Two results about the Sackin and Colless indices for phylogenetic trees and their shapes

Gary Goh · Michael Fuchs · Louxin Zhang

Received: 12 February 2022 / Revised: 27 August 2022 / Accepted: 23 October 2022 / Published online: 23 November 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
The Sackin and Colless indices are two widely-used metrics for measuring the balance of trees and for testing evolutionary models in phylogenetics. This short paper contributes two results about the Sackin and Colless indices of trees. One result is the asymptotic analysis of the expected Sackin and Colless indices of tree shapes (which are full binary rooted unlabelled trees) under the uniform model where tree shapes are sampled with equal probability. Another is a short direct proof of the closed formula for the expected Sackin index of phylogenetic trees (which are full binary rooted trees with leaves being labelled with taxa) under the uniform model.

Keywords
Phylogenetics · Tree balance · Sackin index · Colless index · Asymptotic analysis

Mathematics Subject Classification 05A16 · 05C30 · 92D15

1 Introduction
The Sackin (1972), Shao and Sokal (1990) and Colless (1982) indices are two widely-used metrics for measuring the balance of phylogenetic trees and testing evolutionary models (Avino et al. 2019; Xue et al. 2020; Mooers and Heard 1997; Scott et al. 2020;
Blum et al. 2006; Kirkpatrick and Slatkin 1993). Phylogenetic trees are binary rooted trees in which each internal node has two children and only the leaves are labelled one-to-one with taxa. For a phylogenetic tree, its Sackin index is defined as the sum over its internal nodes of the number of leaves below that node, whereas its Colless index is defined as the sum over its internal nodes of the balance of that node, where the balance of a node is defined to be the difference in the number of leaves below the two children of that node. Because of their wide applications, the two tree balance metrics have been studied in the past decades (see the recent comprehensive survey Fischer et al. 2021).

The Sackin and Colless indices of a random phylogenetic tree have been investigated under the Yule-Harding model (where ordered tree shapes, i.e., tree shapes with the children of each non-leaf node having a left-to-right order, are generated using a birth process, the leaves of these ordered tree shapes are labeled according to a permutation on the taxa which is chosen uniformly at random, and then the left-to-right order of the children of non-leaf nodes is disregarded) and the uniform model (where trees are sampled with equal probability) (Kirkpatrick and Slatkin 1993; Heard 1992; Blum and François 2005; Blum et al. 2006). The expected Sackin and Colless indices of a phylogenetic tree are proved to be asymptotic to $\sqrt{\pi n^{3/2}}$ under the uniform model and $n \log n$ under the Yule-Harding model (Blum and François 2005; Blum et al. 2006). Recently, Mir and Rosselló (2013) discovered surprisingly that the expected Sackin index of a phylogenetic tree is simply

$$\frac{4^n - 1}{(2n-2)!} - n \quad (1)$$

under the uniform model. An alternative proof of this closed formula was given by King and Rosenberg (2021). Both asymptotic and exact results on the variances of the Sackin and Colless indices have also been reported (Kirkpatrick and Slatkin 1993; Blum and François 2005; Blum et al. 2006; Coronado et al. 2020).

It is not hard to see that the Sackin index of a binary tree is actually equal to the sum of the depths of all its leaves (Steel 2016). Therefore, the Sackin index and the tree height have also been studied for other types of trees in the combinatorics and theoretical computer science literature (Flajolet and Odlyzko 1982; Broutin and Flajolet 2012; Fill and Kapur 2004; Fuchs and Jin 2015).

In this paper, we focus on two questions about the Sackin and Colless indices. The first question is what the expected Sackin and Colless indices of a random binary tree shape are under the uniform model (Rogers 1996). Here, tree shapes (also called Otter or Polya trees) are binary rooted trees with unlabeled leaves where each internal node has two children. Although there is increasing interest in tree balance indices for tree shapes in the study of phylodynamic problems (Colijn and Plazzotta 2018; Kim et al. 2020), to the best of our knowledge, the statistical properties of these two indices and other tree balance indices have not been formally studied for tree shapes (Fischer et al. 2021). Here, we prove that the expected Sackin and Colless indices of a tree shape with $n$ leaves are asymptotic to $\sqrt{\pi} \lambda^{-1} n^{3/2}$ under the uniform model, where $\lambda \approx 1.1300337163$. 

\[\text{Springer}\]
Given that the closed formula (1) for the expected Sackin index of a phylogenetic tree under the uniform model is rather simple, the second question is whether a direct proof exists for the formula or not. We answer this question by using a simple recurrence for the Sackin index that is derived using the fact that all the phylogenetic trees on $n$ taxa can be enumerated by inserting the $n$-th taxon into every edge of the phylogenetic trees on $n - 1$ taxa (Felsenstein 2004). Recently, this technique was used by Zhang for computing the sum over all nodes of the number of the descendants of that node and counting the number of tree-child networks with one reticulation (Zhang 2019).

2 Basic definitions and notation

2.1 Phylogenetic trees and shapes

A tree shape is a full binary rooted tree in which all nodes are unlabeled. A phylogenetic tree on $n$ taxa is a full binary rooted tree with $n$ leaves in which its leaves are uniquely labeled with a taxon and each of the $n - 1$ non-leaf nodes has two children.

Let $T$ be a phylogenetic tree on $n$ taxa or a tree shape. We use $V_0(T)$ to denote the set of all non-leaf nodes of $T$ and $V(T)$ to denote the set of all nodes. A leaf $x$ is said to be below a node $u$ in $T$ if the unique path from the root to $x$ passes through $u$. We use $\ell_T(u)$ to denote the number of leaves below $u$ in $T$. Also, we set $\ell_T(u) := 1$ if $u$ is a leaf.

Let $u \in V_0(T)$. The balance of $u$ is defined to be $|\ell_T(v) - \ell_T(w)|$, where $v$ and $w$ are the two children of $u$. We use $\delta_T(u)$ to denote the balance of $u$.

For each non-root $u \in V(T)$, we use $p(u)$ to denote the parent of $u$ in $T$.

2.2 Sackin and Colless indices

**Definition 1** The Sackin index of a tree shape or a phylogenetic tree $T$ is defined to be $\sum_{u \in V_0(T)} \ell_T(u)$, and denoted by $S(T)$.

**Definition 2** The Colless index of a tree shape or a phylogenetic tree $T$ is defined to be $\sum_{u \in V_0(T)} \delta_T(u)$, and denoted by $C(T)$.

The expected Sackin and Colless indices of a tree shape under the uniform model are respectively defined as:

$$ESI_{sh}(n) := \frac{1}{b_n} \sum_{T \in T(n)} S(T)$$

and

$$ECI_{sh}(n) := \frac{1}{b_n} \sum_{T \in T(n)} C(T).$$
where $T(n)$ denotes the set of all tree shapes with $n$ leaves and $b_n := |T(n)|$. Although there does not exist a closed formula for $b_n$, $b_n$ can be computed using the following recurrence formulas for $n > 1$ (A001190 in the On-Line Encyclopedia of Integer Sequences\(^1\)):

\[
 b_n = \sum_{1 \leq k < n/2} b_k b_{n-k} + \begin{cases} 
 0, & \text{if } n \text{ is odd;} \\
 \frac{1}{2} b_{n/2}(b_{n/2} + 1), & \text{if } n \text{ is even.}
\end{cases}
\] (2)

Equivalently, the generating function $B(z) := \sum_i b_i z^i$ satisfies the following equation:

\[
 B(z) = z + \frac{1}{2} \left( B(z)^2 + B(z^2) \right).
\] (3)

The expected Sackin index of a phylogenetic tree under the uniform model is defined similarly, that is,

\[
 \text{ESI}_p(n) := \frac{1}{a_n} \sum_{P \in \mathcal{P}(n)} S(P),
\]

where $\mathcal{P}(n)$ denotes the set of all phylogenetic trees on $n$ taxa and $a_n := |\mathcal{P}(n)| = \frac{(2n-2)!}{2^{n-1}(n-1)!}$ (see Steel 2016).

3 Asymptotic analysis of the expected Sackin and Colless indices for tree shapes

Recall that $T(n)$ denotes the set of all possible tree shapes with $n$ leaves. Let $S_n := \sum_{T \in T(n)} S(T)$, which is the sum of the Sackin index over all tree shapes with $n$ leaves. Obviously, $S_1 = 0$ and $S_2 = 2$.

For $n > 2$, $T(n)$ can be obtained by combining every pair of tree shapes $T' \in T(k)$ and $T'' \in T(n-k)$, where $k$ can range from 1 to $\lfloor n/2 \rfloor$. For a specific integer $k \leq n/2$, $T \in T(k)$ and $T' \in T(n-k)$, $S(T) = n + S(T') + S(T'')$ for the tree shape $T$ obtained by combining $T'$ and $T''$, as there are $n$ leaves below the root of $T$.

\(^1\) https://oeis.org/
Using the facts mentioned in the previous paragraph and Eq. (2), we obtain that:

$$S_n = \sum_{1 \leq k < n/2} \left( \sum_{T \in T(k)} \sum_{T' \in T(n-k)} (n + S(T) + S(T')) \right)$$

$$= \sum_{1 \leq k < n/2} \left( nb_k b_{n-k} + \sum_{T \in T(k)} \sum_{T' \in T(n-k)} (S(T) + S(T')) \right)$$

$$= n \sum_{1 \leq k < n/2} b_k b_{n-k}$$

$$+ \sum_{1 \leq k < n/2} \left( \sum_{T \in T(k)} \sum_{T' \in T(n-k)} S(T) + \sum_{T \in T(k)} \sum_{T' \in T(n-k)} S(T') \right)$$

$$= nb_n + \sum_{1 \leq k < n/2} (b_{n-k} S_k + b_k S_{n-k})$$

$$= nb_n + \sum_{1 \leq k < n} S_k b_{n-k}, \quad (4)$$

for odd $n$ and

$$S_n = nb_n + \sum_{1 \leq k < n/2} \left( \sum_{T \in T(k)} \sum_{T' \in T(n-k)} (S(T) + S(T')) \right)$$

$$+ \sum_{T, T' \in T(n/2) : T \neq T'} (S(T) + S(T')) + \sum_{T \in T(n/2)} 2S(T)$$

$$= nb_n + \sum_{1 \leq k < n/2} (b_{n-k} S_k + b_k S_{n-k}) + \left( \sum_{T \in T(n/2)} (b_n/2 - 1) S(T) \right) + 2S_{n/2}$$

$$= nb_n + S_{n/2} + \sum_{1 \leq k < n} S_k b_{n-k}, \quad (5)$$

for even $n$.

### 3.1 The asymptotic value of $ESI_{sh}(n)$

It is unknown whether or not one can derive a closed formula for $S_n$ from Eqs. (4–5). However, an asymptotic analysis of $S_n$ follows from the classical asymptotic analysis of $b_n$ from Eq. (2). In order to recall the latter, we need the notion of $\Delta$-analyticity. First, a $\Delta$-domain with parameters $\delta$ and $\phi$ is a domain in the complex plane of the form:

$$\Delta := \{ z \in \mathbb{C} : |z| < 1 + \delta, \ |\arg(z - 1)| > \phi \}$$
with \( \delta > 0 \) and \( 0 < \phi < \pi / 2 \); see Definition VI.1 in Flajolet and Sedgewick (2009). A function is called \( \Delta \)-analytic if it is analytic in such a \( \Delta \)-domain.

**Lemma 1** (Broutin and Flajolet 2012) The convergence radius \( \rho \) of the generating function \( B(z) \) of \( b_n \) in Eq. (3) satisfies \( 1/4 \leq \rho \leq 1/2 \), where \( \rho + B(\rho^2)/2 = 1/2 \). Moreover, \( B(z) \) is \( \Delta \)-analytic and satisfies as \( z \to \rho \) in a \( \Delta \)-domain:

\[
B(z) = 1 - \lambda \sqrt{1 - z/\rho} + \mathcal{O}(1 - z/\rho), \quad \lambda := \sqrt{2\rho + 2\rho^2 B'(\rho^2)}.
\]

Thus,

\[
b_n \sim \frac{\lambda}{2\sqrt{\pi n^{3/2} \rho^n}}, \quad (n \to \infty).
\]

**Remark 1** \( \rho \) and \( \lambda \) can be computed up to very high precision, e.g.,

\[
\rho = 0.40269750367 \ldots \quad \text{and} \quad \lambda = 1.1300337163 \ldots.
\]

The computation is done as follows: first, Eq. (2) is used to compute a finite number of terms \( \tilde{b}_n \) of \( b_n \) that are used to set up a polynomial \( \tilde{B}(z) \) which approximates \( B(z) \); then, find \( \tilde{\rho} \) such that \( \tilde{\rho} + \tilde{B}(\rho^2)/2 = 1/2 \). Clearly, \( \tilde{\rho} \) approximates \( \rho \) and this approximation can be made arbitrarily precise; also an approximation of \( \lambda \) can be derived from it via Eq. (6).

**Remark 2** The asymptotic expansion in Eq. (7) follows from the singularity expansion in Eq. (6) by the transfer theorems (see Theorem VI.3 and Corollary VI.1 in Flajolet and Sedgewick (2009)) which assert that if \( A(z) \) is \( \Delta \)-analytic with \( A(z) \sim c(1 - z/\rho)^{-\alpha} \), where \( c, \rho \in \mathbb{R} \setminus \{0\} \) and \( \alpha \in \mathbb{C} \setminus \{0, -1, -2, \ldots\} \), then \([z^n]A(z) \sim [z^n]c(1 - z/\rho)^{-\alpha} \sim c\rho^{-n}\alpha^{-1}/\Gamma(\alpha)\), where \([z^n]f(z)\) denotes the \( n \)-th coefficient in the Maclaurin series of \( f(z) \) and \( \Gamma(z) \) is the gamma function. Indeed, set \( A(z) := B(z) - 1 \). Then, by (6), we have \( A(z) \sim -\lambda \sqrt{1 - z/\rho} \) and thus

\[
b_n = [z^n]A(z) \sim \frac{-\lambda}{\Gamma(-1/2)n^{3/2} \rho^n} = \frac{\lambda}{2\sqrt{\pi n^{3/2} \rho^n}}
\]

which is (7). (In the last step, we used that \( \Gamma(-1/2) = -2\sqrt{\pi} \).)

**Remark 3** More generally, the process of showing that a genearting function \( A(z) \) is \( \Delta \)-analytic, deriving the expansion \( A(z) \sim c(1 - z/\rho)^{-\alpha} \) as \( z \to \rho \) and then using the transfer theorems to obtain the asymptotics of \([z^n]A(z)\) is called singularity analysis; see Chapter VI in Flajolet and Sedgewick (2009).

**Remark 4** Singularity analysis is closed under several operations on functions; see Section VI.10 in Flajolet and Sedgewick (2009). For instance, if singularity analysis can be applied to \( A(z) \), it can also be applied to \( A'(z) \), where the singularity expansion
of $A'(z)$ is obtained from the one of $A(z)$ by term-by-term differentiation. E.g., $B'(z)$ from the previous lemma is also $\Delta$-analytic with singularity expansion as $z \to \rho$

$$B'(z) \sim \frac{\lambda}{2\rho} \cdot \frac{1}{\sqrt{1 - z/\rho}}, \quad (8)$$

from which the asymptotic expansion of $[z^n] B'(z)$ follows by the transfer theorems. (Of course, since $[z^n] B'(z) = (n + 1)[z^{n+1}] B(z)$, this expansion is just the expansion in Eqn. (7) multiplied by $n/\rho$.)

**Theorem 2** Under the uniform model, the expected Sackin index of a tree shape with $n$ leaves, $\text{ESI}_{\text{sh}}(n)$, is asymptotic to $\pi^{1/2} \lambda^{-1} n^{3/2}$, where $\lambda$ is given in Eq. (6).

**Proof** The recurrence formulas in Eqs. (4–5) translate into the following equation for the generating function $S(z) := \sum S_i z^i$ of $S_n$:

$$S(z) = zB'(z) + SB(z) + S(z^2) \quad (9)$$

since the generating function of $\sum_{1 \leq k < n} S_k b_{n-k}$ is the product $S(z)B(z)$ and

$$\sum_{n \geq 1} nb_n z^n = zB'(z), \quad \sum_{n \text{ even}} S_{n/2} z^n = S(z^2).$$

Indeed,

$$S(z) = \sum_{n \geq 1} S_n z^n = \sum_{n \geq 1} nb_n z^n + \sum_{n \geq 1} \left( \sum_{1 \leq k < n} S_k b_{n-k} \right) z^n + \sum_{n \text{ even}} S_{n/2} z^n$$

$$= zB'(z) + SB(z) + S(z^2)$$

which gives (9).

Next, by rewriting Eq. (9) into

$$S(z) = \frac{zB'(z) + S(z^2)}{1 - B(z)},$$

we see that the radius of convergence of $S(z)$ is equal to $\rho$. (Because $B(z)$ and $B'(z)$ both have radius of convergence equal to $\rho$ and $S(z^2)$ is analytic at $\rho$ since $0 < \rho < 1$.) Moreover, from Eq. (6) and the closure properties of singularity analysis (Remark 4 above), we obtain that $S(z)$ is $\Delta$-analytic and satisfies as $z \to \rho$ in a $\Delta$-domain:

$$S(z) \sim \frac{\rho(\lambda/2\rho)(1 - z/\rho)^{-1/2} + S(\rho^2)}{\lambda \sqrt{1 - z/\rho} + O(1 - z/\rho)} \sim \frac{1}{2} \cdot \frac{1}{1 - z/\rho},$$

where we used Eq. (8) and that $S(z^2)$ is analytic at $\rho$. 
By the transfer theorems (see Remark 2), we obtain:

\[ S_n \sim \frac{1}{2} [z^n] (1 - z/\rho)^{-1} \sim \frac{1}{2\rho^n}, \quad (n \to \infty) \tag{10} \]

and thus

\[ \text{ESI}_{sh}(n) = \frac{S_n}{b_n} \sim \frac{1}{\lambda} \frac{1/(2\rho^n)}{(2\sqrt{n}^{3/2}\rho^n)} = \sqrt{\pi}^{-1} n^{-3/2}, \quad (n \to \infty) \]

using Eqn. (7). This proves the claim. \( \square \)

### 3.2 The asymptotic value of ECI\(_{sh}(n)\)

Next, we derive the asymptotic value of ECI\(_{sh}(n)\). First, for each internal node \( u \) of a tree, we use \( c_1(u) \) and \( c_2(u) \) to denote the two children of \( u \). We have that \( \ell(u) = \ell(c_1(u)) + \ell(c_2(u)) \) and thus \( \delta(u) = |\ell(c_1(u)) - \ell(c_2(u))| = \ell(u) - 2 \min(\ell(c_1(u)), \ell(c_2(u))). \) From this, it follows that for each tree shape \( T \),

\[ D(T) := S(T) - C(T) = 2 \sum_{u \in V_0(T)} \min(\ell(c_1(u)), \ell(c_2(u))). \]

Defining

\[ D_n := \frac{1}{2} \sum_{T \in T(n)} D(T), \]

we obtain:

\[ C_n = \sum_{T \in T(n)} C(T) = S_n - 2D_n. \tag{11} \]

Since the integer subsets \( \{k : 1 \leq k < n/2\} \) and \( \{k : 1 \leq k \leq n/2\} \) are the same and \( \sum_{1 \leq k < n/2} b_{n-k}D_k = \sum_{n/2 \leq n-k < n} b_{n-k}D_{n-(n-k)} = \sum_{n/2 \leq k < n} b_kD_{n-k} \) for odd \( n \), we have the following recurrence formula:

\[ D_n = \sum_{1 \leq k < n/2} \left( \sum_{T \in T(k)} \sum_{T^s \in T(n-k)} (D(T) + D(T^s) + k) \right) \]

\[ = \sum_{1 \leq k < n/2} kb_k b_{n-k} + \sum_{1 \leq k < n/2} (b_{n-k}D_k + b_kD_{n-k}) \]

\[ = \sum_{1 \leq k \leq n/2} kb_k b_{n-k} + \sum_{1 \leq k < n} D_k b_{n-k}, \quad \text{for odd } n. \tag{12} \]
We now consider the case when \( n \) is even. Since each shape with \( n/2 \) leaves can form exactly \( b_{n/2} - 1 \) shape pairs with all \( b_{n/2} - 1 \) other shapes, we have that

\[
\sum_{T, T' \in \mathcal{T}(n/2); T \neq T'} (D(T) + D(T')) = \sum_{T \in \mathcal{T}(n/2)} (b_{n/2} - 1)D(T) = (b_{n/2} - 1)D(n/2).
\]

We also have that, for even \( n \),

\[
\sum_{1 \leq k < n/2} (b_{n-k}D_k + b_kD_{n-k}) + b_{n/2}D_{n/2} = \sum_{1 \leq k < n} b_{n-k}D_k.
\]

Therefore, for even \( n \),

\[
D_n = \sum_{1 \leq k < n/2} \left( \sum_{T \in \mathcal{T}(k)} \sum_{T' \in \mathcal{T}(n-k)} (D(T) + D(T') + k) \right) + \sum_{T, T' \in \mathcal{T}(n/2); T \neq T'} (D(T) + D(T') + n/2) + \sum_{T \in \mathcal{T}(n/2)} (2D(T) + n/2)
\]

\[
= \sum_{1 \leq k < n/2} kb_kb_{n-k} + \sum_{1 \leq k < n/2} (b_{n-k}D_k + b_kD_{n-k}) + (b_{n/2} - 1)D_{n/2} + \left(\frac{b_{n/2}}{2}\right)n + 2D_{n/2} + \frac{n}{2}b_{n/2}
\]

\[
= \sum_{1 \leq k \leq n/2} kb_kb_{n-k} + \sum_{1 \leq k < n} D_kb_{n-k} - \frac{n}{2}\left(\frac{b_{n/2}}{2}\right) + D_{n/2}.
\]

(13)

We first need a technical lemma for:

\[
F_n := \sum_{1 \leq k \leq n/2} kb_kb_{n-k} + \begin{cases} 0, & \text{if } n \text{ is odd;} \\ -\frac{n}{2}\left(\frac{b_{n/2}}{2}\right), & \text{if } n \text{ is even.} \end{cases}
\]

**Lemma 3** We have \( F_n = \mathcal{O}(n^{-1} \rho^{-n}) \).

**Proof** By using Eq. (7),

\[
F_n = \mathcal{O}\left(\rho^{-n}\sum_{1 \leq k \leq n/2} k^{-1/2}(n-k)^{-3/2} + n^{-2} \rho^{-n}\right)
\]

\[
= \mathcal{O}\left(n^{-1} \rho^{-n}\int_0^{1/2} x^{-1/2}(1-x)^{-3/2}dx + n^{-2} \rho^{-n}\right)
\]

\[
= \mathcal{O}\left(n^{-1} \rho^{-n} + n^{-2} \rho^{-n}\right) = \mathcal{O}(n^{-1} \rho^{-n}),
\]

where in the second step, we approximated the sum by an integral. \( \square \)
Now, define:

\[
\tilde{D}_n := Kn^{-1} \rho^{-n} + \sum_{1 \leq k < n} \tilde{D}_k b_{n-k} + \begin{cases} 0, & \text{for } n \text{ is odd;} \\ \tilde{D}_{n/2}, & \text{for } n \text{ is even}, \end{cases}
\] (14)

where \( K \) is the implied \( O \)-constant from the last lemma. The reason for considering this sequence is that it (a) majorizes \( D_n \), namely, \( D_n \leq \tilde{D}_n \) (which is easily proved by induction) and (b) its asymptotics can derived with similar tools as used in the proof of Theorem 5.

**Lemma 4** We have,

\[
\tilde{D}_n \sim \frac{K}{\lambda \sqrt{\pi}} n^{-1/2} (\log n) \rho^{-n}, \quad (n \to \infty).
\]

Consequently, \( D_n = O (n^{-1/2} (\log n) \rho^{-n}) \).

**Proof** Let \( \tilde{D}(z) := \sum_i \tilde{D}_i z^i \) be the generating function of \( \tilde{D}_n \). Then, the recurrence in Eqn. (14) translates into

\[
\tilde{D}(z) = K \log \frac{1}{1 - z/\rho} + \tilde{D}(z) B(z) + \tilde{D}(z^2) \tag{15}
\]

since

\[
\sum_{n \geq 1} K n^{-1} \rho^{-n} z^n = K \log \frac{1}{1 - z/\rho}
\]

and the rest of the terms in (15) are explained as in the derivation of Eq. (9). Solving for \( D(z) \) gives:

\[
\tilde{D}(z) = \frac{K \log \frac{1}{1 - z/\rho} + \tilde{D}(z^2)}{1 - B(z)}.
\]

Thus, from Eq. (6), \( \tilde{D}(z) \) satisfies as \( z \to \rho \) in a \( \Delta \)-domain:

\[
\tilde{D}(z) \sim \frac{K \log \frac{1}{1 - z/\rho} + \tilde{D}(\rho^2)}{\lambda \sqrt{1 - z/\rho} + O(1 - z/\rho)} \sim \frac{K}{\lambda} \cdot \frac{\log \frac{1}{1 - z/\rho}}{\sqrt{1 - z/\rho}}
\]

from which the claimed result follows by the transfer theorems (which also work with log-factors; see Theorem VI.3 in Flajolet and Sedgewick 2009).

Now from Eqs. (10), (11) and Lemma 4, we have the following result.

**Theorem 5** Under the uniform model, the expected Colless index of a tree shape with \( n \) leaves, \( \text{ECI}_{sh}(n) \), is asymptotic to \( \pi^{1/2} \lambda^{-1} n^{3/2} \).
3.3 Visualization on the asymptotic analyses

The exact and asymptotic values of \( ESI_{sh}(n) \) and \( ECI_{sh}(n) \) were computed and compared for \( n \) up to 700 (Fig. 1). Here, For \( ESI_{sh}(n) \), the exact values were computed using the formulas in Eqs. (4–5) and the asymptotic values were computed using the formula in Theorem 2. For \( ECI_{sh}(n) \), the exact values were computed using the formulas in Eqs. (11–13) and the asymptotic values were computed using Theorem 5. The comparison indicates that the asymptotic value \( \sqrt{\pi \lambda} \frac{1}{n^{3/2}} \) is a very good approximation to the Sackin index even for a small number \( n \). However, the asymptotic value overestimates the Colless index with a relatively large margin. The large margin is due to the fact that \( ESI_{sh}(n) - ECI_{sh}(n) = 2D(n)/b(n) \) is of the order \( n \log n \) according to our proof; however, the relative error will tend to 0 with a speed of at least \( \log n / \sqrt{n} \).

4 The expected Sackin index for phylogenetic trees

Mir et al. discovered the following simple closed formula for the expected Sackin index for a phylogenetic tree under the uniform model.

**Theorem 6** *(Mir and Rosselló 2013)* For any \( n \), \( ESI_p(n) = \frac{4^{n-1} n! (n-1)!}{(2n-2)!} - n \).

An alternative proof was presented in King and Rosenberg (2021) recently. Here, we will present a short direct proof using the following enumeration of phylogenetic trees (see Felsenstein 2004 for example):

Assume that there is an open edge entering the root of each phylogenetic tree. \( \mathcal{P}(n+1) \) can be obtained from \( \mathcal{P}(n) \) by attaching Leaf \( n+1 \) on each of the \( 2n-1 \) edges of every tree of \( \mathcal{P}(n) \) (Fig. 2a).

Let \( S_n^{(p)} := \sum_{P \in \mathcal{P}(n)} S(P) \). Note that \( S_n^{(p)} = ESI_p(n) \times a_n \), where \( a_n = |\mathcal{P}(n)| \). For each \( P \in \mathcal{P}(n) \), we use \( \mathcal{A}(P) \) to denote the set of \( 2n-1 \) phylogenetic trees on \( n+1 \) taxa that are obtained from \( P \) by attaching Leaf \( n+1 \) on each of the \( 2n-1 \) tree edges of \( P \). Then,

\[
S_n^{(p)} = \sum_{P \in \mathcal{P}(n)} \sum_{Q \in \mathcal{A}(P)} S(Q). \tag{16}
\]
Consider a tree $Q \in \mathcal{A}(P)$. Note that Leaf $n + 1$ and its parent are the only nodes of $Q$ that are not found in $P$. Assume that $Q$ is obtained by attaching Leaf $n + 1$ to the edge $e$ that enters $u$ in $P$. The number of leaves below the parent of Leaf $n + 1$ is $1 + \ell_P(u)$ in $Q$ (Fig. 2b). Therefore, the amount contributed by the parents of Leaf $n + 1$ to the sum $\sum_{Q \in \mathcal{A}(P)} S(Q)$ is:

$$\sum_{u \in V(P)} (1 + \ell_P(u)) = (2n - 1) + \sum_{u \in V(P)} \ell_P(u)$$

$$= (2n - 1) + n + \sum_{u \in V_0(P)} \ell_P(u)$$

$$= 3n - 1 + S(P),$$

where the $n$ in the second expression is the sum of $\ell_P(u)$ (which is 1) over all the $n$ leaves $u$ in $P$ and $V(P)$ and $V_0(P)$ are the set of nodes and non-leaf nodes, respectively, in $P$; see Sect. 2.1.

For $w \in V(P)$, we have either $\ell_P(w) = \ell_Q(w)$ or $\ell_P(w) = \ell_Q(w) + 1$. Furthermore, the latter holds if and only if $Q$ is obtained by attaching Leaf $n + 1$ to an edge below $w$ in $P$. Since there are $2\ell_P(w) - 2$ edges below $w$ in $P$, thus $\ell_Q(w) = \ell_P(w) + 1$ for exactly $2\ell_P(w) - 2$ trees $Q$ of $\mathcal{A}(P)$. Therefore,

$$\sum_{Q \in \mathcal{A}(P)} S(Q) = (2n - 1)S(P) + (S(P) + (3n - 1)) + \sum_{w \in V_0(P)} (2\ell_P(w) - 2)$$

$$= 2nS(P) + (3n - 1) + 2S(P) - 2|V_0(P)|$$

$$= 2(n + 1)S(P) + (n + 1).$$
Adding $n + 1$ to each term in the left-hand side of the above equality, which can be considered as the contribution of the $n + 1$ leaves, we further have:

$$
\sum_{Q \in A(P)} (S(Q) + (n + 1)) = 2(n + 1)S(P) + (n + 1) + (2n - 1)(n + 1)
= 2(n + 1)(S(P) + n).
$$

By Eq. (16), we obtain the following simple recurrence formula:

$$
S_{n+1}^{(p)} + (n + 1)a_{n+1} = \sum_{P \in \mathcal{P}(n)} \sum_{Q \in A(P)} (S(Q) + (n + 1))
= \sum_{P \in \mathcal{P}(n)} 2(n + 1)(S(P) + n)
= 2(n + 1)\left(S_{n}^{(p)} + na_{n}\right).
$$

(18)

Since $S_{2}^{(p)} = 2$ and $a_{2} = 1$, Eq. (18) implies that $S_{n}^{(p)} = 2^{n-1}n! - na_{n}$ and

$$
\text{ESI}_{p}(n) = \frac{S_{n}^{(p)}}{a_{n}} = \frac{4^{n-1}n!(n - 1)!}{(2n - 2)!} - n
$$

Theorem 6 is proved.

5 Conclusion

In this short paper, we contributed two results to the study of the Sackin and Colless indices. We have proved that the asymptotic value of Sackin and Colless indices are the same for tree shapes under the uniform model. The same phenomenon was also observed for phylogenetic trees under the uniform model; see Blum et al. (2006). Thus, our result is expected since tree shapes under the uniform model are known to behave similar to phylogenetic trees under the uniform model; see the discussion in the introduction of Broutin and Flajolet (2012). In particular, the average height of phylogenetic trees and binary tree shapes with $n$ leaves are both asymptotically equal to $2\lambda^{-1}\sqrt{\pi n}$ (see Flajolet and Odlyzko 1982 and Broutin and Flajolet 2012).

We also presented a short direct proof of the closed formula for the expected Sackin index of phylogenetic trees under the uniform model. The proof is based on a tree enumeration approach that is different from one used in Mir and Rosselló (2013) and King and Rosenberg (2021). This technique was also used by Goh (2022) to derive a short proof of the closed formula for the expected total cophenetic index of a phylogenetic tree under the uniform model that was introduced in Mir and Rosselló (2013) (see also Fischer et al. 2021). It is an interesting problem whether or not the proof technique in Section 4 can be used to investigate other tree balance indices (such as those given in the survey paper Fischer et al. 2021).
Acknowledgements The authors thank the two anonymous reviewers for useful suggestions and comments for preparing the final version of this paper. LZ was supported by MOE Tier 1 grant R-146-000-318-114; MF was supported by MOST-109-2115-M-004-003-MY2.

Author Contributions GG: Recurrence formulas; LZ: Recurrence formulas, writing; MF: Asymptotic analysis, writing.

Declarations

Conflicts of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Avino M, Ng GT, He Y, Renaud MS, Jones BR, Poon AF (2019) Tree shape-based approaches for the comparative study of cophylogeny. Ecol Evol 9(12):6756–6771
Blum MG, François O (2005) On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. Math Biosci 195(2):141–153
Blum MG, François O, Janson S (2006) The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. Ann Appl Probab 16(4):2195–2214
Blum MGB, Heyer E, François O, Austerlitz F (2006) Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. PLoS Genet 2(8):122
Brouin N, Flajolet P (2012) The distribution of height and diameter in random non-plane binary trees. Random Struct Algorithms 41(2):215–252
Colijn C, Plazzotta G (2018) A metric on phylogenetic tree shapes. Syst Biol 67(1):113–126
Colless DH (1982) Review of “phylogenetics: the theory and practice of phylogenetic systematics”. Syst Zool 31(1):100–104
Coronado TM, Mir A, Rosselló F, Rotger L (2020) On Sackin’s original proposal: the variance of the leaves’ depths as a phylogenetic balance index. BMC Bioinform 21(1):1–17
Felsenstein J (2004) Inferring Phylogenies. Sinauer Assoc Inc, Sunderland
Fill JA, Kapur N (2004) Limiting distributions for additive functionals on Catalan trees. Theoret Comput Sci 326(1–3):69–102
Fischer M, Herbst L, Kersting S, Kühn L, Wicke K (2021) Tree balance indices: a comprehensive survey. arXiv preprint arXiv:2109.12281
Flajolet P, Odlyzko A (1982) The average height of binary trees and other simple trees. J Comput Syst Sci 25(2):171–213
Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge
Fuchs M, Jin EY (2015) Equality of Shapley value and fair proportion index in phylogenetic trees. J Math Biol 71(5):1133–1147
Goh G (2022) Metrics for measuring the shape of phylogenetic trees. Honors Thesis, National University of Singapore
Heard SB (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46(6):1818–1826
Kim J, Rosenberg NA, Palacios JA (2020) Distance metrics for ranked evolutionary trees. Proc Natl Acad Sci 117(46):28876–28886
King MC, Rosenberg NA (2021) A simple derivation of the mean of the Sackin index of tree balance under the uniform model on rooted binary labeled trees. Math Biosci 342:108688
Kirkpatrick M, Slatkin M (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47(4):1171–1181
Mir A, Rosselló F et al (2013) A new balance index for phylogenetic trees. Math Biosci 241(1):125–136
Mooers AO, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 72(1):31–54
Rogers JS (1996) Central moments and probability distributions of three measures of phylogenetic tree imbalance. Syst Biol 45(1):99–110

Springer
Sackin MJ (1972) “Good” and “Bad” Phenograms. Syst Biol 21(2):225–226. https://doi.org/10.1093/sysbio/21.2.225
Scott JG, Maini PK, Anderson AR, Fletcher AG (2020) Inferring tumor proliferative organization from phylogenetic tree measures in a computational model. Syst Biol 69(4):623–637
Shao K-T, Sokal RR (1990) Tree balance. Syst Zool 39(3):266–276
Steel M (2016) Phylogeny: discrete and random processes in evolution. SIAM, Philadelphia
Xue C, Liu Z, Goldenfeld N (2020) Scale-invariant topology and bursty branching of evolutionary trees emerge from niche construction. Proc Natl Acad Sci 117(14):7879–7887
Zhang L (2019) Generating normal networks via leaf insertion and nearest neighbor interchange. BMC Bioinform 20(20):1–9

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.