Distribution of external branch lengths in Yule histories

Filippo Disanto[a]   Michael Fuchs[b]

Submitted: Aug 10, 2022; Accepted: Jul 17, 2023; Published: Aug 25, 2023

Abstract

The Yule branching process is a classical model for the random generation of gene tree topologies in population genetics. It generates binary ranked trees—also called *histories*—with a finite number n of leaves. We study the lengths $\ell_1 > \ell_2 > \cdots > \ell_k > \cdots$ of the external branches of a Yule generated random history of size n, where the length of an external branch is defined as the rank of its parent node. When $n \to \infty$, we show that the random variable ℓ_k, once rescaled as $\frac{n-\ell_k}{\sqrt{n/2}}$, follows a χ-distribution with $2k$ degrees of freedom, with mean $\mathbb{E}(\ell_k) \sim n$ and variance $\mathbb{V}(\ell_k) \sim n\big(k - \frac{\pi k^2}{16^k}\binom{2k}{k}^2\big)$. Our results contribute to the study of the combinatorial features of Yule generated gene trees, in which external branches are associated with singleton mutations affecting individual gene copies.

Mathematics Subject Classifications: 05A15, 05A16, 05C05, 60C05

1 Introduction

Tree models of speciation are crucial in biological studies for testing hypotheses about evolution. From the spectrum of mutations observed across a set of genes, statistical methods [10] enable the inference of a tree representing the ancestry relationships among the sampled genetic sequences. The comparison of the inferred tree with model predictions can assist in the analysis of the biological forces that have driven the evolution of the considered genes.

The reconstruction of the gene tree from genome data can be subject to several types of errors. Measuring branches in proper units of time, one problem is estimating the exact edge lengths of the tree from the polymorphism observed along the considered chromosomes. For example, assuming that molecular differences have accumulated at a constant rate, the human-chimpanzee divergence is estimated to date back to 4.3 millions

[a] Department of Mathematics, University of Pisa, Pisa, Italy (filippo.disanto@unipi.it).

[b] Department of Mathematical Sciences, National Chengchi University, Taipei 116, Taiwan (mfuchs@nctu.edu.tw).

years ago, while—at the moment—the oldest fossil with human-like features is 100,000 years older (pag. 31 of [19]). A less informative but more robust inference approach can proceed by restricting the tree search space to infer only the "topology"—i.e., the branching pattern and the relative temporal order of the speciation (splitting) events—of the gene tree, which will be then compared with tree topologies considered under a proper neutral model.

The Yule distribution [14, 25] is a fundamental probability model of tree topologies, also called "histories", used in evolutionary analyses. Histories are full binary rooted trees, with a ranking of internal nodes that divides the tree in different layers (Fig. 1A). The probabilistic features of Yule distributed histories have been subject of numerous investigations (see, e.g., [2, 18, 20, 21, 22]), with a particular interest on combinatorial properties that affect the frequency spectrum of mutations in population genetic tree models. Our focus is on the length distribution of tree branches. Branch length can be seen as a discrete parameter—when only the number of tree layers spanned by a branch is considered—or as a time related quantity—when each tree layer is in turn considered with a length given by a continuous random variable. In the latter case, histories are called "coalescent" trees. While branch length of coalescent trees has been widely studied (see, e.g., [1, 5, 6, 7, 11, 12, 16]), the discrete length of the edges of a random history has received less attention.

In this paper, extending previous results [9], we investigate the distribution of the different lengths of the external branches—i.e., those branches ending with a leaf—of random histories of given size selected under the Yule model. External branch length is an important parameter to study as it relates to singleton mutations in the site frequency spectrum of population genetic trees. Denoting by ℓ_k the kth largest length of an external branch in a Yule distributed random history of n leaves, our main finding is that, for every $k \geqslant 1$, the rescaled variable $\frac{n-\ell_k}{\sqrt{n/2}}$ follows asymptotically a χ-distribution with $2k$ degrees of freedom, with convergence of all moments (Theorem 9).

The paper is organized as follows. We introduce terminology and some useful properties of histories in Section 2, showing in particular that external branch lengths in random histories can also be analyzed in terms of peaks of random permutations. In Section 3, we refine calculations of [9] finding a closed formula for the probability of the length, ℓ_1, of the longest external branch in a random history of given size n and a recurrence for computing the probability of the kth largest length, ℓ_k, of an external branch. For increasing n, the asymptotic distribution of the variables $\ell_1, \ell_2, \ldots, \ell_k, \ldots$ is finally examined in Section 4. Our results on the discrete variables ℓ_k parallel those obtained by Bocharov *et al.* [3] on the distribution of the *time* length of the kth longest external branch of a random tree of depth t generated under the Yule pure-birth process.

2 Yule histories, external branches and non-peak values of permutations

For a given positive integer n, a *history* [20] of size n is a full binary rooted tree with n leaves and $n-1$ ranked internal nodes (Fig. 1A). The rank of each internal node is defined by an integer label in $[1, n-1]$ bijectively associated with the node. The labeling decreases along any path from the root toward a leaf of the tree, determining a temporal ordering of the coalescent events—the merging of two edges—that characterize the branching structure of the tree. In a history of size n, there are $2n-1$ edges, or *branches*. A branch connecting an internal node and a leaf is said to be an *external* branch. The *length* of a branch is the difference between the rank of the nodes it connects. If the branch is external, then its length is simply the rank of its parent node.

In population genetics, histories are tree structures that represent the evolution of individual genes from a common ancestor. Conditioning on a given history, an infinite sites model [19] produces a set of mutations across the genes associated with the leaves of the tree. Roughly speaking, mutations occur as random events along the branches of the history (Fig. 1B), with each branch containing a number of mutations that depends on its length, and with each mutation affecting only the set of gene copies descended from the branch it belongs to. In particular, a history with one or more "long" external branches will be associated with a biological scenario in which one or more gene copies will possess a "large" number of singleton mutations—i.e., mutations affecting only one individual. A random history of size n selected under a proper null model distribution describes the evolutionary relationships of n individual genes randomly sampled from a population under neutral evolution, and the length of the longest external branches in the random history relates to the largest number of singleton mutations that characterize single individuals in the sample.

In this paper, we focus on distributive properties of external branch length for random histories considered under a well known model of neutral evolution. More precisely, we will study external branch lengths ordered by size over random histories of size n selected under the *Yule* probability model [14, 25], or, equivalently, over random ordered histories of size n selected uniformly at random. An *ordered* history of size n is a plane embedding of a history of size n in which subtrees carry a left-right orientation. In other words, flipping the two subtrees stemming from a given node of an ordered history yields a different ordered history (unless the flipped subtrees consist of only one node). The number of ordered histories of size n is thus $(n-1)!$, and the Yule distribution over the set of histories of size n is induced by the uniform distribution over the set of ordered histories of size n by summing the probability $1/(n-1)!$ of each ordered history with the same underlying (un-ordered) history [8]. In particular, if $\mathrm{c}(t)$ is the number of cherries (i.e., pendant subtrees with exactly two leaves) in a history t of n leaves, then $2^{n-1-\mathrm{c}(t)}$ is the number of different plane embeddings of t, and therefore $2^{n-1-\mathrm{c}(t)}/(n-1)!$ is the Yule probability of the history t [20].

A series of combinatorial results on the lengths of external branches of uniformly distributed ordered histories (or Yule distributed histories) has been obtained in [9] in

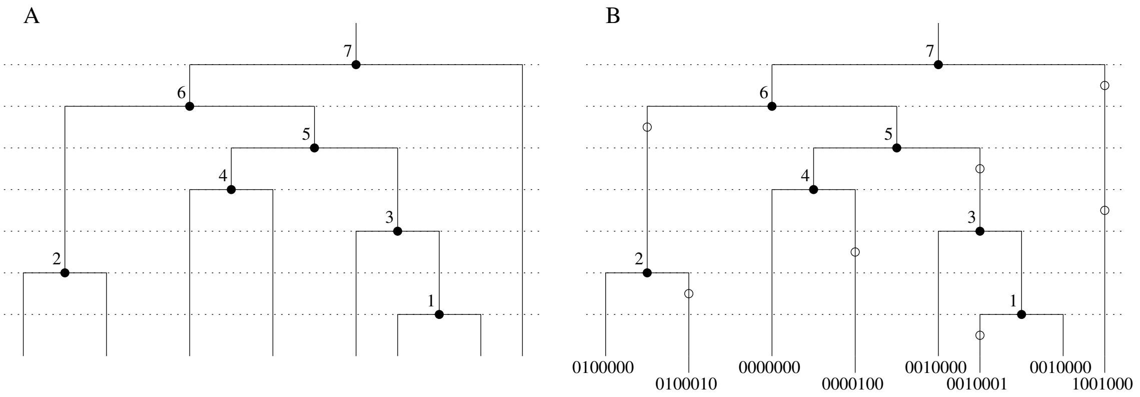

Figure 1: Histories and gene sequences. **(A)** A history of size $n = 8$. The ranking of internal nodes decreases along any path going from the root to the leaves of the tree. The length of an external branch is the rank of its parent node. The different lengths of the external branches ordered by size are $\ell_1 = 7 > \ell_2 = 4 > \ell_3 = 3 > \ell_4 = 2 > \ell_5 = 1$. **(B)** The history depicted in A with leaves associated with genes represented as binary sequences with ancestral alleles of type 0 and derived alleles of type 1. A mutation (white circle) affects only the gene sequences associated with the leaves descending from the branch where it occurs. In this example, there is a mutation for each layer of the tree: the ith mutation (looking from top to bottom) changes the allele at the ith locus (position) of the gene.

relationship with a study [4] of the number of permutations of fixed size with a given set of peak values, where the entry $\pi(i)$ is a *peak value* in the permutation $(\pi(1), \dots, \pi(i), \dots, \pi(n))$ when $i \neq 1, i \neq n$ and $\pi(i-1) < \pi(i) > \pi(i+1)$. Indeed, there exists a well known [13] bijection that associates an ordered history t of size n with a permutation π_t of the first $n-1$ positive integers. The mapping $t \to \pi_t$ can be described recursively by setting $\pi_t = (\pi_{t_L}, r(t), \pi_{t_R})$, where $r(t)$ is the (label of the) root of t, and t_L, t_R are respectively the left and right subtrees stemming from the root of t (if any). In particular, ordering by size the different lengths $\ell_1 > \ell_2 > \cdots > \ell_k > \cdots$ of the external branches of t, the kth length, ℓ_k, is easily seen to correspond to the kth largest non-peak value in the permutation π_t. For example, if t is the ordered history of size $n = 8$ depicted in Fig. 1, then $\pi_t = (2, 6, 4, 5, 3, 1, 7)$ has the following non-peak values: $2, 4, 3, 1, 7$, which correspond to the different lengths $\ell_1 = 7 > \ell_2 = 4 > \ell_3 = 3 > \ell_4 = 2 > \ell_5 = 1$ of the external branches of t. By using the correspondence with non-peak values of permutations, in the next section we calculate the probability of the variable ℓ_k in an ordered history of size n selected uniformly at random.

3 The probability of the kth external branch length

Given an ordered history t of size n, consider the different external branch lengths of t ordered by size as $\ell_1 > \ell_2 > \cdots > \ell_k > \cdots$, where $\ell_k \leqslant n - k$. As observed above, the value of ℓ_k corresponds to the kth largest non-peak value in the associated permutation π_t.

In this section, we study the number $h_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k)$ of ordered histories of size n in which $\ell_j = s_j$ for $j = 1, \ldots, k$, which determines the probability $p_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k) = h_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k)/(n-1)!$.

We start our calculations by using the following result of [4] for the number $\Pi_n(Q)$ of permutations of size n with peak values matching the elements of a given set Q :

Lemma 1 (Lemma 3.3 of [4])**.** *Let $n \geqslant 3$, $S \subseteq [3, n]$, and $r = \max S$ if $S \neq \emptyset$, 1 otherwise. For any $0 \leqslant j \leqslant n - r - 1$, we have*

$$\Pi_n(S \cup [n-j+1, n]) = 2(j+1)\Pi_{n-1}(S \cup [n-j, n-1]) + j(j+1)\Pi_{n-2}(S \cup [n-j, n-2]),$$

where $[a, b] = \{x \in \mathbb{Z} : a \leqslant x \leqslant b\}$.

We use the lemma as follows. Fix $s_1, s_2, \ldots, s_{k-1}, s_k$ such that $n \geqslant s_1 > s_2 > \cdots > s_{k-1} > s_k$, and let Z be a subset of the integers in the interval $[3, s_k - 1]$. Then, by replacing $S = Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cdots \cup [s_2+1, s_1-1]$ and $j = n - s_1$ in the formula above, we find

$$\begin{aligned}
& \Pi_n(Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cup [s_2+1, s_1-1] \cup [s_1+1, n]) \\
= \; & \Pi_n(S \cup [n-j+1, n]) \\
= \; & 2(j+1)\Pi_{n-1}(S \cup [n-j, n-1]) + j(j+1)\Pi_{n-2}(S \cup [n-j, n-2]) \\
= \; & 2(n-s_1+1) \\
\times \; & \Pi_{n-1}(Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cdots \cup [s_2+1, s_1-1] \cup [s_1, n-1]) \\
+ \; & (n-s_1)(n-s_1+1) \\
\times \; & \Pi_{n-2}(Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cdots \cup [s_2+1, s_1-1] \cup [s_1, n-2]).
\end{aligned}$$

If we sum both sides of the latter equation over the possible subsets Z of $[3, s_k - 1]$, then we obtain

$$\begin{aligned}
& \sum_Z \Pi_n(Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cdots \cup [s_2+1, s_1-1] \cup [s_1+1, n]) \qquad (1) \\
& = 2(n-s_1+1) \\
& \times \sum_Z \Pi_{n-1}(Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cdots \cup [s_2+1, s_1-1] \cup [s_1, n-1]) \\
& + (n-s_1)(n-s_1+1) \\
& \times \sum_Z \Pi_{n-2}(Z \cup [s_k+1, s_{k-1}-1] \cup [s_{k-1}+1, s_{k-2}-1] \cup \cdots \cup [s_2+1, s_1-1] \cup [s_1, n-2]),
\end{aligned}$$

where the first sum counts the permutations of size n in which the first largest non-peak value is $\ell_1 = s_1$, the second largest non-peak value is $\ell_2 = s_2, dots$, and the kth largest non-peak value is $\ell_k = s_k$. Similarly, the second and third sums count respectively the permutations of size $n-1$ and $n-2$ in which $\ell_1 = s_2, \ell_2 = s_3, dots$, and $\ell_{k-1} = s_k$. Note that when we set $k = 1$ and $s_1 = s$, we have $S = Z \subseteq [3, s-1]$ and the calculation above yields

$$\begin{aligned}\sum_Z \Pi_n(Z \cup [s+1,n]) &= 2(n-s+1) \sum_Z \Pi_{n-1}(Z \cup [s,n-1]) \\ &\quad +(n-s)(n-s+1) \sum_Z \Pi_{n-2}(Z \cup [s,n-2]),\end{aligned} \tag{2}$$

where the first sum counts the permutations of size n in which the largest non-peak value is $\ell_1 = s$, while the second and third sums count respectively the permutations of size $n-1$ and $n-2$ in which the largest non-peak value is strictly smaller than s, that is, $\ell_1 < s$. By rewriting (1) and (2) in terms of ordered histories, we find

$$\begin{aligned}h_{n+1}(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k) &= 2(n-s_1+1) \\ &\quad \times h_n(\ell_1 = s_2, \ell_2 = s_3, \ldots, \ell_{k-1} = s_k) \\ &\quad +(n-s_1)(n-s_1+1) \\ &\quad \times h_{n-1}(\ell_1 = s_2, \ell_2 = s_3, \ldots, \ell_{k-1} = s_k)\end{aligned} \tag{3}$$

and

$$h_{n+1}(\ell_1 = s) = 2(n-s+1)\, h_n(\ell_1 < s) + (n-s)(n-s+1)\, h_{n-1}(\ell_1 < s), \tag{4}$$

where $h_i(\ell_1 < s) \equiv \sum_{j<s} h_i(\ell_1 = j)$.

Because $h_{n+1}(\ell_1 = s) = h_{n+1}(\ell_1 < s+1) - h_{n+1}(\ell_1 < s)$, Eq. (4) yields the recurrence $h_{n+1}(\ell_1 < s+1) = h_{n+1}(\ell_1 < s) + 2(n-s+1)\, h_n(\ell_1 < s) + (n-s)(n-s+1)\, h_{n-1}(\ell_1 < s)$, which, by replacing $n+1$ by n and $s+1$ by s, reads as

$$h_n(\ell_1 < s) = h_n(\ell_1 < s-1) + 2(n-s+1)\, h_{n-1}(\ell_1 < s-1) + (n-s)(n-s+1)\, h_{n-2}(\ell_1 < s-1), \tag{5}$$

where $h_n(\ell_1 < s) = 0$ if $s = \lceil n/2 \rceil$ (ℓ_1 is at least $\lceil n/2 \rceil$), and $h_n(\ell_1 < s) = (n-1)!$ if $s = n$ (ℓ_1 is at most $n-1$). In particular, when $\lceil n/2 \rceil \leqslant s \leqslant n \geqslant 3$, we have

$$h_n(\ell_1 < s) = \frac{(s-1)!\,(s-2)!\,(2s-n)\,(2s-n-1)}{(2s-n)!} \tag{6}$$

as the right-hand side—say $r(n,s)$—of the latter equation satisfies the same recurrence (5) given for $h_n(\ell_1 < s)$. Indeed, $r(n, \lceil n/2 \rceil) = 0$ and $r(n,n) = (n-1)!$. Furthermore, assuming $\lceil n/2 \rceil < s < n$, a simple calculation shows that $r(n,s) = r(n,s-1) + 2(n-s+1)\, r(n-1,s-1) + (n-s)(n-s+1)\, r(n-2,s-1)$, where we note that all the factorials in $r(n,s-1), r(n-1,s-1)$, and $r(n-2,s-1)$ are well defined being of the form $m!$ with $m \geqslant 0$.

The next proposition summarizes our enumerative results from a probability point of view.

Proposition 2. *Let $n \geqslant 3$. If $p_n(\ell_1 = s)$ denotes the probability of $\ell_1 = s$ in an ordered history of size n selected uniformly at random, then*

$$p_n(\ell_1 = s) = \frac{(s-1)!(s-2)!(4ns + s - n^2 - n - 3s^2)}{(2s-n)!\,(n-1)!}, \tag{7}$$

where $\lceil n/2\rceil \leqslant s \leqslant n-1$. Furthermore, the joint probability $p_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k)$ of $\ell_1 = s_1, \ell_2 = s_2, dots,$ and $\ell_k = s_k$ in an ordered history of size n selected uniformly at random satisfies the recurrence

$$\begin{aligned} p_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k) &= \frac{2(n-s_1)}{n-1} p_{n-1}(\ell_1 = s_2, \ell_2 = s_3, \ldots, \ell_{k-1} = s_k) \\ &+ \frac{(n-s_1)(n-s_1-1)}{(n-1)(n-2)} \\ &\times p_{n-2}(\ell_1 = s_2, \ell_2 = s_3, \ldots, \ell_{k-1} = s_k), \end{aligned} \tag{8}$$

with initial condition given by (7).

Proof. Equation (7) follows from (6) as $p_n(\ell_1 = s) = [h_n(\ell_1 < s+1) - h_n(\ell_1 < s)]/(n-1)!$. The recurrence in (8) is obtained by replacing $n+1$ by n in (3) and dividing both sides of the resulting equation by $(n-1)!$. □

By summing over the possible values of $\ell_1, \ldots, \ell_{k-1}$ the joint probability $p_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k)$ yields for $k \geqslant 2$ the probability of $\ell_k = s_k$ in random ordered history of n leaves:

$$p_n(\ell_k = s_k) = \sum_{s_1 = s_k+k-1}^{n-1} \sum_{s_2 = s_k+k-2}^{s_1-1} \cdots \sum_{s_i = s_k+k-i}^{s_{i-1}-1} \cdots \sum_{s_{k-1} = s_k+1}^{s_{k-2}-1} p_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k). \tag{9}$$

For instance, if $k=2$, then we obtain

$$\begin{aligned} p_n(\ell_2 = s_2) &= \sum_{s_1=s_2+1}^{n-1} p_n(\ell_1 = s_1, \ell_2 = s_2) \\ &= \sum_{s_1=s_2+1}^{n-1} \frac{2(n-s_1)}{n-1} p_{n-1}(\ell_1 = s_2) + \frac{(n-s_1)(n-s_1-1)}{(n-1)(n-2)} p_{n-2}(\ell_1 = s_2) \\ &= \frac{2p_{n-1}(\ell_1 = s_2)}{n-1} \sum_{s_1=s_2+1}^{n-1} (n-s_1) + \frac{p_{n-2}(\ell_1 = s_2)}{(n-1)(n-2)} \sum_{s_1=s_2+1}^{n-1} (n-s_1)(n-s_1-1) \\ &= \frac{(s_2-2)!(s_2-1)!(n-s_2-1)(n-s_2)}{3(n-1)!(2s_2-n+2)!} \\ &\quad \times \left(2n^3 - n^2(13s_2+4) + n(s_2(26s_2+21)-2) - s_2((15s_2+23)s_2+2)+4\right), \end{aligned} \tag{10}$$

which can be used when $n \geqslant 5$ and s_2 is in the range $\lceil n/2\rceil - 1 \leqslant s_2 \leqslant n-2$. Similarly, if $k=3$, then we have

$$p_n(\ell_3 = s_3) = \sum_{s_1=s_3+2}^{n-1} \sum_{s_2=s_3+1}^{s_1-1} p_n(\ell_1 = s_1, \ell_2 = s_2, \ell_3 = s_3) \tag{11}$$

$$\begin{aligned}
&= \sum_{s_1=s_3+2}^{n-1} \sum_{s_2=s_3+1}^{s_1-1} \frac{2(n-s_1)}{n-1} p_{n-1}(\ell_1 = s_2, \ell_2 = s_3) \\
&\quad + \frac{(n-s_1)(n-s_1-1)}{(n-1)(n-2)} p_{n-2}(\ell_1 = s_2, \ell_2 = s_3) \\
&= \frac{4p_{n-2}(\ell_1 = s_3)}{(n-1)(n-2)} \sum_{s_1=s_3+2}^{n-1} \sum_{s_2=s_3+1}^{s_1-1} (n-s_1)(n-1-s_2) \\
&\quad + \frac{2p_{n-3}(\ell_1 = s_3)}{(n-1)(n-2)(n-3)} \\
&\quad \times \sum_{s_1=s_3+2}^{n-1} \sum_{s_2=s_3+1}^{s_1-1} (n-s_1)(n-s_2-2)(2n-2-s_2-s_1) \\
&\quad + \frac{p_{n-4}(\ell_1 = s_3)}{(n-1)(n-2)(n-3)(n-4)} \\
&\quad \times \sum_{s_1=s_3+2}^{n-1} \sum_{s_2=s_3+1}^{s_1-1} (n-s_1)(n-s_1-1)(n-2-s_2)(n-s_2-3),
\end{aligned}$$

which can be coupled with (7), when $n \geqslant 7$ and $\lceil n/2 \rceil - 2 \leqslant s_3 \leqslant n-3$.

4 Asymptotic distribution of the kth external branch length

In this section, we derive distributive properties of the random variable ℓ_k—the kth largest external branch length—considered over ordered histories of size n selected under the uniform distribution. We start by considering the case $k = 1$, and then generalize to arbitrary values of k.

By dividing Eq. (6) by the number $(n-1)!$ of ordered histories of size n, we obtain the probability

$$p_n(\ell_1 < s) = \frac{(s-1)!(s-2)!}{(2s-n-2)!(n-1)!}, \ \lceil n/2 \rceil < s \leqslant n,$$

or alternatively, with $u = s-1$,

$$p_n(\ell_1 \leqslant u) = \frac{u!(u-1)!}{(2u-n)!(n-1)!}, \ \lceil n/2 \rceil \leqslant u \leqslant n-1. \tag{12}$$

Our first result is the following local limit theorem.

Lemma 3. *When $n \to \infty$,*

(a) the probability $p_n(\ell_1 = \lfloor n - x\sqrt{n/2} \rfloor)$ admits an asymptotic expansion of the form

$$p_n(\ell_1 = \lfloor n - x\sqrt{n/2} \rfloor) = \frac{x}{\sqrt{n/2}} e^{-x^2/2}(1 + o(1)) + \mathcal{O}\left(\frac{e^{-x^2/2}}{n}\right)$$

uniformly for $0 \leqslant x \leqslant x^* \equiv n^{1/7}$.

(b) Furthermore,

$$p_n(\ell_1 \leqslant n - x^* \sqrt{n/2}) = \mathcal{O}\left(e^{-n^{2/7}/2}\right),$$

with x^* as defined in part (a).

Proof. For part (a), first assume that $x \leqslant x^*$ is such that $u \equiv n - x\sqrt{n/2}$ is a non-negative integer smaller than n. Then, Eq. (12) yields

$$p_n(\ell_1 = u) = p_n(\ell_1 \leqslant u) - p_n(\ell_1 \leqslant u-1) = \frac{(u-1)!(u-2)!(4nu + u - n^2 - n - 3u^2)}{(2u-n)(n-1)!};$$

see also Eq. (7). Using Stirling's formula $z! \sim z^z e^{-z}\sqrt{2\pi z}(1 + \frac{1}{12z} + \frac{1}{288z^2} - \frac{139}{51840z^3} - \cdots)$ and some tedious computation (which is best done with a computer algebra system) gives

$$p_n(\ell_1 = u) = \frac{x}{\sqrt{n/2}} e^{-x^2} 2 \left(1 + \mathcal{O}\left(\frac{|x| + |x|^3}{\sqrt{n}}\right)\right)$$

uniformly as $x = \mathcal{O}(n^{1/6})$. Thus, for the given range of x

$$\mathcal{O}\left(\frac{|x| + |x|^3}{\sqrt{n}}\right) = \mathcal{O}(n^{3/7-1/2}) = o(1).$$

This shows that the claimed expansion (without the last term) holds for this case. Note that the case $u = n$, i.e., $x = 0$, is trivially covered as $p_n(\ell_1 = n) = 0$.

Next, if u is not an integer, then $\lfloor u \rfloor = u + \mathcal{O}(1) = n - x\sqrt{n/2} + \mathcal{O}(1) = n - (x + \mathcal{O}(1/\sqrt{n}))\sqrt{n/2}$, and thus we are in the first case with x replaced by $\tilde{x} = x + \mathcal{O}(1/\sqrt{n})$. Hence,

$$\begin{aligned} p_n(\ell_1 = \lfloor u \rfloor) &= \frac{\tilde{x}}{\sqrt{n/2}} e^{-\tilde{x}^2/2}(1 + o(1)) = \frac{x + \mathcal{O}(1/\sqrt{n})}{\sqrt{n/2}} e^{-x^2/2 + o(1)}(1 + o(1)) \\ &= \frac{x}{\sqrt{n/2}} e^{-x^2/2}(1 + o(1)) + \mathcal{O}\left(\frac{e^{-x^2/2}}{n}\right), \end{aligned}$$

which establishes the claim also in this case.

For part (b), we are interested in $p_n(\ell_1 \leqslant \lfloor n - x^*\sqrt{n/2} \rfloor)$. Starting from (12), we use Stirling's approximation $\log(z) = z\log(z) - z + (1/2)\log(2\pi z) + o(1)$ to expand $\log(p_n(\ell_1 \leqslant u)) = \log(u!) + \log((u-1)!) - \log((2u-n)!) - \log((n-1)!)$ as $\frac{1}{2}(2(n-2u)\log(2u-n) - \log(2u-n) - 2n\log(n-1) + \log(n-1) + (2u-1)\log(u-1) + 2u\log(u) + \log(u)) + o(1)$. Then, we plug in $u = \lfloor n - x^*\sqrt{n/2} \rfloor = n - n^{1/7}\sqrt{n/2} - c_n$, where c_n is the fractional part of $n - n^{1/7}\sqrt{n/2}$, and replace the resulting terms of the form $\log(n + f(n))$ by $\log(n) + f(n)/n - f(n)^2/n^2$ (where $f(n)/n \to 0$). Simple algebraic manipulations finally give

$$\log(p_n(\ell_1 \leqslant \lfloor n - x^*\sqrt{n/2} \rfloor)) = -\frac{n^{2/7}}{2} + o(1),$$

which shows the claim. □

Let us denote by *Rayleigh*(λ) the Rayleigh distribution with parameter λ and the weak convergence of the sequence of random variables (X_n) to the variable X by the symbol $X_n \xrightarrow{d} X$. From the previous lemma, we obtain the following proposition that describes the asymptotic distribution of the random variable ℓ_1 considered over ordered histories of size n selected uniformly at random.

Proposition 4. *As $n \to \infty$,*

$$\frac{n-\ell_1}{\sqrt{n/2}} \xrightarrow{d} \text{Rayleigh}(1)$$

with convergence of all moments. In particular, the mean and the variance of ℓ_1 satisfy respectively

$$\mathbb{E}(\ell_1) \sim n \qquad \textit{and} \qquad \mathbb{V}(\ell_1) \sim \left(1-\frac{\pi}{4}\right) n. \tag{13}$$

Proof. Fix an $x \geqslant 0$. In order to prove the limit law, we have to show that, when $n \to \infty$, the probability of $(n-\ell_1)/\sqrt{n/2} \leqslant x$ converges to $1-e^{-x^2/2}$, which is the cumulative function of the Rayleigh distribution with parameter 1. We first write

$$\begin{aligned} p_n\left(\frac{n-\ell_1}{\sqrt{n/2}} \leqslant x\right) &= p_n(n - x\sqrt{n/2} \leqslant \ell_1) = p_n(\lceil n - x\sqrt{n/2}\rceil \leqslant \ell_1) \\ &= \sum_{s=\lceil n-x\sqrt{n/2}\rceil}^{n} p_n(\ell_1 = s) = \sum_{t=0}^{\tilde{x}} p_n(\ell_1 = n - t\sqrt{n/2}), \end{aligned} \tag{14}$$

where the latter sum is in steps of size $\sqrt{2/n}$ and $\tilde{x} = x + \mathcal{O}(1/\sqrt{n})$ is such that $n - \tilde{x}\sqrt{n/2} = \lceil n - x\sqrt{n/2}\rceil$. For n sufficiently large, we can assume $\tilde{x} \leqslant x \leqslant n^{1/7}$ and thus use part (a) of the lemma writing (14) as

$$\sum_{t=0}^{\tilde{x}} \frac{t}{\sqrt{n/2}} e^{-t^2/2}(1+o(1)) + \mathcal{O}\left(\frac{e^{-t^2/2}}{n}\right) = \sum_{t=0}^{\tilde{x}} \frac{t}{\sqrt{n/2}} e^{-t^2/2}(1+o(1)) + \sum_{t=0}^{\tilde{x}} \mathcal{O}\left(\frac{e^{-t^2/2}}{n}\right). \tag{15}$$

Because the $1+o(1)$ factor in the second sum of (15) holds uniformly, it can be put in front of the sum obtaining

$$\sum_{t=0}^{\tilde{x}} \frac{t}{\sqrt{n/2}} e^{-t^2/2}(1+o(1)) = (1+o(1)) \sum_{t=0}^{\tilde{x}} \frac{t}{\sqrt{n/2}} e^{-t^2/2} = (1+o(1)) \sum_{t=0}^{x} \frac{t}{\sqrt{n/2}} e^{-t^2/2} + o(1),$$

where the upper limit in the last sum is now x. Moreover, the third sum in (15) can be bounded as

$$\sum_{t=0}^{\tilde{x}} \mathcal{O}\left(\frac{e^{-t^2/2}}{n}\right) = \mathcal{O}\left(\sum_{t=0}^{\infty} \frac{e^{-t^2/2}}{n}\right) = o(1).$$

Hence, for $n \to \infty$, the probability $p_n\left(\frac{n-\ell_1}{\sqrt{n/2}} \leqslant x\right)$ converges to the Riemann sum $\sum_{t=0}^{x} \frac{t}{\sqrt{n/2}} e^{-t^2/2}$ with step size $dt = \sqrt{2/n}$, which can be approximated by the integral $\int_0^x te^{-t^2/2}\mathrm{d}t = 1 - e^{-x^2/2}$, as claimed.

By a similar approach, one can also show that all moments converge. Starting from

$$\mathbb{E}\left(\frac{n-\ell_1}{\sqrt{n/2}}\right)^m = \sum_{s=0}^{n}\left(\frac{n-s}{\sqrt{n/2}}\right)^m p_n(\ell_1 = s),$$

we replace s by $s = n - x\sqrt{n/2}$ and break the sum into two parts obtaining

$$\begin{aligned}&\sum_{x=0}^{\sqrt{2n}} x^m p_n(\ell_1 = n - x\sqrt{n/2})\\ &= \sum_{0 \leqslant x < n^{1/7}} x^m p_n(\ell_1 = n - x\sqrt{n/2}) + \sum_{n^{1/7} \leqslant x \leqslant \sqrt{2n}} x^m p_n(\ell_1 = n - x\sqrt{n/2}) \equiv \Sigma_1 + \Sigma_2,\end{aligned}$$

where all the sums proceed in steps of size $\sqrt{2/n}$. For Σ_2, by part (b) of the lemma, we have

$$\Sigma_2 = \mathcal{O}\left(n^{m/2} e^{-n^{2/7}/2}\right) = o(1).$$

For Σ_1, by part (a) of the lemma, we have

$$\Sigma_1 = (1 + o(1)) \sum_{0 \leqslant x < n^{1/7}} \frac{x^{m+1}}{\sqrt{n/2}} e^{-x^2/2} + \mathcal{O}\left(n^{-1} \sum_{0 \leqslant x < n^{1/7}} e^{-x^2/2}\right).$$

Here, the Riemann sum in Σ_1 can be approximated by the integral $\int_0^{n^{1/7}} x^{m+1} e^{-x^2/2}\mathrm{d}x$, which converges to $\int_0^\infty x^{m+1} e^{-x^2/2}\mathrm{d}x$. Overall,

$$\mathbb{E}\left(\frac{n-\ell_1}{\sqrt{n/2}}\right)^m \overset{n\to\infty}{\longrightarrow} \int_0^\infty x^{m+1} e^{-x^2/2}\mathrm{d}x$$

which proves the claimed convergence of moments. Finally, (13) follows from this convergence by straightforward computation. □

Note that when the limit distribution is uniquely determined by its moment sequence (which is the case for the Rayleigh distribution), convergence of all moments implies weak convergence. Although the second part of the proof of the latter proposition suffices to show that also the first claim holds true, we decided to provide the calculations for the convergence in distribution with the aim of improving the readability of the remaining part of the proof.

In the following, our goal is to show that, for an arbitrary fixed value of $k \geqslant 1$, the random variable ℓ_k follows asymptotically a χ distribution with $2k$ degrees of freedom.

Indeed, note that the Rayleigh distribution found for the case $k = 1$ is a χ distribution with 2 parameters.

The next lemma describes the solution to the recurrence (8) for the joint probability $p_n(\ell_1 = s_1, \ell_2 = s_2, \dots, \ell_k = s_k)$ and a formula for the probability $p_n(\ell_k = s_k)$ given in (9) in terms of the probability of $\ell_1 = s_k$ in trees of size smaller than or equal to n.

Lemma 5. *By setting $\mu_n(x) \equiv \frac{2x}{n-1}$ and $\nu_n(x) \equiv \frac{x(x-1)}{(n-1)(n-2)}$, we have*

$$\begin{aligned} &p_n(\ell_1 = s_1, \ell_2 = s_2, \dots, \ell_k = s_k) \\ &= \sum_{\omega} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell} \, (n - n_{\omega,\ell} - \ell - s_{\ell+1}) \right) p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k), \end{aligned} \tag{16}$$

where the sum runs over all words $\omega = \omega^{[0]} \cdots \omega^{[k-2]}$ of length $k-1$ with letters from the alphabet $\{\mu, \nu\}$, and $n_{\omega,\ell}$ is the number of ν in the first ℓ letters of ω (with $n_{\omega,0} = 0$). With the same notation, we also have

$$p_n(\ell_k = s_k) = \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \sum_{\omega} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1} - n_{\omega,\ell}) \right) p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k), \tag{17}$$

where $s_k^* \equiv n - k + 1 - s_k$.

Proof. For a fixed n and k, set $p'_i(j) \equiv p_{n-i}(\ell_1 = s_j, \dots, \ell_{k-j+1} = s_k)$, $\mu'_i(j) \equiv \frac{2(n-i-s_j)}{n-i-1}$, and $\nu'_i(j) \equiv \frac{(n-i-s_j)(n-i-1-s_j)}{(n-i-1)(n-i-2)}$. The recurrence (8) finds $p'_0(1) = p_n(\ell_1 = s_1, \ell_2 = s_2, \dots, \ell_k = s_k)$ by iteratively computing

$$p'_i(j) = \mu'_i(j) \, p'_{i+1}(j+1) + \nu'_i(j) \, p'_{i+2}(j+1). \tag{18}$$

The procedure ends after $k-1$ steps, that is, when we obtain terms of the form $p_{n-x}(\ell_1 = s_k) = p'_x(k)$, for a certain value of x. For $k = 4$, the diagram in Fig. 2 depicts the three iterations needed for evaluating $p'_0(1)$. The latter quantity is calculated as the sum of the probabilities at the bottom of the diagram, each multiplied by the sum of the words of length $k-1$ over the alphabet $\{\mu', \nu'\}$ that encode the different paths connecting the corresponding leaf node to the root of the diagram. More precisely, for arbitrary values of n and k, we have

$$p'_0(1) = \sum_{\omega} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n_{\omega,\ell}+\ell} \, (\ell+1) \right) p'_{n_{\omega,k-1}+k-1}(k),$$

where the sum runs over all words $\omega = \omega^{[0]} \cdots \omega^{[k-2]}$ of length $k-1$ with letters from the alphabet $\{\mu', \nu'\}$, and $n_{\omega,\ell}$ is the number of ν' in the first ℓ letters of ω (with $n_{\omega,0} = 0$). By replacing indices, the latter formula is equivalent to that claimed in (16).

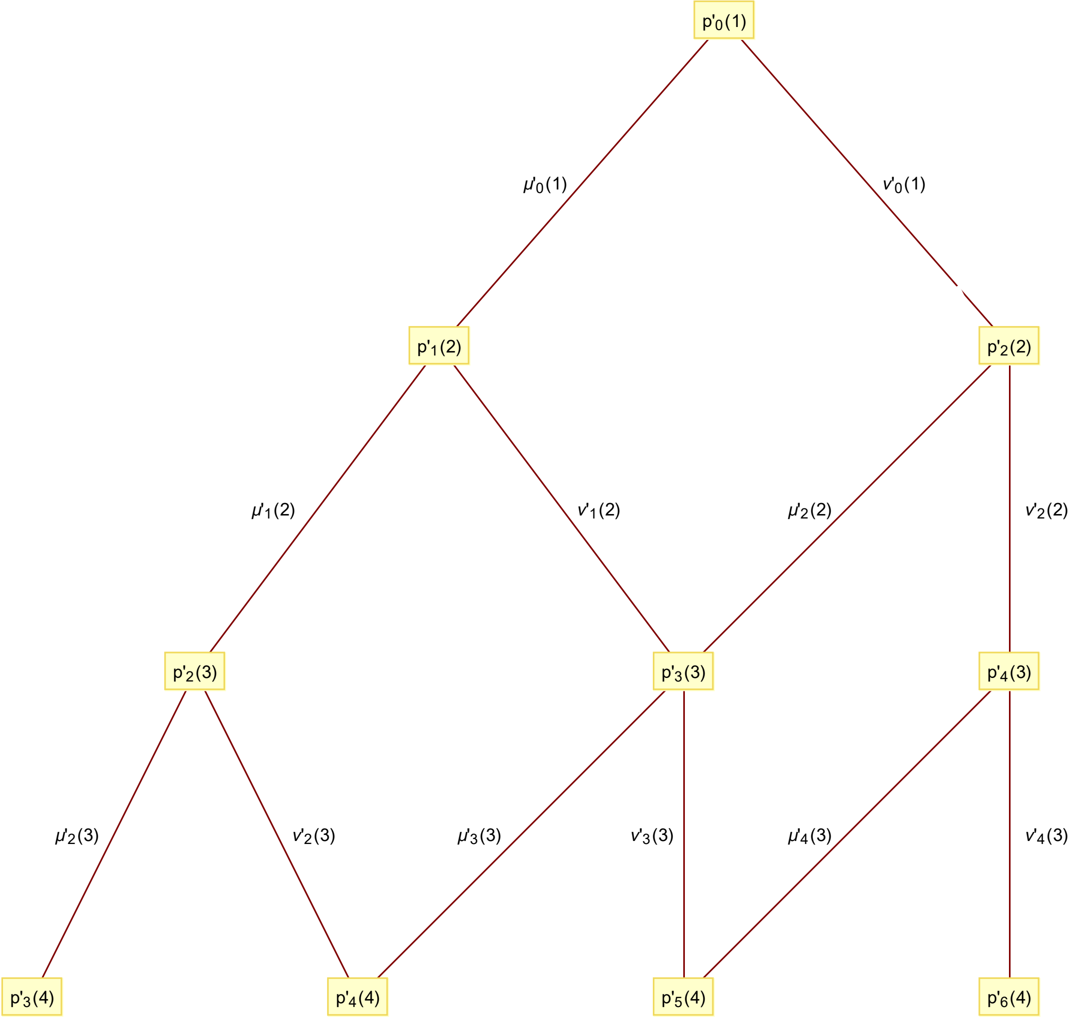

Figure 2: Schematic diagram of the first three iterative steps of the procedure (18) for calculating $p_0'(1) = p_n(\ell_1 = s_1, \ell_2 = s_2, \ldots, \ell_k = s_k)$.

Finally, plugging (16) into (9) yields

$$p_n(\ell_k = s_k) = \sum_{s_1=s_k+k-1}^{n-1} \sum_{s_2=s_k+k-2}^{s_1-1} \cdots \sum_{s_{k-1}=s_k+1}^{s_{k-2}-1} \sum_{\omega} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(n - n_{\omega,\ell} - \ell - s_{\ell+1}) \right) \\ \times p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k).$$

By setting $s^*_\ell = n - \ell + 1 - s_\ell$ for $\ell = 1, \ldots, k$, the right-hand side can be written as

$$\sum_{s^*_1=1}^{s^*_k} \sum_{s^*_2=s^*_1}^{s^*_k} \cdots \sum_{s^*_{k-1}=s^*_{k-2}}^{s^*_k} \sum_{\omega} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s^*_{\ell+1} - n_{\omega,\ell}) \right) p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k),$$

which gives (17). □

With the same notation used above, we now provide two more useful lemmas.

Lemma 6. *For $s_k = \lfloor n - x\sqrt{n/2} \rfloor$, we have*

$$\sum_{s_1=1}^{s^*_k} \sum_{s_2=s_1}^{s^*_k} \cdots \sum_{s_{k-1}=s_{k-2}}^{s^*_k} \prod_{\ell=0}^{k-2} \mu_{n-\ell}(s_{\ell+1}) = \frac{x^{2k-2}}{2^{k-1}(k-1)!} + \mathcal{O}\left(\frac{1 + x^{2k-3}}{\sqrt{n}} \right)$$

uniformly for $0 \leqslant x \leqslant \sqrt{2n}$.

Proof. Note that

$$\begin{aligned} \sum_{s_1=1}^{s^*_k} \sum_{s_2=s_1}^{s^*_k} \cdots \sum_{s_{k-1}=s_{k-2}}^{s^*_k} \prod_{\ell=0}^{k-2} \mu_{n-\ell}(s_{\ell+1}) &= \frac{2^{k-1} \sum_{s_1=1}^{s^*_k} s_1 \sum_{s_2=s_1}^{s^*_k} s_2 \cdots \sum_{s_{k-1}=s_{k-2}}^{s^*_k} s_{k-1}}{(n-1)\cdots(n-k+1)} \\ &= \frac{2^{k-1} r(s^*_k)}{n^{k-1}} + \mathcal{O}\left(\frac{r(s^*_k)}{n^k} \right), \end{aligned}$$

where $r(z)$ is the polynomial $r(z) \equiv \sum_{s_1=1}^{z} s_1 \sum_{s_2=s_1}^{z} s_2 \cdots \sum_{s_{k-1}=s_{k-2}}^{z} s_{k-1}$. In order to determine the asymptotic behavior of $r(z)$, we rely on Faulhaber's formula:

$$\sum_{m=1}^{N} m^t = \frac{1}{t+1} \sum_{k=0}^{t} \binom{t+1}{k} B_k (N+1)^{t+1-k} \overset{N\to\infty}{\sim} \frac{N^{t+1}}{t+1} \overset{N\to\infty}{\sim} \int_1^N x^t \mathrm{d}x, \tag{19}$$

where B_k denotes the k-th Bernoulli number. In particular, we use the fact that, for a given polynomial $p(u) = \alpha_k u^k + \cdots + \alpha_1 u + \alpha_0$, the polynomial $\sum_{u=a}^{b} p(u) = \sum_{u=1}^{b} p(u) - \sum_{u=1}^{a-1} p(u)$ has its term $\frac{\alpha_k b^{k+1}}{k+1}$ with the highest power in b and its term $-\frac{\alpha_k a^{k+1}}{k+1}$ with the highest power in a matching those that appear in the integral $\int_a^b p(z)\mathrm{d}z$. As a consequence, if we substitute each sum in $r(z)$ by an integral sign, we then find a polynomial

$\int_1^z s_1 \mathrm{d}s_1 \int_{s_1}^z s_2 \mathrm{d}s_2 \cdots \int_{s_{k-2}}^z s_{k-1}\, \mathrm{d}s_{k-1}$ with the same leading term of $r(z)$. Furthermore, by a simple induction on k one can show that

$$\int_{z_{k+1}}^z z_k \mathrm{d}z_k \cdots \int_{z_3}^z z_2 \mathrm{d}z_2 \int_{z_2}^z z_1 \mathrm{d}z_1 = \frac{1}{2^k} \sum_{i=0}^{k} \frac{(-1)^i z^{2k-2i} z_{k+1}^{2i}}{i!(k-i)!},$$

and therefore the leading term of $r(z)$ is that of $\frac{1}{2^{k-1}} \sum_{i=0}^{k-1} \frac{(-1)^i z^{2k-2-2i}}{i!(k-1-i)!}$, that is, $\frac{x^{2k-2}}{2^{k-1}(k-1)!}$. Hence,

$$\sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \prod_{\ell=0}^{k-2} \mu_{n-\ell}(s_{\ell+1}) = \frac{(s_k^*)^{2k-2}}{n^{k-1}(k-1)!} + \mathcal{O}\left(\frac{(s_k^*)^{2k-3}}{n^{k-1}} + \frac{r(s_k^*)}{n^k} \right)$$

By plugging $s_k^* = x\sqrt{n/2} + \mathcal{O}(1)$ into the latter asymptotic formula and performing a straightforward expansion, we obtain the claimed result. □

The next result shows that Lemma 6 gives the main term of the multiple sum in (17).

Lemma 7. *For $s_k = \lfloor n - x\sqrt{n/2} \rfloor$, we have*

$$\sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \prod_{\ell=0}^{k-2} \omega_{n-n_{\omega,\ell}-\ell}^{[\ell]}(s_{\ell+1} - n_{\omega,\ell}) = \mathcal{O}\left(\frac{1 + x^{2k-1}}{\sqrt{n}} \right)$$

uniformly for $0 \leqslant x \leqslant \sqrt{2n}$ and for all words $\omega = \omega^{[0]} \cdots \omega^{[k-2]}$ of length $k-1$ with letters from the alphabet $\{\mu, \nu\}$ different from the word whose letters are all equal to μ.

Proof. Assume that ω has $m \geqslant 1$ letters equal to ν. Then, since $\nu_n(x)$ is a quadratic polynomial, by again using Faulhaber's formula (19), we obtain that

$$\sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \prod_{\ell=0}^{k-2} \omega_{n-n_{\omega,\ell}-\ell}^{[\ell]}(s_{\ell+1} - n_{\omega,\ell}) = \frac{r(s_k^*)}{q(n)},$$

where $r(z)$ is a polynomial of degree $m+2k-2$ and $q(z)$ is a polynomial of degree $m+k-1$. Thus, by setting $s_k^* = x\sqrt{n/2} + \mathcal{O}(1)$, we obtain that

$$\sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \prod_{\ell=0}^{k-2} \omega_{n-n_{\omega,\ell}-\ell}^{[\ell]}(s_{\ell+1} - n_{\omega,\ell}) = \frac{r(s_k^*)}{q(n)} = \mathcal{O}\left(\frac{1 + x^{m+2k-2}}{n^{m/2}} \right).$$

From this the result follows by observing that $x \leqslant \sqrt{2n}$. □

From the last three lemmas, we can now deduce the following generalization of Lemma 3.

Corollary 8. *When $n \to \infty$,*

(a) the probability $p_n(\ell_k = \lfloor n - x\sqrt{n/2}\rfloor)$ admits an asymptotic expansion of the form

$$p_n(\ell_k = \lfloor n - x\sqrt{n/2}\rfloor) = \frac{x^{2k-1}}{2^{k-1}(k-1)!\sqrt{n/2}} e^{-x^2/2}(1+o(1)) + \mathcal{O}\left(\frac{e^{-x^2/2}}{n}\right)$$

uniformly for $0 \leqslant x \leqslant x^* \equiv n^{1/7}$.

(b) Furthermore,

$$p_n(\ell_k \leqslant n - x^*\sqrt{n/2}) = \mathcal{O}\left(n^{k-1}e^{-n^{2/7}/2}\right)$$

with x^* as defined in part (a).

Proof. First, note that for any given word ω of length $k-1$ over the alphabet $\{\mu,\nu\}$ (in the sense of Lemma 5), we have

$$p_{n-n_\omega-k+1}(\ell_1 = \lfloor n - x\sqrt{n/2}\rfloor) = p_{n-n_\omega-k+1}(\ell_1 = \lfloor n - n_\omega - k + 1 - \tilde{x}\sqrt{(n-n_\omega-k+1)/2}\rfloor),$$

where $\tilde{x} = x + \mathcal{O}(1/\sqrt{n})$. As a consequence, by applying part (a) of Lemma 1 with x replaced by $\tilde{x}$ and n replaced by $n - n_\omega - k + 1$, it follows that part (a) of Lemma 3 also holds when p_n is replaced by $p_{n-n_\omega-k+1}$. Moreover, also part (b) of Lemma 3 holds true when p_n is replaced by $p_{n-n_\omega-k+1}$. Indeed, from (12), we find

$$\begin{aligned} p_{n-n_\omega-k+1}(\ell_1 \leqslant n^*) &= \frac{n^*!(n^*-1)!}{(2n^*-n+n_\omega+k-1)!(n-n_\omega-k)!} \\ &= \frac{(n-1)\cdots(n-n_\omega-k+1)}{(2n^*-n+n_\omega+k-1)\cdots(2n^*-n+1)} \cdot \frac{n^*!(n^*-1)!}{(2n^*-n)!(n-1)!} \\ &= \mathcal{O}(p_n(\ell_1 \leqslant n^*)), \end{aligned}$$

where $n_\omega \equiv n_{\omega,k-1}$ and $n^* \equiv \lfloor n - x^*\sqrt{n/2}\rfloor$.

In order to prove part (a) of the corollary, assume $0 \leqslant x \leqslant x^*$ and set $s_k = \lfloor n - x\sqrt{n/2}\rfloor$. From (17), we find

$$\begin{aligned} p_n(\ell_k = s_k) = \sum_{s_1=1}^{s_k^*}\sum_{s_2=s_1}^{s_k^*}\cdots\sum_{s_{k-1}=s_{k-2}}^{s_k^*}\Bigg[&\prod_{\ell=0}^{k-2}\mu_{n-\ell}(s_{\ell+1})p_{n-k+1}(\ell_1 = s_k) \\ &+ \sum_{\omega\neq\mu\mu\cdots\mu}\left(\prod_{\ell=0}^{k-2}\omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1}-n_{\omega,\ell})\right)p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k)\Bigg]. \end{aligned}$$

Then, the expansion of Lemma 1 for the factors $p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k)$ coupled with Lemmas 6 and 7 yield

$$p_n(\ell_k = s_k) = \left[\frac{x}{\sqrt{n/2}}e^{-x^2/2}(1+o(1)) + \mathcal{O}\left(\frac{e^{-x^2/2}}{n}\right)\right]$$

$$
\begin{aligned}
&\times \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \left[\prod_{\ell=0}^{k-2} \mu_{n-\ell}(s_{\ell+1}) + \sum_{\omega\neq\mu\mu\cdots\mu} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1}-n_{\omega,\ell})\right)\right] \\
&= \left[\frac{x}{\sqrt{n/2}} e^{-x^2/2}(1+o(1)) + \mathcal{O}\left(\frac{e^{-x^2/2}}{n}\right)\right] \\
&\quad \times \left[\frac{x^{2k-2}}{2^{k-1}(k-1)!} + \mathcal{O}\left(\frac{1+x^{2k-3}}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{1+x^{2k-1}}{\sqrt{n}}\right)\right] \\
&= \frac{x^{2k-1}}{2^{k-1}(k-1)!\sqrt{n/2}} e^{-x^2/2}(1+o(1)) + \mathcal{O}\left(\frac{e^{-x^2/2}}{n}\right),
\end{aligned}
$$

as claimed in (a).

For part (b) we can write $p_n(\ell_k \leqslant n - x^*\sqrt{n/2}) = \sum_x p_n(\ell_k = \lfloor n - x\sqrt{n/2}\rfloor) = \sum_x p_n(\ell_k = s_k)$, where the sum proceeds in steps of $\sqrt{2/n}$ over the range $x^* \leqslant x \leqslant \sqrt{2n}$ and we set $s_k = \lfloor n - x\sqrt{n/2}\rfloor$. Hence, by using (17) together with Lemmas 6 and 7, we obtain

$$
\begin{aligned}
&p_n(\ell_k \leqslant n - x^*\sqrt{n/2}) \\
&= \sum_x \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \sum_{\omega} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1}-n_{\omega,\ell})\right) p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k) \\
&= \sum_{\omega} \sum_x \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1}-n_{\omega,\ell})\right) p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k) \\
&= \sum_x \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \left(\prod_{\ell=0}^{k-2} \mu_{n-\ell}(s_{\ell+1})\right) p_{n-k+1}(\ell_1 = s_k) \\
&\quad + \sum_{\omega\neq\mu\mu\cdots\mu} \sum_x \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1}-n_{\omega,\ell})\right) p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k) \\
&= \sum_x p_{n-k+1}(\ell_1 = s_k) \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \left(\prod_{\ell=0}^{k-2} \mu_{n-\ell}(s_{\ell+1})\right) \\
&\quad + \sum_{\omega\neq\mu\mu\cdots\mu} \sum_x p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k) \sum_{s_1=1}^{s_k^*} \sum_{s_2=s_1}^{s_k^*} \cdots \sum_{s_{k-1}=s_{k-2}}^{s_k^*} \left(\prod_{\ell=0}^{k-2} \omega^{[\ell]}_{n-n_{\omega,\ell}-\ell}(s_{\ell+1}-n_{\omega,\ell})\right) \\
&= \sum_x p_{n-k+1}(\ell_1 = s_k) \left[\frac{x^{2k-2}}{2^{k-1}(k-1)!} + \mathcal{O}\left(\frac{1+x^{2k-3}}{\sqrt{n}}\right)\right] \\
&\quad + \sum_{\omega\neq\mu\mu\cdots\mu} \sum_x p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k) \left[\mathcal{O}\left(\frac{1+x^{2k-1}}{\sqrt{n}}\right)\right].
\end{aligned}
$$

Finally, since $x \leqslant \sqrt{2n}$, we have

$$\begin{aligned}
p_n(\ell_k \leqslant n - x^*\sqrt{n/2}) &= \mathcal{O}\left(n^{k-1}\right) \sum_{x} p_{n-k+1}(\ell_1 = s_k) \\
&\quad + \mathcal{O}\left(n^{k-1}\right) \sum_{\omega \neq \mu\mu\cdots\mu} \sum_{x} p_{n-n_{\omega,k-1}-k+1}(\ell_1 = s_k) \\
&= \mathcal{O}\left(n^{k-1}\right) p_{n-k+1}(\ell_1 \leqslant n^*) \\
&\quad + \mathcal{O}\left(n^{k-1}\right) \sum_{\omega \neq \mu\mu\cdots\mu} p_{n-n_{\omega,k-1}-k+1}(\ell_1 \leqslant n^*) \\
&= \mathcal{O}\left(n^{k-1}\right) \mathcal{O}\left(e^{-n^{2/7}/2}\right) = \mathcal{O}\left(n^{k-1} e^{-n^{2/7}/2}\right). \qquad \square
\end{aligned}$$

The next theorem, which extends Proposition 4, is our main result.

Theorem 9. *For a fixed $k \geqslant 1$, let ℓ_k be the kth largest external branch length in a random ordered history of size n selected uniformly at random and denote by $\chi(2k)$ the χ-distribution with $2k$ degrees of freedom. Then, as $n \to \infty$,*

$$\frac{n - \ell_k}{\sqrt{n/2}} \xrightarrow{d} \chi(2k),$$

with convergence of all moments. In particular, the mean and the variance of ℓ_k satisfy respectively

$$\mathbb{E}(\ell_k) \sim n \qquad \text{and} \qquad \operatorname{Var}(\ell_k) \sim \left(k - \frac{\pi k^2}{16^k}\binom{2k}{k}^2\right) n. \tag{20}$$

Proof. Following the proof of Proposition 4, we show that all moments converge, which implies convergence in distribution. Starting from

$$\mathbb{E}\left(\frac{n - \ell_k}{\sqrt{n/2}}\right)^m = \sum_{s=0}^{n} \left(\frac{n - s}{\sqrt{n/2}}\right)^m p_n(\ell_k = s),$$

we replace s by $s = n - x\sqrt{n/2}$ and break the sum into two parts obtaining

$$\begin{aligned}
\sum_{x=0}^{\sqrt{2n}} x^m p_n(\ell_k = n - x\sqrt{n/2}) &= \sum_{0 \leqslant x < n^{1/7}} x^m p_n(\ell_k = n - x\sqrt{n/2}) \\
&\quad + \sum_{n^{1/7} \leqslant x \leqslant \sqrt{2n}} x^m p_n(\ell_k = n - x\sqrt{n/2}) \equiv \Sigma_1 + \Sigma_2,
\end{aligned}$$

where all the sums proceed in steps of size $\sqrt{2/n}$. For Σ_2, by part (b) of the latter corollary, we have

$$\Sigma_2 = \mathcal{O}\left(n^{m/2+k-1} e^{-n^{2/7}/2}\right) = o(1).$$

For Σ_1, by part (a) of Corollary 8, we have

$$\Sigma_1 = \frac{1+o(1)}{2^{k-1}(k-1)!} \cdot \sum_{0 \leqslant x < n^{1/7}} \frac{x^{m+2k-1}}{\sqrt{n/2}} e^{-x^2/2} + \mathcal{O}\left(n^{-1} \sum_{0 \leqslant x < n^{1/7}} e^{-x^2/2} \right).$$

Hence, the Riemann sum in Σ_1 can be approximated by the integral $\int_0^{n^{1/7}} x^{m+2k-1} e^{-x^2/2} \mathrm{d}x$, which converges to $\int_0^\infty x^{m+2k-1} e^{-x^2/2} \mathrm{d}x$. Overall,

$$\begin{aligned} \mathbb{E}\left(\frac{n-\ell_k}{\sqrt{n/2}} \right)^m &\overset{n\to\infty}{\longrightarrow} \frac{1}{2^{k-1}(k-1)!} \int_0^\infty x^{m+2k-1} e^{-x^2/2} \mathrm{d}x \\ &= \frac{1}{2^{k-1}(k-1)!} \cdot 2^{m/2+k-1}\, \Gamma\left(\frac{m+2k}{2} \right) = 2^{m/2} \frac{\Gamma\left(\frac{m}{2}+k\right)}{\Gamma(k)} \end{aligned}$$

which proves the claimed convergence of moments. Finally, (20) follows from this convergence by straightforward computation. For instance, setting $m = 1$ we obtain

$$\frac{n - \mathbb{E}(\ell_k)}{\sqrt{n/2}} \overset{n\to\infty}{\longrightarrow} \frac{\sqrt{2\pi}k\binom{2k}{k}}{4^k}, \tag{21}$$

and similarly for the variance. □

5 Conclusions

For random histories of fixed size n selected under the Yule probability model, or, equivalently, for ordered histories of size n selected uniformly at random, we have studied the variable ℓ_k defined as the kth largest length of an external branch. Measuring the length of an external branch as the rank of its parent node, Theorem 9 shows that the rescaled variable $\mathcal{L}_k \equiv \frac{n-\ell_k}{\sqrt{n/2}}$ follows asymptotically a χ-distribution with $2k$ degrees of freedom (Fig. 3), with convergence of all moments. The mean of ℓ_k is shown to be asymptotically equivalent to n, independently of k. More precisely, by plugging the approximation $\binom{2k}{k} \approx \frac{4^k}{\sqrt{\pi k}}$ into (21), we find that $\mathbb{E}(\ell_k)$ behaves like $n - \sqrt{k\,n}$ for increasing n. The variance of ℓ_k is asymptotically equivalent to $\left(k - \frac{\pi k^2}{16^k}\binom{2k}{k}^2\right) n$.

Our approach has used a well known correspondence between trees and permutations, in which the kth largest length of an external branch of an ordered history of size n is the kth largest non-peak value in the associated permutation of size $n-1$ (Section 2). Thus, Proposition 2 and Theorem 9 also contribute to the study of the probabilistic properties of the value-peaks of permutations investigated in [4].

In this paper we focused only on the *discrete* length of the external branches of random trees. Nevertheless, our results can also find applications in the analysis of the *time* length of the external branches of trees generated under the "coalescent" [15, 17, 24] and "Yule" [14, 25] processes. A coalescent tree of size n is a pair consisting of a random Yule history

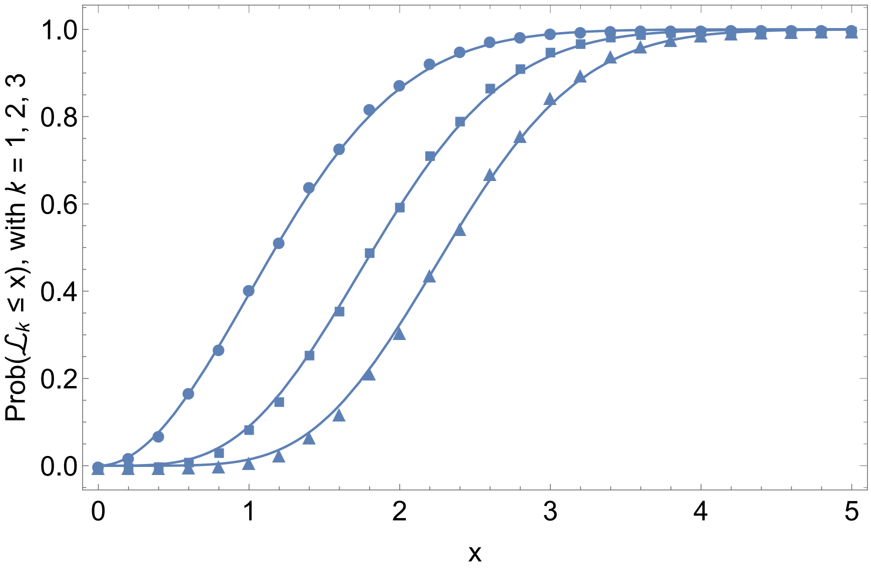

Figure 3: Probability that for $n = 1000$ the rescaled variable $\mathcal{L}_k \equiv \frac{n-\ell_k}{\sqrt{n/2}}$ is less than or equal to $x \in [0,5]$ (in steps of 0.2), when $k = 1$ (dots), $k = 2$ (squares), and $k = 3$ (triangles). Values are calculated from Eqs. (7), (10), and (11). Solid lines give the cumulative function for the χ-distribution with $2k$ degrees of freedom, with $k = 1,2,3$ from left to right.

t of n leaves and a sequence $(\tau_2, \ldots, \tau_n)$ of independent exponentially distributed random variables assigning a time length to the different layers of t (Fig. 1). The variable τ_i gives the time length of the layer in which exactly i branches of t coexist, and its mean is $\mathbb{E}(\tau_i) = 1/\lambda_i$, with $\lambda_i = \binom{i}{2}$. Hence, the expected value of the time length of an external branch of t of discrete length s can be calculated as $\sum_{i=n+1-s}^{n} \mathbb{E}(\tau_i) = \frac{2}{n-s} - \frac{2}{n}$. By using our finding that $\mathbb{E}(\ell_k) \approx n - \sqrt{k\,n}$, we thus see that, in a random coalescent tree of large size n, the mean of the kth time length of an external branch will behave roughly like $\frac{2}{\sqrt{k\,n}}$. In a Yule generated tree with speciation rate λ and time depth t, the variable τ_i is exponentially distributed with mean $1/(\lambda \cdot i)$ and the expected number of leaves in the tree is $n = e^{\lambda\, t}$. Under this setting, Bocharov *et al.* [3] study the time length L_t of the longest external branch showing (among other things) that L_t/t converges to $1/2$ in mean and probability when $t \to \infty$. This is in agreement with our observation that the discrete length ℓ_1 is on average close to $n - \sqrt{n}$. Indeed, if an external branch spans $n - \sqrt{n}$ layers of the tree, then the mean of its time length can be calculated as above by summing the expectations of the variables $\tau_n, \tau_{n-1}, \ldots, \tau_{n+1-n+\sqrt{n}}$, which gives $L_t \approx \frac{1}{\lambda}(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{\sqrt{n}+1}) \approx \frac{1}{\lambda}(\log n - \log\sqrt{n}) = \frac{1}{\lambda}\log\sqrt{n} \approx \frac{1}{\lambda}\log\sqrt{e^{\lambda t}} = \frac{t}{2}$.

In Section 6 of [9], the number of mutations seen in a single individual of different human and zebrafish populations is analyzed within the neutral scenario of coalescent trees by means of Yule histories. Singleton mutations—i.e. mutations affecting single individuals—can be modeled as random events occurring along the external branches of the tree. Doubleton mutations—which affect pairs of individuals—take place along those branches of the tree from which exactly two leaves descend. It would be of interest to broaden the calculations of this article to investigate the length of the longest tree edge connecting the root node of a cherry subtree to its parent node.

It would also be nice to extend the results of [23] on the time length of a random branch and a random external branch in a tree generated under the Yule process to the discrete setting of Yule histories.

Acknowledgements

Support to MF was provided by the NSTC (National Science and Technology Council, Taiwan) grant NSTC-111-2115-M-004-002-MY2.

References

[1] M. G. B. Blum, O. François, *Minimal clade size and external branch length under the neutral coalescent*, Adv. Appl. Probab. 37 (2005): 647–662.

[2] M. G. B. Blum, O. François, *On statistical tests for phylogenetic tree imbalance: The Sackin and other indices revisited*, Math. Biosci. 195 (2005): 141–153.

[3] S. Bocharov, S. Harris, E. Kominek, A. Mooers, M. Steel, *Predicting long pendant edges in model phylogenies, with applications to biodiversity and tree inference*, Syst. Biol. (2022): article syac059.

[4] P. Bouchard, H. Chang, J. Ma, J. Yeh, Y. N. Yeh, *Value-peaks of permutations*, Electron. J. Comb. 17 (2010): # R46.

[5] A. Caliebe, R. Neininger, M. Krawczak, U. Rösler, *On the length distribution of external branches in coalescence trees: genetic diversity within species*, Theor. Popul. Biol. 72 (2007): 245–252.

[6] I. Dahmer, G. Kersting, *The internal branch length of the Kingman coalescent*, Ann. Appl. Probab. 25 (2015): 1325–1348.

[7] C. Diehl, G. Kersting, *External branch lengths of Λ-coalescents without a dust component*, Electron. J. Probab. 24 (2019): 1–36.

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

[9] F. Disanto, T. Wiehe, *Measuring the external branches of a Kingman tree: A discrete approach*, Theor. Popul. Biol. 134 (2020): 92–105.

[10] J. Felsenstein J, *Inferring phylogenies*, Sinauer, Sunderland (2004).

[11] F. Freund, M. Möhle, *On the time back to the most recent common ancestor and the external branch length of the Bolthausen-Sznitman coalescent*, Markov Process. Related Fields 15 (2009): 387–416.

[12] Y. X. Fu, *Statistical tests of neutrality of mutations*, Genetics 133 (1993): 693–709.

[13] I. P. Goulden, D. M. Jackson, *Combinatorial Enumeration*, Wiley, Chichester (1983).

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

[15] R. R. Hudson, *Gene genealogies and the coalescent process*, Oxf. Surv. Evol. Biol. 7 (1990): 1–44.

[16] S. Janson, G. Kersting, *On the total external length of the Kingman coalescent*, Electron. J. Probab. 16 (2011): 2203–2218.

[17] J. F. C. Kingman, *The coalescent*, Stoch. Proc. Appl. 13 (1982): 235–248.

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

[19] R. Nielsen, M. Slatkin, *An Introduction to Population Genetics: Theory and Applications* Sinauer Associates, Sunderland, Massachusetts (2013).

[20] 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 (2006): 129–146.

[21] M. J. Sanderson, *How many taxa must be sampled to identify the root node of a large clade?*, Syst. Biol. 45 (1996): 168–173.

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

[23] M. Steel, A. Mooers, *The expected length of pendant and interior edges of a Yule tree*, Appl. Math. Lett. 23 (2010): 1315–1319.

[24] F. Tajima, *Evolutionary relationship of DNA sequences in finite populations*, Genetics 105 (1983): 437–460.

[25] G. U. Yule, *A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S.*, Philos. Trans. Roy. Soc. Lond. Ser. B 213 (1924): 21–87.