disanto, M. Fuchs, A. R. Paningbatan, N. A. Rosenberg (2023). The distributions under two species-tree models of the number of root ancestral configurations for matching gene trees and species trees, *Ann. Appl. Probab.*, **32:6**, 4426–4458.

[10] M. Fischer, L. Herbst, S. Kersting, L. Kühn, K. Wicke. *Tree Balance Indices: A Comprehensive Survey*, Springer, 1st edition, 2023.

[11] P. Flajolet, X. Gourdon, C. Martinez (1997). Patterns in random binary search trees, *Random Struc. Algor.*, **11:3**, 223–244.

[12] P. Flajolet and R. Sedgewick. *An Introduction to the Analysis of Algorithms*, Addison-Wesley, 2nd edition, 2013.

[13] E. F. Harding (1971). The probabilities of rooted tree-shapes generated by random bifurcation, *Adv. App. Probab.*, **3:1**, 44–77.

[14] H.-K. Hwang and R. Neininger (2002). Phase change of limit laws in the Quicksort recurrence under varying toll functions, *SIAM J. Comput.*, **31**, 1687–1722.

[15] S. Janson (2022). Central limit theorems for additive functionals and fringe trees in tries, *Electron. J. Probab.*, **27**, Paper No. 47.

[16] J. F. C. Kingman (1982). The coalescent, *Stochastic Process. Appl.*, **13:3**, 235–248.

[17] D. E. Knuth. *The Art of Computer Programming, Volume III, Sorting and Searching*, Addison-Wesley, 2nd edition, 1998.

[18] H. M. Mahmoud. *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[19] A. McKenzie and M. Steel (2000). Distributions of cherries for two models of trees, *Math. Biosci.*, **164**, 81–92.

[20] A. Mir, F. Rosselló, L. Rotger (2013). A new balance index for phylogenetic trees, *Math. Biosci.*, **241:1**, 125–136.

[21] R. Neininger and L. Rüschendorf (2004). A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.*, **14:1**, 378–418.

[22] C. Semple and M. Steel. *Phylogenetics*, Oxford University Press, 2003.

[23] M. Steel. *Phylogeny—Discrete and Random Processes in Evolution*, CBMS-NSF Regional Conference Series in Applied Mathematics, 89, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

[24] S. Wagner (2015). Central limit theorems for additive tree parameters with small toll functions, *Combin. Probab. Comput.*, **24:1**, 329–353.

[25] G. U. Yule (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S, *Philos. Trans. R. Soc. B, Biol. Sci.*, **213**, 21–87.