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

Filippo Disanto[*], Michael Fuchs[†], Ariel R. Paningbatan[‡], Noah A. Rosenberg[§]

September 17, 2021

Abstract

For a pair consisting of a gene tree and a species tree, the *ancestral configurations* at an internal node of the species tree are the distinct sets of gene lineages that can be present at that node. Ancestral configurations appear in computations of gene tree probabilities under evolutionary models conditional on fixed species trees, and the enumeration of *root* ancestral configurations—ancestral configurations at the root of the species tree—assists in describing the complexity of these computations. In the case that the gene tree matches the species tree in topology, we study the distribution of the number of root ancestral configurations of a random labeled tree topology under each of two models. We use analytic combinatorics to perform the calculations, considering ancestral configurations in the context of additive tree parameters and using singularity analysis of generating functions to evaluate the asymptotic growth of their coefficients. First, choosing a tree uniformly at random from the set of labeled topologies with n leaves, we extend an earlier computation of the asymptotic exponential growth of the mean and variance of the number of root ancestral configurations, showing that the number of root ancestral configurations of a random tree asymptotically follows a lognormal distribution; the logarithm has mean $\sim 0.272n$ and variance $\sim 0.034n$. The asymptotic mean of the logarithm of the number of root ancestral configurations produces $e^{0.272n} \approx 1.313^n$ when exponentiated, numerically close to the previously obtained mean of $(4/3)^n$ for the exponential growth of the number of root ancestral configurations. Next, considering labeled topologies selected according to the Yule–Harding model, we obtain the asymptotic mean and variance of the number of root ancestral configurations of a random tree and the asymptotic distribution of its logarithm. The asymptotic mean follows $\sim 1.425^n$ and the variance follows $\sim 2.045^n$; the random variable has an asymptotic lognormal distribution, and its logarithm has mean $\sim 0.351n$ and variance $\sim 0.008n$. The asymptotic mean of the logarithm produces $e^{0.351n} \approx 1.420^n$, close to 1.425^n. With the higher probabilities assigned by the Yule–Harding model to balanced trees in comparison with those assigned under the uniform model, a larger asymptotic exponential growth $\sim 1.425^n$ of the mean number of root ancestral configurations for the Yule–Harding model compared to $(4/3)^n$ in the uniform model suggests an effect of increasing tree balance in increasing the number of root ancestral configurations. A methodological innovation of our approach is that to calculate the Yule–Harding asymptotic variance $\sim 2.045^n$, singularity analysis of a generating function to obtain asymptotic growth is conducted from the Riccati differentiation that the generating function satisfies—without possessing the exact form for the generating function.

[*]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.

[‡]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.

- **Keywords:** analytic combinatorics, gene trees, lognormal distribution, phylogenetics, Riccati equation, species trees.

- **Mathematics subject classification (2010):** 05A15 · 05A16 · 60C05 · 92B10 · 92D15

- **Running title:** Root ancestral configurations

1 Introduction

In the study of combinatorial properties of species trees (trees that describe evolutionary relationships among species) and gene trees (trees that describe evolutionary relationships among gene lineages for members of the species), one useful concept is that of an ancestral configuration. Given a gene tree, a species tree, and a node of the species tree, an ancestral configuration is a list of the gene lineages that are present at the node of the species tree (Fig. 1). Looking backward in time, or from the leaves of trees to the root, the fact that gene lineages only find their common ancestors once their associated species have found common ancestors produces conditions describing which ancestral configurations are present at a species tree node. These conditions enable the enumeration of the configurations. Ancestral configurations appear in recursive evaluations of the probabilities of gene tree topologies conditional on species tree topologies [46], so that enumerations of ancestral configurations assist in assessing the complexity of the computation.

When the node at which an ancestral configuration is considered is the root node of the species tree, ancestral configurations are termed *root ancestral configurations*, or root configurations for short. For matching gene trees and species trees—that is, if the species tree and gene tree have the same labeled topology—the number of root configurations is greater than or equal to the number of ancestral configurations for any other species tree node. This property can be used to show that as the number of leaves increases, the total number of ancestral configurations for the gene tree and species tree—the sum of the number of ancestral configurations across all species tree nodes—has the same exponential growth as the number of root configurations [14, Section 2.3.2]. Hence, it suffices for investigations of the exponential growth of the total number of ancestral configurations for matching gene trees and species trees to focus on root configurations.

Disanto & Rosenberg [14] studied the number of root configurations for matching gene trees and species trees, considering the number of root configurations of families of increasingly large trees. They characterized the labeled tree topologies with the largest number of root configurations among trees with n leaves, showing that this number of root configurations lies between $k_0^{n-1/4}-1$ and k_0^n-1, where k_0 is a constant approximately equal to 1.5028 [14, Proposition 4]. They then studied the number of root configurations in trees selected uniformly at random from the set of labeled topologies with n leaves. Using techniques of analytic combinatorics, they showed that the mean number of root configurations grows with $(4/3)^n$, and the variance with $\sim 1.8215^n$ [14, Propositions 5 and 6].

Here, we extend these results on the distribution of the number of root configurations under a model imposing a uniform distribution on the set of labeled topologies. We review background results in Section 2. In Section 3, we describe correspondences between classes of trees, which we use in Section 4 to obtain an asymptotic normal distribution for the logarithm of the number of root configurations under the uniform model—and find that its mean, approximately $0.272n$, generates exponential growth $e^{0.272n} \approx 1.313^n$. In Section 5, we obtain similar results under the Yule–Harding model, including the asymptotic mean and variance of the number of root configurations and the asymptotic distribution of its logarithm. This set of computations also makes use of a correspondence between tree classes. The calculation of the asymptotic variance additionally employs a novel approach, in which asymptotic growth of the coefficients of a generating function that solves a Riccati equation is obtained without having the exact form of the generating function itself. We discuss the results in Section 6.

2 Preliminaries

We study ancestral configurations for rooted binary leaf-labeled trees. In Section 2.1, we introduce results on various classes of trees. In Section 2.2, we discuss the Yule–Harding distribution on labeled topologies. In Section 2.3, we recall properties of generating functions and analytic combinatorics. Following Wu [46], in Section 2.4 we define ancestral configurations, and we review enumerative results from Disanto & Rosenberg [14]. In Section 2.5, we relate ancestral configurations to additive tree parameters, which have been widely studied in the literature [27; 45].

2.1 Classes of trees

We will need to consider many classes of trees: labeled topologies, unlabeled topologies, ordered unlabeled topologies, labeled histories, unlabeled histories, and ordered unlabeled histories. Many terms in the setting of evolutionary trees can be connected to concepts from settings that do not have a biological context [1; 4; 7]; our terminology generally follows that typical of mathematical studies of evolutionary trees [39].

2.1.1 Labeled topologies

We refer to a bifurcating rooted tree t with $|t| = n$ labeled leaves as a *labeled topology* of size $|t| = n$, or a "tree" for short (Fig. 1A); these trees are sometimes called phylogenetic trees or Schröder trees. They are *unordered* or *non-plane* in the sense that if left–right positions of two child nodes are exchanged in a labeled topology, then the same labeled topology is obtained. For the set $\{a, b, c, \dots\}$ of possible labels for the leaves of a tree, we impose an alphabetical linear order $a \prec b \prec c \prec \dots$ The leaf labels of a tree of size n are the first n labels in order $\prec$.

We denote by T_n the set of trees of size n, with $T = \bigcup_{n=1}^{\infty} T_n$ denoting the set of all trees. The number of trees of size $n \geq 2$ is $|T_n| = (2n-3)!! = 1 \times 3 \times 5 \times \dots \times (2n-3)$ [19], or, for $n \geq 1$,

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

The exponential generating function for $|T_n|$ is

$$T(z) = \sum_{t \in T} \frac{z^{|t|}}{|t|!} = \sum_{n=1}^{\infty} \frac{|T_n| z^n}{n!} = z + \frac{z^2}{2} + \frac{3z^3}{6} + \frac{15z^4}{24} + \dots,$$

given by Flajolet & Sedgewick [23, Example II.19]

$$T(z) = 1 - \sqrt{1-2z}. \tag{2}$$

2.1.2 Ordered unlabeled topologies

An *orientation* of an unlabeled topology t is a planar embedding of t in which subtrees descending from the internal nodes of t are considered with a left–right orientation. For instance, the *unlabeled topology* underlying the labeled topology depicted in Fig. 1A has exactly two different orientations, which are depicted in Fig. 2A. An orientation of an unlabeled topology is called an *ordered* unlabeled topology, or a *plane* unlabeled topology. The set of all possible ordered unlabeled topologies of size n is enumerated by the Catalan number C_{n-1} [38, Exercise 6.19d], where

$$C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{3}$$

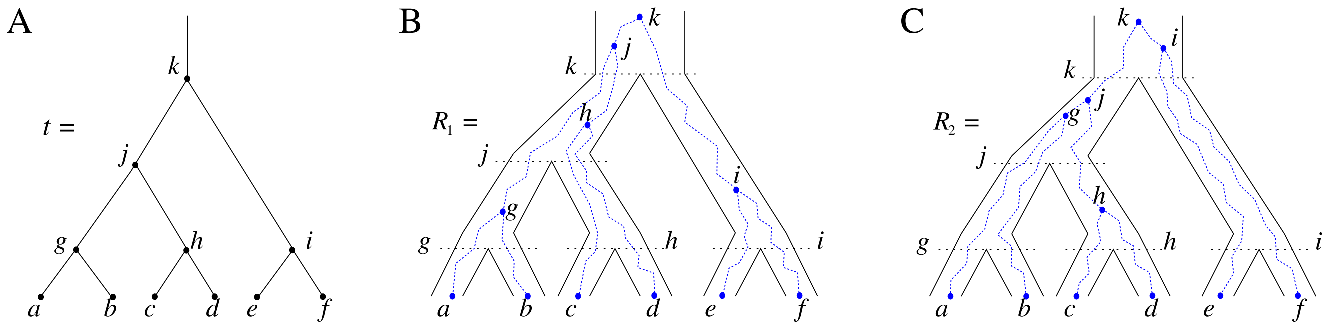

Figure 1: A gene tree and species tree with matching labeled topology t. **(A)** A tree t of size 6, characterized by its shape and leaf labels. For convenience, we label the internal nodes of t, by g, h, i, j, k in this case, identifying each lineage (edge) by its immediate descendant node. For example, lineage h results from coalescence of lineages c and d. **(B)** A possible realization R_1 of the gene tree in (A) (dotted lines) in the matching species tree (solid lines). The ancestral configurations at species tree nodes j and k are $\{g, c, d\}$ and $\{g, h, i\}$, respectively. **(C)** A different realization R_2 of the gene tree in (A) in the species tree. At species tree nodes j and k, the configurations are $\{a, b, h\}$ and $\{j, e, f\}$, respectively. The figure is modified from Figure 1 of Disanto & Rosenberg [14] and Figure 1 of Disanto & Rosenberg [16].

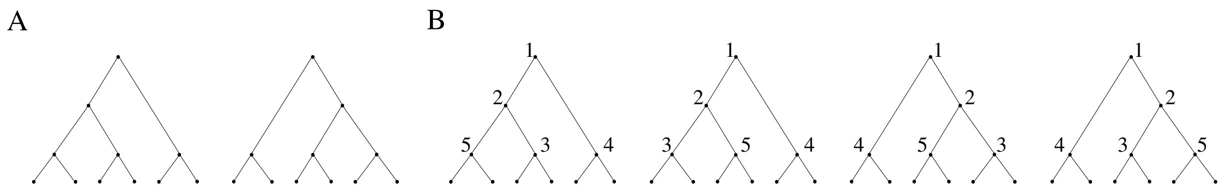

Figure 2: Ordered unlabeled topologies and histories. **(A)** The two orientations of the unlabeled topology that underlies the labeled topology of Fig. 1A. **(B)** The four orientations of the unlabeled history underlying the labeled history in Fig. 3A.

The ordinary generating function is

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

With the leaves and associated incident edges stripped away so that only the tree connecting the internal nodes remains, an ordered unlabeled topology is also called a *Catalan tree* or *pruned binary tree*, for example by Wagner [45] (see also Flajolet & Sedgewick [23], Example I.13).

2.1.3 Labeled histories

A *labeled history* is a labeled topology together with a temporal (linear) ordering of its internal nodes (Fig. 3). Like a labeled topology, a labeled history is left–right unordered, or non-plane: if the left–right positions of two child nodes are interchanged in a labeled history, then the same labeled history is obtained. If t is a labeled history of size n, then we represent the time ordering of its $n-1$ bifurcations by bijectively associating each internal node of t with an integer label in the interval $[1, n-1]$. The labeling is increasing in the sense that each internal node other than the root has a larger label than its parent node.

For a given label set of size n, the set of labeled histories is denoted H_n. Its cardinality is [39, p. 46]

$$|H_n| = \frac{n!\,(n-1)!}{2^{n-1}}. \tag{4}$$

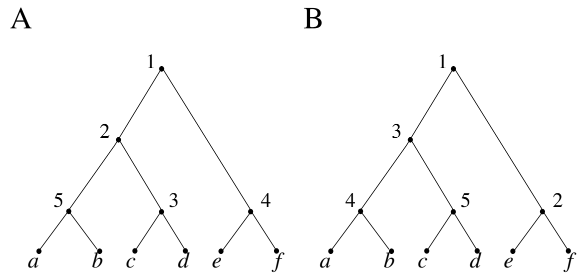

Figure 3: Labeled histories. **(A)** The labeled history of the labeled gene tree topology depicted in Fig. 1B. The temporal ordering of the coalescence events in the gene tree is determined by the integer labeling of the internal nodes of the associated labeled topology. **(B)** The labeled history of the labeled gene tree topology depicted in Fig. 1C.

2.1.4 Ordered unlabeled histories

By removing leaf labels of a labeled history t, we obtain the unlabeled history underlying t. As we did for unlabeled topologies, we define an orientation of an unlabeled history t as a planar embedding of t in which child nodes are considered with a left–right orientation. Fig. 2B shows the orientations of the unlabeled history underlying the labeled history of Fig. 3A. We call each object so oriented an *ordered* unlabeled history, or a *plane* unlabeled history. The ordered unlabeled histories of size n are enumerated by F_{n-1} [39, p. 47],

$$F_n = n!. \tag{5}$$

Ordered unlabeled histories are also called *binary increasing trees* [3; 45] or *ranked oriented trees* [39].

2.2 The Yule–Harding distribution

Different labeled histories can share the same underlying labeled topology. For example, the labeled histories of Fig. 3 have the underlying labeled topology depicted in Fig. 1A. The number of labeled histories of size n with the same labeled topology t is

$$\frac{(n-1)!}{\prod_{r=3}^{n}(r-1)^{d_r(t)}}, \tag{6}$$

where $d_r(t)$ is the number of internal nodes of t from which exactly r leaves descend [39, p. 46]. Eq. (6) also appears as the so-called "shape functional" of binary search trees [20].

By summing the probability $1/|H_n|$ of each uniformly distributed labeled history of size n with a given underlying labeled topology, the uniform distribution over the set H_n induces the Yule–Harding (or Yule) distribution over the set T_n of labeled topologies [6; 7; 17; 18; 25; 31; 32; 40; 48]. The probability of a labeled topology t is

$$\mathrm{P_{YH}}(t) = \frac{2^{n-1}}{n!\prod_{r=3}^{n}(r-1)^{d_r(t)}}. \tag{7}$$

Under this distribution, among all labeled topologies with size n, those with the largest number of labeled histories have the highest probability. For balanced labeled topologies, the product in the denominator of Eq. (7) tends to be smaller than for unbalanced topologies, resulting in a greater probability.

2.3 Asymptotic growth and analytic combinatorics

Our study concerns the growth of increasing sequences. A sequence of non-negative numbers a_n is said to have exponential growth k^n or, equivalently, to be of exponential order k, if $a_n = k^n s(n)$, where s is subexponential, that is, $\limsup_{n\to\infty}[s(n)^{1/n}] = 1$. Sequence a_n grows exponentially in n if its exponential order exceeds 1.

If (a_n) has exponential order k_a and (b_n) has exponential order $k_b < k_a$, then the sequence of ratios b_n/a_n converges to 0 exponentially fast as $(k_b/k_a)^n$. If sequences a_n and b_n have the same exponential order, then we write $a_n \bowtie b_n$. If in addition the ratio b_n/a_n converges to 1, then we write $a_n \sim b_n$ and say that (a_n) and (b_n) have the same asymptotic growth.

Some results make use of techniques of analytic combinatorics (Flajolet & Sedgewick [23], Sections IV and VI). In particular, the entries of a sequence of integers $(a_n)_{n\geq 0}$ can be interpreted as coefficients of the power series expansion $A(z) = \sum_{n=0}^{\infty} a_n z^n$ at $z = 0$ of a function $A(z)$, the generating function of the sequence. Considering z as a complex variable, the behavior of $A(z)$ near its singularities—the points in the complex plane where $A(z)$ is not analytic—can provide information on the growth of its coefficients. Under suitable conditions, a correspondence exists between the expansions $A_\alpha(z)$, $\alpha \in S$, of the generating function $A(z)$ near singularities in its set S of dominant singularities—that is, its singularities of smallest modulus—and the asymptotic growth of the coefficients a_n. In the simplest case, if α is the only dominant singularity of $A(z)$, then the nth coefficient a_n of $A(z)$ has asymptotic growth $[z^n]A_\alpha(z)$, that is, the nth coefficient of $A_\alpha(z)$ (Theorem VI.4 of Flajolet & Sedgewick [23]). In symbols,

$$a_n \sim [z^n]A_\alpha(z).$$

The exponential order of sequence (a_n) is the inverse of the modulus of the dominant singularity α of $A(z)$ (Theorem IV.7 of Flajolet & Sedgewick [23]). That is,

$$a_n \bowtie \alpha^{-n}.$$

As an example, sequence $|T_n|/n!$, with $|T_n|$ as in Eq. (1), has exponential order 2 because $\alpha = \frac{1}{2}$ is the only dominant singularity of the associated generating function in Eq. (2). Thus, as $n \to \infty$, $|T_n|/n!$ increases with a subexponential multiple of 2^n.

2.4 Ancestral configurations for matching gene trees and species trees

In this section, following Disanto & Rosenberg [14], we review features of the objects on which our study focuses: the ancestral configurations of a gene tree G in a species tree S.

2.4.1 Gene trees and species trees

A *species tree* is a tree of evolutionary relationships among a set of species. A *gene tree* is a tree of evolutionary relationships among individual genetic lines of descent, or lineages, at a specific genomic site. Gene trees and species trees are typically viewed as objects evolving forward in time, from the root to the leaves, or backward in time, from the leaves to the root. They consist of both a labeled topology and a set of edge lengths, positive values that describe the lengths of time separating pairs of nodes.

In studies of gene trees and species trees, the leaf label set of a gene tree G is often taken to be a subset of the leaf label set of a species tree S, so that a gene tree evolves conditionally on the species tree. Here, because we consider only the combinatorial structure of gene trees and species trees, we are not concerned with numerical values of edge lengths. Hence, it is convenient to identify a gene tree or a species tree with its associated labeled topology; for ease of understanding, however, it is still said that a gene tree or species tree "has" a labeled topology rather than that it "is" a labeled topology. Because we are concerned with ancestor–descendant relationships, it is also convenient to retain a perspective that gene trees and species trees unfold over time.

We here examine the case that the leaf label sets of G and S are bijectively associated. In other words, a single genetic lineage is sampled from each species corresponding to a leaf of the species tree. We further restrict attention to the case in which G and S have the same labeled topology, so that the gene tree and species tree are said to be *matching*. With the perspective that a gene tree unfolds over time conditionally on a species tree, an instance of the evolutionary process that produces gene tree G on species tree S is a *realization* of G on S.

Looking backward in time, the lineages of G are traced back past nodes of S until the root of G is reached; at a given point in time, a lineage of G is associated with a label that contains information about which leaves descend from it. For convenience, a node of a gene tree or species tree is associated with its immediate ancestral edge, so that a node and its immediate ancestral edge are assigned the same label.

2.4.2 Ancestral configurations

An ancestral configuration can be viewed as a certain function of a realization of G on S, with G and S representing a gene tree and a species tree, respectively, and of a node of S. Suppose R is a realization of a gene tree G on a species tree S, where $G = S = t$ (Fig. 1). Looking backward in time, for node η of S, consider the set $C(\eta, R)$ of genetic lineages—edges of G—that are present in S at the point in time just before node η is reached.

The set $C(\eta, R)$ is the *ancestral configuration* of G at node η of S. For example, for tree t in Fig. 1A, with the realization R_1 of gene tree $G = t$ in the species tree $S = t$ in Fig. 1B, just before the root node k, the gene lineages present in the species tree are lineages g, h, and i. Hence, at species tree node k, the ancestral configuration is the set of gene lineages $C(k, R_1) = \{g, h, i\}$. Similarly, the ancestral configuration of the gene tree at species tree node j is $C(j, R_1) = \{g, c, d\}$. In Fig. 1C, with a different realization R_2 of the same gene tree, the ancestral configuration at the species tree root k is $C(k, R_2) = \{j, e, f\}$. The ancestral configuration at node j is $C(j, R_2) = \{a, b, h\}$.

Let $\Re(G, S)$ be the set of realizations of gene tree $G = t$ in species tree $S = t$. For a given node η of t, considering all possible elements $R \in \Re(G, S)$, the set of ancestral configurations is

$$C(\eta) = \{C(\eta, R) : R \in \Re(G, S)\}. \tag{8}$$

The associated number of ancestral configurations is

$$c_\eta = |C(\eta)|. \tag{9}$$

The quantity c_η counts the ways the lineages of G can reach the timepoint right before node η in S, considering all possible realizations of gene tree G in species tree S. Choosing t as in Fig. 1A, we have $C(g) = \{\{a, b\}\}$, $C(h) = \{\{c, d\}\}$, $C(i) = \{\{e, f\}\}$, $C(j) = \{\{a, b, c, d\}, \{g, c, d\}, \{a, b, h\}, \{g, h\}\}$, and

$$C(k) = \{\{j, i\}, \{j, e, f\}, \{g, h, i\}, \{g, h, e, f\}, \{a, b, h, i\}, \{a, b, h, e, f\}, \{g, c, d, i\}, \{g, c, d, e, f\}, \{a, b, c, d, i\}, \{a, b, c, d, e, f\}\}. \tag{10}$$

For different realizations $R_1, R_2 \in \Re(G, S)$ and an internal node η, it need not be true that $C(\eta, R_1) \neq C(\eta, R_2)$.

We say that a leaf or a 1-leaf tree has no ancestral configurations. The definition of an ancestral configuration at node η, by considering the timepoint right before node η in the species tree, excludes the case in which all gene tree lineages descended from gene tree node η have coalesced at species tree node η. Thus, $\{\eta\} \notin C(\eta)$.

Because we consider the case of $G = S = t$, the set $C(\eta)$ and the quantity c_η in Eqs. (8) and (9) depend only on node η and tree t. We use the term *configurations at node* η *of* t to denote elements of $C(\eta)$.

2.4.3 Root and total configurations

Our focus is on configurations at the root of t. Let $N(t)$ be the set of nodes of a tree t, including both leaf nodes and internal nodes. With $|t|$ leaf nodes and $|t| - 1$ internal nodes in t, $|N(t)| = 2|t| - 1$. Define the *total* number of configurations in t by

$$c = \sum_{\eta \in N(t)} c_\eta.$$

Let c_r be the number of configurations at the root r of t, or *root configurations* for short. Because $c_r \geq c_\eta$ for each node η of t, we have

$$c_r \leq c \leq (2|t| - 1)c_r. \tag{11}$$

Quantities c and c_r are equal up to a factor that is at most polynomial in $|t|$, and they have the same exponential order when measured across families of trees of increasing size.

Selecting a tree of size n at random from the set of labeled topologies, inequality (11) gives $\mathbb{E}_n[c_r] \leq \mathbb{E}_n[c] \leq 2n\mathbb{E}_n[c_r]$ and $\mathbb{E}_n[c_r^2] \leq \mathbb{E}_n[c^2] \leq 4n^2\mathbb{E}_n[c_r^2]$. In expectation $\mathbb{E}$ and variance $\mathbb{V}$, exponential growth for total configurations follows that for root configurations:

$$\mathbb{E}_n[c] \bowtie \mathbb{E}_n[c_r] \tag{12}$$
$$\mathbb{E}_n[c^2] \bowtie \mathbb{E}_n[c_r^2] \tag{13}$$
$$\mathbb{V}_n[c] = \mathbb{E}_n[c^2] - \mathbb{E}_n[c]^2 \bowtie \mathbb{E}_n[c_r^2] - \mathbb{E}_n[c_r]^2 = \mathbb{V}_n[c_r]. \tag{14}$$

Eq. (14) follows from the fact that the exponential growth of $\mathbb{E}_n[c^2]$ is faster than that of $\mathbb{E}_n[c]^2$, as can be demonstrated from results in the next section (Eqs. (17) and (19)), and the exponential growth of $\mathbb{E}_n[c_r^2]$ is faster than that of $\mathbb{E}_n[c_r]^2$ (Eqs. (16) and (18)); we then have $\mathbb{V}_n[c] \sim \mathbb{E}_n[c^2]$ and $\mathbb{V}_n[c_r] \sim \mathbb{E}_n[c_r^2]$, and Eq. (14) follows from Eq. (13).

2.4.4 Known results

We recall some results of Disanto & Rosenberg [14] on the number of configurations possessed by a tree.

(i) For a given tree t with $|t| > 1$, let r denote the root node of t, with r_L and r_R being the two child nodes of r. The number c_r of possible configurations at r can be recursively computed as

$$c_r = (c_{r_L} + 1)(c_{r_R} + 1), \tag{15}$$

where we set $c_r = 0$ if $|t| = 1$. Fig. 4 illustrates the application of Eq. (15) successively from the leaves to the root of each of three labeled topologies of size $n = 15$.

(ii) Consider a representative labeling of each unlabeled topology of size n. Among these trees, the largest number of root configurations and the largest total number of configurations have exponential order k_0, where $k_0 \approx 1.5028$. The smallest number of root configurations and the smallest total number of configurations have polynomial growth with the tree size n. Furthermore, consider the balanced family of unlabeled topologies defined recursively by $|t_1| = 1$ and $t_n = (t_d, t_{n-d})$, where d denotes the power of 2 nearest to $\frac{n}{2}$. Among the unlabeled topologies with n leaves, t_n has the largest number of root configurations. The maximally asymmetric caterpillar unlabeled topology has the smallest number of root configurations.

(iii) For a labeled topology of given size n selected uniformly at random, the mean number of root configurations c_r and the mean total number of configurations c grow asymptotically like

$$\mathbb{E}_n[c_r] \sim \sqrt{\frac{3}{2}}\left(\frac{4}{3}\right)^n, \tag{16}$$
$$\mathbb{E}_n[c] \bowtie \left(\frac{4}{3}\right)^n. \tag{17}$$

The variances of c_r and c satisfy the asymptotic relations

$$\mathbb{V}_n[c_r] \sim \sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n, \tag{18}$$
$$\mathbb{V}_n[c] \bowtie \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n. \tag{19}$$

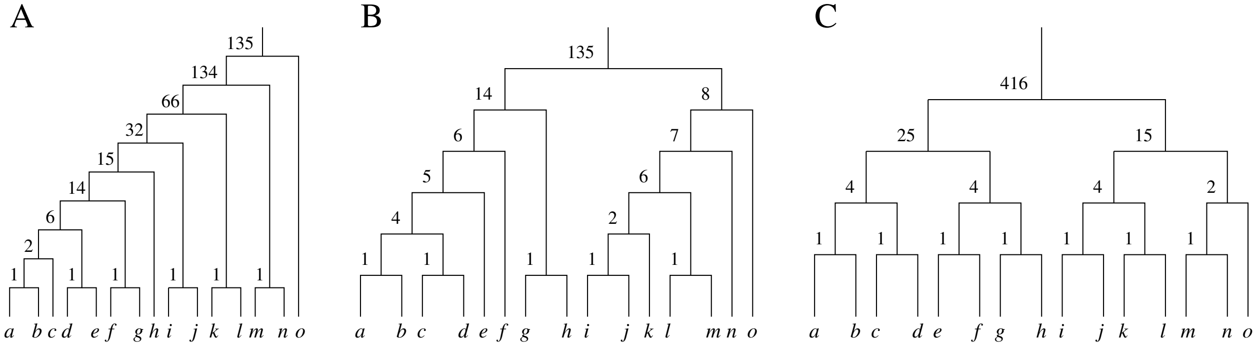

Figure 4: The number of ancestral configurations at the internal nodes of three labeled topologies of size $n = 15$. **(A, B)** Two labeled topologies in which the number of root configurations is the mean number $c_r = 135$ of root configurations calculated across the set of representative labelings of the unlabeled topologies of size 15. In this set, the labeled topologies in (A) and (B) have respectively the largest number 61776 and smallest number 14400 of labeled histories. **(C)** The labeled topology with 15 leaves that has the most root configurations (416) and the most labeled histories (2745600).

2.5 Additive tree parameters and root configurations

A quantity $F(t)$ that can be computed for trees t and whose value can be calculated as

$$F(t) = F(t_L) + F(t_R) + f(t),$$

where t_L and t_R are the two root subtrees of t, is called an *additive tree parameter* with *toll function* $f(t)$ [21; 27; 45]. Additive tree parameters and toll functions have been widely investigated [27, Remark 1.16]. We make use of results from Wagner [45]. For various tree families, Wagner [45] showed that an additive tree parameter $F(t)$ is asymptotically normally distributed if the toll function $f(t)$ is bounded and the mean of $|f(t)|$, considered over uniformly distributed trees of fixed size, goes to 0 exponentially fast as the tree size increases.

For a tree t, consider the quantity $\log(c_r + 1)$, that is, the natural logarithm of one more than the number of root configurations of t. From Eq. (15), a simple calculation yields for $|t| \geq 2$

$$\log(c_r + 1) = \log(c_{r_L} + 1) + \log(c_{r_R} + 1) + \log\left(1 + \frac{1}{c_r}\right). \tag{20}$$

In Eq. (20), if we set

$$F(t) = \log[c_r(t) + 1],$$

then the associated toll function is given for $|t| \geq 2$ by

$$f(t) = \log\left[1 + \frac{1}{c_r(t)}\right].$$

We set $f(t) = F(t) = \log(1) = 0$ if $|t| = 1$. We can therefore consider root configurations in the context of additive tree parameters.

3 Equivalences for the distribution of the number of root configurations

We prove a series of equivalences needed for analyzing distributional properties of the number of root configurations. In Section 3.1, we show that the distribution of the number of root configurations over uniformly distributed labeled topologies or labeled histories can be analyzed by considering equivalently the distribution of the number of root configurations over uniformly distributed ordered unlabeled topologies or ordered unlabeled histories, respectively. In Section 3.2, we obtain a correspondence between antichains of pruned binary trees and root configurations of ordered unlabeled topologies.

3.1 Equivalences with ordered unlabeled topologies and histories

Distributional properties of a tree parameter defined over the set of labeled topologies can in some cases be investigated by studying the same parameter over a different tree family. In particular, if the tree parameter under consideration depends only on tree topology, then its distribution can be equivalently analyzed over a different tree set taken under a probability model that induces or is induced by the probability model assumed for labeled topologies. In this direction, Blum *et al.* [4] derived a general framework for analyzing tree parameters of labeled topologies under a variety of probabilistic models defined over binary search trees.

In this section, we obtain results analogous to those of Blum *et al.* [4]. We show that the number of root configurations—or any other tree parameter that depends only on the branching structure of the tree—has the same distribution when considered over uniformly distributed labeled topologies or over uniformly distributed ordered unlabeled topologies of the same size (Lemma 1). Similarly, the number of root configurations has the same distribution over uniformly distributed labeled histories of size n as for uniformly distributed ordered unlabeled histories of size n (Lemma 2).

Moreover, because the uniform distribution over the set of labeled histories of size n induces the Yule–Harding distribution over the set of labeled topologies of size n (Section 2.2), as a direct consequence of Lemma 2 we have that the number of root configurations has the same distribution when considered over Yule–Harding-distributed labeled topologies or over uniformly distributed ordered unlabeled histories (Lemma 3). By using these facts, Propositions 1 and 2 give recursive formulas for the probabilities under the uniform and Yule–Harding probability models, respectively, that a random labeled topology of size n has $c_r = \rho$ root configurations.

Lemma 1 *The distribution of the number of root configurations over labeled topologies of size n selected uniformly at random matches the distribution of the number of root configurations over ordered unlabeled topologies of size n selected uniformly at random.*

Proof. First, we note that the number of root configurations of a labeled topology or ordered unlabeled topology depends only on the underlying unlabeled topology. Thus, to prove the claim, it suffices to show that for each unlabeled topology t of size n, we have

$$\frac{\mathrm{or}(t)}{C_{n-1}} = \frac{\mathrm{lab}(t)}{|T_n|}, \tag{21}$$

where $\mathrm{or}(t)$ and $\mathrm{lab}(t)$ are the number of orientations of t and the number of leaf labelings of t, respectively. Note from Eqs. (3) and (1) that $\mathrm{or}(t)/C_{n-1}$ and $\mathrm{lab}(t)/|T_n|$ give the probability of the unlabeled topology t induced by the uniform distribution over the set of ordered unlabeled topologies and labeled topologies of n leaves, respectively.

By using $C_{n-1} = \binom{2n-2}{n-1}/n$ and $|T_n| = (2n-2)!/[2^{n-1}(n-1)!]$ from Eqs. (3) and (1), Eq. (21) can be rewritten

$$\mathrm{lab}(t) = \mathrm{or}(t)\frac{n!}{2^{n-1}},$$

which we demonstrate by induction on the size of t. Let t_L and t_R be the two root subtrees of t, with sizes $|t_L| = L$ and $|t_R| = R$. Thus, for $n \geq 2$,

$$\mathrm{lab}(t) = \mathrm{lab}(t_L)\,\mathrm{lab}(t_R)\binom{n}{L}\frac{1}{1+\delta_{t_L=t_R}} \tag{22}$$

$$\mathrm{or}(t) = \mathrm{or}(t_L)\,\mathrm{or}(t_R)\,\frac{2}{1+\delta_{t_L=t_R}}, \tag{23}$$

where $\delta_{t_L=t_R} = 1$ if $t_L = t_R$, and $\delta_{t_L=t_R} = 0$ otherwise. If we insert $\mathrm{lab}(t_L) = \mathrm{or}(t_L)L!/2^{L-1}$ and $\mathrm{lab}(t_R) =$

$\mathrm{or}(t_R)R!/2^{R-1}$ into Eq. (22), then we find

$$\mathrm{lab}(t) = \mathrm{or}(t_L)\,\mathrm{or}(t_R)\,\frac{L!\,R!}{2^{n-2}}\binom{n}{L}\frac{1}{1+\delta_{t_L=t_R}} \tag{24}$$

$$= \mathrm{or}(t_L)\,\mathrm{or}(t_R)\,\frac{n!}{2^{n-1}}\,\frac{2}{1+\delta_{t_L=t_R}} = \mathrm{or}(t)\frac{n!}{2^{n-1}}, \tag{25}$$

as desired. □

The proof shows that the ratio of orderings to labelings for an unlabeled topology is independent of the unlabeled topology. Hence, because the number of root configurations of a labeled topology or ordered unlabeled topology depends only on the underlying unlabeled topology, the probability that a labeled topology chosen uniformly at random has ρ root configurations equals the probability that an ordered unlabeled topology chosen uniformly at random has ρ root configurations. We use Lemma 1 to calculate the probability that a labeled topology of size n selected under the uniform distribution has ρ root configurations as the probability that an ordered unlabeled topology of size n selected under the uniform distribution has ρ root configurations.

Proposition 1 *Let R_n be the random variable that represents the number of root configurations in an ordered unlabeled topology of size n selected uniformly at random. (i) We have $R_1 = 0$, and for $n \geq 2$,*

$$R_n \overset{d}{=} (R_{I_n} + 1)(R^*_{n-I_n} + 1), \tag{26}$$

where I_n is distributed over the interval $[1, n-1]$ with Catalan probability $\mathbb{P}[I_n = j] = C_{j-1}C_{n-j-1}/C_{n-1}$, R^*_j is an independent copy of R_j for each $j \in [1, n-1]$, and both R_j and R^*_j are independent of I_j for $j \in [1, n-1]$. Furthermore, (ii) the probability that a random labeled topology of size n selected under the uniform distribution has $c_r = \rho$ root configurations can be calculated as $\mathbb{P}[c_r = \rho] = \mathbb{P}[R_n = \rho]$, where $\mathbb{P}[R_n = \rho]$ has recursive formula

$$\mathbb{P}[R_n = \rho] = \sum_{d \in Div(\rho)} \sum_{j=1}^{n-1} \mathbb{P}[I_n = j]\,\mathbb{P}[R_j = d-1]\,\mathbb{P}\left[R_{n-j} = \frac{\rho}{d} - 1\right], \tag{27}$$

where $Div(\rho)$ denotes the set of positive integers that divide ρ, $\mathbb{P}[I_n = j] = C_{j-1}C_{n-j-1}/C_{n-1}$, and $\mathbb{P}[R_n = 0] = \delta_{n,1}$.

Proof. The recurrence in Eq. (26) follows from Eq. (15). Observe that for a random uniform ordered unlabeled topology t of n leaves, the probability that the left (or right) root subtree of t has size $I_n = j$ is given by $\mathbb{P}[I_n = j] = C_{j-1}C_{n-j-1}/C_{n-1}$, where C_{j-1}, C_{n-j-1}, and C_{n-1} give the numbers of ordered unlabeled topologies of size j, $n-j$, and n, respectively (Section 2.1.2). This establishes (i).

For (ii), Eq. (27) is a direct consequence of Lemma 1 and Eq. (26). □

We now consider the equivalence between uniformly distributed labeled histories and uniformly distributed ordered unlabeled histories.

Lemma 2 *The distribution of the number of root configurations over labeled histories of size n selected uniformly at random matches the distribution of the number of root configurations over ordered unlabeled histories of size n selected uniformly at random.*

Proof. The proof is similar to that of Lemma 1: we show that for each unlabeled history t of size n, we have

$$\frac{\mathrm{or}(t)}{F_{n-1}} = \frac{\mathrm{lab}(t)}{|H_n|}, \tag{28}$$

where $\mathrm{or}(t)$ and $\mathrm{lab}(t)$ are the number of orientations of t and the number of leaf labelings of t, respectively. In other words, we prove that the uniform distribution over the set of ordered unlabeled histories of size n and the uniform distribution over the set of labeled histories of size n both induce the same probability distribution over the set of unlabeled histories of n leaves. The same property has already been shown by Lambert & Stadler [28, p. 116] following a slightly different approach.

Using $F_{n-1} = (n-1)!$ and $|H_n| = n!(n-1)!/2^{n-1}$ from Eqs. (5) and (4), Eq. (28) can be rewritten

$$\mathrm{lab}(t) = \mathrm{or}(t)\frac{n!}{2^{n-1}},$$

which we verify by induction on $|t|$. Let t_L and t_R denote the two root subtrees of t, with sizes $|t_L| = L$ and $|t_R| = R$. Hence, for $n \geq 2$ we have

$$\begin{aligned} \mathrm{lab}(t) &= \mathrm{lab}(t_L)\,\mathrm{lab}(t_R)\binom{n}{L} && (29)\\ \mathrm{or}(t) &= 2\,\mathrm{or}(t_L)\,\mathrm{or}(t_R). && (30)\end{aligned}$$

By setting $\mathrm{lab}(t_L) = \mathrm{or}(t_L)L!/2^{L-1}$ and $\mathrm{lab}(t_R) = \mathrm{or}(t_R)R!/2^{R-1}$ in Eq. (29), we find

$$\begin{aligned} \mathrm{lab}(t) &= \mathrm{or}(t_L)\,\mathrm{or}(t_R)\,\frac{L!\,R!}{2^{n-2}}\binom{n}{L} && (31)\\ &= \mathrm{or}(t_L)\,\mathrm{or}(t_R)\,\frac{2\,n!}{2^{n-1}} = \mathrm{or}(t)\frac{n!}{2^{n-1}}, && (32)\end{aligned}$$

as desired. □

Next, we describe implications of Lemma 2 for Yule–Harding-distributed labeled topologies.

Lemma 3 *The distribution of the number of root configurations over labeled topologies of size n selected according to the Yule–Harding distribution matches the distribution of the number of root configurations over ordered unlabeled histories of size n selected uniformly at random.*

Proof. The equivalence follows from Lemma 2 and the fact that the uniform distribution over labeled histories of size n induces the Yule–Harding distribution on the set of labeled topologies of size n (Section 2.2). □

By Lemma 3, we can calculate the probability that a labeled topology of size n selected under the Yule–Harding distribution has ρ root configurations as the probability that a random uniform ordered unlabeled history of size n has ρ root configurations. In particular, we have the following proposition.

Proposition 2 *Let R_n be the random variable that represents the number of root configurations in an ordered unlabeled history of size n selected uniformly at random. (i) We have $R_1 = 0$, and for $n \geq 2$,*

$$R_n \stackrel{d}{=} (R_{I_n} + 1)(R^*_{n-I_n} + 1), \tag{33}$$

where I_n is uniformly distributed over the interval $[1, n-1]$, R^*_j is an independent copy of R_j for each $j \in [1, n-1]$, and both R_j and R^*_j are independent of I_j for $j \in [1, n-1]$. Furthermore, (ii) the probability that a random labeled topology of size n selected under the Yule–Harding distribution has $c_r = \rho$ root configurations can be calculated as $\mathbb{P}[c_r = \rho] = \mathbb{P}[R_n = \rho]$, where $\mathbb{P}[R_n = \rho]$ has recursive formula

$$\mathbb{P}[R_n = \rho] = \sum_{d \in Div(\rho)} \sum_{j=1}^{n-1} \mathbb{P}[I_n = j]\,\mathbb{P}[R_j = d-1]\,\mathbb{P}\left[R_{n-j} = \frac{\rho}{d} - 1\right], \tag{34}$$

where $Div(\rho)$ denotes the set of positive integers that divide ρ, $\mathbb{P}[I_n = j] = \frac{1}{n-1}$, and $\mathbb{P}[R_n = 0] = \delta_{n,1}$.

Proof. The formula in Eq. (33) follows directly from Eq. (15) when we observe that, for a random uniform ordered unlabeled history t of n leaves, the probability that the left (or right) root subtree of t has size $I_n = j$ is

$$\mathbb{P}[I_n = j] = \frac{F_{j-1}F_{n-j-1}\binom{n-2}{j-1}}{F_{n-1}} = \frac{1}{n-1}.$$

Eq. (34) is a direct consequence of Lemma 3 and Eq. (33). □

3.2 Equivalences with antichains of pruned binary trees

To use results of Wagner [45] to obtain probability distributions for root configurations, we must translate between root configurations for labeled topologies and non-empty antichains for pruned binary trees.

A pruned binary tree is an ordered unlabeled topology in which the external branches—those terminating in a leaf—have been removed. If a node of the initial ordered unlabeled topology has one incident external branch, then pruning renders the node of the pruned binary tree with only one immediate descendant; a node with two incident external branches is pruned to possess no immediate descendants. To illustrate the pruning operation, consider the ordered unlabeled topology depicted on the left of Fig. 2A and assign arbitrary labels to all its nodes, as in Fig. 1A. The leaf labels of the pruned binary tree resulting from this process can be described by the Newick format $((g,h),i)$. Note that pruned binary trees have their left–right orientation induced by the overlying ordered unlabeled topology.

If t is an ordered unlabeled topology of size n and $\tilde{t}$ is its associated pruned binary tree of $n-1$ nodes, then we can consider $\tilde{t}$ as the Hasse diagram of a partially ordered set with ground set given by the nodes of $\tilde{t}$—the *internal* nodes of t—and order relation determined by the descendant–ancestor relationship in $\tilde{t}$. An antichain of $\tilde{t}$ is a subset of its nodes such that no two elements in the subset are comparable by the order relation. For instance, the two-element antichains of pruned binary tree $((g,h),i)$ in Fig. 1A are $\{g,h\}, \{g,i\}, \{h,i\}$, and $\{j,i\}$.

The non-empty antichains of the pruned binary tree $\tilde{t}$ bijectively correspond to the root configurations of the overlying ordered unlabeled topology t: omitting leaves from a root configuration of t yields an antichain of $\tilde{t}$, and adding leaves to an antichain of $\tilde{t}$ so that each leaf of t is either represented or has one of its ancestral nodes represented yields a root configuration of t.

For instance, consider the set in Eq. (10) of the root configurations of the ordered unlabeled topology in Fig. 1A. By omitting leaves from each configuration, we obtain the antichains of $\tilde{t}$:

$$\{\{j,i\}, \{j\}, \{g,h,i\}, \{g,h\}, \{h,i\}, \{h\}, \{g,i\}, \{g\}, \{i\}, \emptyset\}.$$

We make a substitution of the empty antichain $\emptyset$ that emerges from the root configuration consisting of all the leaves by the antichain $\{k\}$ consisting only of the root of $\tilde{t}$; we have then bijectively paired all root configurations of t and all non-empty antichains of $\tilde{t}$. Using this correspondence, we have the next result.

Lemma 4 *The distribution of the number of root configurations over labeled topologies of size n selected uniformly at random matches the distribution of the number of non-empty antichains over the set of $(n-1)$-node pruned binary trees selected uniformly at random.*

Proof. By Lemma 1, the number of root configurations has the same distribution when considered over uniformly distributed labeled topologies of size n or over uniformly distributed ordered unlabeled topologies of size n. By the correspondence between antichains of pruned binary trees with $n-1$ nodes and root configurations of associated ordered unlabeled topologies of size n, the distribution of the number of root configurations over uniformly distributed ordered unlabeled topologies of size n matches the distribution of the number of non-empty antichains over uniformly distributed pruned binary trees with $n-1$ nodes. □

4 Root configurations under the uniform distribution on labeled topologies

Disanto & Rosenberg [14] determined the mean and variance of the number of root configurations for uniformly distributed labeled topologies of size n (Section 2.4.4). In this section, we use the correspondence with antichains given in Section 3.2 to show that the logarithm of the number of root configurations for uniformly distributed labeled topologies of size n, suitably rescaled, converges to a normal distribution.

Wagner [45, Section 2.3.2] studied the number $a(t)$ of non-empty antichains of a randomly selected pruned binary tree t of given size. For a pruned binary tree of n nodes selected uniformly at random, he considered $\log a(t)$, showing that $\big(\log a - \mathbb{E}_n[\log a]\big)/\sqrt{\mathbb{V}_n[\log a]}$ converges to a standard normal distribution as $n \to \infty$, where $\mathbb{E}_n[\log a] \sim \mu n$ and $\mathbb{V}_n[\log a] \sim \sigma^2 n$, with constants $(\mu, \sigma^2) \approx (0.272, 0.034)$.

By Lemma 4, Wagner's variable $\log a$ asymptotically has the same distribution as the variable $\log c_r$ considered over uniformly distributed labeled topologies of size $n+1$. We thus have the following result.

Proposition 3 *The logarithm of the number of root configurations in a labeled topology of size n selected uniformly at random, rescaled as $\big(\log c_r - \mathbb{E}_n[\log c_r]\big)/\sqrt{\mathbb{V}_n[\log c_r]}$, converges to a standard normal distribution, where $\mathbb{E}_n[\log c_r] \sim \mu n$ and $\mathbb{V}_n[\log c_r] \sim \sigma^2 n$, $(\mu, \sigma^2) \approx (0.272, 0.034)$.*

The result gives an asymptotic lognormal distribution for the number of root configurations of a labeled topology of size n selected uniformly at random. Although we do not expect $e^{\mathbb{E}_n[\log c_r]}$ and $e^{\sigma_n[\log c_r]}$ to agree with $\mathbb{E}_n[c_r]$ and $\sigma_n[c_r]$, for the mean we see that in the $n \to \infty$ limit, $e^{\mathbb{E}_n[\log c_r]} \approx e^{0.272n} \approx 1.313^n$, numerically close to the exponential growth of $\mathbb{E}_n[c_r]$, or $(4/3)^n$ (Eq. (16)). For, the standard deviation $e^{\sigma_n[\log c_r]} \approx e^{\sqrt{0.034}n} \approx 1.202^n$ is not as close to the exponential growth of $\sigma_n[c_r]$ from Eq. (18), which gives $\Big[2/\sqrt{7(8\sqrt{2}-11)}\Big]^n \approx 1.350^n$.

For fixed n, we can compute the exact distribution of c_r and $\log c_r$ under a uniform distribution across labeled topologies of size n, as described in Proposition 1ii. Fig. 5 shows the cumulative distribution $\mathbb{P}\big[\log c_r \leq \mathbb{E}[\log c_r] + y\sigma[\log c_r]\big]$ as a function of y, when labeled topologies are selected uniformly at random among the 2.13×10^{14} labeled topologies with 15 leaves. To obtain the distribution, we can count root configurations for arbitrary labelings of each of the 4850 unlabeled topologies with 15 leaves, and then count labelings for each unlabeled topology [39, p. 47]. Already for small tree size, the figure shows that the exact cumulative distribution is close to the cumulative distribution of a Gaussian random variable with mean 0 and variance 1.

5 Root configurations under the Yule–Harding distribution on labeled topologies

We next study distributional properties of the number of root configurations for labeled topologies selected under the Yule–Harding probability model. Section 2.2 noted that this model assigns higher probability to trees with a high degree of balance compared to that assigned by the uniform model; Section 2.4.4 noted that balanced trees have high numbers of root configurations relative to unbalanced trees. We therefore find that the mean number of root configurations for labeled topologies of size n grows exponentially faster under the Yule–Harding model than under the uniform model. The variance of the number of root configurations also has faster growth.

Note that in the main results of the section—Propositions 4, 6, and 7—expectations $\mathbb{E}_n$ and variances $\mathbb{V}_n$ are taken with respect to the Yule–Harding distribution.

5.1 Lognormal distribution of the number of root configurations

We begin the analysis of the number of root configurations under the Yule–Harding distribution by showing that the logarithm of the number of root configurations of a Yule–Harding random labeled topology of size n, when suitably rescaled, converges to a standard normal distribution.

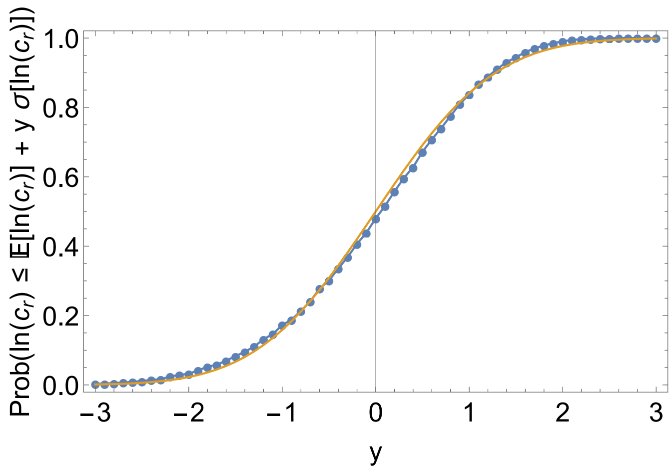

Figure 5: Cumulative distribution of the natural logarithm of the number of root configurations for uniformly distributed labeled topologies of size $n = 15$ (dotted line). Each dot has its abscissa determined by a value of y ranging in the interval $y \in [-3, 3]$ in steps of 0.1. Given y, the quantity plotted is the probability that a labeled topology with $n = 15$ chosen uniformly at random has a number of root configurations less than or equal to $\exp\left(\mathbb{E}[\log c_r] + y\sigma[\log c_r]\right)$, where $\mathbb{E}[\log c_r]$ and $\sigma[\log c_r]$ are respectively the mean and standard deviation of the logarithm of the number of root configurations for uniformly distributed labeled topologies with $n = 15$ leaves (Proposition 3). The solid line is the cumulative distribution of a Gaussian random variable with mean 0 and variance 1.

The results in this section are obtained by considering root configurations over ordered unlabeled histories of given size selected under the uniform distribution. Owing to Lemma 3, we can demonstrate that the number of root configurations in a Yule–Harding random labeled topology of size n asymptotically follows a lognormal distribution by showing that the number of root configurations is asymptotically lognormally distributed when considered over the set of uniformly distributed ordered unlabeled histories of n leaves. We use a result of Wagner [45] for additive tree parameters of ordered unlabeled histories. We first must verify a technical condition for the mean of the random variable $\log(1 + 1/c_r)$, considered over uniformly distributed ordered unlabeled histories. This verification proceeds by considering cherry nodes [31], internal nodes whose two immediate descendant nodes are leaves.

Lemma 5 *For uniformly distributed ordered unlabeled histories of size n, the mean value $\mathbb{E}_n\left[\log(1 + 1/c_r)\right]$ of the random variable $\log(1 + 1/c_r)$ converges to 0 exponentially fast as n increases. In particular,*

$$\mathbb{E}_n\left[\log\left(1 + \frac{1}{c_r}\right)\right] = \mathcal{O}(0.9^n). \tag{35}$$

Proof. To show that $\mathbb{E}_n\left[\log(1 + 1/c_r)\right]$ has exponential growth $\mathcal{O}(0.9^n)$ for an ordered unlabeled history t of size n selected uniformly at random, we consider the mean value $\mathbb{E}_n[2^{-\mathrm{ch}}]$ of the random variable $2^{-\mathrm{ch}}$—where ch is the number of cherries in t. We claim that

$$\mathbb{E}_n[2^{-\mathrm{ch}}] = \mathcal{O}(0.9^n). \tag{36}$$

For a tree t with $|t| \geq 3$, $c_r(t) \geq 2^{\mathrm{ch}(t)}$, as each cherry node generates a pair of ancestral configurations: the configuration corresponding to the node, and the configuration corresponding to its pair of leaves. At the root node, a root configuration can be obtained by choosing ancestral configurations at each of the cherry nodes and augmenting the configuration with leaves that do not descend from cherry nodes.

Noting $\log(1 + x) \leq x$ for $x > 0$, for each ordered unlabeled history t with size $|t| \geq 3$, we have

$$\log\left[1 + \frac{1}{c_r(t)}\right] \leq \frac{1}{c_r(t)} \leq 2^{-\mathrm{ch}(t)}.$$

By taking expectations, we see that Eq. (36) implies Eq. (35):

$$\mathbb{E}_n \left[\log \left(1 + \frac{1}{c_r} \right) \right] \leq \mathbb{E}_n[2^{-\mathrm{ch}}].$$

It remains to verify Eq. (36). In their Theorem 2, Disanto & Wiehe [18] studied the generating function $F(x, z)$ counting the number of unlabeled histories t of size n with a given number of cherries, where each unlabeled history t is weighted by its probability $2^{n-1-\mathrm{ch(t)}}/(n-1)!$ under the Yule–Harding distribution:

$$F(x, z) = \sum_t \frac{2^{n-1-\mathrm{ch(t)}}}{(n-1)!} x^{\mathrm{ch}(t)} z^n.$$

The sum proceeds over unlabeled histories ("ranked trees" in Disanto & Wiehe [18]). The coefficient of $x^h z^n$ in $F(x, z)$ gives the probability of h cherries in unlabeled histories of size n under the Yule–Harding distribution, or equivalently, the probability of h cherries in ordered unlabeled histories of size n selected uniformly at random. Hence, the expectation $\mathbb{E}_n[2^{-\mathrm{ch}}]$ is obtained from the coefficient of z^n in $F(\frac{1}{2}, z)$. From Disanto & Wiehe [18],

$$F\left(\frac{1}{2}, z\right) = f(z) = \frac{ze^{z\sqrt{2}} - z}{(\sqrt{2}-2)e^{z\sqrt{2}} + 2 + \sqrt{2}}.$$

By Theorem IV.7 of Flajolet & Sedgewick [23] (see also Section 2.3), $\mathbb{E}_n[2^{-\mathrm{ch}}]$ grows exponentially like $[z^n]f(z) \bowtie \alpha^{-n}$, where α is the dominant singularity of $f(z)$. The value of α is the solution of smallest modulus of the equation $(\sqrt{2}-2)e^{z\sqrt{2}} + 2 + \sqrt{2} = 0$, whose left-hand side is the denominator of $f(z)$. Because

$$\alpha = \frac{1}{\sqrt{2}} \log \left(\frac{2+\sqrt{2}}{2-\sqrt{2}} \right) = \frac{\sqrt{2}\log(3+2\sqrt{2})}{2} \approx 1.246,$$

$\alpha^{-1} \approx 0.802$ and thus, conservatively, $\mathbb{E}_n[2^{-\mathrm{ch}}] = \mathcal{O}(0.9^n)$. Hence, $\mathbb{E}_n[\log(1 + \frac{1}{c_r})]$ also decays to 0 as $\mathcal{O}(0.9^n)$. $\square$

Considering as in Section 2.5 the additive tree parameter $F(t) = \log[c_r(t) + 1]$, by Lemma 5 we have demonstrated that the associated toll function $f(t) = \log[1 + 1/c_r(t)]$ satisfies

$$\frac{\sum_t f(t)}{F_{n-1}} = \mathbb{E}_n \left[\log \left(1 + \frac{1}{c_r} \right) \right] = \mathcal{O}(0.9^n), \tag{37}$$

where the sum proceeds over all $(n-1)!$ ordered unlabeled histories t of size n (Eq. (5)). Eq. (37), together with the fact that $f(t)$ is bounded because $c_r(t) \geq 1$ for $|t| \geq 2$, show that the hypotheses of Theorem 4.2 of Wagner [45] are satisfied. By applying the theorem, we can conclude that for an ordered unlabeled history t of size n selected uniformly at random, the standardized version of the random variable $F(t) = \log[c_r(t) + 1]$ converges asymptotically to a normal distribution with mean 0 and variance 1. By the same theorem, the mean and variance of $F(t) = \log[c_r(t) + 1]$ grow respectively like μn and $\sigma^2 n$, for two constants

$$\mu = \sum_t \frac{2f(t)}{(|t|+1)!} \approx 0.351, \tag{38}$$

$$\begin{aligned} \sigma^2 = {} & \sum_t \frac{2f(t)[2F(t) - f(t)]}{(|t|+1)!} - \mu^2 + \sum_{t_1} \sum_{t_2} \frac{4f(t_1)f(t_2)}{(|t_1|+1)!(|t_2|+1)!} \\ & \times \left[\frac{(|t_1|-1)(|t_2|-1)}{|t_1|+|t_2|-1} - |t_1| - |t_2| + 2 + \frac{(|t_1|-1)(|t_2|-1)}{(|t_1|+|t_2|)(|t_1|+|t_2|+1)} \right. \\ & \left. + \frac{(|t_1|-1)^2(|t_2|-1)^2}{(|t_1|+|t_2|-1)(|t_1|+|t_2|)(|t_1|+|t_2|+1)} \right] \approx 0.008. \end{aligned} \tag{39}$$

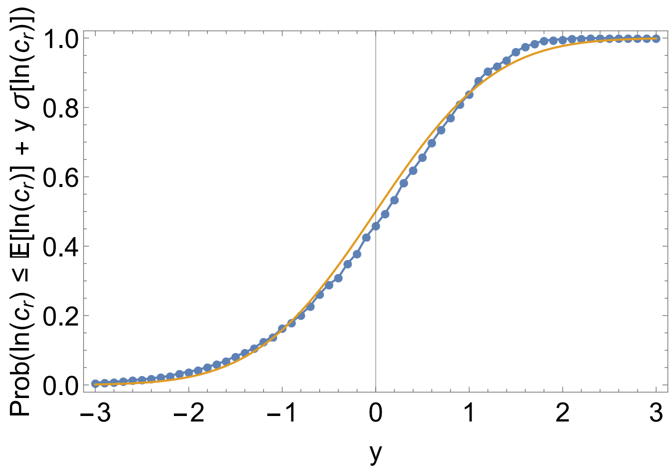

Figure 6: Cumulative distribution of the natural logarithm of the number of root configurations for labeled topologies of size $n = 15$ considered under the Yule–Harding distribution (dotted line). Each dot has its abscissa determined by a value of y ranging in the interval $y \in [-3, 3]$ in steps of 0.1. Given y, the quantity plotted is the probability that a labeled topology with $n = 15$ chosen at random under the Yule–Harding distribution has a number of root configurations less than or equal to $\exp(\mathbb{E}[\log c_r] + y\sigma[\log c_r])$, where $\mathbb{E}[\log c_r]$ and $\sigma[\log c_r]$ are respectively the mean and the standard deviation of the logarithm of the number of root configurations for Yule–Harding distributed labeled topologies of $n = 15$ leaves (Proposition 4). The solid line is the cumulative distribution of a Gaussian random variable with mean 0 and variance 1.

Note that the sums in Eqs. (38) and (39) are defined over all ordered unlabeled histories, but that the approximations have been calculated by disregarding histories of size strictly larger than 15 and 12 in the sums for μ and σ^2, respectively. The equivalence of Lemma 3 between the distribution of the number of root configurations over uniformly distributed ordered unlabeled histories and the distribution of the number of root configurations over Yule–Harding distributed labeled topologies, coupled with the fact that the difference $\log(c_r+1) - \log c_r = \log(1+1/c_r)$ is small, finally yields the following proposition.

Proposition 4 *The logarithm of the number of root configurations in a labeled topology of size n selected under the Yule–Harding distribution, rescaled as $(\log c_r - \mathbb{E}_n[\log c_r])/\sqrt{\mathbb{V}_n[\log c_r]}$, converges to a standard normal distribution, where $\mathbb{E}_n[\log c_r] \sim \mu n$ and $\mathbb{V}_n[\log c_r] \sim \sigma^2 n$ for $(\mu, \sigma^2) \approx (0.351, 0.008)$.*

For fixed n, we can compute the exact distribution of c_r (and $\log c_r$) under the Yule–Harding distribution across all labeled topologies of size n as in Proposition 2ii. Similarly to the computations in Fig. 5, we can weight the counts of root configurations for unlabeled topologies by their Yule–Harding probabilities [39, p. 47]. Fig. 6 shows the cumulative distribution $\mathbb{P}[\log c_r \leq \mathbb{E}[\log c_r] + y\sigma[\log c_r]]$ plotted as a function of y, when labeled topologies of size $n = 15$ are selected under the Yule–Harding distribution. The distribution is close to the cumulative distribution of a Gaussian random variable with mean 0 and variance 1.

5.2 Mean number of root configurations

In Section 5.1, we have analyzed distributional properties of the logarithm of the number of root configurations considered over labeled topologies of given size selected under the Yule–Harding distribution. In this section, we study the mean number of root configurations under the Yule–Harding distribution.

From Lemma 3, the mean number of root configurations in a random labeled topology of size n selected under the Yule–Harding distribution is also the mean number of root configurations in a uniform random ordered unlabeled history of n leaves. To calculate this mean, we use the distributional recurrence in Proposition 2 for the variable R_n and, by applying generating functions and singularity analysis, we obtain the following result.

Proposition 5 *The mean number of root configurations in an ordered unlabeled history of size n selected uniformly at random satisfies the asymptotic relation $\mathbb{E}[R_n] \sim k_e^n$, where $k_e = 1/(1 - e^{-2\pi\sqrt{3}/9})$.*

Proof. Set $e_n \equiv \mathbb{E}[R_n]$. Then $\mathbb{E}[R_{I_n}\, R^*_{n-I_n}] = \sum_{j=1}^{n-1} \mathbb{P}[I_n = j]\, \mathbb{E}[R_j\, R^*_{n-j}] = \frac{1}{n-1} \sum_{j=1}^{n-1} \mathbb{E}[R_j]\, \mathbb{E}[R^*_{n-j}]$. Proposition 2 yields for $n \geq 2$ the recurrence

$$e_n = 1 + \frac{1}{n-1} \sum_{j=1}^{n-1} e_j e_{n-j} + \frac{2}{n-1} \sum_{j=1}^{n-1} e_j, \tag{40}$$

with initial condition $e_1 = 0$.

Defining the generating function

$$E(z) \equiv \sum_{n=1}^{\infty} e_n z^n = z^2 + 2z^3 + \frac{10}{3} z^4 + \frac{31}{6} z^5 + \ldots, \tag{41}$$

the recurrence in Eq. (40) translates into the Riccati differential equation

$$zE'(z) = E(z)^2 + \frac{1+z}{1-z}\, E(z) + \frac{z^2}{(1-z)^2}, \tag{42}$$

with initial condition $E(0) = 0$. To obtain the differential equation, we have multiplied both sides of Eq. (40) by $(n-1)z^n$, summed for $n \geq 1$, and then used the facts that $\sum_{n=1}^{\infty} (n-1) e_n z^n = zE'(z) - E(z)$, $\sum_{n=1}^{\infty} (n-1) z^n = z^2\, [1/(1-z)]' = z^2/(1-z)^2$, $\sum_{n=1}^{\infty} (\sum_{j=1}^{n-1} e_j e_{n-j}) z^n = E(z)^2$, and $\sum_{n=1}^{\infty} (\sum_{j=1}^{n-1} e_j) z^n = E(z)[1/(1-z) - 1]$.

Solving the differential equation yields

$$E(z) = \frac{2z \sin\left(\frac{\sqrt{3}}{2} \log(1-z)\right)}{(z-1)\left[\sqrt{3} \cos\left(\frac{\sqrt{3}}{2} \log(1-z)\right) + \sin\left(\frac{\sqrt{3}}{2} \log(1-z)\right)\right]}. \tag{43}$$

$E(z)$ has infinitely many singularities. The singularity of $E(z)$ with smallest modulus occurs at $z = \alpha \equiv 1 - e^{-2\pi\sqrt{3}/9} \approx 0.702$. The singularity of smallest modulus is obtained by setting to 0 the factor

$$\sqrt{3} \cos\left[\frac{\sqrt{3}}{2} \log(1-z)\right] + \sin\left[\frac{\sqrt{3}}{2} \log(1-z)\right] \tag{44}$$

appearing in the denominator of Eq. (43). The expansion of $E(z)$ at its dominant singularity $z = \alpha$ looks like

$$E(z) \overset{z \to \alpha}{\sim} \frac{1}{1 - \frac{z}{\alpha}},$$

which can be obtained by plugging the Taylor expansion $-\sqrt{3} e^{+2\pi\sqrt{3}/9} (z - \alpha)$ of the factor (44) in the denominator of Eq. (43). By Theorem VI.4 of Flajolet & Sedgewick [23] (see also Section 2.3), we finally obtain

$$[z^n] E(z) \sim [z^n] \left(\frac{1}{1 - \frac{z}{\alpha}}\right) = \alpha^{-n},$$

as $n \to \infty$. □

The next proposition follows immediately from Proposition 5 and Lemma 3.

Proposition 6 *The mean number of root configurations in a labeled topology of size n selected at random under the Yule–Harding distribution has asymptotic growth $\mathbb{E}_n[c_r] \sim k_e^n$, where $k_e = 1/(1 - e^{-2\pi\sqrt{3}/9}) \approx 1.42538682$. Furthermore, the mean total number of configurations has asymptotic growth $\mathbb{E}_n[c] \bowtie \mathbb{E}_n[c_r]$.*

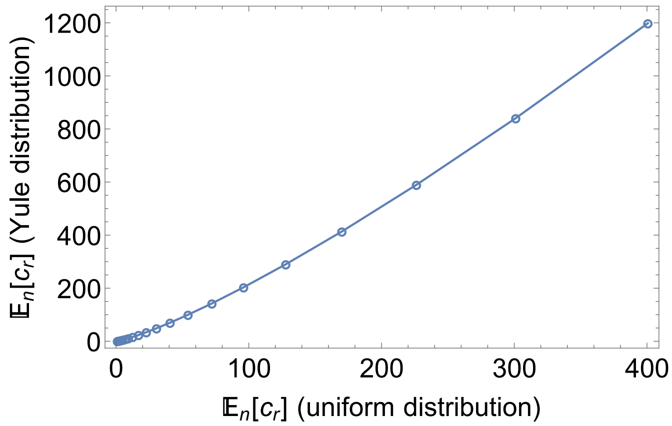

Figure 7: Mean number of root configurations of labeled topologies of size n under the Yule–Harding and uniform distributions, for $2 \leq n \leq 20$. Values for the uniform distribution are computed from the power series expansion of Eq. (33) of Disanto & Rosenberg [14]; values for Yule–Harding are computed from the power series expansion of Eq. (43).

For small tree size ($n \leq 20$), Fig. 7 plots the mean number of root configurations for a random tree of size n selected under the Yule–Harding distribution as a function of the corresponding mean under the uniform distribution. The plot provides a numerical visualization of the similar behavior of the numbers of root configurations under the Yule–Harding and uniform distributions. The mean is greater for the Yule–Harding distribution, but the two quantities are highly correlated, with Pearson's correlation coefficient approximately 0.995.

5.3 Variance of the number of root configurations

In this section, we analyze the asymptotic growth of the variance of the number of root configurations under the Yule–Harding distribution. In particular, by using Lemma 3, we study the variance of the number of root configurations in a uniform random ordered unlabeled history of size n.

Following Section 5.2 and squaring Eq. (33), we obtain a recurrence for $s_n \equiv \mathbb{E}[R_n^2]$. For $n \geq 2$,

$$s_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} s_j\, s_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} s_j + \frac{4}{n-1}\sum_{j=1}^{n-1} s_j\, e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j\, e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j, \tag{45}$$

with initial condition $s_1 = 0$.

Starting from this recurrence, a symbolic calculation similar to that used to derive Eq. (42) shows that the generating function $S(z) \equiv \sum_{n=1}^{\infty} s_n z^n = z^2 + 4z^3 + \frac{34}{3}z^4 + \frac{55}{2}z^5 \ldots$ satisfies the Riccati differential equation

$$z\, S'(z) = S(z)^2 - S(z)\left[\frac{1+z}{z-1} - 4E(z)\right] + \frac{[z - 2(z-1)E(z)]^2}{(z-1)^2}. \tag{46}$$

This equation can be written

$$S'(z) = g_2(z)\, S(z)^2 + g_1(z)\, S(z) + g_0(z) \tag{47}$$

by setting

$$\Big(g_2(z), g_1(z), g_0(z)\Big) \equiv \left(\frac{1}{z}, \left(4E(z) - \frac{1+z}{z-1}\right)\frac{1}{z}, \frac{[z-2(z-1)E(z)]^2}{z(z-1)^2}\right).$$

By substituting $U(z) \equiv \exp[\int_0^z S(x)/(-x)\, dx]$, we obtain $S(z) = -zU'(z)/U(z)$, and Eq. (47) can be rewritten as a second-order linear differential equation equation

$$U''(z) - \left(g_1(z) + \frac{g_2'(z)}{g_2(z)}\right)U'(z) + g_2(z)\, g_0(z)\, U(z) = 0. \tag{48}$$

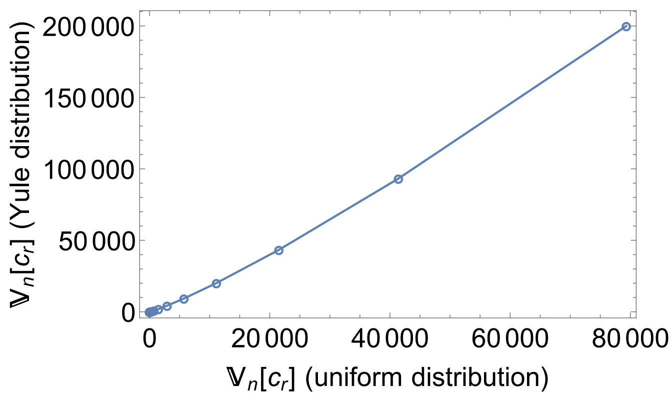

Figure 8: Variance of the number of root configurations of labeled topologies of size n under the Yule–Harding and uniform distributions, for $2 \leq n \leq 20$. Values for the uniform distribution are computed from the power series expansion of Eq. (39) of Disanto & Rosenberg [14]; values for Yule–Harding are computed from Eqs. (45) and (40).

The coefficients of Eq. (48) are analytic functions for $|z| < 0.702$, with a removable singularity at $z = 0$ as the expansion (41) of $E(z)$ starts with a quadratic non-zero term. Using existence results for the solutions of second-order ordinary differential equations, $U(z)$ must be analytic for $|z| < 0.702$, the constant being the radius of convergence of $E(z)$ as determined in the proof of Proposition 5. Therefore, also $U'(z)$ is analytic for $|z| < 0.702$, and thus $S(z)$ is a meromorphic function on this domain, being a quotient of two analytic functions. To analyze the singularities of a meromorphic function, one must locate the possible roots of its denominator function. In our case, the set of singularities of $S(z)$ consists of the roots of $U(z)$. In particular, by studying in the Appendix the function $U(z)$ in $\mathcal{B} \equiv \{z \in \mathbb{C} \ : \ |z| \leq \frac{1}{2}\}$, we find that $S(z)$ has a unique dominant singularity $\alpha \approx 0.4889986317$, the unique and simple root of $U(z)$ within $\mathcal{B}$ (Proposition 8).

As a consequence, we can write $U(z) = (z - \alpha)\tilde{U}(z)$, with $\tilde{U}(\alpha) \neq 0$ and $U'(\alpha) = (-\alpha)\tilde{U}(\alpha) \neq 0$. Therefore, for $z \to \alpha$ the generating function $S(z)$ admits the expansion

$$S(z) = \frac{-zU'(z)}{U(z)} \overset{z \to \alpha}{\sim} \frac{(-\alpha)[U'(\alpha) + U''(\alpha)(z-\alpha) + \ldots]}{U(\alpha) + U'(\alpha)(z-\alpha) + \ldots} \overset{z \to \alpha}{\sim} \frac{(-\alpha)U'(\alpha)}{U'(\alpha)(z-\alpha)} = \frac{-\alpha}{z-\alpha} = \frac{1}{1-\frac{z}{\alpha}}.$$

From Theorem VI.4 of Flajolet & Sedgewick [23] (see also Section 2.3), we can thus recover the asymptotic growth of the associated coefficients

$$\mathbb{E}[R_n^2] = [z^n]S(z) \sim [z^n]\left(\frac{1}{1-\frac{z}{\alpha}}\right) = \alpha^{-n}, \tag{49}$$

and hence derive the asymptotic growth of the variance $\mathbb{V}[R_n]$. In particular, we have the following result.

Proposition 7 *The variance of the number of root configurations in a labeled topology of size n selected at random under the Yule–Harding distribution has asymptotic growth $\mathbb{V}_n[c_r] \sim k_v^n$, where $k_v \approx 2.0449954971$. Furthermore, the variance of the total number of configurations has asymptotic growth $\mathbb{V}_n[c] \bowtie \mathbb{V}_n[c_r]$.*

Proof. For uniformly distributed ordered unlabeled histories of size n, Eq. (49) yields $\mathbb{E}[R_n^2] \sim k_v^n$, $k_v \equiv 1/\alpha \approx 2.0449954971$. From Proposition 5, $\mathbb{E}[R_n]^2 \sim (k_e^2)^n$, with $k_e^2 \approx 2.03$. Because $k_v > k_e^2$, as $n \to \infty$ we obtain

$$\mathbb{V}[R_n] = \mathbb{E}[R_n^2] - \mathbb{E}[R_n]^2 \sim k_v^n.$$

By Lemma 3, the variance of the variable R_n is the variance of the number of root configurations considered over labeled topologies of n leaves selected under the Yule–Harding distribution. $\square$

Table 1: Distributional properties of the number of root and total configurations.

Results		Uniform model		Yule–Harding model	
Root configurations	Mean	$\mathbb{E}_n[c_r] \sim 1.225 \cdot 1.333^n$	Eq. (16)	$\mathbb{E}_n[c_r] \sim 1.425^n$	Proposition 6
	Variance	$\mathbb{V}_n[c_r] \sim 1.405 \cdot 1.822^n$	Eq. (18)	$\mathbb{V}_n[c_r] \sim 2.045^n$	Proposition 7
	Lognormal distribution	$\mathbb{E}_n[\log c_r] \sim 0.272 \cdot n$	Proposition 3	$\mathbb{E}_n[\log c_r] \sim 0.351 \cdot n$	Proposition 4
		$\mathbb{V}_n[\log c_r] \sim 0.034 \cdot n$	Proposition 3	$\mathbb{V}_n[\log c_r] \sim 0.008 \cdot n$	Proposition 4
Total configurations	Mean	$\mathbb{E}_n[c] \bowtie 1.333^n$	Eq. (17)	$\mathbb{E}_n[c] \bowtie 1.425^n$	Proposition 6
	Variance	$\mathbb{V}_n[c] \bowtie 1.822^n$	Eq. (19)	$\mathbb{V}_n[c] \bowtie 2.045^n$	Proposition 7

As we did for the mean, we numerically visualize the similarity in variance of the number of root configurations for trees of size n selected at random under the Yule–Harding and uniform distributions. For small tree size ($n \leq 20$), we plot in Fig. 8 the variance of the number of root configurations for a random tree of size n selected under the Yule–Harding distribution as a function of the variance of the number of root configurations for a random uniform tree of the same size. As was true of the mean, the Yule–Harding and uniform distributions on labeled topologies give correlated variances (correlation coefficient 0.997).

6 Discussion

Considering gene trees and species trees with a matching labeled topology $G = S = t$, we have studied distributional properties of the number c_r of root ancestral configurations for labeled topologies t of fixed size under two probability models, the uniform model and the Yule–Harding model (Table 1). We have made use of techniques of analytic combinatorics, relying on equivalences across tree types (Section 3), and making particular use of results of Wagner [45] on distributional properties of additive tree parameters for several families of trees.

Extending results of Disanto & Rosenberg [14], for the uniform model we have shown that the logarithm of the number of root configurations, when standardized, converges asymptotically to a standard normal distribution (Proposition 3). Under the Yule–Harding distribution, as is the case for uniformly distributed labeled topologies, the logarithm of the number of root configurations, when standardized, converges to a standard normal distribution (Proposition 4). The study produces the first results on asymptotic distributions under the uniform or Yule–Harding models for ancestral configurations, and further, for any of the recently studied combinatorial quantities that require consideration of both gene trees and species trees—ancestral configurations [14; 46], coalescent histories [2; 10; 11; 12; 13; 26; 33; 34; 35; 42], compact coalescent histories [15; 47], deep coalescence costs [29; 30; 41; 43; 44], history classes [36], non-equivalent ancestral configurations [16; 46], and ranked histories [8; 9; 37].

We have also determined the asymptotic growth of the mean and the variance of the number of root configurations, finding that under the Yule–Harding model, $\mathbb{E}_n[c_r] \sim 1.425^n$ (Proposition 6) and $\mathbb{V}_n[c_r] \sim 2.045^n$ (Proposition 7). As $\mathbb{E}_n[c] \bowtie \mathbb{E}_n[c_r]$ and $\mathbb{V}_n[c] \bowtie \mathbb{V}_n[c_r]$, we also recover the exponential growth rate of the mean and the variance of the total number of configurations under the Yule–Harding model. These results were obtained by use of recursions to obtain Riccati differential equations for generating functions (Eqs. (42) and (46)). For the case of the mean, the Riccati equation was solvable (Eq. (43)); for the variance, although the equation was not solvable, the asymptotic growth was nevertheless possible to obtain. Our method introduced for this case has potential for broader application, as many problems involving various types of trees and other combinatorial structures can lead to related Riccati equations [5; 22; 24].

Both the mean and the variance across labeled topologies of the number of ancestral configurations are empirically highly correlated between the uniform and Yule–Harding models (Figs. 7 and 8). Alongside the results of Disanto & Rosenberg [14] for the uniform case, the larger values for Yule–Harding (Table 1) suggest a role for tree balance in predicting the number of root configurations. By considering a representative labeling for each unlabeled topology of size $n = 15$, in Fig. 9 we plot on a logarithmic scale the number of root configurations

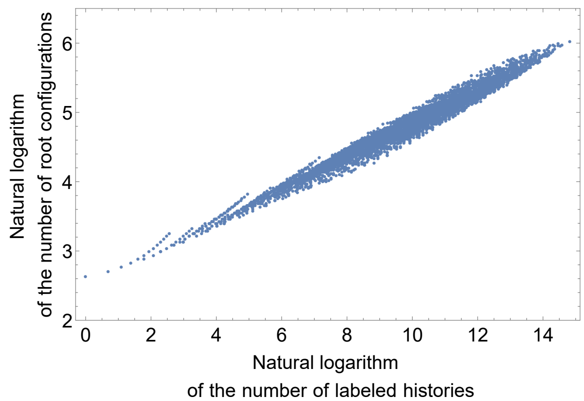

Figure 9: Natural logarithm of the number of root configurations and natural logarithm of the number of labeled histories for a representative labeling of each unlabeled topology of size $n = 15$. The number of points plotted is 4850, the number of unlabeled topologies with $n = 15$ leaves. The Pearson correlation is approximately 0.987 (0.784 without log scaling).

as a function of the number of labeled histories, the latter calculated with Eq. (6). The numerical illustration in the figure shows that empirically, the two quantities are correlated: highly balanced labeled topologies—which tend to have larger numbers of labeled histories (Section 2.2)—in general have larger numbers of root configurations.

In particular, the largest number of root configurations is possessed by the balanced labeled topology depicted in Fig. 4C, which also has the largest number of labeled histories, 2745600. The trend in this example is confirmed by our asymptotic results. Under the Yule–Harding probability model, which gives more weight to balanced labeled topologies than does the uniform model, the mean number of root configurations and the mean total number of configurations grow exponentially faster than under the uniform distribution (Table 1). This differing behavior also accords with the proof of Disanto & Rosenberg [14] that balanced and caterpillar trees respectively possess the largest and smallest numbers of root configurations for fixed tree size (Section 2.4.3).

Several directions and extensions naturally arise from our work. First, we focused on root rather than total configurations; although some results for total configurations follow quickly (Table 1), we did not consider total configurations in detail. Second, we assumed that the gene tree and species tree had the same labeled topology, and we did not study nonmatching gene trees and species trees. The nonmatching case merits further analysis, as a nonmatching gene tree labeled topology can have more root and total configurations than the topology that matches the species tree [14]. Third, ancestral configurations can be considered up to an equivalence relationship that accounts for symmetries in gene trees [46]. The resulting equivalence classes—the nonequivalent ancestral configurations—are used for calculating probabilities of gene trees in STELLS [46], with computational complexity that depends on the number of these classes. Some investigation of this number has been carried out by Disanto & Rosenberg [16] for uniformly distributed matching gene trees and species trees. It would be of interest to see whether the techniques we have used could derive distributional properties of the number of nonequivalent ancestral configurations under the uniform and Yule–Harding probability models.

Appendix. The function $U(z)$ has a unique and simple root of smallest modulus

In this appendix, we prove that the function $U(z) \equiv \sum_{n=0}^{\infty} u_n z^n$, which is analytic in the region $|z| < 0.702$ and there satisfies the differential equation in Eq. (48), has a unique and simple root α of smallest modulus. We also calculate the first ten digits of $\alpha \approx 0.4889986317$. The calculation is performed without first solving the differential equation to obtain the function $U(z)$.

We start in Lemma 6 by providing a recurrence for u_n, which is then used to find an upper bound of $|u_n|$ in Lemma 8. Next, we consider the set $\mathcal{B} \equiv \{z \in \mathbb{C} \ : \ |z| \leq \frac{1}{2}\}$ in the complex plane and decompose $U(z)$ into a sum $U(z) = U_1(z) + U_2(z)$, where $U_1(z) = \sum_{n=0}^{100} u_n z^n$ is a polynomial and $U_2(z) = \sum_{n=101}^{\infty} u_n z^n$. The bound for

$|u_n|$ in Lemma 8 yields a bound for $|U_1(z)|$ (Lemma 9), which in turn implies that $|U_1(z)| > |U_2(z)|$ if $z \in \partial\mathcal{B}$. Hence, by Rouché's theorem we have that inside $\mathcal{B}$, the function $U(z)$ has the same number of roots—considered with their multiplicity—as the polynomial $U_1(z)$. Lemma 10 shows that $U_1(z)$ has a unique and simple root inside $\mathcal{B}$, and in Proposition 8 we conclude the proof of our claim by finding an approximation of the unique and simple root α of $U(z)$ inside $\mathcal{B}$—which turns out to be very close to the root of $U_1(z)$ inside $\mathcal{B}$.

In $U(z) = \sum_{n=0}^{\infty} u_n z^n$, we have $u_n \equiv [z^n]U(z)$. From Eq. (48), we derive a recurrence for u_n. Recall that e_n gives the mean number of root configurations in an ordered unlabeled history of size $n \geq 1$.

Lemma 6 *For $n \geq 2$, we have*

$$u_n = \frac{1}{n(n-1)} \sum_{k=0}^{n-1} (3n-k-3)u_k - \frac{4}{n(n-1)} \sum_{k=0}^{n-1} (n-2k-1)e_{n-k}u_k + \frac{4}{n(n-1)} \sum_{k=0}^{n-1} \left(\sum_{j=0}^{n-k-1} e_j \right) u_k, \tag{50}$$

with $u_0 = 1$ and $u_1 = 0$.

Proof. First notice that for $n \geq 0$, the coefficient of z^n in each term of Eq. (48) can be written as

$$[z^n]U''(z) = (n+2)(n+1)u_{n+2}$$

$$-[z^n]\left(g_1 + \frac{g_2'}{g_2}\right) U'(z) = -\sum_{k=0}^{n} (n-k+1)(4e_{k+1}+2)u_{n-k+1}$$

$$[z^n]g_2 g_0 U(z) = \sum_{k=0}^{n} \left[(k+1) + 4\sum_{j=0}^{k} e_{j+1} + 4\sum_{j=0}^{k+2} e_j e_{k-j+2} \right] u_{n-k},$$

where for convenience we set $e_0 = 0$.

Making a substitution to the index of summation, we have

$$-4\sum_{k=0}^{n} (n-k+1)e_{k+1}u_{n-k+1} = -4\sum_{k=0}^{n+1} k e_{n-k+2} u_k.$$

Hence, the sum for $-[z^n](g_1 + g_2'/g_2)U'(z)$ can be simplified as

$$-[z^n]\left(g_1 + \frac{g_2'}{g_2}\right) U'(z) = -4\sum_{k=0}^{n+1} k e_{n-k+2} u_k - 2\sum_{k=0}^{n} (n-k+1)u_{n-k+1}.$$

The second sum in this equation together with the first sum $\sum_{k=0}^{n}(k+1)u_{n-k}$ of $[z^n]g_2 g_0 U(z)$ give

$$-2\sum_{k=0}^{n} (n-k+1)u_{n-k+1} + \sum_{k=0}^{n} (k+1)u_{n-k} = \sum_{k=0}^{n+1} (n-3k+1)u_k.$$

Furthermore, by setting $n = k+2$ in Eq. (40), the inner sums of $[z^n]g_2 g_0 U(z)$ can be rewritten as

$$4\sum_{j=0}^{k} e_{j+1} + 4\sum_{j=0}^{k+1} e_j e_{k-j+2} = 4(k+1)e_{k+2} - 4(k+1) - 4\sum_{j=1}^{k+1} e_j.$$

Hence, the coefficient of z^n in Eq. (48) becomes

$$(n+2)(n+1)u_{n+2} - 4\sum_{k=0}^{n+1} k e_{n-k+2} u_k + \sum_{k=0}^{n+1} (n-3k+1)u_k + \sum_{k=0}^{n} \left[4(k+1)e_{k+2} - 4(k+1) - 4\sum_{j=1}^{k+1} e_j \right] u_{n-k}.$$

In this expression, we make two substitutions:

$$\begin{aligned}\sum_{k=0}^{n} 4(k+1)e_{k+2}u_{n-k} &= \sum_{k=0}^{n+1} 4(n-k+1)e_{n-k+2}u_k \qquad (51)\\ \sum_{k=0}^{n+1}(n-3k+1)u_k - 4\sum_{k=0}^{n}(k+1)u_{n-k} &= \sum_{k=0}^{n+1}(n-3k+1)u_k - 4\sum_{k=0}^{n}(n-k+1)u_k\\ &= \sum_{k=0}^{n+1}(-3n+k-3)u_k, \qquad (52)\end{aligned}$$

obtaining

$$(n+2)(n+1)u_{n+2} - 4\sum_{k=0}^{n+1} ke_{n-k+2}u_k + \sum_{k=0}^{n+1} 4(n-k+1)e_{n-k+2}u_k + \sum_{k=0}^{n+1}(-3n+k-3)u_k + \sum_{k=0}^{n}\left(-4\sum_{j=1}^{k+1} e_j\right)u_{n-k},$$

and thus

$$(n+2)(n+1)u_{n+2} + \sum_{k=0}^{n+1} 4(n-2k+1)e_{n-k+2}u_k + \sum_{k=0}^{n+1}(-3n+k-3)u_k + \sum_{k=0}^{n}\left(-4\sum_{j=1}^{k+1} e_j\right)u_{n-k}.$$

Finally, because $e_0 = 0$, in this expression we can substitute

$$\sum_{k=0}^{n}\left(-4\sum_{j=1}^{k+1} e_j\right)u_{n-k} = \sum_{k=0}^{n}\left(-4\sum_{j=0}^{k+1} e_j\right)u_{n-k} = \sum_{k=0}^{n}\left(-4\sum_{j=0}^{n-k+1} e_j\right)u_k = \sum_{k=0}^{n+1}\left(-4\sum_{j=0}^{n-k+1} e_j\right)u_k,$$

obtaining for $n \geq 0$

$$(n+2)(n+1)u_{n+2} + \sum_{k=0}^{n+1} 4(n-2k+1)e_{n-k+2}u_k - \sum_{k=0}^{n+1}(3n-k+3)u_k - 4\sum_{k=0}^{n+1}\left(\sum_{j=0}^{n-k+1} e_j\right)u_k = 0,$$

which rescaled is recurrence (50). The starting conditions $u_0 = 1$ and $u_1 = 0$, follow from the fact that $U(0) = 1$ and $U'(0) = 0$ as $U(z) = \exp[\int_0^z S(x)/(-x)\,dx]$. $\square$

In Lemma 8, we use the recurrence to find an upper bound for $|u_n|$. First, we need an upper bound for e_n.

Lemma 7 *For $n \geq 0$, we have $e_n \leq (\frac{9}{10})(\frac{3}{2})^n$.*

Proof. Using the recurrence (40), with the help of computing software we have shown that the inequality holds for $0 \leq n \leq 41$. We proceed by induction. Suppose the inequality holds for all $k < n$ with $n > 41$. By Eq. (40),

$$\begin{aligned} e_n &\leq 1 + \frac{81}{100(n-1)}\sum_{j=1}^{n-1}\left(\frac{3}{2}\right)^n + \frac{9}{5(n-1)}\sum_{j=1}^{n-1}\left(\frac{3}{2}\right)^j\\ &= 1 + \frac{81}{100}\left(\frac{3}{2}\right)^n + \frac{18}{5(n-1)}\left(\frac{3}{2}\right)^n - \frac{27}{5(n-1)}\\ &= \frac{9}{10}\left(\frac{3}{2}\right)^n - \frac{9}{10}\left(\frac{1}{10} - \frac{4}{n-1}\right)\left(\frac{3}{2}\right)^n - \frac{27}{5(n-1)} + 1.\end{aligned}$$

In the last step, we can see that a positive number is subtracted from $\frac{9}{10}(\frac{3}{2})^n$ for $n > 41$, as

$$\frac{9}{10}\left(\frac{1}{10} - \frac{4}{n-1}\right)\left(\frac{3}{2}\right)^n + \frac{27}{5(n-1)} - 1 > \frac{9}{10}\frac{1}{400}\left(\frac{3}{2}\right)^{42} - 1 > 0.$$

Thus, the claim is proved. $\square$

Lemma 8 *For $n \geq 0$, we have $|u_n| \leq (\frac{9}{5})^n$.*

Proof. Using recurrence (50), computing software verifies the inequality for $0 \leq n \leq 25$. We proceed by induction. Suppose that the inequality holds for all $k < n$ with $n > 25$. For simplicity of computation, instead of the bound in Lemma 7, we use the more conservative $(\frac{3}{2})^n$ as a bound for e_n. With Eq. (50), we get

$$
\begin{aligned}
|u_n| &\leq \frac{3}{n}\sum_{k=0}^{n-1}\left(\frac{9}{5}\right)^k + \frac{4}{n}\sum_{k=0}^{n-1}\left(\frac{3}{2}\right)^{n-k}\left(\frac{9}{5}\right)^k + \frac{4}{n(n-1)}\sum_{k=0}^{n-1}\left(\sum_{j=0}^{n-k-1}\left(\frac{3}{2}\right)^j\right)\left(\frac{9}{5}\right)^k \\
&= \frac{15}{4n}\left(\frac{9}{5}\right)^n - \frac{15}{4n} + \frac{20}{n}\left(\frac{9}{5}\right)^n - \frac{20}{n}\left(\frac{3}{2}\right)^n + \frac{30}{n(n-1)}\left(\frac{9}{5}\right)^n - \frac{40}{n(n-1)}\left(\frac{3}{2}\right)^n + \frac{10}{n(n-1)} \\
&= \frac{5(19n+5)}{4n(n-1)}\left(\frac{9}{5}\right)^n - \frac{20(n+1)}{n(n-1)}\left(\frac{3}{2}\right)^n - \frac{5(3n-11)}{4n(n-1)}.
\end{aligned}
$$

In the last step, we have $|u_n| \leq (\frac{9}{5})^n$, as for $n > 25$, the following two inequalities hold:

$$
\begin{aligned}
\frac{5(19n+5)}{4n(n-1)} &\leq 1 \\
-\frac{20(n+1)}{n(n-1)}\left(\frac{3}{2}\right)^n - \frac{5(3n-11)}{4n(n-1)} &\leq 0.
\end{aligned}
$$

Thus, the claim is proved. □

We now consider the set $\mathcal{B} \equiv \{z \in \mathbb{C} \ : \ |z| \leq \frac{1}{2}\}$, and the partition $U(z) = \sum_{k=0}^{\infty} u_k z^k = U_1(z) + U_2(z)$, $U_1(z) \equiv \sum_{k=0}^{100} u_k z^k$ and $U_2(z) \equiv \sum_{k=101}^{\infty} u_k z^k$. Using the bound for $|u_n|$ from Lemma 8, for each $z \in \mathcal{B}$ we have

$$
|U_2(z)| \leq \sum_{k=101}^{\infty} |u_k|\,|z|^k \leq \sum_{k=101}^{\infty}\left(\frac{9}{5}\right)^k\left(\frac{1}{2}\right)^k = 10\left(\frac{9}{10}\right)^{101} \approx 0.0002390525900. \tag{53}
$$

Next, we need a lower bound for $|U_1(z)|$.

Lemma 9 *We have* $\min_{z\in\partial\mathcal{B}} |U_1(z)| \geq \dfrac{3}{1000}$.

Proof. We obtain the result by considering a function

$$
G(t) \equiv \left[\sum_{k=0}^{100} u_k \cos(kt)\left(\frac{1}{2}\right)^k\right]^2 + \left[\sum_{k=0}^{100} u_k \sin(kt)\left(\frac{1}{2}\right)^k\right]^2.
$$

$G(t)$ has period 2π, with $G(\pi - t) = G(\pi + t)$, if $t \in [0, \pi]$. For $|z| \in \partial\mathcal{B}$ we can write $z = \frac{1}{2}[\cos t + i \sin t]$ for $t \in [0, 2\pi)$, and thus

$$
|U_1(z)| \;=\; \left|\sum_{k=0}^{100} u_k \left[\left(\frac{1}{2}\right)[\cos t + i\sin t]\right]^k\right| = \left|\sum_{k=0}^{100} u_k \cos(kt)\left(\frac{1}{2}\right)^k + i\sum_{k=0}^{100} u_k \sin(kt)\left(\frac{1}{2}\right)^k\right| = \sqrt{G(t)}.
$$

By using the bound in Lemma 8, we have the following inequality

$$\begin{aligned}
|G'(t)| &= \Bigg| 2\left[\sum_{k=0}^{100} u_k \cos(kt)\left(\frac{1}{2}\right)^k\right]\left[-\sum_{k=0}^{100} k u_k \sin(kt)\left(\frac{1}{2}\right)^k\right] \\
&\qquad + 2\left[\sum_{k=0}^{100} u_k \sin(kt)\left(\frac{1}{2}\right)^k\right]\left[\sum_{k=0}^{100} k u_k \cos(kt)\left(\frac{1}{2}\right)^k\right]\Bigg| \\
&\le 2\left|\sum_{k=0}^{100} u_k \cos(kt)\left(\frac{1}{2}\right)^k\right|\left|\sum_{k=0}^{100} k u_k \sin(kt)\left(\frac{1}{2}\right)^k\right| \\
&\qquad + 2\left|\sum_{k=0}^{100} u_k \sin(kt)\left(\frac{1}{2}\right)^k\right|\left|\sum_{k=0}^{100} k u_k \cos(kt)\left(\frac{1}{2}\right)^k\right| \\
&\le 2\left[\sum_{k=0}^{100} |u_k||\cos(kt)|\left(\frac{1}{2}\right)^k\right]\left[\sum_{k=0}^{100} k|u_k||\sin(kt)|\left(\frac{1}{2}\right)^k\right] \\
&\qquad + 2\left[\sum_{k=0}^{100} |u_k||\sin(kt)|\left(\frac{1}{2}\right)^k\right]\left[\sum_{k=0}^{100} k|u_k||\cos(kt)|\left(\frac{1}{2}\right)^k\right] \\
&\le 4\left[\sum_{k=0}^{100}\left(\frac{9}{10}\right)^k\right]\left[\sum_{k=0}^{100} k\left(\frac{9}{10}\right)^k\right] \approx 3598.862135. \qquad (54)
\end{aligned}$$

We set $\mathcal{I} = \{\frac{k\pi}{1000000} \ : \ k \in \mathbb{Z}, 0 \le k \le 1000000\}$. A numerical calculation shows that

$$\min_{t\in\mathcal{I}} G(t) = G(0) \approx 0.01949528529. \qquad (55)$$

With these preparations complete, we prove our claim by showing that

$$\min_{t\in[0,\pi]} G(t) \ge \frac{9}{1000000}. \qquad (56)$$

We prove Eq. (56) by contradiction. Suppose there exists $t_0 \in [0,\pi]$ such that $G(t_0) < \frac{9}{1000000}$. Then we can find $t_1 \in \mathcal{I}$ such that

$$|t_1 - t_0| \le \frac{\pi}{2000000}. \qquad (57)$$

By the Mean Value Theorem, we can find $c \in (t_0, t_1)$ such that $G(t_1) - G(t_0) = G'(c)(t_1 - t_0)$. From Eqs. (54) and (57),

$$\frac{1800\pi}{1000000} \ge |G'(c)(t_1 - t_0)| = |G(t_1) - G(t_0)| \ge G(t_1) - G(t_0). \qquad (58)$$

However, because $t_1 \in \mathcal{I}$, by Eq. (55), we have

$$G(t_1) - G(t_0) \ge G(0) - G(t_0) \ge \frac{1}{100} - \frac{9}{1000000} = \frac{9991}{1000000}.$$

This result contradicts the upper bound in Eq. (58). Thus, Eq. (56) holds and the claim has been proven. □

Next, we study the root of $U_1(z)$ inside $\mathcal{B}$.

Lemma 10 *The polynomial $U_1(z)$ has a unique (simple) root β inside $\mathcal{B}$, with $\beta \approx 0.4889986317$.*

Proof. First, by the Intermediate Value Theorem, there exists a real root β with $0 < \beta < \frac{1}{2}$, as we can numerically compute $U_1(0)\,U_1(\frac{1}{2}) < 0$ for the polynomial $U_1(z)$. Thus, we must prove

$$\frac{U_1(z)}{z-\beta} = \frac{U_1(z) - U_1(\beta)}{z-\beta} = \sum_{k=0}^{100} u_k \frac{z^k - \beta^k}{z-\beta} = \sum_{k=0}^{100} u_k \sum_{\ell=0}^{k-1} \beta^{k-1-\ell} z^\ell = \sum_{\ell=0}^{99} \left(\sum_{k=\ell+1}^{100} u_k \beta^{k-1-\ell} \right) z^\ell$$

satisfies $|U_1(z)/(z-\beta)| > 0$ in $\mathcal{B}$.

To do so, we first use the bisection method for root-finding to numerically approximate β by

$$\tilde{\beta} = \frac{1101127027820569}{2251799813685248} \approx 0.4889986317,$$

with the approximation error

$$|\beta - \tilde{\beta}| \le \frac{1}{2^{50}}. \tag{59}$$

Then, we define the polynomial

$$Q(z) \equiv \sum_{\ell=0}^{99} a_\ell z^\ell, \text{ with } a_\ell \equiv \sum_{k=\ell+1}^{100} u_k \tilde{\beta}^{k-1-\ell},$$

through which we can write

$$\begin{aligned} \frac{U_1(z)}{z-\beta} &= Q(z) + (\beta - \tilde{\beta}) R(z), \\ R(z) &\equiv \sum_{\ell=0}^{99} \left(\sum_{k=\ell+1}^{100} u_k \frac{\beta^{k-1-\ell} - \tilde{\beta}^{k-1-\ell}}{\beta - \tilde{\beta}} \right) z^\ell = \sum_{\ell=0}^{99} \left(\sum_{k=\ell+2}^{100} u_k \sum_{j=0}^{k-2-\ell} \beta^j \tilde{\beta}^{k-2-\ell-j} \right) z^\ell. \end{aligned}$$

Note that on $\mathcal{B}$,

$$|R(z)| \le \sum_{\ell=0}^{99} \sum_{k=\ell+2}^{100} \sum_{j=0}^{k-2-\ell} |u_k| |\beta|^j |\tilde{\beta}|^{k-2-\ell-j} |z|^\ell \le \sum_{\ell=0}^{99} \sum_{k=\ell+2}^{100} \sum_{j=0}^{k-2-\ell} \left(\frac{9}{5}\right)^k \left(\frac{1}{2}\right)^{k-2} \approx 3234.224489, \tag{60}$$

where we used the bound for $|u_n|$ from Lemma 8 and the fact that $\beta, \tilde{\beta}, |z| \le \frac{1}{2}$.

Next, let us consider the function

$$S(r,\theta) \equiv \sum_{\ell=0}^{99} a_\ell r^\ell \cos(\ell\theta)$$

defined over the rectangle $(r,\theta) \in [0, \frac{1}{2}] \times [0, \pi]$, where $S(r,\theta) = \Re(Q(z))$ if $z = r[\cos(\pm\theta) + i \sin(\pm\theta)] \in \mathcal{B}$. We need the following bound for the gradient of S:

$$\begin{aligned} |\nabla S| &= \left| \left(\sum_{\ell=0}^{99} \ell a_\ell r^{\ell-1} \cos(\ell\theta), \sum_{\ell=0}^{99} -\ell a_\ell r^\ell \sin(\ell\theta) \right) \right| = \left| \sum_{\ell=0}^{99} \left(\ell a_\ell r^{\ell-1} \cos(\ell\theta), -\ell a_\ell r^\ell \sin(\ell\theta) \right) \right| \\ &= \left| \sum_{\ell=0}^{99} \ell a_\ell r^{\ell-1} \left(\cos(\ell\theta), -r\sin(\ell\theta) \right) \right| \le \sum_{\ell=0}^{99} \ell |a_\ell| |r|^{\ell-1} |\left(\cos(\ell\theta), -r\sin(\ell\theta) \right)| \\ &\le \sum_{\ell=0}^{99} \ell |a_\ell| |r|^{\ell-1} \le \sum_{\ell=0}^{99} \ell |a_\ell| \left(\frac{1}{2}\right)^{\ell-1} \approx 89.628949. \end{aligned} \tag{61}$$

Here, we have made use of $|r| < \frac{1}{2}$ and for $|r| < 1$, $\sqrt{\cos^2 x + r^2 \sin^2 x} \leq \sqrt{\cos^2 x + \sin^2 x} = 1$.

A numerical calculation shows that over the grid $\mathcal{I} \equiv \{(\frac{k}{2000}, \frac{j\pi}{1000}) \ : \ (k,j) \in \mathbb{Z}^2, 0 \leq k, j \leq 1000\}$, we have

$$\min_{(r,\theta)\in\mathcal{I}} |S(r,\theta)| = \left| S\left(\frac{1}{2}, \frac{502\pi}{1000}\right) \right| \approx 0.9518894218. \tag{62}$$

We now show—with a similar method to that used to prove Lemma 9—that

$$\min_{(r,\theta)\in[0,\frac{1}{2}]\times[0,\pi]} |S(r,\theta)| \geq \frac{3235}{2^{50}}. \tag{63}$$

Suppose for contradiction that there exists $z_0 = (r_0, \theta_0) \in [0, \frac{1}{2}] \times [0, \pi]$ such that $|S(r_0, \theta_0)| < 3235/2^{50}$. Then let us take $z_1 = (r_1, \theta_1) \in \mathcal{I}$ such that

$$|z_1 - z_0| < \sqrt{\frac{1}{16} + \frac{\pi^2}{4}} \left(\frac{1}{1000} \right) \leq \frac{1}{500}. \tag{64}$$

By the Mean Value Theorem, there exists a point (r, θ) on the line segment from (r_0, θ_0) to (r_1, θ_1) such that

$$\nabla S(r,\theta) \cdot (z_1 - z_0) = S(r_1, \theta_1) - S(r_0, \theta_0),$$

where $\cdot$ is the inner product of $\mathbb{R}^2$. By using the Cauchy-Schwarz inequality together with (61), (62) and (64), the assumption $|S(r_0, \theta_0)| < 3235/2^{50}$ would thus give

$$\begin{aligned} \frac{90}{500} &\geq |\nabla S(r,\theta)||z_1 - z_0| \geq |\nabla S(r,\theta) \cdot (z_1 - z_0)| = |S(r_1,\theta_1) - S(r_0,\theta_0)| \\ &\geq |S(r_1,\theta_1)| - |S(r_0,\theta_0)| \geq \frac{9}{10} - \frac{3235}{2^{50}} > 0.89, \end{aligned}$$

which is a contradiction. Hence, Eq. (63) holds.

Finally, because for $z \in \mathcal{B}$ we have

$$|Q(z)| \geq |\Re(Q(z))| \geq \min_{(r,\theta)\in[0,\frac{1}{2}]\times[0,\pi]} |S(r,\theta)|,$$

by using Eqs. (59), (60), and (63) it follows that in $\mathcal{B}$,

$$\begin{aligned} \left| \frac{U_1(z)}{z - \beta} \right| &= \left| Q(z) + (\beta - \tilde{\beta}) R(z) \right| \geq \left| |Q(z)| - |(\tilde{\beta} - \beta) R(z)| \right| \geq \frac{3235}{2^{50}} - |(\tilde{\beta} - \beta)||R(z)| \\ &\geq \frac{3235}{2^{50}} - \frac{|R(z)|}{2^{50}} > \frac{3235}{2^{50}} - \frac{3234.224489\ldots}{2^{50}} > 0. \end{aligned}$$

This concludes the proof. □

Combining Lemmas 9 and 10 with the inequality in Eq. (53), we obtain the following proposition.

Proposition 8 *The function $U(z)$ has a unique (simple) root α inside $\mathcal{B}$, where $\alpha \approx 0.4889986317$.*

Proof. For the decomposition $U(z) = U_1(z) + U_2(z)$, Eq. (53) together with Lemma 9 gives for $z \in \partial\mathcal{B}$

$$|U_1(z)| \geq \frac{3}{1000} > 0.00025 > |U_2(z)|.$$

Hence, from Rouché's theorem, inside $\mathcal{B}$ the function $U(z)$ has the same number of roots (considered with multiplicity) as polynomial $U_1(z)$. From Lemma 10, we know that $U_1(z)$ has one (simple) root inside $\mathcal{B}$.

The only remaining step is the numerical computation of α, whose first ten digits turn out to coincide with the constant β found in Lemma 10 as the root of $U_1(z)$ inside $\mathcal{B}$. We again decompose $U(z)$:

$$U(z) = \sum_{k=0}^{\infty} u_k z^k = \sum_{k=0}^{500} u_k z^k + \sum_{k=501}^{\infty} u_k z^k = \tilde{U}_1(z) + \tilde{U}_2(z).$$

Note that from our bound for $|u_k|$ (Lemma 8), for each $z \in \mathcal{B}$ we have

$$|\tilde{U}_2(z)| \leq \sum_{k=501}^{\infty} |u_k|\,|z|^k \leq \sum_{k=501}^{\infty} \left(\frac{9}{5}\right)^k \left(\frac{1}{2}\right)^k = 10\left(\frac{9}{10}\right)^{501} \leq 10^{-21}. \tag{65}$$

Let us now consider

$$\begin{aligned} \alpha' &= \frac{550563513910285}{1125899906842624} \approx 0.48899863172938484723 \\ \alpha'' &= \frac{1101127027820571}{2251799813685248} \approx 0.48899863172938529132. \end{aligned}$$

These values were chosen using the bisection method such that

$$\tilde{U}_1(\alpha') = 2.708185805\ldots \cdot 10^{-16} \quad \text{and} \quad \tilde{U}_1(\alpha'') = -4.953373282\ldots \cdot 10^{-15}.$$

From the bound of $|\tilde{U}_2(z)|$ in Eq. (65), it is clear that $U(\alpha') > 0$ and $U(\alpha'') < 0$. Let α be the unique root of $U(z)$ in $\mathcal{B}$, which by the Intermediate Value Theorem must be a real root in (α', α''), and let $\epsilon \equiv \alpha - \alpha' \leq 10^{-14}$. Note that

$$\frac{1}{\alpha'} - \frac{1}{\alpha} = \frac{\epsilon}{\alpha'(\alpha' + \epsilon)} \leq \frac{\epsilon}{(\alpha')^2} \leq 5 \cdot 10^{-14}.$$

Thus, we can use

$$\begin{aligned} \alpha' &= 0.48899863172938484723 \\ (\alpha')^{-1} &= 2.0449954971518340953 \end{aligned}$$

to approximate α and α^{-1}, respectively. □

Acknowledgments This work developed from discussions at the Banff International Research Station. Support was provided by a Rita Levi-Montalcini grant from the Ministero dell'Istruzione, dell'Università e della Ricerca (FD), grants MOST-104-2923-M-009-006-MY3 and MOST-107-2115-M-009-010-MY2 (MF, ARP), and National Institutes of Health grant R01 GM131404 (NAR).

References

[1] D. Aldous. Probability distributions on cladograms. In D. Aldous and R. Pemantle, editors, *Random Discrete Structures*, pages 1–18. Springer-Verlag, New York, 1996.

[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] F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. *Lect. Notes Comput. Sc.*, 581:24–48, 1992.

[4] M. G. B. Blum, O. François, and S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Adv. Appl. Prob.*, 16:2195–2214, 2006.

[5] O. Bodini, J. Courtiel, S. Dovgal, and H.-K. Hwang. Asymptotic distribution of parameters in random maps. In J. A. Fill and M. D. Ward, editors, *29th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, pages 13:1–13:12. Dagstuhl, Germany, 2018.

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

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

[8] J. H. Degnan, N. A. Rosenberg, and T. Stadler. A characterization of the set of species trees that produce anomalous ranked gene trees. *IEEE/ACM Trans. Comp. Biol. Bioinf.*, 9:1558–1568, 2012.

[9] J. H. Degnan, N. A. Rosenberg, and T. Stadler. The probability distribution of ranked gene trees on a species tree. *Math. Biosci.*, 235:45–55, 2012.

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

[11] 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.

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

[13] 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. Comp. Biol. Bioinf.*, 13:913–925, 2016.

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

[15] 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.

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

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

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

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

[20] J. A. Fill. On the distribution of binary search trees under the random permutation model. *Random Struct. Algor.*, 8:1–25, 1996.

[21] J. A. Fill and N. Kapur. Limiting distributions for additive functionals on Catalan trees. *Theor. Comp. Sci.*, 326:69–102, 2004.

[22] P. Flajolet, X. Gourdon, and C. Martinez. Patterns in random binary search trees. *Random Struct. Algor.*, 11:223–244, 1997.

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

[24] M. Fuchs, C. Holmgren, D. Mitsche, and R. Neininger. A note on the independence number, domination number and related parameters of random binary search trees and random recursive trees. *Discr. Appl. Math.*, 292:64–71, 2021.

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

[26] 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.

[27] C. Holmgren and S. Janson. Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.*, 20:1–51, 2015.

[28] A. Lambert and T. Stadler. Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theor. Pop. Biol.*, 90:113–128, 2013.

[29] W. P. Maddison. Gene trees in species trees. *Syst. Biol.*, 46:523–536, 1997.

[30] W. P. Maddison and L. L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.*, 55:21–30, 2006.

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

[32] 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.

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

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

[35] 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.

[36] N. A. Rosenberg and R. Tao. Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst. Biol.*, 57:131–140, 2008.

[37] T. Stadler and J. H. Degnan. A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Alg. Mol. Biol.*, 7:7, 2012.

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

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

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

[41] C. Than and L. Nakhleh. Species tree inference by minimizing deep coalescences. *PLoS Comp. Biol.*, 5:e1000501, 2009.

[42] C. Than, D. Ruths, H. Innan, and L. Nakhleh. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.*, 14:517–535, 2007.

[43] C. V. Than and N. A. Rosenberg. Mathematical properties of the deep coalescence cost. *IEEE/ACM Trans. Comp. Biol. Bioinformat.*, 10:61–72, 2013.

[44] C. V. Than and N. A. Rosenberg. Mean deep coalescence cost under exchangeable probability distributions. *Discr. Appl. Math.*, 174:11–26, 2014.

[45] S. Wagner. Central limit theorems for additive tree parameters with small toll functions. *Combinator. Prob. Comput.*, 24:329–353, 2015.

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

[47] 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.

[48] 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.