The distributions under two species-tree models of the total number of ancestral configurations for matching gene trees and species trees

Filippo Disanto[*], Michael Fuchs[†], Chun-Yen Huang[‡], Ariel R. Paningbatan[§], Noah A. Rosenberg[¶]

July 4, 2023

Abstract

Given a gene-tree labeled topology G and a species tree S, the *ancestral configurations* at an internal node k of S represent the combinatorially different sets of gene lineages that can be present at k when all possible realizations of G in S are considered. Ancestral configurations have been introduced as a data structure for evaluating the conditional probability of a gene-tree labeled topology given a species tree, and their enumeration assists in describing the complexity of this computation. In the case that the gene-tree labeled topology $G = t$ matches that of the species tree S, by techniques of analytic combinatorics, we study distributional properties of the *total* number of ancestral configurations measured across the different nodes of a random labeled topology t selected under the uniform and the Yule probability models. Under both of these probabilistic scenarios, we show that the total number T_n of ancestral configurations of a random labeled topology of n taxa asymptotically follows a lognormal distribution. Over uniformly distributed labeled topologies, the asymptotic growth of the mean and the variance of T_n are found to satisfy $\mathbb{E}_{\mathrm{U}}[T_n] \sim 2.449{\cdot}1.333^n$ and $\mathbb{V}_{\mathrm{U}}[T_n] \sim 5.050{\cdot}1.822^n$, respectively. Under the Yule model, which assigns higher probabilities to more balanced labeled topologies, we obtain the mean $\mathbb{E}_{\mathrm{Y}}[T_n] \sim 1.425^n$ and the variance $\mathbb{V}_{\mathrm{Y}}[T_n] \sim 2.045^n$.

Keywords: analytic combinatorics, ancestral configurations, gene trees, phylogenetics, species trees.

Mathematics subject classification (2010): 05A15 · 05A16 · 05C05 · 92D15

1 Introduction

Ancestral configurations are lists that describe for a given gene-tree topology G and a species tree S the sets of gene lineages that can be present at a given node of S (Fig. 1). They have been introduced by Wu [35] as a data structure in the calculation of the probabilities of gene-tree topologies conditional on species trees under the multispecies coalescent model. In particular, for a given species tree S, Wu's algorithm "STELLS" evaluates the probability of a gene-tree topology G by recursively computing the probabilities of the ancestral configurations of G at the different nodes of S, proceeding from the tips towards the root of S [35]. The running time of STELLS depends on the total number $c(G, S)$ of ancestral configurations of G in S, that is, on the sum of the number of ancestral configurations of G across the different nodes of S.

[*] Department of Mathematics, University of Pisa, Pisa 56126, Italy. Email: filippo.disanto@unipi.it.

[†] Department of Mathematical Sciences, National Chengchi University, Taipei 116, Taiwan. Email: mfuchs@nctu.edu.tw.

[‡] Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan.

[§] Institute of Mathematics, University of the Philippines Diliman, Quezon City 1101, Philippines. Email: arpaningbatan@math.upd.edu.ph.

[¶] Department of Biology, Stanford University, Stanford, CA 94305, USA. Email: noahr@stanford.edu.

If the topology $G = t$ of the gene tree matches the topology of species tree S, then the total number of configurations of G in S becomes a function $c(G, S) = c(t)$ of t, whose behavior over tree families of increasing size can be analyzed by tools of enumerative and analytic combinatorics [18]. In initial studies [7, 11], by examining the number of ancestral configurations at the root of a randomly selected topology t with number of leaves n, we derived theorems that determine the exponential growth—indicated here by the symbol "$\bowtie$"—of the mean $\mathbb{E}[c]$ and the variance $\mathbb{V}[c]$ of the total number of configurations. In particular, we found that the mean grows exponentially like $\mathbb{E}_{\mathrm{U}}[c] \bowtie 1.333^n$ and $\mathbb{E}_{\mathrm{Y}}[c] \bowtie 1.425^n$ for uniformly and Yule-distributed topologies of size n, respectively. Under the same distributions, the exponential growth of the variance satisfies $\mathbb{V}_{\mathrm{U}}[c] \bowtie 1.822^n$ and $\mathbb{V}_{\mathrm{Y}}[c] \bowtie 2.045^n$.

These results, however, do not fully characterize the *sub*exponential growth of the mean and variance of the random variable c, and the problem of describing the asymptotic distribution of the total number of configurations has remained open. Here, we solve these problems by using generating functions to count the total number of ancestral configurations in random tree topologies. Surprisingly, we find that up to a constant factor—which we calculate exactly—the exponential growth of $\mathbb{E}_{\mathrm{U}}[c]$, $\mathbb{E}_{\mathrm{Y}}[c]$, $\mathbb{V}_{\mathrm{U}}[c]$, and $\mathbb{V}_{\mathrm{Y}}[c]$ determines the full asymptotic behavior of the associated quantities. More precisely, for random topologies of increasing size n selected under the uniform and Yule distributions, we show that $\mathbb{E}_{\mathrm{U}}[c] \sim 2.449 \cdot 1.333^n$, $\mathbb{E}_{\mathrm{Y}}[c] \sim 1.425^n$, $\mathbb{V}_{\mathrm{U}}[c] \sim 5.050 \cdot 1.822^n$, and $\mathbb{V}_{\mathrm{Y}}[c] \sim 2.045^n$. Furthermore, under both the uniform and Yule models, we obtain an asymptotic lognormal distribution of the total number of ancestral configurations. We study the correlation between the total number of configurations and the closely related number of root ancestral configurations.

Our approach uses standard techniques of analytic combinatorics for deriving the asymptotic growth of integer sequences coupled with a key observation that enables the study of distributional properties of the number of ancestral configurations of random tree topologies selected under the uniform and Yule models by equivalently considering uniformly distributed classes of plane trees often known as *Catalan trees* and *increasing binary trees*. The results contribute to the enumerative study of combinatorial structures in the relationship between gene trees and species trees [2, 6, 8, 9, 10, 12, 13, 20, 25, 26, 27, 28], and they can assist in relating the complexity of algorithms for computing gene-tree probabilities with ancestral configurations to algorithms that use an evaluation based on other structures [6, 33, 36].

2 Preliminaries

We start with some definitions, preliminary results, and basic principles of enumerative combinatorics. In Section 2.1, we introduce labeled topologies and their uniform and Yule probability distributions. In Section 2.2, we present generating function techniques for use in analyzing the asymptotic growth of integer sequences.

2.1 Labeled and unlabeled topologies

A *labeled topology* t, also called a *phylogenetic tree*, of size $|t| = n$ is a binary rooted tree whose n external nodes—its leaves—possess distinct labels, often for small n the first n letters alphabetically (Fig. 1A). In this study, lower-case t always denotes a tree. Labeled topologies are *non-plane*, or *unordered*, in the sense that each pair of child nodes carries no left-right orientation; we obtain the same labeled topology by transposing the two subtrees stemming from an internal node.

It is convenient to denote the internal nodes of a labeled topology t by letters different from those associated with the leaves (Fig. 1B). We identify each edge of t by (the label of) its immediate descendant node, i.e., by the node closer to the leaves that is adjacent to the edge. We describe the descendant–ancestor order relation defined over the set of nodes of t by the symbol $\preceq$. More precisely, for distinct nodes x and y, we write $x \prec y$ in t if y is a node belonging to the path connecting node x to the root of t. The subtree of t rooted at node k, which contains those nodes x of t with $x \preceq k$, is denoted by t^k. Hence, in particular, $t^k = t$ if k is the root of t, and $t^k = \bullet_k$ if k is a leaf of t, where $\bullet_k$ is a subtree that contains only node k.

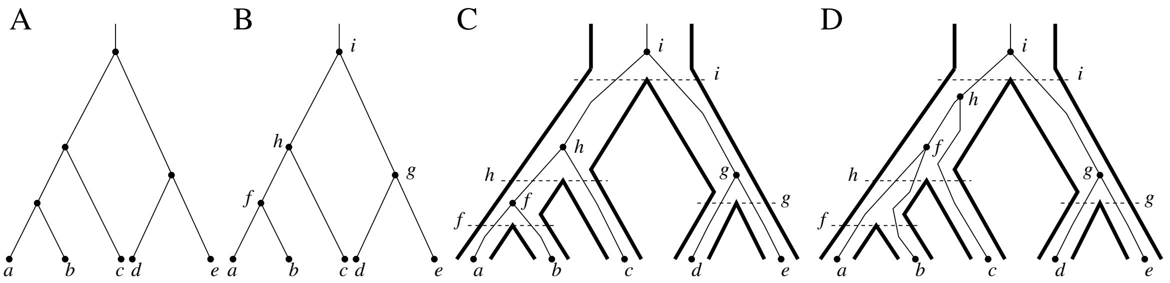


Figure 1: Labeled topologies, gene trees, and species trees. **(A)** A labeled topology of size 5. **(B)** The labeled topology in (A) with its internal nodes labeled. We identify each edge of the tree by its immediate descendant node; for example, lineage h results from the coalescence of lineages c and f. **(C)** A possible realization (thin lines) of the gene-tree labeled topology of (A) in a species tree with a matching labeled topology (thick lines). The ancestral configuration at species-tree node i is $\{g, h\}$. The configuration at node h is $\{c, f\}$. **(D)** A different realization of the gene-tree labeled topology in (A) in a matching species tree. The ancestral configurations at species-tree nodes i and h are $\{g, h\}$ and $\{a, b, c\}$, respectively.

By removing labels from a labeled topology t, we obtain the tree *shape* or *unlabeled topology* underlying t. Unlabeled topologies are also called *Otter trees* [23]; for increasing numbers of leaves $n \geqslant 1$, they are enumerated by the Wedderburn-Etherington numbers, $1, 1, 1, 2, 3, 6, 11, 23, 46, \ldots$ [17, 34].

Distinct labeled topologies t_1 and t_2 can possibly share the same unlabeled topology. For instance, in *Newick format*, labeled topologies $t_1 = ((a, b), c)$ and $t_2 = (a, (b, c))$ share unlabeled topology $((\bullet, \bullet), \bullet) = (\bullet, (\bullet, \bullet))$. The number $\mathrm{lab}(t)$ of labeled topologies with shape t is obtained recursively by eq. (22) of [7],

$$\mathrm{lab}(t) = \mathrm{lab}(t_L)\,\mathrm{lab}(t_R)\binom{|t|}{|t_L|}\frac{1}{1 + \delta_{t_L = t_R}}. \tag{1}$$

Here, t_L and t_R are the two subtrees stemming from the root of t (*root subtrees*, for short) and $\delta_{t_L = t_R}$ is the Kronecker delta that equals 1 if t_L and t_R are the same unlabeled topology. We set lab(t) = 1 if $|t| = 1$.

Let L_n denote the set of labeled topologies of size n. For $n \geqslant 2$, the cardinality of L_n is $|L_n| = (2n-3)!! = 1 \times 3 \times 5 \times \ldots \times (2n-3)$ [16, 17], which can be written

$$|L_n| = \frac{(2n)!}{2^n(2n-1)n!}. \tag{2}$$

Different probability models can be considered over the set L_n of labeled topologies of fixed size n [1]. Under the uniform model, each labeled topology $t \in L_n$ has equal probability

$$\mathbb{P}_U[t] = \frac{1}{|L_n|} = \frac{2^n(2n-1)n!}{(2n)!}.$$

The Yule model is a generative model in which each lineage is equally likely to be the next to bifurcate forward in time, or equivalently, each pair of lineages is equally likely to be next to merge back in time [19, 31, 37]. Many of its combinatorial features have been studied [5, 14, 15, 22, 24]; a labeled topology t of n leaves has probability

$$\mathbb{P}_Y[t] = \frac{2^{n-1}}{n! \prod_{r=3}^{n}(r-1)^{d_r(t)}} \tag{3}$$

under the Yule model, where $d_r(t)$ is the number of nodes of t with r descending leaves [4, 21, 32].

Owing to the product appearing in the denominator, under the Yule distribution, more balanced labeled topologies tend to have larger probabilities [19]. For example, among the labeled topologies of size 5, the one depicted in Fig. 1A has maximal Yule probability $\mathbb{P}[(((a, b), c), (d, e))] = \frac{1}{60}$; taking $((((a, b), c), d), e)$ and $(((a, b), (c, d)), e)$ as representative labeled topologies for their unlabeled shapes, we have $\mathbb{P}[((((a, b), c), d), e)] = \frac{1}{180}$ and $\mathbb{P}[(((a, b), (c, d)), e)] = \frac{1}{90}$.

2.2 Asymptotic growth and generating functions

This article studies the growth of non-negative integer sequences. We use the following notation. For two sequences $(a_n)_{n\geqslant 0}$ and $(b_n)_{n\geqslant 0}$, we write $a_n \sim b_n$ when the ratio b_n/a_n converges to 1 for $n \to \infty$. If $a_n \sim b_n$, then we say that, asymptotically, sequences a_n and b_n have the same growth. The sequence a_n is said to have exponential growth k^n or, equivalently, to be of exponential order k, when $a_n \sim k^n s(n)$, where k is a constant and $s(n)$ is a subexponential factor. We write $a_n \bowtie b_n$ if a_n and b_n have the same exponential growth.

The generating function of a sequence $(a_n)_{n\geqslant 0}$ is the power series $A(z) = \sum_{n=0}^{\infty} a_n z^n$. Multiplying $A(z)$ by a generating function $B(z) = \sum_{n=0}^{\infty} b_n z^n$ gives the generating function $A(z)\,B(z) = \sum_{n=0}^{\infty} \sum_{j=0}^{n} a_j z^j \cdot b_{n-j} z^{n-j} = \sum_{n=0}^{\infty} (\sum_{j=0}^{n} a_j b_{n-j}) z^n$, whose nth coefficient is the convolution $\sum_{j=0}^{n} a_j b_{n-j}$. Also, if k is a constant, then $A(z) + B(z) = \sum_{n=0}^{\infty} (a_n + b_n) z^n$ and $kA(z) = \sum_{n=0}^{\infty} (k a_n) z^n$.

If $A(z) = \sum_{n=0}^{\infty} a_n z^n$ is considered as a function of the complex variable z, then the analysis of $A(z)$ near its singularities—the points in the complex plane where $A(z)$ is not analytic—can assist in the study of the asymptotic growth of the coefficients $a_n = [z^n]A(z)$. The simplest scenario is when $A(z)$ has a unique dominant singularity $\alpha > 0$, that is, when $z = \alpha > 0$ is the only singularity of $A(z)$ of smallest modulus. In this case, under the fairly general conditions of Theorems IV.7 and VI.4 of [18], the singular expansion $A_\alpha(z)$ of the generating function $A(z)$ at $z = \alpha$ determines the asymptotic growth of the coefficients a_n as

$$a_n \sim [z^n]A_\alpha(z) \bowtie \alpha^{-n}. \tag{4}$$

In other words, the nth coefficient of $A(z)$ has for increasing n the same growth as the nth coefficient of the expansion $A_\alpha(z)$, where, in particular, $\frac{1}{\alpha}$ is the exponential order of sequence a_n. For instance, as given by Example II.19 of [18], $L(z) = 1 - \sqrt{1-2z}$ is the generating function associated with the sequence $|L_n|/n!$—where L_n is the number of labeled topologies of size n (eq. 2). The dominant singularity of $L(z)$ is $\alpha = \frac{1}{2}$, and indeed, in agreement with eq. (4), we have $|L_n|/n! = \binom{2n}{n}/[2^n(2n-1)] \bowtie 2^n = \alpha^{-n}$.

In the following sections, we apply the asymptotic relation in eq. (4) to generating functions $A(z)$ with a unique dominant singularity $\alpha > 0$ and singular expansion given by either $A_\alpha(z) = 1/(1 - \frac{z}{\alpha})$ or $A_\alpha(z) = k_1 - k_2\sqrt{1 - z/\alpha}$, where k_1 and $k_2 > 0$ are constants. Using the "$\sim$" equivalence in eq. (4) with $[z^n]\big(1/(1-z)\big) = 1$ and $[z^n]\big(-\sqrt{1-z}\big) \sim 1/(2\sqrt{\pi n^3})$ [18, p. 388], these expansions yield

$$A_\alpha(z) = \frac{1}{1-\frac{z}{\alpha}} \quad \Rightarrow \quad a_n \sim \alpha^{-n} \tag{5}$$

$$A_\alpha(z) = k_1 - k_2\sqrt{1-\frac{z}{\alpha}} \quad \Rightarrow \quad a_n \sim k_2\,\frac{\alpha^{-n}}{2\sqrt{\pi n^3}}. \tag{6}$$

If $A'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}$ and $\int_0^z A(t)\,dt = \sum_{n=0}^{\infty} [a_n/(n+1)] z^{n+1}$ are generating functions obtained by differentiating and integrating generating function $A(z)$, then we use Theorems VI.8 and VI.9(ii) of [18] to calculate the singular expansions of $A'(z)$ and $\int_0^z A(t)\,dt$ at their dominant singularity α as

$$A'(z) \overset{z\to\alpha}{\sim} \Big(A_\alpha(z)\Big)' \tag{7}$$

$$\int_0^z A(t)\,dt \overset{z\to\alpha}{\sim} \int_0^z A_\alpha(t)\,dt. \tag{8}$$

In particular, both $A'(z)$ and $\int_0^z A(t)\,dt$ have the same dominant singularity α as $A(z)$, and their singular expansions are obtained by respectively differentiating and integrating the singular expansion $A_\alpha(z)$ of $A(z)$. Note that we apply eq. (8) to functions $A(z)$ with singular expansion $A_\alpha(z) = 1/(1-\frac{z}{\alpha})^2$, in agreement with the hypothesis of case (i) of Theorem VI.9 of [18]. Upon integrating this $A_\alpha(z)$, we recover an expansion of the form $\alpha/(1-\frac{z}{\alpha})$ for generating function $\int_0^z A(t)\,dt$, whose asymptotic coefficients can be examined by use of eq. (5).

3 Ancestral configurations for matching gene trees and species trees

In this section, we define ancestral configurations for matching gene trees and species trees (Section 3.1) and explain how distributional properties of the number of ancestral configurations can be equivalently studied over labeled topologies and over other tree families (Section 3.2). Section 3.3 reviews results of [7] and [11] on ancestral configurations at the root of randomly selected labeled topologies.

3.1 Definitions and examples

We first introduce gene trees, species trees, and realizations of a gene tree in a species tree. Following [35], we define ancestral configurations for pairs of gene trees and species trees that share the same labeled topology.

3.1.1 Gene trees and species trees

A *species-tree* labeled topology represents the evolutionary relationships of a set of populations or species identified with the leaves of the tree. A *gene-tree* labeled topology describes the evolution of (genomic regions of) individuals sampled from a set of populations or species.

If individuals are sampled from the populations considered at the leaves of the species tree, then the gene tree can be viewed from a biological perspective as a set of gene lineages (Fig. 1C, thin lines) that have spread by evolutionary forces within the branching structure of the species tree (Fig. 1C, thick lines). We assume that exactly one individual is sampled for each population at the leaves of the species tree. We only examine pairs of *matching* species trees and gene trees, that is, pairs with the same labeled topology (Fig. 1C and D).

For a fixed species tree, the same gene tree can result from different instances, or realizations, of the evolutionary process. In panels C and D of Fig. 1, the gene-tree labeled topology of panel A—with internal nodes denoted as in panel B—is depicted within the species tree as an outcome of two realizations that differ in the choice of the branches (edges) of the species tree where the coalescent events (internal nodes) of the gene tree occur. In particular, switching from panel C to panel D, we find gene-tree coalescent event f in two different species-tree branches. In mathematical terms, a *realization* of gene tree G in species tree S with matching labeled topology $S = G = t$ is a function R mapping the set of internal nodes of t onto itself such that two conditions hold: (i) for all internal nodes k, $k \preceq R(k)$, and (ii) for all internal nodes k_1 and k_2, $k_1 \preceq k_2 \Rightarrow R(k_1) \preceq R(k_2)$. By identifying each species-tree branch by its immediate descendant node, the coalescent event corresponding to internal node k of the gene tree G is specified by the realization R to occur in branch $R(k)$ of the species tree. For example, the realization that encodes the evolutionary scenario in Fig. 1C is $R(k) = k$ for all $k \in \{f, g, h, i\}$, whereas in Fig. 1D, the realization instead has $R(f) = R(h) = h$, $R(g) = g$, and $R(i) = i$.

3.1.2 Ancestral configurations

When species trees are equipped with branch lengths that measure the time separating pairs of adjacent nodes, the conditional probability of a gene-tree labeled topology for a given species tree can be calculated under the multispecies coalescent model [6]. Ancestral configurations were introduced by Wu [35] as a data structure for the recursive calculation of this conditional probability, with each node of the species tree associated with a given set of ancestral configurations depending on the possible realizations of the gene tree. At each step, Wu's algorithm computes the probability under the coalescent model that an ancestral configuration at a given node of the species tree has "evolved" from the ancestral configurations at its child nodes, proceeding recursively from the leaves to the root. The cost of Wu's algorithm is affected by the total number of ancestral configurations measured across all nodes of the species tree.

In our setting, where the gene-tree labeled topology matches that of the species tree, ancestral configurations are defined as follows. Given a gene-tree labeled topology $G = t$ and a matching species tree S, let R be a realization of G in S. For a given node k of S, consider the set $C(k) = C(k, R)$ of gene lineages (edges of G)

present in S at the time point right before node k, when time flows from the leaves toward the species-tree root. The set $C(k)$ is called the *ancestral configuration* of the gene tree at species-tree node k under realization R. For example, in the realization of Fig. 1C, the ancestral configurations at the species-tree internal nodes are $C(f) = \{a, b\}$, $C(g) = \{d, e\}$, $C(h) = \{c, f\}$, and $C(i) = \{g, h\}$, where each gene lineage is identified by its immediate descendant node. In the realization of Fig. 1D, the ancestral configuration at the internal node of S denoted by h is $C(h) = \{a, b, c\}$; at the other nodes, the ancestral configurations follow the previous case.

Let $\mathbf{R}(t)$ be the set of possible realizations of the gene-tree labeled topology $G = t$ in the matching species tree S. For a given node k of S, by considering all possible realizations $R \in \mathbf{R}(t)$, we define the set

$$C_k = C_k(t) = \{C(k, R) : R \in \mathbf{R}(t)\}, \tag{9}$$

with cardinality

$$c_k = c_k(t) = |C_k|. \tag{10}$$

Thus, c_k counts ways that the gene lineages of G can reach the time point right before node k in S, when all realizations of the gene-tree labeled topology $G = t$ in S are considered. For instance, taking t as in Fig. 1A, $C_f = \{\{a, b\}\}$, $C_g = \{\{d, e\}\}$, $C_h = \{\{a, b, c\}, \{c, f\}\}$, and $C_i = \{\{a, b, c, d, e\}, \{c, d, e, f\}, \{d, e, h\}, \{a, b, c, g\}, \{c, f, g\}, \{g, h\}\}$, for a total of 10 ancestral configurations.

Note that from the definition of ancestral configuration, $\{k\} \notin C_k$. Indeed, gene lineages can coalesce to produce node k of G only in the part of the species tree above node k. For consistency with this observation, we set $c_k = 0$ if node k is a leaf. In the original paper of Wu [35], for a leaf node k, the conventions $C_k = \{k\}$ and $c_k = 1$ were chosen. Other than that choice, our definition of the set of ancestral configurations C_k at a node k matches that of Wu, and the sum of the number of ancestral configurations across the nodes of a tree differs only by a constant term given by the tree size. The asymptotic results that we derive in later sections also hold under the original setting of Wu [35].

The set $C_k \cup \{\{k\}\}$ can be viewed as the set of maximal antichains of the subtree t^k of t rooted at node k. In particular, if r is the root of t, then $C_r \cup \{\{r\}\}$ corresponds to the set of maximal antichains of t. An antichain of subtree t^k is indeed a subset of its nodes—possibly including the leaves—whose elements are pairwise incomparable with respect to the descendant–ancestor order relation $\preceq$ defined in t. An ancestral configuration of C_k is a "maximal" antichain of t^k in the sense that it is not properly contained in any other antichain of t^k.

By summing the number c_k for k ranging over the set $N(t)$ of nodes of a labeled topology t, we find the total number of ancestral configurations of t, which we denote by

$$c = c(t) = \sum_{k \in N(t)} c_k. \tag{11}$$

Equivalently, $c + 2|t| - 1$ is the total number of maximal antichains across subtrees of t, including the $|t|$ leaves in $N(t)$ and for counts at each of the $|t| - 1$ internal nodes, including as a maximal antichain the node itself.

For a gene-tree labeled topology $G = t$ and matching species tree S, the total number of ancestral configurations $c(t)$ of t is computed recursively by decomposing t in its left and right root subtrees t_L and t_R (once we fix an embedding of t into the plane). If $c_r(t)$ denotes the number of ancestral configurations at the root of t, then

$$\begin{aligned} c(t) &= c(t_L) + c(t_R) + c_r(t) && (12)\\ c_r(t) &= [c_r(t_L) + 1][c_r(t_R) + 1], && (13) \end{aligned}$$

where $c(t) = c_r(t) = 0$ for $|t| = 1$ [11].

For example, suppose t is the labeled topology of Fig. 1A, with internal nodes denoted as in Fig. 1B. Recalling

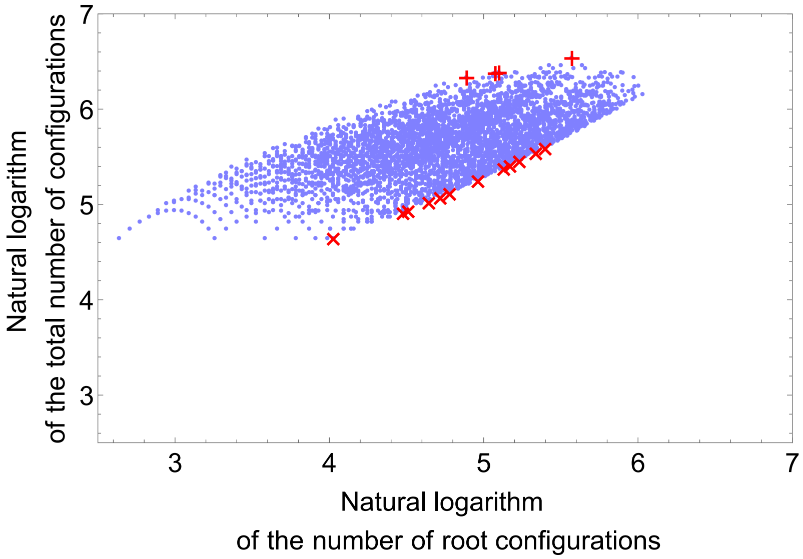


Figure 2: Natural logarithms of the total number of ancestral configurations and the number of root ancestral configurations for representative labelings of each of the 4850 unlabeled topologies of size $n = 15$ leaves: Topologies at which the difference between the total number of configurations and the number of root configurations attains its maximal value of 436 and minimal value of 49 are indicated by + and × coordinates, respectively; the topologies themselves appear in Table 1.

that t^k refers to the subtree of t rooted at node k, by applying eqs. (12) and (13), we find

$$\begin{aligned}
c(t) &= c(t^h) + c(t^g) + c_r(t) \\
&= [c(t^f) + \cancel{c(t^c)} + c_r(t^h)] + [\cancel{c(t^d)} + \cancel{c(t^e)} + c_r(t^g)] + [c_r(t^h) + 1][c_r(t^g) + 1] \\
&= c(t^f) + c_r(t^h) + c_r(t^g) + ([c_r(t^f) + 1][\cancel{c_r(t^c)} + 1] + 1)([\cancel{c_r(t^d)} + 1][\cancel{c_r(t^e)} + 1] + 1) \\
&= [\cancel{c(t^a)} + \cancel{c(t^b)} + c_r(t^f)] + [c_r(t^f) + 1][\cancel{c_r(t^c)} + 1] + [\cancel{c_r(t^d)} + 1][\cancel{c_r(t^e)} + 1] + [c_r(t^f) + 2]\cdot 2 \\
&= 4c_r(t^f) + 6 \\
&= 4[\cancel{c_r(t^a)} + 1][\cancel{c_r(t^b)} + 1] + 6 \\
&= 10.
\end{aligned}$$

When the labeled topology t has size n, the total number c of ancestral configurations can be bounded by means of the number c_r of root ancestral configurations as

$$c_r \leqslant c \leqslant (2n - 1)c_r. \tag{14}$$

Indeed, there are $|N(t)| = 2n - 1$ nodes in t and, for every node k of t, we have $c_k \leqslant c_r$.

Because c and c_r differ by a factor that is at most polynomial in the tree size n, they have the same exponential order when measured across tree families of increasing size. Based on this observation, the studies of [7] and [11] of the asymptotic growth of the number of root ancestral configurations in random trees obtained the exponential order of the mean and variance of the total number of ancestral configurations in labeled topologies of size n selected at random under the uniform and Yule distributions. In Section 4, we refine these results, obtaining full asymptotic distributions of the total number of ancestral configurations. We also study the correlation between the total number of ancestral configurations and the number of root ancestral configurations in random labeled topologies of increasing size.

Fig. 2 shows on a log scale the total number of ancestral configurations and the number of root ancestral configurations for representative labelings of each unlabeled topology of size $n = 15$. The figure illustrates that

Table 1: Representative labeled topologies that achieve the maximal and minimal values of the difference $c(t)-c_r(t)$ between the total number of configurations and the number of root configurations, among trees t with size $n=15$ leaves.

Labeled topology	Total number of	Number of root	
	configurations	configurations	Difference
t	$c(t)$	$c_r(t)$	$c(t)-c_r(t)$
$(((((((a,b),(c,d)),((e,f),(g,h))),((i,j),(k,\ell))),m),n),o)$	569	133	436
$(((((((a,b),(c,d)),((e,f),(g,h))),((i,j),k)),(\ell,m)),n),o)$	596	160	436
$(((((((a,b),(c,d)),((e,f),g)),((h,i),(j,k))),(\ell,m)),n),o)$	600	164	436
$((((((a,b),(c,d)),((e,f),(g,h))),((i,j),(k,\ell))),(m,n)),o)$	699	263	436
$((((((((a,b),c),d),e),f),g),h),((((((i,j),k),\ell),m),n),o))$	105	56	49
$((((((((a,b),c),d),e),f),g),h),(((((i,j),k),\ell),m),(n,o)))$	137	88	49
$(((((((a,b),c),d),e),f),(g,h)),((((((i,j),k),\ell),m),n),o))$	140	91	49
$((((((((a,b),c),d),e),f),g),h),((((i,j),k),\ell),((m,n),o)))$	153	104	49
$((((((a,b),c),d),e),((f,g),h)),((((((i,j),k),\ell),m),n),o))$	161	112	49
$(((((a,b),c),d),(((e,f),g),h)),((((((i,j),k),\ell),m),n),o))$	168	119	49
$(((((((a,b),c),d),e),f),(g,h)),(((((i,j),k),\ell),m),(n,o)))$	192	143	49
$(((((((a,b),c),d),e),f),(g,h)),((((i,j),k),\ell),((m,n),o)))$	218	169	49
$((((((a,b),c),d),e),((f,g),h)),(((((i,j),k),\ell),m),(n,o)))$	225	176	49
$(((((a,b),c),d),(((e,f),g),h)),(((((i,j),k),\ell),m),(n,o)))$	236	187	49
$((((((a,b),c),d),e),((f,g),h)),((((i,j),k),\ell),((m,n),o)))$	257	208	49
$(((((a,b),c),d),(((e,f),g),h)),((((i,j),k),\ell),((m,n),o)))$	270	221	49

Each pair $\big(c_r(t),c(t)\big)$ is associated (on a log scale) with a point in Fig. 2.

the total number of ancestral configurations exceeds the number of root ancestral configurations. It also shows that the two quantities are positively correlated across trees. The total number of ancestral configurations and the number of root ancestral configurations are often close; their maximal difference occurs for labeled topologies that are relatively balanced at many nodes (Table 1). The corresponding minimum occurs for trees that are balanced at the root but caterpillar-like in the two root subtrees.

3.2 Ordered tree families and equivalent probability models of ancestral configurations

The definition in eq. (9) of the set of ancestral configurations at a node of a labeled topology t depends only on the shape of t. Ancestral configurations as well as the quantities in eqs. (10) and (11) can be defined in the same way for many types of bifurcating rooted trees t (e.g. labeled, unlabeled, ordered, unordered). This section explains that probabilistic properties of the number of ancestral configurations considered over random labeled topologies can be equivalently analyzed over different tree families. In Sections 3.2.1 and 3.2.2, we introduce the families of ordered unlabeled topologies and ordered unlabeled histories. Next, in Section 3.2.3, we recall some equivalence results of [7] on the distribution of the number of ancestral configurations. In particular, Lemma 1 states that the number of ancestral configurations has the same distribution when considered over uniformly

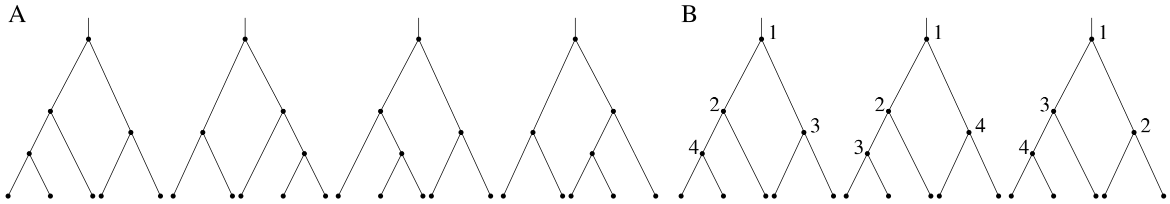


Figure 3: Ordered unlabeled topologies and ordered unlabeled histories. **(A)** The four possible ordered unlabeled topologies whose underlying unordered unlabeled topology is $(((\bullet,\bullet),\bullet),(\bullet,\bullet))$. **(B)** The three possible ordered unlabeled histories whose underlying ordered unlabeled topology matches the leftmost shape in (A).

distributed labeled topologies and over uniformly distributed ordered unlabeled topologies of the same size. The equivalence extends to the distribution of the number of ancestral configurations over Yule-distributed labeled topologies and uniformly distributed ordered unlabeled histories. We conclude the section with Lemma 2, a preliminary result to the calculations of Section 4.

3.2.1 Ordered unlabeled topologies

An *ordered* unlabeled topology is a binary rooted plane tree, that is, an unlabeled topology t equipped with a left–right orientation of the subtrees descending from its internal nodes. Note that the term "ordered" does not refer here to a temporal ordering of the internal nodes of the tree; it only specifies that these nodes have two different types of descendants: left and right. The tree *shape* of an ordered unlabeled topology is the underlying unordered unlabeled topology. Each ordered unlabeled topology is an embedding of its shape into the plane.

Fig. 3A depicts the four ordered unlabeled topologies with shape $(((\bullet,\bullet),\bullet),(\bullet,\bullet))$. Denoting by $\mathrm{out}(t)$ the number of ordered unlabeled topologies with shape t, eq. (23) of [7] gives

$$\mathrm{out}(t) = 2\,\mathrm{out}(t_L)\,\mathrm{out}(t_R)\,\frac{1}{1+\delta_{t_L=t_R}}, \tag{15}$$

where $\mathrm{out}(t) = 1$ if $|t| = 1$.

Ordered unlabeled topologies are also called *Catalan trees* as they are enumerated, with respect to the size n, by the $(n-1)$-th Catalan number C_{n-1} ([30], Exercise 6.19d), where

$$C_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n^3}}. \tag{16}$$

The generating function of the sequence C_n is

$$C(z) = \sum_{n=0}^{\infty} C_n z^n = \frac{1-\sqrt{1-4z}}{2z}. \tag{17}$$

$C(z)$ has singular expansion $C(z) \overset{z\to\alpha}{\sim} 2 - 2\sqrt{1-4z}$ at its dominant singularity $\alpha = \frac{1}{4}$, as can be seen by setting $z = \frac{1}{4}$ in the denominator; by eq. (6), we obtain the asymptotic expression in eq. (16).

A decomposition provides a useful formula for the probability that the left and right root subtrees of a uniformly selected ordered unlabeled topology of n leaves have sizes j and $n-j$, respectively $(1 \leqslant j \leqslant n)$. Each ordered unlabeled topology t of $n \geqslant 2$ leaves results from the following recursive construction: (a) take two ordered unlabeled topologies t_L and t_R of sizes j and $n-j$, respectively, and (b) append t_L and t_R to the left and right, respectively, of a common root node. For example, the leftmost ordered unlabeled topology of Fig. 3A is obtained by appending $t_L = ((\bullet,\bullet),\bullet)$ to the left and $t_R = (\bullet,\bullet)$ to the right of the shared root.

For the third ordered unlabeled topology of Fig. 3A, we take instead $t_L = (\bullet, (\bullet, \bullet)))$ and $t_R = (\bullet, \bullet)$. Because C_{j-1} possible choices exist for t_L and C_{n-1-j} exist for t_R, the probability that a uniformly distributed ordered unlabeled topology t of size n has a left root subtree t_L with size j and a right root subtree t_R with size $n-j$ is

$$\mathbb{P}[|t_L| = j \,\&\, |t_R| = n-j] = \frac{C_{j-1}\, C_{n-1-j}}{C_{n-1}}. \tag{18}$$

3.2.2 Ordered unlabeled histories

An ordered unlabeled *history* of size n is a plane embedding of an unlabeled topology of n leaves whose internal nodes are bijectively labeled by the integers from the interval $[1, n-1]$ in such a way that each non-root internal node has a larger label than its parent node (Fig. 3B). From a biological standpoint, the labels at the internal nodes define a temporal ordering of the coalescent events in the history.

In the language of computer science, ordered unlabeled histories—with leaves and their incident edges stripped away—correspond to the so-called *increasing binary trees* (Example II.17 of [18]), with the term "increasing" referring to the labels of the nodes that increase along any path from the root to the leaves of the tree. To specify the linear ordering of the internal nodes, we write the Newick format of an ordered unlabeled history by adding as a subscript next to a closed parenthesis the label of the corresponding internal node. For instance, $(((\bullet, \bullet)_4, \bullet)_2, (\bullet, \bullet)_3)_1$ indicates the first ordered unlabeled history depicted in Fig. 3B.

The *shape* of an ordered unlabeled history is the underlying unordered unlabeled topology obtained by removing labels at internal nodes and ignoring left–right orientation. With the same notation used in eqs. (1) and (15), the number $\mathrm{ouh}(t)$ of $\underline{\text{o}}$rdered $\underline{\text{u}}$nlabeled $\underline{\text{h}}$istories with tree shape t is calculated recursively as

$$\mathrm{ouh}(t) = 2\, \mathrm{ouh}(t_L)\, \mathrm{ouh}(t_R) \binom{|t|-2}{|t_L|-1} \frac{1}{1+\delta_{t_L=t_R}}, \tag{19}$$

where $\mathrm{ouh}(t) = 1$ if $|t| = 1$. Each ordered unlabeled history with shape t is constructed by appending two ordered unlabeled histories h_1 and h_2 with shapes t_L and t_R, respectively, to the left and right of a common root node, while choosing one of $\binom{|t|-2}{|t_L|-1}$ possibilities for merging the linear ordering of the internal nodes of h_1 with that of the internal nodes of h_2. The factor $1/(1+\delta_{t_L=t_R})$ accounts for possible symmetries in this process.

The set of all possible ordered unlabeled histories of size n is enumerated by F_{n-1} ([31], p. 47), where

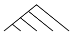
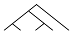
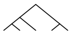
$$F_n = n!. \tag{20}$$

A formula analogous to eq. (18) can be found for ordered unlabeled histories by extending the recursive construction of ordered unlabeled topologies. To construct an ordered unlabeled history t of size n, we do the following: (a) take two ordered unlabeled histories t_L and t_R of sizes j and $n-j$, respectively, (b) append t_L and t_R to the left and to the right, respectively, of a shared root node, and (c) merge the linear ordering of the internal nodes of t_L with the linear ordering of the internal nodes of t_R to define a linear ordering of the internal nodes of t.

For example, the leftmost ordered unlabeled history of Fig. 3B is obtained by appending $t_L = ((\bullet, \bullet)_2, \bullet)_1$ to the left and $t_R = (\bullet, \bullet)_1$ to the right of the same root, and then merging the orderings of the internal nodes of t_L and t_R by putting the root node of t_R between the internal nodes of t_L. Because there are F_{j-1} choices for t_L, F_{n-1-j} choices for t_R, and $\binom{n-2}{j-1}$ ways to merge the ordering of the $j-1$ internal nodes of t_L with the ordering of the $n-1-j$ internal nodes of t_R, the probability that a uniformly distributed ordered unlabeled history t of size n has its left root subtree t_L of size j and its right root subtree t_R of size $n-j$ is

$$\mathbb{P}[|t_L| = j \,\&\, |t_R| = n-j] = \frac{F_{j-1}\, F_{n-1-j} \binom{n-2}{j-1}}{F_{n-1}} = \frac{1}{n-1}. \tag{21}$$

Table 2: Induced probabilities of unlabeled topologies of size 5.

Unlabeled topology	Probability induced by uniform ordered unlabeled topologies	Probability induced by uniform labeled topologies	Probability induced by uniform ordered unlabeled histories	Probability induced by Yule labeled topologies
t	$\frac{\text{out}(t)}{C_4}$	$\frac{\text{lab}(t)}{\lvert L_5\rvert}$	$\frac{\text{ouh}(t)}{F_4}$	$\text{lab}(t) \times \mathbb{P}_Y[t]$
	$\frac{8}{14} = \frac{4}{7}$	$\frac{60}{105} = \frac{4}{7}$	$\frac{8}{24} = \frac{1}{3}$	$60 \times \frac{1}{180} = \frac{1}{3}$
	$\frac{2}{14} = \frac{1}{7}$	$\frac{15}{105} = \frac{1}{7}$	$\frac{4}{24} = \frac{1}{6}$	$15 \times \frac{1}{90} = \frac{1}{6}$
	$\frac{4}{14} = \frac{2}{7}$	$\frac{30}{105} = \frac{2}{7}$	$\frac{12}{24} = \frac{1}{2}$	$30 \times \frac{1}{60} = \frac{1}{2}$

For each unlabeled topology t of size 5, the probability of t induced by the uniform distribution over ordered unlabeled topologies is the ratio of out(t) to the total number $C_4 = 14$ of ordered unlabeled topologies. Similar calculations appear for uniform labeled topologies ($|L_5| = 105$), ordered unlabeled histories ($F_4 = 24$), and Yule labeled topologies. Quantities out(t), lab(t), and ouh(t) are recursively computed from eqs. (15), (1), and (19), respectively. The probability of t induced by the Yule distribution over labeled topologies is obtained by multiplying lab(t) by the Yule probability in eq. (3) of each labeled topology with t as underlying unlabeled topology. The second and third columns agree, as do the fourth and fifth.

3.2.3 Equivalent models for ancestral configurations

We previously noticed [7] that ordered unlabeled topologies and ordered unlabeled histories can be used to study the number of ancestral configurations of uniformly and Yule-distributed labeled topologies, respectively. We observed that the number of ancestral configurations of a given tree structure depends only on the underlying unlabeled topology. Second, as shown in the proofs of Lemmas 1, 2 and 3 of [7], the uniform distribution over the set of ordered unlabeled topologies of size n and the uniform distribution over the set of labeled topologies of size n induce the same distribution over the set of underlying unlabeled topologies of size n; the uniform distribution over ordered unlabeled histories of size n and the Yule distribution over labeled topologies of size n induce the same distribution over the unlabeled topologies of size n. In other words, for each unlabeled topology t, the sum of the probabilities of the uniformly distributed ordered unlabeled topologies (resp. histories) having the shape of t equals the sum of the probabilities of the uniformly (resp. Yule) distributed labeled topologies with tree shape t. These two facts yield the next lemma. Table 2 shows the case $n = 5$.

Lemma 1 *The distribution of the number of ancestral configurations over uniformly (resp. Yule) distributed labeled topologies of size n is the distribution of the number of ancestral configurations over uniformly distributed ordered unlabeled topologies (resp. histories) of size n.*

From this lemma, probabilistic properties of the number of ancestral configurations of uniformly and Yule-distributed labeled topologies can be equivalently studied over uniformly distributed ordered unlabeled topologies and uniformly distributed ordered unlabeled histories, respectively.

To use these equivalences, we require the following lemma.

Lemma 2 *Let R_n and T_n be the random variables that represent the number of root ancestral configurations and the total number of ancestral configurations in a random ordered unlabeled topology (resp. history) of size n selected under the uniform distribution. Equivalently, by Lemma 1, R_n and T_n represent the numbers of root ancestral configurations and the total number of ancestral configurations in a random labeled topology of size n selected under the uniform (resp. Yule) distribution. Then we have $R_1 = T_1 = 0$, and for $n \geqslant 2$,*

$$T_n \overset{d}{=} T_{I_n} + T^*_{n-I_n} + R_n, \tag{22}$$

$$R_n \overset{d}{=} R_{I_n} R^*_{n-I_n} + R_{I_n} + R^*_{n-I_n} + 1, \tag{23}$$

where I_n is distributed over the interval $[1, n-1]$ with probability $\mathbb{P}[I_n = j] = C_{j-1}\, C_{n-1-j}/C_{n-1}$ (resp. $\mathbb{P}[I_n = j] = \frac{1}{n-1}$), R_j^* and T_j^* are independent copies of R_j and T_j, respectively, for each $j \in [1, n-1]$, and both R_j and R_j^* as well as T_j and T_j^* are independent of I_j for $j \in [1, n-1]$.

Proof. The distributional recurrences follow directly from eqs. (12) and (13). $\mathbb{P}[I_n = j]$ follows eqs. (18) and (21), giving the probability that the left root subtree of an ordered unlabeled topology or history of n taxa selected uniformly at random has size $I_n = j$. □

3.3 Known results on the distribution of ancestral configurations

For the random variables R_n and T_n, the asymptotic behavior of the moments $\mathbb{E}[R_n]$, $\mathbb{E}[R_n^2]$, and $\mathbb{E}[T_n]$ and variances $\mathbb{V}[R_n]$ and $\mathbb{V}[T_n]$ were studied under the uniform model of labeled topologies by [11] (Propositions 5 and 6), and under the Yule model by [7] (Propositions 5.4 and 5.5):

$$\mathbb{E}[T_n] \bowtie \mathbb{E}[R_n] \sim \begin{cases} \sqrt{\frac{3}{2}} \left(\frac{4}{3}\right)^n, & \text{Uniform model,} \\ \left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n, & \text{Yule model,} \end{cases} \tag{24}$$

$$\mathbb{V}[T_n] \bowtie \mathbb{V}[R_n] \sim \mathbb{E}[R_n^2] \sim \begin{cases} \sqrt{\frac{7(11-\sqrt{2})}{34}} \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n, & \text{Uniform model,} \\ (2.0449954971\ldots)^n, & \text{Yule model.} \end{cases} \tag{25}$$

The exponential order $2.0449954971\ldots$ of $\mathbb{V}[T_n]$ under the Yule model was approximated by a numerical procedure described in the Appendix of [7].

Under both the uniform and Yule models, the logarithm of the number of root configurations of a randomly selected labeled topology of size n was shown to asymptotically follow a normal distribution. Propositions 4.1 and 5.2 of [7] state that the rescaled random variable

$$\frac{\log R_n - \mathbb{E}[\log R_n]}{\sqrt{\mathbb{V}[\log R_n]}}$$

converges to a standard normal distribution, where $\mathbb{E}[\log R_n] \sim \mu n$, $\mathbb{V}[\log R_n] \sim \sigma^2 n$, and

$$(\mu, \sigma^2) \approx \begin{cases} (0.272, 0.034), & \text{Uniform model,} \\ (0.351, 0.008), & \text{Yule model.} \end{cases} \tag{26}$$

4 Distributional properties of the total number of ancestral configurations

Previous results on ancestral configurations focused on the number of root configurations of labeled topologies selected under the uniform and Yule distributions, while providing only partial analysis of the asymptotic growth of the total number of configurations. We now study in more complete detail the random variable *total number of ancestral configurations* under the same two probability models. In particular, we determine the asymptotic growth of its mean and variance. In agreement with eqs. (24) and (25), we find that the mean and variance of the total number of configurations differ from the mean and variance of the number of root configurations only in their subexponential terms, which turn out to be constants. Moreover, we find that, as is true of the number of root configurations, the total number of configurations follows an asymptotically lognormal distribution.

4.1 Uniform ordered unlabeled topologies and uniform labeled topologies

By Lemma 1, the distribution of the number of ancestral configurations over random labeled topologies of size n selected uniformly at random is the distribution of the number of ancestral configurations over uniformly distributed ordered unlabeled topologies of size n. We use this equivalence to derive the results of this section, denoting by R_n and T_n, respectively, the number of root ancestral configurations and the total number of ancestral configurations in a random ordered unlabeled topology of size n selected under the uniform distribution.

Our first proposition uses the system of distributional recurrences of Lemma 2 to determine the asymptotic behavior of the mean of T_n.

Proposition 3 *The mean total number of ancestral configurations in an ordered unlabeled topology of size n selected uniformly at random satisfies the asymptotic relation $\mathbb{E}[T_n] \sim 2\mathbb{E}[R_n] \sim \sqrt{6}(4/3)^n$.*

Proof. By eq. (23) in Lemma 2 coupled with $\mathbb{E}[R_{I_n}\, R^*_{n-I_n}] = \sum_{j=1}^{n-1} \mathbb{P}[I_n = j]\, \mathbb{E}[R_j\, R^*_{n-j}] = \sum_{j=1}^{n-1} \mathbb{P}[I_n = j]\, \mathbb{E}[R_j]\, \mathbb{E}[R^*_{n-j}]$, we find that for $n \geqslant 1$, the expectation of R_n satisfies

$$C_{n-1}\mathbb{E}[R_n] = \sum_{j=1}^{n-1} C_{j-1}\, C_{n-1-j} \left(\mathbb{E}[R_j]\, \mathbb{E}[R_{n-j}] + \mathbb{E}[R_j] + \mathbb{E}[R_{n-j}] + 1\right), \tag{27}$$

which holds also for $n = 1$ as $\mathbb{E}[R_1] = 0$. Similarly, with $\mathbb{E}[T_1] = 0$, for $n \geqslant 1$, eq. (22) in Lemma 2 gives

$$C_{n-1}\mathbb{E}[T_n] = 2\left(\sum_{j=1}^{n-1} C_{j-1}C_{n-1-j}\mathbb{E}[T_j]\right) + C_{n-1}\mathbb{E}[R_n]. \tag{28}$$

Define the generating functions

$$R(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n \tag{29}$$

$$T(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n]z^n, \tag{30}$$

whose coefficients $C_{n-1}\mathbb{E}[R_n] = C_{n-1}\sum_{i=0}^{\infty} |\{t : c_r(t) = i\}| i/C_{n-1}$ and $C_{n-1}\mathbb{E}[T_n] = C_{n-1}\sum_{i=0}^{\infty} |\{t : c(t) = i\}| i/C_{n-1}$ give respectively the sum of the number of root configurations and the sum of the total number of configurations over all ordered unlabeled topologies t of n taxa.

The recurrences in eqs. (27) and (28) translate into a system of equations for $R(z)$ and $T(z)$:

$$S_1 \equiv \begin{cases} R(z) = R(z)^2 + 2zC(z)\, R(z) + z^2C(z)^2 \\ T(z) = 2zC(z)\, T(z) + R(z), \end{cases}$$

where $C(z)$ is the Catalan generating function (eq. 17). Indeed, multiplying eq. (27) by z^n, we have

$$\begin{aligned} C_{n-1}\mathbb{E}[R_n]z^n \;=\; & \sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}z^{n-1-j} \\ & + z\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z^2\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}z^{n-1-j}. \end{aligned}$$

The first equation of S_1 is obtained by summing over $n \geqslant 1$ and simplifying:

$$\begin{aligned}
R(z) &= \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n = \sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}z^{n-1-j} \\
&\quad + z\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z^2\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}z^{n-1-j} \\
&= R(z)^2 + zR(z)\,C(z) + zC(z)\,R(z) + z^2C(z)^2.
\end{aligned}$$

Similarly, from eq. (28), we obtain the second equation of S_1:

$$\begin{aligned}
T(z) = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n]z^n &= 2z\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[T_j]z^j \cdot C_{n-1-j}z^{n-1-j} + \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n \\
&= 2zT(z)\,C(z) + R(z).
\end{aligned}$$

Solving system S_1 for $T(z)$ yields

$$T(z) = \frac{R(z)}{1-2zC(z)} = \frac{\sqrt{1-4z} - \sqrt{2\sqrt{1-4z}-1}}{2\sqrt{1-4z}}, \tag{31}$$

which has dominant singularity $\alpha \equiv \frac{3}{16}$ at the root of $2\sqrt{1-4z}-1$, with $\frac{3}{16}$ being smaller than the root $\frac{1}{4}$ of $1-4z$. The singular expansion is

$$T(z) \overset{z\to\alpha}{\sim} k_1 - \sqrt{\frac{3}{2}} \cdot \sqrt{1-\frac{16z}{3}},$$

for a certain constant k_1. Eq. (6) thus yields

$$[z^n]T(z) \sim \sqrt{\frac{3}{2}}\frac{(16/3)^n}{2\sqrt{\pi n^3}}.$$

By using the fact that $C_{n-1} \sim 4^{n-1}/\sqrt{\pi n^3}$ (eq. 16), we obtain

$$\mathbb{E}[T_n] = \frac{[z^n]T(z)}{C_{n-1}} \sim \frac{\sqrt{\frac{3}{2}}\frac{(16/3)^n}{2\sqrt{\pi n^3}}}{\frac{4^{n-1}}{\sqrt{\pi n^3}}} = \sqrt{6}\left(\frac{4}{3}\right)^n,$$

which is twice the asymptotic value of $\mathbb{E}[R_n]$ given for the uniform case in eq. (24). □

Fig. 4 plots the exact ratio $\mathbb{E}[T_n]/\mathbb{E}[R_n]$ with increasing n. In agreement with Proposition 3, the ratio $\mathbb{E}[T_n]/\mathbb{E}[R_n]$ approaches 2 as n increases.

We now consider the variance $\mathbb{V}[T_n]$ of the total number of ancestral configurations and its correlation coefficient $\rho[T_n, R_n]$ with the number of root configurations in uniformly distributed ordered unlabeled topologies of fixed size n. The next lemma provides a series of distributional recurrences.

Lemma 4 *Consider the random variables $\tilde{R}_n \equiv R_n + 1$ and T_n. We have $\tilde{R}_1 = 1$, $T_1 = 0$, and for $n \geqslant 2$,*

$$\begin{aligned}
\tilde{R}_n &\overset{d}{=} \tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1, && (32)\\
(\tilde{R}_n)^2 &\overset{d}{=} (\tilde{R}_{I_n})^2(\tilde{R}^*_{n-I_n})^2 + 2\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1, && (33)\\
T_n\tilde{R}_n &\overset{d}{=} T_{I_n}\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + T^*_{n-I_n}\tilde{R}^*_{n-I_n}\tilde{R}_{I_n} + T_{I_n} + T^*_{n-I_n} + (\tilde{R}_n)^2 - \tilde{R}_n, && (34)\\
(T_n)^2 &\overset{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2T_nR_n - (R_n)^2, && (35)
\end{aligned}$$

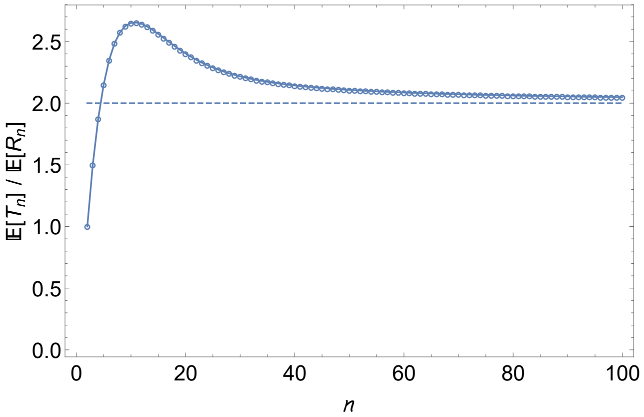


Figure 4: Ratio of the mean total number $\mathbb{E}[T_n]$ of ancestral configurations and mean number $\mathbb{E}[R_n]$ of root configurations for uniformly distributed ordered unlabeled topologies (or uniformly distributed labeled topologies) of size $2 \leqslant n \leqslant 100$. Values of $\mathbb{E}[R_n]$ and $\mathbb{E}[T_n]$ are computed from the recurrences in eqs. (27) and (28), respectively.

where I_n is distributed over the interval $[1, n-1]$ with probability $\mathbb{P}[I_n = j] = C_{j-1}C_{n-1-j}/C_{n-1}$, R_j^*, $\tilde{R}_j^*$, and T_j^* are independent copies of R_j, $\tilde{R}_j$ and T_j, respectively, for every $j \in [1, n-1]$, and both R_j and R_j^* as well as $\tilde{R}_j$, $\tilde{R}_j^*$, T_j, and T_j^* are independent of I_j for $j \in [1, n-1]$.

Proof. Eq. (32) follows directly from eq. (23) in Lemma 2. Eq. (33) is obtained by squaring eq. (32). For eq. (34), eq. (22) in Lemma 2 with eq. (32) yields

$$T_n\tilde{R}_n = (T_{I_n} + T^*_{n-I_n} + R_n)\tilde{R}_n = (T_{I_n} + T^*_{n-I_n})\tilde{R}_n + R_n\tilde{R}_n = (T_{I_n} + T^*_{n-I_n})(\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1) + (\tilde{R}_n - 1)\tilde{R}_n.$$

Finally, by squaring eq. (22) in Lemma 2, we obtain

$$\begin{aligned}(T_n)^2 &\overset{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + (R_n)^2 + 2T_{I_n}T^*_{n-I_n} + 2T_{I_n}R_n + 2T^*_{n-I_n}R_n \\ &\overset{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2(R_n)^2 + 2T_{I_n}R_n + 2T^*_{n-I_n}R_n - (R_n)^2 \\ &\overset{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2R_n(R_n + T_{I_n} + T^*_{n-I_n}) - (R_n)^2,\end{aligned}$$

which gives eq. (35), because $R_n + T_{I_n} + T^*_{n-I_n} = T_n$ again by eq. (22). □

To proceed with the asymptotic analysis of the variance $\mathbb{V}[T_n]$ and correlation $\rho[T_n, R_n]$, we now determine the asymptotic behavior of expectations $\mathbb{E}[T_n^2]$ and $\mathbb{E}[T_nR_n]$. We define the following generating functions:

$$\begin{aligned}
\tilde{R}(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n + \sum_{n=1}^{\infty} C_{n-1}z^n = R(z) + zC(z), && (36)\\
\tilde{S}(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n, && (37)\\
S(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n^2]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2 - 2\tilde{R}_n + 1]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2 - 2(R_n+1) + 1]z^n \\
&= \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n - 2\sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n - \sum_{n=1}^{\infty} C_{n-1}z^n = \tilde{S}(z) - 2R(z) - zC(z), && (38)\\
\tilde{V}(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n\tilde{R}_n]z^n, && (39)\\
V(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_nR_n]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n(\tilde{R}_n - 1)]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n\tilde{R}_n]z^n - \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n]z^n \\
&= \tilde{V}(z) - T(z), && (40)\\
U(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n^2]z^n. && (41)
\end{aligned}$$

Here, $R(z)$ is from eq. (29), $C(z)$ is given in eq. (17), and $T(z)$ is given in eq. (31).

The distributional recurrences of Lemma 4 determine recurrences for the expectations $\mathbb{E}[\tilde{R}_n^2]$, $\mathbb{E}[T_n\tilde{R}_n]$, and $\mathbb{E}[T_n^2]$, which then translate into a system of functional equations:

$$S_2 \equiv \begin{cases} \tilde{S}(z) - z = \tilde{S}(z)^2 + 2\tilde{R}(z)^2 + z^2C(z)^2 \\ \tilde{V}(z) = 2\tilde{V}(z)\,\tilde{R}(z) + 2zC(z)\,T(z) + \tilde{S}(z) - \tilde{R}(z) \\ U(z) = 2zC(z)\,U(z) + 2T(z)^2 + 2V(z) - S(z). \end{cases}$$

Focusing on the first equation in S_2, we observe that eq. (33) gives for $n \geqslant 2$

$$C_{n-1}\mathbb{E}[\tilde{R}_n^2] = \sum_{j=1}^{n-1} C_{j-1}C_{n-1-j}\Big(\mathbb{E}[\tilde{R}_j^2]\,\mathbb{E}[\tilde{R}_{n-j}^2] + 2\mathbb{E}[\tilde{R}_j]\,\mathbb{E}[\tilde{R}_{n-j}] + 1\Big),$$

which multiplied by z^n can be rewritten as

$$\begin{aligned}
C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n &= \sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j^2]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}^2]z^{n-j} + 2\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}]z^{n-j} \\
&\quad + z^2\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}z^{n-1-j}.
\end{aligned}$$

Summing over $n \geqslant 2$, we obtain

$$\begin{aligned}
\tilde{S}(z) &= z + \sum_{n=2}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n = z + \sum_{n=2}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j^2]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}^2]z^{n-j} \\
&\quad + 2\sum_{n=2}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}]z^{n-j} + z^2\sum_{n=2}^{\infty}\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}z^{n-1-j} \\
&= z + \tilde{S}(z)^2 + 2\tilde{R}(z)^2 + z^2C(z)^2.
\end{aligned}$$

Similarly, the second equation of S_2 follows from eq. (34), and the third equation from eq. (35).

To solve system S_2, we first find $R(z)$ from the first equation of system S_1 from the proof of Proposition 3. We then find $\tilde{R}(z)$ by using eq. (36). From the first equation of system S_2, we can obtain $\tilde{S}(z)$ and also $S(z)$ from eq. (38). $\tilde{S}(z)$ is then used together with $T(z)$ (eq. 31) for calculating $\tilde{V}(z)$ from the second equation of S_2. Once we have a formula for $\tilde{V}(z)$, we obtain $V(z)$ from eq. (40), and we finally compute $U(z)$ from the third equation of S_2. Writing $r \equiv \sqrt{1-4z}$, we find

$$V(z) = \frac{-\sqrt{2r-1} + r\left(-r + \sqrt{2r-1} - \sqrt{-2r+4\sqrt{2r-1}-1} + 3\right) - 1}{2r\sqrt{2r-1}}, \tag{42}$$

$$U(z) = \frac{1}{2}\left(-\frac{1}{r^3} + \frac{-\frac{2\sqrt{-2r+4\sqrt{2r-1}-1}}{\sqrt{2r-1}} + \sqrt{-2r+4\sqrt{2r-1}-1} + \frac{4}{\sqrt{2r-1}} + 3}{r} - \frac{6}{\sqrt{2r-1}} + 1\right). \tag{43}$$

The dominant singularity of the generating functions $V(z)$ and $U(z)$ is at $\alpha \equiv 7(8\sqrt{2}-11)/16$, which is the dominant singularity of the square root $\sqrt{-2r+4\sqrt{2r-1}-1}$ appearing in eqs. (42) and (43). We obtain the expansion of $V(z)$ and $U(z)$ at their dominant singularity α by plugging the expansion

$$\sqrt{-2r+4\sqrt{2r-1}-1} \overset{z\to\alpha}{\sim} \sqrt{\frac{7(11-\sqrt{2})}{34}} \cdot \sqrt{1 - \frac{16z}{7(8\sqrt{2}-11)}}$$

in eqs. (42) and (43), while setting $z = \alpha$ elsewhere. Algebraic manipulations then lead to

$$V(z) \overset{z\to\alpha}{\sim} k_1 - \frac{1}{2}\left(1 + \frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}} \cdot \sqrt{1 - \frac{16z}{7(8\sqrt{2}-11)}},$$

$$U(z) \overset{z\to\alpha}{\sim} k_1' - \frac{1}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}} \cdot \sqrt{1 - \frac{16z}{7(8\sqrt{2}-11)}},$$

for certain constants k_1 and k_1'. Eq. (6), together with the asymptotic expansion in eq. (16), finally yields

$$\mathbb{E}[T_n R_n] = \frac{[z^n]V(z)}{C_{n-1}} \sim \left(1 + \frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}} \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n, \tag{44}$$

$$\mathbb{E}[T_n^2] = \frac{[z^n]U(z)}{C_{n-1}} \sim \frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}} \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n. \tag{45}$$

By using these calculations, we obtain the following result.

Proposition 5 *The variance of the total number T_n of ancestral configurations in an ordered unlabeled topology of size n selected uniformly at random satisfies the asymptotic relation $\mathbb{V}[T_n] \sim \mathbb{E}[T_n^2]$, where $\mathbb{E}[T_n^2]$ grows as in eq. (45). For increasing values of n, the correlation coefficient $\rho[T_n, R_n]$ between the total number T_n of ancestral configurations and the number R_n of root configurations converges to a constant $\rho[T_n, R_n] \to 0.9004\ldots$.*

Proof. First, for the variance we have $\mathbb{V}[T_n] = \mathbb{E}[T_n^2] - \mathbb{E}[T_n]^2 \sim \mathbb{E}[T_n^2]$. Indeed, from Proposition 3, $\mathbb{E}[T_n]^2 \bowtie [(4/3)^2]^n = (16/9)^n$, whereas from eq. (45), $\mathbb{E}[T_n^2] \bowtie \left[4/[7(8\sqrt{2}-11)]\right]^n$, and $4/[7(8\sqrt{2}-11)] > \frac{16}{9}$.

Second, the covariance $\mathrm{Cov}[T_n, R_n]$ grows like $\mathrm{Cov}[T_n, R_n] = \mathbb{E}[T_n R_n] - \mathbb{E}[T_n]\,\mathbb{E}[R_n] \sim \mathbb{E}[T_n R_n]$. Indeed, by Proposition 3 and eq. (24) for the uniform model, we have $\mathbb{E}[T_n]\,\mathbb{E}[R_n] \bowtie (4/3)^n \cdot (4/3)^n = (16/9)^n$; from eq. (44), we have $\mathbb{E}[T_n R_n] \bowtie \left[4/[7(8\sqrt{2}-11)]\right]^n$, with $4/[7(8\sqrt{2}-11)] > \frac{16}{9}$. Hence, for the correlation coefficient,

$$\rho[T_n, R_n] = \frac{\mathrm{Cov}[T_n, R_n]}{\sqrt{\mathbb{V}[T_n]}\sqrt{\mathbb{V}[R_n]}} \sim \frac{1+\dfrac{\sqrt{2}}{2}}{\sqrt{\dfrac{2}{17}(15+11\sqrt{2})}} \approx 0.9004,$$

where we used the asymptotic formula for $\mathbb{V}[R_n]$ given for the uniform model in eq. (25) as well as the asymptotics in eqs. (44) and (45) for $\mathrm{Cov}[T_n, R_n]$ and $\mathbb{V}[T_n]$, respectively. □

To conclude this section, we show that the total number T_n of ancestral configurations of an ordered unlabeled topology of size n selected uniformly at random has an asymptotic lognormal distribution. From Lemma 1 and Section 3.3, we know that the logarithm of the number R_n of root configurations in an ordered unlabeled topology of size n selected uniformly at random converges asymptotically to a standard normal distribution, that is,

$$\frac{\log R_n - \mathbb{E}[\log R_n]}{\sqrt{\mathbb{V}[\log R_n]}} \xrightarrow{d} N(0,1), \tag{46}$$

with $\mathbb{E}[\log R_n] \sim 0.272 \cdot n$ and $\mathbb{V}[\log R_n] \sim 0.034 \cdot n$ (where the constants are approximate).

From eq. (14), the variables R_n and T_n measured over the same random ordered unlabeled topology of size n satisfy $R_n \leqslant T_n \leqslant (2n-1)R_n$. This inequality gives $\log T_n = \log R_n + \epsilon_n$, where the random variable ϵ_n has values in $[0, \log(2n-1)]$. Thus, we have

$$\mathbb{E}[\log T_n] \;=\; \mathbb{E}[\log R_n] + \mathbb{E}[\epsilon_n] = \mathbb{E}[\log R_n] + \mathcal{O}(\log n) \sim \mathbb{E}[\log R_n] \tag{47}$$

$$\mathbb{V}[\log T_n] \;=\; \mathbb{V}[\log R_n] + \mathbb{V}[\epsilon_n] + 2\mathrm{Cov}[\log R_n, \epsilon_n] \sim \mathbb{V}[\log R_n], \tag{48}$$

where we use $\mathbb{V}[\epsilon_n] \leqslant [\log(2n-1)]^2/4$ from Popoviciu's inequality on the maximal variance for a bounded random variable [3], so that the comparison with the linearly increasing $\mathbb{V}[\log R_n]$, gives $\lim_{n\to\infty} \mathbb{V}[\epsilon_n]/\mathbb{V}[\log R_n] = 0$; we also use $\mathrm{Cov}[\log R_n, \epsilon_n] \leqslant \sqrt{\mathbb{V}[\log R_n]}\,\sqrt{\mathbb{V}[\epsilon_n]}$ from the Cauchy-Schwarz inequality.

Next, we write

$$\frac{\log T_n - \mathbb{E}[\log T_n]}{\sqrt{\mathbb{V}[\log T_n]}} \;=\; \frac{\log R_n - \mathbb{E}[\log R_n]}{\sqrt{\mathbb{V}[\log R_n]}} \cdot \frac{\sqrt{\mathbb{V}[\log R_n]}}{\sqrt{\mathbb{V}[\log T_n]}} + \frac{\epsilon_n - \mathbb{E}[\epsilon_n]}{\sqrt{\mathbb{V}[\log T_n]}}. \tag{49}$$

The expression $(\log R_n - \mathbb{E}[\log R_n])/\sqrt{\mathbb{V}[\log R_n]}$ converges in distribution to a normal random variable with mean 0 and variance 1 (eq. 46). The ratio $\sqrt{\mathbb{V}[\log R_n]}/\sqrt{\mathbb{V}[\log T_n]}$ is a number sequence that by eq. (48) converges to a finite constant, 1. The expression $(\epsilon_n - \mathbb{E}[\epsilon_n]/\sqrt{\mathbb{V}[\log T_n]})$ converges in mean square to 0, as

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{\epsilon_n - \mathbb{E}[\epsilon_n]}{\sqrt{\mathbb{V}[\log T_n]}} - 0\right)^2\right] = \lim_{n\to\infty} \frac{\mathbb{V}[\epsilon_n]}{\mathbb{V}[\log T_n]};$$

the denominator $\mathbb{V}[\log T_n]$ increases linearly with n (Section 3.3), and again by Popoviciu's inequality, the numerator is bounded above by $[\log(2n-1)]^2/4$, so that $\lim_{n\to\infty} \mathbb{V}[\epsilon_n]/\mathbb{V}[\log T_n] = 0$.

As convergence in mean square implies convergence in probability [29, p. 10], we can apply Slutsky's theorem on perturbation of random variables that converge in distribution by random variables that converge in probability [29, p. 19] to eq. (49). In particular, the convergence in distribution of $(\log R_n - \mathbb{E}[\log R_n])/\sqrt{\mathbb{V}[\log R_n]}$, trivial convergence in probability of $\sqrt{\mathbb{V}[\log R_n]}/\sqrt{\mathbb{V}[\log T_n]}$, and convergence in probability of $(\epsilon_n - \mathbb{E}[\epsilon_n])/\sqrt{\mathbb{V}[\log T_n]}$ allow us to conclude $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathbb{V}[\log T_n]}$ converges in distribution to a normal random variable with mean 0 and variance 1.

Fig. 5 shows the cumulative distribution $\mathbb{P}[\log T_n \leqslant \mathbb{E}[\log T_n] + y\sqrt{\mathbb{V}[\log T_n]}]$ as a function of y, when ordered unlabeled topologies of size 15 are selected uniformly at random. To obtain the distribution, we count total

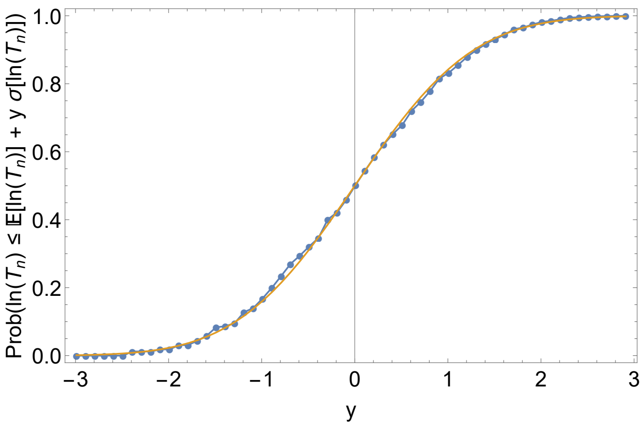


Figure 5: Cumulative distribution of the natural logarithm of the total number of configurations for uniformly distributed ordered unlabeled topologies (or uniformly distributed labeled topologies) of size $n = 15$ (dotted line). For each $y \in [-3, 3]$ in steps of 0.1, the quantity plotted is the probability that an ordered unlabeled topology (or labeled topology) with $n = 15$ chosen uniformly at random has total number of configurations less than or equal to $\exp(\mathbb{E}[\log T_n] + y\sigma[\log T_n])$, where $\mathbb{E}[\log T_n]$ and $\sigma[\log T_n] = \sqrt{\mathbb{V}[\log T_n]}$ are respectively the mean and standard deviation of the logarithm of the total number of configurations for uniformly distributed ordered unlabeled topologies (or labeled topologies) with $n = 15$ taxa. The solid line is the cumulative distribution of a Gaussian random variable with mean 0 and variance 1.

configurations for each unlabeled topology t with 15 leaves, and then count the number of ordered unlabeled topologies having the shape of t (eq. 15). The figure illustrates the agreement between the exact cumulative distribution of ancestral configurations and the standard normal distribution.

By the equivalence between ordered unlabeled topologies and labeled topologies reported in Lemma 1, we can state the main results of this section as follows.

Theorem 6 *For a labeled topology of size n selected at random under the uniform distribution, the mean and the variance of the total number T_n of ancestral configurations grow asymptotically like*

$$\mathbb{E}[T_n] \sim \sqrt{6}\left(\frac{4}{3}\right)^n, \tag{50}$$

$$\mathbb{V}[T_n] \sim \frac{2}{17}(15 + 11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n. \tag{51}$$

Furthermore, the logarithm of the total number of ancestral configurations in a labeled topology of size n selected uniformly at random, rescaled as $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathrm{Var}[\log T_n]}$, converges to a standard normal distribution, where $\mathbb{E}[\log T_n] \sim \mu n$ and $\mathbb{V}[\log T_n] \sim \sigma^2 n$, $(\mu, \sigma^2) \approx (0.272, 0.034)$.

4.2 Uniform ordered unlabeled histories and Yule labeled topologies

By Lemma 1, the distribution of the total number of ancestral configurations over random labeled topologies of size n selected under the Yule probability model is the distribution of the total number of configurations over uniformly distributed ordered unlabeled histories of size n. We exploit this equivalence for this section, now denoting by R_n and T_n, respectively, the number of root ancestral configurations and the total number of ancestral configurations in a random ordered unlabeled history of size n under the uniform distribution.

Lemma 7 *Consider the random variables R_n and T_n. We have $R_1 = T_1 = 0$, and for $n \geqslant 2$,*

$$\begin{aligned}
R_n &\overset{d}{=} R_{I_n}\, R^*_{n-I_n} + R_{I_n} + R^*_{n-I_n} + 1, && (52)\\
T_n &\overset{d}{=} T_{I_n} + T^*_{n-I_n} + R_n, && (53)\\
T_nR_n &\overset{d}{=} T_{I_n}R_{I_n}R^*_{n-I_n} + T_{I_n}R_{I_n} + T_{I_n}R^*_{n-I_n} + T_{I_n} \\
&\quad + T^*_{n-I_n}R_{I_n}R^*_{n-I_n} + T^*_{n-I_n}R_{I_n} + T^*_{n-I_n}R^*_{n-I_n} + T^*_{n-I_n} + (R_n)^2, && (54)\\
(T_n)^2 &\overset{d}{=} (T_{I_n})^2 + (T^*_{I_{n-I_n}})^2 + 2T_{I_n}T^*_{n-I_n} + 2T_nR_n - (R_n)^2, && (55)
\end{aligned}$$

where I_n is a uniformly distributed variable over the interval $[1, n-1]$, R^*_j and T^*_j are independent copies of R_j and T_j, respectively, for every $j \in [1, n-1]$, and both R_j and R^*_j as well as T_j, and T^*_j are independent of I_j for $j \in [1, n-1]$.

Proof. Eqs. (52) and (53) are from Lemma 2. By expanding $T_nR_n \overset{d}{=} (T_{I_n} + T^*_{n-I_n})(R_{I_n} + 1)(R^*_{n-I_n} + 1) + (R_n)^2$, we have eq. (54). Finally, eq. (55) is obtained by squaring eq. (53); it also copies eq. (35) from Lemma 4. □

The distributional recurrences in Lemma 7 can be used to determine recurrences for the expectations $\mathbb{E}[R_n]$, $\mathbb{E}[T_n]$, $\mathbb{E}[T_nR_n]$, and $\mathbb{E}[T_n^2]$. For $n \geqslant 1$, we can write

$$\begin{aligned}
(n-1)\mathbb{E}[R_n] &= \Big(\sum_{j=1}^{n-1} \mathbb{E}[R_j]\,\mathbb{E}[R_{n-j}]\Big) + 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[R_j]\Big) + (n-1), && (56)\\
(n-1)\mathbb{E}[T_n] &= 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\Big) + (n-1)\mathbb{E}[R_n], && (57)\\
(n-1)\mathbb{E}[T_nR_n] &= 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_jR_j]\,\mathbb{E}[R_{n-j}]\Big) + 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_jR_j]\Big) + 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\,\mathbb{E}[R_{n-j}]\Big) \\
&\quad +2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\Big) + (n-1)\mathbb{E}[R_n^2], && (58)\\
(n-1)\mathbb{E}[T_n^2] &= 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_j^2]\Big) + 2\Big(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\,\mathbb{E}[T_{n-j}]\Big) + 2(n-1)\mathbb{E}[T_nR_n] - (n-1)\mathbb{E}[R_n^2]. && (59)
\end{aligned}$$

Define the following generating functions

$$\begin{aligned}
R(z) &\equiv \sum_{n=1}^{\infty} \mathbb{E}[R_n]z^n, && (60)\\
T(z) &\equiv \sum_{n=1}^{\infty} \mathbb{E}[T_n]z^n, && (61)\\
S(z) &\equiv \sum_{n=1}^{\infty} \mathbb{E}[R_n^2]z^n, && (62)\\
V(z) &\equiv \sum_{n=1}^{\infty} \mathbb{E}[T_nR_n]z^n, && (63)\\
U(z) &\equiv \sum_{n=1}^{\infty} \mathbb{E}[T_n^2]z^n. && (64)
\end{aligned}$$

The recurrences in eqs. (57), (58), and (59) translate into a system of differential equations:

$$S_3 \equiv \begin{cases} T'(z) - \frac{z+1}{z-z^2}T(z) = R'(z) - \frac{R(z)}{z} \\ V'(z) - \left(\frac{2R(z)}{z} + \frac{z+1}{z-z^2}\right)V(z) = \frac{2T(z)\,R(z) + \frac{2z}{1-z}T(z) + zS'(z) - S(z)}{z} \\ U'(z) - \frac{z+1}{z-z^2}U(z) = \frac{2T(z)^2 + 2zV'(z) - 2V(z) - zS'(z) + S(z)}{z}, \end{cases}$$

The derivatives $R'(z)$, $T'(z)$, $S'(z)$, $V'(z)$, and $U'(z)$ appear in S_3 due to the factor $n-1$ in eqs. (57), (58), and (59). We derive the third equation in S_3 as an example. First, multiplying both sides of eq. (59) by z^n yields

$$\begin{aligned} zn\mathbb{E}[T_n^2]z^{n-1} - \mathbb{E}[T_n^2]z^n &= 2\sum_{j=1}^{n-1}\mathbb{E}[T_j^2]z^j \cdot z^{n-j} + 2\sum_{j=1}^{n-1}\mathbb{E}[T_j]z^j \cdot \mathbb{E}[T_{n-j}]z^{n-j} + 2zn\mathbb{E}[T_nR_n]z^{n-1} \\ &\quad -2\mathbb{E}[T_nR_n]z^n - zn\mathbb{E}[R_n^2]z^{n-1} + \mathbb{E}[R_n^2]z^n. \end{aligned}$$

Summing over $n \geqslant 1$, we obtain

$$\begin{aligned} z\sum_{n=1}^{\infty} n\mathbb{E}[T_n^2]z^{n-1} - \sum_{n=1}^{\infty}\mathbb{E}[T_n^2]z^n &= 2\sum_{n=1}^{\infty}\sum_{j=1}^{n-1}\mathbb{E}[T_j^2]z^j \cdot z^{n-j} + 2\sum_{n=1}^{\infty}\sum_{j=1}^{n-1}\mathbb{E}[T_j]z^j \cdot \mathbb{E}[T_{n-j}]z^{n-j} \\ &\quad +2z\sum_{n=1}^{\infty} n\mathbb{E}[T_nR_n]z^{n-1} - 2\sum_{n=1}^{\infty}\mathbb{E}[T_nR_n]z^n - z\sum_{n=1}^{\infty} n\mathbb{E}[R_n^2]z^{n-1} + \sum_{n=1}^{\infty}\mathbb{E}[R_n^2]z^n. \end{aligned}$$

To complete the derivation, we note that this equation can be rewritten:

$$zU'(z) - U(z) = 2U(z)\left(\frac{z}{1-z}\right) + 2T(z)^2 + 2zV'(z) - 2V(z) - zS'(z) + S(z),$$

as $U'(z) = \left(\sum_{n=1}^{\infty}\mathbb{E}[T_n^2]z^n\right)' = \sum_{n=1}^{\infty} n\mathbb{E}[T_n^2]z^{n-1}$, $V'(z) = \left(\sum_{n=1}^{\infty}\mathbb{E}[T_nR_n]z^n\right)' = \sum_{n=1}^{\infty} n\mathbb{E}[T_nR_n]z^{n-1}$, $S'(z) = \left(\sum_{n=1}^{\infty}\mathbb{E}[R_n^2]z^n\right)' = \sum_{n=1}^{\infty} n\mathbb{E}[R_n^2]z^{n-1}$, and $\frac{z}{1-z} = \sum_{n=1}^{\infty} z^n$.

We also observe that the generating functions $R(z)$ and $S(z)$ (eqs. 60 and 62) were studied in the analysis of root configurations under the Yule model for labeled topologies in Section 5 of [7]. In particular, eq. (39) in the proof of Proposition 5.3 of [7] found that $R(z)$—there denoted by $E(z)$—has explicit form

$$R(z) = \frac{2z\sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)}{(z-1)\left[\sqrt{3}\cos\left(\frac{\sqrt{3}}{2}\log(1-z)\right) + \sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)\right]}. \tag{65}$$

The dominant singularity is $\alpha_1 \equiv 1 - e^{-2\pi\sqrt{3}/9}$, and the singular expansion at the dominant singularity is

$$R(z) \overset{z\to\alpha_1}{\sim} \frac{1}{1-\frac{z}{\alpha_1}}. \tag{66}$$

The generating function $S(z)$ was found in Section 5.3 of [7] to have singular expansion

$$S(z) \overset{z\to\alpha_2}{\sim} \frac{1}{1-\frac{z}{\alpha_2}}, \tag{67}$$

where the dominant singularity $\alpha_2 \equiv 0.4889986317\ldots$ was approximated in the Appendix. By singularity analysis (eq. 5), the expansions in eqs. (66) and (67) yield the asymptotic relations in eqs. (24) and (25):

$$\mathbb{E}[R_n] \sim \alpha_1^{-n} \quad \text{and} \quad \mathbb{E}[R_n^2] \sim \alpha_2^{-n}. \tag{68}$$

Note indeed, that the asymptotic constant $2.0449954971\ldots$ appearing in eq. (25) is obtained as α_2^{-1}.

We now observe that eq. (66) and the first equation of S_3 yield the asymptotic growth of the mean number $\mathbb{E}[T_n]$ of ancestral configurations in an ordered unlabeled history of size n selected uniformly at random.

Proposition 8 *The mean total number of ancestral configurations in an ordered unlabeled history of size n selected uniformly at random satisfies the asymptotic relation $\mathbb{E}[T_n] \sim \mathbb{E}[R_n] \sim \alpha_1^{-n} = [1/(1-e^{-2\pi\sqrt{3}/9})]^n$.*

Proof. We start by rewriting the first equation of S_3 as

$$T'(z)\,M(z) - \frac{z+1}{z-z^2}M(z)\,T(z) = \left[R'(z) - \frac{R(z)}{z}\right] M(z), \tag{69}$$

where $M(z) \equiv (z-1)^2/z$ is the integrating factor.

Since $M'(z) = -\frac{z+1}{z-z^2}M(z)$, the left-hand side of eq. (69) can be rewritten $[T(z)\,M(z)]'$, yielding

$$\left[T(z)\frac{(z-1)^2}{z}\right]' = \left[R'(z) - \frac{R(z)}{z}\right]\frac{(z-1)^2}{z}. \tag{70}$$

Because $T_1 = 0$, the expansion of $T(z)$ starts with a non-zero quadratic term. Hence, we have

$$\left[T(z)\frac{(z-1)^2}{z}\right]_{z=0} = 0,$$

and the differential equation in eq. (70) thus gives $T(z)\frac{(z-1)^2}{z} = \int_0^z [R'(t) - R(t)/t]\frac{(t-1)^2}{t}\,dt$, that is,

$$T(z) = \frac{z}{(z-1)^2}\int_0^z \left[R'(t) - \frac{R(t)}{t}\right]\frac{(t-1)^2}{t}\,dt.$$

To obtain the singular expansion of $T(z)$, we must analyze functions $R'(t)$, $[R'(t)-R(t)/t]\frac{(t-1)^2}{t}$, and $\int_0^z [R'(t)-R(t)/t]\frac{(t-1)^2}{t}\,dt$ at their dominant singularity. Because $\alpha_1 = 1-e^{-2\pi\sqrt{3}/9}$ is the dominant singularity of $R(t)$ and $R(t) \overset{t\to\alpha_1}{\sim} 1/(1-\frac{t}{\alpha_1})$ (eq. 66), from eq. (7), $R'(t)$ has dominant singularity at $t = \alpha_1$. Its singular expansion is

$$R'(t) \overset{t\to\alpha_1}{\sim} \frac{1}{\alpha_1(1-\frac{t}{\alpha_1})^2},$$

obtained by differentiating the expansion of $R(t)$. It follows that α_1 is also the dominant singularity of the function $[R'(t) - R(t)/t]\frac{(t-1)^2}{t}$, whose singular expansion follows

$$\left[R'(t) - \frac{R(t)}{t}\right]\frac{(t-1)^2}{t} \overset{t\to\alpha_1}{\sim} \left[\frac{1}{\alpha_1(1-\frac{t}{\alpha_1})^2}\right]\frac{(\alpha_1-1)^2}{\alpha_1}. \tag{71}$$

Finally, by eq. (8) $\int_0^z [R'(t) - R(t)/t]\frac{(t-1)^2}{t}\,dt$ can be expanded at its dominant singularity α_1 by integrating the singular expansion of the integrand function (eq. 71). Consequently, the expansion of $T(z)$ at its dominant singularity α_1 satisfies

$$T(z) \overset{z\to\alpha_1}{\sim} \frac{\alpha_1}{(\alpha_1-1)^2}\int_0^z \frac{1}{\alpha_1(1-\frac{t}{\alpha_1})^2}\frac{(\alpha_1-1)^2}{\alpha_1}\,dt = \frac{1}{1-\frac{z}{\alpha_1}} - 1 \overset{z\to\alpha_1}{\sim} \frac{1}{1-\frac{z}{\alpha_1}}.$$

By eqs. (5) and (68), we conclude

$$\mathbb{E}[T_n] = [z^n]T(z) \sim \alpha_1^{-n} \sim \mathbb{E}[R_n]. \quad \square$$

In Fig. 6, we show a numerical plot of the ratio $\mathbb{E}[T_n]/\mathbb{E}[R_n]$ as a function of n. Following the proposition, as n increases, the numerical ratio approaches 1.

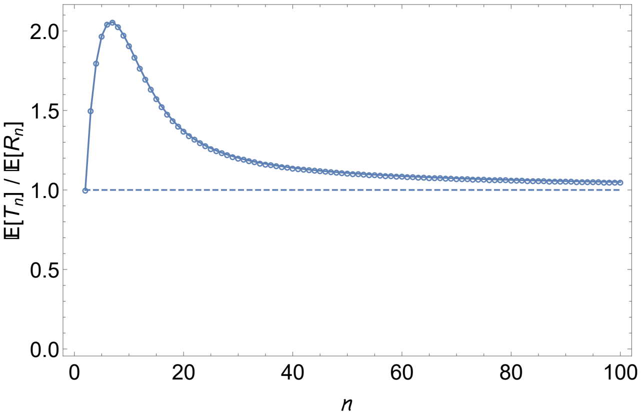


Figure 6: Ratio of the mean total number $\mathbb{E}[T_n]$ of ancestral configurations and mean number $\mathbb{E}[R_n]$ of root configurations for uniformly distributed ordered unlabeled histories (or Yule-distributed labeled topologies) of size $2 \leqslant n \leqslant 100$. Values of $\mathbb{E}[R_n]$ and $\mathbb{E}[T_n]$ are computed from the recurrences in eqs. (56) and (57), respectively.

We now study the variance $\mathbb{V}[T_n]$ of the total number of ancestral configurations and the correlation coefficient $\rho[T_n, R_n]$ between the numbers of total and root configurations in uniformly distributed ordered unlabeled histories of fixed size. By again using properties of singular expansions under differentiation and integration, we determine the asymptotic growth of the expectations $\mathbb{E}[T_n R_n]$ and $\mathbb{E}[T_n^2]$. We start with $\mathbb{E}[T_n R_n]$. We abbreviate a term in the second equation of S_3 by $P(z)$:

$$P(z) \equiv \frac{2T(z)R(z) + \frac{2z}{1-z}T(z) + zS'(z) - S(z)}{z}.$$

The second equation of S_3 becomes $V'(z) - V(z)\,[\frac{2R(z)}{z} + \frac{z+1}{z-z^2}] = P(z)$. We introduce integration factor $M(z)$,

$$M(z) \equiv \left[e^{\int_0^z -\frac{2R(t)}{t}\,dt}\right] \cdot \frac{(z-1)^2}{z},$$

such that $M'(z) = -[\frac{2R(z)}{z} + \frac{z+1}{z-z^2}]M(z)$. We find $[V(z)\,M(z)]' = P(z)\,M(z)$, and thus

$$V(z) = \frac{1}{M(z)} \int_0^z P(t)\,M(t)\,dt.$$

To determine the singular expansion of $V(z)$ at its dominant singularity, we observe that $P(t)\,M(t)$ is a function of $T(t)$, $R(t)$, $S(t)$, and $S'(t)$. As demonstrated in Proposition 8, $T(t)$ and $R(t)$ have the same dominant singularity $\alpha_1 \approx 0.702$ (eq. 66), a value larger than the dominant singularity $\alpha_2 \approx 0.489$ of $S(t)$ and $S'(t)$ (eq. 67). Hence, $R(t)$ and $T(t)$ are analytic functions in a neighborhood of 0, say $|t| \leqslant \frac{1}{2}$, that contains α_2. As a consequence, we can obtain the singular expansion of $P(t)\,M(t)$ at its dominant singularity α_2 by replacing $S(t)$ and $S'(t)$ with their expansions $S(t) \overset{t\to\alpha_2}{\sim} 1/(1-\frac{t}{\alpha_2})$ (eq. 67) and $S'(t) \overset{t\to\alpha_2}{\sim} [1/(1-\frac{t}{\alpha_2})]' = 1/[\alpha_2(1-\frac{t}{\alpha_2})^2]$, while substituting $t = \alpha_2$ elsewhere. We find

$$P(t)\,M(t) \overset{t\to\alpha_2}{\sim} \left[\frac{2T(\alpha_2)\,R(\alpha_2) + \frac{2\alpha_2}{1-\alpha_2}T(\alpha_2)}{\alpha_2} + \frac{1}{\alpha_2(1-\frac{t}{\alpha_2})^2}\right] M(\alpha_2) \overset{t\to\alpha_2}{\sim} \frac{M(\alpha_2)}{\alpha_2(1-\frac{t}{\alpha_2})^2}.$$

The singular expansion under integration (eq. 8) thus gives $\int_0^z P(t)\,M(t)\,dt \overset{z\to\alpha_2}{\sim} \int_0^z M(\alpha_2)/[\alpha_2(1-\frac{t}{\alpha_2})^2]\,dt \overset{z\to\alpha_2}{\sim} M(\alpha_2)/(1-\frac{z}{\alpha_2})$, from which the singular expansion of $V(z)$ at its dominant singularity α_2 is

$$V(z) \overset{z\to\alpha_2}{\sim} \frac{1}{M(\alpha_2)} \frac{M(\alpha_2)}{1-\frac{z}{\alpha_2}} = \frac{1}{1-\frac{z}{\alpha_2}}.$$

Hence, by applying eq. (5) together with eq. (68), we have

$$\mathbb{E}[T_n R_n] = [z^n]V(z) \sim \alpha_2^{-n} \sim \mathbb{E}[R_n^2]. \tag{72}$$

We follow the same approach to determine the asymptotic growth of $\mathbb{E}[T_n^2]$. Multiplying both sides of the third equation of S_3 by the integrating factor $M(z) \equiv (z-1)^2/z$ used in the proof of Proposition 8, we find

$$U(z) = \frac{z}{(z-1)^2} \int_0^z \left[\frac{2T(t)^2}{t} + 2\left(V'(t) - \frac{V(t)}{t}\right) - \left(S'(t) - \frac{S(t)}{t}\right)\right] \frac{(t-1)^2}{t}\, dt.$$

We abbreviate

$$G(t) \equiv \frac{2T(t)^2}{t} + 2\left(V'(t) - \frac{V(t)}{t}\right) - \left(S'(t) - \frac{S(t)}{t}\right).$$

The singular expansions $V(t) \overset{t \to \alpha_2}{\sim} 1/(1-\frac{t}{\alpha_2})$, $V'(t) \overset{t \to \alpha_2}{\sim} 1/[\alpha_2(1-\frac{t}{\alpha_2})^2]$, $S(t) \overset{t \to \alpha_2}{\sim} 1/(1-\frac{t}{\alpha_2})$, and $S'(t) \overset{t \to \alpha_2}{\sim} 1/[\alpha_2(1-\frac{t}{\alpha_2})^2]$ yield the expansion

$$G(t) \overset{t \to \alpha_2}{\sim} \frac{1}{\alpha_2(1-\frac{t}{\alpha_2})^2}.$$

Consequently, at its dominant singularity α_2, $U(z)$ satisfies

$$U(z) = \frac{z}{(z-1)^2} \int_0^z G(t) \frac{(t-1)^2}{t}\, dt \overset{z \to \alpha_2}{\sim} \frac{\alpha_2}{(\alpha_2-1)^2} \int_0^z \frac{1}{\alpha_2(1-\frac{t}{\alpha_2})^2} \frac{(\alpha_2-1)^2}{\alpha_2}\, dt \overset{z \to \alpha_2}{\sim} \frac{1}{1-\frac{z}{\alpha_2}}.$$

By applying eq. (5) together with eq. (72), we finally have

$$\mathbb{E}[T_n^2] = [z^n]U(z) \sim \alpha_2^{-n} \sim \mathbb{E}[T_n R_n] \sim \mathbb{E}[R_n^2]. \tag{73}$$

From these calculations, we obtain the next result.

Proposition 9 *The variance of the total number T_n of ancestral configurations in an ordered unlabeled history of size n selected under the uniform distribution satisfies the asymptotic relation $\mathbb{V}[T_n] \sim \mathbb{E}[T_n^2] \sim \mathbb{V}[R_n] \sim \alpha_2^{-n} = (2.0449954971\ldots)^n$. For increasing values of n, the correlation coefficient $\rho[T_n, R_n]$ between the total number T_n of ancestral configurations and the number R_n of root configurations converges to 1, $\rho[T_n, R_n] \to 1$.*

Proof. For the variance, we observe that $\mathbb{V}[T_n] = \mathbb{E}[T_n^2] - \mathbb{E}[T_n]^2 \sim \mathbb{E}[T_n^2]$. Indeed, from Proposition 8, we find $\mathbb{E}[T_n]^2 \bowtie (\alpha_1^{-2})^n$, where $\alpha_1^{-2} < \alpha_2^{-1}$, where α_2^{-1} is the exponential order of $\mathbb{E}[T_n^2]$, as in eq. (73). Also, $\mathbb{E}[T_n^2] \sim \mathbb{V}[R_n]$, because $\mathbb{E}[T_n^2] \sim \mathbb{E}[R_n^2] \sim \mathbb{V}[R_n]$ follows from eqs. (73) and (25).

Similarly, for the covariance between T_n and R_n we obtain

$$\mathrm{Cov}[T_n, R_n] = \mathbb{E}[T_n R_n] - \mathbb{E}[T_n]\,\mathbb{E}[R_n] \sim \mathbb{E}[T_n R_n] \sim \mathbb{E}[T_n^2].$$

Indeed, from Proposition 8 we have $\mathbb{E}[T_n]\,\mathbb{E}[R_n] \bowtie (\alpha_1^{-2})^n$, while from eq. (73), $\mathbb{E}[T_n R_n] \sim \mathbb{E}[T_n^2] \bowtie \alpha_2^{-n}$, with $\alpha_2^{-1} > \alpha_1^{-2}$. Hence, the correlation coefficient between T_n and R_n is

$$\rho[T_n, R_n] = \frac{\mathrm{Cov}[T_n, R_n]}{\sqrt{\mathbb{V}[T_n]}\sqrt{\mathbb{V}[R_n]}} \sim \frac{\mathbb{E}[T_n^2]}{\mathbb{E}[T_n^2]} = 1. \quad \square$$

By the same argument of eqs. (47), (48), and (49), for uniformly distributed ordered unlabeled histories, the total number T_n of ancestral configurations can be shown to follow an asymptotic lognormal distribution. In particular, the variables $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathbb{V}[\log T_n]}$ and $(\log R_n - \mathbb{E}[\log R_n])/\sqrt{\mathbb{V}[\log R_n]}$ for

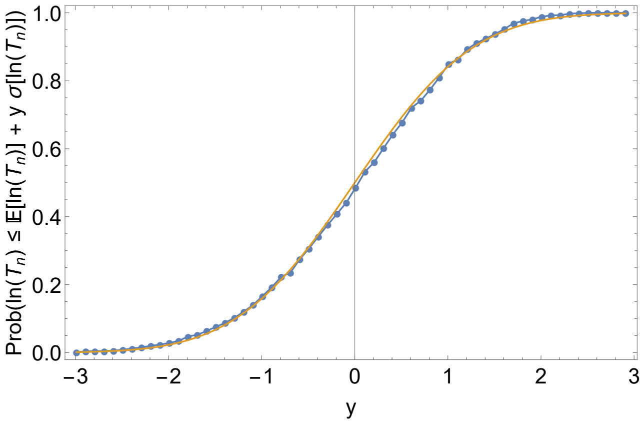


Figure 7: Cumulative distribution of the natural logarithm of the total number of configurations for uniformly distributed ordered unlabeled histories (or Yule-distributed labeled topologies) of size $n = 15$ (dotted line). For each $y \in [-3, 3]$ in steps of 0.1, the quantity plotted is the probability that an ordered unlabeled history (or Yule-distributed labeled topology) with $n = 15$ chosen at random has a total number of configurations less than or equal to $\exp(\mathbb{E}[\log T_n] + y\sigma[\log T_n])$, where $\mathbb{E}[\log T_n]$ and $\sigma[\log T_n] = \sqrt{\mathbb{V}[\log T_n]}$ are respectively the mean and standard deviation of the logarithm of the total number of configurations for uniformly distributed ordered unlabeled histories (or Yule-distributed labeled topologies) with $n = 15$ taxa. The solid line is the cumulative distribution of a Gaussian random variable with mean 0 and variance 1.

random ordered unlabeled histories of size n converge asymptotically to standard normal distributions, where $\mathbb{E}[\log T_n] \sim \mathbb{E}[\log R_n] \sim 0.351n$ and $\mathbb{V}[\log T_n] \sim \mathbb{V}[\log R_n] \sim 0.008n$ (eq. 26).

In Fig. 7, we plot the cumulative distribution $\mathbb{P}[\log T_n \leqslant \mathbb{E}[\log T_n] + y\sqrt{\mathbb{V}[\log T_n]}]$ as a function of y, when ordered unlabeled histories of size 15 are selected uniformly at random. To obtain the distribution, we can count total configurations for each unlabeled topology t with 15 leaves, and then count the number (eq. 19) of ordered unlabeled histories having t as tree shape. The figure illustrates the agreement between $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathbb{V}[\log T_n]}$ and the standard normal distribution.

By the equivalence in Lemma 1 between uniformly distributed ordered unlabeled histories and Yule-distributed labeled topologies, we summarize the results of this section.

Theorem 10 *For a labeled topology of size n selected at random under the Yule distribution, the mean and the variance of the total number T_n of ancestral configurations grow asymptotically like*

$$\mathbb{E}[T_n] \sim \left(\frac{1}{1 - e^{-2\pi\sqrt{3}/9}}\right)^n, \tag{74}$$

$$\mathbb{V}[T_n] \sim (2.0449954971\ldots)^n. \tag{75}$$

Furthermore, the logarithm of the total number of ancestral configurations in a Yule-distributed labeled topology of size n selected at random, rescaled as $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathrm{Var}[\log T_n]}$, converges to a standard normal distribution, where $\mathbb{E}[\log T_n] \sim \mu n$ and $\mathbb{V}[\log T_n] \sim \sigma^2 n$, $(\mu, \sigma^2) \approx (0.351, 0.008)$.

5 Conclusions

For a gene tree and species tree with matching labeled topology t of size n selected at random under the uniform and Yule probability models, we have studied the asymptotic distribution of the total number T_n of ancestral configurations of t. By using techniques of analytic combinatorics, we have extended results of [7] and [11], where the focus was on the number R_n of root ancestral configurations of t.

We have found that under both the uniform and Yule models, the total number of configurations has an asymptotic lognormal distribution, as was also demonstrated for the number of root configurations by [7]; the

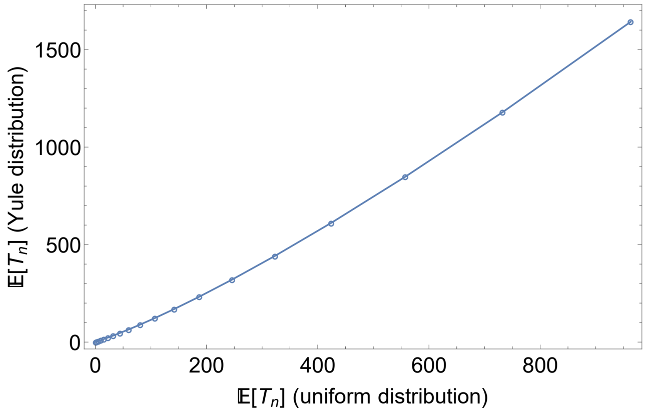


Figure 8: Mean total number of configurations of labeled topologies of size n under the Yule and uniform distributions, for $2 \leqslant n \leqslant 20$. Values for the uniform distribution are computed by combining the recurrences in eqs. (27) and (28); values for the Yule distribution are computed by combining the recurrences in eqs. (56) and (57).

lognormal distributions for total configurations ultimately trace to the fact that total configurations are bounded above by a linear multiple of root configurations. In Theorems 6 and 10, we have shown that the mean and the variance of the total number of ancestral configurations grow like $\mathbb{E}[T_n] \sim \sqrt{6}(4/3)^n$ and $\mathbb{V}[T_n] \sim 5.050 \cdot 1.822^n$, for uniformly distributed labeled topologies, and like $\mathbb{E}[T_n] \sim 1.425^n$ and $\mathbb{V}[T_n] \sim 2.045^n$, when labeled topologies of size n are selected under the Yule probability model. In particular, we observe that the mean total number of configurations is twice the mean number of root configurations under the uniform distribution for labeled topologies, with a correlation coefficient between T_n and R_n close to 0.9 for large n. For labeled topologies under the Yule distribution, the mean total number of configurations behaves asymptotically like the mean number of root configurations, with a correlation coefficient between T_n and R_n that approaches 1 for increasing n. A summary appears in Table 3.

That T_n and R_n have the same asymptotic growth under the Yule distribution on labeled topologies, and a correlation that approaches 1, is somewhat remarkable. R_n tabulates ancestral configurations only at the root, whereas T_n sums configurations across all $n-1$ internal nodes, *including* the root. The correlation result indicates that under the Yule distribution, the configurations at non-root nodes contribute negligibly to the total.

The difference in results for the uniform and Yule models suggests a correlation between tree balance and total configurations. Indeed, [11] suggested such a relationship for root configurations. A similar relationship might exist for total configurations; we find indeed that under the Yule model, which gives more weight to balanced topologies, the mean *total* number of configurations grows faster than under the uniform model (Fig. 8).

Several directions naturally arise from our work. For instance, we did not characterize the labeled topologies of given size that have the largest total number of ancestral configurations. Section 4 of [11] described the recursive structure of labeled topologies with the maximal number of root ancestral configurations. However, as shown in Fig. 9, the number of root and total configurations do not necessarily attain their maximal values at the same labeled topology. We also did not consider non-matching gene trees and species trees. The non-matching case, in which the gene tree and species tree have different labeled topologies, merits further analysis, as a non-matching gene-tree labeled topology can have more total configurations than the topology that matches the species tree [11]. It is of interest to see if techniques used in this article can be extended to derive distributional properties of the number of ancestral configurations when the gene tree and species tree differ in topology.

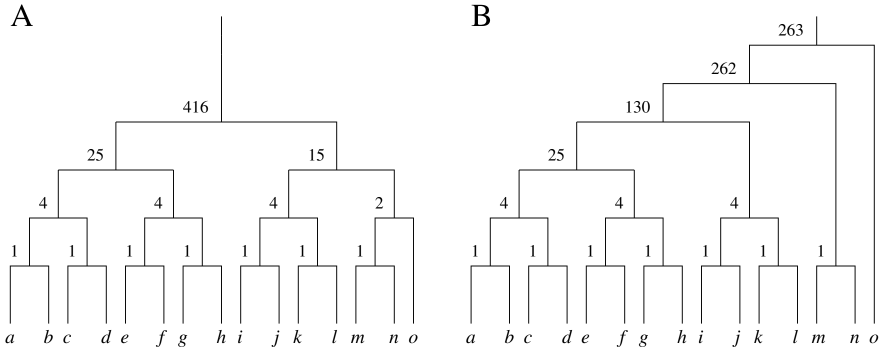


Figure 9: The labeled topologies of size $n = 15$ that (up to permutation of the labels) have maximal numbers of configurations. (A) Root configurations (416). (B) Total configurations (699). Numbers of configurations appear at each internal node. The rightmost point in Fig. 2 has coordinates $(\log 416, \log 477)$; the topmost point has coordinates $(\log 263, \log 699)$.

Table 3: Summary of asymptotic equivalences for the number of root configurations and the total number of configurations under the uniform and Yule models on labeled topologies.

	Uniform model		Yule model	
Quantity	Result	Reference	Result	Reference
$\mathbb{E}[R_n]$	$\sqrt{\frac{3}{2}}\left(\frac{4}{3}\right)^n$	eq. (24)	$\left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n$	eq. (24)
$\mathbb{E}[R_n^2]$	$\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	eq. (25)	$(2.0449954971\ldots)^n$	eq. (25)
$\mathbb{V}[R_n]$	$\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	eq. (25)	$(2.0449954971\ldots)^n$	eq. (25)
$\mathbb{E}[T_n]$	$\sqrt{6}\left(\frac{4}{3}\right)^n$	Prop. 3	$\left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n$	Prop. 8
$\mathbb{E}[T_n^2]$	$\frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	eq. (45)	$(2.0449954971\ldots)^n$	Prop. 9
$\mathbb{V}[T_n]$	$\frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	Prop. 5	$(2.0449954971\ldots)^n$	Prop. 9
$\mathbb{E}[T_nR_n]$	$\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	eq. (44)	$(2.0449954971\ldots)^n$	Prop. 9
$\mathrm{Cov}[T_n,R_n]$	$\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	Prop. 5	$(2.0449954971\ldots)^n$	Prop. 9
$\rho[T_n,R_n]$	$\frac{1+\frac{\sqrt{2}}{2}}{\sqrt{\frac{2}{17}(15+11\sqrt{2})}}$	Prop. 5	1	Prop. 9

We note the numerical values of recurring constants: $1/(1-e^{-2\pi\sqrt{3}/9}) \approx 1.4253868277$, $1-e^{-2\pi\sqrt{3}/9} \approx 0.70156394081$, $4/[7(8\sqrt{2}-11)] \approx 1.8215272244$, $7(8\sqrt{2}-11)/4 \approx 0.5489898732$. Constant 2.0449954971 was evaluated in the Appendix of [7], and its reciprocal is 0.4889986317.

Acknowledgments This project developed from discussions at a workshop at the Banff International Research Station. Support was provided by National Institutes of Health grant R01 GM117590 to NAR, by grants MOST-107-2115-M-009-010-MY2 and MOST-109-2115-M-004-003-MY2 from Ministry of Science and Technology (MOST), Taiwan to MF, CYH and ARP, and by a Rita Levi-Montalcini grant from the Ministero dell'Istruzione, dell'Università e della Ricerca to FD.

References

[1] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.*, 16:23–34, 2001.

[2] E. Alimpiev and N. A. Rosenberg. Enumeration of coalescent histories for caterpillar species trees and p-pseudocaterpillar gene trees. *Adv. Appl. Math.*, 131:102265, 2021.

[3] R. Bhatia and C. Davis. A better bound on the variance. *Amer. Math. Monthly*, 107:353–357, 2000.

[4] J. K. M. Brown. Probabilities of evolutionary trees. *Syst. Biol.*, 43:78–91, 1994.

[5] H. Chang and M. Fuchs. Limit theorems for patterns in phylogenetic trees. *J. Math. Biol.*, 60:481–512, 2010.

[6] J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.

[7] F. Disanto, M. Fuchs, A. R. Paningbatan, and N. A. Rosenberg. The distributions under two species-tree models of the number of root ancestral configurations for matching gene trees and species trees. *Ann. Appl. Prob.*, 32:4426–4458, 2022.

[8] F. Disanto and E. Munarini. Local height in weighted Dyck models of random walks and the variability of the number of coalescent histories for caterpillar-shaped gene trees and species trees. *SN Appl. Sci.*, 1:578, 2019.

[9] F. Disanto and N. A. Rosenberg. Coalescent histories for lodgepole species trees. *J. Comput. Biol.*, 22:918–929, 2015.

[10] F. Disanto and N. A. Rosenberg. Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 13:913–925, 2016.

[11] F. Disanto and N. A. Rosenberg. Enumeration of ancestral configurations for matching gene trees and species trees. *J. Comput. Biol.*, 24:831–850, 2017.

[12] F. Disanto and N. A. Rosenberg. Enumeration of compact coalescent histories for matching gene trees and species trees. *J. Math. Biol.*, 78:155–188, 2019.

[13] F. Disanto and N. A. Rosenberg. On the number of non-equivalent ancestral configurations for matching gene trees and species trees. *Bull. Math. Biol.*, 81:384–407, 2019.

[14] F. Disanto, A. Schlizio, and T. Wiehe. Yule-generated trees constrained by node imbalance. *Math. Biosci.*, 246:139–147, 2013.

[15] F. Disanto and T. Wiehe. Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. *Math. Biosci.*, 242:195–200, 2013.

[16] J. Felsenstein. The number of evolutionary trees. *Syst. Zool.*, 27:27–33, 1978.

[17] J. Felsenstein. *Inferring Phylogenies.* Sinauer, Sunderland, MA, 2004.

[18] P. Flajolet and R. Sedgewick. *Analytic Combinatorics.* Cambridge University Press, Cambridge, 2009.

[19] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Appl. Prob.*, 3:44–77, 1971.

[20] Z. M. Himwich and N. A. Rosenberg. Roadblocked monotonic paths and the enumeration of coalescent histories for non-matching caterpillar gene trees and species trees. *Adv. Appl. Math.*, 113:101939, 2020.

[21] M. C. King and N. A. Rosenberg. A mathematical connection between single-elimination sports tournaments and evolutionary trees. *Math. Mag.*, 96:in press, 2023.

[22] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Math. Biosci.*, 164:81–92, 2000.

[23] R. Otter. The number of trees. *Ann. Math.*, 49:583–599, 1948.

[24] N. A. Rosenberg. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. *Ann. Comb.*, 10:129–146, 2006.

[25] N. A. Rosenberg. Counting coalescent histories. *J. Comput. Biol.*, 14:360–377, 2007.

[26] N. A. Rosenberg. Coalescent histories for caterpillar-like families. *IEEE/ACM Trans. Comp. Biol. Bioinf.*, 10:1253–1262, 2013.

[27] N. A. Rosenberg. Enumeration of lonely pairs of gene trees and species trees by means of antipodal cherries. *Adv. Appl. Math.*, 102:1–17, 2019.

[28] N. A. Rosenberg and J. H. Degnan. Coalescent histories for discordant gene trees and species trees. *Theor. Pop. Biol.*, 77:145–151, 2010.

[29] R. J. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley, New York, 1980.

[30] R. P. Stanley. *Enumerative Combinatorics Volume 2.* Cambridge University Press, New York, 1999.

[31] M. Steel. *Phylogeny: Discrete and Random Processes in Evolution.* Society for Industrial and Applied Mathematics, Philadelphia, 2016.

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

[33] J. Truszkowski, C. Scornavacca, and F. Pardi. Computing the probability of gene trees concordant with the species tree in the multispecies coalescent. *Theor. Pop. Biol.*, 137:22–31, 2021.

[34] J. H. M. Wedderburn. The functional equation $g(x^2) = 2\alpha x + [g(x)]^2$. *Ann. Math.*, 24:121–140, 1922.

[35] Y. Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66:763–775, 2012.

[36] Y. Wu. An algorithm for computing the gene tree probability under the multispecies coalescent and its application in the inference of population tree. *Bioinformatics*, 32:i225–i233, 2016.

[37] G. U. Yule. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F. R. S. *Phil. Trans. R. Soc. Lond. B*, 213:21–87, 1925.