PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Check for updates

Article submitted to journal

**Subject Areas:**

xxxxx, xxxxx, xxxx

**Keywords:**

Yule model, binary search tree, shape parameter, moment-transfer approach, contraction method, $\beta$-splitting model

**Author for correspondence:**

Michael Fuchs

e-mail: mfuchs@nccu.edu.tw

# Shape Parameters of Evolutionary Trees in Theoretical Computer Science

Michael Fuchs[1]

[1] Department of Mathematical Sciences
National Chengchi University
Taipei 116
Taiwan

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 the Yule and binary search tree 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* or *YHK model* for short) 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 an important method 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. In Section 5, we explain a generalization of the Yule model which also includes other random tree models from combinatorics and computer science. We conclude the paper by a summary in Section 6.

THE ROYAL SOCIETY
PUBLISHING

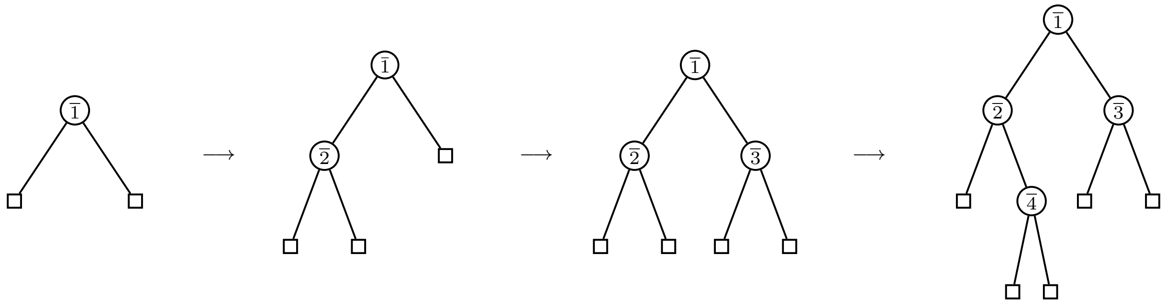


**Figure 1.** A ranked plane tree obtained in steps (i) & (ii) of the Yule model with probability $1/(5-1)! = 1/24$.

## 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 1; the former two are called *internal nodes* and the latter *leaves*. Also, we call a tree *plane* (or *ordered*) if the children (the tips of the outgoing edges) of any internal node have a left-right order and *non-plane* (or *unordered*) 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 $\tau$, is an evolutionary tree with the labels of the leaves discarded, i.e., $\tau$ is merely a non-plane tree.

Evolutionary trees are basic objects in phylogenetics; see [38,39]. One of the simplest random models for them arose from the seminal paper of Yule [41] and is consequently called the *Yule model*. However, note that the model was not explicitly defined in [41] but rather in a paper of Harding [25]. 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[1] 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 $\bar{1},\bar{2},\dots$ 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*. Note that there are exactly $(n-1)!$ such trees of size $n$. (Another name for ranked trees is *histories*; see [29] and references therein.)
- 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 the second part of Remark 1. 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*

$$\mathbb{P}(t) = \frac{2^{n-1}}{n!}\prod_{v\in I(t)}\frac{1}{|t_v|-1}, \tag{2.1}$$

*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.

[1] Recall that a *depth-first search* of a tree starts from the root and traverses every node by branching always along right edges until this is not possible anymore in which case the algorithm backtracks to the first left edge that has not been traversed yet along which the search is continued.

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:

$$\mathrm{rank}(\tilde{t}) = \binom{n-2}{|\tilde{t}^{\ell}|-1}\mathrm{rank}(\tilde{t}^{\ell})\mathrm{rank}(\tilde{t}^{r}), \tag{2.2}$$

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 excluding the label of the root to the subtrees $\tilde{t}^{\ell}$ and $\tilde{t}^{r}$ and then ranking these two subtrees. By iterating (2.2),

$$\mathrm{rank}(\tilde{t}) = (n-2)! \prod_{v\in I(t)\setminus\{r\}} \frac{1}{|\tilde{t}_v|-1} = (n-1)! \prod_{v\in I(t)} \frac{1}{|t_v|-1}, \tag{2.3}$$

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 $\mathbb{P}(t) = 2^{n-1}\mathrm{rank}(\tilde{t})/(n!(n-1)!)$, where $2^{n-1}$ and $1/(n!(n-1)!)$ are the number of embeddings and the probability, respectively, of a random instance which produces $t$. ∎

Note that (2.3) 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 (2.1). 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 [30].

*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.

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., balance indices which are shape parameters that measure the balance (or imbalance) of a tree; see [19]. We introduce two of them, where the second strictly speaking does not satisfy the definition of an *(im)balance index* from [19] 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 $\tau$ 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 $\tau$ of size $n$ is defined as the number of cherries of $\tau$, i.e., number of internal nodes with both children being leaves.

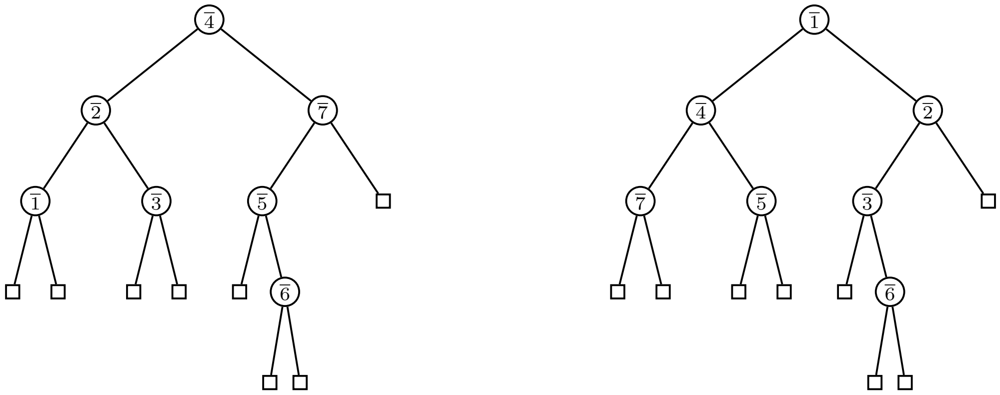
Another (quite different) shape parameter, which arises in the context of gene trees and species trees, is the number of *ancestral configurations*. This parameter was investigated in [15,16] where it was shown that its distribution under the Yule model coincides with its distribution for ranked plane 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

$$\frac{\mathrm{pla}(t)}{(n-1)!} = \frac{\mathrm{lab(t)}}{n!(n-1)!2^{1-n}}, \tag{3.1}$$

where $\mathrm{pla}(t)$ and $\mathrm{lab}(t)$ are the number of different ranked *pla*ne trees and ranked non-plane leaf-*lab*eled trees, respectively, which correspond to $t$ (after forgetting the ordering of the children of the internal nodes resp. discarding leaf labels). Note that the shape parameter has the same value for the latter two classes of trees and the denominators in (3.1) are the number of ranked plane trees



(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.

(see item (i) of Remark 1) and the number of ranked non-plane leaf-labeled trees (see paragraph after Theorem 1) of size $n$. Thus, (3.1) is an identity for two probabilities.

We can re-formulate (3.1) as follows

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

which is proved by taking any ranked non-plane leaf-labeled tree corresponding to $t$ ( counted by lab(t) on the right-hand side), choosing an order of the children of every internal node (counted by the factor $2^{n-1}$), and then discarding the leaf labels (which gives the factor $1/n!$). This yields all ranked plane trees corresponding to $t$, i.e., the left-hand side of the claim. ∎

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., Section 6.2.2 in [31], Chapter 2 in [32], and Figure 2, (a) for an example.

**Definition 4** (Binary Search Tree)**.** A *binary search tree* of size $n$ is a plane tree built from a permutation of $\{\overline{1},\dots,\overline{n-1}\}$ 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 or 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 $\{\overline{1},\dots,\overline{n-1}\}$.

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 Section 6.2.2 in [31] and Section 2.4 in [32]. Also, since binary search trees are closely related to the *quicksort algorithms* from computer science, the Sackin index is also closely related to one of the most widely studied complexity measures of quicksort, namely, the *number of key comparisons*; see Section 1.5 in [22] and Example 2 below. The number of cherries and more general *patterns* have been studied for binary search trees as well; see, e.g., [20].

An important observation is that shape parameters under the Yule model have the same distribution as shape parameters for random binary search trees; see, e.g., [8] or [5].

**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.* By Theorem 2, the claim follows by constructing a map that maps a permutation of the set $\{\overline{1},\dots,\overline{n-1}\}$ 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). ∎

## 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 $\tau$ 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 $\tau^{\ell}$ and $\tau^{r}$, respectively, as follows:

$$X(\tau) = X(\tau^{\ell}) + X(\tau^{r}) + T(\tau), \tag{4.1}$$

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

*Example* 2. (i) The Sackin index[2] 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., [26,27,40]. 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, (4.1) 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 [39].)

**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 $\{\overline{1}, \dots, \overline{n-1}\}$. The left subtree of the root has size $j$ if and only if the first element of this random permutation is $\overline{j}$ which 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,*

$$X_n \overset{d}{=} X_{I_n} + X^{*}_{n-I_n} + T_n, \qquad (n \ge 2), \tag{4.2}$$

*where $I_n$ is as in the last lemma, $X_n^{*} \overset{d}{=} X_n$, 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 [26,27,40]. Here, we are going to explain a powerful and (very) general methods for deriving moments and limit laws, namely, the *moment-transfer approach*. Moreover, at the end of this section, we briefly comment on another method, namely, the *contraction method*. (Both methods also work if $T_n$ is random and satisfies suitable conditions.)

The major steps of the moment-transfer approach are summarized in Figure 1 in [10]. We briefly recall them, adjusted to the way they are used below. First, it is shown that all moments and central moments satisfy the same type of recurrence. Then, this recurrence is solved and mean and variance are derived from the solution. Next, a so-called *asymptotic transfer result* is obtained for the recurrence which is used to derive the first-order asymptotics of all central moments. Finally, the limit law is identified from the asymptotic moment sequence.

Now, to carry out these steps, we first observe that from (4.2), we see that all (centered or non-centered) moments of $X_n$ indeed satisfy the same recurrence (for details, see the proof of Lemma 6 and Proposition 8 below):

$$a_n = \frac{2}{n-1}\left(\sum_{j=1}^{n-1} a_j\right) + b_n, \qquad (n \ge 2), \tag{4.3}$$

where (for the sake of simplicity) $a_1 = 0$. This recurrence has the solution:

$$a_n = 2n\left(\sum_{j=2}^{n-1} \frac{b_j}{j(j+1)}\right) + b_n, \qquad (n \ge 2). \tag{4.4}$$

From this, e.g., for the cherry index $C_n$, we immediately obtain the following result; see, e.g., [19,33] (phylogenetics) and [20,26] (computer science).

**Lemma 6.** *We have,*

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

[2] Note that this toll function subtracted by two gives also the *cost* of building a binary search tree of size $n$ as after the first element of the permutation (of length $n-1$) is placed into the root exactly $n-2$ elements need to be compared with the first element in order to build the subtrees. Equivalently, this gives also the number of key comparisons of quicksort.

*Proof.* Note that the mean satisfies (4.3) with $b_n = T_n$ where $T_n$ is given in item (ii) of Example 2. Thus, from (4.4), for $n \geq 3$,

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

as claimed. For the variance, the result follows similarly but with more computations that in particular rely on the result for the mean. (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 more useful in general to aim for asymptotic results. We first recall some asymptotic notation. First, for a sequences $a_n$ and a positive sequence $b_n$, we write $a_n = \mathcal{O}(b_n)$ if $|a_n/b_n|$ is bounded. Moreover, we use $a_n = o(b_n)$ if $a_n/b_n$ tends to 0 and $a_n \sim b_n$ if $a_n/b_n$ tends to 1. Using these notations, we state now an asymptotic transfer result for (4.3); see, Lemma 2 in [26]. The result says, in a nutshell, that $a_n$ behaves linearly if $b_n$ is sublinear and $a_n$ grows as $b_n$ if $b_n$ is superlinear. (The linear case, which is not needed here, exhibits a different behavior; see [26].)

**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 $a_n \sim (\alpha+1)cn^{\alpha}/(\alpha-1)$.*

*Remark* 3. According to our definitions, we need to require that $c > 0$. However, the above result still holds for $c = 0$ if $f(n) \sim 0 \cdot g(n)$ is interpreted as $f(n) = o(g(n))$. We will use this (convenient but non-standard) convention throughout the rest of the paper.

*Proof. (i)* This follows from (4.4) by extending the range of summation to infinity which introduces an error of $o(1)$ as the resulting series converges by the assumption that $b_n = \mathcal{O}(n^{1-\epsilon})$. Thus,

$$a_n = 2n\left(\sum_{j\geq 2}\frac{b_j}{j(j+1)} + o(1)\right) + b_n \sim 2\left(\sum_{j\geq 2}\frac{b_j}{j(j+1)}\right)n$$

as claimed.

*(ii)* By plugging the assumption that $b_n \sim cn^{\alpha}$ into (4.4), 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 (4.3) where $b_n$ is a function of moments of smaller order; see (4.5) and (4.6) below. (This method has been nicknamed *moment pumping* in computer science; see, e.g., Section VII.10.1 in [22].)

**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 $N(0,1)$, i.e.,*

$$g_m = \begin{cases} (2k)!2^{-k}/k!, & \textit{if } m = 2k; \\ 0, & \textit{if } m = 2k+1. \end{cases}$$

*Proof.* We prove the claim by induction on $m$. Note that the result holds for $m = 0$ and $m = 1$ (with the convention from Remark 3 since $g_1 = 0$), 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 lengthy computation (which proceeds by plugging (4.2) for $C_n$ into the definition of $\phi_n^{[m]}$ and expanding the $m$-th power with the help of the multinomial theorem):

$$\phi_n^{[m]} = \frac{2}{n-1}\sum_{j=1}^{n-1}\phi_j^{[m]} + \psi_n^{[m]}, \tag{4.5}$$

where

$$\psi_n^{[m]} = \frac{1}{n-1}\sum_{j=1}^{n-1}\sum_{\substack{a+b+c=m\\ a,b<m}}\binom{m}{a,b,c}\phi_j^{[a]}\phi_{n-j}^{[b]}\Delta(j,n)^c \tag{4.6}$$

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 plugging the induction hypothesis into (4.6),

$$\begin{aligned}
\psi_n^{[m]} &= \frac{1}{n-1}\sum_{j=1}^{n-1}\sum_{\substack{a+b+c=m\\ a,b<m}}\binom{m}{a,b,c}\phi_j^{[a]}\phi_{n-j}^{[b]}\Delta(j,n)^c\\
&\sim \frac{1}{n-1}\sum_{j=1}^{n-1}\sum_{\substack{a+b=m\\ a,b<m}}\binom{m}{a,b}g_a\left(\frac{2j}{45}\right)^{a/2}g_b\left(\frac{2(n-j)}{45}\right)^{b/2}\\
&\sim \left(\frac{2n}{45}\right)^{m/2}\sum_{\substack{a+b=m\\ a,b<m}}\binom{m}{a,b}g_ag_b\frac{1}{n-1}\sum_{j=1}^{n-1}\left(\frac{j}{n}\right)^{a/2}\left(1-\frac{j}{n}\right)^{b/2}\\
&\sim \left(\frac{2n}{45}\right)^{m/2}\sum_{\substack{a+b=m\\ a,b<m}}\binom{m}{a,b}g_ag_b\int_0^1 x^{a/2}(1-x)^{b/2}\mathrm{d}x, \qquad (4.7)
\end{aligned}$$

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 Riemann sum by the integral.

As for the sum in (4.7), first note that it is 0 for $m$ odd (as either $a$ or $b$ is odd and thus either $g_a = 0$ or $g_b = 0$). Likewise, for even $m = 2k$, we only need to consider even $a$ and $b$ and thus obtain

$$\begin{aligned}
\sum_{\ell=1}^{k-1}\binom{2k}{2\ell}g_{2\ell}g_{2k-2\ell}\int_0^1 x^{\ell}(1-x)^{k-\ell}\mathrm{d}x &= \frac{(2k)!}{2^k}\sum_{\ell=1}^{k-1}\frac{1}{\ell!(k-\ell!)}\int_0^1 x^{\ell}(1-x)^{k-\ell}\mathrm{d}x\\
&= \frac{(2k)!}{2^k}\sum_{\ell=1}^{k-1}\frac{1}{\ell!(k-\ell!)}\cdot\frac{\ell!(k-\ell)!}{(k+1)!} = \frac{k-1}{k+1}g_{2k},
\end{aligned}$$

where we have plugged in the expression for $g_m$ and simplified in the first step and evaluated the integral (either directly by using integration by parts or by appealing to the beta function) and again simplified in the second step.

Overall this shows that the sum in (4.7) equals $(m-1)g_m/(m+1)$ and 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. ▮

Finally, from the method of moments in probability theory (see Chapter 30 in [6]), 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 [33] (with a completely different method).

*Remark* 4. In fact, much more is known, namely, the number of occurrences of any *pattern* is asymptotically normal; see [26] where this was proved with the same approach as above and [20] where the authors used a mathematically more advanced method (which is based on complex analysis). Also, this result can be extended to patterns whose size is even allowed to moderately grow with $n$; see [9][3].

The moment-transfer method can also be applied to the Sackin index $S_n$. Here, we first obtain from (4.4) the following result for the mean and variance; see, e.g., [19] (phylogenetics) and Section 5.7 in [22] (computer science).

**Lemma 9.** *We have,*

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

*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 [26].

[3] The patterns considered in these papers should be more accurately called *fringe patterns* as they consist of a vertex in the tree and all its descendants. In [9], an even more general definition of pattern is used, namely, every finite set of fringe patterns of the same size is considered to be a pattern

**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*

$$c_m = \frac{m+1}{m-1} \sum_{\substack{a+b+c=m\\ a,b<m}} \binom{m}{a,b,c} c_a c_b \int_0^1 x^a(1-x)^b \Lambda(x)^c \mathrm{d}x, \qquad (m \geq 2) \tag{4.8}$$

*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 results from probability theory (see Chapter 30 in [6]), 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 (4.8). 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

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

a result first established in [36,37].

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, where the toll-sequence is actually random, we consider the *total cophenetic index* $\Phi_n$ 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 [19], was introduced in [34]. For evolutionary trees under the Yule model, the moment-transfer approach was used in [11] to derive the following limit distribution result.

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

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

*where $\Phi$ 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 (4.2). 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 [35]). 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 [26]. On the other hand, if mean and/or variance are not needed for the application of the method, then the contraction method yields asymptotic expansions of these quantities as a consequence.

## 5. A Generalization of the Yule Model

In this section, we explain a generalization of the Yule model which contains other models from computer science; see [14] for a generalization different to the one discussed below.

We start with a re-formulation of Definition 2. For this, 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 generated in steps (i) & (ii) of the Yule process is uniformly distributed on $\{1, \ldots, n-1\}$.*

*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 $1 \leq j \leq n-1$

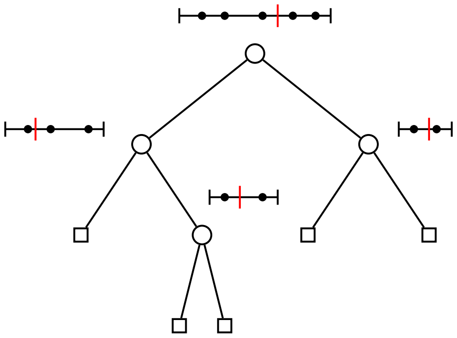


**Figure 3.** A plane tree constructed by step (i') & (ii'); the interval with dots are next to the node and the red vertical line indicates where the interval is cut.

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;
(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 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 $\beta$-function. However, $j$ is not allowed to be $0$ or $n$. Thus, the probability that the left subtree has size $j$ with $1\leq j\leq n-1$ 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] (see also [2]), it was suggested to use a $\beta$-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 $\Gamma$-function. The resulting random model on evolutionary trees of size $n$ is called the *$\beta$-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:

$$p_{n,j}=\frac{1}{c(\beta)}\frac{\Gamma(\beta+j+1)\Gamma(\beta+n-j+1)}{j!(n-j)!},\qquad(1\leq j\leq n-1),\tag{5.1}$$

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 $\beta$ to $\beta>-2$.

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

(i) $\beta=\infty$: This gives the random trie model from computer science (whose name comes from the word data re*trie*val); see Chapter 5 in [32];
(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 [7]. We have

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

and additive shape parameters still satisfy (4.2) but with $P(I_n = j) = p_{n,j}$. However, extending the tools from Section 4 has so far turned out to be largely elusive. Nevertheless, there has recently been some progress on this case; see [3,4].

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

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

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; e.g., see Section 2.5 in [38]. (This model has also been studied in combinatorics where it is called the *Catalan model*; see, e.g., Section 5.5 in [22].)

# 6. Conclusion

The main purpose of this survey was to draw attention to the fact that shape parameters for evolutionary trees under the Yule model and binary search trees under the binary search tree model share the same distribution (Theorem 3). For the former (phylogenetics), these shape parameters are often used to measure the balance of trees; for the latter (theoretical computer science), they describe the complexity of algorithms performed on trees. In phylogenetics, mainly mean and variance of shape parameters were considered; in theoretical computer science, also more sophisticated stochastic results such as limit laws have been proved. We presented a general and powerful method from the latter area, namely, the moment-transfer approach and discussed in detail two shape parameters which have been analyzed in both areas. More examples could be given, e.g., the product on the right-hand side of (2.1) has also been studied in theoretical computer science where it is called the *shape functional*; see, e.g., [17].

The connection between the Yule model and the binary search tree model extends to other models in phylogenetics and theoretical computer science. More precisely, we explained that the $\beta$-splitting model for $\beta = \infty$ corresponds to the trie model in theoretical computer science and $\beta = -3/2$ gives the PDA model in phylogenetics and the Catalan model in combinatorics (which has also some significance in theoretical computer science; see, e.g., Section 5.6 in [22]). Moments of shape parameters under the $\beta$-splitting model satisfy the recurrence:

$$a_n = \sum_{j=1}^{n-1} p_{n,j}(a_j + a_{n-j}) + b_n,$$

where $p_{n,j}$ is given by (5.1). There exist results towards an asymptotic transfer theorem for this recurrence; see [23]. However, no asymptotic transfer result of the same strength as Lemma 7 has so far been established for the $\beta$-splitting model. It is an interesting open problem to prove such a result as this would one allow to obtain moments and limit distribution results of shape parameters for the $\beta$-splitting model, in particular, for $\beta = -1$ which seems to be the most relevant value of $\beta$ in applications.

We conclude with some pointers on further reading about the topic of this survey. (Some of the references below have already appeared in the main text.) First, concerning literature about phylogenetics, the survey [19], which however mainly discusses combinatorial properties, is an excellent resource. A subset of the authors of this survey published an accompanying study that also addresses probabilistic models and results; see [28]. This study, in particular the software package which comes with it, should lead to many interesting conjectures for whose proofs the tools from theoretical computer science might turn out to be useful. As for suggested literature from computer science, we recommend the introductory texts [22,32] and [21,31] for a more advanced treatment. The book [21] also contains a brief introduction into the moment-transfer approach (e.g. from Page 532 on), however, for further reading on this method, we refer the interested reader to the research literature (e.g., [12,13,18,24]).

# Acknowledgements

We thank the reviewers and the editor for positive feedback and helpful comments. Moreover, we 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 (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Statist. Sci.*, **16:1**, 23–34.
3. D. Aldous. The critical beta-splitting random tree model II: Overview and open problems, arXiv:2303.02529.
4. D. Aldous and B. Pittel. The critical beta-splitting random tree: Heights and related results, arXiv:2302.05066.
5. F. Bergeron, P. Flajolet, B. Salvy (1992). Varieties of increasing trees, *Lecture Notes in Comput. Sci.*, **581**, 24–48.
6. P. Billingsley. *Probability and Measure*, 3rd edition, John Wiley & Sons, New York, 1995.
7. M. G. B. Blum and O. Francọis (2006). Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance, *Syst. Biol.*, **55**, 685–691.
8. M. G. B. Blum, O. Francọis, S. Janson (2006). The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance, *Ann. Appl. Probab.*, **16:4**, 2195–2214.
9. H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

10. C.-H. Chen and M. Fuchs (2011). On the moment-transfer approach for random variables satisfying a one-sided distributional recurrence, *Electron. J. Probab.*, **16:30**, 903–928.
11. L.-A. Chen. *Probabilistic Analysis of the Total Cophenetic Index in Phylogenetic Trees*, master thesis, National Chiao Tung Univesity (NCTU), 2015.
12. W.-M. Chen and H.-K. Hwang (2003). Analysis in distribution of two randomized algorithms for finding the maximum in a broadcast communication model, *J. Algor.*, **46:2**, 140–177.
13. H.-H. Chern and H.-K. Hwang (2001). Phase changes in random $m$-ary search trees and generalized quicksort, *Random Struc. Algor.*, **19:3-4**, 316–358.
14. E. H. Dickey and N. A. Rosenberger. Labeled histories with multifurcation and simultaneity, *Philos. Trans. R. Soc. B: Biol. Sci.*, submitted.
15. 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.
16. 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.
17. J. A. Fill (1996). On the distribution of binary search trees under the random permutation model, *Random Struc. Algor.*, **8:1**, 1-25.
18. J. A. Fill and N. Kapur (2005). Transfer theorems and asymptotic distributional results for $m$-ary search trees, *Random Struc. Algor.*, **26:4**, 359–391.
19. M. Fischer, L. Herbst, S. J. Kersting, L. Kühn, K. Wicke. *Tree Balance Indices: A Comprehensive Survey*, Springer, 1st edition, 2023.
20. P. Flajolet, X. Gourdon, C. Martinez (1997). Patterns in random binary search trees, *Random Struc. Algor.*, **11:3**, 223–244.
21. P. Flajolet and R. Sedgewick. *Analytic Combinatorics*, Cambridge University Press, 1st edition, 2009.
22. P. Flajolet and R. Sedgewick. *An Introduction to the Analysis of Algorithms*, Addison-Wesley, 2nd edition, 2013.
23. M. Fuchs and A. R. Paningbatan (2020). Correlation between Shapley values of rooted phylogenetic trees under the beta-splitting model, *J. Math. Biol.*, **80:3**, 627–653.
24. M. Fuchs, R. Neininger, H.-K. Hwang (2006). Profiles of random trees: limit theorems for random recursive trees and binary search trees, *Algorithmica*, **46:3-4**, 367–407.
25. E. F. Harding (1971). The probabilities of rooted tree-shapes generated by random bifurcation, *Adv. App. Probab.*, **3:1**, 44–77.
26. 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.
27. S. Janson (2022). Central limit theorems for additive functionals and fringe trees in tries, *Electron. J. Probab.*, **27**, Paper No. 47.
28. S. J. Kersting, K. Wicke, M. Fischer. Tree balance in phylogenetic models, *Philos. Trans. R. Soc. B: Biol. Sci.*, submitted.
29. M. C. King and N. A. Rosenberg (2023). A mathematical connection between single-elimination sports tournaments and evolutionary trees, *Math. Mag.*, **96:5**, 484–497.
30. J. F. C. Kingman (1982). The coalescent, *Stochastic Process. Appl.*, **13:3**, 235–248.
31. D. E. Knuth. *The Art of Computer Programming, Volume III, Sorting and Searching*, Addison-Wesley, 2nd edition, 1998.
32. H. M. Mahmoud. *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.
33. A. McKenzie and M. Steel (2000). Distributions of cherries for two models of trees, *Math. Biosci.*, **164**, 81–92.
34. A. Mir, F. Rosselló, L. Rotger (2013). A new balance index for phylogenetic trees, *Math. Biosci.*, **241:1**, 125–136.
35. R. Neininger and L. Rüschendorf (2004). A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.*, **14:1**, 378–418.
36. M. Régnier (1989). A limiting distribution for Quicksort, *RAIRO Inform. Théor. Appl.*, **23**, 335–343.
37. U. Rösler (1991). A limit theorem for "Quicksort", *RAIRO Inform. Théor. Appl.*, **25**, 85–100.
38. C. Semple and M. Steel. *Phylogenetics*, Oxford University Press, 2003.
39. 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.
40. S. Wagner (2015). Central limit theorems for additive tree parameters with small toll functions, *Combin. Probab. Comput.*, **24:1**, 329–353.
41. 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.