Revisiting Shao and Sokal's B_2 index of phylogenetic balance

François Bienvenu[1,2], Gabriel Cardona[1,3], and Celine Scornavacca[1]

[1] *Institut des Sciences de l'Evolution de Montpellier (Université de Montpellier, CNRS, IRD, EPHE), F-34095 Montpellier, France*
[2] *UMR AGAP (Université de Montpellier, CIRAD, INRAE, L'institut Agro), F-34398 Montpellier, France*
[3] *Department of Mathematics and Computer Science, University of the Balearic Islands, Ctra. Valldemossa km 7.5, E-07120 Palma, Spain*

November 18, 2021

Abstract

Measures of phylogenetic balance, such as the Colless and Sackin indices, play an important role in phylogenetics. Unfortunately, these indices are specifically designed for phylogenetic trees, and do not extend naturally to phylogenetic networks (which are increasingly used to describe reticulate evolution). This led us to consider a lesser-known balance index, whose definition is based on a probabilistic interpretation that is equally applicable to trees and to networks. This index, known as the B_2 index, was first proposed by Shao and Sokal in 1990. Surprisingly, it does not seem to have been studied mathematically since. Likewise, it is used only sporadically in the biological literature, where it tends to be viewed as arcane. In this paper, we study mathematical properties of B_2 such as its expectation and variance under the most common models of random trees and its extremal values over various classes of phylogenetic networks. We also assess its relevance in biological applications, and find it to be comparable to that of the Colless and Sackin indices. Altogether, our results call for a reevaluation of the status of this somewhat forgotten measure of phylogenetic balance.

Contents

1 Introduction

1.1 Biological context

Whether it is to compare them, to perform simple statistical tests or to identify general trends or patterns, it is often useful for biologists to study trees through the lens of one or more summary statistics. In phylogenetics, many of the most prominent summary statistics aim at capturing the same intuitive idea: that some trees look more "symmetric" than others. These statistics are collectively known as *balance indices*.

Among the multitude of balance indices that have been proposed over the years (see e.g. [18, Chapter 33]), two stand out by their historical importance – and are still by far the most widely used today: the Colless index and the Sackin index. The Colless index, introduced by Colless in [14], is specific to rooted binary trees. It is defined as

$$\mathrm{Colless}(T) \;=\; \sum_{i\in I} |\lambda_1(i) - \lambda_2(i)|\,, \tag{1}$$

where the sum runs over the internal vertices of T and $\{\lambda_1(i), \lambda_2(i)\}$ is the number of leaves in each of the two subtrees descended from i. The Sackin index – which, contrary to what its name suggests, was introduced by Shao and Sokal in [41] – is defined for any rooted tree by the formula

$$\mathrm{Sackin}(T) \;=\; \sum_{\ell\in L} \delta_\ell\,, \tag{2}$$

where the sum runs over the leaves of T and δ_ℓ denotes the depth of ℓ (i.e. the number of edges of the path joining it to the root). Note that although we refer to them as balance indices, the Colless and Sackin indices actually measure the *im*balance of a phylogeny: the higher they are, the less balanced the phylogeny.

One of the problems of the Colless and Sackin indices, which was the starting point of this work, is that there is no single, natural way to extend their definition to phylogenetic networks. Although hardly a concern until recently, this is bound to become a major issue as the mounting evidence of the major roles played by phenomena such as gene transfers and hybridization forces biologists to abandon trees in favor of networks [5, 24].

While it is relatively easy to come up with network statistics that reduce to the Colless / Sackin index in trees, it is hard to favor one over the other – or even to see why they should conform with our intuition of what "balance" is. This led us to use a different approach and consider a lesser-known – and to some extent forgotten – measure of balance known as Shao and Sokal's B_2 index. Surprisingly given its very natural interpretation and the abundance of mathematical papers studying the properties of other balance indices [7, 9, 12, 15, 16, 36, 37, to name but a few], it seems that the mathematical properties of this balance index have never been studied before. Meanwhile, the current consensus in the biological literature seems to be that B_2 is not as useful as other balance indices; but on closer inspection this idea can mostly be traced to a single study [3]. Moreover, the specific assumptions made in that study limit the scope of its conclusions.

The aim of this paper is to fill in the current gap in the literature around B_2, in particular concerning its mathematical properties. Our main contribution is therefore a series of propositions and theorems about B_2, but we also include a statistical analysis that strongly suggests that the biological relevance of this index may have been underestimated when compared to that of the Colless and Sackin indices – and thus calls for more empirical work on the subject.

1.2 Definition and basic properties of B_2

In this section, we recall the intuition behind Shao and Sokal's B_2 index and give its formal definition in the context of phylogenetic networks. We then list some of its elementary properties. For this, we first need to introduce some vocabulary.

Definition 1.1. A *rooted phylogeny* is a directed acyclic graph that has exactly one vertex with no incoming edges. This vertex is called the *root* of the phylogeny, and the vertices with no outgoing edges are called its *leaves*. ◇

An intuitive idea in order to measure the "balance" of a rooted phylogeny is to send water from its root, then let that water trickle down the edges and accumulate in the leaves: the more evenly the water ends up being distributed among the leaves, the more balanced the phylogeny.

In order to formalize this idea, we consider a simple forward random walk started from the root – that is, at each step we follow one of the outgoing edges of the current vertex, uniformly at random, until we get trapped in a leaf. In a finite phylogeny, the stationary distribution of this random walk is a probability distribution $(p_\ell)_{\ell \in L}$ on the leaf-set L of the phylogeny. To quantify the uniformity of this distribution, we compute its Shannon entropy. This gives us the following definition, which is due to Shao and Sokal [41].

Definition 1.2. The B_2 index of a finite rooted phylogeny N is defined as

$$B_2(N) = -\sum_{\ell \in L} p_\ell \log_2 p_\ell \,,$$

where the sum runs over the leaves and p_ℓ is the probability that the simple forward random walk started from the root ends in ℓ. ◇

Remark 1.3. Using base-2 logarithms in this definition is more of a convention than a mathematical necessity. However, we will see that this turns out to be convenient when working with binary trees, which are prominent in biology. ◇

Note that, although technically valid, Definition 1.2 is not relevant in the case of infinite rooted phylogenies. Indeed, the random walk can then follow infinite paths without ever reaching a leaf. As a result, the parts of the phylogeny that do not subtend any leaf are not accounted for by this definition (think of the infinite binary tree, whose B_2 index would be 0).

Although this may not seem relevant for biological applications, from a mathematical point of view it is useful to define B_2 for infinite phylogenies (for instance, to give a meaning to its expected value under models that can produce infinite phylogenies, such as Galton–Watson trees; or to simplify the study of its limiting behaviour in large phylogenies). We do so in the context of locally finite phylogenies, i.e. phylogenies where every vertex has a finite degree.

Definition 1.4. The B_2 index of a locally finite rooted phylogeny N is defined as

$$B_2(N) = \lim_{k\to\infty} B_2(N_{[k]})\,,$$

where $N_{[k]}$ denotes the ball of radius k centered at the root in N, that is, the subgraph of N induced by the vertices whose distance to the root is at most k. ◇

In particular, Definition 1.4 ensures that if (N_i) is a sequence of rooted phylogenies that converges in distribution to N (in the local topology – see e.g. [17]), then $B_2(N_i)$ converges in distribution to $B_2(N)$.

Example 1.5. Let $\mathrm{CB}(h)$ be the complete binary tree with height h, i.e. the fully symmetric binary tree with $n = 2^h$ leaves. Then,

$$B_2(\mathrm{CB}(h)) \;=\; \log_2 n \;=\; h\,. \qquad \diamond$$

Example 1.6. Let $\mathrm{Cat}(n)$ be the caterpillar with n leaves (sometimes also known as the comb), i.e. the rooted binary tree with n leaves where every internal node has at least one child that is a leaf. Then,

$$B_2(\mathrm{Cat}(n)) \;=\; 2 - 2^{-n+2}\,. \qquad \diamond$$

Before closing this section and listing our main results, let us already point out some properties of B_2 that follow immediately from its definition.

Proposition 1.7. *Let N be a finite rooted phylogeny with n leaves. Then,*

$$0 \;\leqslant\; B_2(N) \;\leqslant\; \log_2 n\,.$$

Proof. This follows immediately from the definition of B_2 as a Shannon entropy. These bounds are tight, but they can be slightly improved when considering more restricted classes of phylogenies. This is discussed in Section 2. □

Proposition 1.8. *If T is a rooted binary tree, then letting δ_ℓ denote the depth of ℓ (i.e. the number of edges of the path joining it to the root) we have*

$$B_2(T) \;=\; \sum_{\ell\in L} \delta_\ell\, 2^{-\delta_\ell}\,.$$

Remark 1.9. This was already pointed out by Shao and Sokal in [41], and in fact in subsequent works using B_2 this expression is almost invariably used as its definition, without reference to its probabilistic interpretation. ◇

Proof. To obtain this expression from Definition 1.2, it suffices to note that, in a binary tree, the random walk has exactly δ_ℓ "left / right" decisions to make in order to get to ℓ. As a result, $p_\ell = 2^{-\delta_\ell}$ and the proposition follows. □

The next propositions are simple observations which we state as formal propositions to avoid having to re-detail them several times. We group them here because they are of constant use throughout the various sections of this document, and because we think they give some useful intuition about B_2. Readers who are less interested in the technical details may skip the rest of this section and go directly to Section 1.3, where we outline our main results.

Proposition 1.10. *Let N and N' be two rooted phylogenies, and let N'' be the rooted phylogeny obtained by grafting N' on a leaf $\ell^* \in N$, i.e. by making the vertices of N that point to ℓ^* point to the root of N' instead. Then,*

$$B_2(N'') \;=\; B_2(N) \;+\; p_{\ell^*} B_2(N')\,,$$

where p_{ℓ^*} denotes the probability of reaching ℓ^* in N.

Proof. Let L and L' be the respective leaf-sets of N and N', and let p_ℓ and p'_ℓ denote the probability of reaching a leaf ℓ in each of these phylogenies. Then,

$$\begin{aligned} B_2(N'') \;&=\; -\sum_{\substack{\ell\in L\\ \ell\neq\ell^*}} p_\ell \log_2 p_\ell \;-\sum_{\ell\in L'} p_{\ell^*}p'_\ell \log_2(p_{\ell^*}p'_\ell) \\ &=\; -\sum_{\ell\in L} p_\ell \log_2 p_\ell + p_{\ell^*}\log_2 p_{\ell^*} - p_{\ell^*}\sum_{\ell\in L'} p'_\ell \log_2 p'_\ell - p_{\ell^*}\log_2 p_{\ell^*}\sum_{\ell\in L'} p'_\ell \\ &=\; B_2(N) \;+\; p_{\ell^*} B_2(N')\,. \end{aligned}$$

□

Let us point out two particularly useful consequences of Proposition 1.10.

Corollary 1.11. *Let N^* be the rooted phylogeny obtained by grafting a cherry (that is, two leaves with the same parent) on a leaf ℓ of a rooted phylogeny N. Then, $B_2(N^*) = B_2(N) + p_\ell$.*

Corollary 1.12. *Let N' and N'' be two rooted phylogenies, and let $N = N' \oplus N''$ be the rooted phylogeny obtained by creating a new root and making it point to the roots of N' and N''. Then,*

$$B_2(N) \;=\; \frac{1}{2}\Big(B_2(N') + B_2(N'')\Big) \;+\; 1\,.$$

Proof. Use Proposition 1.10 twice to graft N' and N'' on the leaves of the rooted binary tree with two leaves. □

Proposition 1.13. *Let N and N' be two rooted phylogenies on the same leaf-set such that the probabilities p_ℓ and p'_ℓ of reaching ℓ are the same in N and in N' for every leaf ℓ, with the possible exception of two fixed leaves ℓ_1 and ℓ_2. Then,*

$$\operatorname{sgn}\Big(B_2(N') - B_2(N)\Big) \;=\; \operatorname{sgn}\Big((p_{\ell_1} - p'_{\ell_1})(p_{\ell_1} - p'_{\ell_2})\Big)\,,$$

where $\operatorname{sgn}(x) \in \{-1, 0, +1\}$ denotes the sign of x.

Remark 1.14. Perhaps a more intuitive way to understand Proposition 1.13 is to note that $(p_{\ell_1} - p'_{\ell_1})(p_{\ell_1} - p'_{\ell_2})$ has the same sign as $|p_{\ell_1} - p_{\ell_2}| - |p'_{\ell_1} - p'_{\ell_2}|$. Thus, $B_2(N') < B_2(N)$ if and only if p'_{ℓ_1} and p'_{ℓ_2} are more spread out than p_{ℓ_1} and p_{ℓ_2}. $\diamond$

Proof. Letting $f\colon x \mapsto -x \log x$ and $\Delta = p'_{\ell_1} - p_{\ell_1} = p_{\ell_2} - p'_{\ell_2}$, we have

$$\begin{aligned} B_2(N') - B_2(N) &= f(p'_{\ell_1}) + f(p'_{\ell_2}) - f(p_{\ell_1}) - f(p_{\ell_2}) \\ &= \big(f(p_{\ell_1} + \Delta) - f(p_{\ell_1})\big) - \big(f(p'_{\ell_2} + \Delta) - f(p'_{\ell_2})\big), \end{aligned}$$

and the proposition follows from the strict concavity of f (recall that f is strictly concave if and only if $(x, y) \mapsto (f(x) - f(y))/(x - y)$ is decreasing in x and in y). $\square$

1.3 Main results

In Section 2, we study the range of B_2 over several classes of rooted phylogenies. This basic information is particularly relevant if one wants to compare the B_2 index of phylogenies that have a different number of leaves, or belong to different classes (e.g, comparing reticulated and non-reticulated phylogenies). We show in Theorem 2.7 that for every temporal tree-child network N with n leaves (and in particular for every binary tree),

$$2 - 2^{-n+2} \leqslant B_2(n) \leqslant \lfloor \log_2 n \rfloor + \frac{n - 2^{\lfloor \log_2 n \rfloor}}{2^{\lfloor \log_2 n \rfloor}}.$$

Moreover, in the special case of binary trees, we fully characterize the trees that attain these bounds. Notably, the only binary tree that minimizes B_2 is the caterpillar tree – in agreement with the conventional idea that the caterpillar tree should be the least balanced tree.

Although the range of B_2 on binary trees is more narrow than that of other balances indices, such as the Colless and the Sackin indices (whose range length is asymptotically $n^2/2$; see [15, 19]), this should not give the impression that B_2 is a "coarser" measure of balance. In fact, it is exactly the opposite: we show in Proposition A.2.2 of Appendix A.2 that B_2 takes at least $2^{\lfloor n/2 \rfloor - 1}$ distinct values on the set of binary trees with n leaves – whereas the Colless and Sackin index, being integer-valued, can take at most $\Theta(n^2)$ different values. As a result, the average number of trees of size n that have the same B_2 index is exponentially smaller than the average number of trees of size n that have the same Colless / Sackin index, meaning that B_2 is better able to discriminate between trees.

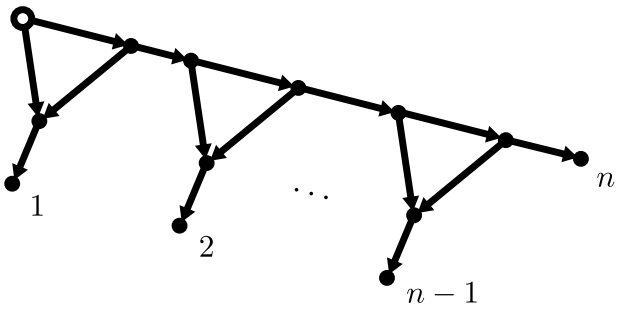


Figure 1: The fat caterpillar with n leaves, $\mathrm{FCat}(n)$. $\mathrm{FCat}(n+1)$ is obtained by grafting $\mathrm{FCat}(2)$ on the n-th leaf of $\mathrm{FCat}(n)$.

To close Section 2, we show in Theorem 2.8 that the minimum of B_2 on the set of tree-child network is $(\frac{8}{3} - \log_2(3))(1 - 4^{-n+1})$, and that the only tree-child network that attains it is the so-called "fat caterpillar", represented in Figure 1. This result is a relatively straightforward consequence of a series of lemmas of independent interest that characterize the effect of various local modifications of a rooted phylogeny on its B_2 index.

Section 3 is devoted to the study of the mean and variance of B_2 for general families of random trees. Theorem 3.1 gives an explicit, simple expression for the expected value of the B_2 index of a time-inhomogeneous Galton–Watson tree $T_{[k]}$ stopped after k generations. In the time-homogeneous case, this expression reduces to

$$\mathbb{E}\Big(B_2(T_{[k]})\Big) \;=\; \frac{\eta}{1-\alpha}\Big(1-\alpha^k\Big)\,,$$

with $\alpha = \mathbb{P}(Y > 0)$ and $\eta = \mathbb{E}(\log_2(Y)\,\mathbb{1}_{\{Y>0\}})$, where Y is the offspring distribution and α should be less than 1. In particular, for Galton–Watson trees with $2\,\mathrm{Bernoulli}(p)$ offspring distribution (that is, two offspring with probability p and 0 with probability $1-p$), which are the most relevant Galton–Watson trees in phylogenetics, this gives

$$\mathbb{E}\Big(B_2(T_{[k]})\Big) \;=\; \frac{p}{1-p}\Big(1-p^k\Big)\,.$$

For those Galton–Watson trees it is also possible to get an explicit expression for the variance of B_2 (see e.g. Proposition 3.3).

In the second part of Section 3, we study Markov branching trees. Theorem 3.4 gives recurrence relations to study the mean and variance of B_2 under any Markov branching model. Applying these allows us to show in Theorems 3.7 and 3.9 that under the ERM / Yule model with n leaves,

$$\mathbb{E}(B_2(T_n^{\mathsf{ERM}})) \;=\; \sum_{k=1}^{n-1}\frac{1}{k} \quad\text{and}\quad \mathrm{Var}(B_2(T_n^{\mathsf{ERM}})) \;\xrightarrow[n\to\infty]{}\; 2-\pi^2/6$$

and that under the PDA model with n leaves,

$$\mathbb{E}(B_2(T_n^{\mathsf{PDA}})) \;=\; 3\,\frac{n-1}{n+1} \quad\text{and}\quad \mathrm{Var}(B_2(T_n^{\mathsf{PDA}})) \;\xrightarrow[n\to\infty]{}\; \frac{4}{9}\,.$$

In particular, since $\mathbb{E}(B_2(T_n^{\mathsf{ERM}})) \sim \ln n$ and $\mathbb{E}(B_2(T_n^{\mathsf{PDA}})) \sim 3$, this shows that the ERM model generates trees that are very balanced whereas the PDA model generates trees that are very unbalanced – in agreement with what we obtain using the Colless and Sackin indices [9].

To complement this theoretical study and to evaluate the relevance of B_2 in real-world applications, in Section 4 we estimate the "statistical power" of B_2 when it comes to distinguishing various types of trees from one another. We then compare it to that of other balance indices – following in fact the very approach that led Agapow and Purvis to dismiss B_2 as a relevant measure of phylogenetic balance [3]. Our conclusions, however, are very different: our analysis shows that the performance of each balance index varies widely depending on the specific context in which it is used, and none of the balance indices that we test stands out as consistently better than the others. Neither did B_2 perform significantly worse

than other indices: in fact, averaged over all the scenarios that we consider (which were selected independently of the output of our analysis), B_2 happens to be the index with the best overall performance.

Our analysis also includes a comparison of the statistical power of pairs of balance indices. This comparison strongly supports the idea that B_2 better complements the Colless and Sackin indices than they complement each other. In fact, it even suggests that, taken jointly, the Colless and Sackin indices might the least informative pair of balance statistics – presumably because of their strong correlation.

A synthetic comparison of B_2 to other balance indices and a discussion of its current status in phylogenetics are given in Section 5.

2 Extremal values of B_2

We have seen in Proposition 1.7 that, for any rooted phylogeny, B_2 is non-negative and at most $\log_2 n$, where n is the number of leaves of the phylogeny. Moreover, if we consider the whole class of rooted phylogenies, then these bounds cannot be improved, as shown by the phylogenies depicted in Figure 2. But what about more restricted classes of rooted phylogenies? In this section, we answer this question for biologically relevant classes of rooted phylogenies: binary trees; temporal tree-child networks; and general tree-child networks.

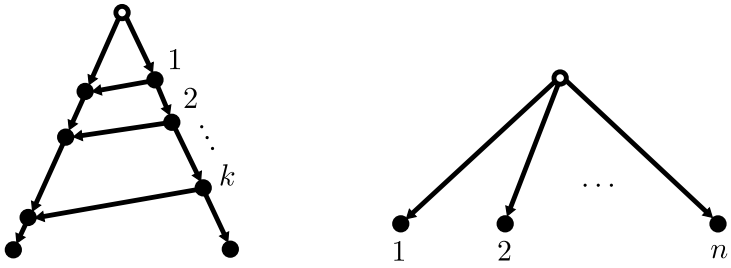


Figure 2: Examples of phylogenies showing that the bounds $0 \leqslant B_2 \leqslant \log_2 n$ are tight: the phylogeny on the left has $B_2 \to 0$ as $k \to +\infty$ and the one on the left has $B_2 = \log_2 n$. Note that for $n = 1$, we always have $B_2 = 0$.

Tree-child networks are a class of rooted phylogenies that were introduced in [11]. As of today, they are arguably the most widely studied class of phylogenetic networks. Let us briefly recall their definition.

Definition 2.1. A rooted phylogeny is said to be *binary* if the root has outdegree 2 and every other internal vertex has either:

- indegree 1 and outdegree 2, in which case it is called a *tree vertex*;
- indegree 2 and outdegree 1, in which case it is called a *reticulation*. ◇

Definition 2.2. A (binary) *tree-child network* is a binary rooted phylogeny such that every internal vertex has at least one child that is a tree vertex or a leaf. ◇

Before studying the range of B_2 over the class of tree-child networks, let us focus on specific subclasses that are both particularly relevant from a biological point of view and easier to tackle from a mathematical one.

2.1 Binary trees and temporal tree-child networks

Let us start with the simple case of rooted binary trees.

Theorem 2.3. *Let T be a rooted binary tree with n leaves. Then,*

$$2 - 2^{-n+2} \leqslant B_2(T) \leqslant \lfloor \log_2 n \rfloor + \frac{n - 2^{\lfloor \log_2 n \rfloor}}{2^{\lfloor \log_2 n \rfloor}}.$$

Moreover, these bounds are sharp and

(i) *The caterpillar tree* $\mathrm{Cat}(n)$ *is the only rooted binary tree with n leaves that minimizes B_2.*

(ii) *The rooted binary trees that maximize B_2 are exactly the trees such that* $\max|\delta_\ell - \delta_{\ell'}| \leqslant 1$*, where the maximum is taken over every pair of leaves and δ_ℓ denotes the depth of leaf ℓ.*

Remark 2.4. The caterpillar tree is also the only binary tree that maximizes the Sackin index and the only binary tree that maximizes the Colless index.

The binary trees that maximize B_2 are also exactly the binary trees that minimize the Sackin index, see [19]. A binary tree that minimizes the Colless index also maximizes the B_2 index, but the converse is not true: there are binary trees that maximize B_2 but do not minimize the Colless index. See [15] for more on this. ⋄

Proof of Theorem 2.3. Let T be a rooted binary tree. Consider a cherry of T with parent v (that is, both children of v are leaves) and a leaf $\ell \in T$ that is not a child of v. Then, by Corollary 1.11, moving the cherry from v to ℓ yields a binary tree T' such that

$$B_2(T') - B_2(T) = 2^{-\delta_\ell} - 2^{-\delta_v}.$$

Since it is possible to turn any binary tree into any other binary tree by repeatedly moving cherries, it follows that:

(i) T minimizes B_2 if and only if it does not have a cherry with parent v and a leaf ℓ not in that cherry such that $\delta_\ell > \delta_v$. The caterpillar is the only such binary tree.

(ii) T maximizes B_2 if and only if it does not have a cherry with parent v and leaf ℓ such that $\delta_\ell < \delta_v$, i.e. if and only if the maximum difference of depth between any two leaves is at most 1.

Finally, to compute B_2 for the trees that maximize it, note that if n is not a power of 2 then these trees are obtained by grafting a cherry on $k = n - 2^{\lfloor \log_2 n \rfloor}$ of the leaves of the complete binary tree with $2^{\lfloor \log_2 n \rfloor}$ leaves. The upper bound then follows from Corollary 1.11. □

Let us now turn to temporal tree-child networks. A temporal phylogeny is a phylogeny that is constrained to be compatible with the output of a time-embedded evolutionary process. This idea is formalized as follows.

Definition 2.5. A rooted binary phylogeny is *temporal* if there exists a function t on its vertex set such that, for every edge $\vec{uv}$, if v is a reticulation then $t(u) = t(v)$; otherwise, $t(u) < t(v)$. This function t is then known as a *temporal labeling*. ⋄

Remark 2.6. An alternative, perhaps more intuitive way to define temporal tree-child networks is through the notion of *ranked tree-child network*, or RTCNs [6]. As explained in Figure 3, RTCNs are the phylogenies generated by sequentially grafting cherries on leaves and tridents on pairs of leaves, keeping track of the step of the construction at which each internal vertex was added and making it an integral part of the resulting object. Discarding this information and keeping only the underlying rooted phylogeny always yields a temporal tree-child network. Moreover, every temporal tree-child network is the underlying rooted phylogeny of some RTCN. As a result, temporal tree-child networks and RTCNs are interchangeable for most purposes, and one can think about them in terms of the diagrams represented in Figure 3. ◇

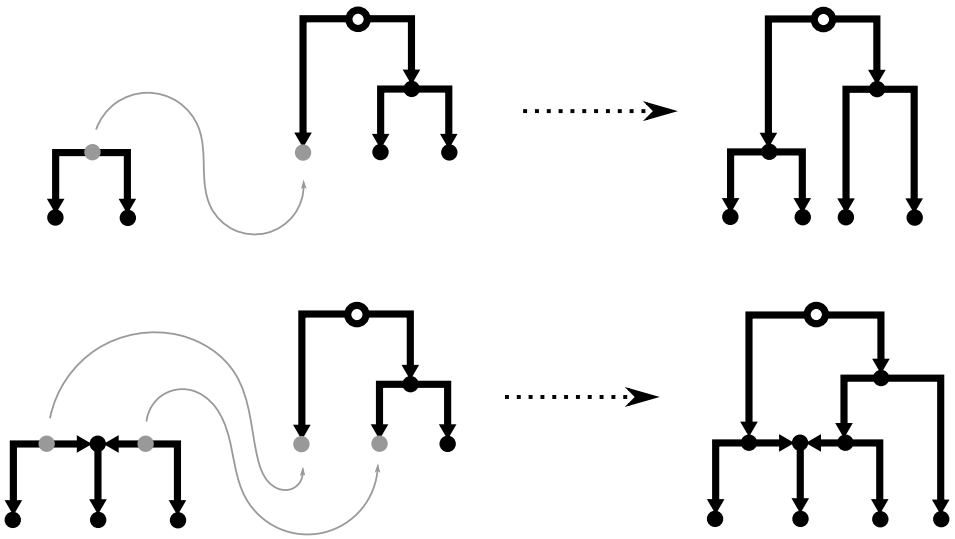


Figure 3: The operations of grafting a cherry (top) and grafting a trident (bottom). Sequentially performing these operations and keeping track of all the information of the construction, as done here through the vertical layout of the vertices, yields RTCNs. Temporal tree-child networks are obtained by discarding this vertical layout.

As it turns out, the range of B_2 is the same in binary trees and in temporal tree-child networks, as the next theorem shows. The difference with binary trees, however, is that the caterpillar (resp. the binary trees such that $\max|\delta_\ell - \delta_{\ell'}| \leqslant 1$) are not the only temporal tree-child networks that minimize (resp. maximize) B_2, as shown by the examples given in Figure 4.

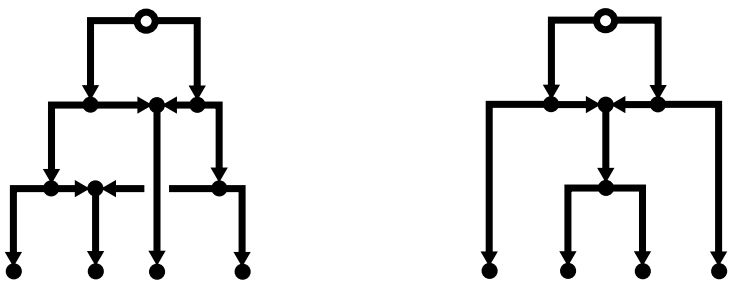


Figure 4: Examples of temporal tree-child networks that minimize (left) or maximize (right) B_2 despite being different from the trees described in Theorem 2.3.

Theorem 2.7. *For every temporal tree-child network with n leaves,*

$$2 - 2^{-n+2} \leqslant B_2(T) \leqslant \lfloor \log_2 n \rfloor + \frac{n - 2^{\lfloor \log_2 n \rfloor}}{2^{\lfloor \log_2 n \rfloor}}.$$

Proof. We will show that for any temporal tree-child network N it is possible to find two binary trees T' and T'' with the same number of leaves as N and such that $B_2(T') \leqslant B_2(N) \leqslant B_2(T'')$.

Assume that N is not a tree, and let then r be a reticulation with maximal temporal labeling. Denote the siblings of r (that is, the two other children of each of its two parents) by u and v, and its child by w. Note that, since no reticulation has a greater temporal labeling than r, the phylogenies N_u, N_v and N_w subtended by u, v and w do not contain any reticulations – and therefore are disjoint. This situation is represented in Figure 5. Let us show that it is possible, by removing r, to obtain two temporal tree-child networks N' and N'' such that $B_2(N') \leqslant B_2(N) \leqslant B_2(N'')$.

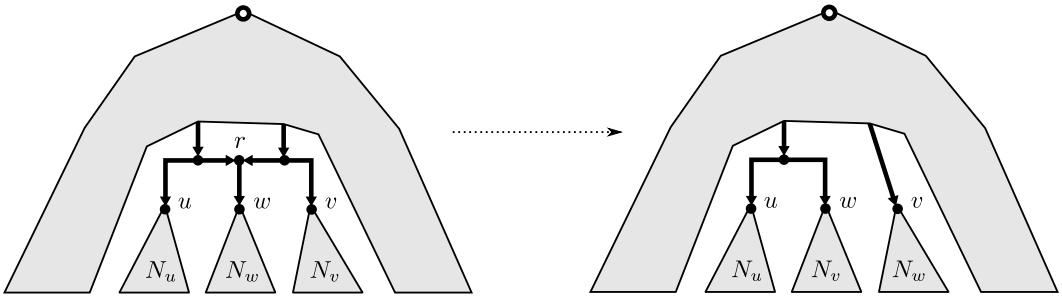


Figure 5: Left, the configuration described in the proof of Theorem 2.7, where the (possibly empty) sub-phylogenies N_u, N_v and N_w are cut from each other and from the rest of the phylogeny. Right, the result of the transformation that consists in removing the reticulation and swapping N_v and N_w.

Let us start with N'. Let p_u, p_v and $p_w = p_u + p_v$ be the probabilities that the simple forward random walk goes through u, v and w, respectively, and assume that $p_u \leqslant p_v$. Also assume without loss of generality that $B_2(N_w) \leqslant B_2(N_v)$. Indeed, if that is not the case, then swap N_v and N_w, and let $\tilde{N}$ be the resulting temporal tree-child network. By Proposition 1.10,

$$B_2(\tilde{N}) - B_2(N) = (p_w - p_v)(B_2(N_v) - B_2(N_w)) < 0,$$

and we can carry on with $\tilde{N}$ instead of N. Now, remove the edge going from the parent of v to the parent of w, and merge v and its parent; then swap N_v and N_w, and let N' be the resulting temporal tree-child network. This transformation is depicted in Figure 5. It can be seen as a succession of three steps: ungrafting N_v and N_w; removing the reticulation; and regrafting N_v and N_w. As a result,

$$B_2(N') - B_2(N) = \Delta_{\text{ungraft}} + \Delta_{\text{remove}} + \Delta_{\text{regraft}},$$

where, by Proposition 1.10,

$$\Delta_{\text{ungraft}} + \Delta_{\text{regraft}} = (p_w - 2p_v)(B_2(N_v) - B_2(N_w)) \leqslant 0$$

and, by Proposition 1.13, $\Delta_{\text{remove}} \leqslant 0$ since $(p_w - p_v)(p_w - 2p_v) \leqslant 0$.

To obtain N'', we use the same transformation but swapping the roles of u and v. Still with $p_u \leqslant p_v$, this time we can assume that $B_2(N_w) \leqslant B_2(N_u)$ before the transformation – since, otherwise, swapping N_u and N_w would increase B_2. Replacing v by u in the expressions above, we see that this time we get a temporal tree-child network N'' such that $B_2(N'') \geqslant B_2(N)$.

Finally, to obtain two binary trees T' and T'' such that $B_2(T') \leqslant B_2(N) \leqslant B_2(T'')$, it suffices to apply the transformation described above repeatedly, until all reticulations have been removed. This concludes the proof. □

Let us now see what happens when we relax the temporal constraint and consider the whole class of tree-child networks.

2.2 General tree-child networks

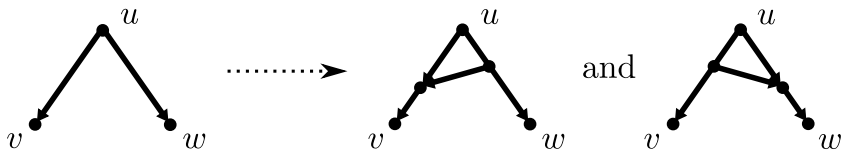
Theorem 2.8. *Let N be a tree-child network with n leaves. Then,*

$$B_2(N) \;\geqslant\; \left(\tfrac{8}{3} - \log_2(3)\right)\left(1 - 4^{-n+1}\right).$$

Moreover, this bound is sharp and the only tree-child network that minimizes B_2 is the so-called "fat caterpillar" represented in Figure 1.

The first ingredient in our proof of Theorem 2.8 is the following observation about rooted phylogenies.

Lemma 2.9. *Let N be a rooted phylogeny and let u, v and w be three vertices of N such that u is a parent of v and w, and neither of these two vertices is an ancestor of the other. Denote by N' and N'' the rooted phylogenies obtained by adding an edge between $\vec{uv}$ and $\vec{uw}$, in one direction for N' and in the other one for N'', as shown below.*



Then,

$$B_2(N') \;+\; B_2(N'') \;\leqslant\; 2\,B_2(N).$$

In particular, $\min\{B_2(N'), B_2(N'')\} \leqslant B_2(N)$. *Moreover, these inequalities are strict if and only if there exists a leaf ℓ such that $P_v(\ell) \neq P_w(\ell)$, where $P_x(\ell)$ denote the probability that the simple random walk started from x ends in ℓ.*

Remark 2.10. If N is a tree-child network, then for every vertex x there exists a leaf ℓ such that every internal vertex of the path from x to ℓ is a tree vertex. Letting v and w be the vertices of the statement of the proposition, there is thus always at least one leaf subtended by v that cannot be reached from w (and vice-versa). As a result, the inequalities of Lemma 2.9 are always strict. Note however that for N' and N'' to be tree-child networks, v and w should not be reticulations. ◇

Proof. Let N' be the phylogeny obtained by making the new edge point towards the incoming edge of v. For any vertex x and any leaf ℓ, let p_x denote the probability that the simple random walk started from the root of N goes through x and $Q(\ell)$

the probability that it ends in ℓ without going through v or w. Finally, let $P_x(\ell)$ denote the probability that the simple random walk started from x ends in ℓ. Since neither of v and w is an ancestor of the other, the random walk either goes through v, or it goes through w, or it avoids both of them. Therefore, the probability of reaching ℓ in N is

$$p_\ell \;=\; p_v P_v(\ell) \;+\; p_w P_w(\ell) \;+\; Q(\ell)\,.$$

Similarly, the probabilities of reaching ℓ in N' and in N'' are respectively

$$p'_\ell \;=\; (p_v + p_u/4) P_v(\ell) \;+\; (p_w - p_u/4) P_w(\ell) \;+\; Q(\ell)$$

and

$$p''_\ell \;=\; (p_v - p_u/4) P_v(\ell) \;+\; (p_w + p_u/4) P_w(\ell) \;+\; Q(\ell)\,.$$

As a result, we have

$$p'_\ell \;+\; p''_\ell \;=\; 2\,p_\ell\,,$$

and it follows from the strict concavity of $f : x \mapsto -x \log x$ that

$$f(p'_\ell) \;+\; f(p''_\ell) \;\leqslant\; 2\,f(p_\ell)\,,$$

where the inequality is strict if and only if $p'_\ell \neq p''_\ell$, i.e. if and only if $P_v(\ell) \neq P_w(\ell)$. Summing these inequalities over the leaves, we thus get

$$B_2(N') \;+\; B_2(N'') \;\leqslant\; 2\,B_2(N)\,,$$

with a strict inequality if and only there exist a leaf ℓ such that $P_v(\ell) \neq P_w(\ell)$. This concludes the proof. □

Before giving the second ingredient of our proof of Theorem 2.8, let us recall a standard fact about tree-child networks. We also recall its proof for the sake of completeness.

Lemma 2.11. *A tree-child network with n leaves has at most $n-1$ reticulations. If it does not have $n-1$ reticulations, then it has a vertex with two non-reticulation children.*

Proof. By the hand-shaking lemma, every rooted binary phylogeny satisfies

$$r + n - 1 = t\,,$$

where r is the number of reticulations, n the number of leaves and t the number of tree vertices, including the root. In the case of a tree-child network, since every reticulation has two tree-vertex-or-root parents that it does not share with any other reticulation, we also have

$$2\,r \;\leqslant\; t\,,$$

and it follows that $r \leqslant n - 1$.

Now, assume that a tree-child network has less that $n-1$ reticulations. It then has $t > 2\,r$ vertices with two children. Since a reticulation is shared by two of these t vertices, at least one of them has no reticulation child. □

The second ingredient in our proof of Theorem 2.8 is the following property, which is specific to tree-child networks.

Lemma 2.12. *Let N be a tree-child network. If N has a reticulation whose child is not a leaf, then there exists a tree-child network $N^\star$ with the same number of leaves as N and such that $B_2(N^\star) < B_2(N)$.*

Proof. Assume that r is a reticulation whose child u is a tree vertex. Let v and w be the children of u, and e and e' the two incoming edges of r. Finally, let N' and N'' be the tree-child networks obtained by removing r and u from N, and:

- in the case of N', making e point to v and e' point to w;
- in the case of N'', making e point to w and e' point to v.

This construction is illustrated in Figure 6. Note that N' and N'' are indeed tree-child networks, because N is a tree-child network and, therefore, neither the parents nor the siblings of r are reticulations.

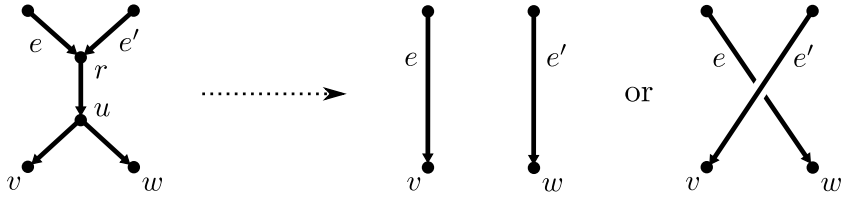


Figure 6: The transformation of N in to N' and N''

Now, let $X = (X_k)$ denote the simple random walk started from the root of N. Since N has no directed cycles, X visits u at most one time. When this happens, it goes through $u\vec{v}$ with probability $1/2$ and through $u\vec{w}$ with probability $1/2$. Thus, if we let $\tilde{X}$ be the random walk induced by X on $N \setminus \{r, u\}$; X' and X'' be simple random walks started from the root of N' and N'', respectively; and $Y \sim \text{Bernoulli}(1/2)$ be independent of (X', X''), then

$$\tilde{X} \;\sim\; YX' \,+\, (1-Y)X''\,.$$

In particular, the probabilities p_ℓ, p'_ℓ and p''_ℓ of reaching ℓ in N, N' and N'' satisfy

$$p_\ell \;=\; \frac{1}{2}\big(p'_\ell \,+\, p''_\ell\big)\,.$$

Therefore, by the same concavity argument as in the proof of Lemma 2.9,

$$B_2(N') \,+\, B_2(N'') \;\leqslant\; 2\,B_2(N)\,.$$

Note however that this time the inequality is not strict, because if the probability of going through e is the same as the probability of going through e', then $p'_\ell = p''_\ell$ for every leaf. In order to get a strict inequality, choose $\tilde{N} \in \{N', N''\}$ such that $B_2(\tilde{N}) \leqslant B_2(N)$, and note that $\tilde{N}$ has strictly less than $n-1$ reticulations, since it has one reticulation less than N. As a result, by Lemma 2.11 $\tilde{N}$ has at least one vertex with two non-reticulation children; and we can apply Lemma 2.9 to get a tree-child network $N^\star$ such that

$$B_2(N^\star) \;<\; B_2(\tilde{N}) \;\leqslant\; B_2(N)\,,$$

finishing the proof. □

With Lemmas 2.9 and 2.12, we can now prove Theorem 2.8 – i.e. show that the fat caterpillar is the only tree-child network that minimizes n.

Proof of Theorem 2.8. Let N be a tree-child network with n leaves that minimizes B_2. Then, N has $n-1$ reticulations (otherwise by Lemma 2.11 it would have a vertex with two non-reticulated children and we could use Lemma 2.9 to contradict its minimality). Moreover, by Lemma 2.12 the children of each of these reticulations are all leaves. As a result, the tree vertices of N are aligned on a single path, as represented in Figure 7.A (to see this, start from the root and follow the edges that point to tree vertices until a leaf is reached; since no reticulation subtends a tree vertex, the path thus obtained contains all tree vertices).

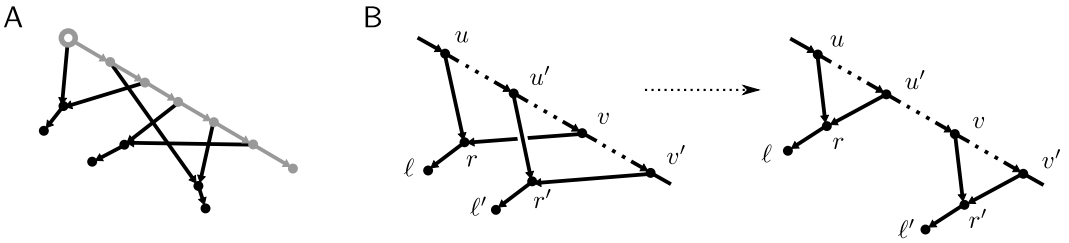


Figure 7: A, an example of a tree-child network that has $n-1$ reticulations for n leaves and where the child of every reticulation is a leaf (such tree-child networks are sometimes known as *one-component tree-child networks* [10]). The path containing all tree vertices is highlighted. B, the swapping of edges described in the proof, for one particular configuration of u, u', v and v' where u is an ancestor of v' and u' is an ancestor of v.

Now, assume that N has two reticulations r and r' with parents u, v and u', v', respectively, such that u is an ancestor of v' and u' is an ancestor of v. In that case, let N' be the tree-child network obtained by replacing the edges $v\vec{r}$ and $u'\vec{r'}$ by $u'\vec{r}$ and $v\vec{r'}$. Letting ℓ denote the child of r and ℓ' that of r', we have: in N, $(p_\ell, p_{\ell'}) = (p_u + p_v, p_{u'} + p_{v'})$; and, in N', $(p'_\ell, p'_{\ell'}) = (p_u + p_{u'}, p_v + p_{v'})$. Thus, by Proposition 1.13, $B_2(N') - B_2(N)$ has the same sign as

$$(p_v - p_{u'})(p_u - p_{v'}) \;<\; 0\,.$$

This shows that if N minimizes B_2 then it does not have two reticulations r and r' such that one parent of r is an ancestor of r' and one parent of r' is an ancestor of r – in other words, it is the fat caterpillar $\mathrm{FCat}(n)$.

Finally, the B_2 index of $\mathrm{FCat}(n)$ can be computed by noting that $B_2(\mathrm{FCat}(2)) = 2 - \frac{3}{4}\log_2(3)$ and that $\mathrm{FCat}(n+1)$ is obtained by grafting $\mathrm{FCat}(2)$ on the leaf with probability 4^{-n+1} of $\mathrm{FCat}(n)$. Using Proposition 1.10 and a little algebra then yields the lower bound in Theorem 2.8. □

3 Properties of B_2 in random trees

In this section, we study the mean and variance of B_2 for two of the most prominent families of random trees: Galton–Watson trees and Markov branching trees. To put the results on Markov branching trees in context, the reader is referred to Chapter I of Lucía Rotger's PhD dissertation [38], which contains a discussion of similar results for other balance indices.

3.1 Galton–Watson trees

We consider general Galton–Watson trees whose generations are indexed by $t \geqslant 0$, generation 0 being the root, and where the number of offspring of individual i from generation t is $Y_t(i) \sim Y_t$. The number Z_t of individuals in generation t is therefore given by $Z_0 = 1$ and

$$Z_{t+1} \;=\; \sum_{i=1}^{Z_t} Y_t(i)\,.$$

The natural filtration of the process is $\mathcal{F}_t = \sigma(Y_s(i) : i \geqslant 1, 0 \leqslant s \leqslant t-1)$.

Theorem 3.1. *Let T be a time-inhomogeneous Galton–Watson tree with offspring distribution Y_t in generation t, and let*

$$\alpha_t = \mathbb{P}(Y_t > 0) \quad \textit{and} \quad \eta_t = \mathbb{E}\Big(\log_2(Y_t)\,\mathbb{1}_{\{Y_t>0\}}\Big).$$

Then, letting $T_{[k]}$ denote the restriction of T to generations $t = 0, \dots, k$,

$$\mathbb{E}\Big(B_2(T_{[k]})\Big) \;=\; \sum_{t=0}^{k-1} \eta_t \prod_{s=0}^{t-1} \alpha_s\,.$$

In particular, if T is time-homogeneous, so that $\alpha_t = \alpha$ and $\eta_t = \eta$ for every t,

$$\mathbb{E}\Big(B_2(T_{[k]})\Big) \;=\; \begin{cases} \dfrac{\eta}{1-\alpha}\Big(1-\alpha^k\Big) & \text{if } \alpha < 1 \\ \eta k & \text{if } \alpha = 1. \end{cases}$$

Remark 3.2. Recall that, by Definition 1.4 of B_2 for infinite trees, we have $B_2(T) = \lim_k B_2(T_{[k]})$. In time-homogeneous Galton–Watson trees, there is thus a dichotomy between $\alpha < 1$, where $\mathbb{E}(B_2(T)) = \eta/(1-\alpha)$ is always finite, including in the supercritical case (where this implies that $\mathbb{E}(B_2(T) \mid |T| = +\infty)$ is finite); and $\alpha = 1$, where we always have $\mathbb{E}(B_2(T)) = +\infty$ (except in the degenerate case $Y = 1$ a.s., where T is an infinite path and $B_2(T) = 0$). $\diamond$

Proof. Let $P_t(i)$ be the probability that the random walk goes through individual i in generation t (note that this is a random variable). A bit of book-keeping shows that

$$B_2(T_{[t+1]}) \;=\; B_2(T_{[t]}) \;+\; \sum_{i=1}^{Z_t} P_t(i) \log_2\Big(Y_t(i)\Big)\,\mathbb{1}_{\{Y_t(i)>0\}}\,. \tag{3}$$

Moreover, since Z_t and $(P_t(i) : i \geqslant 1)$ are $\mathcal{F}_t$-measurable and $(Y_t(i) : i \geqslant 1)$ is independent of $\mathcal{F}_t$, with $Y_t(i) \sim Y_t$ for all i,

$$\mathbb{E}\left(\sum_{i=1}^{Z_t} P_t(i) \log_2\Big(Y_t(i)\Big)\,\mathbb{1}_{\{Y_t(i)>0\}} \,\middle|\, \mathcal{F}_t\right) \;=\; \mathbb{E}\Big(\log_2(Y_t)\,\mathbb{1}_{\{Y_t>0\}}\Big) \sum_{i=1}^{Z_t} P_t(i)\,.$$

As a result, taking expectations in (3) we get

$$\mathbb{E}\Big(B_2(T_{[t+1]})\Big) \;=\; \mathbb{E}\Big(B_2(T_{[t]})\Big) \;+\; \mathbb{E}\Big(\log_2(Y_t)\,\mathbb{1}_{\{Y_t>0\}}\Big)\,\mathbb{E}\left(\sum_{i=1}^{Z_t} P_t(i)\right).$$

Now, observe that $\sum_{i=1}^{Z_t} P_t(i)$ is the probability that the random walk reaches generation t, conditional on $\mathcal{F}_t$. Its expected value is therefore the total probability that the random walk reaches generation t, which is also the probability that it does not get trapped in a leaf for some generation $s < t$. As a result,

$$\mathbb{E}\left(\sum_{i=1}^{Z_t} P_t(i)\right) \;=\; \prod_{s=0}^{t-1} \mathbb{P}(Y_s > 0)\,.$$

Writing $\alpha_t = \mathbb{P}(Y_t > 0)$ and $\eta_t = \mathbb{E}\Big(\log_2(Y_t)\,\mathbb{1}_{\{Y_t>0\}}\Big)$, we therefore have

$$\mathbb{E}\Big(B_2(T_{[t+1]})\Big) \;=\; \mathbb{E}\Big(B_2(T_{[t]})\Big) \;+\; \eta_t \prod_{s=0}^{t-1} \alpha_s\,,$$

and the theorem follows from the fact that $\mathbb{E}(B_2(T_{[0]})) = 0$. □

To close this section, we give an expression for the variance of B_2 in binary Galton–Watson trees. In order to present a simple expression – and because this is what we need in the rest of this document – we focus on the critical case.

Proposition 3.3. *Let T be a critical binary Galton–Watson tree, that is, assume that the offspring distribution is $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 2) = 1/2$. Then,*

$$\mathrm{Var}\Big(B_2(T_{[k]})\Big) \;=\; \frac{4}{3} - 2^{-k+2} + 4^{-k}\left(k + \frac{8}{3}\right).$$

As a result, $\mathrm{Var}(B_2(T)) = 4/3$.

The proof of this proposition relies on standard calculations similar to that of Theorem 3.1. It can be found in Section A.1 of the Appendix.

3.2 Markov branching trees

Markov branching trees are a general class of random binary trees that were introduced by Aldous in [4]. They have since become prominent in phylogenetics, where they are mainly known through the β-splitting models, a one-parameter family of models that can generate a wide variety of random tree shapes. In particular, the β-splitting models include the ERM model (which generates trees that have the same shape as Yule trees), the PDA model (which generates uniform leaf-labeled rooted binary trees) and the AB model (which generates trees that resemble real-world phylogenies [8]).

A Markov branching tree is described by a family $\mathbf{q} = (q_n)_{n\geqslant 2}$ of probability distributions, known as the *root-split distributions*, such that q_n is symmetric on $\{1, \ldots, n-1\}$, that is, $q_n(k) = q_n(n-k)$. If we do not worry about labels, which are irrelevant for our purposes, a Markov branching tree T with n leaves can be generated as follows:

1. Sample a random variable K according to q_n.
2. Let the two subtrees of T be independent Markov branching trees with K and $n - K$ leaves, respectively.

For a complete introduction to Markov branching trees, see e.g. [29].

Theorem 3.4. *Let T_n be a Markov branching tree with n leaves and root-split distributions $\mathbf{q} = (q_n)$. Let $\mu_n = \mathbb{E}(B_2(T_n))$, $s_n = \mathbb{E}(B_2(T_n)^2)$ and $v_n = \mathrm{Var}(B_2(T_n))$. Then, letting $K_n \sim q_n$, we have the following recurrence relations:*

(i) $\mu_n = \mathbb{E}(\mu_{K_n}) + 1$

(ii) $s_n = \frac{1}{2}\,\mathbb{E}(s_{K_n}) + \frac{1}{2}\,\mathbb{E}(\mu_{K_n}\,\mu_{n-K_n}) + 2\,\mathbb{E}(\mu_{K_n}) + 1$

(iii) $v_n = \frac{1}{4}\,\mathrm{Var}(\mu_{K_n} + \mu_{n-K_n}) + \frac{1}{2}\,\mathbb{E}(v_{K_n})$

with the initial conditions $\mu_1 = s_1 = v_1 = 0$.

Proof. Let (T'_k) and (T''_k) be two independent sequences of Markov branching trees with root-split distributions $\mathbf{q}$ and such that, for all $k \geqslant 1$, T'_k and T''_k have k leaves, T'_1 and T''_1 being the tree that consists only of the root. Letting $T = T' \oplus T''$ denote the tree obtained by creating a new root and making it point to the roots of T' and T'', we thus have

$$T_n \;=\; T'_{K_n} \oplus T''_{n-K_n},$$

where $K_n \sim q_n$ is independent of (T'_k) and (T''_k). As a result, by Corollary 1.12,

$$B_2(T_n) \;=\; \frac{1}{2}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n})\Big) + 1. \tag{4}$$

Taking expectations and using that $K_n \sim n - K_n$ and that K_n is independent of (T'_k) and (T''_k), we get

$$\mathbb{E}(B_2(T_n)) \;=\; \mathbb{E}(B_2(T'_{K_n})) \;+\; 1,$$

which, since $\mathbb{E}(B_2(T'_{K_n}) \mid K_n = k) = \mu_k$ for all k, is point (i).

Point (ii) is proved similarly: from Equation (4), we get

$$\begin{aligned}\mathbb{E}\Big(B_2(T_n)^2\Big) \;&=\; \tfrac{1}{4}\mathbb{E}\Big(B_2(T'_{K_n})^2\Big) \;+\; \tfrac{1}{4}\mathbb{E}\Big(B_2(T''_{n-K_n})^2\Big) \;+\\ &\quad\; \tfrac{1}{2}\mathbb{E}\Big(B_2(T'_{K_n})B_2(T''_{n-K_n})\Big) \;+\; \mathbb{E}\Big(B_2(T'_{K_n})\Big) \;+\; \mathbb{E}\Big(B_2(T''_{n-K_n})\Big) \;+\; 1,\end{aligned}$$

which, by the symmetry of K_n and its independence from (T'_k) and (T''_k), gives

$$\mathbb{E}\Big(B_2(T_n)^2\Big) \;=\; \tfrac{1}{2}\mathbb{E}\Big(B_2(T'_{K_n})^2\Big) \;+\; \tfrac{1}{2}\mathbb{E}\Big(B_2(T'_{K_n})B_2(T''_{n-K_n})\Big) \;+\; 2\,\mathbb{E}\Big(B_2(T'_{K_n})\Big) \;+\; 1.$$

Writing these expectations as $\mathbb{E}(\mathbb{E}(\,\cdot \mid K_n))$ and using the notation from the statement of the theorem, this gives point (ii).

Finally, point (iii) is obtained by applying the law of total variance to Equation (4):

$$\begin{aligned}\mathrm{Var}(B_2(T_n)) \;&=\; \tfrac{1}{4}\,\mathrm{Var}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n})\Big)\\ &=\; \tfrac{1}{4}\,\mathrm{Var}\Big(\mathbb{E}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n}) \;\Big|\; K_n\Big)\Big)\\ &\quad\; + \tfrac{1}{4}\,\mathbb{E}\Big(\mathrm{Var}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n}) \;\Big|\; K_n\Big)\Big).\end{aligned}$$

In this last expression,

$$\mathbb{E}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n}) \;\Big|\; K_n\Big) \;=\; \mu_{K_n} \;+\; \mu_{n-K_n}\,.$$

Likewise, since $B_2(T'_{K_n})$ and $B_2(T''_{n-K_n})$ are independent conditional on K_n,

$$\begin{aligned}\mathrm{Var}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n}) \,\Big|\, K_n\Big) &= \mathrm{Var}\Big(B_2(T'_{K_n}) \,\Big|\, K_n\Big) + \mathrm{Var}\Big(B_2(T''_{n-K_n}) \,\Big|\, K_n\Big) \\ &= v_{K_n} + v_{n-K_n},\end{aligned}$$

and, taking expectations and using the symmetry of K_n, we get

$$\mathbb{E}\Big(\mathrm{Var}\Big(B_2(T'_{K_n}) + B_2(T''_{n-K_n}) \,\Big|\, K_n\Big)\Big) = 2\,\mathbb{E}(v_{K_n})\,.$$

Putting the pieces together, this yields the required expression for $\mathrm{Var}(B_2(T_n))$ and finishes the proof. □

Remark 3.5. Equation (4) is known as a random recursive equation. These can be studied with the so-called *contraction method*, which often makes it possible to characterize (the appropriately rescaled) limit of a random sequence as the solution of a distributional equation – see e.g. [35]. This has been done to study the limiting distribution of the Sackin and Colless indices under the ERM and PDA model [9]. Doing something similar with B_2 might be possible – although, as the simulations of Section 1.1 show, the situation will be more complex and there will be no central limit theorem. ◇

In the rest of this section, we use the recurrence relations of Theorem 3.4 to study the expected value and the variance of B_2 under what are undoubtedly the two most important models of random trees in mathematical phylogenetics: the ERM / Yule model and the PDA / uniform model.

Definition 3.6. The ERM model is the Markov branching model whose root-split distributions are given by

$$\forall k \in \{1, \dots, n-1\}, \quad q_n(k) = \frac{1}{n-1}\,. \qquad \diamond$$

What makes the ERM so central to mathematical biology is that it generates trees that have the same shape as the genealogical tree associated to the Yule process, i.e. the pure-birth process where every individual gives birth at constant rate 1. This is also the random tree shape associated to the Kingman coalescent (where every pair of lineages coalesces at rate 1; [26]) and to the genealogy of extant individuals in the Moran model (where at each step an ordered pair of individuals is sampled uniformly at random for the first one to be replaced by a copy of the second [33]). It corresponds to the uniform distribution on the set of ranked binary trees with n labeled leaves. This variety of constructions explains why this model arises in a wide range of biological applications as well as in many mathematical problems.

Theorem 3.7. *Let T_n be a tree with n leaves sampled under the ERM / Yule model.*

(i) $\mathbb{E}(B_2(T_n)) = \displaystyle\sum_{k=1}^{n-1} \frac{1}{k}$.

(ii) $\mathrm{Var}(B_2(T_n)) \xrightarrow[n\to\infty]{} 2 - \pi^2/6$.

Proof. One way to prove point (i) is to show that $\mu_n = \sum_{k=1}^{n-1} 1/k$ is indeed the solution of the recurrence for the expected value of B_2 given in Theorem 3.4.

However, a perhaps more intuitive way to obtain $\mathbb{E}(B_2(T_n))$ it is to note that, in the Yule model, T_n is obtained by grafting a cherry on a leaf $L \in T_{n-1}$, sampled uniformly and independently of the shape of T_{n-1}. Let P_ℓ denote the probability of reaching a fixed leaf $\ell \in T_{n-1}$ (note that this is a random variable). Then, by Corollary 1.11, $B_2(T_n) = B_2(T_{n-1}) + P_L$ and therefore

$$\mathbb{E}(B_2(T_n)) \;=\; \mathbb{E}(B_2(T_{n-1})) \;+\; \mathbb{E}(P_L)\,.$$

Since L is independent of (P_ℓ), $\mathbb{E}(P_L) = \mathbb{E}(p_L)$, where $p_\ell = \mathbb{E}(P_\ell) = 1/(n-1)$, by exchangeability. As a result,

$$\mathbb{E}(B_2(T_n)) \;=\; \mathbb{E}(B_2(T_{n-1})) \;+\; \frac{1}{n-1}\,,$$

and we use that $\mathbb{E}(B_2(T_1)) = 0$ to conclude.

Let us now turn to point (ii). Perhaps surprisingly given the simplicity of the derivation of the expected value of $B_2(T_n)$, we could not find a way to obtain the limit of its variance other than solving the recurrence relations of Theorem 3.4 explicitly.

Let $H_n = \sum_{k=1}^{n} 1/k$ denote the harmonic numbers, and $H_n^{(m)} = \sum_{k=1}^{n} 1/k^m$ the generalized harmonic numbers of order m. Letting $v_n = \mathrm{Var}(B_2(T_n))$, point (iii) of Theorem 3.4 can be written

$$v_n \;=\; \alpha_{n-1} \;+\; \frac{1}{2(n-1)} \sum_{k=1}^{n-1} v_k\,, \tag{5}$$

where $\alpha_{n-1} = \frac{1}{4}\,\mathrm{Var}(H_{K_n-1} + H_{n-1-K_n})$, with $K_n \sim \mathrm{Uniform}(\{1,\dots,n-1\})$. Note that this already shows that if v_n has a finite limit ℓ, then $\ell = 2\lim_n \alpha_n$. Indeed, in that case $\frac{1}{n-1}\sum_{k=1}^{n-1} v_k \to \ell$, by Cesàro's lemma, and therefore ℓ satisfies $\ell = \lim_n \alpha_n + \ell/2$.

Let us begin by computing α_{n-1} explicitly. First, since $K_n \sim n - K_n$,

$$\begin{aligned} &\mathrm{Var}(H_{K_n-1} + H_{n-K_n-1}) \\ =\;& 2\Big(\mathbb{E}(H_{K_n-1}^2) \;+\; \mathbb{E}(H_{K_n-1}\,H_{n-1-K_n}) \;-\; 2\,\mathbb{E}(H_{K_n-1})^2\Big). \end{aligned} \tag{6}$$

Moreover, the following identities for sums of harmonic numbers are well-known; see e.g. Section 1.2.7 of [28]:

- $\mathbb{E}(H_{K_n-1}^2) = H_{n-1}H_{n-2} + 2 - 2H_{n-1}$. [28, §1.2.7, Eq. (8)]
- $\mathbb{E}(H_{K_n-1})^2 = (H_{n-1} - 1)^2$. [28, §1.2.7, Exercise 15]
- $\mathbb{E}(H_{K_n-1}H_{n-1-K_n}) = 1 - H_{n-1}^{(2)} + (H_{n-1} - 1)^2$. [43, Theorem 1]

Plugging these in (6), after some simplifications we get

$$\alpha_{n-1} \;=\; \frac{1}{2}\Big(2 \;-\; H_{n-1}^{(2)} \;-\; \frac{H_{n-1}}{n-1}\Big)\,. \tag{7}$$

Let us now solve the recurrence (5) in order to get an explicit expression for v_n. We start by rearranging the terms in order to get a first-order recurrence:

$$\begin{aligned} v_{n+1} &= \alpha_n + \frac{1}{2n}\sum_{k=1}^{n} v_k \\ &= \alpha_n + \frac{v_n}{2n} + \frac{n-1}{n}\cdot\frac{1}{2(n-1)}\sum_{k=1}^{n-1} v_k \\ &= \alpha_n + \frac{v_n}{2n} + \frac{n-1}{n}\Big(v_n - \alpha_{n-1}\Big)\,. \end{aligned}$$

As a result, letting $\beta_n = \alpha_n - \alpha_{n-1}(n-1)/n$, we see that v_n is the solution of

$$v_{n+1} = \frac{2n-1}{2n}\, v_n + \beta_n\,,$$

with the initial condition $v_1 = 0$. Solving this first-order recurrence then yields

$$v_n = \frac{(2n-1)!!}{(2n)!!}\sum_{k=2}^{n-1}\beta_k\frac{(2k)!!}{(2k-1)!!}\,, \tag{8}$$

where $n!!$ denotes the double factorial. Finally, to see that $v_n \to c = 2 - \pi^2/6$, note that $\beta_k \sim c/(2k)$ and recall that $(2k)!!/(2k-1)!! \sim \sqrt{\pi k}$. The summands in the expression of v_n therefore are asymptotically equivalent to $(c\sqrt{\pi})/2\sqrt{k})$, and the result follows from a standard application of the integral test for convergence, since $\int \frac{1}{\sqrt{x}}\,dx = 2\sqrt{x}$. □

Definition 3.8. The PDA model is the Markov branching model whose root-split distributions are given by

$$\forall k \in \{1,\dots,n-1\}, \quad q_n(k) = \frac{1}{2}\binom{n}{k}\frac{t_k\, t_{n-k}}{t_n}\,,$$

where $t_n = (2n-3)!!$ is the number of rooted binary trees with n labeled leaves. ◇

What makes the PDA model stand out is that it generates trees that are uniformly distributed on the set of rooted binary trees with n labeled leaves. For this reason, it is sometimes referred to as the "uniform model" and, in phylogenetics, it is the standard alternative to the Yule model when in need of a null model.

Theorem 3.9. *Let T_n be a tree with n leaves sampled under the PDA model or, equivalently, uniformly on the set of rooted binary trees with n labeled leaves.*

(i) $\mathbb{E}(B_2(T_n)) = 3\,(n-1)/(n+1)$.

(ii) $\mathrm{Var}(B_2(T_n)) \xrightarrow[n\to\infty]{} 4/9$.

Proof. By Theorem 3.4, to prove (i) it suffices to show that $\mu_n = 3\,(n-1)/(n+1)$ is indeed the solution of $\mu_1 = 0$ and

$$\mu_n = 1 + \sum_{k=1}^{n-1}\mu_k\, q_n(k)\,, \tag{9}$$

where

$$q_n(k) = \frac{1}{2}\binom{n}{k}\frac{(2k-3)!!\,(2(n-k)-3)!!}{(2n-3)!!}\,.$$

Assume that $\mu_k = 3\,(k-1)/(k+1)$ for all $k \in \{1, \dots n-1\}$. Note that this can be written as $1 + \mu_k = 2(2k-1)/(k+1)$, so that

$$\Big(1 \ + \ \mu_k\Big)\, q_n(k) \ = \ 2\,\frac{2n-1}{n+1}\, q_{n+1}(k+1)\,.$$

Plugging this in Equation (9), we get

$$\begin{aligned}\mu_n \ = \ \sum_{k=1}^{n-1} \Big(1+\mu_k\Big)\, q_n(k) = \ 2\,\frac{2n-1}{n+1} \sum_{k=1}^{n-1} q_{n+1}(k+1) \\ = \ 2\,\frac{2n-1}{n+1}\,\Big(1 - q_{n+1}(1)\Big)\end{aligned}$$

which, since $q_{n+1}(1) = \frac{n+1}{2(2n-1)}$, yields $\mu_n = 3(n-1)/(n+1)$.

To prove point (ii), recall that, as $n \to \infty$, the uniform rooted binary tree with n labeled leaves converges in distribution to the size-biased Galton–Watson tree $\hat{T}$ obtained by grafting independent critical binary Galton–Watson trees on each leaf of the infinite caterpillar, as illustrated in Figure 8. See [25] for a complete introduction to the subject.

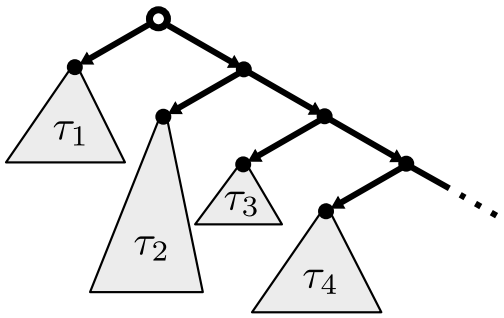


Figure 8: Construction of the size-biased Galton–Watson tree (for the critical binary Galton–Watson tree). Let the root be the end of a one-way infinite path known as the *spine* of the tree, and let every vertex from the spine point to the root of an independent Galton–Watson tree with 2 Bernoulli(2) offspring distribution.

Letting τ_k denote the Galton–Watson tree grafted on the leaf at depth k of the infinite caterpillar, by Proposition 1.10 we have

$$B_2(\hat{T}) \ = \ \sum_{k \geqslant 1} 2^{-k} B_2(\tau_k)\,.$$

Because the variables $B_2(\tau_k)$ are independent, this gives

$$\mathrm{Var}(B_2(\hat{T})) \ = \ \sum_{k \geqslant 1} 4^{-k}\, \mathrm{Var}(B_2(\tau_k))$$

and, since we have seen in Proposition 3.3 that $\mathrm{Var}(B_2(\tau_k)) = 4/3$,

$$\mathrm{Var}(B_2(\hat{T})) = \frac{4}{9}\,.$$

Since $\mathrm{Var}(B_2(T_n)) \to \mathrm{Var}(B_2(\hat{T}))$, this concludes the proof. □

Remark 3.10. Recalling that T_n has the same distribution as the Galton–Watson tree with $2\,\mathrm{Bernoulli}(p)$ offspring distribution conditioned on having n leaves, and that the probability that the $2\,\mathrm{Bernoulli}(p)$-Galton–Watson tree has n leaves is

$$\varpi_n \;=\; 2^{n-1}(2n-3)!!\,p^{n-1}(1-p)^n/n!\,,$$

point (i) from Theorem 3.9 can be used to give an alternative derivation of the expected value of B_2 in binary Galton–Watson trees (which we already have as a special case of Theorem 3.1). Indeed, it is readily checked that

$$\sum_{n\geqslant 1} 3\,\frac{n-1}{n+1}\varpi_n \;=\; \frac{p}{1-p}\,. \qquad \diamond$$

4 Biological relevance: empirical study

4.1 Design of the study

We follow the approach originally developed by Kirkpatrick and Slatkin (henceforth K&S) in [27] and later extended by Agapow and Purvis (A&P) in [3]. The idea is to compare how good are various balance indices at distinguishing between trees of different origins. For this, we pick a null model to generate random trees and use it to build empirical confidence intervals for each balance index. We then consider sets of trees that were not generated by this null model and, for each balance index, compute the percentage of those trees whose balance index does not fall in the confidence interval obtained under the null model. This percentage is used as a direct measure of the "power" of the balance index: the higher it is, the more efficient the balance index was at distinguishing the trees of the test set from those generated by the null model.

Although the general strategies are the same, our analysis differs from those of K&S and A&P in three important respects:

1. We use several, varied null models.
2. We use real-world phylogenies for our test sets.
3. We consider a wider range of tree sizes.

Null models: Both K&S and A&P use the ERM model as their null model. This is standard, but debatable: indeed, this model is well-known to produce trees that are far more balanced than real-world phylogenies [8] (in fact, A&P acknowledge this and consider one-sided confidence intervals as a result). Thus, both studies focus on rejecting a single, relatively unrealistic null hypothesis.

To address this issue, we consider 5 different null models which, together, cover a large variety of tree shapes: the ERM model; the PDA model; the Aldous branching model (AB); and two versions of a multiplicative fitness landscape birth process (MFL). In this last model, each lineage has it own, evolving birth rate. The birth rates do not spontaneously change over time, but when a birth occurs the birth rate of each of the two newly formed lineages is inherited from their mother and then multiplied by $(1+s)$ with probability p, independently of each

other and of everything else. We consider two variants MFL1 and MFL2 of this model, where the parameters are $s = 0.25$ and $p = 0.5$ for MFL1, and $s = 0.5$ and $p = 0.5$ for MFL2. The choice of these parameters is somewhat arbitrary: our goal is mainly to try less conventional null models. We did check that the trees generated by these models seem biologically relevant; in fact for the range of the number of leaves considered, they produce trees that are either slightly less (MFL1) or slightly more (MFL2) balanced than the trees produced by the AB model. However, in order not to risk biasing our study, we did not try to optimize the parameters s and p based on a specific balance index.

For each null model and each value of the number of leaves, two-sided (symmetric) 95% confidence intervals were built empirically using 10^5 replicates. The code for doing so and its output can be downloaded from [2].

Test sets: The test sets used by A&P consist of trees generated by various ad hoc models of random trees. This is ideal to isolate the effect of a specific biological mechanism on the performance of the balance indices. However, this also leaves the biological relevance of the trees open to debate.

Since we are more interested in assessing the potential relevance of various balance indices in real-world applications than in precisely characterizing the situations in which each of them should be used, we use real phylogenetic trees to constitute our test sets. Conceptually, these can be seen as samples from "black box" models of random trees. Thus, instead of having models whose inner workings we understand but whose relevance is questionable, we have models that we know nothing about but whose relevance is undeniable.

In total, we used trees from 4 databases: 4378 trees from TreeBASE [42]; 14509 trees from OrthoMaM [40]; 77843 trees from PhylomeDB [23]; and 85581 trees from HOGENOM [34]. More details on how these trees were obtained can be found in [2], where the complete list of trees that we used can also be downloaded in Newick format.

Range of the number of leaves: K&S consider trees with up to 50 leaves, and A&P with up to 64 leaves. While these were considered to be large trees at the time, this is not so much the case today. In fact, 12% of the phylogenetic trees that we collected have between 50 and 100 leaves; and 20% have more than 100 leaves. Moreover, larger trees are bound to become more and more common as sequencing data and reconstitution methods improve.

In order to see how the size of the trees affects the performance of each balance index, we grouped the trees of each dataset in categories based on their number of leaves: very small trees ($6 \leqslant n \leqslant 12$); small trees ($12 \leqslant n \leqslant 24$); intermediate trees ($25 \leqslant n \leqslant 50$); large trees ($50 \leqslant n \leqslant 100$); and very large trees ($n > 100$). Note that these size categories are not related to the estimation of the confidence intervals (which is done for every possible value of n): the point of these categories is to have enough trees of similar sizes to compute a "by-size" average of the proportion of trees rejected by each balance index. Also note that the slight overlap between some of the categories is not a problem (these categories can be thought of as playing the role of windows in a moving average).

In the case of the OrthoMaM database, only the *large* and *very large* categories were considered because the other categories did not contain enough trees to get a reliable estimate of the proportion of trees rejected by each balance index. All categories that were included in our analysis contain at least 500 trees – usually several thousands. Our estimation of the size-specific proportions of trees rejected by the balance indices should therefore be reasonably reliable.

Balance indices: The set of balance indices that we compare is very similar to that used by K&S and by A&P, as well as other studies that compare balance indices [21, 30, 31]. We consider: the Colless index; the Sackin index; the number of cherries, as suggested in [32]; σ^2_N, the variance of the depths of the leaves; the B_1 index; and the B_2 index. Let us briefly recall the definition of the balance indices that we have not yet mentioned in this paper.

The variance of the depths of the leaves was originally suggested as a measure of phylogenetic balance by Sackin in [39], but formally introduced and first used by K&S in [27]. It is defined as

$$\sigma^2_N(T) \;=\; \frac{1}{n}\sum_{\ell\in L}(\delta_\ell - \bar{\delta}(T))^2, \tag{10}$$

where the sum runs over the leaves of the tree; n is the number of leaves; δ_ℓ the depth of leaf ℓ; and $\bar{\delta}(T) = \frac{1}{n}\sum_\ell \delta_\ell$ the average leaf depth. Note that $\bar{\delta}(T)$ is a rescaled version of the Sackin index, and that various rescaled versions are frequently used instead of $\text{Sackin}(T) = \sum_\ell \delta_\ell$. Since we estimate our confidence intervals for each value of n, such scalings by a function of n are irrelevant.

The B_1 index was introduced by Shao and Sokal in [41] and is defined as

$$B_1(T) \;=\; \sum_{i\in I^*}\Big(\max\{\delta_\ell : \ell \text{ is subtended by } i\}\Big)^{-1}, \tag{11}$$

where the sum runs over internal vertices, excluding the root.

Summary: Altogether, this gives us 85 (null model; test set; size category) scenarios under which to compare 6 balance indices. The complete list of null models, test sets, size categories and balance indices that we use in our analysis is summarized in Table 1. For each null model, we have the confidence interval of each balance index; and for each test set and each size category, we have at least 500 trees. We estimate the proportion of trees rejected by each balance index in each specific scenario and use it as a direct measure of the "statistical power" of that index in that scenario. The results are presented in Section 4.2.

In addition to this, we also test whether some pairs of statistics work better than others. The motivation for this is that, although biologists often restrict themselves to the Colless and Sackin indices, it is well documented that these indices are extremely correlated – both under theoretical models and in practice (see e.g. [7] as well as Figure 10 below). As a result, to some extent they contain the same information and it is unclear whether using them jointly provides a real advantage over using only one of them with another balance index such as B_2.

Null models	Test sets	Range of n	Balance indices
ERM (Yule model)	TreeBASE [42] 4378 trees	$n \in 6$–12	Sackin index [41] see Eq. (2) here
PDA (uniform model)	OrthoMaM [40] 14509 trees	$n \in 12$–24 $n \in 25$–50	Colless index [14] see Eq. (1)
AB (β-splitting with $\beta = -1$)	PhylomeDB [23] 77843 trees HOGENOM [34] 85581 trees	$n \in 50$–100 $n > 100$	# of cherries [32] σ_N^2 index [27, 39] see Eq. (11)
MFL1 (see main text) $p = 0.5, s = 0.25$			B_1 index [41] see Eq. (11)
MFL2 $p = 0.5, s = 0.5$			B_2 index [41] see Def. 1.2

Table 1: Summary of the null models, test sets, size categories and balance indices used in this study. Note that the columns are independent (that is, there is no correspondence between the lines of different columns).

To assess this, we estimate the statistical power of each pair of balance indices. The methodology for doing this is exactly the same as when testing the power of a single balance index – except that this time we need 2D confidence regions to decide whether to keep or to reject each tree. Because there is no standard way to build such 2D confidence regions for non-gaussian data (in particular when the joint distribution of the data is not symmetric, as is the case here; see Figure 10), we had to choose one such method somewhat arbitrarily. We chose to decompose the observations into convex layers and pick the largest layer that contains less than 95% of the observations. Compared to fitting a density and using its level sets, this method has the advantage of being less computationally intensive and more robust when used on data that take discrete values (when fitting a density, e.g. with a kernel density estimation, the smoothing can unpredictably impact the shape of the resulting level sets).

The convex layers of a set of points are the nested convex polygons obtained through the following procedure: let S_0 be the complete set of points and H_0 be the vertices of its convex hull. Let then $S_{i+1} = S_i \setminus H_i$ and H_{i+1} be the vertices of the convex hull of S_{i+1}, and iterate until S_{i+1} is empty. This construction is illustrated in Figure 9. See e.g. [13] for more on 2D convex layers and how to compute them.

Note that because a null model can generate the same tree several times, and also because different trees can have the same balance indices, the points that we consider are not necessarily distinct; and this relevant information should be taken into account. We thus treat the set of sampled points as a multiset. When several points are superposed (and therefore correspond to the same vertex of a convex hull H_i), we only remove one of them from S_i. Our convex layers can therefore overlap at their vertices, and each point that is present k times in the data is the vertex of k convex layers.

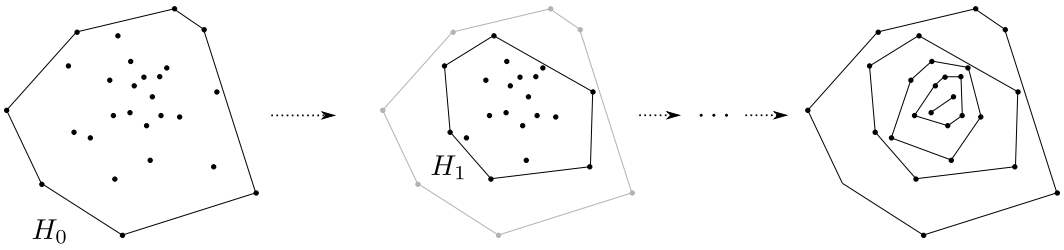


Figure 9: Example of construction of the convex layers decomposition of a set of points (here the innermost convex layer is a degenerate polygon with two vertices).

Finally, since we stop the construction as soon as a convex layer contains less than 95% of the points, our confidence regions do not contain exactly 95% of the points generated by the null model. However, because of the large ($= 10^5$) number of points that we use, they always contain between 95% and 94.5% of the points. Examples of the joint distributions of some balance statistics and of the corresponding 95% confidence regions are given in Figure 10. As before, the code used to compute the confidence regions and its output can be found in [2].

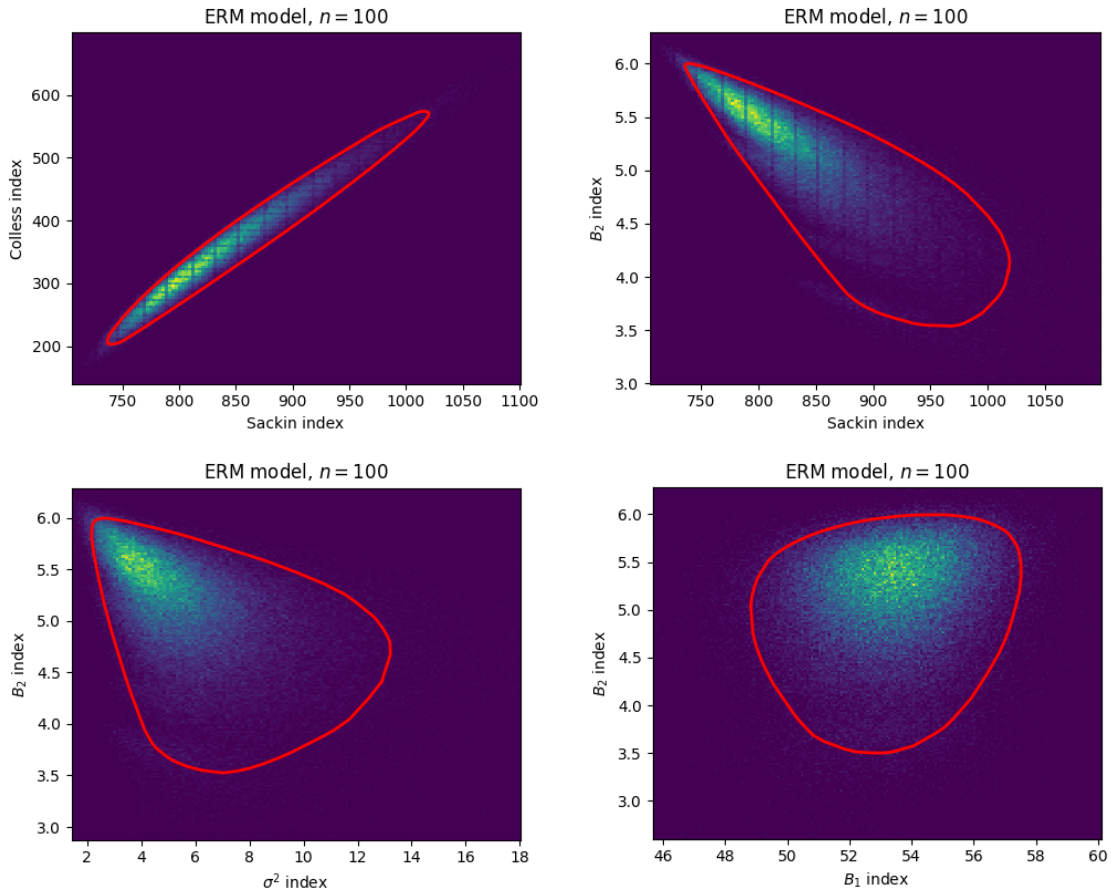


Figure 10: Examples of 2D histograms depicting the joint distributions of several balance indices, here under the ERM model for $n = 100$ leaves. The color scale is not given because it is not relevant. The 95% confidence regions based on the convex layers are represented in red. These histograms and confidence regions were build using the same 10^5 independent realizations of the ERM model. Note that the vertical (resp. horizontal) stripes visible in the distribution of the Sackin (resp. Colless) index are artifacts that result from the binning used to plot the histograms. This occurs only for the Sackin and Colless indices because they take fewer distinct values than the σ_N^2, B_1 or B_2 indices.

4.2 Results

Table 2 is an excerpt from the full output of our analysis. The complete table of results (where the performance of each balance index and each pair of balance indices is given for each of the 85 scenarios that we considered) can be downloaded from [2]. Here we have selected a handful of relevant scenarios to illustrate specific points; a more systematic analysis will follow.

Model	Test set	Range n	Trees	Sackin	Colless	Cherries	σ^2_N	B_1	B_2
ERM	HOGENOM	6–12	39763	4.51	4.77	1.59	5.59	3.40	4.60
ERM	HOGENOM	12–24	24493	17.64	17.91	5.35	18.09	16.05	13.71
ERM	HOGENOM	25–50	14371	35.57	35.11	11.33	34.37	29.27	18.55
ERM	HOGENOM	50–100	7020	53.13	52.45	20.98	50.84	46.85	23.80
ERM	HOGENOM	>100	3305	73.31	72.28	33.77	71.07	65.93	35.46
ERM	PhylomeDB	6–12	11632	31.04	30.75	9.13	26.68	19.11	32.25
ERM	PhylomeDB	12–24	20022	71.97	69.87	12.87	58.95	41.93	77.07
ERM	PhylomeDB	25–50	15523	94.00	92.66	21.81	82.30	59.06	97.04
ERM	PhylomeDB	50–100	11302	98.76	98.37	29.07	90.88	66.71	99.85
ERM	PhylomeDB	>100	21643	99.81	99.72	39.02	95.01	70.59	100
PDA	OrthoMaM	50-100	2700	35.48	41.07	18.52	66.00	67.19	4.26
PDA	OrthoMaM	>100	11741	51.40	63.09	32.92	84.43	91.76	2.84
AB	OrthoMaM	50-100	2700	8.26	7.44	4.04	1.22	4.85	13.85
AB	OrthoMaM	>100	11741	4.64	3.88	1.88	0.49	1.57	10.33
MFL1	OrthoMaM	50-100	2700	2.89	2.44	20.70	0.30	23.30	16.96
MFL1	OrthoMaM	>100	11741	0.68	0.56	15.19	0.06	12.61	11.50
MFL2	OrthoMaM	50-100	2700	3.15	5.70	6.89	16.00	1.63	4.48
MFL2	OrthoMaM	>100	11741	4.34	7.69	3.58	38.11	0.54	2.5

Table 2: Estimated statistical power of each balance index, as measured by the percentage of trees of the test set rejected against the null model. The scenarios given in this table were hand-picked to illustrate specific points. The highlighted values correspond to the best index for a given scenario and the greyed out ones to indices with a power $\leqslant 5\%$.

First, one can ask whether some balance indices are consistently better when working with a specific null model. In particular, in light of the study of Agapow and Purvis we might expect the Sackin and Colless indices to be markedly more powerful than other balance indices when testing against the ERM model. While this general trend is somewhat confirmed by our study, this is by no means a golden rule: as shown in Table 2, B_2 is better at distinguishing the trees of PhylomeDB from ERM trees than any other balance index.

Second, one can wonder whether some balance indices are consistently better with some test sets. Here, comparing the performance of the σ^2_N, B_1 and B_2 indices on OrthoMaM is informative: indeed, these indices perform completely differently depending on the null model used. For instance, B_2 is the only index that fails to distinguish the trees of OrthoMaM from the PDA model; but it is also the only index that is somewhat able to distinguish those same trees from the AB model. Similarly, σ^2_N is the only index able to distinguish OrthoMaM from MFL2, and B_1 is the best index to distinguish OrthoMaM from PDA or MFL1. The Sackin and the Colless index stand out by their poor performance when testing OrthoMaM against the AB, MFL1 or MFL2 models.

Table 2 shows that there are no absolute rules when it comes to ranking the performance of the various balance indices, and looking at the full output of our analysis only reinforces this conclusion. However, general trends might still emerge when looking across the 85 scenarios that we considered.

In order to assess this, we rank the indices from best (1) to worst (6) for each scenario. We then we average these ranks over all scenarios that share a common factor: over all scenarios where the null model is the ERM model; over all scenarios where the test set is TreeBASE; etc. These average ranks are given in Table 3. We use average ranks rather than average powers because when averaging the powers an excellent performance in a single scenario might compensate poor performances in several other scenarios. Using ranks ensures that each scenario gets the same importance.

Factor	Scenarios	Sackin	Colless	Cherries	σ^2_N	B_1	B_2
ERM	17	1.71	2.29	6.00	3.59	4.65	2.76
PDA	17	4.47	3.38	4.18	2.74	1.35	4.88
AB	17	2.76	2.88	5.47	4.00	3.53	2.35
MFL1	17	3.41	3.82	4.29	4.47	2.65	2.35
MFL2	17	3.47	3.00	4.97	2.88	3.88	2.79
TreeBASE	25	3.34	3.44	4.58	3.68	2.80	3.16
OrthoMaM	10	3.00	3.10	4.00	4.00	3.40	3.50
PhylomeDB	25	2.94	3.46	5.12	4.14	3.68	1.66
HOGENOM	25	3.28	2.32	5.64	2.60	3.08	4.08
$n \in 6$–12	15	3.17	2.77	6.00	2.17	4.33	2.57
$n \in 12$–24	15	3.27	2.87	5.73	3.33	3.60	2.20
$n \in 25$–50	15	2.93	3.07	5.07	4.00	2.93	3.00
$n \in 50$–100	20	3.22	3.35	4.47	4.00	2.65	3.30
$n > 100$	20	3.20	3.20	4.10	3.90	2.85	3.75
All	85	3.16	3.08	4.98	3.54	3.21	3.03

Table 3: Average ranks of the balance indices. For each scenario, the indices are ranked from 1 (the best) to 6 (the worse). The ranks are then averaged over all scenarios that share the factor indicated in the left-most column.

Although a few general trends can be identified (such as the fact that the Sackin index seems to be consistently better when testing against the ERM model; or that B_1 seems to be the best index when testing against the PDA model or working with large trees), the main takeaway from Table 3 is that there does not seem to be a "silver bullet" index that would work well in every situation: with the exception of the number of cherries, each index works well in some situations and poorly in some others. This is made apparent by the fact that, when the ranks are averaged over all 85 scenarios, they all come out roughly equal to 3.5 – which is what we expect in the absence of any difference of performance between the indices.

All things considered, the only reasonably certain conclusions that emerge from our analysis are the following:

- The number of cherries is consistently worse than the other balances indices.

- Neither the Colless nor the Sackin index are consistently better than other balance indices.
- The B_2 index is not consistently worse than other indices. In fact, its overall performance seems comparable to that of the Colless and of the Sackin index.

Let us now turn to the results of our bivariate analysis. Recall that, since the Colless and the Sackin index are currently the standard measures of phylogenetic balance and are frequently used together, our specific goal is to test whether other balance indices might complement the Colless / Sackin indices better than they complement each other. Table 4 (resp. 5) gives the average rank of each of the pairs of indices that include the Colless (resp. Sackin) index, using the same methodology as previously (note however that this time the ranks go from 1 to 5, so the expected rank when all indices have the same overall performance is 3).

Factor	Scenarios	(C, Sackin)	(C, cherries)	(C, σ^2_N)	(C, B_1)	(C, B_2)
ERM	17	3.79	3.24	2.41	2.97	2.59
PDA	17	4.41	2.91	2.74	1.44	3.50
AB	17	3.91	3.00	2.00	2.85	3.24
MFL1	17	4.35	3.50	2.00	2.74	2.41
MFL2	17	3.74	3.18	2.06	3.11	2.91
TreeBASE	25	4.40	3.56	1.94	2.54	2.56
OrthoMaM	10	2.65	4.00	2.20	3.45	2.70
PhylomeDB	25	4.28	3.36	1.52	3.56	2.28
HOGENOM	25	4.00	2.24	3.28	1.44	4.04
$n \in$ 6–12	15	4.40	3.07	1.07	3.00	3.47
$n \in$ 12–24	15	4.00	3.80	2.23	2.83	2.13
$n \in$ 25–50	15	4.20	2.80	2.73	2.40	2.87
$n \in$ 50–100	20	3.92	3.15	2.25	2.52	3.15
$n > 100$	20	3.80	3.05	2.75	2.45	2.95
All	85	4.04	3.16	2.24	2.62	2.93

Table 4: Average ranks of the pairs of balance indices that include the Colless index.

One clear conclusion emerges from Tables 4 and 5: that the Sackin index is the worst possible balance index to complement the Colless index, and vice versa. In fact, even complementing them with the number of cherries (which was the only index to perform significantly worse than the others when considered alone) almost invariably gives better results. As mentioned above, this result is not surprising given the strong correlation of the Sackin and Colless indices. It is nevertheless of significant importance, considering how these indices are used in practice.

Another, more unexpected conclusion to be drawn from Tables 4 and 5 is that σ^2_N seems to be consistently better than any other index at complementing the Colless index (with the possible exception of the B_1 index). This also seems to be the case when it comes to complementing the Sackin index, even though in that case things are a bit more nuanced and both the B_1 and B_2 index can be useful.

Factor	Scenarios	(S, Colless)	(S, cherries)	(S, σ^2_N)	(S, B_1)	(S, B_2)
ERM	17	4.62	2.94	2.35	2.59	2.50
PDA	17	4.29	2.82	2.82	1.35	3.71
AB	17	4.12	3.03	2.21	2.65	3.00
MFL1	17	4.53	3.59	2.12	2.70	2.06
MFL2	17	3.94	3.35	2.21	2.79	2.70
TreeBASE	25	4.70	3.48	2.38	2.24	2.20
OrthoMaM	10	2.80	3.80	1.80	3.30	3.30
PhylomeDB	25	4.76	3.26	1.94	3.30	1.74
HOGENOM	25	4.04	2.44	2.92	1.36	4.24
$n \in$ 6–12	15	4.63	3.53	1.47	2.40	2.97
$n \in$ 12–24	15	4.53	3.67	2.27	2.60	1.93
$n \in$ 25–50	15	4.40	2.73	2.87	2.33	2.67
$n \in$ 50–100	20	4.15	3.02	2.40	2.42	3.00
$n > 100$	20	3.95	2.90	2.60	2.35	3.20
All	85	4.30	3.15	2.34	2.42	2.79

Table 5: Average ranks of the pairs of balance indices that include the Sackin index.

Lastly, although this is not the aim of this study, having at our disposal a measure of the power of every pair of balance indices also makes it tempting to adopt a more exploratory approach and assess whether some more exotic combinations of balance indices might be useful. It turns out that this is the case: out of the 15 pairs of indices (expected rank under uniformity: 8), (B_1, B_2) and (σ^2_N, B_2) stand out for having an average rank that is slightly better than other pairs. Moreover, the fact that B_2 is part of these seemingly optimal pairs of balance indices reinforces the idea that its relevance may have been underestimated. Finally, let us point out that, strikingly, (Sackin, Colless) turns out to have the worst performance of all possible pairs of indices.

	Sackin	Colless	Cherries	σ^2_N	B_1	B_2
Sackin	—	11.44	8.77	6.52	6.26	7.59
Colless		—	9.02	6.71	7.26	8.76
Cherries			—	9.91	11.12	7.01
σ^2_N				—	7.93	6.08
B_1					—	5.63

Table 6: Average rank (over all 85 scenarios) of each pair of balance indices. The two best performing pairs are highlighted.

5 Concluding comments

Given its intuitive definition, it is legitimate to wonder why B_2 is not more prominent in phylogenetics; or why it has in fact not been studied in its own right before. To conclude this article, we speculate as to why this might be the case.

One first reason could be that the intuition behind the definition of B_2 is not always well understood. Indeed, although its probabilistic interpretation is very clearly laid out in Shao and Sokal's original paper, it is almost never mentioned in subsequent works that use B_2. For instance, to motivate its definition Kirkpatrick and Slatkin merely say that *"The statistic B_2 was suggested by the Shannon–Wiener statistic. The Shannon–Wiener statistic was developed as an index of information content, and so a measure of tree shape related to it might be a useful statistic for detecting patterns."* [27]. In most sources, B_2 is only described, somewhat vaguely, either as an information theoretic measure of balance or as a weighted variant of the Sackin index. In our opinion this is unfortunate, because the probabilistic interpretation of B_2 is precisely what sets it apart from other balance indices. For instance, it has already been pointed out in the literature that it is not entirely clear why the Sackin index happens to correlate with our intuition of phylogenetic (im)balance – as illustrated by the fact that Sackin's original idea had little to do with the index that came to bear his name [16].

A second possible reason for the lack of popularity of B_2 is that it might be perceived as a mathematically unwieldy quantity. For instance, in the only mention of B_2 that we could find in the mathematical literature [9], after studying the asymptotic distribution of the Colless and Sackin indices under the ERM model Blum, François and Janson conclude by saying that *"In the same spirit, we believe that the B_1 index of Shao and Sokal could be studied without difficulties. Studying the remaining statistics (B_2 and σ_N^2) would nevertheless require considerably more effort."* The results of Sections 2 and 3, as well as the simplicity of their proofs, show that the idea that B_2 is untractable is unjustified. In fact, in some respects B_2 seems *more* tractable than other classic balance indices (compare for instance its expected value under the ERM model to that of the Colless index [22], and consider the fact that the expected value of the Colless index under the PDA model is currently not known).

Finally, the third – and most likely main – reason for the current status of B_2 is probably the reputation of being less useful than other balance indices that it earned from Agapow and Purvis's 2002 study, which they conclude by saying: *"B_2 never performs well and should not be used."* [3]. While this conclusion is justified in their particular setting, it does not hold when their hypotheses are relaxed – in particular when using other null models than the ERM model, or when working with some specific types of phylogenies that occur in the real world. In fact, that B_2 could be useful in some contexts could already be observed in some other studies comparing the performance of various balance indices – in particular in [30], where B_2 is found to be the second most powerful statistic considered, beating the Colless index. The study by Agapow and Purvis also does not take into account the possibility of using balance indices jointly. As shown in Section 4, this completely changes the relevance of the indices, as some that perform well on their own, such as the Sackin and Colless indices, can perform very poorly when used jointly.

In conclusion, none of the reasons that currently make B_2 a "second-rate" balance index seems justified. Our work calls for a reevaluation of the status of B_2 in phylogenetics – in particular in the age of phylogenetic networks, where alternatives will have to be found to the classical measures of phylogenetic balance that are used on trees.

Acknowledgements

The authors thank Simon Penel for his help with the HOGENOM database and Roberto Bacilieri for helpful discussions.

FB and CS were funded by grants ANR-16-CE27-0013 and ANR-19-CE45-0012 from the Agence Nationale de la Recherche, respectively; GC was funded by FEDER / Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación project PGC2018-096956-B-C43.

References

[1] The On-Line Encyclopedia of Integer Sequences, published electronically at http://oeis.org, 2020.

[2] Data and code for "Revisiting Shao and Sokal's B_2 index of phylogenetic balance". *Zenodo*, 2020. DOI:10.5281/zenodo.4088651.

[3] P.-M. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51(6):866–872, 2002. DOI:10.1080/10635150290102564.

[4] D. Aldous. Probability distributions on cladograms. In *Random discrete structures*, pages 1–18. Springer, 1996. DOI:10.1007/978-1-4612-0719-1_1.

[5] E. Bapteste, L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie, and J. Whitfield. Networks: expanding evolutionary thinking. *Trends in Genetics*, 29(8):439–441, 2013. DOI:10.1016/j.tig.2013.05.007.

[6] F. Bienvenu, A. Lambert, and M. Steel. Combinatorial and stochastic properties of ranked tree-child networks. *arXiv preprint*, 2020. arXiv:2007.09701.

[7] M. G. Blum and O. François. On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. *Mathematical Biosciences*, 195 (2):141–153, 2005. DOI:10.1016/j.mbs.2005.03.003.

[8] M. G. Blum and O. François. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, 2006. DOI:10.1080/10635150600889625.

[9] M. G. Blum, O. François, and S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Annals of Applied Probability*, 16(4):2195–2214, 2006. DOI:10.1214/105051606000000547.

[10] G. Cardona and L. Zhang. Counting and enumerating tree-child networks and their subclasses. *Journal of Computer and System Sciences*, 114:84–104, 2020. DOI:10.1016/j.jcss.2020.06.001.

[11] G. Cardona, F. Rosselló, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):552–569, 2009. DOI:10.1109/TCBB.2007.70270.

[12] G. Cardona, A. Mir, and F. Rosselló. Exact formulas for the variance of several balance indices under the Yule model. *Journal of Mathematical Biology*, 67 (6-7):1833–1846, 2013. DOI:10.1007/s00285-012-0615-9.

[13] B. Chazelle. On the convex layers of a planar set. *IEEE Transactions on Information Theory*, 31:509–517, 1985. DOI:10.1109/TIT.1985.1057060.

[14] D. H. Colless. Review of "Phylogenetics: the theory and practice of phylogenetic systematics". 1982. DOI:10.2307/2413419.

[15] T. M. Coronado, M. Fischer, L. Herbst, F. Rosselló, and K. Wicke. On the minimum value of the Colless index and the bifurcating trees that achieve it. *Journal of Mathematical Biology*, 80(7):1993–2054, 2020. DOI:10.1007/s00285-020-01488-9.

[16] T. M. Coronado, A. Mir, F. Rosselló, and L. Rotger. On Sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index. *BMC Bioinformatics*, 21(1):1–17, 2020. DOI:10.1186/s12859-020-3405-1.

[17] N. Curien. Random graphs: the local convergence point of view. *Lecture notes*, 2018. https://www.imo.universite-paris-saclay.fr/~curien/cours/cours-RG.pdf.

[18] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2nd edition, 2003.

[19] M. Fischer. Extremal values of the sackin balance index for rooted binary trees. *arXiv preprint*, 2018. arXiv:1801.10418.

[20] P. Flajolet and H. Prodinger. Level number sequences for trees. *Discrete Mathematics*, 65(2):149–156, 1987. DOI:10.1016/0012-365X(87)90137-3.

[21] M. Hayati, B. Shadgar, and L. Chindelevitch. A new resolution function to evaluate tree shape statistics. *PloS ONE*, 14(11), 2019.

[22] S. B. Heard. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46(6):1818, 1992. DOI:10.2307/2410033.

[23] J. Huerta-Cepas, S. Capella-Gutiérrez, L. P. Pryszcz, M. Marcet-Houben, and T. Gabaldón. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic acids research*, 42(D1):D897–D902, 2014. DOI:10.1093/nar/gkt1177.

[24] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006. DOI:10.1093/molbev/msj030.

[25] S. Janson. Simply generated trees, conditioned Galton–Watson trees, random allocations and condensation. *Probability Surveys*, 9:103–252, 2012. DOI:10.1214/11-PS188.

[26] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982. DOI:10.1016/0304-4149(82)90011-4.

[27] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171, 1993. DOI:10.2307/2409983.

[28] D. E. Knuth. *The Art of Computer Programming: Volume 1: Fundamental Algorithms*. Addison-Wesley Professional, 3rd edition, 1997.

[29] A. Lambert. Probabilistic models for the (sub)tree(s) of life. *Brazilian Journal of Probability and Statistics*, 31(3):415–475, 2017. DOI:10.1214/16-BJPS320.

[30] L. P. Maia, A. Colato, and J. F. Fontanari. Effect of selection on the topology of genealogical trees. *Journal of theoretical biology*, 226(3):315–320, 2004.

[31] F. A. Matsen. A geometric approach to tree shape statistics. *Systematic biology*, 55(4):652–661, 2006. DOI:10.1080/10635150600889617.

[32] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164(1):81 – 92, 2000. ISSN 0025-5564. DOI:10.1016/S0025-5564(99)00060-7.

[33] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958. DOI:10.1017/S0305004100033193.

[34] S. Penel, A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière. Databases of homologous gene families for comparative genomics. In *BMC bioinformatics*, volume 10, 2009. DOI:10.1186/1471-2105-10-S6-S3.

[35] U. Roesler and L. Rüschendorf. The contraction method for recursive algorithms. *Algorithmica*, 29(1):3–33, 2001.

[36] J. S. Rogers. Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution*, 48(6):2026–2036, 1994. DOI:10.1111/j.1558-5646.1994.tb02230.x.

[37] J. S. Rogers. Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology*, 45(1):99, 1996. DOI:10.2307/2413515.

[38] L. Rotger. *New balance indices and metrics for phylogenetic trees*. PhD thesis, Universitat de les Illes Balears, 2019.

[39] M. J. Sackin. "good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972. DOI:10.1093/sysbio/21.2.225.

[40] C. Scornavacca, K. Belkhir, J. Lopez, R. Dernat, F. Delsuc, E. J. P. Douzery, and V. Ranwez. OrthoMaM v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4):861–862, 2019. DOI:10.1093/molbev/msz015.

[41] K. T. Shao and R. R. Sokal. Tree balance. *Systematic Zoology*, 39(3):266–276, 1990. DOI:10.2307/2992186.

[42] R. A. Vos, J. P. Balhoff, J. A. Caravas, M. T. Holder, H. Lapp, W. P. Maddison, P. E. Midford, A. Priyam, J. Sukumaran, X. Xia, and A. Stoltzfus. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Systematic biology*, 61(4):675–689, 2012. DOI:10.1093/sysbio/sys025.

[43] C. Wei, D. Gong, and Q. Wang. Chu–Vandermonde convolution and harmonic number identities. *Integral Transforms and Special Functions*, 24(4):324–330, 2013. DOI:10.1080/10652469.2012.689762.

Appendix

A.1 Variance of B_2 in binary Galton–Watson trees

In this section, we prove Proposition 3.3 concerning the variance of B_2 in binary Galton–Watson trees. Let us start with a standard result, which we recall and prove for the sake of completeness.

Lemma A.1.1. *Let (Z_t) be a Galton–Watson process with offspring distribution $Y \sim 2\,\mathrm{Bernoulli}(1/2)$. Then, $\mathbb{E}(Z_t^2) = t+1$.*

Proof. Since $Z_{t+1} = \sum_{i=1}^{Z_t} Y_t(i)$, we have

$$Z_{t+1}^2 \;=\; \sum_{i=1}^{Z_t} Y_t(i)^2 \;+\; \sum_{i=1}^{Z_t} \sum_{j \neq i}^{Z_t} Y_t(i)\, Y_t(j)\,.$$

Letting $\mathcal{F}_t$ be the natural filtration of the process, we thus have

$$\mathbb{E}\Big(Z_{t+1}^2 \,\Big|\, \mathcal{F}_t\Big) \;=\; \mathbb{E}(Y^2)\, Z_t \;+\; \mathbb{E}(Y)^2\, Z_t(Z_t-1)$$

and, as a result,

$$\mathbb{E}\Big(Z_{t+1}^2\Big) \;=\; \mathbb{E}(Y^2)\,\mathbb{E}(Z_t) \;+\; \mathbb{E}(Y)^2\,\mathbb{E}(Z_t^2) \;-\; \mathbb{E}(Y)^2\,\mathbb{E}(Z_t)\,.$$

Since here $\mathbb{E}(Y) = 1$, $\mathbb{E}(Y^2) = 2$ and $\mathbb{E}(Z_t) = 1$, this simplifies to

$$\mathbb{E}\Big(Z_{t+1}^2\Big) \;=\; \mathbb{E}(Z_t^2) \;+\; 1\,.$$

The lemma then follows by induction, since $\mathbb{E}(Z_0^2) = 1$. □

Let us now recall Proposition 3.3 and prove it.

Proposition 3.3. *Let T be a critical binary Galton–Watson tree, that is, assume that the offspring distribution is $\mathbb{P}(Y=0) = \mathbb{P}(Y=2) = 1/2$. Then,*

$$\mathrm{Var}\Big(B_2(T_{\lfloor t \rfloor})\Big) \;=\; \frac{4}{3} - 2^{-t+2} + 4^{-t}\Big(t + \frac{8}{3}\Big)\,.$$

Proof. To alleviate the notation, let us write $B_t = B_2(T_{\lfloor t \rfloor})$. With this notation, since in the case of a binary Galton–Watson tree $\mathbb{1}_{\{Y_t>0\}} = Y_t/2$, Equation (3) from the proof of Theorem 3.1 becomes

$$B_{t+1} \;=\; B_t \;+\; 2^{-(t+1)} \sum_{i=1}^{Z_t} Y_t(i)\,. \tag{A.1}$$

Therefore, letting $\mathcal{F}_t$ denote the natural filtration of the process and using that $\mathbb{E}(Y) = 1$, we have

$$\mathbb{E}\Big(B_{t+1}^2 \,\Big|\, \mathcal{F}_t\Big) \;=\; B_t^2 \;+\; 2^{-t} B_t\, Z_t \;+\; 2^{-2(t+1)}\,\mathbb{E}\Big(Z_{t+1}^2 \,\Big|\, \mathcal{F}_t\Big).$$

As a result,

$$\mathbb{E}(B_{t+1}^2) \;=\; \mathbb{E}(B_t^2) \;+\; 2^{-t}\,\mathbb{E}(B_t Z_t) \;+\; 2^{-2(t+1)}\,\mathbb{E}(Z_{t+1}^2). \tag{A.2}$$

Let us now turn our attention to $\mathbb{E}(B_t Z_t)$. Using Equation (A.1), we get

$$\begin{aligned} B_{t+1}Z_{t+1} &= \left(B_t + 2^{-(t+1)}Z_{t+1}\right)Z_{t+1} \\ &= B_t \sum_{i=1}^{Z_t} Y_t(i) \;+\; 2^{-(t+1)}Z_{t+1}^2\,, \end{aligned}$$

and since $\mathbb{E}(B_t \sum_{i=1}^{Z_t} Y_t(i) \mid \mathcal{F}_t) = B_t Z_t$ this yields

$$\mathbb{E}(B_{t+1}Z_{t+1}) \;=\; \mathbb{E}(B_t Z_t) \;+\; 2^{-(t+1)}\mathbb{E}(Z_{t+1}^2)\,.$$

By Lemma A.1.1, $\mathbb{E}(Z_{t+1}^2) = t+2$. Since $\mathbb{E}(B_0 Z_0) = 0$, we thus have

$$\mathbb{E}(B_t Z_t) \;=\; \sum_{s=0}^{t-1} 2^{-(s+1)}(s+2) \;=\; 3 - 2^{-t}(t+3)\,.$$

Plugging this and $\mathbb{E}(Z_{t+1}^2) = t+2$ in Equation (A.2), we get the following closed recurrence relation for $\mathbb{E}(B_t^2)$:

$$\mathbb{E}(B_{t+1}^2) \;=\; \mathbb{E}(B_t^2) \;+\; 2^{-t}\left(3 - 2^{-t}(t+3)\right) \;+\; 2^{-2(t+1)}\,(t+2).$$

Solving this recurrence relation with the initial condition $\mathbb{E}(B_0^2) = 0$ then yields

$$\mathbb{E}(B_t^2) \;=\; \frac{7}{3} - 6\cdot 2^{-t} + 4^{-t}\left(t + \frac{11}{3}\right).$$

Finally, since by Theorem 3.1, $\mathbb{E}(B_t) = 1 - 2^{-t}$, we have

$$\mathrm{Var}(B_t) \;=\; \frac{4}{3} - 4\cdot 2^{-t} + 4^{-t}\left(t + \frac{8}{3}\right),$$

concluding the proof. □

A.2 Bounds on the number of distinct values of B_2

Proposition A.2.2. *Let $\mathcal{T}_n$ denote the set of rooted binary trees with n leaves, labeled or unlabeled. Then,*

$$2^{\lfloor n/2 \rfloor - 1} \;\leqslant\; \#B_2(\mathcal{T}_n) \;\leqslant\; a_n$$

where a_n is sequence A002572 in the Online Encyclopedia of Integer Sequences [1] and satisfies $a_n \sim K\rho^n$, with $\rho \approx 1.7941$ and $K \approx 0.2545$ the Flajolet–Prodinger constant.

Proof. The upper bound is obtained by noting that $B_2(T)$ is a function of the multiset of the depths of the leaves of the binary tree T. Therefore, B_2 cannot take more values than the number a_n of such multisets, whose asymptotics were characterized by Flajolet and Prodinger in [20].

To obtain the lower bound, we exhibit $2^{\lfloor n/2 \rfloor - 1}$ rooted binary trees with n leaves whose B_2 indices are different. For any integer m and any $\mathbf{x} \in \{0,1\}^m$, let $T(\mathbf{x})$ denote the ordered (that is, embedded in the plane) rooted binary tree obtained by the following sequential construction: starting from the binary tree with two leaves, for $k = 1, \ldots, m$,

- If $x_k = 0$, graft a cherry on the left-most leaf with depth k.
- If $x_k = 1$, graft a cherry on each of the two left-most leaves with depth k.

This construction is illustrated in Figure 11.

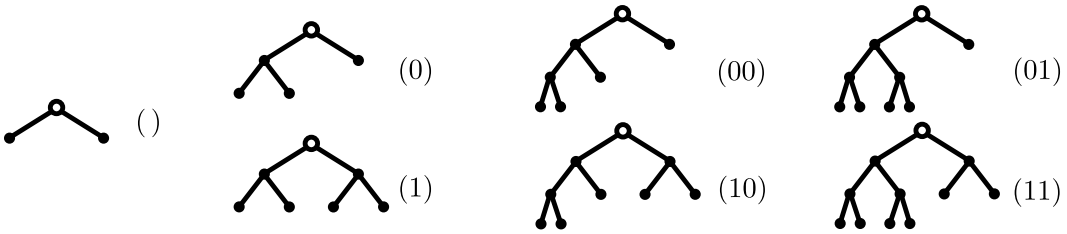


Figure 11: The trees $T(\mathbf{x})$ for $m \in \{0,1,2\}$. Each tree is represented with the corresponding vector $\mathbf{x} \in \{0,1\}^m$ on the right.

Clearly, $T(\mathbf{x})$ has $2 + \sum_{k=1}^{m}(x_k + 1)$ leaves and, by Corollary 1.11,

$$B_2(T(\mathbf{x})) \;=\; 1 \;+\; \sum_{k=1}^{m}(x_k + 1)\, 2^{-k}\,.$$

As a result,

$$\Big\{B_2(T(\mathbf{x})) : \mathbf{x} \in \{0,1\}^m\Big\} \;=\; \Big\{u_m + i\, 2^{-m} : i = 0, \ldots, 2^m - 1\Big\}$$

(to see this, note that $B_2(T(\mathbf{x})) = u_m + \underline{\mathbf{x}}_{(2)}$, where $\underline{\mathbf{x}}_{(2)} = \sum_k x_k 2^{-k}$ denotes the dyadic rational of $[0,1[$ whose binary expansion is $\mathbf{x}$). Thus, this construction generates 2^m trees whose B_2 indices differ by at least 2^{-m}. However, these trees do not have the same number of leaves.

Let us first assume that n is even. Then, with $m = n/2 - 1$, $T(1 \cdots 1)$ has n leaves and every other tree $T(\mathbf{x})$ with $\mathbf{x} \in \{0,1\}^m$ has:

(i) $n - k_{\mathbf{x}}$ leaves, with $1 \leqslant k_{\mathbf{x}} \leqslant m$;

(ii) its left-most leaf at depth $m + 1$.

Now, if we graft a caterpillar with $k_{\mathbf{x}} + 1$ leaves on the left-most leaf at depth $m+1$ and let $T'(\mathbf{x})$ denote the resulting tree, then:

(i′) $T'(\mathbf{x})$ has n leaves;

(ii′) by Proposition 1.10, $B_2(T'(\mathbf{x})) - B_2(T(\mathbf{x})) = 2^{-(m+1)}(2 - 2^{-k+1}) \in]0, 2^{-m}[$.

Since the B_2 indices of the trees $T(\mathbf{x})$ differ by at least 2^{-m}, by point (ii′) the B_2 indices of the trees $T'(\mathbf{x})$ are all different, thereby proving the proposition in the case where n is even.

If n is odd, do the same construction, again with $m = \lfloor n/2 \rfloor - 1$, to get 2^m trees with $n - 1$ leaves. Then, for each of these trees, graft a cherry on the sibling of the left-most vertex at depth $m+1$ (which exists and is always a leaf). This extra step increases B_2 by the same amount $2^{-(m+1)}$ for every tree, concluding the proof. □