The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance

Michael G. B. Blum, Olivier François and Svante Janson

Grenoble University and Uppsala University

Michael Blum and Olivier François
Laboratory TIMC-TIMB
Pavillon Taillefer
Faculté de médecine
F38706 La Tronche cedex, France
E-mail: {michael.blum,olivier.francois}@imag.fr

Svante Janson
Uppsala University
Department of Mathematics
P. O. Box 480
S-751 06 Uppsala, Sweden
E-mail: svante@math.uu.se

Abstract: Two statistics called the Sackin's and Colless' indices are frequently used for capturing the balance of phylogenetic trees. In this article, these statistics are studied under the Yule and uniform model of phylogenetic trees. Asymptotics for the mean, variance, covariance of these indices are obtained, as well as their limiting joint distribution for large phylogenies. Under the Yule model the limiting distribution arises as a solution of a functional fixed point equation. Under the uniform model the limiting distribution is the Airy distribution. The cornerstone of this study is that the probabilistic models for phylogenetic trees are strongly related to the random permutation and the Catalan models for binary search trees.

AMS 2000 subject classifications: Primary 05C05; secondary 60F05, 60C05, 92D15..

Keywords and phrases: Random phylogenetic trees, Yule process, Catalan trees, shape statistics, contraction method, central limit theorem, Airy distribution..

1. Introduction

Phylogenetic trees (PT) represent the shared history of extant species. The idea of using trees to model evolution dates to Darwin (1859, see his diagram p.117). In a (rooted) PT, there is a common ancestral species called the root and each branching represents the time of divergence. A PT is usually reconstructed using data from n different species (or taxa) which are located at the leaves. The tree has $n-1$ internal nodes that correspond to the ancestors of the sample. There are two distinct features of rooted PT's. First is the branching structure or topology of the tree. Second is the branch lengths which indicate the time separating major evolutionary events. The shape of such trees carries useful information about the history of diversification rates among species by reflecting the footprint left by evolutionary processes.

Biologists have widely investigated the way by which the shape of PT's can be measured (Kirkpatrick and Slatkin, 1993). Mooers and Heard (1997) wrote an exhaustive review about tree balance in systematic biology, and Aldous (2001) gave a introduction in a more mathematical setting. How these measures are related to macroevolution processes has been studied by Rogers (1994) and Agapow and Purvis (2002) relying upon heavy computer simulations. So far several statistics have been introduced to measure the shape of PT's (see Agapow and Purvis, 2002, for eight of them). Among these statistics, the most widespread are the Sackin's and the Colless' indices. Sackin's index (Sackin, 1972, Shao and Sokal, 1990) computes the sum of depths for all leaves in the tree. Colless' index (Colless, 1982) inspects the internal nodes, partitioning the leaves that descend from them into groups of sizes L and R, and computes the sum of absolute values $|L - R|$ for all ancestors.

The probabilistic distribution of the Sackin's and Colless' statistics have been investigated for various models of biologically plausible random trees. Two random models of PT are often considered in the literature. The most famous is the Yule model (Yule, 1924). The Yule model is a branching process with constant speciation rate where the number of extant species is specified. The assumption of constant speciation rate may be weakened by assuming that the diversification rate could vary in time but is the same for all species at any time. This assumption does not modify the distribution of PT shape. An alternative model considered by biologists is called the uniform model. It assumes that all PT's are equally likely. This model is biologically motivated as it arises from a large family of Galton-Watson processes conditioned by the total size of the trees (see Aldous, 1991a). In addition, McKenzie and Steel (2001) have shown that when speciation events are constrained to occur before a time τ after their previous speciation event, then the resulting process converges to the uniform model as τ tends to zero. In both models, Rogers (1994) studied the joint distribution of the Sackin's and the Colless' statistics using numerical computations. He concluded that these statistics were strongly correlated in large PT's. The limiting distribution of the Colless' statistic was also conjectured to be non Gaussian (Rogers, 1993).

This article describes the mean, the (co)variance and the limiting joint distribution of the Sackin's and Colless' indices for large PT's under the Yule and uniform models. Because this study is mainly concerned with the topology of PT's, branch lengths can be ignored. A PT is then a cycle-free connected graph, with vertices of degree one (the leaves), two (the root), or three (all ancestors except the root). Leaves are usually labeled whereas ancestors are not. This simplified model of phylogeny without branch lengths is sometimes called a cladogram (see Aldous, 1995). Our proofs use the connection to recent results in theoretical computer science, as well as the correspondence between PT's and random binary search trees (BST). This approach extends results by Blum and François (2005) who showed that the Sackin's index has the same limit distribution as the number of comparisons used by the quicksort algorithm (Hoare, 1961). More specifically, we deal with the Yule and uniform models separately. For the Yule model, our analysis relies on the recursive structure of the tree, and makes use of the fixed point method (see for instance Hwang and Neininger, 2002). This method was introduced in the probabilistic analysis of algorithms by Rösler (1991). In the uniform model, the results are based on the connection between uniform trees and Bernoulli excursions (Takacs, 1991). A large family of statistics similar to the Sackin's and Colless' indices have been studied by Fill and Kapur (2004) under the Catalan model for BST's.

In Section 2, we shall present our main results. Section 3 explains how probabilistic models for PT's are related to probabilistic models for BST's. Section 4 is dedicated to the Yule model while Section 5 deals with the uniform model.

2. The Sackin and the Colless for large phylogenetic trees

Consider a PT with n leaves. Sackin's statistic adds the number of internal nodes between each leaf and the root of the tree to form the following index

$$S_n = \sum_{i=1}^{n} d_i \,,$$

where the sum runs over the n leaves of the tree and d_i is the number of ancestors crossed in the path from i to the root (including the root). Colless' statistic inspects the internal nodes, partitioning the leaves that descend from them into groups of sizes L_j and R_j and computes

$$C_n = \sum_{j=1}^{n-1} |L_j - R_j| \,,$$

where the sum runs over the internal nodes, and L_j (resp. R_j) corresponds to the number of leaves in the left (resp. right) subtree under node j.

Denote $\mathcal{M}_2$ the space of all bivariate, centered probability measures with finite second moments, and $\mathcal{L}(X)$ an element of $\mathcal{M}_2$. We have the following result.

Theorem 1. *Assume the Yule model of PT. Consider the map $\mathcal{T} : \mathcal{M}_2 \to \mathcal{M}_2$ such that for all $\nu \in \mathcal{M}_2$ we have*

$$\mathcal{T}(\nu) = \mathcal{L}\left(\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \begin{pmatrix} S \\ C \end{pmatrix} + \begin{bmatrix} 1-U & 0 \\ 0 & 1-U \end{bmatrix} \begin{pmatrix} S' \\ C' \end{pmatrix} + \begin{pmatrix} b_S \\ b_C \end{pmatrix} \right),$$

with

$$\begin{pmatrix} b_S \\ b_C \end{pmatrix} = \begin{pmatrix} 2U \log U + 2(1-U)\log(1-U) + 1 \\ U \log U + (1-U)\log(1-U) + 1 - 2\min(U, 1-U) \end{pmatrix}$$

where (S, C)*,* (S', C') *and* U *are independent random variables such that* $\mathcal{L}(S, C) = \mathcal{L}(S', C') = \nu$ *and* U *is uniform over the interval* $(0, 1)$*. Then, we have*

$$\left(\frac{S_n - \mathbb{E}[S_n]}{n}, \frac{C_n - \mathbb{E}[C_n]}{n}\right) \xrightarrow{d} (S, C), \quad n \to \infty$$

where the convergence holds in distribution, and the limiting probability distribution is the unique fixed-point of the map $\mathcal{T}$*.*

Remark 1. The convergence in Theorem 1 will actually be proved for a stronger topology than convergence in distribution. As can be seen from section 4, it indeed holds for the Wasserstein-Mallows d_2-metric (Rachev and Rüschendorf, 1995) which warrants the existence and convergence of the second moments.

Remark 2. This result extends the fact that the normalized Sackin index

$$\bar{S}_n = \frac{S_n - \mathbb{E}[S_n]}{n} \tag{1}$$

converges in distribution to the same limit as the number of comparisons in the quicksort algorithm. According to Rösler (1991), the limit S satisfies a (functional) fixed-point equation of the following type

$$S \stackrel{d}{=} US + (1-U)S' + 2U \log U + 2(1-U)\log(1-U) + 1 \tag{2}$$

where S, S', U are independent random variables, S and S' are identically distributed, U is uniformly distributed over the inverval $(0, 1)$, and the equality holds for distributions. Regarding the Colless' index, the functional fixed point equation becomes

$$C \overset{d}{=} UC + (1-U)C' + U\log U + (1-U)\log(1-U) + 1 - 2\min(U, 1-U)\,. \tag{3}$$

This is a well-known result in systematic biology that the expectation of S_n is of order $2n\log n$. More precisely, Kirkpatrick and Slatkin (1993) showed that

$$\mathbb{E}[S_n] = 2n\sum_{j=2}^{n}\frac{1}{j}$$

and

$$\mathbb{E}[S_n] = 2n\ln n + (2\gamma - 2)n + o(n)\,,$$

where γ is the Euler constant. Using the connection to the quicksort algorithm, the variance of the limiting distribution can be obtained according to Knuth (1973) as

$$\mathrm{Var}[S_n] \sim \left(7 - 2\frac{\pi^2}{3}\right)n^2, \quad n \to \infty\,.$$

These results can be extended to the case of the Colless' index as follows, with the remark that S_n and C_n are strongly correlated for large PT's.

Theorem 2. *Assume the Yule model of PT. Then, we have*

$$\mathbb{E}[C_n] = n\log n + (\gamma - 1 - \log 2)n + o(n) \tag{4}$$

$$\mathrm{Var}[C_n] \sim (3 - \frac{\pi^2}{6} - \log 2)n^2 \tag{5}$$

$$\mathrm{Cor}[S_n, C_n] \sim \frac{27 - 2\pi^2 - 6\log 2}{\sqrt{2(18 - \pi^2 - 6\log 2)(21 - 2\pi^2)}} \approx 0.98 \tag{6}$$

as n goes to infinity.

Regarding the uniform model of PT, mathematical results have deserved less attention than for the Yule model. After an appropriate rescaling we prove the convergence of both S_n and C_n toward the same marginal probability distribution, and we identify this distribution as $\sqrt{8}$ times the integral of the standard Brownian excursion $e(t)$

$$\omega = \int_0^1 e(t)\, dt\,.$$

The distribution of random variable $\mathcal{A} = \sqrt{8}\omega$ is known as the Airy distribution. A formula for the moments of $\mathcal{A}$ is given by Flajolet and Louchard (2001). In particular, we have

$$\mathbb{E}[\mathcal{A}] = \sqrt{\pi}\,,$$

and

$$Var[\mathcal{A}] = \frac{10-3\pi}{3}\,.$$

Theorem 3. *Assume the uniform model of PT. Then, we have*

$$\left(\frac{S_n}{n^{3/2}}, \frac{C_n}{n^{3/2}}\right) \xrightarrow{d} (\mathcal{A}, \mathcal{A}), \quad n \to \infty \tag{7}$$

where the convergence holds in distribution.

Remark 3. Regarding S_n the connection to the internal path length of a BST enables us to state that

$$\frac{S_n}{n^{3/2}} \xrightarrow{d} \mathcal{A}$$

immediately. This result was actually established by Takacs (1991) using the method of moments. In addition we find that

$$\mathbb{E}[S_n] \sim \sqrt{\pi} n^{3/2}$$

and

$$\mathrm{Var}[S_n] \sim (\frac{10}{3} - \pi)n^3 .$$

The moments of C_n follow from the next Theorem.

Theorem 4. *Assume the uniform model of PT. Then, we have*

$$\mathbb{E}[C_n] \sim \sqrt{\pi} n^{3/2},$$

and

$$\mathrm{Var}[C_n] \sim \frac{10 - 3\pi}{3} n^3$$

as n goes to infinity. In addition, the variables S_n and C_n are asymptotically correlated

$$\mathrm{Cor}[S_n, C_n] \sim 1,$$

and we have for any $k, \ell \geq 0$

$$\mathbb{E}[C_n^k \, S_n^\ell] \sim n^{3(k+\ell)/2} \mathbb{E}[\mathcal{A}^{k+\ell}]$$

as n goes to infinity.

3. Phylogenetic and binary search trees

Trees are often encountered in theoretical computer science as data structures connected with *divide and conquer* algorithms. In this section, we explain how binary search trees can be mapped onto phylogenetic trees univoquely, and how probabilistic models for BST's are transported on probabilistic models for PT's.

Mapping binary search trees A binary tree can be defined recursively. It is either empty or it is a node (the root) with left and right subtrees. A binary search tree is a binary tree where labels are associated to the vertices. These labels are constrained: the label of a vertex is greater or equal than all labels contained in the left subtree and lower or equal than all labels contained in the right subtree (Sedgewick and Flajolet, section 5.5, 1996). The transformation that maps BST's into PT's is explained by Aldous (1995). Given a BST with $n-1$ vertices, the structure is modified as follows. Vertices in the BST become ancestors in the PT. Do do so, two leaves are connected to each vertex of degree one, and one leaf is connected to each vertex of degree two. The root has a special status. If the degree is 0, 1 or 2, then 2, 1 or 0 leaves are added. The labels of leaves are chosen arbitrarily amongst the $n!$ possible orders.

Two obtained PT's are equivalent if their left and right subtrees can be swapped recursively (see Figure 1). The set of PT's is the set of equivalence classes for this equivalence relation. Figure 2 gives a graphical representation of these transformations.

A PT may therefore arise from the construction of 2^{n-1} equivalent modified BST's. Because there are $\mathcal{C}_{n-1}$ BST's with $(n-1)$ vertices, we obtain $\mathcal{C}_{n-1}n!/2^{n-1}$ possible PT's. This number coincides with the total number of PT's which equals $(2n-3)!!$.

Probabilistic models The mapping described in the above paragraph also transfers probabilistic models for BST's to probabilistic models for PT's. For instance, there will be an equivalence between the random permutation model for BST's (Mahmoud, 1992) and the Yule model for PT's. Probabilistic models for BST's (with $n-1$ vertices) can be described as a general class of models called *branching Markov processes*. A definition of branching Markov processes can be found in (Aldous 1995). We recall this definition here. Let $\hat{q}_{n-1}$ be a symmetric probability distribution on $\{0,\dots,n-2\}$

$$\hat{q}_{n-1}(i) = \hat{q}_{n-1}(n-2-i), \quad i = 0,\dots,n-2\,.$$

In the branching Markov process, the size of the left subtree is chosen according to the probability distribution $\hat{q}_{n-1}$. This procedure is repeated recursively in subtrees assuming local independence. The probability distribution $\hat{q}_{n-1}$ is called the splitting distribution. In the same way, probability distributions on PT's with n leaves can be associated with splitting probability distributions q_n on $\{1,\dots,n-1\}$. At each step, the labels of the left subtree, of size i, are sampled uniformly amongst all the $\begin{pmatrix} n \\ i \end{pmatrix}$ possible labels. At the end of the construction, distinctions left-right are merely suppressed for building the PT.

Lemma 1. *Assume that $\hat{T}_{n-1}$ is a BST sampled according to a branching*

Markov process with splitting probability $\hat{q}_{n-1}$. Denote by Φ_n the transformation which consists of extending a BST with $n-1$ vertices into a PT with n leaves. Then $T_n = \Phi_n(\hat{T}_{n-1})$ is a PT sampled according to a branching Markov process with splitting probability q_n such that

$$q_n(i) = \hat{q}_{n-1}(i-1), \quad i = 1, \ldots, n-1\,.$$

Proof. This is a consequence of the basic properties of Φ_n. If a BST has i vertices in its left subtree, the resulting PT has $i+1$ leaves in one of the two subtrees of the root. The symmetry property of $\hat{q}_n$ insures that all members of the same equivalence class have the same probability of occurrence. $\square$

Lemma 1 has the interesting consequence that well-studied models of BST's can be shifted into models on PT's. The Yule and uniform models on PT's fall then into special cases of branching Markov processes.

On one hand the random permutation model for BST's with $n-1$ vertices is a branching Markov process with splitting probability

$$\hat{q}_{n-1}(i) = \frac{1}{n-1}, \quad i = 0, \ldots, n-2\,.$$

This model is mapped by Φ_n into the Yule model for PT's with n leaves with splitting probability

$$q_n(i) = \frac{1}{n-1}, \quad i = 1, \ldots, n-1\,.$$

The splitting distribution for Yule trees was found by Harding (1971). Note that the same splitting property also holds for coalescent tree topologies (Kingman, 1982).

On the other hand, the Catalan model for binary BST with $n-1$ vertices assumes that all $\mathcal{C}_{n-1}$ binary trees have the same probability of occurrence. The number of trees with left subtree of size is i is equal to $\mathcal{C}_i\mathcal{C}_{n-2-i}$. The Catalan model is a branching Markov process where the splitting distribution is given by

$$\hat{q}_{n-1}(i) = \frac{\mathcal{C}_i\mathcal{C}_{n-2-i}}{\mathcal{C}_{n-1}}, \quad i = 0, \ldots, n-2\,.$$

The transformation Φ_n maps the Catalan model into the uniform model for PT with n leaves. The splitting distribution for PT is then

$$q_n(i) = \frac{1}{2}\left(\begin{array}{c} n \\ i \end{array}\right)\frac{(2i-3)!!(2(n-i)-3)!!}{(2n-3)!!}, \quad i = 1, \ldots, n-1\,.$$

This formula can also be found in (Aldous, 1995).

Lemma 2. *Let q_n be a splitting distribution on $\{1, \ldots, n-1\}$. Let h be a function of pairs of integral numbers. Denote by X_n an additive random variable defined recursively as*

$$X_n \stackrel{d}{=} X_{I_n} + X_{J_n} + h(I_n, J_n)$$

where the I_ℓ's are sampled under a branching Markov process of splitting distribution q_n and $I_n + J_n = n$. Define $\hat{X}_n$ by

$$\hat{X}_n \stackrel{d}{=} \hat{X}_{\hat{I}_n} + \hat{X}_{\hat{J}_n} + h(\hat{I}_n + 1, \hat{J}_n + 1)$$

where the $\hat{I}_\ell$'s are sampled under the branching Markov process of splitting distribution $\hat{q}_n$ with

$$\hat{q}_n(i) = q_{n+1}(i+1), \quad i = 0, \ldots, n-1$$

and $\hat{I}_n + \hat{J}_n = n - 1$. *Then, we have*

$$X_n \stackrel{d}{=} \hat{X}_{n-1} \, .$$

Proof. Note that $\hat{I}_n \stackrel{d}{=} I_{n+1} - 1$, i.e. the distribution of $\hat{I}_n$ is given by $\hat{q}_n$. Similarly, we have

$$X_n \stackrel{d}{=} X_{\hat{I}_{n-1}+1} + X_{\hat{J}_{n-1}+1} + h(\hat{I}_{n-1} + 1, \hat{J}_{n-1} + 1).$$

Setting $\hat{X}_{n-1} = X_n$, we prove the result. □

Remark 4. This lemma states that for additive random variables built from a Markov branching PT of splitting distribution q_n there exist additive random variables $\hat{X}_n$ built from a Markov branching BST of distribution $\hat{q}_n$. In addition, X_n and $\hat{X}_{n-1}$ have the same distribution. This lemma can obviously be generalized to multivariate random variables. In the next sections, all random variables are studied in the context of BST's. Applying Lemma 2, the results can be shifted to PT's without difficulties.

4. Yule model

The Yule model is a branching Markov process for PT's with splitting probability

$$q_n(i) = \frac{1}{n-1}, \quad \text{for } i = 1 \ldots (n-1)\,.$$

The Sackin's index S_n has been defined as a sum of depths over the leaves. It can also be expressed as a sum over the ancestors (Rogers, 1994)

$$S_n = \sum_{j=1}^{n-1} N_j$$

where N_j is the number of leaves descending from the ancestor j. Applying Lemma 2 we obtain that S_n has the same distribution as $\hat{S}_{n-1} + 2(n-1)$ where $\hat{S}_n$ is defined by

$$\hat{S}_n \stackrel{d}{=} \hat{S}_{I_n} + \hat{S}'_{J_n} + n - 1\,. \tag{8}$$

where I_n is distributed uniformly over $\{0, \ldots, n-1\}$ and $J_n = n-1-I_n$. The recursion satisfied by $\hat{S}_n$ is well-studied as it arises from the analysis of the quicksort algorithm or the internal path length of a BST under the random permutation model. Similarly, C_n has the same distribution as $\hat{C}_{n-1}$, where

$$\hat{C}_n \stackrel{d}{=} \hat{C}_{I_n} + \hat{C}'_{J_n} + |I_n - J_n|\,. \tag{9}$$

In order to describe the joint distribution of $(\hat{S}_n, \hat{C}_n)$ under the random permutation model, we shall follow the same lines of proof as Neininger (2002) who studied the joint convergence of the Wiener index and the internal path length of a BST.

Proof of Theorem 1. Step 1 (Computing expectations). Denote $\hat{c}_n = \mathbb{E}[\hat{C}_n]$ and $\hat{s}_n = \mathbb{E}[\hat{S}_n]$. We have

$$\hat{s}_n = 2n \log n + (2\gamma - 4)n + o(n)\,. \tag{10}$$

We rewrite equation (9) as

$$\hat{C}_n \stackrel{d}{=} \hat{C}_{I_n} + \hat{C}'_{J_n} + n - 1 - 2\min(I_n, J_n)\,. \tag{11}$$

Conditioning on I_n in the above equation, we find that

$$\hat{c}_n = (n - 1 - 2t_n) + \frac{2}{n}\sum_{k=0}^{n-1} \hat{c}_k$$

where

$$t_n = \mathbb{E}\left[\min(I_n, J_n)\right] = \begin{cases} \frac{n-2}{4} & \text{if } n \text{ is even} \\ \frac{(n-1)^2}{4n} & \text{if } n \text{ is odd.} \end{cases}$$

Applying Lemma 1 of Hwang and Neininger (p. 1691, 2002) we obtain that

$$\hat{c}_n = (n - 1 - 2t_n) + 2(n+1)\sum_{k=1}^{n-1} \frac{k - 1 - 2t_k}{(k+1)(k+2)}\,.$$

An asymptotic expansion of the above expression leads to the following result

$$\hat{c}_n = n \log n + (\gamma - 1 - \log 2)n + o(n)\,. \tag{12}$$

Step 2. Limit distribution. Let us consider the rescaled quantities

$$\hat{X}_n = \begin{pmatrix} \frac{\hat{S}_n - \hat{s}_n}{n} \\ \frac{\hat{C}_n - \hat{c}_n}{n} \end{pmatrix},$$

and $\hat{X}'_n$ an independent copy of $\hat{X}_n$. From equations (8) and (11) we have

$$\hat{X}_n = A_1^{(n)} \hat{X}_{I_n} + A_2^{(n)} \hat{X}'_{J_n} + b^{(n)} \tag{13}$$

where

$$A_1^{(n)} = \frac{1}{n}\begin{pmatrix} I_n & 0 \\ 0 & I_n \end{pmatrix},$$

$$A_2^{(n)} = \frac{1}{n}\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix},$$

and

$$\begin{pmatrix} b_S^{(n)} \\ b_C^{(n)} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \hat{s}_{I_n} + \hat{s}_{J_n} - \hat{s}_n + n - 1 \\ \hat{c}_{I_n} + \hat{c}_{J_n} - \hat{c}_n + n - 1 - 2\min(I_n, J_n) \end{pmatrix}.$$

Since I_n/n converges in L^2 toward U a uniform variable over $(0,1)$, we obtain that

$$A_1^{(n)} \xrightarrow{L^2} A_1^* = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix},$$

and

$$A_2^{(n)} \xrightarrow{L^2} A_2^* = \begin{pmatrix} 1-U & 0 \\ 0 & 1-U \end{pmatrix}.$$

Using the asymptotic expansion of $\hat{c}_n$ given by equation (12) and the asymptotic expansion of $\hat{s}_n$ given by (10), we find that

$$b_S^{(n)} = \frac{1}{n}\left(2I_n \log\left(\frac{I_n}{n}\right) + 2J_n \log\left(\frac{J_n}{n}\right) + n\right) + o(1)$$

and

$$b_C^{(n)} = \frac{1}{n}\left(I_n \log\left(\frac{I_n}{n}\right) + J_n \log\left(\frac{J_n}{n}\right) + n - 2\min(I_n, J_n)\right) + o(1)\,.$$

Thus, we obtain that

$$\begin{pmatrix} b_S^{(n)} \\ b_C^{(n)} \end{pmatrix} \xrightarrow{L^2} \begin{pmatrix} 2U\log U + 2(1-U)\log(1-U) + 1 \\ U\log U + (1-U)\log(1-U) + 1 - 2\min(U, 1-U) \end{pmatrix}.$$

Assuming that $\hat{X}_n$ converges in distribution, the limiting distribution $\mathcal{L}(\hat{X})$ must satisfy the following condition

$$\hat{X} \stackrel{d}{=} A_1^* \hat{X} + A_2^* \hat{X}' + b^* \tag{14}$$

where $\hat{X}$, $\hat{X}'$,and (A_1^*, A_2^*, b^*) are independent and $\hat{X}' \stackrel{d}{=} \hat{X}$.

The multivariate contraction theorem (Neininger, 2001) states that there is a unique probability distribution $\mathcal{L}(\hat{X})$ satisfying (14) in $\mathcal{M}_2$. Moreover it states that the distribution of $\hat{X}_n$ converges toward the distribution of $\hat{X}$ in the Wasserstein-Mallow d_2-metric. The convergence in this metric is the same as the convergence in distribution and the convergence of second moments. Neininger's Theorem can be applied provided that the four following conditions hold

(i) $(A_1^{(n)}, A_2^{(n)}, b^{(n)}) \stackrel{L^2}{\to} (A_1^*, A_2^*, b^*),\ n \to \infty,$

(ii) $\mathbb{E}\left[\|(A_1^*)^t A_1^*\|_{op}\right] + \mathbb{E}\left[\|(A_2^*)^t A_2^*\|_{op}\right] < 1,$

(iii) $\mathbb{E}\left[1_{\{I_n \le \ell\} \cup \{I_n = n\}} \|(A_1^*)^t A_1^*\|_{op}\right] \to 0,\ \forall \ell \in \mathbb{N},\ n \to \infty,$

(iv) $\mathbb{E}\left[1_{\{J_n \le \ell\} \cup \{J_n = n\}} \|(A_2^*)^t A_2^*\|_{op}\right] \to 0,\ \forall \ell \in \mathbb{N},\ n \to \infty,$

where $\|A\|_{op} = sup_{\|x\|=1} \|Ax\|$ is the operator norm of A. For symmetric matrices, as here, this equals the spectral radius.

(i) has already been proved. (ii) holds because

$$\mathbb{E}\left[\left\|(A_1^*)^t A_1^*\right\|_{op}\right] + \mathbb{E}\left[\left\|(A_2^*)^t A_2^*\right\|_{op}\right] = \mathbb{E}[U^2 + (1-U)^2] = \frac{2}{3} < 1\,.$$

(iii) and (iv) are obvious because $\|(A_r^*)^t A_r^*\|_{op} \le 1$, for $r = 1, 2$, and

$$P(\{I_n \le \ell\} \cup \{I_n = n\}) = P(\{J_n \le \ell\} \cup \{J_n = n\}) \le \frac{\ell + 1}{n} \to 0$$

for all $\ell \in \mathbb{N}$ and $n \to \infty$. $\square$

Proof of Theorem 2. According to Theorem 1, equation (14) has a unique solution, so we can consider (S, C), (S', C'), two independent copies with

$\mathcal{L}(S,C) = \mathcal{L}(S',C')$ being the fixed point of (14) and U uniform over $(0,1)$. By definition, we have

$$\mathcal{L}\begin{pmatrix} S \\ C \end{pmatrix} \stackrel{d}{=} \mathcal{T}\mathcal{L}\begin{pmatrix} S \\ C \end{pmatrix}$$

Using the fact that all random variables (except U) are centered, we find that

$$\begin{aligned} \mathbb{E}[C^2] &= \mathbb{E}[U^2C^2] + \mathbb{E}[(1-U)^2C'^2] \\ &\quad + \mathbb{E}[(U\log U + (1-U)\log(1-U) + 1 - 2\min(U,1-U))^2]\,. \end{aligned}$$

Thus, we have

$$\mathrm{Var}[C^2] = \mathbb{E}[C^2] = (3 - \frac{\pi^2}{6} - \log 2)\,.$$

In the same way, we find that

$$\begin{aligned} \mathbb{E}[SC] &= \mathbb{E}[U^2SC] + \mathbb{E}[(1-U)^2S'C'] \\ &\quad + \mathbb{E}[(2U\log U + 2(1-U)\log(1-U) + 1) \\ &\quad * (U\log U + (1-U)\log(1-U) + 1 - 2\min(U,1-U))]\,. \end{aligned}$$

This leads to

$$\mathrm{Cov}(S,C) = \mathbb{E}[SC] = \frac{9}{2} - \frac{\pi^2}{3} - \log 2\,.$$

Using that $\mathbb{E}[S^2] = 7 - 2\pi^2/3$ (Rösler, 1991), we find that

$$\mathrm{Cor}(S,C) = \frac{27 - 2\pi^2 - 6\log 2}{\sqrt{2(18 - \pi^2 - 6\log 2)(21 - 2\pi^2)}}\,.$$

Theorem 1 holds in the Wasserstein-Mallows d_2-metric which implies the convergence of second moments. This leads to

$$\mathrm{Var}[S_n] \sim \mathrm{Var}[S]n^2, \qquad \mathrm{Var}[C_n] \sim \mathrm{Var}[C]n^2,$$

and

$$\mathrm{Cov}(S_n, C_n) \sim \mathrm{Cov}(S, C)n^2, \qquad \mathrm{Cor}(S_n, C_n) \sim \mathrm{Cor}(S, C)\,.$$

□

5. Uniform model

For a given n, the uniform model assumes that all PT's with n leaves are equally likely. Again we use the fact that the Sackin's and the Colless' indices for a PT with n leaves drawn according to the uniform model have the same probability distribution as $\hat{S}_{n-1} + 2(n-1)$ and $\hat{C}_{n-1}$ which are respectively defined by equation (8) and (9). Under the Catalan model for BST, I_n is distributed according to $\hat{q}_n$, where

$$\hat{q}_n(i) = \frac{\mathcal{C}_i \mathcal{C}_{n-1-i}}{\mathcal{C}_n}, \quad i = 0, \ldots, n-1\,.$$

Conditional on I_n, $(\hat{S}_{J_n}, \hat{C}_{I_n})$ is independent of $(\hat{S}'_{J_n}, \hat{C}'_{J_n})$.

Clearly, $\hat{S}_n$ has the same distribution as the internal path length of a BST under the Catalan model, and $\hat{C}_n$ has the same distribution as the random variable $\sum_{j=1}^{n-1} |\hat{L}_j - \hat{R}_j|$ where the sum is over the $n-1$ vertices of a BST drawn under the Catalan model. Note that $\hat{C}_n$ can be rewritten as

$$\sum_{j=1}^{n-1} (\hat{N}_j - 1) - 2\min(\hat{L}_j, \hat{R}_j)$$

where $\hat{N}_j$ is the number of vertices of the subtree rooted at j (including j), and $\hat{L}_j$ ($\hat{R}_j$) is the number of vertices of the left (right) subtree. Then we have

$$\hat{C}_n = \hat{S}_n - 2\sum_{j=1}^{n-1} \min(\hat{L}_j, \hat{R}_j)\,.$$

It is well known (Takacs, 1991) that $\hat{S}_n/n^{3/2}$ converges in distribution to the Airy distribution. The proof relies on the one-to-one correspondence between binary trees and Bernoulli excursions. Takacs (1991) computed the moments of $\hat{S}_n/n^{3/2}$ and their limiting values to establish convergence from the method of moments. The goal of this section is to prove the convergence of

$$\mathbb{E}\left[\frac{\sum_{j=1}^{n-1}\min(\hat{L}_j,\hat{R}_j)}{n^{3/2}}\right] \to 0, \quad n \to 0\,.$$

This implies that $(\hat{C}_n - \hat{S}_n)/n^{3/2}$ converges in probability to 0. By the Slutsky's Theorem, this argument completes the proof of Theorem 3.

Lemma 3. *Let $n \geq 2$, and consider a BST with n vertices under the Catalan model. Denote j_0 the root of the tree. We have*

$$\mathbb{E}[\min(\hat{L}_{j_0},\hat{R}_{j_0})] \leq K\sqrt{n}$$

for some constant K.

Proof. The Stirling's formula yields to a well known asymptotic expansion for the Catalan number $\mathcal{C}_n$

$$\mathcal{C}_n = \frac{4^n}{\sqrt{\pi n^3}}\left(1 + O(\frac{1}{n})\right)\,. \tag{15}$$

The expectation of $\min(\hat{L}_{j_0},\hat{R}_{j_0})$ is given by

$$\mathbb{E}[\min(\hat{L}_{j_0},\hat{R}_{j_0})] = \sum_{k=1}^{\lfloor n/2\rfloor -1} 2k\frac{\mathcal{C}_k\mathcal{C}_{n-1-k}}{\mathcal{C}_n} + \mathbb{1}_{\{n\in 2\mathbb{N}+1\}}\frac{(n-1)\mathcal{C}^2_{(n-1)/2}}{2\mathcal{C}_n}$$

Using (15), we find that

$$\sum_{k=1}^{\lfloor n/2\rfloor -1} 2k\frac{\mathcal{C}_k\mathcal{C}_{n-1-k}}{\mathcal{C}_n} = \frac{4^{n-1}}{\pi\mathcal{C}_n(n-1)}\frac{1}{n-1}\sum_{k=1}^{\lfloor n/2\rfloor -1} 2\frac{(1+O(1/k)+O(1/n))}{\sqrt{k/(n-1)}(1-k/(n-1))^{3/2}}$$

The sum in the right hand side of the above equality is a Riemann sum. Using that

$$\int_0^{1/2} \frac{1}{\sqrt{x}(1-x)^{3/2}} = 2\,,$$

we have

$$\sum_{k=1}^{\lfloor n/2 \rfloor -1} 2k \frac{\mathcal{C}_k \mathcal{C}_{n-1-k}}{\mathcal{C}_n} = \frac{4^n}{\pi \mathcal{C}_n (n-1)} \left(1 + o(1)\right).$$

Using (15) again leads to

$$\mathbb{E}[\min(\hat{L}_{j_0}, \hat{R}_{j_0})] \sim \sqrt{n/\pi}\,. \tag{16}$$

which concludes the proof of the result. □

By conditioning on the sizes of the two subtrees of the root, and using induction on n, it follows that

$$\mathbb{E}\left[\frac{\sum_{j=1}^{n-1} \min(\hat{L}_j, \hat{R}_j)}{n^{3/2}}\right] \leq K \mathbb{E}\left[\frac{\sum_j \sqrt{\hat{N}_j}}{n^{3/2}}\right].$$

In the following, we prove that the right-hand side of the above inequality converges to 0 as n grows to ∞. A lemma which is interesting in its own gives the key argument. In the following, we use the standard convention that $\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 1$.

Lemma 4. *Let $n \geq 1$. Consider a BST with n vertices sampled according to the Catalan model. Pick a vertex V at random amongst the n vertices. Denote $\hat{K}_n = \hat{N}_V$, we have*

$$\mathbb{P}(\hat{K}_n = k) = \frac{\mathcal{C}_k \begin{pmatrix} 2n-2k \\ n-k \end{pmatrix}}{n \mathcal{C}_n}, \qquad \textit{if } k = 1, \ldots, n\,.$$

Proof. The proof relies on combinatorial arguments. Let us denote $\nu_k(T)$ the number of subtrees with k vertices in the BST T with $|T| = n$ vertices. For $k = 1, \ldots, n$, we have

$$\begin{aligned}\mathbb{P}(\hat{K}_n = k) &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{P}(\hat{N}_j = k) \\ &= \frac{1}{n}\mathbb{E}[\sum_{j=1}^{n}\mathbb{1}_{\{\hat{N}_j = k\}}] = \frac{1}{n}\mathbb{E}[\nu_k(T)]\,.\end{aligned}$$

$\nu_k(T)$ satisfies the following linear recursion

$$\nu_k(T) = \delta_{|T|,k} + \sum_{S}\nu_k(S) \tag{17}$$

where δ denotes the Kronecker symbol and the sum is over the subtrees of the root of T. Let us denote by $\mathcal{B}$ the set of all BST's. We introduce the cumulative generating function defined by

$$F_k(z) = \sum_{T\in\mathcal{B}}\nu_k(T)z^{|T|},$$

and

$$G_k(z) = \sum_{T\in\mathcal{B}}\delta_{|T|,k}z^{|T|} = \mathcal{C}_k z^k.$$

From the linear recurrence equation (17), Theorem 5.7 in (Sedgewick and Flajolet, 1996) establishes a relationship between cumulative generating functions F_k and G_k

$$F_k(z) = \frac{G_k(z)}{\sqrt{1-4z}}\,.$$

Using that

$$\frac{1}{\sqrt{1-4z}} = \sum_{i\geq 0}\binom{2i}{i}z^i,$$

we find that

$$F_k(z) = \sum_{i \geq k} \binom{2i-2k}{i-k} \mathcal{C}_k z^i \,.$$

Since the expectation of $\nu_k(z)$ is given by

$$\mathbb{E}[\nu_k(T)] = [z^n] F_k(z)/\mathcal{C}_n \,.$$

This completes the proof of the Lemma. □

Corollary 1. *Let $n \geq 2$. Consider a PT with n leaves sampled from the uniform model. Let K_n denotes the number of leaves descending from a uniformly chosen random ancestor. We have*

$$\mathbb{P}(K_n = k) = \frac{\mathcal{C}_{k-1} \binom{2n-2k}{n-k}}{(n-1)\mathcal{C}_{n-1}}, \quad \textit{for } k = 2, \ldots, n \,.$$

Proof. It is a direct consequence of Lemma 2. □

Remark 5. As n grows to infinity, the distribution of $\hat{K}_n$ converges to

$$P(\hat{K} = k) = 4^{-k} \mathcal{C}_k \sim \frac{1}{\sqrt{\pi} k^{3/2}} \quad \text{for large } k.$$

The tail of the distribution of $\hat{K}$ has a power law of parameter $3/2$. This can be compared to a similar result in the context of BST's (Martinez et al., 1998) under the random permutation model. In this case, $\hat{K}_n$ has power law distribution of parameter 2. Since $3/2$ is lower than 2, large random subtrees are more likely in the Catalan model than in the random permutation model. It was an expected result since Catalan binary trees are known to be more unbalanced than BST's under the random permutation model (Sedgewick and Flajolet, section 5.6, 1996).

Remark 6. Actually, the limiting distribution in the preceding comment is equal to the size of the critical Galton-Watson process with a binomial Bi(2,1/2) offspring distribution (Aldous, Lemma 9, 1991b). This is the Galton-Watson process corresponding to binary trees.

Now we are ready to prove that $\mathbb{E}[\sum_j \sqrt{\hat{N}_j}]/n^{3/2}$ converges to 0 as n goes to ∞. Let $\alpha \in]0,1[$ and split the sum $\mathbb{E}[\sum_j \sqrt{\hat{N}_j}]$ in two parts

$$\mathbb{E}\left[\sum_j \sqrt{\hat{N}_j}\right] = \mathbb{E}\left[\sum_{j,\hat{N}_j<n^\alpha} \sqrt{\hat{N}_j} + \sum_{j,\hat{N}_j\geq n^\alpha} \sqrt{\hat{N}_j}\right]$$

Obviously, we have

$$\begin{aligned}\frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j<n^\alpha} \sqrt{\hat{N}_j}\right] &\leq \frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j<n^\alpha} \sqrt{n^\alpha}\right] \\ &\leq n^{\alpha/2-1/2} \to 0\end{aligned}$$

when n grows to ∞. For the second term, we have

$$\frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j\geq n^\alpha} \sqrt{\hat{N}_j}\right] \leq \frac{1}{n}\mathbb{E}\left[\sum_{j,\hat{N}_j\geq n^\alpha} 1\right].$$

The right hand-side of the inequality is equal to $P(\hat{K}_n \geq n^\alpha)$. Applying Lemma 4, we find that

$$\frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j\geq n^\alpha} 1\right] \sim \kappa\, n^{-\alpha/2}$$

for some constant κ. This expression converges to 0 when n grows to ∞. This completes the proof of Theorem 3.

Remark 7. Fill and Kapur (2004) established more precise results concerning $\mathbb{E}[\sum_j \sqrt{\hat{N}_j}]$. They proved that

$$\mathbb{E}[\sum_j \sqrt{\hat{N}_j}] \sim \frac{1}{\sqrt{\pi}} n \log n \, . \tag{18}$$

Their results rely on Hadamard products. Note that our proof uses elementary arguments, and is instructive in its own as concerns the shape of Catalan trees. Besides, equation (18) follows easily from Lemma 4, by direct estimation of the sum by an integral.

Proof of Theorem 4. In the following, we prove the convergence of the mixed moments of $\hat{S}_n$ and $\hat{C}_n$ which states that for $k, l \geq 0$, we have

$$\mathbb{E}[\hat{S}_n^k \hat{C}_n^\ell] \sim n^{3(k+\ell)/2} \mathbb{E}[\mathcal{A}^{k+\ell}] \, . \tag{19}$$

The argument is similar to the argument given by Janson (Remark 3.5, 2003) to establish the convergence of the mixed moments of the internal path length and the Wiener index. Since the convergence in distribution has been established in Theorem 3, the above equation is equivalent to uniform integrability of $n^{-3(k+\ell)/2} \hat{S}_n^k \hat{C}_n^\ell$ for $n \geq 1$ and any fixed k, ℓ. Since $\hat{C}_n \leq \hat{S}_n$, the result follows from the fact that $n^{-3k/2} (\hat{S}_n)^k$ is uniformly integrable for every fixed k. This is true because Takacs (1991) obtained the convergence of the moments of $\hat{S}_n$. $\square$

References

[1] Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical functions*. Dover, New York.

[2] Aldous, D. J. (1991a). The continuum random tree II: an overview. In *Stochastic analysis*, eds. N.T. Barlow and N.H. Bingham, 23-70. Cambridge University Press.

[3] Aldous, D. J. (1991b). Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.* **1** 228-266.

[4] Aldous, D. J. (1995). Probability distributions on cladograms. In *Random Discrete Structures*, eds. D. Aldous and R. Pemantle, 1-18. Springer, Berlin.

[5] Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.* **16** 23-34.

[6] Agapow, P-M. and Purvis, A. (2002). Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology* **51** 866-872.

[7] Blum, M. G. B. and François, O. (2005). On statistical tests of phylogeny imbalance: the Sackin and other indices revisited. *Math. Biosci.* **195** 141-153.

[8] Colless, D. H. (1982). Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology* **31** 100-104.

[9] Darwin, C. (1859). *The origin of species*. Reprinted by Penguin books, London, UK.

[10] Fill, J. A. and Kapur, N. (2004). Limiting distributions for additive functionals on Catalan trees. *Theoret. Comput. Sci.* **326** 69-102.

[11] Flajolet, P. and Louchard, G. (2001). Analytic variations on the airy distribution. *Algorithmica* **31** 361-377.

[12] Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. in Appl. Probab.* **3** 44-77.

[13] Hoare, C. A. R. (1962). Quicksort. *Computer Journal* **5** 10-15.

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

[15] Janson, S. (2003). The Wiener index of simply generated random trees. *Random Structures Algorithms* **22** 337-358.

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

[17] Kirkpatrick, M. and Slatkin, M. (1993). Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* **47** 1171-1181.

[18] Knuth, D. E. (1973). *The Art of Computer Programming*, Volume 3: Sorting and Searching. Addison-Wesley, Reading, MA.

[19] Mahmoud, H. (1992). *Evolution of Random Search Trees.* Wiley, New York.

[20] Martinez, C., Panholzer, A. and Prodinger, H. (1998). The number of descendants and ascendants in random search trees. *Electron. J. Combin.* **5**.

[21] McKenzie, A. and Steel, M. (2001). Properties of phylogenetic trees generated by Yule-type speciation models. *Math. Biosci.* **170** 91-112.

[22] Mooers, A. and Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quart. Rev. Biol.* **72** 31-54.

[23] Neininger, R. (2001). On a multivariate contraction method for random recursive structures with applications to quicksort. *Random Structures Algorithms* **19** 498-524.

[24] Neininger, R. (2002). The Wiener index of random trees. *Combin. Probab. Comput.* **11** 587-597.

[25] Rachev, S. T. and Rüschendorf, L. (1995). Probability metrics and recursive algorithms. *Adv. in Appl. Probab.* **27** 770-799.

[26] Rogers, J. S. (1993). Response of Colless's tree imbalance to number of terminal taxa. *Systematic Biology* **42** 102-105.

[27] Rogers, J. S. (1994). Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution* **48** 2026-2036.

[28] Rösler, U. (1991). A limit theorem for "Quicksort". *Theor. Inform. and Appl.* **25** 85-100.

[29] Sackin, M. J. (1972). "Good" and "bad" phenograms. *Systematic Zoology* **21** 225-226.

[30] Sedgewick, R. and Flajolet, P. (1996). *An Introduction to the Analysis of Algorithms.* Addison Wesley, Reading, MA.

[31] Shao, K. and Sokal, R. R. (1990). Tree balance. *Systematic Zoology* **39** 66-276.

[32] Takacs, L. (1991). A Bernoulli excursion and its various applications. *Adv. in Appl. Probab.* **23** 557-585.

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

FIG 1. *Two graphical representations of the same PT. They are identical by swapping left and right subtrees.*

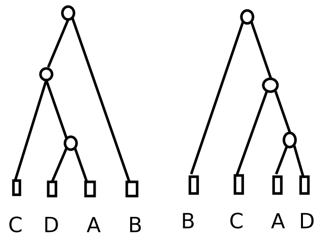

FIG 2. *The transformation of 2 binary search trees to the same PT. The extension consists of connecting two leaves to vertices with outdegree 0 and one leaf to vertices with outdegree 1. The two obtained trees represent the same PT.*