

## A balance index for phylogenetic trees based on rooted quartets

Tomás M. Coronado<sup>a</sup>, Arnau Mir<sup>a</sup>, Francesc Rosselló<sup>a</sup>, Gabriel Valiente<sup>b</sup>

<sup>a</sup>*Balearic Islands Health Research Institute (IdISBa) and Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma, Spain*

<sup>b</sup>*Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain*

---

### Abstract

We define a new balance index for rooted phylogenetic trees based on the symmetry of the evolutive history of every set of 4 leaves. This index makes sense for multifurcating trees and it can be computed in time linear in the number of leaves. We determine its maximum and minimum values for arbitrary and bifurcating trees, and we provide exact formulas for its expected value and variance on bifurcating trees under Ford's  $\alpha$ -model and Aldous'  $\beta$ -model and on arbitrary trees under the  $\alpha$ - $\gamma$ -model.

---

### 1. Introduction

One of the most broadly studied properties of the topology of rooted phylogenetic trees is their balance, that is, the tendency of the subtrees rooted at all children of any given node to have a similar shape. The main reason for this interest is that the balance of a rooted tree embodies the symmetry of the evolutive history it describes, and hence it reflects, at least to some extent, a feature of the forces that drove the evolution of the set of species represented in the tree; see Chapter 33 of [12].

The balance of a tree is usually quantified by means of *balance indices*. The two most popular such indices are *Colless' index* [8] for bifurcating trees, which is defined as the sum, over all internal nodes  $v$ , of the absolute value of the difference between the number of descendant leaves of the pair of children of  $v$ , and *Sackin's index* [25, 26], which is defined, for arbitrary trees, as the sum of the depths of all leaves in the tree. But many other balance indices have been proposed in the literature, like for instance, for bifurcating trees, the variance of the depths of the leaves [16, 25], the sum of the reciprocals of the orders of the rooted subtrees [26], and the number of cherries [19], and, for arbitrary trees, the *total cophenetic index* [20] and a generalization of the Colless index [21]. For more indices, see again Chapter 33 in the book by Felsenstein [12]. All these balance indices depend only on the topology of the trees, not on the branch lengths or the actual labels on their leaves, although the balance of time-stamped trees has also been considered by Dearlove and Frost [11]. This abundance of balance indices is partly motivated by the advice given by Shao and Sokal [26] to use more than one such index to quantify the balance of a tree, as well as by their use as tools to test stochastic models of evolution [4, 15, 16, 22, 26]; other properties of the shapes of phylogenetic trees used in this connection include the distribution of clades' sizes [30, 31] and the joint distribution of the numbers of rooted subtrees of different types [28].

In this paper we propose a new balance index, the *rooted quartet index*. To define it, we associate to each 4-tuple of different leaves of the tree  $T$  a value that quantifies the symmetry of the joint evolution of the species they represent, in the sense that it grows with the number

---

*Email addresses:* [t.martinez@uib.eu](mailto:t.martinez@uib.eu) (Tomás M. Coronado), [arnau.mir@uib.eu](mailto:arnau.mir@uib.eu) (Arnau Mir), [cesc.rossello@uib.eu](mailto:cesc.rossello@uib.eu) (Francesc Rosselló), [valiente@cs.upc.edu](mailto:valiente@cs.upc.edu) (Gabriel Valiente)

of isomorphisms of the restriction of  $T$  to them (the *rooted quartet* they define), and then we add up these values over all 4-tuples of different leaves of  $T$ . The idea behind the definition of this balance index is that a highly symmetrical evolutive process should give rise to symmetrical evolutive histories of many small subsets of taxa. In terms of phylogenetic trees, this leads us to expect that, the most symmetrical a phylogenetic tree is, the most symmetrical will be its restrictions to subsets of leaves of a fixed cardinality. Since the smallest number of leaves yielding enough different tree topologies to allow a meaningful comparison of their symmetry is 4, we assess the balance of a tree by measuring the symmetry of all its rooted quartets and adding up the results. And indeed, in Section 4 below we shall find the trees with maximum and minimum values of our rooted quartet index in both the arbitrary and the bifurcating cases, and it will turn out that the minimum value is reached exactly at the combs (see Fig. 1.(a)), which are usually considered the least balanced trees, and the maximum value is reached, in the arbitrary case, exactly at the rooted stars (see Fig. 1.(b)) and, in the bifurcating case, exactly at the maximally balanced trees (cf. Fig. 3), which in both cases are considered the most balanced trees.

Besides taking its maximum and minimum values at the expected trees, other important features of our index are that it can be easily computed in linear time and that its mean value and variance can be explicitly computed on any probabilistic model of phylogenetic trees satisfying two natural conditions: independence under relabelings and sampling consistency. This allows us to provide these values for two well-known probabilistic models of bifurcating phylogenetic trees, Ford’s  $\alpha$ -model [13] and Aldous’  $\beta$ -model [2], which include as specific instances the Yule [14, 29] and the uniform [6, 24, 19] models, as well as for Chen-Ford-Winkel’s  $\alpha$ - $\gamma$ -model of multifurcating trees [7]. To our knowledge, this is the first shape index for which closed formulas for the expected value and the variance under the  $\alpha$ - $\gamma$ -model have been provided.

The rest of this paper is organized as follows. In the next section we introduce the basic notations and facts on phylogenetic trees that will be used in the rest of the paper, and we recall several preliminary results on probabilistic models of phylogenetic trees, proving those results for which we have not been able to find a suitable reference in the literature. Then, in Section 3, we define our rooted quartet index  $rQI$  and we establish its basic properties. In Section 4 we compute its maximum and minimum values, and finally, in Section 5, we compute its expected value and variance under different probabilistic models. This paper is accompanied by the GitHub page [https://github.com/biocom-uib/Quartet\\_Index](https://github.com/biocom-uib/Quartet_Index) containing a set of Python scripts that perform several computations related to this index.

## 2. Preliminaries

### 2.1. Notations and conventions

In this paper, by an (*unlabeled*) *tree* we mean a rooted tree without out-degree 1 nodes. As it is usual, we understand such a tree as a directed graph, with its arcs pointing away from the root. A tree is *bifurcating* when all its internal nodes have out-degree 2; when we want to emphasize that a tree need not be bifurcating, we shall call it *multifurcating*. We shall denote by  $L(T)$  the set of leaves of a tree  $T$ , by  $V_{int}(T)$  its set of internal nodes, and by  $child(u)$  the set of *children* of an internal node  $u$ , that is, those nodes  $v$  such that  $(u, v)$  is an arc in  $T$ . We shall always consider two isomorphic trees as equal, and we shall denote by  $\mathcal{T}_n^*$  and  $\mathcal{BT}_n^*$  the sets of (isomorphism classes of) multifurcating trees and of bifurcating trees with  $n$  leaves, respectively.

A *phylogenetic tree* on a set  $\Sigma$  is a tree with its leaves bijectively labeled in  $\Sigma$ . An *isomorphism* of phylogenetic trees is an isomorphism of trees that preserves the leaves’ labels. To simplify the language, we shall always identify a leaf of a phylogenetic tree with its label and we shall say that two isomorphic phylogenetic trees “are the same”. We shall denote by  $\mathcal{T}(\Sigma)$  and  $\mathcal{BT}(\Sigma)$  the sets of (isomorphism classes of) multifurcating phylogenetic trees and of bifurcating phylogenetic trees on  $\Sigma$ , respectively. If  $\Sigma$  and  $\Sigma'$  are any two sets of labels of the same cardinality, say  $n$ , then any bijection  $\Sigma \leftrightarrow \Sigma'$  extends in a natural way to bijections  $\mathcal{T}(\Sigma) \leftrightarrow \mathcal{T}(\Sigma')$  and  $\mathcal{BT}(\Sigma) \leftrightarrow \mathcal{BT}(\Sigma')$ .

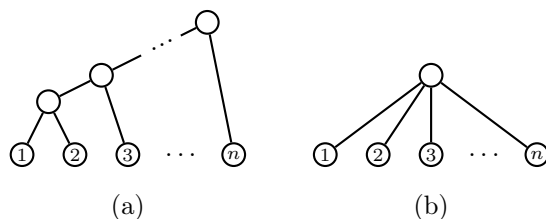
When the specific set of labels  $\Sigma$  is irrelevant and only its cardinality matters, we shall write  $\mathcal{T}_n$  and  $\mathcal{BT}_n$  (with  $n = |\Sigma|$ ) instead of  $\mathcal{T}(\Sigma)$  and  $\mathcal{BT}(\Sigma)$ , and we shall identify  $\Sigma$  with the set  $[n] = \{1, 2, \dots, n\}$ . If  $|\Sigma| = n$ , there exists a forgetful mapping  $\pi : \mathcal{T}(\Sigma) \rightarrow \mathcal{T}_n^*$  that sends every phylogenetic tree  $T$  on  $\Sigma$  to its underlying unlabeled tree: we shall call  $\pi(T)$  the *shape* of  $T$ . We shall write  $T_1 \equiv T_2$  to denote that two phylogenetic trees  $T_1, T_2$  (possibly on different sets of labels of the same cardinality) have the same shape.

We shall represent trees and phylogenetic trees by means of their usual Newick format,<sup>1</sup> although we shall omit the ending mark “;” in order not to confuse it in the text with a semicolon punctuation mark. In the case of (unlabeled) trees, we shall denote the leaves with  $*$  symbols.

Given two nodes  $u, v$  in a tree  $T$ , we say that  $v$  is a *descendant* of  $u$ , and also that  $u$  is an *ancestor* of  $v$ , when there exists a path from  $u$  to  $v$  in  $T$ ; this, of course, includes the case of the stationary path from a node  $u$  to itself, and hence, in this context, we shall use the adjective *proper* to mean that  $u \neq v$ . Given a node  $v$  of a tree  $T$ , the *subtree*  $T_v$  of  $T$  rooted at  $v$  is the subgraph of  $T$  induced by the descendants of  $v$ . We shall denote by  $\kappa_T(v)$ , or simply by  $\kappa(v)$  if  $T$  is implicitly understood, the number of leaves of  $T_v$ .

Given a tree  $T$  and a subset  $X \subseteq L(T)$ , the *restriction*  $T(X)$  of  $T$  to  $X$  is the tree obtained by first taking the subgraph of  $T$  induced by all the ancestors of leaves in  $X$  and then suppressing its out-degree 1 nodes. By *suppressing* a node  $u$  with out-degree 1 we mean that if  $u$  is the root, we remove it together with the arc incident to it, and, if  $u$  is not the root and if  $u'$  and  $u''$  are, respectively, its parent and its child, then we remove the node  $u$  and the arcs  $(u', u), (u, u'')$  and we replace them by a new arc  $(u', u'')$ . For every  $Y \subseteq L(T)$ , the tree  $T(-Y)$  obtained by *removing*  $Y$  from  $T$  is nothing but the restriction  $T(L(T) \setminus Y)$ . If  $T$  is a phylogenetic tree on a set  $\Sigma$  and  $X \subseteq \Sigma$ , the restrictions  $T(X)$  and  $T(-X)$  are phylogenetic trees on  $X$  and  $\Sigma \setminus X$ , respectively.

A *comb* is a bifurcating phylogenetic tree such that all its internal nodes have a leaf child: see Fig. 1.(a). All combs with the same number  $n$  of leaves have the same shape, and we shall generically denote them (as well as their shape in  $\mathcal{T}_n^*$ ) by  $K_n$ . A *rooted star* is a phylogenetic tree all of whose leaves are children of the root: see Fig. 1.(b). For every set  $\Sigma$ , there is only one rooted star on  $\Sigma$ , and if  $|\Sigma| = n$ , we shall generically denote it (as well as its shape) by  $S_n$ .

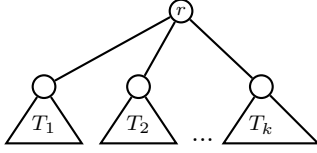


**Figure 1:** (a) A comb  $K_n$ . (b) A rooted star  $S_n$ .

Given  $k \geq 2$  phylogenetic trees  $T_1, \dots, T_k$ , with every  $T_i \in \mathcal{T}(\Sigma_i)$  and the sets of labels  $\Sigma_i$  pairwise disjoint, their *root join* is the phylogenetic tree  $T_1 \star T_2 \star \dots \star T_k$  on  $\bigcup_{i=1}^k \Sigma_i$  obtained by connecting the roots of (disjoint copies of)  $T_1, \dots, T_k$  to a new common root  $r$ ; see Fig. 2. If  $T_1, \dots, T_k$  are unlabeled trees, a similar construction yields a tree  $T_1 \star \dots \star T_k$ .

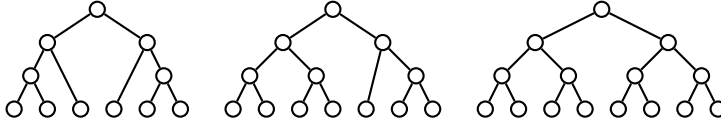
Let  $T$  be a bifurcating tree. For every  $v \in V_{int}(T)$ , say with children  $v_1, v_2$ , the *balance value* of  $v$  is  $bal_T(v) = |\kappa(v_1) - \kappa(v_2)|$ . An internal node  $v$  of  $T$  is *balanced* when  $bal_T(v) \leq 1$ . So, a node  $v$  with children  $v_1$  and  $v_2$  is balanced if, and only if,  $\{\kappa(v_1), \kappa(v_2)\} = \{\lfloor \kappa(v)/2 \rfloor, \lceil \kappa(v)/2 \rceil\}$ . We shall say that a bifurcating tree  $T$  is *maximally balanced* when all its internal nodes are balanced. Recursively, a bifurcating tree is maximally balanced when its root is balanced and

<sup>1</sup>See <http://evolution.genetics.washington.edu/phylip/newicktree.html>



**Figure 2:** The root join  $T_1 \star \dots \star T_k$ .

the subtrees rooted at the children of the root are both maximally balanced. This implies that, for any fixed number  $n$  of nodes, there is only one maximally balanced tree in  $\mathcal{BT}_n^*$ ; see Section 2.1 in [20]. Fig. 3 depicts the maximally balanced trees with  $n = 6, 7, 8$  leaves. When  $n$  is a power of 2, the maximally balanced bifurcating tree with  $n$  leaves is the *fully symmetric bifurcating tree*, where, for each internal node, the pair of subtrees rooted at its children are isomorphic; see again Fig. 3 for  $n = 8$ .



**Figure 3:** The shapes of the maximally balanced trees with 6, 7, and 8 leaves. The tree with 8 leaves is fully symmetric.

## 2.2. Probabilistic models

A *probabilistic model of phylogenetic trees*  $P_n$ ,  $n \geq 1$ , is a family of probability mappings  $P_n : \mathcal{T}_n \rightarrow [0, 1]$ , each one sending each phylogenetic tree in  $\mathcal{T}_n$  to its probability under this model. Every such a probabilistic model of phylogenetic trees  $P_n$  induces a *probabilistic model of trees*, that is, a family of probability mappings  $P_n^* : \mathcal{T}_n^* \rightarrow [0, 1]$ , by defining the probability of a tree as the sum of the probabilities of all phylogenetic trees in  $\mathcal{T}_n$  with that shape:

$$P_n^*(T^*) = \sum_{\substack{T \in \mathcal{T}_n \\ \pi(T) = T^*}} P_n(T).$$

If  $|\Sigma| = n$ , then  $P_n : \mathcal{T}_n \rightarrow [0, 1]$  induces also a probability mapping  $P_\Sigma$  on  $\mathcal{T}(\Sigma)$  through the bijection  $\mathcal{T}_\Sigma \leftrightarrow \mathcal{T}_n$  induced by a given bijection  $\Sigma \leftrightarrow [n]$ .

A *probabilistic model of bifurcating phylogenetic trees* is a probabilistic model of phylogenetic trees  $P_n$  such that  $P_n(T) = 0$  for every  $T \in \mathcal{T}_n \setminus \mathcal{BT}_n$ .

A probabilistic model of phylogenetic trees  $P_n$  is *shape invariant* (or *exchangeable*, according to Aldous [2]) when, for every  $T, T' \in \mathcal{T}_n$ , if  $T \equiv T'$ , then  $P_n(T) = P_n(T')$ . In this case, for every  $T^* \in \mathcal{T}_n^*$  and for every  $T \in \pi^{-1}(T^*)$ ,

$$P_n^*(T^*) = |\{T' \in \mathcal{T}_n : \pi(T') = T^*\}| \cdot P_n(T).$$

Conversely, every probabilistic model of trees  $P_n^*$  defines a shape invariant probabilistic model of phylogenetic trees  $P_n$  by means of

$$P_n(T) = \frac{P_n^*(\pi(T))}{|\{T' \in \mathcal{T}_n : T' \equiv T\}|}. \quad (1)$$

Notice that if  $P_n$  is shape invariant, then, for every set of labels  $\Sigma$ , say, with  $|\Sigma| = n$ , the probability mapping  $P_\Sigma : \mathcal{T}(\Sigma) \rightarrow [0, 1]$  induced by the mapping  $P_n$  does not depend on the specific bijection  $\Sigma \leftrightarrow [n]$  used to define it.

A probabilistic model of phylogenetic trees  $P_n$  is *sampling consistent* [2] (or also *deletion stable*, according to Ford [13]) when, for every  $n \geq 2$ , if we choose a tree  $T \in \mathcal{T}_n$  with probability distribution  $P_n$  and we remove its leaf  $n$ , the resulting tree is obtained with probability distribution  $P_{n-1}$ ; formally, when, for every  $n \geq 2$  and for every  $T_0 \in \mathcal{T}_{n-1}$ ,

$$P_{n-1}(T_0) = \sum_{\substack{T \in \mathcal{T}_n \\ T(-n)=T_0}} P_n(T).$$

It is straightforward to prove, by induction on  $n - m$  and using that, for every  $T \in \mathcal{T}_n$  and for every  $1 \leq m < n$ , the restriction of  $T(-n)$  to  $[m]$  is simply  $T([m])$ , that this condition is equivalent to the following:  $P_n$  is sampling consistent when, for every  $n \geq 2$ , for every  $1 \leq m < n$ , and for every  $T_0 \in \mathcal{T}_m$ ,

$$P_m(T_0) = \sum_{\substack{T \in \mathcal{T}_n \\ T([m])=T_0}} P_n(T). \quad (2)$$

It is also easy to prove that if  $P_n$  is sampling consistent *and* shape invariant, so that the probability of a phylogenetic tree is not affected by permutations of its leaves, then, for every  $n \geq 2$ , for every  $\emptyset \neq X \subsetneq [n]$ , say, with  $|X| = m$ , and for every  $T_0 \in \mathcal{T}(X)$ ,

$$P_X(T_0) = \sum_{\substack{T \in \mathcal{T}_n \\ T(X)=T_0}} P_n(T).$$

(where  $P_X$  stands for the probability mapping on  $\mathcal{T}(X)$  induced by  $P_m$  through any bijection  $X \leftrightarrow [m]$ ).

A probabilistic model of trees  $P_n^*$  is *sampling consistent* when, for every  $n \geq 2$ , if we choose a tree  $T \in \mathcal{T}_n^*$  with probability distribution  $P_n^*$  and a leaf  $x \in L(T)$  equiprobably and if we remove  $x$  from  $T$ , the resulting tree is obtained with probability distribution  $P_{n-1}^*$ : formally, when, for every  $n \geq 2$  and for every  $T_0 \in \mathcal{T}_{n-1}^*$ ,

$$P_{n-1}^*(T_0) = \sum_{T \in \mathcal{T}_n^*} \frac{|\{x \in L(T) : T(-x) = T_0\}|}{n} \cdot P_n^*(T).$$

We prove now several lemmas on probabilistic models that will be used in Section 5. The first lemma provides an extension of equation (2) to trees; we include it because we have not been able to find a suitable reference for it in the literature. In it, and henceforth,  $\mathcal{P}_k(X)$  denotes the set of all subsets of cardinality  $k$  of  $X$ .

**Lemma 1.** *A probabilistic model of trees  $P_n^*$  is sampling consistent if, and only if, for every  $n \geq 2$ , for every  $1 \leq m < n$ , and for every  $T_0 \in \mathcal{T}_m^*$ ,*

$$P_m^*(T_0) = \sum_{T \in \mathcal{T}_n^*} \frac{|\{X \in \mathcal{P}_m(L(T)) : T(X) = T_0\}|}{\binom{n}{m}} \cdot P_n^*(T).$$

*Proof.* The “if” implication is obvious. As far as the “only if” implication goes, we prove by induction on  $n - m$  that if  $P_n^*$  is sampling consistent, then, for every  $T_0 \in \mathcal{T}_m^*$ ,

$$P_m^*(T_0) = \sum_{T_n \in \mathcal{T}_n^*} \frac{|\{X \in \mathcal{P}_{n-m}(L(T_n)) : T_n(-X) = T_0\}|}{\binom{n}{m}} \cdot P_n^*(T_n).$$

The starting case  $m = n - 1$  is the sampling consistency property. Assume now that this equality

holds for  $m$  and let  $T_0 \in \mathcal{T}_{m-1}^*$ . Then

$$\begin{aligned}
P_{m-1}^*(T_0) &= \sum_{T_m \in \mathcal{T}_m^*} \frac{|\{x \in L(T_m) : T_m(-x) = T_0\}|}{m} \cdot P_m^*(T_m) \\
&\quad (\text{by the sampling consistency}) \\
&= \sum_{T_m \in \mathcal{T}_m^*} \left( \frac{|\{x \in L(T_m) : T_m(-x) = T_0\}|}{m} \right. \\
&\quad \left. \cdot \sum_{T_n \in \mathcal{T}_n^*} \frac{|\{X \in \mathcal{P}_{n-m}(L(T_n)) : T_n(-X) = T_m\}|}{\binom{n}{m}} \cdot P_n^*(T_n) \right) \\
&\quad (\text{by the induction hypothesis}) \\
&= \sum_{T_m \in \mathcal{T}_m^*} \sum_{T_n \in \mathcal{T}_n^*} \left( \frac{|\{x \in L(T_m) : T_m(-x) = T_0\}|}{m} \right. \\
&\quad \left. \cdot \frac{|\{X \in \mathcal{P}_{n-m}(L(T_n)) : T_n(-X) = T_m\}|}{\binom{n}{m}} \right) \cdot P_n^*(T_n) \\
&= \sum_{T_n \in \mathcal{T}_n^*} \frac{|\{(X, x) \in \mathcal{P}_{n-m}(L(T_n)) \times (L(T_n) \setminus X) : (T_n(-X))(-x) = T_0\}|}{m \cdot \binom{n}{m}} \cdot P_n^*(T_n) \\
&= \sum_{T_n \in \mathcal{T}_n^*} \frac{|\{(X, x) \in \mathcal{P}_{n-m}(L(T_n)) \times (L(T_n) \setminus X) : (T_n(-(X \cup \{x\}))) = T_0\}|}{m \cdot \binom{n}{m}} \cdot P_n^*(T_n) \\
&= \sum_{T_n \in \mathcal{T}_n^*} \frac{(n-m+1) |\{Y \in \mathcal{P}_{n-m+1}(L(T_n)) : T_n(-Y) = T_0\}|}{m \cdot \binom{n}{m}} \cdot P_n^*(T_n) \\
&= \sum_{T_n \in \mathcal{T}_n^*} \frac{|\{Y \in \mathcal{P}_{n-m+1}(L(T_n)) : T_n(-Y) = T_0\}|}{\binom{n}{m-1}} \cdot P_n^*(T_n)
\end{aligned}$$

which proves the inductive step.  $\square$   $\square$

**Lemma 2.** *Let  $P_n$  be a shape invariant probabilistic model of phylogenetic trees. For every  $T_{n-1}, T'_{n-1} \in \mathcal{T}_{n-1}$ , if  $T_{n-1} \equiv T'_{n-1}$ , then*

$$\sum_{\substack{T_n \in \mathcal{T}_n \\ T_n(-n) = T_{n-1}}} P_n(T_n) = \sum_{\substack{T'_n \in \mathcal{T}_n \\ T'_n(-n) = T'_{n-1}}} P_n(T'_n).$$

*Proof.* Let  $T_{n-1}^* = \pi(T_{n-1}) = \pi(T'_{n-1})$  and let  $f : T_{n-1} \rightarrow T'_{n-1}$  be an isomorphism of unlabeled trees, which exists because  $T_{n-1}$  and  $T'_{n-1}$  are both isomorphic as unlabeled trees to their shape  $T_{n-1}^*$ . For every  $T \in \mathcal{T}_{n-1}$ , let

$$E_n(T) = \{T_n \in \mathcal{T}_n : T_n(-n) = T\}.$$

Each  $T_n$  in  $E_n(T_{n-1})$  is obtained by adding a leaf  $n$  to  $T_{n-1}$  as a new child either to an internal node, or to a new node obtained by splitting an arc into two consecutive arcs, or to a new bifurcating root (whose other child would be the old root). This entails the existence of a shape preserving bijection

$$\Phi : E_n(T_{n-1}) \rightarrow E_n(T'_{n-1})$$

that sends each  $T_n \in E_n(T_{n-1})$  to the phylogenetic tree  $\Phi(T_n)$  obtained by adding the leaf  $n$  to  $T'_{n-1}$  at the place corresponding through the isomorphism  $f$  to the place where it has been added to  $T_{n-1}$ . Then, since  $P_n$  is shape invariant,

$$\sum_{T_n \in E(T_{n-1})} P_n(T_n) = \sum_{T_n \in E(T_{n-1})} P_n(\Phi(T_n)) = \sum_{T'_n \in E(T'_{n-1})} P_n(T'_n)$$

as we claimed.  $\square$   $\square$

Next lemma generalizes Cor. 40 of [13]. For the sake of completeness, we provide a direct complete proof of it.

**Lemma 3.** *Let  $P_n$  be a shape invariant probabilistic model of phylogenetic trees and let  $P_n^*$  be the corresponding probabilistic model of trees. Then,  $P_n$  is sampling consistent if, and only if,  $P_n^*$  is sampling consistent.*

*Proof.* Let us prove first the “only if” implication. Let  $P_n$  be sampling consistent. Then, for every  $T_{n-1}^* \in \mathcal{T}_{n-1}^*$  and for every  $\widehat{T}_{n-1} \in \pi^{-1}(T_{n-1}^*)$ ,

$$\begin{aligned}
P_{n-1}^*(T_{n-1}^*) &= |\{T_{n-1} \in \mathcal{T}_{n-1} : \pi(T_{n-1}) = T_{n-1}^*\}| \cdot P_{n-1}(\widehat{T}_{n-1}) \\
&\quad \text{(by the shape invariance of } P_n) \\
&= |\{T_{n-1} \in \mathcal{T}_{n-1} : \pi(T_{n-1}) = T_{n-1}^*\}| \cdot \sum_{\substack{T_n \in \mathcal{T}_n \\ T_n(-n) = \widehat{T}_{n-1}}} P_n(T_n) \\
&\quad \text{(by the sampling consistency of } P_n) \\
&= \sum_{T_{n-1} \in \pi^{-1}(T_{n-1}^*)} \sum_{\substack{T_n \in \mathcal{T}_n \\ T_n(-n) = T_{n-1}}} P_n(T_n) \\
&\quad \text{(by Lemma 2)} \\
&= \sum_{\substack{T_n \in \mathcal{T}_n \\ \pi(T_n(-n)) = T_{n-1}^*}} P_n(T_n) = \sum_{\substack{T_n \in \mathcal{T}_n \\ \pi(T_n(-i)) = T_{n-1}^*}} P_n(T_n) \quad \text{for every } i = 1, \dots, n \\
&\quad \text{(by the shape invariance of } P_n).
\end{aligned}$$

Therefore

$$\begin{aligned}
n \cdot P_{n-1}^*(T_{n-1}^*) &= \sum_{i=1}^n \sum_{\substack{T_n \in \mathcal{T}_n \\ \pi(T_n(-i)) = T_{n-1}^*}} P_n(T_n) \\
&= \sum_{T_n \in \mathcal{T}_n} |\{i \in [n] : \pi(T_n(-i)) = T_{n-1}^*\}| \cdot P_n(T_n) \\
&= \sum_{T_n^* \in \mathcal{T}_n^*} \sum_{T_n \in \pi^{-1}(T_n^*)} |\{i \in [n] : \pi(T_n(-i)) = T_{n-1}^*\}| \cdot P_n(T_n) \\
&= \sum_{T_n^* \in \mathcal{T}_n^*} \left( |\{x \in L(T_n^*) : T_n^*(-x) = T_{n-1}^*\}| \cdot \sum_{T_n \in \pi^{-1}(T_n^*)} P_n(T_n) \right) \\
&= \sum_{T_n^* \in \mathcal{T}_n^*} |\{x \in L(T_n^*) : T_n^*(-x) = T_{n-1}^*\}| \cdot P_n^*(T_n^*)
\end{aligned}$$

and hence

$$P_{n-1}^*(T_{n-1}^*) = \sum_{T_n^* \in \mathcal{T}_n^*} \frac{|\{x \in L(T_n^*) : T_n^*(-x) = T_{n-1}^*\}|}{n} \cdot P_n^*(T_n^*)$$

as we wanted to prove.

The proof on the “if” implication consists in carefully running backwards the sequence of equalities in the proof of the “only if” implication. Indeed, assume that  $P_n^*$  is sampling consistent

and let  $T_{n-1} \in \mathcal{T}_{n-1}$  and  $T_{n-1}^* = \pi(T_{n-1}) \in \mathcal{T}_{n-1}^*$ . Then

$$\begin{aligned}
P_{n-1}^*(T_{n-1}^*) &= \sum_{T_n^* \in \mathcal{T}_n^*} \frac{|\{x \in L(T_n^*) : T_n^*(-x) = T_{n-1}^*\}|}{n} \cdot P_n^*(T_n^*) \\
&\quad (\text{by the sampling consistency of } P_n^*) \\
&= \frac{1}{n} \sum_{T_n^* \in \mathcal{T}_n^*} \left( |\{x \in L(T_n^*) : T_n^*(-x) = T_{n-1}^*\}| \cdot \sum_{T_n \in \pi^{-1}(T_n^*)} P_n(T_n) \right) \\
&= \frac{1}{n} \sum_{T_n^* \in \mathcal{T}_n^*} \sum_{T_n \in \pi^{-1}(T_n^*)} |\{i \in [n] : \pi(T_n(-i)) = T_{n-1}^*\}| \cdot P_n(T_n) \\
&= \frac{1}{n} \sum_{T_n \in \mathcal{T}_n} |\{i \in [n] : \pi(T_n(-i)) = T_{n-1}^*\}| \cdot P_n(T_n) \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{\substack{T_n \in \mathcal{T}_n \\ \pi(T_n(-i)) = T_{n-1}^*}} P_n(T_n) = \sum_{\substack{T_n \in \mathcal{T}_n \\ \pi(T_n(-n)) = T_{n-1}^*}} P_n(T_n) \\
&\quad (\text{by the shape invariance of } P_n) \\
&= \sum_{T_{n-1}' \in \pi^{-1}(T_{n-1}^*)} \sum_{\substack{T_n \in \mathcal{T}_n \\ T_n(-n) = T_{n-1}'}} P_n(T_n) \\
&= |\{T_{n-1}' \in \mathcal{T}_{n-1} : \pi(T_{n-1}') = T_{n-1}^*\}| \cdot \sum_{\substack{T_n \in \mathcal{T}_n \\ T_n(-n) = T_{n-1}'}} P_n(T_n) \\
&\quad (\text{by Lemma 2})
\end{aligned}$$

and thus, dividing both sides of this equality by  $|\{T_{n-1}' \in \mathcal{T}_{n-1} : \pi(T_{n-1}') = T_{n-1}^*\}|$  and using the shape invariance of  $P_n$ , we obtain

$$\sum_{\substack{T_n \in \mathcal{T}_n \\ T_n(-n) = T_{n-1}'}} P_n(T_n) = \frac{P_{n-1}^*(T_{n-1}^*)}{|\{T_{n-1}' \in \mathcal{T}_{n-1} : \pi(T_{n-1}') = T_{n-1}^*\}|} = P_{n-1}(T_{n-1})$$

as we wanted to prove. □ □

In Section 5 we shall be concerned with three specific parametric probabilistic models of phylogenetic trees: the  $\beta$ -model, the  $\alpha$ -model, and the  $\alpha$ - $\gamma$ -model. To close this section, we provide detailed descriptions of these models and the explicit computation of the probabilities of all trees with 4 leaves under them.

### 2.2.1. Aldous' $\beta$ -model.

The  $\beta$ -splitting model  $P_{\beta,n}^A$  [2, 3] is a probabilistic model of bifurcating phylogenetic trees that depends on one parameter  $\beta \in (-2, \infty)$ . Let us recall its definition. For every  $m \geq 2$  and  $a = 1, \dots, m-1$ , let

$$q_{m,\beta}(a) = \frac{1}{a_m(\beta)} \cdot \frac{\Gamma(\beta + a + 1)\Gamma(\beta + m - a + 1)}{\Gamma(a + 1)\Gamma(m - a + 1)},$$

where  $\Gamma$  stands for the usual Gamma function defined on  $\mathbb{R}^+$ ,

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,$$

and  $a_m(\beta)$  is a suitable normalizing constant so that  $\sum_{a=1}^{m-1} q_{m,\beta}(a) = 1$ . Recall (see, for instance, Chapter 6 in [1]) that  $\Gamma$  satisfies that  $\Gamma(x+1) = x\Gamma(x)$  and that, for every  $n \in \mathbb{N}$ ,  $\Gamma(n+1) = n!$ .



For every  $m \geq 2$  and  $a = 1, \dots, \lfloor m/2 \rfloor$ , let

$$\widehat{q}_{m,\beta}(a) = \begin{cases} q_{m,\beta}(a) + q_{m,\beta}(m-a) = 2q_{m,\beta}(a) & \text{if } a \neq m/2 \\ q_{m,\beta}(a) & \text{if } a = m/2 \end{cases}$$

With these notations, the probabilities under this model are computed as follows. Let  $n \geq 1$  be a given desired number of leaves:

1. Start with a tree  $T'_1$  consisting of a single node labeled  $n$ . Set  $P'_{\beta,1}(T'_1) = 1$ .
2. At each step  $j = 1, \dots, n-1$ , the current tree  $T'_j$  contains leaves with labels greater than 1. Then, choose equiprobably a leaf in  $T'_j$  with a label  $m$  greater than 1, choose a number  $a = 1, \dots, \lfloor m/2 \rfloor$  with probability distribution  $\widehat{q}_{m,\beta}(a)$ , and split this leaf into a cherry with a child labeled  $a$  and a child labeled  $m-a$ . The resulting tree  $T'_{j+1}$  has then probability

$$P'_{\beta,j+1}(T'_{j+1}) = \frac{\widehat{q}_{m,\beta}(a)}{|\{\text{leaves in } T'_j \text{ labeled } > 1\}|} \cdot P'_{\beta,j}(T'_j).$$

3. When the desired number  $n$  of leaves is reached, all leaves are labeled 1 and  $T'_n$  can be understood as a tree. Then, the probability of a given tree is defined as the sum of the probabilities of all ways of obtaining it by means of the previous procedure; that is, for every  $T_n^* \in \mathcal{BT}_n^*$ , its probability under the  $\beta$ -model is

$$P_{\beta,n}^{A,*}(T_n^*) = \sum_{T'_n = T_n^*} P'_{\beta,n}(T'_n).$$

4. Finally, the probability  $P_{\beta,n}^A(T)$  of any phylogenetic tree  $T \in \mathcal{BT}_n$  is obtained from the probability under  $P_{\beta,n}^{A,*}$  of its shape by means of equation (1):

$$P_{\beta,n}^A(T) = \frac{P_{\beta,n}^{A,*}(\pi(T))}{|\{T' \in \mathcal{BT}_n : T' \equiv T\}|}.$$

The last step in the definition of  $P_{\beta,n}^A$  makes it shape invariant by construction, and Aldous [2] proves that it is sampling consistent. Hence, by Lemma 3, the  $\beta$ -model of trees  $P_{\beta,n}^{A,*}$  is also sampling consistent. This  $\beta$ -model includes as specific cases the Yule model [14, 29] (when  $\beta = 0$ ) and the uniform model [6, 23] (when  $\beta = -3/2$ ).

In Section 5 we shall need to know the probability  $P_{\beta,4}^{A,*}$  of the maximally balanced tree with 4 leaves  $((*,*), (*, *))$ , which we denote in this paper by  $Q_3^*$  (see Figure 6 below). We compute this probability in the following lemma, taking the opportunity to provide a detailed example of how this model associates probabilities to trees through their construction.

**Lemma 4.** *For every  $\beta \in (-2, \infty)$ ,*

$$P_{\beta,4}^{A,*}(Q_3^*) = \frac{3\beta + 6}{7\beta + 18}.$$

*Proof.* We start with a single node labeled 4. In order to obtain a maximally balanced tree  $((1,1), (1,1))$  using the previous procedure, in the first step we must split this node into a cherry with both leaves labeled 2. The probability of choosing this split is

$$\widehat{q}_{4,\beta}(2) = q_{4,\beta}(2) = \frac{1}{a_4(\beta)} \cdot \frac{\Gamma(\beta+3)\Gamma(\beta+3)}{\Gamma(3)\Gamma(3)}.$$

Let us compute the normalizing constant  $a_4(\beta)$ : since

$$q_{4,\beta}(1) = q_{4,\beta}(3) = \frac{1}{a_4(\beta)} \cdot \frac{\Gamma(\beta+2)\Gamma(\beta+4)}{\Gamma(2)\Gamma(4)}$$

$$q_{4,\beta}(2) = \frac{1}{a_4(\beta)} \cdot \frac{\Gamma(\beta+3)\Gamma(\beta+3)}{\Gamma(3)\Gamma(3)}$$

imposing that  $q_{4,\beta}(1) + q_{4,\beta}(2) + q_{4,\beta}(3) = 1$  we obtain

$$a_4(\beta) = \frac{2\Gamma(\beta+2)\Gamma(\beta+4)}{6} + \frac{\Gamma(\beta+3)^2}{4} = \frac{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}{12}.$$

Therefore,

$$q_{4,\beta}(2) = \frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}.$$

In the second step, we choose one of the leaves with probability  $1/2$  and we split it into a cherry  $(1,1)$ . Since there is only one way of splitting a leaf labeled 2,  $q_{2,\beta}(1) = 1$ . So, the probability of the tree obtained in this step is

$$\frac{1}{2}q_{4,\beta}(2).$$

Then, in the third step, we are forced to choose the other leaf labeled 2 and to split it into a cherry  $(1,1)$ . We obtain a maximally balanced tree with all its leaves labeled 1 and its probability is still  $q_{4,\beta}(2)/2$ .

Now, there are two ways of obtaining the tree  $((1,1), (1,1))$  with this construction, depending on which leaf of the cherry  $(2,2)$  we choose to split first. So, the probability of the tree  $Q_3^*$  is

$$P_{\beta,4}^{A,*}(Q_3^*) = 2 \cdot \frac{1}{2}q_{4,\beta}(2) = \frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}.$$

Finally, using that  $\Gamma(x+1) = x\Gamma(x)$ , we have that

$$\frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}$$

$$= \frac{3(\beta+2)^2\Gamma(\beta+2)^2}{4(\beta+3)(\beta+2)\Gamma(\beta+2)^2 + 3(\beta+2)^2\Gamma(\beta+2)^2} = \frac{3\beta+6}{7\beta+18}$$

as we claimed. □

### 2.2.2. Ford's $\alpha$ -model.

The  $\alpha$ -model  $P_{\alpha,n}^F$  introduced by Ford [13] is another probabilistic model of bifurcating phylogenetic trees that depends on one parameter  $\alpha \in [0, 1]$ . It is defined as follows. Let  $n \geq 1$  be any desired number of leaves:

1. Start with the tree  $T_1$  consisting of a single node labeled 1. Set  $P'_{\alpha,1}(T_1) = 1$ .
2. For every  $m = 1, \dots, n-1$ , let  $T_{m+1} \in \mathcal{BT}_{m+1}$  be obtained by adding a new leaf labeled  $m+1$  to  $T_m$ . Then:
  - If the new leaf is added to an arc ending in a leaf,

$$P'_{\alpha,m+1}(T_{m+1}) = \frac{1-\alpha}{m-\alpha} \cdot P'_{\alpha,m}(T_m).$$

- If the new leaf is added to an arc ending in an internal node,

$$P'_{\alpha,m+1}(T_{m+1}) = \frac{\alpha}{m-\alpha} \cdot P'_{\alpha,m}(T_m).$$

- If the new leaf is added to a new root,

$$P'_{\alpha,m+1}(T_{m+1}) = \frac{\alpha}{m-\alpha} \cdot P'_{\alpha,m}(T_m).$$

3. When the desired number  $n$  of leaves is reached, the probability of a given tree is defined as the sum of the probabilities of all phylogenetic trees with that shape; that is, for every  $T_n^* \in \mathcal{BT}_n^*$ , its probability under the  $\alpha$ -model is

$$P_{\alpha,n}^{F,*}(T_n^*) = \sum_{\pi(T'_n)=T_n^*} P'_{\alpha,n}(T'_n).$$

4. Finally, the probability  $P_{\alpha,n}^F(T)$  of any phylogenetic tree  $T \in \mathcal{BT}_n$  is obtained from the probability under  $P_{\alpha,n}^{F,*}$  of its shape by means of equation (1):

$$P_{\alpha,n}^F(T) = \frac{P_{\alpha,n}^{F,*}(\pi(T))}{|\{T' \in \mathcal{BT}_n : T' \equiv T\}|}.$$

The  $\alpha$ -model is again shape invariant by construction and sampling consistent by Prop. 42 of [13], and it also includes as specific cases the Yule model (when  $\alpha = 0$ ) and the uniform model (when  $\alpha = 1/2$ ).

In Section 5, we shall also need to know  $P_{\alpha,4}^{F,*}(Q_3^*)$ , where we recall that  $Q_3^*$  stands for the fully symmetric tree with 4 leaves. This value was provided by Ford [13] in Section 7, Fig. 20, as well as by Coronado et al [10]. In the following lemma we compute it directly from the model's definition to illustrate also in this case how the probability of a tree is obtained through its construction.

**Lemma 5.** *For every  $\alpha \in [0, 1]$ ,*

$$P_{\alpha,4}^{F,*}(Q_3^*) = \frac{1-\alpha}{3-\alpha}.$$

*Proof.* To compute this probability, we shall already start with the cherry  $T_2 = (1, 2)$  in  $\mathcal{BT}_2$ , which has probability  $P'_{\alpha,2}(T_2) = 1$ . Every tree in  $\mathcal{BT}_3$  is obtained by adding a leaf labeled 3 to  $T_2$ . These trees are described in Figure 4. Their probabilities are:

- $K^{(1)}$  and  $K^{(2)}$  are obtained by adding the leaf 3 to an arc in  $T_2$  ending in a leaf. Their probability is then

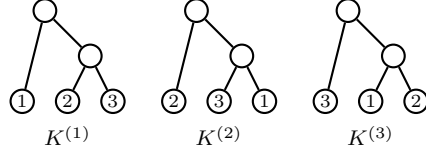
$$P'_{\alpha,3}(K^{(1)}) = P'_{\alpha,3}(K^{(2)}) = \frac{1-\alpha}{2-\alpha}.$$

- $K^{(3)}$  is obtained by adding the leaf 3 to a new root. Its probability is then

$$P'_{\alpha,3}(K^{(3)}) = \frac{\alpha}{2-\alpha}.$$

Now, there are three phylogenetic trees in  $\mathcal{BT}_4$  of shape  $Q_3^*$ , depicted in Figure 5. Each one of them is obtained from the corresponding phylogenetic tree  $K^{(i)}$  by adding the leaf 4 to the arc from the root to its only leaf child. Their probability is, then,

$$P'_{\alpha,4}(Q_3^{(i)}) = \frac{1-\alpha}{3-\alpha} \cdot P'_{\alpha,3}(K^{(i)})$$

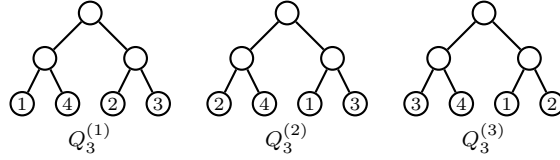


**Figure 4:** The phylogenetic trees in  $\mathcal{BT}_3$ .

and hence, since  $\sum_{i=1}^3 P'_{\alpha,3}(K^{(i)}) = 1$ ,

$$P_{\alpha,4}^*(Q_3^*) = \sum_{i=1}^3 P'_{\alpha,4}(Q_3^{(i)}) = \frac{1-\alpha}{3-\alpha}$$

as we claimed. □ □



**Figure 5:** The fully symmetric phylogenetic trees in  $\mathcal{BT}_4$ .

### 2.2.3. Chen-Ford-Winkel's $\alpha$ - $\gamma$ -model.

The  $\alpha$ - $\gamma$ -model  $P_{\alpha,\gamma,n}$ , defined by Chen et al [7], is a probabilistic model of multifurcating phylogenetic trees that depends on two parameters  $\alpha, \gamma$  with  $0 \leq \gamma \leq \alpha \leq 1$ . It generalizes Ford's  $\alpha$ -model by allowing in the recursive construction of trees to add new leaves not only to arcs or to a new root, but also to internal nodes. More specifically, the probability  $P_{\alpha,\gamma,n}^*(T^*)$  of a tree  $T^* \in \mathcal{T}_n^*$  under this model is defined as follows. Let  $n \geq 1$  be any desired number of leaves:

1. Start with the tree  $T_1 \in \mathcal{T}_1$  consisting of a single node labeled 1. Set  $P_{\alpha,\gamma,1}(T_1) = 1$ .
2. For every  $m = 1, \dots, n-1$ , let  $T_{m+1} \in \mathcal{T}_{m+1}$  be obtained by adding a new leaf labeled  $m+1$  to  $T_m$ . Then:
  - If the new leaf is added to an arc  $e$  ending in a leaf,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{1-\alpha}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

- If the new leaf is added to an arc  $e$  ending in an internal node,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{\gamma}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

- If the new leaf is added to a new root,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{\gamma}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

- If the new leaf is added as a child of an internal node  $u$ ,

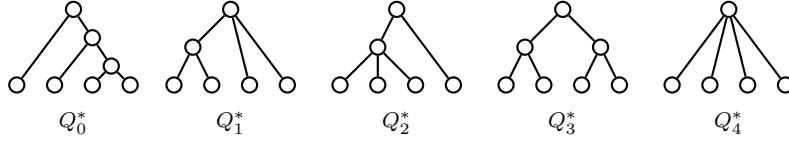
$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{(\deg_{out}(u) - 1)\alpha - \gamma}{m - \alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

3. When the desired number  $n$  of leaves is reached, the probability  $P_{\alpha,\gamma,n}(T_n)$  of the resulting tree  $T_n$  is the one obtained in this way. Then, the probability  $P_{\alpha,\gamma,n}^*(T^*)$  of a given tree  $T^* \in \mathcal{T}_n^*$  is defined as the sum of the probabilities of all phylogenetic trees with that shape:

$$P_{\alpha,\gamma,n}^*(T^*) = \sum_{\pi(T_n)=T^*} P_{\alpha,\gamma,n}(T_n).$$

Notice that if  $\alpha = \gamma$ , this process only produces bifurcating trees and then, for every  $T_n \in \mathcal{BT}_n$ ,  $P_{\alpha,\alpha,n}(T_n) = P'_{\alpha,n}(T_n)$ —the provisional probability of  $T_n$  defined by the recursive application of step 2 in the definition of the  $\alpha$ -model—and, for every  $T_n^* \in \mathcal{BT}_n^*$ ,  $P_{\alpha,\alpha,n}^*(T_n^*) = P_{\alpha,n}^{F,*}(T_n^*)$ .

It turns out that  $P_{\alpha,\gamma,n}$  is not shape invariant in general (see Prop. 1.(b) of [7]), but the corresponding model for trees  $P_{\alpha,\gamma,n}^*$  is sampling consistent by Thm. 2 of *loc. cit.*



**Figure 6:** The 5 trees in  $\mathcal{T}_4^*$ .

Later in this paper we shall need to know the probabilities under  $P_{\alpha,\gamma,4}^*$  of the five different trees in  $\mathcal{T}_4^*$ , described in Figure 6 together with the notations used in this paper to denote them (motivated by Table 1 in the next section). We compute these probabilities in the following lemma, thus providing an example of explicit computation of probabilities also for this model.

**Lemma 6.** *With the notations of Figure 6:*

$$\begin{aligned} P_{\alpha,\gamma,4}^*(Q_0^*) &= \frac{2(1 - \alpha + \gamma)(2(1 - \alpha) + \gamma)}{(3 - \alpha)(2 - \alpha)} \\ P_{\alpha,\gamma,4}^*(Q_1^*) &= \frac{(5(1 - \alpha) + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} \\ P_{\alpha,\gamma,4}^*(Q_2^*) &= \frac{2(1 - \alpha + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} \\ P_{\alpha,\gamma,4}^*(Q_3^*) &= \frac{(1 - \alpha)(2(1 - \alpha) + \gamma)}{(3 - \alpha)(2 - \alpha)} \\ P_{\alpha,\gamma,4}^*(Q_4^*) &= \frac{(2\alpha - \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} \end{aligned}$$

*Proof.* To compute these probabilities, we shall already start with the cherry  $T_2 = (1, 2)$  in  $\mathcal{T}_2$ , which has probability  $P_{\alpha,\gamma,2}(T_2) = 1$ . Every phylogenetic tree in  $\mathcal{T}_3$  is obtained by adding a leaf labeled 3 to  $T_2$ . There are 4 trees in  $\mathcal{T}_3$ : the bifurcating trees  $K^{(i)}$ ,  $i = 1, 2, 3$ , described in Figure 4, and the rooted star  $S_3$ .

- $S_3$  is obtained by adding the leaf 3 to the root of  $T_2$ . Its probability is then

$$P_{\alpha,\gamma,3}(S_3) = \frac{\alpha - \gamma}{2 - \alpha}.$$

- $K^{(1)}$  and  $K^{(2)}$  are obtained by adding the leaf 3 to an arc in  $T_2$  ending in a leaf. Their probability is then

$$P_{\alpha,\gamma,3}(K^{(1)}) = P_{\alpha,\gamma,3}(K^{(2)}) = \frac{1-\alpha}{2-\alpha}.$$

- $K^{(3)}$  is obtained by adding the leaf 3 to a new root. Its probability is then

$$P_{\alpha,\gamma,3}(K^{(3)}) = \frac{\gamma}{2-\alpha}.$$

Let us move finally to  $\mathcal{T}_4^*$ :

- A tree of shape  $Q_4^*$  can only be obtained by adding the leaf 4 to the root of the tree  $S_3$ . Its probability is, then,

$$P_{\alpha,\gamma,4}(Q_4^*) = \frac{2\alpha-\gamma}{3-\alpha} \cdot P_{\alpha,\gamma,3}(S_3) = \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

- A tree of shape  $Q_0^*$  can be obtained by adding the leaf 4 in some tree  $K_3^{(i)}$  either to a new root, to the arc from the root to the other internal node, or to one of the arcs in its cherry. Its probability is, then,

$$\begin{aligned} P_{\alpha,\gamma,4}(Q_0^*) &= \left(2 \cdot \frac{\gamma}{3-\alpha} + 2 \cdot \frac{1-\alpha}{3-\alpha}\right) \sum_{i=1}^3 P_{\alpha,\gamma,3}(K^{(i)}) \\ &= \frac{2(1-\alpha+\gamma)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}. \end{aligned}$$

- A tree of shape  $Q_1^*$  can be obtained by adding the leaf 4 either to one of the three arcs in the tree  $S_3$  or to the root of some tree  $K_3^{(i)}$ . Its probability is, then,

$$\begin{aligned} P_{\alpha,\gamma,4}(Q_1^*) &= 3 \cdot \frac{1-\alpha}{3-\alpha} \cdot P_{\alpha,\gamma,3}(S_3) + \frac{\alpha-\gamma}{3-\alpha} \sum_{i=1}^3 P_{\alpha,\gamma,3}(K^{(i)}) \\ &= \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}. \end{aligned}$$

- A tree of shape  $Q_2^*$  can be obtained by adding the leaf 4 either to a new root in the tree  $S_3$  or to the non-root internal node in some tree  $K_3^{(i)}$ . Its probability is, then,

$$\begin{aligned} P_{\alpha,\gamma,4}(Q_2^*) &= \frac{\gamma}{3-\alpha} \cdot P_{\alpha,\gamma,3}(S_3) + \frac{\alpha-\gamma}{3-\alpha} \sum_{i=1}^3 P_{\alpha,\gamma,3}(K^{(i)}) \\ &= \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}. \end{aligned}$$

- A tree of shape  $Q_3^*$  can only be obtained by adding the leaf 4 to the arc from the root to its only leaf child in some tree  $K_3^{(i)}$ . Its probability is, then,

$$P_{\alpha,\gamma,4}(Q_3^*) = \frac{1-\alpha}{3-\alpha} \sum_{i=1}^3 P_{\alpha,\gamma,3}(K^{(i)}) = \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}.$$

□

□

Notice that, when  $\alpha = \gamma$ ,

$$P_{\alpha,\alpha,4}(Q_3^*) = \frac{1-\alpha}{3-\alpha} = P_{\alpha,4}^{F,*}(Q_3^*)$$

as it should have been expected.

### 3. Rooted quartet indices

Let  $T$  be a phylogenetic tree on a set  $\Sigma$ . For every  $Q \in \mathcal{P}_4(\Sigma)$ , the *rooted quartet on  $Q$  displayed by  $T$*  is the restriction  $T(Q)$  of  $T$  to  $Q$ . A phylogenetic tree  $T \in \mathcal{T}_n$  can contain rooted quartets of five different shapes, namely, those listed in Figure 6. Notice that a bifurcating phylogenetic tree  $T \in \mathcal{BT}_n$  can only contain rooted quartets of two shapes: those denoted by  $Q_0^*$  and  $Q_3^*$  in the aforementioned figure.

We associate to each rooted quartet an *rQI-value*  $q_i$  that increases with the symmetry of the rooted quartet's shape, as measured by means of its number of automorphisms, going from a value  $q_0 = 0$  for the least symmetric tree, the comb  $Q_0^*$ , to a largest value of  $q_4$  for the most symmetric one, the rooted star  $Q_4^*$ ; see Table 1. The specific numerical values can be chosen in order to magnify the differences in symmetry between specific pairs of trees. For instance, one could take  $q_i = i$ , or  $q_i = 2^i$ .

Rooted quartet	$Q_0^*$	$Q_1^*$	$Q_2^*$	$Q_3^*$	$Q_4^*$
# Automorphisms	2	4	6	8	24
<i>rQI</i>	0	$q_1$	$q_2$	$q_3$	$q_4$

**Table 1:** The rooted quartets' *rQI*-values, with  $0 < q_1 < q_2 < q_3 < q_4$ .

Now, for every  $T \in \mathcal{T}(\Sigma)$ , we define its *rooted quartet index*  $rQI(T)$  as the sum of the *rQI*-values of its rooted quartets:

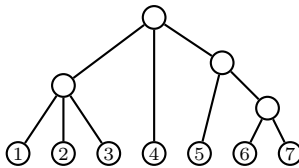
$$\begin{aligned} rQI(T) &= \sum_{Q \in \mathcal{P}_4(\Sigma)} rQI(T(Q)) \\ &= \sum_{i=1}^4 |\{Q \in \mathcal{P}_4(\Sigma) : \pi(T(Q)) = Q_i^*\}| \cdot q_i \end{aligned}$$

In particular, if  $|\Sigma| \leq 3$ , then  $rQI(T) = 0$  for every  $T \in \mathcal{T}(\Sigma)$ . So, *we shall assume henceforth that  $|\Sigma| \geq 4$ .*

It is clear that *rQI* is a *shape index*, in the sense that two phylogenetic trees with the same shape have the same rooted quartet index. It makes sense then to define the rooted quartet index  $rQI(T^*)$  of a tree  $T^* \in \mathcal{T}_n^*$  as the rooted quartet index of any phylogenetic tree of shape  $T^*$ .

**Example 7.** Consider the tree  $T = ((1, 2, 3), 4, (5, (6, 7)))$  depicted in Figure 7. It has: 4 rooted quartets of shape  $Q_0^*$ ; 18 rooted quartets of shape  $Q_1^*$ ; 4 rooted quartets of shape  $Q_2^*$ ; 9 rooted quartets of shape  $Q_3^*$ ; and no rooted quartet of shape  $Q_4^*$ . Therefore

$$rQI(T) = 18q_1 + 4q_2 + 9q_3.$$



**Figure 7:** The tree  $((1, 2, 3), 4, (5, (6, 7)))$ .

**Remark 8.** If we did not take  $q_0 = 0$ , then the resulting index would be

$$\begin{aligned} & \sum_{i=0}^4 q_i \cdot |\{Q \in \mathcal{P}_4(\Sigma) : \pi(T(Q)) = Q_i^*\}| \\ &= q_0 \binom{n}{4} + \sum_{i=1}^4 (q_i - q_0) |\{Q \in \mathcal{P}_4(\Sigma) : \pi(T(Q)) = Q_i^*\}| \end{aligned}$$

which is equivalent (up to the constant addend  $q_0 \binom{n}{4}$ ) to  $rQI$  taking as  $rQI$ -values  $q'_i = q_i - q_0$ .

**Remark 9.** One could also associate other values to the rooted quartet shapes; for instance their Sackin index [25, 26] or their total cophenetic index [20], which measure the imbalance of the rooted quartet's shape, from a smallest value at  $Q_4^*$  to a largest value at  $Q_0^*$ . All results obtained in this paper are easily translated to any other sets of values.

Since a bifurcating tree can only contain rooted quartets of shape  $Q_0^*$  and  $Q_3^*$ , its  $rQI$  index is simply  $q_3$  times its number of rooted quartets of shape  $Q_3^*$ . Therefore, in order to avoid this spurious factor, when dealing only with bifurcating trees we shall use the following alternative *rooted quartet index for bifurcating trees*  $rQIB$ : for every  $T \in \mathcal{BT}(\Sigma)$ ,

$$rQIB(T) = \frac{1}{q_3} rQI(T) = \left| \{Q \in \mathcal{P}_4(\Sigma) : \pi(T(Q)) = Q_3^*\} \right|.$$

The rooted quartet index for bifurcating trees satisfies the following recurrence.

**Lemma 10.** *Let  $T = T_1 \star T_2 \in \mathcal{BT}_n$ , where each  $T_i$  has  $n_i$  leaves. Then,*

$$rQIB(T) = rQIB(T_1) + rQIB(T_2) + \binom{n_1}{2} \cdot \binom{n_2}{2}.$$

*Proof.* For every  $Q \in \mathcal{P}_4([n])$ , there are the following possibilities:

- (1) If  $Q \subseteq L(T_i)$ , for some  $i = 1, 2$ , then  $T(Q) = T_i(Q)$ . Therefore, each  $Q \subseteq L(T_i)$  contributes  $rQIB(T_i)$  to  $rQIB(T)$ .
- (2) If three leaves in  $Q$  belong to one of the subtrees  $T_i$  and the fourth to the other subtree  $T_j$ , then  $T(Q)$  has shape  $Q_0^*$  and thus it does not contribute anything to  $rQIB(T)$ .
- (3) If two leaves in  $Q$  belong to  $T_1$  and the other two to  $T_2$ , then  $T(Q)$  has shape  $Q_3^*$  and thus it contributes 1 to  $rQIB(T)$ . There are  $\binom{n_1}{2} \cdot \binom{n_2}{2}$  such quartets of leaves  $Q$ .

Adding up all these contributions, we obtain the formula in the statement.  $\square$   $\square$

Thus,  $rQIB$  is a *recursive tree shape statistic* in the sense of Matsen [18]. The recurrence in the last lemma implies directly the following explicit formula for  $rQIB$ , which in particular entails that it can be easily computed in time  $O(n)$ , with  $n$  the number of leaves of the tree, by traversing the tree in post-order (cf. the first paragraph in the proof of Proposition 13 below):

**Corollary 11.** *If, for every  $T \in \mathcal{BT}_n$  and for every  $v \in V_{int}(T)$ , we set  $\text{child}(v) = \{v_1, v_2\}$ , then*

$$rQIB(T) = \sum_{v \in V_{int}(T)} \binom{\kappa(v_1)}{2} \cdot \binom{\kappa(v_2)}{2}.$$

Unfortunately,  $rQI$  is not recursive in this sense: there does not exist any family of mappings  $q_m : \mathbb{N}^m \rightarrow \mathbb{R}$ ,  $m \geq 2$ , such that, for every  $T \in \mathcal{T}_n$ , if  $T = T_1 \star \dots \star T_m$ , with each  $T_i$  having  $n_i$  leaves, then

$$rQI(T) = \sum_{i=1}^m rQI(T_i) + q_m(n_1, \dots, n_m).$$



However, next lemma shows that there exists a slightly more involved linear recurrence for  $rQI$ , with its independent term depending on more indices of the trees  $T_i$  than only their numbers of leaves, which still allows its computation in linear time.

For every  $T \in \mathcal{T}_n$ , let  $\Upsilon(T)$  be the number of *non-bifurcating triples* in  $T$  (that is, of restrictions of  $T$  to sets of 3 leaves that have the shape of a rooted star  $S_3$ ; cf. Fig. 1). Notice that if  $T = T_1 \star \cdots \star T_m$  and  $|L(T_i)| = n_i$ , for each  $i = 1, \dots, m$ , then

$$\Upsilon(T) = \sum_{i=1}^m \Upsilon(T_i) + \sum_{1 \leq i_1 < i_2 < i_3 \leq m} n_{i_1} n_{i_2} n_{i_3}$$

and hence

$$\Upsilon(T) = \sum_{v \in V_{int}(T)} \sum_{\{v_1, v_2, v_3\} \subseteq \text{child}(v)} \kappa(v_1) \kappa(v_2) \kappa(v_3).$$

**Lemma 12.** *Let  $T = T_1 \star \cdots \star T_m \in \mathcal{T}_n$ , where each  $T_i$  has  $n_i$  leaves. Then*

$$\begin{aligned} rQI(T) &= \sum_{i=1}^m rQI(T_i) + q_4 \cdot \sum_{1 \leq i_1 < i_2 < i_3 < i_4 \leq m} n_{i_1} n_{i_2} n_{i_3} n_{i_4} \\ &+ q_3 \cdot \sum_{1 \leq i_1 < i_2 \leq m} \binom{n_{i_1}}{2} \binom{n_{i_2}}{2} + q_2 \cdot \sum_{1 \leq i_1 < i_2 \leq m} (n_{i_1} \Upsilon(T_{i_2}) + n_{i_2} \Upsilon(T_{i_1})) \\ &+ q_1 \cdot \sum_{1 \leq i_1 < i_2 < i_3 \leq m} \left( \binom{n_{i_1}}{2} n_{i_2} n_{i_3} + \binom{n_{i_2}}{2} n_{i_1} n_{i_3} + \binom{n_{i_3}}{2} n_{i_1} n_{i_2} \right). \end{aligned}$$

*Proof.* For every  $Q \in \mathcal{P}_4([n])$ , there are the following possibilities:

- (1) If  $Q \subseteq L(T_i)$ , for some  $i$ , then  $T(Q) = T_i(Q)$ . Therefore, each  $Q \subseteq L(T_i)$  contributes  $rQI(T_i)$  to  $rQI(T)$ .
- (2) If 3 leaves, say  $a, b, c$ , in  $Q$  belong to a subtree  $T_i$  and the fourth to another subtree  $T_j$ , then  $T(Q)$ :
  - Has shape  $Q_2^*$  if  $T_i(\{a, b, c\})$  has shape  $S_3$ . For every pair of subtrees  $T_i, T_j$ , there are  $n_j \Upsilon(T_i) + n_i \Upsilon(T_j)$  quartets of leaves  $Q$  of this type, and each one of them contributes  $q_2$  to  $rQI(T)$ .
  - Has shape  $Q_0^*$  if  $T_i(\{a, b, c\})$  is a comb  $K_3$ . These rooted quartets do not contribute anything to  $rQI(T)$ .
- (3) If 2 leaves in  $Q$  belong to a subtree  $T_i$  and the other 2 to another subtree  $T_j$ , then  $T(Q)$  has shape  $Q_3^*$ . For every pair of subtrees  $T_i, T_j$ , there are  $\binom{n_i}{2} \binom{n_j}{2}$  quartets of leaves  $Q$  of this type, and each one of them contributes  $q_3$  to  $rQI(T)$ .
- (4) If 2 leaves in  $Q$  belong to a subtree  $T_i$ , a third leaf to another subtree  $T_j$  and the fourth to a third subtree  $T_k$ , then  $T(Q)$  has shape  $Q_1^*$ . For every triple of subtrees  $T_i, T_j, T_k$ , there are  $\binom{n_i}{2} n_j n_k + \binom{n_j}{2} n_k n_i + \binom{n_k}{2} n_i n_j$  quartets of leaves  $Q$  of this type, and each one of them contributes  $q_1$  to  $rQI(T)$ .
- (5) If each leaf in  $Q$  belongs to a different subtree  $T_i$ , then  $T(Q)$  has shape  $Q_4^*$ . For every four subtrees  $T_i, T_j, T_k, T_l$ , there are  $n_i n_j n_k n_l$  such quartets of leaves  $Q$ , and each one of them contributes  $q_4$  to  $rQI(T)$ .

Adding up all these contributions, we obtain the formula in the statement. □ □

**Proposition 13.** *If  $T \in \mathcal{T}_n$ ,  $rQI(T)$  can be computed in time  $O(n)$ .*

*Proof.* Let  $T$  be a phylogenetic tree in  $\mathcal{T}_n$ . Recall that if a certain mapping  $\phi : V(T) \rightarrow \mathbb{R}$  can be computed in constant time at each leaf of  $T$  and in  $O(\deg(v))$  time at each internal node  $v$  from its value at the children of  $v$ , then the whole vector  $(\phi(v))_{v \in V(T)}$ , and hence also its sum  $\sum_{v \in V(T)} \phi(v)$ , can be computed in  $O(n)$  time by traversing  $T$  in post-order. Indeed, if we

denote by  $m_k$  the number of internal nodes of  $T$  with out-degree  $k$ , then the cost of computing  $(\phi(v))_{v \in V(T)}$  through a post-order traversal of  $T$  is  $O(n + \sum_k m_k \cdot k)$ , and  $\sum_k m_k \cdot k$  is the number of arcs in  $T$ , which is at most  $2n - 2$ . We shall use this remark several times in this proof, and, to begin with, we refer to it to recall that the vector  $(\kappa(v))_{v \in V(T)}$  can be computed in  $O(n)$  time.

Now, in order to simplify the notations, let, for every  $v \in V_{int}(T)$ :

$$\begin{aligned} E_l(v) &= \sum_{\{v_1, \dots, v_l\} \subseteq \text{child}(v)} \kappa(v_1) \cdots \kappa(v_l), \quad l = 2, \dots, \deg(v) \\ F_1(v) &= \sum_{\{v_1, v_2, v_3\} \subseteq \text{child}(v)} \left( \binom{\kappa(v_1)}{2} \kappa(v_2) \kappa(v_3) + \binom{\kappa(v_2)}{2} \kappa(v_1) \kappa(v_3) \right. \\ &\quad \left. + \binom{\kappa(v_3)}{2} \kappa(v_1) \kappa(v_2) \right) \\ F_2(v) &= \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} (\kappa(v_1) \Upsilon(T_{v_2}) + \kappa(v_2) \Upsilon(T_{v_1})) \\ F_3(v) &= \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \binom{\kappa(v_1)}{2} \binom{\kappa(v_2)}{2} \end{aligned}$$

so that

$$\begin{aligned} \Upsilon(T) &= \sum_{v \in V_{int}(T)} E_3(v) \\ rQI(T) &= \sum_{v \in V_{int}(T)} (q_1 F_1(v) + q_2 F_2(v) + q_3 F_3(v) + q_4 E_4(v)) \end{aligned}$$

We want to prove now that each one of the vectors

$$(F_1(v))_{v \in V_{int}(T)}, (F_2(v))_{v \in V_{int}(T)}, (F_3(v))_{v \in V_{int}(T)}, (E_4(v))_{v \in V_{int}(T)}$$

can be computed in  $O(n)$  time, which will clearly entail that  $rQI(T)$  can be computed in  $O(n)$  time.

One of the key ingredients in the proof are the *Newton-Girard formulas* (see, for instance, Section I.2 in [17]): given a (multi)set of numbers  $X = \{x_1, \dots, x_k\}$ , if we let, for every  $l \geq 1$ ,

$$P_l(X) = \sum_{i=1}^k x_i^l, \quad E_l(X) = \sum_{1 \leq i_1 < \dots < i_l \leq k} x_{i_1} \cdots x_{i_l}$$

then

$$E_l(X) = \frac{1}{l!} \begin{vmatrix} P_1(X) & 1 & 0 & \dots & 0 & 0 \\ P_2(X) & P_1(X) & 2 & \dots & 0 & 0 \\ P_3(X) & P_2(X) & P_1(X) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ P_{l-1}(X) & P_{l-2}(X) & P_{l-3}(X) & \dots & P_1(X) & l-1 \\ P_l(X) & P_{l-1}(X) & P_{l-2}(X) & \dots & P_2(X) & P_1(X) \end{vmatrix}$$

If we consider  $l$  as a fixed parameter, every  $P_l(X)$  can be computed in time  $O(k)$  and then this expression for  $E_l(X)$  as an  $l \times l$  determinant allows us also to compute it in time  $O(k)$ .

In particular, if, for every  $v \in V_{int}(V)$ , we consider the multiset  $X_v = \{\kappa(u) : u \in \text{child}(v)\}$ , then every  $E_l(v) = E_l(X_v)$  can be computed in time  $O(\deg(v))$  and hence the whole

vector  $(E_l(v))_{v \in V_{int}(T)}$  can be computed in time  $O(n)$ . In particular,  $(E_3(v))_{v \in V_{int}(T)}$  and  $(E_4(v))_{v \in V_{int}(T)}$  can be computed in linear time.

Then, using the recursion

$$\Upsilon(T_v) = \sum_{v_i \in \text{child}(v)} \Upsilon(T_{v_i}) + E_3(v)$$

we deduce that the whole vector  $(\Upsilon(T_v))_{v \in V_{int}(T)}$  can also be computed in time  $O(n)$ . Now,

$$\begin{aligned} F_2(v) &= \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} (\kappa(v_1)\Upsilon(T_{v_2}) + \kappa(v_2)\Upsilon(T_{v_1})) \\ &= \left( \sum_{v_i \in \text{child}(v)} \kappa(v_i) \right) \left( \sum_{v_j \in \text{child}(v)} \Upsilon(T_{v_j}) \right) - \sum_{v_i \in \text{child}(v)} \kappa(v_i)\Upsilon(T_{v_i}) \\ &= \kappa(v)(\Upsilon(T_v) - E_3(v)) - \sum_{v_i \in \text{child}(v)} \kappa(v_i)\Upsilon(T_{v_i}), \end{aligned}$$

This implies that each  $F_2(v)$  can be computed in time  $O(\deg(v))$  and hence that the whole vector  $(F_2(v))_{v \in V_{int}(T)}$  can be computed in time  $O(n)$ .

Let us focus now on

$$\begin{aligned} F_3(v) &= \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \binom{\kappa(v_1)}{2} \binom{\kappa(v_2)}{2} \\ &= \frac{1}{4} \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \kappa(v_1)^2 \kappa(v_2)^2 + \frac{1}{4} \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \kappa(v_1)\kappa(v_2) \\ &\quad - \frac{1}{4} \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} (\kappa(v_1)^2 \kappa(v_2) + \kappa(v_2)^2 \kappa(v_1)) \end{aligned}$$

In this expression,

$$\sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \kappa(v_1)\kappa(v_2) = E_2(v), \quad \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \kappa(v_1)^2 \kappa(v_2)^2 = E_2(X_v^2),$$

where  $X_v^2 = \{\kappa(u)^2 : u \in \text{child}(v)\}$ , and hence they are computed in time  $O(\deg(v))$ . As far as the subtrahend goes,

$$\begin{aligned} &\sum_{\{v_1, v_2\} \subseteq \text{child}(v)} (\kappa(v_1)^2 \kappa(v_2) + \kappa(v_2)^2 \kappa(v_1)) \\ &= \left( \sum_{v_i \in \text{child}(v)} \kappa(v_i)^2 \right) \left( \sum_{v_j \in \text{child}(v)} \kappa(v_j) \right) - \left( \sum_{v_i \in \text{child}(v)} \kappa(v_i)^3 \right) \end{aligned}$$

and hence it can also be computed in time  $O(\deg(v))$ . Therefore, the whole vector  $(F_3(v))_{v \in V_{int}(T)}$  can be computed in time  $O(n)$ .

Let us consider finally  $F_1(v)$ . We have that

$$\begin{aligned}
F_1(v) &= \sum_{\{v_1, v_2, v_3\} \subseteq \text{child}(v)} \left( \binom{\kappa(v_1)}{2} \kappa(v_2) \kappa(v_3) \right. \\
&\quad \left. + \binom{\kappa(v_2)}{2} \kappa(v_1) \kappa(v_3) + \binom{\kappa(v_3)}{2} \kappa(v_1) \kappa(v_2) \right) \\
&= \frac{1}{2} \sum_{\{v_1, v_2, v_3\} \subseteq \text{child}(v)} \kappa(v_1) \kappa(v_2) \kappa(v_3) (\kappa(v_1) + \kappa(v_2) + \kappa(v_3) - 3) \\
&= \frac{1}{2} \sum_{\{v_1, v_2, v_3\} \subseteq \text{child}(v)} \left( \kappa(v_1)^2 \kappa(v_2) \kappa(v_3) \right. \\
&\quad \left. + \kappa(v_2)^2 \kappa(v_1) \kappa(v_3) + \kappa(v_3)^2 \kappa(v_1) \kappa(v_2) \right) - \frac{3}{2} E_3(v) \\
&= \frac{1}{2} \left( \sum_{\{v_1, v_2, v_3\} \subseteq \text{child}(v)} \kappa(v_1) \kappa(v_2) \kappa(v_3) \right) \left( \sum_{v_i \in \text{child}(v)} \kappa(v_i) \right) \\
&\quad - 2 \left( \sum_{\{v_1, v_2, v_3, v_4\} \subseteq \text{child}(v)} \kappa(v_1) \kappa(v_2) \kappa(v_3) \kappa(v_4) \right) - \frac{3}{2} E_3(v) \\
&= \frac{1}{2} E_3(v) E_1(v) - 2 E_4(v) - \frac{3}{2} E_3(v)
\end{aligned}$$

This expression shows that  $F_1(v)$  can be computed in time  $O(\deg(v))$  and therefore the whole vector  $(F_1(v))_{v \in V_{\text{int}}(T)}$  can be computed in time  $O(n)$ .  $\square$   $\square$

#### 4. Trees with maximum and minimum $rQI$

Let  $n \geq 4$ . In this section we determine which trees in  $\mathcal{T}_n$  and  $\mathcal{BT}_n$  have the largest and smallest corresponding rooted quartet indices. The multifurcating case is easy:

**Theorem 14.** *The minimum value of  $rQI$  in  $\mathcal{T}_n$  is reached exactly at the combs  $K_n$ , and it is 0. The maximum value of  $rQI$  in  $\mathcal{T}_n$  is reached exactly at the rooted star  $S_n$ , and it is  $q_4 \binom{n}{4}$ .*

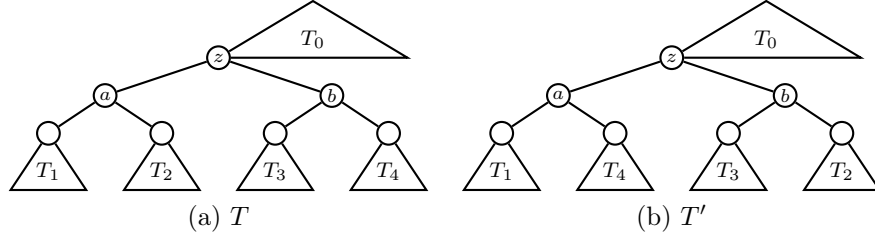
*Proof.* Since the  $rQI$ -value of a rooted quartet goes from 0 to  $q_4$ , we have that  $0 \leq rQI(T) \leq q_4 \binom{n}{4}$ , for every  $T \in \mathcal{T}_n$ . Now, all rooted quartets displayed by a comb  $K_n$  have shape  $Q_0^*$ , and therefore  $rQI(K_n) = 0$ , while all rooted quartets displayed by  $S_n$  have shape  $Q_4^*$ , and therefore  $rQI(S_n) = q_4 \binom{n}{4}$ .

As far as the uniqueness of the trees yielding the maximum and minimum values of  $rQI$  goes, notice that, on the one hand, if  $T$  is not a comb, then it displays some rooted quartet of shape other than  $Q_0^*$ , because it contains either some internal node of out-degree greater than 2, which becomes the root of some multifurcating rooted quartet, or two cherries that determine a rooted quartet of shape  $Q_3^*$ . This implies that if  $T \neq K_n$ , then  $rQI(T) > 0$ . On the other hand, if  $T \neq S_n$ , then its root has some child that is not a leaf and therefore  $T$  displays some rooted quartet of shape other than  $Q_4^*$ , which implies that  $rQI(T) < q_4 \binom{n}{4}$ .  $\square$   $\square$

Therefore, the range of  $rQI$  on  $\mathcal{T}_n$  goes from 0 to  $q_4 \binom{n}{4}$ . This is one order of magnitude wider than the range of the total cophenetic index [20], which, going from 0 to  $\binom{n}{3}$ , was so far the balance index in the literature with the widest range.

We shall now characterize those *bifurcating* phylogenetic trees with largest  $rQI$ , or, equivalently, with largest  $rQIB$ . They turn out to be exactly the maximally balanced trees, as defined at the end of Subsection 2.1. The proof is similar to that of the characterization of the bifurcating phylogenetic trees with minimum total cophenetic index provided in Section 4 of [20].

**Lemma 15.** Let  $T \in \mathcal{BT}_n$  be the bifurcating phylogenetic tree depicted in Fig 8.(a). For every  $i = 1, 2, 3, 4$ , let  $n_i = |L(T_i)|$ , and assume that  $n_1 > n_3$  and  $n_2 > n_4$ . Then,  $rQIB(T)$  is not maximum in  $\mathcal{BT}_n$ .



**Figure 8:** (a) The tree  $T$  in the statement of Lemma 15. (b) The tree  $T'$  in the proof of Lemma 15.

*Proof.* Let  $T'$  be the tree obtained from  $T$  by interchanging  $T_2$  and  $T_4$ ; see Fig 8.(b). We shall prove that  $rQIB(T') > rQIB(T)$ .

Let  $\Sigma_z$  be the set of labels of  $T_z$ , which is also the set of labels of  $T'_z$ . To simplify the language, we shall understand the common subtree  $T_0$  of  $T$  and  $T'$  as a phylogenetic tree on  $([n] \setminus \Sigma_z) \cup \{z\}$ . Then, for every  $Q = \{a, b, c, d\} \in \mathcal{P}_4([n])$ :

- If  $Q \cap \Sigma_z = \emptyset$ , then  $T(Q) = T'(Q) = T_0(Q)$ .
- If  $Q \cap \Sigma_z$  is a single label, say  $d$ , then  $T(Q) = T'(Q) = T_0(\{a, b, c, z\})$ .
- If  $Q \cap \Sigma_z$  consists of two labels, say  $c, d$ , then  $T(Q) = T'(Q)$ . More specifically:  $T(Q) = T'(Q) = ((a, b), (c, d))$  when  $T_0(\{a, b, z\}) = ((a, b), z)$ ;  $T(Q) = T'(Q) = (a, (b, (c, d)))$  when  $T_0(\{a, b, z\}) = (a, (b, z))$ ; and  $T(Q) = T'(Q) = (b, (a, (c, d)))$  when  $T_0(\{a, b, z\}) = (b, (a, z))$ .
- If  $Q \cap \Sigma_z$  consists of three labels, then  $T(Q)$  and  $T'(Q)$  are both combs.

Therefore,  $T(Q)$  and  $T'(Q)$  can only be different when  $Q \subseteq \Sigma_z$ , in which case  $T(Q) = T_z(Q)$  and  $T'(Q) = T'_z(Q)$ . This implies that

$$rQIB(T') - rQIB(T) = rQIB(T'_z) - rQIB(T_z).$$

Now, to compute the difference in the right hand side of this equality, we apply Lemma 10:

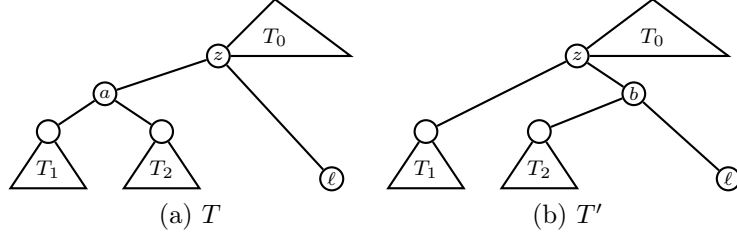
$$\begin{aligned} rQIB(T_z) &= rQIB(T_1) + rQIB(T_2) + rQIB(T_3) + rQIB(T_4) \\ &\quad + \binom{n_1}{2} \binom{n_2}{2} + \binom{n_3}{2} \binom{n_4}{2} + \binom{n_1 + n_2}{2} \binom{n_3 + n_4}{2} \\ rQIB(T'_z) &= rQIB(T_1) + rQIB(T_4) + rQIB(T_2) + rQIB(T_3) \\ &\quad + \binom{n_1}{2} \binom{n_4}{2} + \binom{n_2}{2} \binom{n_3}{2} + \binom{n_1 + n_4}{2} \binom{n_2 + n_3}{2} \end{aligned}$$

and hence

$$rQIB(T'_z) - rQIB(T_z) = \frac{1}{2}(n_1 - n_3)(n_2 - n_4)(n_1 n_3 + n_2 n_4) > 0$$

because  $n_1 > n_3$  and  $n_2 > n_4$  by assumption. □ □

**Lemma 16.** Let  $T \in \mathcal{BT}_n$  be a bifurcating phylogenetic tree containing a leaf whose sibling has at least 3 descendant leaves. Then,  $rQIB(T)$  is not maximum in  $\mathcal{BT}_n$ .



**Figure 9:** (a) The tree  $T$  in the statement of Lemma 16. (b) The tree  $T'$  in the proof of Lemma 16.

*Proof.* Assume that the tree  $T \in \mathcal{BT}_n$  in the statement is the one depicted in Fig 9.(a), with  $\ell$  a leaf such that the subtree  $T_a$  rooted at its sibling  $a$  has  $|L(T_a)| \geq 3$ . Let  $n_1 = |L(T_1)|$  and  $n_2 = |L(T_2)|$  and assume  $n_1 \geq n_2$ : then, since  $n_1 + n_2 \geq 3$ ,  $n_1 \geq 2$ . Let then  $T'$  be the tree depicted in Fig 9.(b): we shall prove that  $rQIB(T') > rQIB(T)$ . Reasoning as in the proof of the last lemma, we deduce that

$$rQIB(T') - rQIB(T) = rQIB(T'_z) - rQIB(T_z).$$

Now, using Lemma 10, we have that

$$\begin{aligned} rQIB(T_z) &= rQIB(T_a) = rQIB(T_1) + rQIB(T_2) + \binom{n_1}{2} \cdot \binom{n_2}{2} \\ rQIB(T'_z) &= rQIB(T_1) + rQIB(T_b) + \binom{n_1}{2} \cdot \binom{n_2+1}{2} \\ &= rQIB(T_1) + rQIB(T_2) + \binom{n_1}{2} \cdot \binom{n_2+1}{2} \end{aligned}$$

and therefore

$$rQIB(T'_z) - rQIB(T_z) = n_2 \binom{n_1}{2} > 0$$

as we wanted to prove.  $\square$   $\square$

**Theorem 17.** *For every  $T \in \mathcal{BT}_n$ ,  $rQIB(T)$  is maximum in  $\mathcal{BT}_n$  if, and only if,  $T$  is maximally balanced.*

*Proof.* Assume that  $rQIB(T)$  is maximum in  $\mathcal{BT}_n$  but that  $T \in \mathcal{BT}_n$  is not maximally balanced, and let us reach a contradiction. Let  $z$  be a non-balanced internal node in  $T$  such that all its proper descendant internal nodes are balanced, and let  $a$  and  $b$  be its children, with  $\kappa(a) \geq \kappa(b) + 2$ .

If  $b$  is a leaf, then, by Lemma 16,  $rQIB(T)$  cannot be maximum in  $\mathcal{BT}_n$ . Therefore,  $a$  and  $b$  are internal, and hence balanced. Let  $v_1, v_2$  be the children of  $a$ ,  $v_3, v_4$  the children of  $b$ , and  $n_i = \kappa(v_i)$ , for  $i = 1, 2, 3, 4$ . Without any loss of generality, we shall assume that  $n_1 \geq n_2$  and  $n_3 \geq n_4$ . Then, since  $a$  and  $b$  are balanced,  $n_1 = n_2$  or  $n_2 + 1$  and  $n_3 = n_4$  or  $n_4 + 1$ . Then,  $n_1 + n_2 = \kappa(a) \geq \kappa(b) + 2 = n_3 + n_4 + 2$  implies that  $n_1 > n_3$ .

Now, by Lemma 15, since by assumption  $rQIB(T)$  is maximum on  $\mathcal{BT}_n$ , it must happen that  $n_1 > n_3 \geq n_4 \geq n_2$ . This forbids the equality  $n_1 = n_2$ , and hence  $n_1 - 1 = n_2 = n_3 = n_4$ . But this contradicts that  $n_1 + n_2 \geq n_3 + n_4 + 2$ .

This implies that a non maximally balanced tree in  $\mathcal{BT}_n$  cannot have maximum  $rQIB$ , and therefore the maximum  $rQIB$  in  $\mathcal{BT}_n$  is reached at the maximally balanced trees, which have all the same shape and hence the same  $rQIB$  index.  $\square$   $\square$

So, the only bifurcating trees with maximum  $rQIB$  (and hence with maximum  $rQI$ ) are the maximally balanced. This maximum value of  $rQIB$  on  $\mathcal{BT}_n$  is given by the following recurrence.

**Lemma 18.** *For every  $n$ , let  $b_n$  be the maximum of  $rQIB$  on  $\mathcal{BT}_n$ . Then,  $b_1 = b_2 = b_3 = 0$  and*

$$b_n = b_{\lceil n/2 \rceil} + b_{\lfloor n/2 \rfloor} + \binom{\lceil n/2 \rceil}{2} \cdot \binom{\lfloor n/2 \rfloor}{2}, \quad \text{for } n \geq 4.$$

*Proof.* This recurrence for  $b_n$  is a direct consequence of Lemma 10 and the fact that the root of a maximally balanced tree in  $\mathcal{BT}_n$  is balanced and the subtrees rooted at their children are maximally balanced. □ □

The sequence  $b_n$  seems to be new, in the sense that it has no relation with any sequence previously contained in Sloane's *On-Line Encyclopedia of Integer Sequences* [27]. Its values for  $n = 4, \dots, 20$  are

$$1, 3, 9, 19, 38, 64, 106, 162, 243, 343, 479, 645, 860, 1110, 1424, 1790, 2237.$$

It is easy to prove, using the Master theorem for solving recurrences [9, Thm. 4.1], that  $b_n$  grows asymptotically in  $O(n^4)$ . Moreover, it is easy to compute  $b_{2^n}$  from this recurrence, yielding

$$b_{2^n} = \left( \frac{4}{7(2^n - 3)} + \frac{3}{7} \right) \binom{2^n}{4}.$$

In particular,  $b_{2^n} / \binom{2^n}{4} \xrightarrow{n \rightarrow \infty} 3/7$ , which is in agreement with the probability of the fully symmetric rooted quartet  $Q_3^*$  under the  $\beta$ -model when  $\beta \rightarrow \infty$ ; cf. Section 4.1 in [2].

**Remark 19.** When the range of values of a shape index  $I$  grows with the number of leaves  $n$  of the phylogenetic trees, as it is the case with  $rQI$  and  $rQIB$ , it makes no sense to compare directly its value on two trees with different numbers of leaves. To overcome this drawback, one usually *normalizes* the index, so that its range becomes independent on  $n$ . A suitable way to do that is to use the generic affine transformation

$$\tilde{I}(T) = \frac{I(T) - \min I(\mathcal{T}_n)}{\max I(\mathcal{T}_n) - \min I(\mathcal{T}_n)}$$

where  $n$  stands for the number of leaves of the tree  $T$  and  $I(\mathcal{T}_n)$  denotes the set of values of  $I$  on  $\mathcal{T}_n$ . In this way, for every number of leaves, the minimum value of the normalized index is always 0 and the maximum value is always 1.

As to our  $rQI$ , its minimum value is always 0, but its maximum depends on whether we are considering multifurcating or bifurcating trees. Therefore, we propose two normalized versions of this index:

- On  $\mathcal{T}_n$ ,  $\widetilde{rQI}(T) = rQI(T) / (q_4 \binom{n}{4})$ .
- On  $\mathcal{BT}_n$ ,  $\widetilde{rQIB}(T) = rQIB(T) / b_n$ , with  $b_n$  computed by means of the recurrence given in Lemma 18.

## 5. The expected value and the variance of $rQI$

Let  $P_n$  be a probabilistic model of phylogenetic trees and  $rQI_n$  the random variable that chooses a phylogenetic tree  $T \in \mathcal{T}_n$  with probability distribution  $P_n$  and computes  $rQI(T)$ . In this section we are interested in obtaining expressions for the expected value  $E_P(rQI_n)$  and the variance  $Var_P(rQI_n)$  of  $rQI_n$  under suitable models  $P_n$ .

Next lemma shows that, to compute these values, we can restrict ourselves to work with unlabeled trees. Let  $P_n^*$  the probabilistic model of trees induced by  $P_n$  and  $rQI_n^*$  the random variable that chooses a tree  $T^* \in \mathcal{T}_n^*$  with probability distribution  $P_n^*$  and computes  $rQI(T^*)$ , defined as  $rQI(T)$  for some phylogenetic tree  $T$  of shape  $T^*$ .

**Lemma 20.** For every  $n \geq 1$ , the distributions of  $rQI_n$  and  $rQI_n^*$  are the same. In particular, their expected values and their variances are the same.

*Proof.* Let  $f_{rQI_n}$  and  $f_{rQI_n^*}$  be the probability density functions of the discrete random variables  $rQI_n$  and  $rQI_n^*$ , respectively. Then, for every  $x_0 \in \mathbb{R}$ ,

$$\begin{aligned} f_{rQI_n}(x_0) &= \sum_{\substack{T \in \mathcal{T}_n \\ rQI(T)=x_0}} P_n(T) = \sum_{\substack{T^* \in \mathcal{T}_n^* \\ rQI(T^*)=x_0}} \sum_{\substack{T \in \mathcal{T}_n \\ \pi(T)=T^*}} P_n(T) \\ &= \sum_{\substack{T^* \in \mathcal{T}_n^* \\ rQI(T^*)=x_0}} P_n^*(T^*) = f_{rQI_n^*}(x_0) \end{aligned}$$

□

□

**Proposition 21.** If  $P_n^*$  is sampling consistent, then

$$E_P(rQI_n) = \binom{n}{4} \sum_{i=1}^4 q_i P_4^*(Q_i^*).$$

*Proof.* By Lemma 20,  $E_P(rQI_n)$  is equal to the expected value  $E_{P^*}(rQI_n^*)$  of  $rQI_n^*$  under  $P_n^*$ , which can be computed as follows:

$$\begin{aligned} E_{P^*}(rQI_n^*) &= \sum_{T^* \in \mathcal{T}_n^*} rQI(T^*) P_n^*(T^*) \\ &= \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{i=1}^4 q_i |\{Q \in \mathcal{P}_4(L(T^*)) : T^*(Q) = Q_i^*\}| \right) P_n^*(T^*) \\ &= \binom{n}{4} \sum_{i=1}^4 q_i \sum_{T^* \in \mathcal{T}_n^*} \frac{|\{Q \in \mathcal{P}_4(L(T^*)) : T^*(Q) = Q_i^*\}|}{\binom{n}{4}} P_n^*(T^*) \\ &= \binom{n}{4} \sum_{i=1}^4 q_i P_4^*(Q_i^*) \end{aligned}$$

because, for every  $i = 1, \dots, 4$ ,

$$\sum_{T^* \in \mathcal{T}_n^*} \frac{|\{Q \in \mathcal{P}_4(L(T^*)) : T^*(Q) = Q_i^*\}|}{\binom{n}{4}} P_n^*(T^*) = P_4^*(Q_i^*)$$

by the sampling consistency of  $P_n^*$ . □

□

This expression for  $E_P(rQI_n)$  should not be surprising: by the sampling consistency property, for each  $i = 1, \dots, 4$ , the expected number of rooted quartets of shape  $Q_i^*$  in a tree of  $n$  leaves is  $\binom{n}{4} P_4^*(Q_i^*)$  and their weight in  $rQI$  value is  $q_i$ .

The  $\alpha$ - $\gamma$ -model for unlabeled trees  $P_{\alpha, \gamma, n}^*$  is sampling consistent [7, Prop. 12]. Therefore, applying the last proposition using the distribution  $P_{\alpha, \gamma, 4}^*$  on  $\mathcal{T}_4^*$  given in Lemma 6, we have the following result.

**Corollary 22.** Let  $P_{\alpha, \gamma, n}$  be the  $\alpha$ - $\gamma$ -model of phylogenetic trees, with  $0 \leq \gamma \leq \alpha \leq 1$ . Then

$$\begin{aligned} E_{P_{\alpha, \gamma}}(rQI_n) &= \binom{n}{4} \left( \frac{(2\alpha - \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} \cdot q_4 + \frac{(1 - \alpha)(2(1 - \alpha) + \gamma)}{(3 - \alpha)(2 - \alpha)} \cdot q_3 \right. \\ &\quad \left. + \frac{2(1 - \alpha + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} \cdot q_2 + \frac{(5(1 - \alpha) + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} \cdot q_1 \right). \end{aligned}$$



If  $P_n$  is a probabilistic model of bifurcating phylogenetic trees, so that  $P_4^*(Q_1^*) = P_4^*(Q_2^*) = P_4^*(Q_4^*) = 0$ , then the expression in Prop. 21 becomes

$$E_P(rQI_n) = \binom{n}{4} q_3 P_4^*(Q_3^*).$$

Taking  $q_3 = 1$ , we obtain the following results.

**Corollary 23.** *If  $P_n$  is a probabilistic model of bifurcating phylogenetic trees such that  $P_n^*$  is sampling consistent, then*

$$E_P(rQIB_n) = \binom{n}{4} P_4^*(Q_3^*).$$

Since the  $\beta$  and  $\alpha$ -models of bifurcating (unlabeled) trees are sampling consistent, this corollary together with the probabilities of  $Q_3^* \in \mathcal{BT}_4^*$  under these models given in Lemmas 4 and 5, respectively, entail the following result.

**Corollary 24.** *Let  $P_{\beta,n}^A$  be Aldous'  $\beta$ -model for bifurcating phylogenetic trees, with  $\beta \in (-2, \infty)$ , and let  $P_{\alpha,n}^F$  be Ford's  $\alpha$ -model for bifurcating phylogenetic trees, with  $\alpha \in [0, 1]$ . Then:*

$$E_{P_{\beta,n}^A}(rQIB_n) = \frac{3\beta + 6}{7\beta + 18} \binom{n}{4}, \quad E_{P_{\alpha,n}^F}(rQIB_n) = \frac{1 - \alpha}{3 - \alpha} \binom{n}{4}.$$

It is easy to check that  $E_{P_{\alpha,n}^F}(rQIB_n)$  agrees with  $E_{P_{\alpha,\gamma}}(rQI_n)$  (up to the factor  $q_3$ ) when  $\alpha = \gamma$ .

In particular, under the Yule model, which corresponds to  $\alpha = 0$  or  $\beta = 0$ , and the uniform model, which corresponds to  $\alpha = 1/2$  or  $\beta = -3/2$ , the expected values of  $rQIB_n$  are, respectively,

$$E_Y(rQIB_n) = \frac{1}{3} \binom{n}{4}, \quad E_U(rQIB_n) = \frac{1}{5} \binom{n}{4}.$$

Let us deal now with the variance of  $rQI_n$ . To simplify the notations, for every  $k = 5, 6, 7, 8$ , for every  $T^* \in \mathcal{T}_k^*$  and for every  $i, j \in \{1, 2, 3, 4\}$ , let

$$\begin{aligned} \Theta_{i,j}(T^*) &= |\{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 : \\ &\quad Q \cup Q' = L(T^*), T^*(Q) = Q_i^*, T^*(Q') = Q_j^*\}| \\ &= |\{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 : \\ &\quad |Q \cap Q'| = 8 - k, T^*(Q) = Q_i^*, T^*(Q') = Q_j^*\}|. \end{aligned}$$

Notice that  $\Theta_{i,j}(T^*) = \Theta_{j,i}(T^*)$ .

**Proposition 25.** *If  $P_n^*$  is sampling consistent, then*

$$\begin{aligned} \text{Var}_P(rQI_n) &= \binom{n}{4} \sum_{i=1}^4 q_i^2 P_4^*(Q_i^*) - \binom{n}{4}^2 \left( \sum_{i=1}^4 q_i P_4^*(Q_i^*) \right)^2 \\ &\quad + \sum_{i=1}^4 \sum_{j=1}^4 q_i q_j \left( \sum_{k=5}^8 \binom{n}{k} \sum_{T^* \in \mathcal{T}_k^*} \Theta_{i,j}(T^*) P_k^*(T^*) \right). \end{aligned}$$

*Proof.* Since, by Lemma 20,  $\text{Var}_P(rQI_n) = \text{Var}_{P^*}(rQI_n^*)$ , we shall compute the latter using the formula  $\text{Var}_{P^*}(rQI_n^*) = E_{P^*}(rQI_n^{*2}) - E_{P^*}(rQI_n^*)^2$ , and therefore we need to compute  $E_{P^*}(rQI_n^{*2})$ .

For every  $T^* \in \mathcal{T}_n^*$ , for every  $Q_i^* \in \mathcal{T}_4^*$  and for every  $Q \in \mathcal{P}_4(L(T^*))$ , set

$$\delta(Q; Q_i^*; T^*) = \begin{cases} 1 & \text{if } T^*(Q) = Q_i^* \\ 0 & \text{if } T^*(Q) \neq Q_i^* \end{cases}$$

Then:

$$\begin{aligned}
E_{P^*}(rQI_n^{*2}) &= \sum_{T^* \in \mathcal{T}_n^*} rQI^*(T^*)^2 P_n^*(T^*) \\
&= \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{Q \in \mathcal{P}_4(L(T^*))} \sum_{i=1}^4 q_i \delta(Q; Q_i^*; T^*) \right)^2 P_n^*(T^*) \\
&= \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{Q \in \mathcal{P}_4(L(T^*))} \sum_{i=1}^4 q_i^2 \delta(Q; Q_i^*; T^*)^2 \right) P_n^*(T^*) \\
&\quad + \sum_{T^* \in \mathcal{T}_n^*} \left[ \sum_{\substack{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 \\ Q \neq Q'}} \sum_{(i, j) \in [4]^2} q_i q_j \delta(Q; Q_i^*; T^*) \delta(Q'; Q_j^*; T^*) \right] P_n^*(T^*)
\end{aligned}$$

Now, since  $\delta(Q; Q_i^*; T^*)^2 = \delta(Q; Q_i^*; T^*)$ ,

$$\begin{aligned}
S_1 &:= \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{Q \in \mathcal{P}_4(L(T^*))} \sum_{i=1}^4 q_i^2 \delta(Q; Q_i^*; T^*)^2 \right) P_n^*(T^*) \\
&= \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{Q \in \mathcal{P}_4(L(T^*))} \sum_{i=1}^4 q_i^2 \delta(Q; Q_i^*; T^*) \right) P_n^*(T^*) \\
&= \sum_{i=1}^4 \left( q_i^2 \sum_{T^* \in \mathcal{T}_n^*} |\{Q \in \mathcal{P}_4(L(T^*)) : T^*(Q) = Q_i^*\}| \cdot P_n^*(T^*) \right) \\
&= \binom{n}{4} \sum_{i=1}^4 \left( q_i^2 \sum_{T^* \in \mathcal{T}_n^*} \frac{|\{Q \in \mathcal{P}_4(L(T^*)) : T^*(Q) = Q_i^*\}|}{\binom{n}{4}} P_n^*(T^*) \right) \\
&= \binom{n}{4} \sum_{i=1}^4 q_i^2 P_4^*(Q_i^*)
\end{aligned}$$

by the sampling consistency of  $P_n^*$ .

As far as the second addend in the previous expression for  $E_{P^*}(rQI_n^{*2})$  goes, we have

$$\begin{aligned}
S_2 &:= \sum_{T^* \in \mathcal{T}_n^*} \sum_{\substack{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 \\ Q \neq Q'}} \left( \sum_{(i, j) \in [4]^2} q_i q_j \delta(Q; Q_i^*; T^*) \delta(Q'; Q_j^*; T^*) \right) P_n^*(T^*) \\
&= \sum_{(i, j) \in [4]^2} q_i q_j \left[ \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{k=0}^3 \sum_{\substack{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 \\ |Q \cap Q'| = k}} \delta(Q; Q_i^*; T^*) \delta(Q'; Q_j^*; T^*) \right) P_n^*(T^*) \right] \\
&= \sum_{(i, j) \in [4]^2} q_i q_j \left[ \sum_{k=0}^3 \sum_{T^* \in \mathcal{T}_n^*} |\{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 : |Q \cap Q'| = k, \right. \\
&\quad \left. T^*(Q) = Q_i^*, T^*(Q') = Q_j^*\}| \cdot P_n^*(T^*) \right]
\end{aligned}$$

Now notice that, for every  $k = 0, \dots, 3$ ,

$$\begin{aligned}
& \sum_{T^* \in \mathcal{T}_n^*} |\{(Q, Q') \in \mathcal{P}_4(L(T^*))^2 : |Q \cap Q'| = k, T^*(Q) = Q_i^*, T^*(Q') = Q_j^*\}| P_n^*(T^*) \\
&= \sum_{T^* \in \mathcal{T}_n^*} \left( \sum_{T_{8-k}^* \in \mathcal{T}_{8-k}^*} |\{X \in \mathcal{P}_{8-k}(L(T^*)) : T^*(X) = T_{8-k}^*\}| \right. \\
&\quad \cdot |\{(Q, Q') \in \mathcal{P}_4(L(T_{8-k}^*))^2 : |Q \cap Q'| = k, T_{8-k}^*(Q) = Q_i^*, T_{8-k}^*(Q') = Q_j^*\}| \Big) P_n^*(T^*) \\
&= \sum_{T_{8-k}^* \in \mathcal{T}_{8-k}^*} |\{(Q, Q') \in \mathcal{P}_4(L(T_{8-k}^*))^2 : |Q \cap Q'| = k, T_{8-k}^*(Q) = Q_i^*, T_{8-k}^*(Q') = Q_j^*\}| \\
&\quad \cdot \binom{n}{8-k} \sum_{T^* \in \mathcal{T}_n^*} \frac{|\{X \in \mathcal{P}_{8-k}(L(T^*)) : T^*(X) = T_{8-k}^*\}|}{\binom{n}{8-k}} P_n^*(T^*) \\
&= \binom{n}{8-k} \sum_{T_{8-k}^* \in \mathcal{T}_{8-k}^*} |\{(Q, Q') \in \mathcal{P}_4(L(T_{8-k}^*))^2 : |Q \cap Q'| = k, \\
&\quad T_{8-k}^*(Q) = Q_i^*, T_{8-k}^*(Q') = Q_j^*\}| P_{8-k}^*(T_{8-k}^*) \\
&= \binom{n}{8-k} \sum_{T_{8-k}^* \in \mathcal{T}_{8-k}^*} |\{(Q, Q') \in \mathcal{P}_4(L(T_{8-k}^*))^2 : Q \cup Q' = L(T_{8-k}^*), \\
&\quad T_{8-k}^*(Q) = Q_i^*, T_{8-k}^*(Q') = Q_j^*\}| P_{8-k}^*(T_{8-k}^*) \\
&= \binom{n}{8-k} \sum_{T_{8-k}^* \in \mathcal{T}_{8-k}^*} \Theta_{i,j}(T_{8-k}^*) P_{8-k}^*(T_{8-k}^*)
\end{aligned}$$

again by the sampling consistency of  $P_n^*$ . Therefore,

$$\begin{aligned}
S_2 &= \sum_{(i,j) \in [4]^2} q_i q_j \left( \sum_{k=0}^3 \binom{n}{8-k} \sum_{T_{8-k}^* \in \mathcal{T}_{8-k}^*} \Theta_{i,j}(T_{8-k}^*) P_{8-k}^*(T_{8-k}^*) \right) \\
&= \sum_{(i,j) \in [4]^2} q_i q_j \left( \sum_{k=5}^8 \binom{n}{k} \sum_{T^* \in \mathcal{T}_k^*} \Theta_{i,j}(T^*) P_k^*(T^*) \right)
\end{aligned}$$

The formula in the statement is then obtained by writing  $Var_{P^*}(IQ_n^*)$  as  $S_1 + S_2 - E_{P^*}(IQ_n^*)^2$  and using the expression for  $E_{P^*}(IQ_n^*) = E_P(IQ_n)$  given in Proposition 21.  $\square$   $\square$

Again, if  $P_n$  is a probabilistic model of bifurcating phylogenetic trees, so that  $P_4^*(Q_1^*) = P_4^*(Q_2^*) = P_4^*(Q_4^*) = 0$ , then, taking  $q_3 = 1$ , this proposition implies that

$$\begin{aligned}
Var_P(rQIB_n) &= \binom{n}{4} P_4^*(Q_3^*) - \binom{n}{4}^2 P_4^*(Q_3^*)^2 \\
&\quad + \sum_{k=5}^8 \binom{n}{k} \left( \sum_{T^* \in \mathcal{BT}_k^*} \Theta_{3,3}(T^*) P_k^*(T^*) \right)
\end{aligned}$$

In this bifurcating case, the figures  $\Theta_{3,3}(T^*)$  appearing in this expression can be easily computed by hand: they are provided in Table 3 in the Appendix A.2. We obtain then the following result.

**Corollary 26.** *If  $P_n$  is a probabilistic model of bifurcating phylogenetic trees such that  $P_n^*$  is*

sampling consistent, then, with the notations for trees given in Table 3 in the Appendix A.2,

$$\begin{aligned}
Var_P(rQIB_n) &= \binom{n}{4} P_4^*(Q_3^*) - \binom{n}{4}^2 P_4^*(Q_3^*)^2 \\
&+ 6 \binom{n}{5} P_5^*(B_{5,3}^*) + \binom{n}{6} (18P_6^*(B_{6,4}^*) + 6P_6^*(B_{6,5}^*) + 36P_6^*(B_{6,6}^*)) \\
&+ \binom{n}{7} (8P_7^*(B_{7,8}^*) + 24P_7^*(B_{7,9}^*) + 36P_7^*(B_{7,10}^*) + 36P_7^*(B_{7,11}^*)) \\
&+ \binom{n}{8} (2P_8^*(B_{8,13}^*) + 6P_8^*(B_{8,14}^*) + 12P_8^*(B_{8,15}^*) + 14P_8^*(B_{8,16}^*) \\
&\quad + 18P_8^*(B_{8,17}^*) + 36P_8^*(B_{8,21}^*) + 36P_8^*(B_{8,22}^*) + 38P_8^*(B_{8,23}^*))
\end{aligned}$$

Proposition 25 and Corollary 26 reduce the computation of  $Var_P(rQI_n)$  or  $Var_P(rQIB_n)$  to the explicit knowledge of  $P_l^*$  for  $l = 4, 5, 6, 7, 8$ . In particular, they allow to obtain explicit formulas for the variance of  $rQIB_n$  under the  $\alpha$  and the  $\beta$ -models, and for the variance of  $rQI_n$  under the  $\alpha$ - $\gamma$ -model.

As far as the bifurcating case goes, on the one hand, the probabilities under the  $\alpha$ -model of the trees appearing explicitly in the formula for the variance of  $rQIB_n$  in Corollary 26 are those given in Table 4 in the Appendix A.2 (they are explicitly computed in the Supplementary Material of [10]). Plugging them in the formula given in Corollary 26 above, we obtain the following result.

**Corollary 27.** *Under the  $\alpha$ -model,*

$$\begin{aligned}
Var_{P_\alpha^F}(rQIB_n) &= \binom{n}{4} \frac{1-\alpha}{3-\alpha} - \binom{n}{4}^2 \frac{(1-\alpha)^2}{(3-\alpha)^2} + 12 \binom{n}{5} \frac{1-\alpha}{4-\alpha} \\
&+ \binom{n}{6} \frac{6(1-\alpha)(112-89\alpha+15\alpha^2)}{(5-\alpha)(4-\alpha)(3-\alpha)} \\
&+ \binom{n}{7} \frac{20(1-\alpha)(74-63\alpha+7\alpha^2)}{(6-\alpha)(5-\alpha)(3-\alpha)} \\
&+ \binom{n}{8} \frac{10(1-\alpha)(506-539\alpha+112\alpha^2-7\alpha^3)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}
\end{aligned}$$

The leading term in  $n$  of  $Var_{P_\alpha^F}(rQIB_n)$  is then

$$\frac{(1-\alpha)(2\alpha+1)}{84(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)^2} \cdot n^8.$$

On the other hand, the probabilities under the  $\beta$ -model of the same trees are given in Table 5 in the Appendix A.2, yielding the following result.

**Corollary 28.** *Under the  $\beta$ -model,*

$$\begin{aligned}
Var_{P_\beta^A}(rQIB_n) &= \binom{n}{4} \frac{3(\beta+2)}{7\beta+18} - \binom{n}{4}^2 \frac{9(\beta+2)^2}{(7\beta+18)^2} + 12 \binom{n}{5} \frac{\beta+2}{3\beta+8} \\
&+ 90 \binom{n}{6} \frac{(\beta+2)(41\beta^2+238\beta+336)}{(31\beta^2+194\beta+300)(7\beta+18)} \\
&+ 60 \binom{n}{7} \frac{(\beta+2)(9\beta^2+53\beta+74)}{(\beta+3)(3\beta+10)(7\beta+18)} \\
&+ 630 \binom{n}{8} \frac{(\beta+2)(127\beta^4+1637\beta^3+7788\beta^2+16084\beta+12144)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}
\end{aligned}$$

So, the leading term in  $n$  of  $Var_{P_\beta^A}(rQIB_n)$  is

$$\frac{(\beta+2)(2\beta^2+9\beta+12)}{2(7\beta+18)^2(127\beta^3+1383\beta^2+4958\beta+5880)} \cdot n^8.$$

When  $\alpha = 0$  or  $\beta = 0$ , which correspond to the Yule model, both formulas for the variance of  $rQIB_n$  reduce to

$$Var_Y(rQIB_n) = \binom{n}{4} \frac{5n^4 + 30n^3 + 118n^2 + 408n + 630}{33075}.$$

In the Appendix A.1 we give an independent derivation of this formula, which provides extra evidence of the correctness of all these computations.

As far as the uniform model goes, when  $\alpha = 1/2$  or  $\beta = 0$ , both formulas yield

$$Var_U(rQIB_n) = \binom{n}{4} \frac{4(2n-1)(2n+1)(2n+3)(2n+5)}{225225}.$$

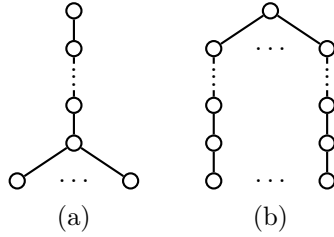
Finally, as far as the  $\alpha$ - $\gamma$ -model goes, we have written a set of Python scripts that compute all  $\Theta_{i,j}(T^*)$ ,  $i, j = 1, 2, 3, 4$ , as well as  $P_{\alpha,\gamma,k}^*(T^*)$  for every  $T^* \in \mathcal{T}_k^*$ ,  $k = 5, 6, 7, 8$ , and combine all these data into an explicit formula for  $Var_{P_{\alpha,\gamma}}(rQI_n)$ . The Python scripts and the resulting formula (in text format and as a Python script that can be applied to any values of  $n$ ,  $\alpha$ , and  $\gamma$ ) can be found in the GitHub page [https://github.com/biocom-uib/Quartet\\_Index](https://github.com/biocom-uib/Quartet_Index) companion to this paper. In particular, the plain text formula (which is too long and uninformative to be reproduced here) is given in the document `variance_table.txt` therein. It can be easily checked using a symbolic computation program that when  $\alpha = \gamma$  it agrees with the variance under the  $\alpha$ -model given in Corollary 27.

## 6. Conclusions

In this paper we have introduced a new balance index for phylogenetic trees, the rooted quartet index  $rQI$ . This index makes sense for multifurcating trees, it can be computed in time linear in the number of leaves, and it has a larger range of values than any other shape index defined so far. We have computed its maximum and minimum values for bifurcating and arbitrary trees, and we have shown how to compute its expected value and variance under any probabilistic model of phylogenetic trees that is sampling consistent and invariant under relabelings. This includes the popular uniform, Yule,  $\alpha$ ,  $\beta$  and  $\alpha$ - $\gamma$ -models. This paper is accompanied by the GitHub page [https://github.com/biocom-uib/Quartet\\_Index](https://github.com/biocom-uib/Quartet_Index) where the interested reader can find a set of Python scripts that perform several computations related to this index.

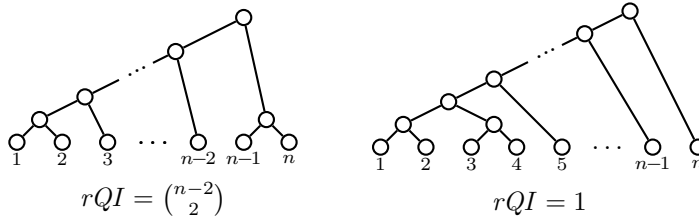
We want to call the reader's attention on a further property of the rooted quartet index: it can be used in a sensible way to measure the balance of *taxonomic trees*, defined as those rooted trees of fixed depth (but with possibly out-degree 1 internal nodes) with their leaves bijectively labeled in a set of taxa. The usual taxonomies with fixed ranks are the paradigm of such taxonomic trees. It turns out that the classical balance indices cannot be used in a sound way to quantify the balance of such trees. For instance, Colless' index cannot be applied to multifurcating trees, and Sackin's index, being the sum of the depths of the leaves in the tree, is constant on all taxonomic trees of fixed depth and number of leaves. As far as the total cophenetic index goes, it is straightforward to check from its very definition that the taxonomic trees with maximum and minimum total cophenetic values among all taxonomic trees of a given depth and a given number of leaves are those depicted in Fig 10. In our opinion, these two trees should be considered as equally balanced. We believe that  $rQI$  can be used to capture the symmetry of a taxonomic tree in a natural way, and we hope to report on it elsewhere.

In a future paper we also plan to study some further properties of  $rQI$ , like for instance its correlation with other balance indices under different probabilistic models. To illustrate the relation between  $rQI$  and other balance indices, in Fig. 12 we provide scatterplots of the values of  $rQI$  (taking  $q_i = i$ ) versus the Sackin index  $S$ , the Colless index  $C$ , the total cophenetic index  $\Phi$  and the number of cherries on  $\mathcal{BT}_{20}$  (which contains more than  $(2 \cdot 20 - 3)!! \geq 8.2 \times 10^{21}$



**Figure 10:** The shapes of the taxonomic trees with maximum (a) and minimum (b) total cophenetic values among all taxonomic trees of given depth and number of leaves.

members) and *versus* the Sackin index  $S$  and the total cophenetic index  $\Phi$  on  $\mathcal{T}_{15}$  (which contains more than  $6.3 \times 10^{15}$  members). The values of the Spearman correlations between these pairs of indices on these classes of trees are given in Table 2. We want to point out the small correlation between  $rQI$  and the number of cherries: although at first sight it could seem that counting the number of fully symmetric rooted quartets in a tree is equivalent to counting pairs of cherries, it is not the case, as the cherries in a rooted quartet may correspond to distant leaves in the tree. For instance, the bifurcating phylogenetic trees in Fig. 11 have both 2 cherries, but their  $rQI$  value is quite different. Notice also that the correlations between  $rQI$  and  $S$ ,  $C$  and  $\Phi$  are negative, because  $rQI$  grows while  $S$ ,  $C$  and  $\Phi$  decrease with the balance of the trees.



**Figure 11:** Two trees with 2 cherries and very different  $rQI$ .

Correlation	Value
$rQI$ vs $S$ on $\mathcal{BT}_{20}$	-0.889
$rQI$ vs $C$ on $\mathcal{BT}_{20}$	-0.893
$rQI$ vs $\Phi$ on $\mathcal{BT}_{20}$	-0.935
$rQI$ vs $Ch$ on $\mathcal{BT}_{20}$	0.165
$rQI$ vs $S$ on $\mathcal{T}_{15}$	-0.787
$rQI$ vs $\Phi$ on $\mathcal{T}_{15}$	-0.827

**Table 2:** Spearman's correlations corresponding to the scatterplots in Fig. 12.

**Acknowledgements.** A preliminary version of this paper was presented at the *Workshop on Algebraic and combinatorial phylogenetics* held in Barcelona (June 26–30, 2017). We thank Mike Steel, Gabriel Riera, Seth Sullivant, the anonymous reviewers and the associate editor for their helpful suggestions on several aspects of this paper. This research was partially supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund through project DPI2015-67082-P (MINECO /FEDER).

## References

- [1] Abramowitz M, Stegun IAS (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover
- [2] Aldous D (1996) Probability distributions on cladograms. In: Random discrete structures, The IMA Volumes in Mathematics and its Applications, vol 76, Springer, pp 1–18
- [3] Aldous D (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* 16:23–34
- [4] Blum MGB, François OF (2005) On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Mathematical Biosciences* 195:141–153
- [5] Cardona G, Mir A, Rosselló F (2013) Exact formulas for the variance of several balance indices under the Yule model. *Journal of Mathematical Biology* 67:1833–1846
- [6] Cavalli-Sforza LL, Edwards A (1967) Phylogenetic analysis: Models and estimation procedures. *Evolution* 21:550–570
- [7] Chen B, Ford D, Winkel M (2009) A new family of Markov branching trees: the alpha-gamma model. *Electronic Journal of Probability* 14:400–430
- [8] Colless DH (1982) Review of “Phylogenetics: The theory and practice of phylogenetic systematics”. *Systematic Zoology* 31:100–104
- [9] Cormen TH, Leiserson CE, Rivest RL, Stein C (2009) *Introduction to Algorithms* (3rd edition). The MIT Press
- [10] Coronado TM, Mir A, Rosselló F (2018) The probabilities of trees and cladograms under Ford’s  $\alpha$ -model. *The Scientific World Journal* 2018:1916094
- [11] Dearlove BL, Frost SD (2015) Measuring asymmetry in time-stamped phylogenies. *PLoS Computational Biology* 11.7 (2015):e1004312
- [12] Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates Inc
- [13] Ford D (2005) Probabilities on cladograms: Introduction to the alpha model. arXiv preprint math/0511246
- [14] Harding E (1971) The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability* 3:44–77
- [15] Keller-Schmidt S, Tuğrul M, Eguíluz VM, Hernández-García E, Klemm K (2015) Anomalous scaling in an age-dependent branching model. *Physical Review E* 91:022803
- [16] Kirkpatrick M, Slatkin M (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47:1171–1181
- [17] Macdonald IG (1995) *Symmetric functions and Hall polynomials* (2nd edition). Oxford University Press
- [18] Matsen F (2007) Optimization over a class of tree shape statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 4:506–512
- [19] McKenzie A, Steel M (2000) Distributions of cherries for two models of trees. *Mathematical Biosciences* 164:81–92

- [20] Mir A, Rosselló F, Rotger L (2013) A new balance index for phylogenetic trees. *Mathematical Biosciences* 241:125–136
- [21] Mir A, Rotger L, Rosselló F (2018) Sound Colless-like balance indices for multifurcating trees. *PLoS ONE* 13(9):e0203401
- [22] Mooers A, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology* 72:31–54
- [23] Pinelis I (2003) Evolutionary models of phylogenetic trees. *Proceedings of the Royal Society of London B: Biological Sciences* 270:1425–1431
- [24] Rosen DE (1978) Vicariant patterns and historical explanation in biogeography. *Systematic Biology* 27:159–188
- [25] Sackin MJ (1972) Good and “bad” phenograms. *Systematic Zoology* 21:225–226
- [26] Shao KT, Sokal R (1990) Tree balance. *Systematic Zoology* 39:226–276
- [27] Sloane NJA (2010) *The On-Line Encyclopedia of Integer Sequences*. URL <http://oeis.org/>
- [28] Wu T, Choi, KP (2016) On joint subtree distributions under two evolutionary models. *Theoretical Population Biology*, 108:13-23
- [29] Yule GU (1924) A mathematical theory of evolution based on the conclusions of Dr J. C. Willis. *Philosophical Transactions of the Royal Society of London Series B* 213:21–87
- [30] Zhu S, Degnan JH, Steel M (2011) Clades, clans and reciprocal monophyly under neutral evolutionary models. *Theoretical Population Biology* 79:220–227
- [31] Zhu S, Than C, Wu T (2015) Clades and clans: a comparison study of two evolutionary models. *Journal of Mathematical Biology*, 71:99–124

## Appendices

### A.1: An alternative derivation of the variance of $rQIB_n$ under the Yule model

In this section we give an alternative proof of the following result.

**Proposition 29.** *Under the Yule model,*

$$\text{Var}_Y(rQIB_n) = \binom{n}{4} \frac{5n^4 + 30n^3 + 118n^2 + 408n + 630}{33075}.$$

*Proof.* By Lemma 10,  $rQIB$  on  $\mathcal{BT}_n$  is a bifurcating recursive tree shape statistic satisfying the recurrence

$$rQIB(T \star T') = rQIB(T) + rQIB(T') + f_{rQIB}(|L(T)|, |L(T')|)$$

with  $f_{rQIB}(a, b) = \binom{a}{2} \binom{b}{2}$ . Then, it satisfies the hypothesis in Cor. 1 of [5] with

$$\begin{aligned} \varepsilon(a, b-1) &= f_{rQIB}(a, b) - f_{rQIB}(a, b-1) \\ &= \binom{a}{2} \binom{b}{2} - \binom{a}{2} \binom{b-1}{2} = (b-1) \binom{a}{2} \\ R(n-1) &= E_Y(rQIB_n) - E_Y(rQIB_{n-1}) \\ &= \frac{1}{3} \binom{n}{4} - \frac{1}{3} \binom{n-1}{4} = \frac{1}{3} \binom{n-1}{3} \end{aligned}$$



Since  $E_Y(rQIB_1) = 0$  and  $f_{rQIB}(n-1, 1) = 0$ , applying the aforementioned result from [5] we have that

$$\begin{aligned}
E_Y(rQIB_n^2) &= \frac{n}{n-1} E_Y(rQIB_{n-1}^2) + \frac{4}{n-1} \sum_{k=1}^{n-2} \varepsilon(k, n-k-1) E_Y(rQIB_k) \\
&\quad + \frac{2}{n-1} \sum_{k=1}^{n-2} R(n-k-1) E_Y(rQIB_k) \\
&\quad + \frac{1}{n-1} \sum_{k=1}^{n-2} (f_{rQIB}(k, n-k)^2 - f_{rQIB}(k, n-k-1)^2) \\
&= \frac{n}{n-1} E_Y(rQIB_{n-1}^2) + \frac{4}{3(n-1)} \sum_{k=1}^{n-2} (n-k-1) \binom{k}{2} \binom{k}{4} \\
&\quad + \frac{2}{9(n-1)} \sum_{k=1}^{n-2} \binom{n-k-1}{3} \binom{k}{4} \\
&\quad + \frac{1}{n-1} \sum_{k=1}^{n-2} \binom{k}{2}^2 \left( \binom{n-k}{2}^2 - \binom{n-k-1}{2}^2 \right) \\
&= \frac{n}{n-1} E_Y(rQIB_{n-1}^2) + \frac{n}{3} \binom{n-2}{4} \frac{15n^2 - 35n + 6}{420} \\
&\quad + \frac{n}{9} \binom{n-2}{4} \frac{n^2 - 13n + 42}{840} \\
&\quad + n \binom{n-2}{2} \frac{3n^4 - 18n^3 + 41n^2 - 42n + 36}{1680} \\
&= \frac{n}{n-1} E_Y(rQIB_{n-1}^2) \\
&\quad + \frac{n(n-2)(n-3)(253n^4 - 2014n^3 + 6119n^2 - 7430n + 3504)}{181440}
\end{aligned}$$

Dividing by  $n$  both sides of this expression for  $E_Y(rQIB_n^2)$  and setting  $y_n = E_Y(rQIB_n^2)/n$ , we obtain the recurrence

$$y_n = y_{n-1} + \frac{(n-2)(n-3)(253n^4 - 2014n^3 + 6119n^2 - 7430n + 3504)}{181440}.$$

Since  $y_0 = y_1 = 0$ , its solution is

$$\begin{aligned}
y_n &= \sum_{k=2}^n \frac{(k-2)(k-3)(253k^4 - 2014k^3 + 6119k^2 - 7430k + 3504)}{181440} \\
&= \frac{(n-3)(n-2)(n-1)(1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040)}{6350400}
\end{aligned}$$

from where we obtain

$$\begin{aligned}
E_Y(rQIB_n^2) &= ny_n \\
&= \binom{n}{4} \frac{1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040}{264600}.
\end{aligned}$$

Finally

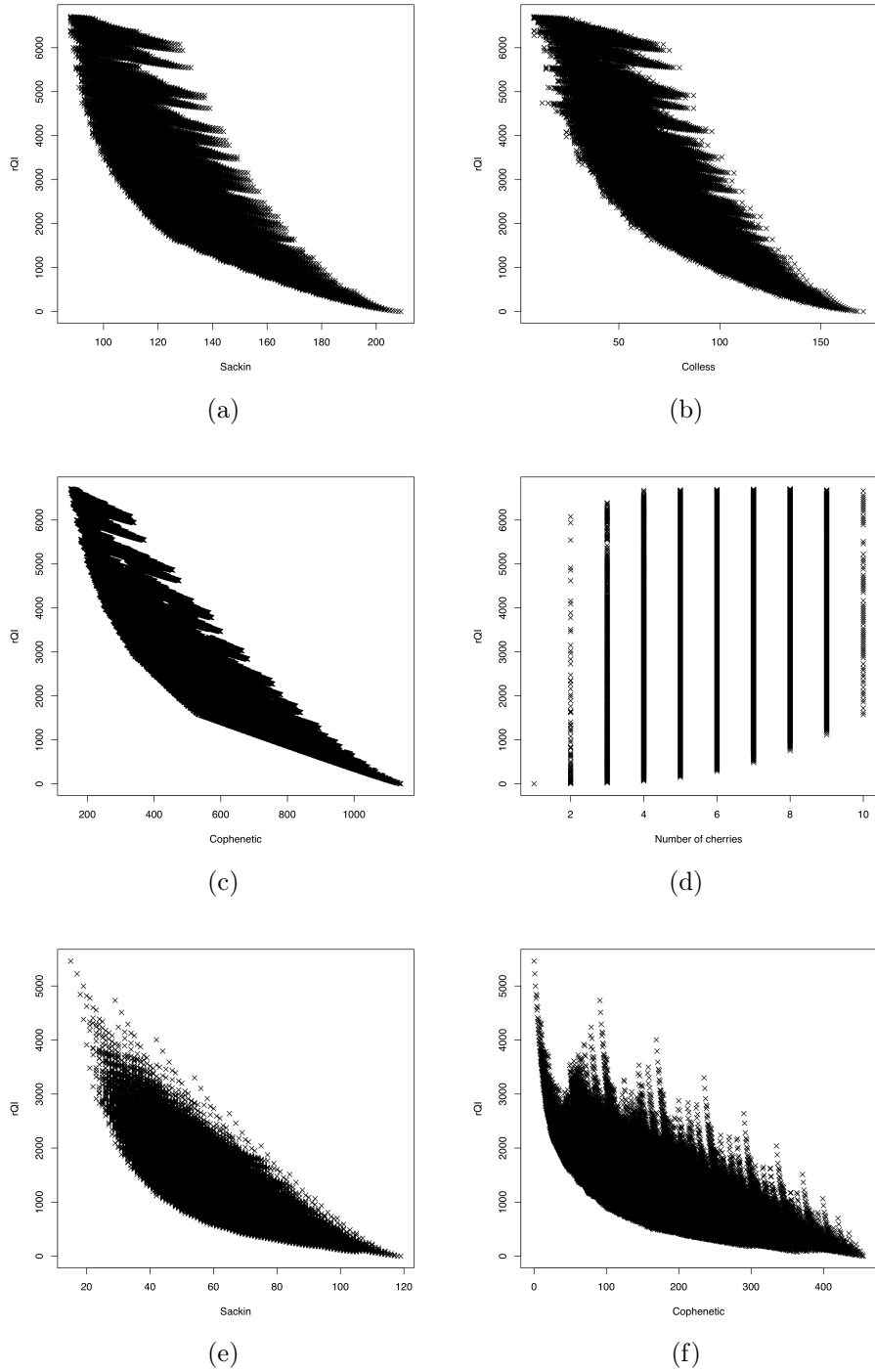
$$\begin{aligned}
\text{Var}_Y(rQIB_n) &= E_Y(rQIB_n^2) - E_Y(rQIB_n)^2 \\
&= \binom{n}{4} \frac{1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040}{264600} - \frac{1}{9} \binom{n}{4}^2 \\
&= \binom{n}{4} \frac{5n^4 + 30n^3 + 118n^2 + 408n + 630}{33075},
\end{aligned}$$

as we claimed.  $\square$

$\square$

$\square$

*A.2: Some tables used in Section 5*



**Figure 12:** Scatterplot of  $rQI$  versus: (a) the Sackin index on  $\mathcal{BT}_{20}$ ; (b) the Colless index on  $\mathcal{BT}_{20}$ ; (c) the total cophenetic index on  $\mathcal{BT}_{20}$ ; (d) the number of cherries on  $\mathcal{BT}_{20}$ ; (e) the Sackin index on  $\mathcal{T}_{15}$ ; (f) the total cophenetic index on  $\mathcal{T}_{15}$ .

Name	Shape	$\Theta_{3,3}(T^*)$
$B_{5,1}^*$	$(*, (*, (*, (*, *))))$	0
$B_{5,2}^*$	$(*, ((*, *), (*, *)))$	0
$B_{5,3}^*$	$((*, *), (*, (*, *)))$	6
$B_{6,1}^*$	$(*, (*, (*, (*, (*, *))))$	0
$B_{6,2}^*$	$(*, (*, ((*, *), (*, *))))$	0
$B_{6,3}^*$	$(*, ((*, *), (*, (*, *))))$	0
$B_{6,4}^*$	$(((*, *), ((*, *), (*, *)))$	18
$B_{6,5}^*$	$(((*, *), (*, (*, (*, *))))$	6
$B_{6,6}^*$	$(((*, (*, *)), (*, (*, *)))$	36
$B_{7,1}^*$	$(*, (*, (*, (*, (*, (*, *))))$	0
$B_{7,2}^*$	$(*, (*, (*, (*, (*, (*, *))))$	0
$B_{7,3}^*$	$(*, (*, ((*, *), (*, (*, *))))$	0
$B_{7,4}^*$	$(*, ((*, *), ((*, *), (*, *))))$	0
$B_{7,5}^*$	$(*, ((*, *), (*, (*, (*, *))))$	0
$B_{7,6}^*$	$(*, ((*, (*, *)), (*, (*, *))))$	0
$B_{7,7}^*$	$(((*, *), (*, (*, (*, (*, *))))$	0
$B_{7,8}^*$	$(((*, *), (*, ((*, *), (*, *))))$	8
$B_{7,9}^*$	$(((*, *), ((*, *), (*, (*, *))))$	24
$B_{7,10}^*$	$(((*, (*, *)), (*, (*, (*, *))))$	36
$B_{7,11}^*$	$(((*, (*, *)), ((*, *), (*, *)))$	36
$B_{8,1}^*$	$(*, (*, (*, (*, (*, (*, (*, *))))$	0
$B_{8,2}^*$	$(*, (*, (*, (*, (*, ((*, *), (*, *))))$	0
$B_{8,3}^*$	$(*, (*, (*, (*, ((*, *), (*, (*, *))))$	0
$B_{8,4}^*$	$(*, (*, ((*, *), ((*, *), (*, (*, *))))$	0
$B_{8,5}^*$	$(*, (*, ((*, *), (*, (*, (*, (*, *))))$	0
$B_{8,6}^*$	$(*, (*, ((*, (*, *)), (*, (*, (*, *))))$	0
$B_{8,7}^*$	$(*, ((*, *), (*, (*, (*, (*, (*, *))))$	0
$B_{8,8}^*$	$(*, ((*, *), (*, (*), (*, (*, *))))$	0
$B_{8,9}^*$	$(*, ((*, *), (*, ((*, *), (*, *))))$	0
$B_{8,10}^*$	$(*, ((*, (*, *)), (*, (*, (*, *))))$	0
$B_{8,11}^*$	$(*, ((*, (*, *)), ((*, *), (*, *))))$	0
$B_{8,12}^*$	$(((*, *), (*, (*, (*, (*, (*, *))))$	0
$B_{8,13}^*$	$(((*, *), (*, (*, (*, *), (*, *))))$	2
$B_{8,14}^*$	$(((*, *), (*, ((*, *), (*, (*, *))))$	6
$B_{8,15}^*$	$(((*, *), ((*, *), (*, (*, (*, *))))$	12
$B_{8,16}^*$	$(((*, *), ((*, *), (*, (*, (*, *))))$	14
$B_{8,17}^*$	$(((*, *), ((*, (*, *)), (*, (*, *))))$	18
$B_{8,18}^*$	$(((*, (*, *)), (*, (*, (*, (*, *))))$	0
$B_{8,19}^*$	$(((*, (*, *)), (*, ((*, *), (*, *))))$	0
$B_{8,20}^*$	$(((*, (*, *)), ((*, *), (*, (*, *))))$	0
$B_{8,21}^*$	$(((*, (*, (*, *))), (*, (*, (*, *))))$	36
$B_{8,22}^*$	$(((*, (*, (*, *))), ((*, *), (*, *)))$	36
$B_{8,23}^*$	$(((*, (*), (*, *)), ((*, *), (*, *)))$	38

**Table 3:** Coefficients of the probabilities of the trees in  $\mathcal{BT}_k^*$ , for  $k = 5, 6, 7, 8$ , in the formula for the variance of  $rQIB_n$ .

Tree	$P_{\alpha,n}^{A,*}$
$Q_3^*$	$\frac{1-\alpha}{3-\alpha}$
$B_{5,3}^*$	$\frac{2(1-\alpha)}{4-\alpha}$
$B_{6,4}^*$	$\frac{(1-\alpha)^2(8-\alpha)}{(5-\alpha)(4-\alpha)(3-\alpha)}$
$B_{6,5}^*$	$\frac{2(1-\alpha)(8-\alpha)}{(5-\alpha)(4-\alpha)(3-\alpha)}$
$B_{6,6}^*$	$\frac{2(1-\alpha)(2-\alpha)}{(5-\alpha)(4-\alpha)}$
$B_{7,8}^*$	$\frac{(1-\alpha)^2(2+\alpha)(10+\alpha)}{(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$
$B_{7,9}^*$	$\frac{2(1-\alpha)^2(10+\alpha)}{(6-\alpha)(5-\alpha)(4-\alpha)}$
$B_{7,10}^*$	$\frac{10(1-\alpha)(2-\alpha)}{(6-\alpha)(5-\alpha)(3-\alpha)}$
$B_{7,11}^*$	$\frac{5(1-\alpha)^2(2-\alpha)}{(6-\alpha)(5-\alpha)(3-\alpha)}$
$B_{8,13}^*$	$\frac{8(1-\alpha)^2(1+\alpha)(2+\alpha)(3+\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$
$B_{8,14}^*$	$\frac{16(1-\alpha)^2(1+\alpha)(3+\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)}$
$B_{8,15}^*$	$\frac{8(1-\alpha)^2(3+\alpha)(8-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$
$B_{8,16}^*$	$\frac{4(1-\alpha)^3(3+\alpha)(8-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$
$B_{8,17}^*$	$\frac{8(1-\alpha)^2(2-\alpha)(3+\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)}$
$B_{8,21}^*$	$\frac{20(1-\alpha)(2-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$
$B_{8,22}^*$	$\frac{20(1-\alpha)^2(2-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$
$B_{8,23}^*$	$\frac{5(1-\alpha)^3(2-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$

**Table 4:** Probabilities under the  $\alpha$ -model of the trees involved in the formula for the variance of  $rQIB_n$

Tree	$P_{\beta,n}^{B,*}$
$Q_3^*$	$\frac{3(\beta+2)}{7\beta+18}$
$B_{5,3}^*$	$\frac{2(\beta+2)}{3\beta+8}$
$B_{6,4}^*$	$\frac{45(\beta+2)^2(\beta+4)}{(31\beta^2+194\beta+300)(7\beta+18)}$
$B_{6,5}^*$	$\frac{60(\beta+2)(\beta+3)(\beta+4)}{(31\beta^2+194\beta+300)(7\beta+18)}$
$B_{6,6}^*$	$\frac{10(\beta+2)(\beta+3)}{31\beta^2+194\beta+300}$
$B_{7,8}^*$	$\frac{3(\beta+2)^2(\beta+4)(\beta+5)}{(\beta+3)(3\beta+8)(3\beta+10)(7\beta+18)}$
$B_{7,9}^*$	$\frac{2(\beta+2)^2(\beta+5)}{(\beta+3)(3\beta+8)(3\beta+10)}$
$B_{7,10}^*$	$\frac{20(\beta+2)(\beta+3)}{3(3\beta+10)(7\beta+18)}$
$B_{7,11}^*$	$\frac{5(\beta+2)^2}{(3\beta+10)(7\beta+18)}$
$B_{8,13}^*$	$\frac{504(\beta+2)^2(\beta+4)^2(\beta+5)^2(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(3\beta+8)(7\beta+18)}$
$B_{8,14}^*$	$\frac{336(\beta+2)^2(\beta+4)(\beta+5)^2(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(3\beta+8)}$
$B_{8,15}^*$	$\frac{1680(\beta+2)^2(\beta+3)(\beta+4)(\beta+5)(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(7\beta+18)}$
$B_{8,16}^*$	$\frac{1260(\beta+2)^3(\beta+4)(\beta+5)(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(7\beta+18)}$
$B_{8,17}^*$	$\frac{280(\beta+2)^2(\beta+3)(\beta+5)(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)}$
$B_{8,21}^*$	$\frac{560(\beta+2)(\beta+3)^3(\beta+4)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}$
$B_{8,22}^*$	$\frac{840(\beta+2)^2(\beta+3)^2(\beta+4)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}$
$B_{8,23}^*$	$\frac{315(\beta+2)^3(\beta+3)(\beta+4)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}$

**Table 5:** Probabilities under the  $\beta$ -model of the trees involved in the formula for the variance of  $rQIB_n$