A balance index for phylogenetic trees based on quartets

Tomás M. Coronado[a], Arnau Mir[a], Francesc Rosselló[a], Gabriel Valiente[b]

[a] *Balearic Islands Health Research Institute (IdISBa) and Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma, Spain*
[b] *Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain*

Abstract

We define a new balance index for phylogenetic trees based on the symmetry of the evolutive history of every quartet of leaves. This index makes sense for polytomic trees and it can be computed in time linear in the number of leaves. We compute its maximum and minimum values for arbitrary and bifurcating trees, and we provide exact formulas for its expected value and variance for bifurcating trees under Ford's α-model and Aldous' β-model and for arbitrary trees under the α-γ-model.

1. Introduction

One of the most broadly studied properties of the topology of 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 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 considered in the tree [9, Chap. 33].

The symmetry of a tree is usually quantified by means of *balance indices*. Such indices include the *Colless' index* [7] 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; the *Sackin's index* [23], which is defined as the sum of the depths of all leaves in the tree, and hence, it measures the average depth of a leaf; and the *total cophenetic index* [17], which is defined as the sum, over all pairs of different leaves of the tree, of the depth of their least common ancestor, and is related to the variance of the leaves' depths (cf. [17, Lem. 2]). But there are many more balance indices [9, Chap. 33], and Shao and Sokal [24, p. 1990] explicitly advise to use more than one such index to quantify tree balance. Such balance indices only depend on the topology of the trees, not on the branch lengths or the actual taxa labeling their leaves, and they have been widely used as tools to test stochastic models of evolution [3, 12, 13, 18, 24].

In this paper we propose a new balance index, the *quartet index*. It is defined as the sum, over all 4-tuples of different leaves of the tree, of a value that quantifies the symmetry of the joint evolution of the species they represent. The value associated with a 4-tuple of leaves is simply required to increase with the number of isomorphisms of the phylogenetic subtree (the *quartet*) they induce. The main features of our index are that it can be easily computed in linear time and that its expected 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 for bifurcating phylogenetic trees, Ford's α-model [10] and Aldous' β-model [1], which

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

include as specific instances the Yule [11, 27] and the uniform [5, 22, 26] models, as well as for Chen-Ford's α-γ-model for polytomic trees [6]. To our knowledge, this is the first shape index for which formulas for the expected value and the variance under the α-γ-model have been provided. We also compute the maximum and minimum values of our quartet index, in both the arbitrary and the bifurcating cases: the minimum value is reached exactly at the combs and the maximum value is reached exactly at the rooted stars in the arbitrary case and at the maximally balanced trees in the bifurcating case.

The rest of this paper is organized as follows. In a first 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 quartet index QI 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` 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. 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 $\mathrm{child}(u)$ 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) trees and of bifurcating trees with n leaves, respectively.

A *phylogenetic tree* on a set Σ is a tree with its leaves bijectively labeled in Σ. 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) phylogenetic trees and of bifurcating phylogenetic trees on Σ, respectively. If Σ and Σ' are any two sets of labels of the same cardinal, 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 Σ is unrelevant and only its cardinal 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 Σ with the set $[n] = \{1, 2, \ldots, n\}$. If $|\Sigma| = n$, there exists a forgetful mapping $\pi : \mathcal{T}(\Sigma) \to \mathcal{T}_n^*$ that sends every phylogenetic tree T on Σ 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 cardinal) have the same shape.

We shall represent trees and phylogenetic trees by means of their usual Newick format [19], 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 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 phylogenetic tree on a set Σ 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 *star* is a phylogenetic tree all whose leaves are children of the root: see Fig. 1.(b). For every set Σ, there is only one star on Σ, and if $|\Sigma| = n$, we shall generically denote it (as well as its shape) by S_n.

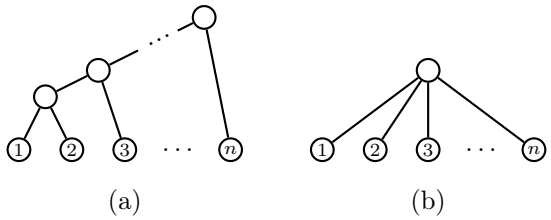


Figure 1: (a) A comb K_n. (b) A star S_n.

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

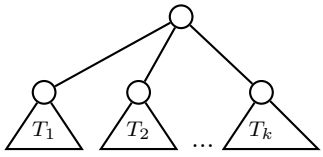


Figure 2: The root join $T_1 \star \cdots \star T_k$.

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 $bal_T(v) \leqslant 1$. 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. Recurrently, a bifurcating tree is maximally balanced when its root is balanced and 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 $\mathcal{BT}_n^*$ [17, §2.1].

2.2. Probabilistic models

A *probabilistic model of phylogenetic trees* $(P_n)_n$ is a family of probability mappings $P_n : \mathcal{T}_n \to [0, 1]$, each one sending each phylogenetic tree in $\mathcal{T}_n$ to its probability under this model. Every such a probabilistic model of phylogenetic trees $(P_n)_n$ induces a *probabilistic model of trees*, that is, a family of probability mappings $P_n^* : \mathcal{T}_n^* \to [0, 1]$, 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 $P_n : \mathcal{T}_n \to [0,1]$ induces also a probability mapping P_Σ 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 $(P_n)_n$ such that $P_n(T) = 0$ for every $T \in \mathcal{T}_n \setminus \mathcal{BT}_n$.

A probabilistic model of phylogenetic trees $(P_n)_n$ is *shape invariant* (or *exchangeable* [1]) 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 $T \in \pi^{-1}(T^*)$,

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

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

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

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

A probabilistic model of phylogenetic trees $(P_n)_n$ is *sampling consistent* [1] (or also *deletion stable* [10]) when, for every $n \geqslant 2$, 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 P_{n-1}; formally, when, for every $n \geqslant 2$ 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 $1 \leqslant m < n$, the restriction of $T(-n)$ to $[m]$ is simply $T([m])$, that this condition is equivalent to the following: $(P_n)_n$ is sampling consistent when, for every $n \geqslant 2$, for every $1 \leqslant m < n$, and for every $T_0 \in \mathcal{T}_m$,

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

It is also easy to prove that if $(P_n)_n$ 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 $n \geqslant 2$, 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 $(P_n^*)_n$ is *sampling consistent*, or *deletion stable*, when, for every $n \geqslant 2$, 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 P_{n-1}^*: formally, when, for every $n \geqslant 2$ and for every $T_0 \in \mathcal{T}_{n-1}^*$,

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

We prove now several lemmas on probabilistic models that will be used in §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 cardinal k of X.

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

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

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

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

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

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

which proves the inductive step. □

Lemma 2. *Let $(P_n)_n$ 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 $f : T_{n-1} \to T'_{n-1}$ 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

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

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 will be the old root). This entails the existence of a shape preserving bijection

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

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 $(P_n)_n$ 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. □

Next lemma generalizes Cor. 40 in [10]. For the sake of completeness, we provide a direct complete proof of it.

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

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

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

Therefore

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

and hence

$$P^*_{n-1}(T^*_{n-1}) = \sum_{T^*_n \in \mathcal{T}^*_n} \frac{\left|\{x \in L(T^*_n) \mid T^*_n(-x) = T^*_{n-1}\}\right|}{n} \cdot P^*_n(T^*_n)$$

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 $(P^*_n)_n$ 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

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

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

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

as we wanted to prove. □

In §5 we shall be concerned with three specific parametric probabilistic models of phylogenetic trees:

1) Aldous' β-model. The β-splitting model $(P^A_{\beta,n})_n$ [1, 2] is a probabilistic model of bifurcating phylogenetic trees that depends on one parameter $\beta \in]-2, \infty[$. Under it, the probability $P^{A,*}_{\beta,n}(T^*)$ of a tree T^* with n leaves is computed through all possible ways of building it from a single node by recurrently splitting leaves into cherries with leaves having two predetermined target numbers of descendant leaves in the final tree; the parameter β controls the probabilities of these target numbers of descendant leaves. Then, the probability $P^A_{\beta,n}(T)$ of a phylogenetic tree T is obtained from the probability of its shape T^* by means of equation (1). In this way, $(P^A_{\beta,n})_n$ is shape invariant by construction, and it turns out to be sampling consistent [1, §6.3];

hence, by Lemma 3, the β-model of trees $(P^{A,*}_{\beta,n})_n$ is also sampling consistent. The β-model includes as specific cases the Yule model [11, 27] (when $\beta = 0$) and the uniform model [5, 21] (when $\beta = -3/2$).

In §5 we shall need to know the probability $P^{A,*}_{\beta,4}$ of the maximally balanced tree with 4 leaves $((*,*),(*,*))$, which we denote in this paper by Q^*_3 (see Figure 3). This probability is

$$P^{A,*}_{\beta,4}(Q^*_3) = \frac{3\beta + 6}{7\beta + 18}. \tag{3}$$

For the convenience of the reader, we compute it in the Appendix A.1 at the end of the paper (see Lemma 25).

2) Ford's α-model. The α-model $(P^F_{\alpha,n})_n$ [10] is another probabilistic model of bifurcating phylogenetic trees that depends on one parameter $\alpha \in [0, 1]$. In it, the probability $P^{F,*}_{\alpha,n}(T^*)$ of a tree T^* is determined through all possible ways of constructing a phylogenetic tree with this shape from a single node (labeled with 1) by adding recurrently new leaves labeled with $2, \ldots, n$ within arcs or to a new root, with the parameter α controlling the probability of adding the new leaf within an arc ending in a leaf or elsewhere. Then, the probability $P^F_{\alpha,n}(T)$ of a phylogenetic tree T is obtained from the probability of its shape by means of equation (1). The α-model is shape invariant by construction and sampling consistent by [10, Prop. 42], and it also includes as specific cases the Yule model (when $\alpha = 0$) and the uniform model (when $\alpha = 1/2$).

In §5, we shall also need to know $P^{F,*}_{\alpha,4}(Q^*_3)$, which is provided by Ford in [10, §7: Fig. 20]:

$$P^{F,*}_{\alpha,4}(Q^*_3) = \frac{1 - \alpha}{3 - \alpha}. \tag{4}$$

3) Chen-Ford-Winkel's α-γ-model. The α-γ-model $(P_{\alpha,\gamma,n})_n$ [6] is a probabilistic model of phylogenetic trees that depends on two parameters $0 \leqslant \gamma \leqslant \alpha \leqslant 1$. In it, the probability of a phylogenetic tree is computed as it is built from a single node (labeled with 1) by adding recurrently new leaves labeled with $2, \ldots, n$ to a new root, to internal nodes, or within arcs, with the parameters α, γ controlling the probabilities of choosing the place where to add a new leaf. It turns out that $(P_{\alpha,\gamma,n})_n$ is not shape invariant in general [6, Prop. 1.(b)], but that the corresponding model for trees $(P^*_{\alpha,\gamma,n})_n$ is sampling consistent [6, Thm. 2]. When $\alpha = \gamma$, $(P^*_{\alpha,\alpha,n})_n$ is simply the α-model $(P^{F,*}_{\alpha,n})_n$ for unlabeled trees.

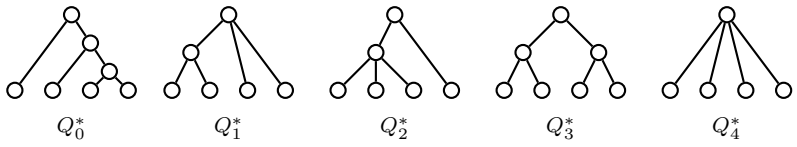


Figure 3: The 5 trees in $\mathcal{T}^*_4$.

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 3. They are:

$$\begin{aligned}
P^*_{\alpha,\gamma,4}(Q^*_0) &= \frac{2(1-\alpha+\gamma)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)} \\
P^*_{\alpha,\gamma,4}(Q^*_1) &= \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \\
P^*_{\alpha,\gamma,4}(Q^*_2) &= \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \\
P^*_{\alpha,\gamma,4}(Q^*_3) &= \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)} \\
P^*_{\alpha,\gamma,4}(Q^*_4) &= \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}
\end{aligned} \tag{5}$$

For the convenience of the reader, we also compute these probabilities in the Appendix A.2 at the end of the paper (see Lemma 26).

3. Quartet indices

Let T be a phylogenetic tree on a set Σ. For every $Q \in \mathcal{P}_4(\Sigma)$, the *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 quartets of five different shapes; they are listed in Figure 3, together with the notations used in this paper to denote them (motivated by Table 1). Notice that a bifurcating phylogenetic tree $T \in \mathcal{BT}_n$ can only contain quartets of two shapes: those denoted by Q^*_0 and Q^*_3 in the aforementioned figure.

We associate to each quartet a *QI-value* q_i that increases with the symmetry of the 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 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 $q_i = 2^i$.

Quartet	Q^*_0	Q^*_1	Q^*_2	Q^*_3	Q^*_4
# Automorphisms	2	4	6	8	24
QI	0	q_1	q_2	q_3	q_4

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

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

$$\begin{aligned}
QI(T) &= \sum_{Q \in \mathcal{P}_4(\Sigma)} QI(T(Q)) \\
&= \sum_{i=1}^{4} \left|\{Q \in \mathcal{P}_4(\Sigma) \mid \pi(T(Q)) = Q^*_i\}\right| \cdot q_i
\end{aligned}$$

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

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

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

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

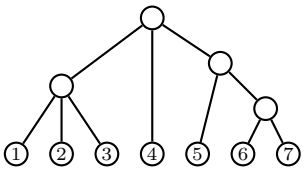


Figure 4: The tree $((1,2,3),4,(5,(6,7)))$.

Remark 5. If we do not take $q_0 = 0$, then the resulting index would be

$$\sum_{i=0}^{4} q_i \cdot \left|\{Q \in \mathcal{P}_4(\Sigma) \mid \pi(T(Q)) = Q_i^*\}\right| = q_0\binom{n}{4} + \sum_{i=1}^{4}(q_i - q_0)\left|\{Q \in \mathcal{P}_4(\Sigma) \mid \pi(T(Q)) = Q_i^*\}\right|\Big)$$

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

Remark 6. One could also associate other values to the quartet shapes; for instance their Sackin index [23, 24] or their total cophenetic index [17], which measure the unbalance of the 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 quartets of shape Q_0^* and Q_3^*, its QI index is simply q_3 times its number of 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 *quartet index for bifurcating trees* QIB: for every $T \in \mathcal{BT}(\Sigma)$,

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

The quartet index for bifurcating trees satisfies the following recurrence.

Lemma 7. *Let $T = T_1 \star T_2 \in \mathcal{BT}_n$, where each T_i has n_i leaves. Then,*

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

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, the quartets with $Q \subseteq L(T_i)$ contribute $QIB(T_i)$ to $QIB(T)$.

(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 $QIB(T)$.

(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 $QIB(T)$. There are $\binom{n_1}{2} \cdot \binom{n_2}{2}$ such quartets.

□

Thus, QIB is a *bifurcating recursive tree shape statistic* in the sense of [16]. The recurrence in the last lemma implies directly the following explicit formula for QIB, 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 10 below):

Corollary 8. *If, for every $T \in \mathcal{BT}_n$ and for every $v \in V_{int}(T)$, we set* $\mathrm{child}(v) = \{v_1, v_2\}$*, then*

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

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

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

However, next lemma shows that there exists a slightly more involved linear recurrence for QI, 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 *triples* in T (that is, of restrictions of T to sets of 3 leaves) of shape S_3. Notice that if $T = T_1 \star \cdots \star T_m$ and $|L(T_i)| = n_i$, for each $i = 1, \ldots, m$, then

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

and hence

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

Lemma 9. *Let $T = T_1 \star \cdots \star T_m \in \mathcal{T}_n$, where each T_i has n_i leaves. Then*

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

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, the quartets with $Q \subseteq L(T_i)$ contribute $QI(T_i)$ to $QI(T)$.

(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 this type, and each one of them contributes q_2 to $QI(T)$
- Has shape Q_0^* if $T_i(\{a, b, c\})$ has shape K_3. These quartets do not contribute anything to $QI(T)$.

(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 this type, and each one of them contributes q_3 to $QI(T)$.

(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 $\binom{n_i}{2}n_jn_k + \binom{n_j}{2}n_kn_i + \binom{n_k}{2}n_in_j$ quartets of this type, and each one of them contributes q_1 to $QI(T)$.

(5) If each leaf in Q belongs to a different subtree T_i, then $T(Q)$ has shape Q_4^*. For every quartet of subtrees T_i, T_j, T_k, T_l, there are $n_in_jn_kn_l$ such quartets Q, and each one of them contributes q_4 to $QI(T)$.

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

Proposition 10. *If $T \in \mathcal{T}_n$, $QI(T)$ can be computed in time $O(n)$.*

Proof. Let $T \in \mathcal{T}_n$. Recall that if a certain mapping $\phi : V(T) \to \mathbb{R}$ 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 $O\big(n + \sum_k m_k \cdot k\big)$, 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 $\big(\kappa(v)\big)_{v\in V(T)}$ can be computed in $O(n)$ time.

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

$$\begin{aligned}
E_l(v) &= \sum_{\{v_1,\dots,v_l\}\subseteq \mathrm{child}(v)} \kappa(v_1)\cdots\kappa(v_l), \qquad l = 2,\dots,\deg(v)\\
F_1(v) &= \sum_{\{v_1,v_2,v_3\}\subseteq \mathrm{child}(v)} \Big(\binom{\kappa(v_1)}{2}\kappa(v_2)\kappa(v_3) + \binom{\kappa(v_2)}{2}\kappa(v_1)\kappa(v_3)\\
&\qquad\qquad + \binom{\kappa(v_3)}{2}\kappa(v_1)\kappa(v_2)\Big)\\
F_2(v) &= \sum_{\{v_1,v_2\}\subseteq \mathrm{child}(v)} \big(\kappa(v_1)\Upsilon(T_{v_2}) + \kappa(v_2)\Upsilon(T_{v_1})\big)\\
F_3(v) &= \sum_{\{v_1,v_2\}\subseteq \mathrm{child}(v)} \binom{\kappa(v_1)}{2}\binom{\kappa(v_2)}{2}
\end{aligned}$$

so that

$$\begin{aligned}
\Upsilon(T) &= \sum_{v\in V_{int}(T)} E_3(v)\\
QI(T) &= \sum_{v\in V_{int}(T)} \big(q_1F_1(v) + q_2F_2(v) + q_3F_3(v) + q_4E_4(v)\big)
\end{aligned}$$

We want to prove now that each one of the vectors

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

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

One of the key ingredients in the proof are the *Newton-Girard formulas* [14, §I.2]: given a (multi)set of numbers $X = \{x_1, \dots, x_k\}$, if we let, for every $l \geqslant 1$,

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

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 also in time $O(k)$.

In particular, if, for every $v \in V_{int}(V)$, we consider the multiset $X_v = \{\kappa(u) \mid u \in \text{child}(v)\}$, then every $E_l(v) = E_l(X_v)$ can be computed in time $O(\deg(v))$ and hence the whole vector $\big(E_l(v)\big)_{v \in V_{int}(T)}$ can be computed in time $O(n)$. In particular, $\big(E_3(v)\big)_{v \in V_{int}(T)}$ and $\big(E_4(v)\big)_{v \in V_{int}(T)}$ 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 $\big(\Upsilon(T_v)\big)_{v \in V_{int}(T)}$ can also be computed in time $O(n)$. And then, since

$$\begin{aligned} F_3(v) &= \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \big(\kappa(v_1)\Upsilon(T_{v_2}) + \kappa(v_2)\Upsilon(T_{v_1})\big) \\ &= \Big(\sum_{v_i \in \text{child}(v)} \kappa(v_i)\Big)\Big(\sum_{v_j \in \text{child}(v)} \Upsilon(T_{v_j})\Big) - \sum_{v_i \in \text{child}(v)} \kappa(v_i)\Upsilon(T_{v_i}) \\ &= \kappa(v)\big(\Upsilon(T_v) - E_3(v)\big) - \sum_{v_i \in \text{child}(v)} \kappa(v_i)\Upsilon(T_{v_i}) \end{aligned}$$

this implies that each $F_3(v)$ can be computed in time $O(\deg(v))$ and hence that the whole vector $\big(F_3(v)\big)_{v \in V_{int}(T)}$ can be computed in time $O(n)$.

Let us focus now on

$$\begin{aligned} F_2(v) &= \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \binom{\kappa(v_1)}{2}\binom{\kappa(v_2)}{2} \\ &= \frac{1}{4} \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \kappa(v_1)^2\kappa(v_2)^2 + \frac{1}{4} \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \kappa(v_1)\kappa(v_2) \\ &\qquad - \frac{1}{4} \sum_{\{v_1, v_2\} \subseteq \text{child}(v)} \big(\kappa(v_1)^2\kappa(v_2) + \kappa(v_2)^2\kappa(v_1)\big) \end{aligned}$$

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

$$\sum_{\{v_1,v_2\}\subseteq \text{child}(v)} \big(\kappa(v_1)^2\kappa(v_2) + \kappa(v_2)^2\kappa(v_1)\big) = \Big(\sum_{v_i\in\text{child}(v)} \kappa(v_i)^2\Big)\Big(\sum_{v_j\in\text{child}(v)} \kappa(v_j)\Big) - \Big(\sum_{v_i\in\text{child}(v)} \kappa(v_i)^3\Big)$$

and hence it can also be computed in time $O(\deg(v))$. Therefore, the whole vector $\big(F_2(v)\big)_{v\in V_{int}(T)}$ can be computed in time $O(n)$.

Let us consider finally

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

Hence, it can also be computed in time $O(\deg(v))$ and therefore, the whole vector $\big(F_1(v)\big)_{v\in V_{int}(T)}$ can be computed in time $O(n)$. □

4. Trees with maximum and minimum *QI*

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

Theorem 11. *The minimum value of QI in $\mathcal{T}_n$ is reached exactly at the combs K_n, and it is 0. The maximum value value of QI in $\mathcal{T}_n$ is reached exactly at the star S_n, and it is $q_4\binom{n}{4}$.*

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

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

□

Therefore, the range of QI 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, which, going from 0 to $\binom{n}{3}$, was so far the balance index in the literature with the widest range [17].

It remains to characterize those *bifurcating* phylogenetic trees with largest QI, or, equivalently, with largest QIB. They turn out to be exactly the maximally balanced trees, as defined at the end of §2.1. The proof is similar to that of the characterization of the bifurcating phylogenetic trees with minimum total cophenetic index provided in [17, §4].

Lemma 12. *Let $T \in \mathcal{BT}_n$ be the bifurcating phylogenetic tree depicted in Fig 5.(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, $QIB(T)$ is not maximum in $\mathcal{BT}_n$.*

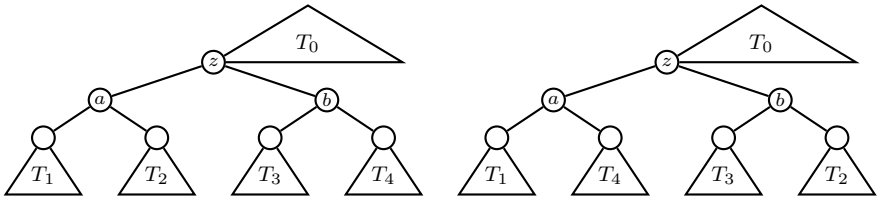


Figure 5: (a) The tree T in the statement of Lemma 12. (b) The tree T' in the proof of Lemma 12.

Proof. Let T' be the tree obtained from T by interchanging T_2 and T_4; see Fig 5.(b). We shall prove that $QIB(T') > QIB(T)$.

Let Σ_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 $Q \cap \Sigma_z$ consist 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 $Q \cap \Sigma_z$ consist 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

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

Now, to compute the difference in the right hand side of this equality, we apply Lemma 7:

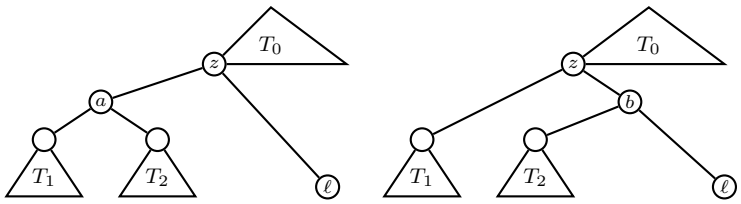
$$\begin{aligned} QIB(T_z) &= QIB(T_1) + QIB(T_2) + QIB(T_3) + QIB(T_4) \\ &\quad + \binom{n_1}{2}\binom{n_2}{2} + \binom{n_3}{2}\binom{n_4}{2} + \binom{n_1+n_2}{2}\binom{n_3+n_4}{2} \\ QIB(T'_z) &= QIB(T_1) + QIB(T_4) + QIB(T_2) + QIB(T_3) \\ &\quad + \binom{n_1}{2}\binom{n_4}{2} + \binom{n_2}{2}\binom{n_3}{2}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

and hence

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

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

Lemma 13. *Let $T \in \mathcal{BT}_n$ be a bifurcating phylogenetic tree containing a leaf whose sibling has at least 3 descendant leaves. Then, $QIB(T)$ is not maximum in $\mathcal{BT}_n$.*



(a) T (b) T'

Figure 6: (a) The tree T in the statement of Lemma 13. (b) The tree T' in the proof of Lemma 13.

Proof. Assume that the tree $T \in \mathcal{BT}_n$ in the statement is the one depicted in Fig 6.(a), with ℓ a leaf such that the subtree T_a rooted at its sibling a has $|L(T_a)| \geqslant 3$. Let $n_1 = |L(T_1)|$ and $n_2 = |L(T_2)|$ and assume $n_1 \geqslant n_2$: then, since $n_1 + n_2 \geqslant 3$, $n_1 \geqslant 2$. Let then T' be the tree depicted in Fig 6.(b): we shall prove that $QIB(T') > QIB(T)$. Reasoning as in the proof of the last lemma, we deduce that

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

Now, using Lemma 7, we have that

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

and therefore

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

as we wanted to prove. □

Theorem 14. *For every $T \in \mathcal{BT}_n$, $QIB(T)$ is maximum in $\mathcal{BT}_n$ if, and only if, T is maximally balanced.*

Proof. Assume that $QIB(T)$ 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 $\kappa(a) \geqslant \kappa(b) + 2$.

If b is a leaf, then, by Lemma 13, $QIB(T)$ 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 $n_1 \geqslant n_2$ and

$n_3 \geqslant n_4$. 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, $n_1 + n_2 = \kappa(a) \geqslant \kappa(b) + 2 = n_3 + n_4 + 2$ implies that $n_1 > n_3$.

Now, by Lemma 12, since by assumption $QIB(T)$ is maximum on $\mathcal{BT}_n$, it must happen that $n_1 > n_3 \geqslant n_4 \geqslant n_2$. This forbids the equality $n_1 = n_2$, and hence $n_1 - 1 = n_2 = n_3 = n_4$. But this contradicts that $n_1 + n_2 \geqslant n_3 + n_4 + 2$.

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

□

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

Lemma 15. *For every n, let q_n be the maximum of QIB on $\mathcal{BT}_n$. Then, $q_1 = q_2 = q_3 = 0$ and*

$$q_n = q_{\lceil n/2 \rceil} + q_{\lfloor n/2 \rfloor} + \binom{\lceil n/2 \rceil}{2} \cdot \binom{\lfloor n/2 \rfloor}{2}, \quad \text{for } n \geqslant 4.$$

Proof. This recurrence for q_n is a direct consequence of Lemma 7 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 $(q_n)_n$ 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* [25]. Its values for $n = 4, \ldots, 20$ 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 [8, Thm. 4.1], that q_n grows asymptotically in $O(n^4)$. Moreover, it is easy to compute q_{2^n} from this recurrence, yielding

$$q_{2^n} = \Big(\frac{4}{7(2^n - 3)} + \frac{3}{7}\Big)\binom{2^n}{4}.$$

In particular, $q_{2^n} / \binom{2^n}{4} \overset{n \to \infty}{\longrightarrow} 3/7$, which is in agreement with the probability of $((*,*),(*,*))$ under the β-model when $\beta \to \infty$; cf. [1, §4.1].

5. The expected value and the variance of *QI*

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

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

Lemma 16. *For every $n \geqslant 1$, the distributions of QI_n and QI_n^* are the same. In particular, their expected values and their variances are the same.*

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

$$f_{QI_n}(x_0) = \sum_{\substack{T \in \mathcal{T}_n \\ QI(T) = x_0}} P_n(T) = \sum_{\substack{T^* \in \mathcal{T}_n^* \\ QI(T^*) = x_0}} \; \sum_{\substack{T \in \mathcal{T}_n \\ \pi(T) = T^*}} P_n(T) = \sum_{\substack{T^* \in \mathcal{T}_n^* \\ QI(T^*) = x_0}} P_n^*(T^*) = f_{QI_n^*}(x_0)$$

□

Proposition 17. *If $\mathcal{M}^*$ is sampling consistent, then*

$$E_{\mathcal{M}}(QI_n) = \binom{n}{4} \sum_{i=1}^{4} q_i P_4^*(Q_i^*).$$

Proof. By Lemma 16, $E_{\mathcal{M}}(QI_n)$ is equal to the expected value $E_{\mathcal{M}^*}(QI_n^*)$ of QI_n^* under $\mathcal{M}^*$, which can be computed as follows:

$$\begin{aligned} E_{\mathcal{M}^*}(QI_n^*) &= \sum_{T^* \in \mathcal{T}_n^*} QI(T^*) P_n^*(T^*) \\ &= \sum_{T^* \in \mathcal{T}_n^*} \Big(\sum_{i=1}^{4} q_i \big| \{Q \in \mathcal{P}_4(L(T^*)) \mid T^*(Q) = Q_i^*\} \big| \Big) P_n^*(T^*) \\ &= \binom{n}{4} \sum_{i=1}^{4} q_i \sum_{T^* \in \mathcal{T}_n^*} \frac{\big| \{Q \in \mathcal{P}_4(L(T^*)) \mid T^*(Q) = Q_i^*\} \big|}{\binom{n}{4}} P_n^*(T^*) = \binom{n}{4} \sum_{i=1}^{4} q_i P_4^*(Q_i^*) \end{aligned}$$

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

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

by the sampling consistency of $\mathcal{M}^*$.

□

The α-γ model for unlabeled trees $\mathcal{M}_{\alpha,\gamma}^*$ is sampling consistent [6, Prop. 12]. Therefore, applying the last proposition using the distribution $\mathcal{M}_{\alpha,\gamma}^*$ on $\mathcal{T}_4^*$ given in §2.2, we have the following result.

Corollary 18. *Let $\mathcal{M}_{\alpha,\gamma} = (P_{\alpha,n})_{n \geqslant 1}$ be the α-γ-model on $(\mathcal{T}_n)_{n \geqslant 1}$, with $\alpha \in [0,1]$. Then*

$$\begin{aligned} E_{\mathcal{M}_{\alpha,\gamma}}(QI_n) \quad &= \binom{n}{4} \Bigg(\frac{(2\alpha - \gamma)(\alpha - \gamma)}{(3-\alpha)(2-\alpha)} \cdot q_4 + \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_3 \\ &\qquad + \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_2 + \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_1 \Bigg). \end{aligned}$$

If $\mathcal{M} = (P_n)_n$ is a probabilistic model of bifurcating phylogenetic trees, so that $P_4^*(Q_4^*) = P_4^*(Q_2^*) = P_4^*(Q_1^*) = 0$, then the expression in Prop. 17 becomes

$$E_{\mathcal{M}}(QI_n) = \binom{n}{4} q_3 P_4^*(Q_3^*).$$

Making $q_3 = 1$, we obtain the following result.

Corollary 19. *If $\mathcal{M}$ is a probabilistic model of bifurcating phylogenetic trees such that $\mathcal{M}^*$ is sampling consistent, then*

$$E_{\mathcal{M}}(QIB_n) = \binom{n}{4} P_4^*(Q_3^*).$$

Since the α and β-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 §2.2 entail the following result.

Corollary 20. *Let $\mathcal{M}_\alpha^F$ be Ford's α-model for bifurcating phylogenetic trees, with $\alpha \in [0,1]$, and let $\mathcal{M}_\beta^A$ be Aldous' β-model for bifurcating phylogenetic trees, with $\beta \in]-2,\infty[$. Then:*

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

It is easy to check that $E_{\mathcal{M}_\alpha^F}(QIB_n)$ agrees with $E_{\mathcal{M}_{\alpha,\gamma}}(QI_n)$ (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 QIB_n are, respectively,

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

Let us deal now with the variance of QI_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

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

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

Proposition 21. *If $\mathcal{M}^*$ is sampling consistent, then*

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

Proof. Since, by Lemma 16, $Var_{\mathcal{M}}(QI_n) = Var_{\mathcal{M}^*}(QI_n^*)$, we shall compute the latter using the formula $Var_{\mathcal{M}^*}(QI_n^*) = E_{\mathcal{M}^*}(QI_n^{*2}) - E_{\mathcal{M}^*}(QI_n^*)^2$, and therefore we need to compute $E_{\mathcal{M}^*}(QI_n^{*2})$.

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:

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

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

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

by the sampling consistency of $\mathcal{M}^*$.

As far as the second addend in the previous expression for $E_{\mathcal{M}^*}(QI_n^{*2})$ goes, we have

$$\begin{aligned}
S_2 &:= \sum_{T^*\in\mathcal{T}_n^*}\sum_{\substack{(Q,Q')\in\mathcal{P}_4(L(T^*))^2\\ Q\neq Q'}}\Big(\sum_{(i,j)\in[4]^2} q_iq_j\delta(Q;Q_i^*;T^*)\delta(Q';Q_j^*;T^*)\Big)P_n^*(T^*)\\
&= \sum_{(i,j)\in[4]^2} q_iq_j\Bigg[\sum_{T^*\in\mathcal{T}_n^*}\Big(\sum_{k=0}^{3}\sum_{\substack{(Q,Q')\in\mathcal{P}_4(L(T^*))^2\\ |Q\cap Q'|=k}}\delta(Q;Q_i^*;T^*)\delta(Q';Q_j^*;T^*)\Big)P_n^*(T^*)\Bigg]\\
&= \sum_{(i,j)\in[4]^2} q_iq_j\Bigg[\sum_{k=0}^{3}\sum_{T^*\in\mathcal{T}_n^*}\big|\{(Q,Q')\in\mathcal{P}_4(L(T^*))^2 \mid |Q\cap Q'|=k,\\
&\qquad\qquad T^*(Q)=Q_i^*, T^*(Q')=Q_j^*\}\big|\cdot P_n^*(T^*)\Bigg]
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{T^*\in\mathcal{T}_n^*}\big|\{(Q,Q')\in\mathcal{P}_4(L(T^*))^2 \mid |Q\cap Q'|=k, T^*(Q)=Q_i^*, T^*(Q')=Q_j^*\}\big|P_n^*(T^*)\\
&\quad= \sum_{T^*\in\mathcal{T}_n^*}\Big(\sum_{T_{8-k}^*\in\mathcal{T}_{8-k}^*}\big|\{X\in\mathcal{P}_{8-k}(L(T^*)) \mid T^*(X)=T_{8-k}^*\}\big|\\
&\qquad\qquad\cdot\big|\{(Q,Q')\in\mathcal{P}_4(L(T_{8-k}^*))^2 \mid |Q\cap Q'|=k, T_{8-k}^*(Q)=Q_i^*, T_{8-k}^*(Q')=Q_j^*\}\big|\Big)P_n^*(T^*)\\
&\quad= \sum_{T_{8-k}^*\in\mathcal{T}_{8-k}^*}\big|\{(Q,Q')\in\mathcal{P}_4(L(T_{8-k}^*))^2 \mid |Q\cap Q'|=k, T_{8-k}^*(Q)=Q_i^*, T_{8-k}^*(Q')=Q_j^*\}\big|\\
&\qquad\qquad\cdot\binom{n}{8-k}\sum_{T^*\in\mathcal{T}_n^*}\frac{\Big(\big|\{X\in\mathcal{P}_{8-k}(L(T^*)) \mid T^*(X)=T_{8-k}^*\}\big|}{\binom{n}{8-k}}P_n^*(T^*)\\
&\quad= \binom{n}{8-k}\sum_{T_{8-k}^*\in\mathcal{T}_{8-k}^*}\big|\{(Q,Q')\in\mathcal{P}_4(L(T_{8-k}^*))^2 \mid |Q\cap Q'|=k,\\
&\qquad\qquad T_{8-k}^*(Q)=Q_i^*, T_{8-k}^*(Q')=Q_j^*\}\big|P_{8-k}^*(T_{8-k}^*)
\end{aligned}$$

$$= \binom{n}{8-k} \sum_{T^*_{8-k} \in \mathcal{T}^*_{8-k}} \big| \{(Q,Q') \in \mathcal{P}_4(L(T^*_{8-k}))^2 \mid Q \cup Q' = L(T^*_{8-k}),$$
$$T^*_{8-k}(Q) = Q^*_i, T^*_{8-k}(Q') = Q^*_j\} \big| P^*_{8-k}(T^*_{8-k})$$
$$= \binom{n}{8-k} \sum_{T^*_{8-k} \in \mathcal{T}^*_{8-k}} \Theta_{i,j}(T^*_{8-k}) P^*_{8-k}(T^*_{8-k})$$

again by the sampling consistency of $\mathcal{M}^*$. Therefore,

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

The formula in the statement is then obtained by writing $Var_{\mathcal{M}^*}(IQ^*_n)$ as $S_1 + S_2 - E_{\mathcal{M}^*}(IQ^*_n)^2$ and using the expression for $E_{\mathcal{M}^*}(IQ^*_n) = E_{\mathcal{M}}(IQ_n)$ given in Proposition 17. □

Again, if $\mathcal{M} = (P_n)_n$ is a probabilistic model of bifurcating phylogenetic trees, so that $P^*_4(Q^*_4) = P^*_4(Q^*_2) = P^*_4(Q^*_1) = 0$, then, taking $q_3 = 1$, this proposition implies that

$$\begin{aligned} Var_{\mathcal{M}}(QIB_n) = \binom{n}{4} P^*_4(Q^*_3) - \binom{n}{4}^2 P^*_4(Q^*_3)^2 \\ + \sum_{k=5}^{8} \binom{n}{k} \Big(\sum_{T^* \in \mathcal{BT}^*_k} \Theta_{3,3}(T^*) P^*_k(T^*) \Big) \end{aligned}$$

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 2 in the Appendix A.4. We obtain then the following result.

Corollary 22. *If $\mathcal{M}$ is a probabilistic model of bifurcating phylogenetic trees such that $\mathcal{M}^*$ is sampling consistent, then, with the notations for trees given in Table 2 in the Appendix A.4,*

$$\begin{aligned} Var_{\mathcal{M}}(QIB_n) &= \binom{n}{4} P^*_4(Q^*_3) - \binom{n}{4}^2 P^*_4(Q^*_3)^2 \\ &+6\binom{n}{5} P^*_5(B^*_{5,3}) + \binom{n}{6} \big(18P^*_6(B^*_{6,4}) + 6P^*_6(B^*_{6,5}) + 36P^*_6(B^*_{6,6})\big) \\ &+\binom{n}{7} \big(8P^*_7(B^*_{7,8}) + 24P^*_7(B^*_{7,9}) + 36P^*_7(B^*_{7,10}) + 36P^*_7(B^*_{7,11})\big) \\ &+\binom{n}{8} \big(2P^*_8(B^*_{8,13}) + 6P^*_8(B^*_{8,14}) + 12P^*_8(B^*_{8,15}) + 14P^*_8(B^*_{8,16}) \\ &\quad +18P^*_8(B^*_{8,17}) + 36P^*_8(B^*_{8,21}) + 36P^*_8(B^*_{8,22}) + 38P^*_8(B^*_{8,23})\big) \end{aligned}$$

Proposition 21 and Corollary 22 reduce the computation of $Var_{\mathcal{M}}(QI_n)$ or $Var_{\mathcal{M}}(QIB_n)$ 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 QIB_n under the α and the β-models, and for the variance of QI_n under the α-γ-model.

As far as the bifurcating case goes, on the one hand, the probabilities under the α-model of the trees appearing explicitly in the formula for the variance of QIB_n in Corollary 22 are those given in Table 3 in the Appendix A.4 (they are explicitly computed in [15, Suppl. Mat.]). Plugging them in the formula given in Corollary 22 above, we obtain the following result.

Corollary 23. *Under the α-model,*

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

In particular, the leading term in n of $Var_{\mathcal{M}_\alpha}(QIB_n)$ is

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

On the other hand, the probabilities under the β-model of the same trees are given in Table 4 in the Appendix A.4, yielding the following result.

Corollary 24. *Under the β-model,*

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

In particular, the leading term in n of $Var_{\mathcal{M}_\beta}(QIB_n)$ 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 QIB_n reduce to

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

In the Appendix A.4 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 $\beta = 0$, these formulas yield

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

Finally, as far as the α-γ-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 $Var_{\mathcal{M}_{\alpha,\gamma}}(QI_n)$. The Python scripts and the resulting formula (in text format and as a Python script that can be applied to any values of n, α, and γ) can be found in the GitHub page `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 α-model given in Corollary 23.

6. Conclusions

In this paper we have introduced a new balance index for phylogenetic trees, the quartet index QI. This index makes sense for polytomic 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, α, β and α-γ-models. This paper is accompanied by the GitHub page `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 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 non-bifurcating 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, 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 7, which, in our opinion, should be considered as equally balanced. We believe that the QI index can be used to capture the symmetry of a taxonomic tree in a natural way, and we hope to report on it elsewhere.

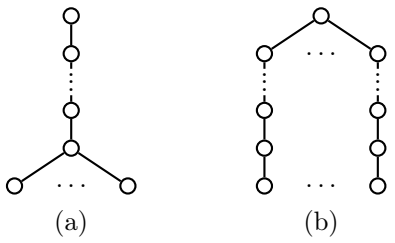


Figure 7: The shapes of the taxonomic trees with maximum (a) and minimum (b) total cophenetic values among all taxonomic trees of given depth and number of leaves.

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 and Seth Sullivant 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] D. Aldous. Probability distributions on cladograms. In *Random Discrete Structures* (D. Aldous and R. Pemantle, eds.), Springer-Verlag (1996), 1–18.

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

[3] M. Blum, O. François. "Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance." Systematic Biology 55 (2006), 685–691.

[4] G. Cardona, A. Mir, F. Rosselló, Exact formulas for the variance of several balance indices under the Yule model. Journal of Mathematical Biology, 67 (2013), 1833–1846.

[5] L. L. Cavalli-Sforza, A. Edwards, Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet., 19 (1967), 233–257.

[6] B. Chen, D. Ford, M. Winkel, A new family of Markov branching trees: the alpha-gamma model. Electron. J. Probab, 14 (2009), 400–430.

[7] D. H. Colless, Review of "Phylogenetics: the theory and practice of phylogenetic systematics". Sys. Zool, 31 (1982), 100–104.

[8] Th. H. Cormen, Ch. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms* (3rd edition). The MIT Press (2009).

[9] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates Inc., 2004.

[10] D. Ford. Probabilities on cladograms: Introduction to the alpha model. arXiv:math/0511246 [math.PR] (2005).

[11] E. Harding, The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3 (1971), 44–77.

[12] S. Keller-Schmidt, M. Tugrul, V. M. Eguiluz, E. Hernandez-Garcia, K. Klemm. "Anomalous scaling in an age-dependent branching model." Physical Review E 91, 022803 (2015).

[13] M. Kirkpatrick, M. Slatkin. "Searching for evolutionary patterns in the shape of a phylogenetic tree." Evolution 47 (1993), 1171–1181.

[14] I. G. Macdonald, *Symmetric functions and Hall polynomials* (2ond ed.). Oxford Mathematical Monographs, Oxford University Press, 1995.

[15] T. Martínez-Coronado, A. Mir, F. Rosselló. On the probabilities of trees and cladograms under Ford's α-model. arXiv:1801.03843 [q-bio.PE]. To appear in The Scientific World Journal (2018).

[16] F. Matsen, Optimization Over a Class of Tree Shape Statistics. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 4 (2007), 506–512.

[17] A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees. Mathematical Biosciences 241 (2013), 125–136.

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

[19] The Newick tree format: `http://evolution.genetics.washington.edu/phylip/newicktree.html`.

[20] M. Petkovsek, H.S. Wilf, D. Zeilberger, $A{=}B$. AK Peters (1996), available at `https://www.math.upenn.edu/~wilf/AeqB.html`.

[21] I. Pinelis. Evolutionary models of phylogenetic trees. Roy. Soc. Lond. Proc. Ser. Biol. Sci., 270 (2003), 1425–1431.

[22] D. E. Rosen, Vicariant Patterns and Historical Explanation in Biogeography. Syst. Biol. 27 (1978), 159–188.

[23] M. J. Sackin, "Good" and "bad" phenograms. Sys. Zool, 21 (1972), 225–226.

[24] K.T. Shao, R. Sokal, Tree balance. Sys. Zool, 39 (1990), 226–276.

[25] N. J. A. Sloane (editor), *The On-Line Encyclopedia of Integer Sequences*, published electronically at `http://oeis.org/`.

[26] M. Steel, A. McKenzie, Distributions of cherries for two models of trees. Math. Biosc. 164 (2000), 81–92.

[27] G. U. Yule, A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. Phil. Trans. Royal Soc. (London) Series B 213 (1924), 21–87.

Appendix: Some computations

A.1: Computation of $P^{A,*}_{\beta,4}(Q^*_3)$

We recall from [1] the definition of the probability $P^{A,*}_{\beta,n}(T)$ of a tree $T \in \mathcal{BT}^*_n$ under the β-model. Let $\beta \in]-2,\infty[$. For every $m \geqslant 2$ 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 $a_m(\beta)$ is a suitable normalizing constant so that $\sum_{a=1}^{m-1} q_{m,\beta}(a) = 1$. For every $m \geqslant 2$ and $a = 1,\dots,\lfloor m/2 \rfloor$, let

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

Then, for a given desired number $n \geqslant 1$ of leaves:

1. Start with a tree T'_1 consisting of a single node labeled n. Let $P'_{\beta,1}(T'_1) = 1$.
2. At each step $j = 1,\dots,n-1$, the current tree T'_j contains leaves with label 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 β-model is
$$P^{A,*}_{\beta,n}(T^*_n) = \sum_{T'_n = T^*_n} P'_{\beta,n}(T'_n)$$

Lemma 25. *For every* $\beta \in]-2,\infty[$,

$$P^{A,*}_{\beta,4}(Q^*_3) = \frac{3\beta+6}{7\beta+18}$$

Proof. To compute $P^{A,*}_{\beta,4}(Q^*_3)$, 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's 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

$$P^{A,*}_{\beta,4}(Q^*_3) = 2 \cdot \frac{1}{2} q_{4,\beta}(2) = \frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}$$

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.

□

A.2: Computation of $P^*_{\alpha,\gamma,4}$

We recall from [6] the definition of the probability $P^*_{\alpha,\gamma,n}(T^*)$ of a tree $T^* \in \mathcal{T}^*_n$ under the α-γ-model. Let $0 \leqslant \gamma \leqslant \alpha \leqslant 1$ and $n \geqslant 1$.

1. Start with the tree $T_1 \in \mathcal{T}_1$ consisting of a single node labeled 1. Let $P_{\alpha,\gamma,1}(T_1) = 1$.

2. For every $m = 1, \ldots, n-1$, let $T_{m+1} \in \mathcal{T}_{m+1}$ be obtained by adding a new leaf labeled $m+1$ to T_m. Then:

   - If T_{m+1} is obtained from T_m by adding the new leaf to an arc e, then:

   $$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{1-\alpha}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m) \text{ if } e \text{ ends in a leaf}$$
   $$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{\gamma}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m) \text{ if } e \text{ ends in an internal node}$$

   - If T_{m+1} is obtained from T_m by adding the new leaf to a new root, then

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

   - If T_{m+1} is obtained from T_m by adding the new leaf as a child of an internal node u, then

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

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

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

Lemma 26. *With the notations of Figure 3:*

$$P^*_{\alpha,\gamma,4}(Q^*_0) = \frac{2(1-\alpha+\gamma)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q^*_1) = \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q^*_2) = \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q^*_3) = \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q^*_4) = \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$

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 tree in $\mathcal{T}_3$ is obtained by adding a leaf labeled 3 to T_2. These trees are described in Figure 8. Their probabilities are:

- 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. 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

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

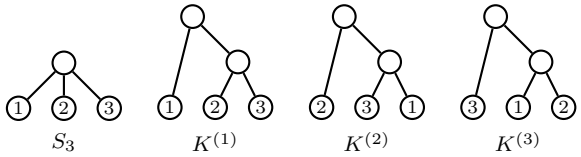


Figure 8: The phylogenetic trees in $\mathcal{T}_3$.

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,

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

- 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,

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

□

A.3: An alternative derivation of the variance of QIB_n under the Yule model

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

Proposition 27. *Under the Yule model,*

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

Proof. By Lemma 7, QIB on $\mathcal{BT}_n$ is a bifurcating recursive tree shape statistic satisfying the recurrence

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

with $f_{QIB}(a,b)=\binom{a}{2}\binom{b}{2}$. Then, it satisfies the hypothesis in [4, Cor. 1] with

$$\begin{aligned}\varepsilon(a,b-1) &= f_{QIB}(a,b)-f_{QIB}(a,b-1)=\binom{a}{2}\binom{b}{2}-\binom{a}{2}\binom{b-1}{2}=(b-1)\binom{a}{2}\\ R(n-1) &= E_{\mathcal{M}}(QIB_n)-E_{\mathcal{M}}(QIB_{n-1})=\frac{1}{3}\binom{n}{4}-\frac{1}{3}\binom{n-1}{4}=\frac{1}{3}\binom{n-1}{3}\end{aligned}$$

Since $E_Y(QIB_1)=0$ and $f_{QIB}(n-1,1)=0$, applying the aforementioned result from [4] we have that

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

Dividing by n both sides of this expression for $E_Y(QIB_n^2)$ and setting $y_n=E_Y(QIB_n^2)/n$, 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

$$E_Y(QIB_n^2)=ny_n=\binom{n}{4}\frac{1265n^4-7110n^3+14419n^2-4086n+5040}{264600}$$

Finally

$$\begin{aligned} Var_{\mathcal{M}}(QIB_n) &= E_Y(QIB_n^2) - E_Y(QIB_n)^2 \\ &= \binom{n}{4} \frac{1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040}{264600} - \frac{1}{9}\binom{n}{4}^2 \\ &= \binom{n}{4} \frac{5n^4 + 30n^3 + 118n^2 + 408n + 630}{33075} \end{aligned}$$

as we claimed. □

A.4: Some tables used in §5

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 2: Coefficients of the probabilities of the trees in $\mathcal{BT}^*_k$, for $k = 5, 6, 7, 8$, in the formula for the variance of QIB_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 3: Probabilities under the α-model of the trees involved in the formula for the variance of QIB_n

Tree	$P^{B,*}_{\beta,n}$
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 4: Probabilities under the β-model of the trees involved in the formula for the variance of QIB_n