Annals of Combinatorics 10 (2006) 129-146
0218-0006/06/010129-18
DOI 10.1007/s00026-006-0278-6

Annals of Combinatorics

The Mean and Variance of the Numbers of r-Pronged Nodes and r-Caterpillars in Yule-Generated Genealogical Trees

Noah A. Rosenberg

Department of Human Genetics and Bioinformatics Program, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave, Ann Arbor, MI 48109-2218, USA
rnoah@umich.edu

Received August 23, 2004

AMS Subject Classification: 05C05, 92D15

Abstract. The Yule model is a frequently-used evolutionary model that can be utilized to generate random genealogical trees. Under this model, using a backwards counting method differing from the approach previously employed by Heard (*Evolution* 46: 1818-1826), for a genealogical tree of n lineages, the mean number of nodes with exactly r descendants is computed ($2 \leq r \leq n-1$). The variance of the number of r-pronged nodes is also obtained, as are the mean and variance of the number of r-caterpillars. These results generalize computations of McKenzie and Steel for the case of $r = 2$ (*Math. Biosci.* 164: 81-92, 2000). For a given n, the two means are largest at $r = 2$, equaling $2n/3$ for $n \geq 5$. However, for $n \geq 9$, the variances are largest at $r = 3$, equaling $23n/420$ for $n \geq 7$. As $n \to \infty$, the fraction of internal nodes that are r-caterpillars for some r approaches $(e^2 - 5)/4 \approx 0.59726$.

Keywords: binary search tree, cherries, coalescent, genealogy, labeled topology, pectinate

1. Introduction

In many contexts in evolutionary biology, it is of interest to investigate the probability distributions of various attributes of genealogical trees. First, predictions about these distributions can be made using models of the evolutionary process that produces the trees. These predictions can then be helpful in understanding which scenarios are possible or likely outcomes of evolution [2,5,16,22,26,29]. By comparison with estimates made from biological data, they can also provide insight into the nature of the processes that generate the data [9,19,20,23,34].

Perhaps the simplest evolutionary model from which predictions about genealogical trees can be made is the Yule or Yule-Harding model [4,12,31,35,36,39]. Under this model, beginning with an ancestral lineage, the genealogical tree for n lineages is formed by successive binary branching events, so that at any point in time, all lineages have equal probability of being the next to branch into two. Equivalently, looking

backwards from n lineages in the present, at any time point, all pairs of lineages have equal probability of being the next to "coalesce" into one. This retrospective viewpoint is typically adopted in population genetics, where the Yule model is used with each lineage corresponding to a distinct copy of a particular genetic site (taken from a set of such copies in a population of individuals). In this context, when combined with a specific model for the times at which "coalescences" occur, the Yule model is termed the *coalescent model* [14, 15, 24, 38].

Here, using a backwards counting approach, we extend known properties of genealogical trees under the Yule model to obtain the mean and variance of the number of nodes with exactly r descendants among the n lineages ($2 \leq r \leq n-1$). These quantities then enable computation of the mean and variance of the number of *r-caterpillars* in genealogical trees.

2. Definitions

The definitions used here are largely based on those of Semple and Steel [32]; Figure 1 illustrates many of the key concepts.

A *genealogical tree* or *genealogy* $\mathcal{G} = (G, X, \psi, t)$ for n leaves is a rooted binary tree G for which (1) ψ is a bijection that associates each leaf of G with a label in a label set X, and (2) t is a map that associates each point p of G (that is, each vertex and each point lying on an edge) with a nonnegative real number $t(p)$, such that (i) for any two distinct points p_1, p_2, if the path from p_1 to the root of G includes p_2, then $t(p_1) < t(p_2)$; (ii) for any two distinct interior vertices $v_1, v_2, t(v_1) \neq t(v_2)$; (iii) for a point $p, t(p) = 0$ if and only if p is a leaf of G.

If the path from a point p_1 to the root includes p_2, then p_1 is *descended* from p_2, which, in turn, is *ancestral* to p_1. Trivially, a point is both ancestral to and descended from itself (however, the point itself is not included when counting its number of descendants). For convenience the label set X is taken to be $\{x_1, x_2, ..., x_n\}$. Interior vertices are alternately termed *internal nodes* (it is assumed in a genealogy that each internal node has exactly two descendants). The value of $t(p)$ for a point p is the *time* of p. Thus, the leaves of genealogies are viewed as existing in the present, with time increasing into the past. The *lineage* of a point p at time $u \geq t(p)$ is the unique point that is both ancestral to p and has time u; this time is usually 0 in uses of the term. The *most recent common ancestor* (MRCA) of a subtree G' of G (or a subset $X' \subset X$) is the node with the smallest time among the collection of nodes that are ancestral to all elements of G' (or to the collection of leaves with label set X').

Let $t(G)$ be the set of values taken by t over all internal nodes of G. Let h be the unique bijection from $t(G)$ into $\{1, 2, ..., n-1\}$ with the property that for any two vertices v_1, v_2, if $t(v_1) < t(v_2)$, then $h(v_1) < h(v_2)$. The *coalescence sequence* or *labeled history* of $\mathcal{G}$ is the sequence of partitions $\pi_0, \pi_1, ..., \pi_{n-1}$ of X such that $\pi_0(\mathcal{G}) = \{\{x_1\}, \{x_2\}, ..., \{x_n\}\}$ and for $i = 1, 2, ..., n-1$, $\pi_i(\mathcal{G})$ is formed from $\pi_{i-1}(\mathcal{G})$ by combining the two blocks in $\pi_{i-1}(\mathcal{G})$ containing leaves descended from the vertex $h^{-1}(i)$ into the same block in $\pi_i(\mathcal{G})$. The labeled history of $\mathcal{G}$ represents the sequence of events that reduce the n leaves to their MRCA; $h^{-1}(i)$ corresponds to the ith *coalescence* or *coalescent event*. The *k-truncated coalescence sequence* or *k-truncated labeled history* of $\mathcal{G}$ ($1 \leq k \leq n$) is the sequence $\pi_0, \pi_1, ..., \pi_{n-k}$.

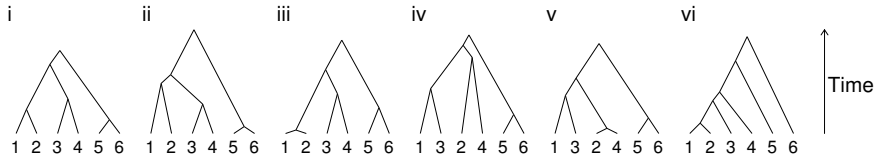

Figure 1: Example genealogies with the label set $\{1, 2, 3, 4, 5, 6\}$. Genealogies (i)–(v) all have the same unlabeled topology. In comparison with (i), (ii) has the same unlabeled history and labeled topology, but a different labeled history; (iii) has the same labeled topology, but different labeled and unlabeled histories; (iv) has the same unlabeled history, but a different labeled history and labeled topology; (v) has different labeled and unlabeled histories and a different labeled topology. Genealogy (vi) has different labeled and unlabeled histories and topologies from the remaining genealogies. It is a 6-caterpillar and is the only one of the genealogies to contain a pitchfork. Genealogies (i)–(v) all contain a symmetric 4-pronged node; the 4-pronged node in (vi) is not symmetric. There are three cherries in each of (i)–(v) and one cherry in (vi).

The *unlabeled history* of $\mathcal{G}$ is a particular sequence of partitions $\pi_0^*, \pi_1^*, ..., \pi_{n-1}^*$ of X obtained in the following manner. We begin with a set of "available" labels, $A = \{1, 2, ..., n\}$, and we equate $B = A$ and $\pi_0^* = \pi_0$. Sequentially, for $i \geq 1$, if one of the blocks in $\pi_i(\mathcal{G})$ containing leaves descended from $h^{-1}(i)$ includes only one leaf, we then reassign the label of that leaf as the label b with the smallest value among those in B, and replace B with $B \setminus \{b\}$ (if both blocks each include only one leaf, reassign the labels of both leaves — the order in which the two reassignments are made is unimportant). We repeat this procedure until each leaf v has been assigned a label $\gamma(v)$ from A. The unlabeled history of $\mathcal{G}$ is then the labeled history of $\mathcal{G}' = (G, A, \gamma, t)$. The *labeled topology* of $\mathcal{G}$ is the tree G with label set X and labeling γ, ignoring the time function t. The *unlabeled topology* of $\mathcal{G}$ is the tree G, ignoring the label set X, the labeling γ and the time t.

The *subgenealogy* $\mathcal{G}_v$ of $\mathcal{G}$ induced by an internal node v is the genealogy $(G_v, X_v, \psi|G_v, t|G_v)$, where G_v is the subtree of G containing all leaves that descend from v, X_v is the label set for G_v, and $\psi|G_v$ and $t|G_v$ denote the restrictions of ψ and t to G_v, respectively. An internal node v of G or its induced subgenealogy is *r-pronged* ($2 \leq r \leq n$) if the subgenealogy contains exactly r descendants among the leaves of G; the node (or subgenealogy) is an *r-caterpillar* if G_v contains exactly r descendants among the leaves of G, and if the internal node with the smallest value of t among all internal nodes in G_v is descended from all other internal nodes in G_v. A genealogy whose root is an r-caterpillar is termed *pectinate*. As special cases, 2-pronged and 3-pronged internal nodes (or their induced subgenealogies) are termed *cherries* and *pitchforks*, respectively. Both cherries and pitchforks are necessarily caterpillars.

For a genealogy $\mathcal{G}$, $d_r(\mathcal{G})$ is the number of r-pronged nodes in G. An internal node with more than two descendants in X is *symmetric* if the two subgenealogies induced by its two immediate descendants have the same unlabeled topology. For a genealogy $\mathcal{G}$, $s(\mathcal{G})$ is its number of symmetric internal nodes. The functions d_r and s can be applied also to labeled or unlabeled histories or topologies.

3. The Yule Model

This section reviews properties of genealogical trees generated by the Yule model (*Yule-generated genealogies*). A Yule-generated genealogy has the property that at any time, each pair of lineages is equally likely to be the next to coalesce (or equivalently, reversing the direction of time, each lineage is equally likely to be the next to divide). It follows directly from this property that the probability distribution of the labeled history of such a genealogy is uniform.

In this section and in those that follow, $\mathcal{G} = (G, X, \psi, t)$ is treated as a random n-leaved Yule-generated genealogy, and "genealogy" implicitly refers to the Yule-generated genealogy $\mathcal{G}$. Henceforth, $a_1, a_2, a_3, i, k, n, n_1, n_2, n_3$, and r are assumed to be positive integers with $n \geq 3$, $r \geq 2$, and except where otherwise specified, $i \leq n-1$ and $k \leq n$. It is assumed in each of (3.1)–(3.5) that T is chosen from the appropriate set of objects to which the result applies (for example, in Theorem 3.3, the collection of possible labeled topologies for the label set X).

Except for (3.3) and (3.4), which are included for completeness, the results in this section are used in proving the new results in the following sections; other than (3.8) and (3.9), they utilize the uniform distribution of labeled histories for Yule-generated genealogies.

Theorem 3.1. [4, 30] *The probability $I(T)$ that a genealogy for n lineages has k-truncated labeled history T is*

$$I(T) = \frac{1}{I_{n,k}} = \frac{2^{n-k}k!(k-1)!}{n!(n-1)!}.$$

Corollary 3.2. [8,21] *The probability $H(T)$ that a genealogy for n lineages has labeled history T is*

$$H(T) = \frac{1}{H_n} = \frac{2^{n-1}}{n!(n-1)!}.$$

Theorem 3.3. [1, 4, 35] *The probability $L(T)$ that a genealogy for n lineages has labeled topology T is*

$$L(T) = \frac{2^{n-1}}{n!\prod_{r=3}^{n}(r-1)^{d_r(T)}}.$$

Theorem 3.4. [25, 38] *The probability $U(T)$ that a genealogy for n lineages has unlabeled history T is*

$$U(T) = \frac{2^{n-1-d_2(T)}}{(n-1)!}.$$

Theorem 3.5. [4, 36] *The probability $Q(T)$ that a genealogy for n lineages has unlabeled topology T is*

$$Q(T) = \frac{2^{n-1-d_2(T)-s(T)}}{\prod_{r=3}^{n}(r-1)^{d_r(T)}}.$$

Corollary 3.6. [33, 37] *The probability $Q(T)$ that a genealogy for n lineages has unlabeled topology T, where T is pectinate, is*

$$Q(T) = \frac{2^{n-2}}{(n-1)!}.$$

Theorem 3.7. [34, 38] *In a genealogy for n lineages, the probability C_n that the two nodes immediately descended from the root have i and $n-i$ descendants is*

$$C_{1n} = 1/(n-1), \quad \text{if } i = n/2 \text{ and } n \text{ is even},$$

$$C_{2n} = 2/(n-1), \quad \text{if } i \neq n/2 \text{ and } i \in \{1, 2, ..., n-1\}.$$

Theorem 3.8. [4, 30] *The number W_{a_1,a_2} of $(n-a_1-a_2)$-truncated coalescence sequences possible for a fixed collection of $n = n_1+n_2$ labels (with $n_1 \geq a_1, n_2 \geq a_2$), such that each coalescence sequence contains two specified subsequences, one of length a_1 that coalesces n_1 lineages to $n_1 - a_1$ and another of length a_2 that coalesces the remaining n_2 lineages to $n_2 - a_2$, is the binomial coefficient*

$$W_{a_1,a_2} = \binom{a_1+a_2}{a_1}.$$

Theorem 3.9. [30] *The number W_{a_1,a_2,a_3} of $(n-a_1-a_2-a_3)$-truncated coalescence sequences possible for a fixed collection of $n = n_1+n_2+n_3$ labels (with $n_1 \geq a_1, n_2 \geq a_2, n_3 \geq a_3$), such that each coalescence sequence contains three specified subsequences, one of length a_1 that coalesces n_1 lineages to $n_1 - a_1$, one of length a_2 that coalesces n_2 lineages to $n_2 - a_2$, and the third of length a_3 that coalesces the remaining n_3 lineages to $n_3 - a_3$, is the trinomial coefficient*

$$W_{a_1,a_2,a_3} = \binom{a_1+a_2+a_3}{a_1, a_2, a_3}.$$

Remark 3.10. A correspondence exists between Yule-generated genealogies with n leaves and the entities generated during construction of *random binary search trees* with $n-1$ vertices (leaves plus internal nodes). A random binary search tree is obtained via sequential addition of descendant vertices to a rooted binary tree so that each vertex with one descendant has another slot in which a descendant can be added, and each vertex with no descendants has two slots in which descendants can be added (interior vertices in binary search trees are allowed to have either 1 or 2 descendants). At any time (reversing the direction of time), all potential slots for insertion of vertices are equally likely to be the next to have a vertex added [17, pp. 68-70]. The corresponding binary search tree for a given Yule-generated genealogy $\mathcal{G} = (G, X, \psi, t)$ is the pair $(G_b, t|G_b)$, where G_b is the subtree of G containing only the internal nodes of G and the edges that connect them, and $t|G_b$ is the restriction of t to G_b.

4. The Number of r-Pronged Nodes

After proving three combinatorial identities as Lemmas 4.1, 4.2, and 4.3, the mean and variance of the number of r-pronged nodes in a Yule-generated genealogy are obtained

in Theorem 4.4, and the properties of the variance function are then examined in Theorem 4.8. The mean had previously been obtained in [13] using a Polya urn method, and the mean and variance were both calculated for the case of $r = 2$ in [18]. Related problems have also been considered in [10].

Because of the correspondence between Yule-generated genealogies and random binary search trees, many results derived in the context of binary search trees [6, 7, 17] can be interpreted as statements about Yule-generated genealogies. For example, using the fact that a node with r descendant nodes in a binary search tree has $r+2$ descendants among the leaves of the corresponding Yule-generated genealogy, Theorem 4.4 (i) and the $r < n/2$ case of Theorem 4.4 (ii) are demonstrated (via different proofs from the ones here) in [6, Theorem 5], which further obtains a limiting distribution in n for the number of r-pronged nodes (see also [3]).

Lemma 4.1. *For positive integers n and r with $n \geq 3$, $r \geq 2$, $n \geq r$,*

$$\sum_{k=1}^{n-r} \binom{k+1}{2} \binom{n-k-2}{r-2} = \binom{n}{r+1}.$$

Proof. The identity follows from [11, Identity 3.3], a statement straightforward to prove using induction on m,

$$\sum_{l=q}^{m-s} \binom{l}{q} \binom{m-l}{s} = \binom{m+1}{q+s+1},$$

substituting $k+1$ for l, 2 for q, $r-2$ for s, and $n-1$ for m. ∎

Lemma 4.2. *For positive integers n, r, and i, with $n - 2r \geq 1$, $r \geq i$, $r \geq 2$,*

$$\sum_{j=1}^{n-2r} \binom{j+1}{2} \binom{i+j+1}{j} \binom{n-2-i-j}{r-2} \binom{n-r-i-j}{r-i}$$
$$= \frac{(i+2)[(i+3)n-(i+1)2r-(i-1)]}{2(n-2r-1)} \binom{2r-i-2}{r-2} \binom{n}{2r+2}. \tag{4.1}$$

Proof. Set $w = n - 2r$ and denote the ratio of the summand in Equation 4.1 to the "right-hand side" by $F(w, j)$:

$$F(w, j) = \frac{\binom{j+1}{2}\binom{i+j+1}{j}\binom{2r+w-2-i-j}{r-2}\binom{r+w-i-j}{r-i}}{\binom{2r-i-2}{r-2}\binom{2r+w+1}{2r+2}} \frac{2(2r+w+1)}{(i+2)[4r+(i+3)w-(i-1)]}.$$

Together with the proof certificate

$$R(w, j) = -\frac{(j-1)(2r+w-1-i-j)(4r+2w-2i+2j+4jr+3jw+ijw)}{(j+1)(w+1-j)(2r+w+1)(4r+3w+4+iw)},$$

$F(w, j)$ satisfies the hypotheses of the Wilf-Zeilberger automatic summation theorem [27, Theorem 7.1.1], from which it follows that $\sum_{j=1}^{w} F(w, j)$ is not dependent on w. Because $\sum_{j=1}^{w} F(w, j) = 1$ when $w = 1$, the result follows. ∎

Lemma 4.3. *For positive integers n and r with $n \geq 3$, $r \geq 2$, $n \geq r$,*

$$\sum_{i=2}^{r} \frac{(i+2)[(i+3)n-(i+1)2r-(i-1)]}{2(n-2r-1)} \binom{2r-i-2}{r-2} \binom{n}{2r+2}$$
$$= \frac{5r-11r^3-2n+8nr^2}{4(r+1)(2r-1)(2r+1)} \binom{2r+1}{r} \binom{n}{2r+1}. \tag{4.2}$$

Proof. Adding terms for $i=0$ and $i=1$ to both sides of Equation 4.2 and setting $u=r-i$, the statement we wish to prove is equivalent to

$$\sum_{u=0}^{r} \alpha_1 u^2 \binom{u+c}{u} + \alpha_2 u \binom{u+c}{u} + \alpha_3 \binom{u+c}{u} = \frac{-5r-8r^2+5n+7nr+1}{2r+1} \binom{2r+1}{r},$$

where $c=r-2$, $\alpha_1 = n-2r-1$, $\alpha_2 = 8r+4r^2-5n-2nr+1$, and $\alpha_3 = (r+2)(-3r-2r^2+3n+nr+1)$. But this identity follows by setting $c=r-2$ in the following three identities concerning nonnegative integers r and c:

$$\sum_{u=0}^{r} \binom{u+c}{u} = \binom{r+c+1}{r} \tag{4.3}$$

$$\sum_{u=0}^{r} u \binom{u+c}{u} = \frac{r(c+1)}{c+2} \binom{r+c+1}{r} \tag{4.4}$$

$$\sum_{u=0}^{r} u^2 \binom{u+c}{u} = \frac{r(c+1)(2r+cr+1)}{(c+2)(c+3)} \binom{r+c+1}{r}. \tag{4.5}$$

Each of equations 4.3-4.5 is straightforward to prove by induction on r. ∎

Theorem 4.4. *In a genealogy for $n \geq 3$ lineages, if $2 \leq r \leq n-1$,*

(i) [13] *the mean number of r-pronged nodes, $M(n,r)$, is*

$$\frac{2n}{r(r+1)},$$

(ii) *the variance of the number of r-pronged nodes, $V(n,r)$, is*

$$V_1(n,r) = \frac{2(4r^2-3r-4)(r-1)n}{r(r+1)^2(2r-1)(2r+1)}, \qquad \text{if} \quad r < n/2, \tag{4.6}$$

$$V_2(n,r) = \frac{(5r-7)(r-1)n}{r(r+1)^2(2r-1)}, \qquad \text{if} \quad r = n/2, \tag{4.7}$$

$$V_3(n,r) = \frac{2(r^2+r-2n)n}{r^2(r+1)^2}, \qquad \text{if} \quad r > n/2. \tag{4.8}$$

Proof. Enumerate the $\binom{n}{r}$ subsets of the label set that contain r elements, denoting the bth such subset by S_b. Let Z_b be the indicator variable for whether there is some internal

node of the genealogy for which S_b is the label set of its induced subgenealogy. The number of r-pronged nodes in the genealogy is

$$Z = \sum_{b=1}^{\binom{n}{r}} Z_b.$$

(i) The mean number of r-pronged nodes in the genealogy is

$$M(n,r) = \mathbb{E}[Z] = \sum_{b=1}^{\binom{n}{r}} \mathbb{E}[Z_b] = \binom{n}{r}\mathbb{E}[Z_1] = \binom{n}{r}\mathbb{P}[Z_1 = 1].$$

$\mathbb{P}[Z_1 = 1]$ can be determined by counting the fraction of coalescence sequences for which S_1 has MRCA at an r-pronged node (Figure 2). In such a sequence, at the time just less than that of the r-pronged node, the r lineages with labels in S_1 have coalesced to 2 lineages and the $n-r$ lineages with labels in $X \setminus S_1$ have coalesced to k lineages ($1 \le k \le n-r$). The next coalescence event is the r-pronged node. The remaining $k+1$ lineages coalesce to the MRCA for X. Thus, applying Theorem 3.1, Corollary 3.2, and Theorem 3.8, the number of coalescence sequences in which S_1 has MRCA at an r-pronged node is $\sum_{k=1}^{n-r} I_{r,2} I_{n-r,k} W_{r-2,n-r-k} H_{k+1}$. Consequently, using Lemma 4.1,

$$\begin{aligned} M(n,r) &= \binom{n}{r}\frac{1}{H_n}\sum_{k=1}^{n-r} I_{r,2} I_{n-r,k} W_{r-2,n-r-k} H_{k+1} \\ &= \frac{2(n-r-1)!\,(r-1)!}{(n-1)!}\sum_{k=1}^{n-r}\binom{k+1}{2}\binom{n-k-2}{r-2} \\ &= \frac{2n}{r(r+1)}. \end{aligned} \tag{4.9}$$

(ii) The variance of the number of r-pronged nodes is

$$\begin{aligned} V(n,r) &= \mathbb{E}\left[\left(\sum_{b=1}^{\binom{n}{r}} Z_b\right)^2\right] - \mathbb{E}\left[\sum_{b=1}^{\binom{n}{r}} Z_b\right]^2 \\ &= \mathbb{E}\left[\sum_{b=1}^{\binom{n}{r}} Z_b\right] - \mathbb{E}\left[\sum_{b=1}^{\binom{n}{r}} Z_b\right]^2 + \mathbb{E}\left[\sum_{\substack{b,b' \\ b'\neq b}}^{\binom{n}{r}} Z_b Z_{b'}\right] \\ &= M(n,r) - M(n,r)^2 + \mathbb{E}\left[\sum_{\substack{b,b' \\ b'\neq b}}^{\binom{n}{r}} Z_b Z_{b'}\right]. \end{aligned} \tag{4.10}$$

Case 1. $r < n/2$. If $S_b \cap S_{b'} \neq \phi$, then $\mathbb{E}[Z_b Z_{b'}] = 0$. For all disjoint S_b and $S_{b'}$, $\mathbb{E}[Z_b Z_{b'}] = \mathbb{P}[Z_b Z_{b'} = 1]$ has the same value. The number of ordered pairs $(S_b, S_{b'})$ for which $S_b, S_{b'} \subset X$ and $S_b \cap S_{b'} = \phi$ is $2\binom{n}{r}\binom{n-r}{r}$. Therefore, supposing that (S_1, S_2) is such a pair,

$$\mathbb{E}\left[\sum_{\substack{b,b' \\ b'\neq b}}^{\binom{n}{r}} Z_b Z_{b'}\right] = 2\binom{n}{r}\binom{n-r}{r}\mathbb{P}[Z_1 Z_2 = 1]. \tag{4.11}$$

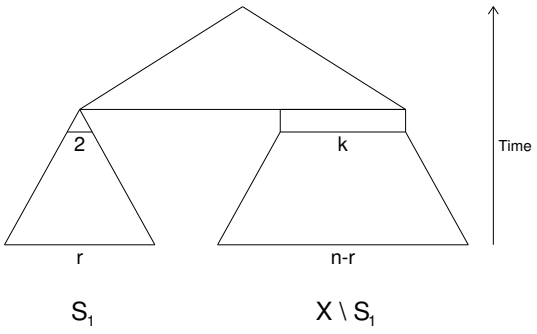

Figure 2: Counting the fraction of coalescence sequences for which S_1 has MRCA at an r-pronged node. Using Theorem 3.1, the number of ways that r lineages can coalesce to 2 lineages is $I_{r,2}$, and for $1 \leq k \leq n-r$, the number of ways that $n-r$ lineages can coalesce to k lineages is $I_{n-r,k}$. By Theorem 3.8, the number of ways of interweaving these sequences of $r-2$ and $n-r-k$ coalescences is $W_{r-2,n-r-k}$. By Corollary 3.2, the number of ways that $k+1$ lineages can coalesce to 1 lineage is H_{k+1}. Consequently, the total number of coalescence sequences for which S_1 has MRCA at an r-pronged node is $\sum_{k=1}^{n-r} I_{r,2} I_{n-r,k} W_{r-2,n-r-k} H_{k+1}$. By Corollary 3.2, the total number of coalescence sequences for n lineages is H_n. Thus, the desired quantity is $(1/H_n)\sum_{k=1}^{n-r} I_{r,2} I_{n-r,k} W_{r-2,n-r-k} H_{k+1}$.

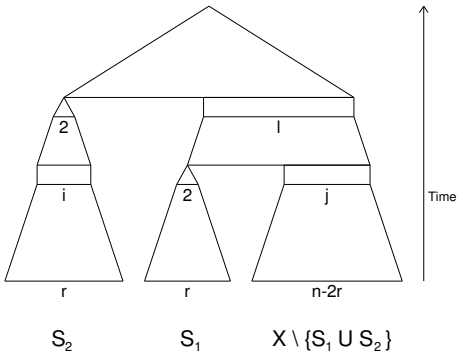

Figure 3: Counting the fraction of coalescence sequences for which both S_1 and S_2 are label sets for subgenealogies, with the time of the MRCA of S_1 smaller than that of S_2. Using Theorem 3.1, the number of ways that r lineages can coalesce to 2 lineages is $I_{r,2}$; for $2 \leq i \leq r$, the number of ways that r lineages can coalesce to i lineages is $I_{r,i}$; for $1 \leq j \leq n-2r$, the number of ways that $n-2r$ lineages can coalesce to j lineages is $I_{n-2r,j}$. By Theorem 3.9, the number of ways of interweaving these sequences of $r-2$, $r-i$ and $n-2r-j$ coalescences is $W_{r-2,r-i,n-2r-j}$. Counting the number of ways that the remaining lineages can coalesce so that S_2 is the label set for an induced subgenealogy follows the same argument as in Figure 2.

In order to have $Z_1Z_2 = 1$, both S_1 and S_2 must be label sets for subgenealogies. Without loss of generality, suppose that the time of the MRCA of S_1 is less than that of S_2. $\mathbb{P}[Z_1Z_2 = 1]$ can be determined by counting the fraction of coalescence sequences for which S_1 and S_2 are the label sets for induced subgenealogies (Figure 3). At the time just less than that of the MRCA of S_1, the r lineages with labels in S_2 have coalesced to i lineages ($2 \leq i \leq r$), and the $n-2r$ lineages with labels in $X \setminus \{S_1 \cup S_2\}$ have coalesced to j lineages ($1 \leq j \leq n-2r$). The next coalescence produces the r-pronged node for which S_1 is the label set of the induced subgenealogy. At the time just less than that of the MRCA of S_2, the remaining i lineages ancestral to S_2 have coalesced to 2 lineages, and the remaining $j+1$ lineages ancestral to $X \setminus S_2$ have coalesced to l lineages ($1 \leq l \leq j$). The next coalescence produces the r-pronged node for which S_2 is the induced subgenealogy. Finally, the remaining $l+1$ lineages ancestral to X coalesce to 1 lineage. The total number of possible coalescence sequences for all n lineages with labels in X is H_n. Using Theorem 3.1, Corollary 3.2 and Theorems 3.8 and 3.9,

$$\mathbb{P}[Z_1Z_2 = 1] = \frac{\sum_{i=2}^{r}\sum_{j=1}^{n-2r} I_{r,2}I_{r,i}I_{n-2r,j}W_{r-2,r-i,n-2r-j}\sum_{l=1}^{j+1} I_{i,2}I_{j+1,l}W_{i-2,j+1-l}H_{l+1}}{H_n}.$$

Simplifying this expression and using Equation 4.11,

$$\mathbb{E}\left[\sum_{\substack{b,b'\\ b'\neq b}}^{\binom{n}{r}} Z_bZ_{b'}\right]$$
$$= \frac{8(r-1)!^2(n-2r-1)!\sum_{i=2}^{r}\sum_{j=1}^{n-2r}\sum_{l=1}^{j+1}\binom{j+1}{2}\binom{l+1}{2}\binom{i+j-l-1}{i-2}\binom{n-2-i-j}{r-2}\binom{n-r-i-j}{r-i}}{(n-1)!}.$$

Sequentially applying Lemmas 4.1, 4.2, and 4.3 to sum over indices l, j and i,

$$\mathbb{E}\left[\sum_{\substack{b,b'\\ b'\neq b}}^{\binom{n}{r}} Z_bZ_{b'}\right] = \frac{2(5r-11r^3-2n+8nr^2)n}{r^2(r+1)^2(2r-1)(2r+1)},$$

and the result follows from Equation 4.10.

Case 2. $r = n/2$. In this case, a genealogy can have either zero or two r-pronged nodes, so that $Z_b = 1$ for either zero or two values of b. Consequently, the sum in Equation 4.10 can have either no terms equal to 1, or exactly two terms equal to 1:

$$\mathbb{E}\left[\sum_{\substack{b,b'\\ b'\neq b}}^{\binom{n}{r}} Z_bZ_{b'}\right] = 2\mathbb{P}\left[\sum_{\substack{b,b'\\ b'\neq b}}^{\binom{n}{r}} Z_bZ_{b'} = 2\right].$$

A genealogy can have two distinct sets each with $r = n/2$ labels and each with MRCA at an r-pronged node if and only if the two sets are disjoint and both r-pronged nodes

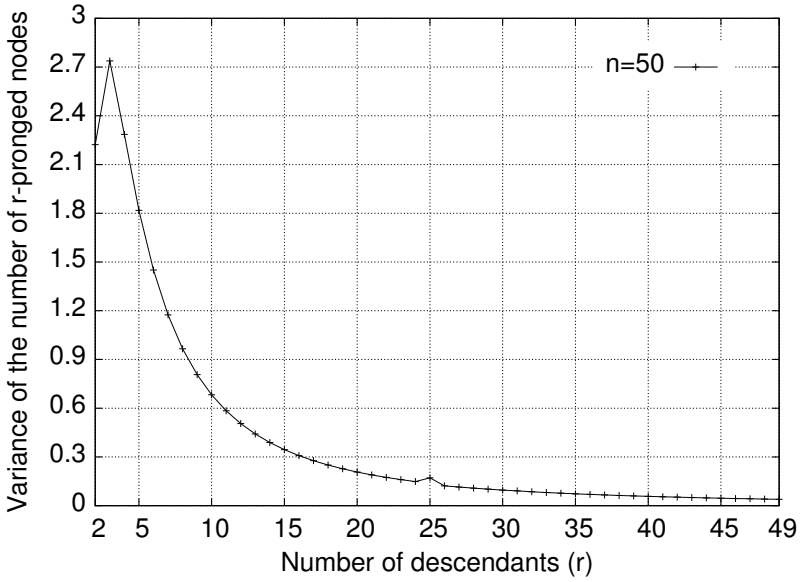

Figure 4: The variance $V(n, r)$ of the number of r-pronged nodes for $n = 50$.

are immediately descended from the root. By Theorem 3.7, the probability of having two such nodes is $1/(n-1)$. Therefore, using Equation 4.10,

$$V(n,r) = \frac{2n}{r(r+1)} \frac{r(r+1)-2n}{r(r+1)} + \frac{2}{n-1},$$

and the result follows using $n = 2r$.

Case 3. $r > n/2$. In this case, at most one set of r labels can be the label set of an induced subgenealogy, so that for any distinct b, b', $Z_b Z_{b'} = 0$. The result then follows from Equation 4.10. ∎

Corollary 4.5. [18] In a genealogy for $n \geq 5$ lineages, the mean number of cherries is $n/3$ and the variance of the number of cherries is $2n/45$.

Corollary 4.6. *In a genealogy for $n \geq 7$ lineages, the mean number of pitchforks is $n/6$ and the variance of the number of pitchforks is $23n/420$.*

Remark 4.7. For a given n, the mean number of r-pronged nodes is largest for $r = 2$, and declines monotonically as r increases. The variance, however, exhibits a more complicated pattern. Figure 4 displays a typical example, namely $n = 50$. The global maximum of $V(50, r)$ occurs at $r = 3$ rather than at $r = 2$, and the decline after $r = 3$ is interrupted by a small peak at $r = 50/2 = 25$. The following theorem summarizes the general shape of $V(n, r)$ as a function of r.

Theorem 4.8. *For a fixed $n \geq 15$ as well as for $n = 11, 13$, as r ranges over integers from 2 to $n-1$,* (i) *the four highest values of $V(n,r)$ are, from greatest to smallest, at $r = 3, 4, 2,$ and 5;* (ii) *for $3 \leq r \leq n-2$, $V(n,r) > V(n,r+1)$, unless n is even and $r+1 = n/2$, in which case $V(n,r) < V(n,r+1)$.*

Proof. We will need the following six inequalities, each of which is straightforward to prove from Equations 4.6-4.8 using elementary methods.

(a) For integers n, r with $n \geq 3$ and $2 \leq r \leq n-1$, $V_2(n,r) > V_1(n,r) > V_3(n,r)$.
(b) For integers n, r with $n \geq 3$ and $3 \leq r \leq n-1$, $V_1(n,r) > V_1(n,r+1)$.
(c) For positive n, $V_1(n,3) > V_1(n,4) > V_1(n,2) > V_1(n,5)$.
(d) For even integers $n \geq 12$, $V_1(n,5) > V_2(n,n/2)$.
(e) For integers n, r with $n \geq 10$ and $n/2 < r \leq n-1$, $V_3(n,r) > V_3(n,r+1)$.
(f) For even integers $n \geq 16$, $V_2(n,n/2) > V_1(n,n/2-1)$.

Using (b) and (c), V_1 has its four highest values at 3, 4, 2, and 5, respectively. For $n \geq 11$, $V(n,5) = V_1(n,5)$. By (a) and (b), $V_1(n,5) > V_3(n,r)$ for all integers n, r with $n \geq 11$, $n/2 < r < n$. Applying (d), it follows that for $n \geq 11$, (i) holds.

Using (a) and (b), $V_1(n, \lceil n/2-1 \rceil) > V_3(n, \lfloor n/2+1 \rfloor)$. For $n \geq 3$, $V(n,r) = V_1(n,r)$ for $2 \leq r \leq \lceil n/2-1 \rceil$ and $V(n,r) = V_3(n,r)$ for $\lfloor n/2+1 \rfloor \leq r \leq n-1$. Applying (a), (b), and (e), (ii) holds for odd $n > 10$. For even n it must also be verified that $V(n,r) < V(n,r+1)$ for $r+1 = n/2$; this follows from (f), but only for $n \geq 16$. ∎

Remark 4.9. The theorem shows that for large enough n, the variance follows a particular pattern as a function of r. For very large r, the variance approaches $2n/r^2$, provided n remains larger than $2r$ ($2.5n/r^2$ if $n = 2r$). For large n, starting with $r = 2$, $V(n,r)/n$ follows the sequence $2/45, 23/420, 8/175, 2/55, 610/21021, 171/7280, \ldots$. At small n ($n \leq 14$), however, the behavior of the variance function can differ from that specified by the theorem (Figure 5).

5. The Number of r-Caterpillars

The mean and variance of the number of r-pronged nodes can be used to obtain the mean and variance of the number of subgenealogies with any given unlabeled topology. Theorem 5.1 gives the general formulas for an arbitrary unlabeled topology T_r; the case of a pectinate T_r is considered in Corollary 5.2, and the properties of the variance function are then explored in Theorem 5.4. Theorem 4.4, Theorem 5.1, and Corollary 5.2 can all be considered generalizations of the formulas of [18] for the mean and variance of the number of cherries.

Theorem 5.1. *In a genealogy with $n \geq 3$ lineages, if $2 \leq r \leq n-1$ and T_r is an unlabeled topology with r leaves,* (i) *the mean number of nodes that have induced subgenealogy T_r, $M_{T_r}(n,r)$, is $Q(T_r)M(n,r)$, and* (ii) *the variance of the number of nodes that have induced subgenealogy T_r, $V_{T_r}(n,r)$, is*

$$V_{T_r}(n,r) = Q(T_r)[1-Q(T_r)]M(n,r) + Q(T_r)^2 V(n,r).$$

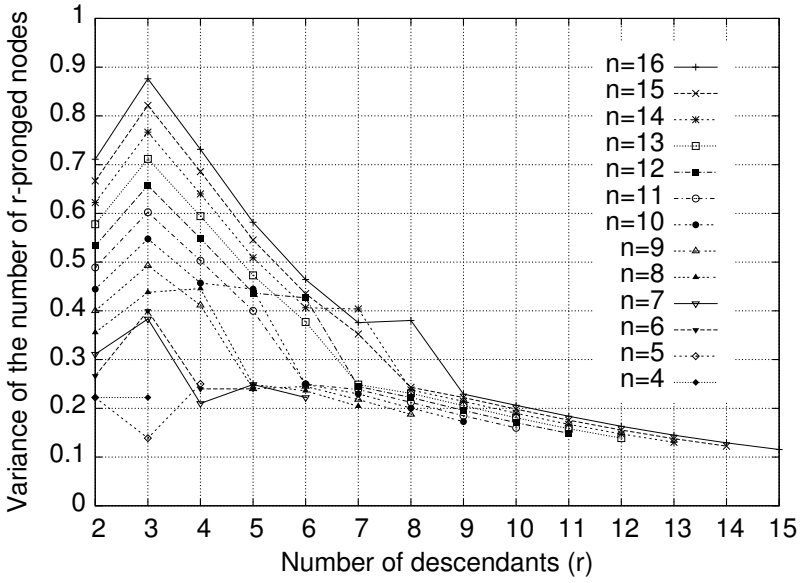

Figure 5: The variance $V(n, r)$ of the number of r-pronged nodes for $n = 4$ through 16. For $n = 3$, $V(3, 2) = 0$.

Proof. Let Y_b be the indicator variable for whether there is some internal node of the genealogy for which both (a) S_b is the label set of its induced subgenealogy, and (b) this induced subgenealogy has unlabeled topology T_r. The number of nodes whose induced subgenealogies have unlabeled topology T_r is

$$Y = \sum_{b=1}^{\binom{n}{r}} Y_b.$$

Note that if $Z_b = 0$, then $Y_b = 0$. Applying Theorem 3.5, if $Z_b = 1$, then $Y_b = 1$ with probability $Q(T_r)$.

(i) The mean number of nodes that have induced subgenealogy T_r is

$$\begin{aligned} M_{T_r}(n, r) = \mathbb{E}[Y] &= \sum_{b=1}^{\binom{n}{r}} \mathbb{P}[Z_b = 0]\,\mathbb{E}[Y_b|Z_b = 0] + \mathbb{P}[Z_b = 1]\,\mathbb{E}[Y_b|Z_b = 1] \\ &= \sum_{b=1}^{\binom{n}{r}} Q(T_r)\,\mathbb{P}[Z_b = 1] \\ &= Q(T_r)M(n, r). \end{aligned}$$

(ii) Using the conditional variance formula [28, p. 138] with the fact that a Bernoulli

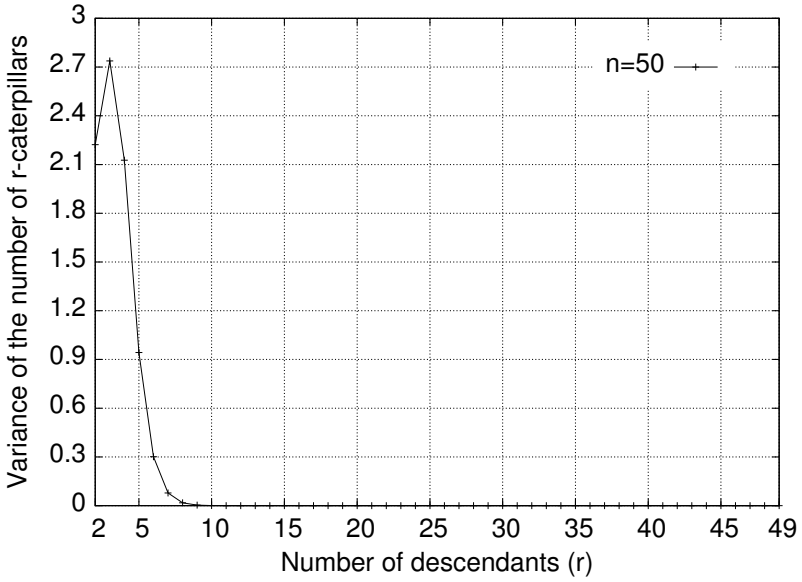

Figure 6: The variance $V_{cat}(n, r)$ of the number of r-caterpillars for $n = 50$.

random variable with parameter $Q(T_r)$ has variance $Q(T_r)[1-Q(T_r)]$,

$$\begin{aligned} V_{T_r}(n,r) = Var[Y] &= \mathbb{E}[Var[Y|Z]] + Var[\mathbb{E}[Y|Z]] \\ &= \mathbb{E}[Q(T_r)[1-Q(T_r)]Z] + Var[Q(T_r)Z] \\ &= Q(T_r)[1-Q(T_r)]M(n,r) + Q(T_r)^2 V(n,r). \end{aligned}$$ ∎

Corollary 5.2. *In a genealogy with $n \geq 3$ lineages, if $2 \leq r \leq n-1$, then* (i) *the mean number of r-caterpillars, $M_{cat}(n, r)$, is*

$$\frac{2^{r-1}n}{(r+1)!},$$

and (ii) *the variance of the number of r-caterpillars, $V_{cat}(n, r)$, is*

$$V_{1cat}(n,r) = \frac{2^{r-1}n[(2r-1)(2r+1)(r+1)! + 2^{r-2}r(-11r^2+5)]}{(r+1)!^2(2r-1)(2r+1)}, \quad \text{if} \quad r < n/2,$$

$$V_{2cat}(n,r) = \frac{2^{r-1}n[(2r-1)(r+1)! + 2^{r-3}r(r^2-14r+9)]}{(r+1)!^2(2r-1)}, \quad \text{if} \quad r = n/2,$$

$$V_{3cat}(n,r) = \frac{2^{r-1}n[(r+1)! - 2^{r-1}n]}{(r+1)!^2}, \quad \text{if} \quad r > n/2.$$

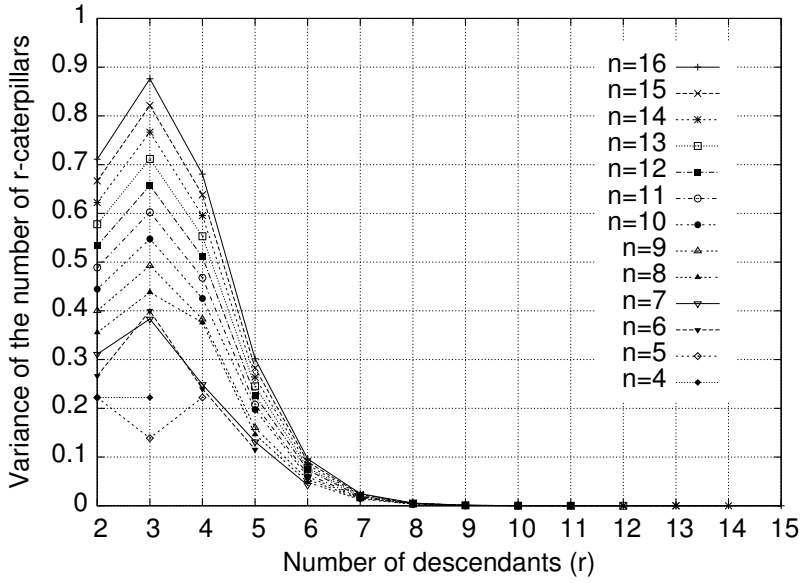

Figure 7: The variance $V_{cat}(n, r)$ of the number of r-caterpillars for $n = 4$ through 16. For $n = 3$, $V(3, 2) = 0$.

Proof. Applying Corollary 3.6, $Q(T_r) = 2^{r-2}/(r-1)!$ if T_r is pectinate. Inserting this quantity along with the values of $M(n, r)$ and $V(n, r)$ from Theorem 4.4 into Theorem 5.1, the result follows. ∎

Remark 5.3. Similarly to the case of r-pronged nodes, the mean number of r-caterpillars is largest for $r = 2$, and the variance is largest at $r = 3$. Figure 6 displays $V_{cat}(50, r)$ as a function of r. In the case of caterpillars, unlike the r-pronged node case, the monotonic decline in variance after $r = 3$ is not interrupted at $r = n/2$, as is demonstrated in the following theorem.

Theorem 5.4. *For a fixed $n \geq 11$ as well as for $n = 6, 7, 9, 10$, as r ranges over integers from 2 to $n-1$,* (i) *the four highest values of $V_{cat}(n, r)$ are from greatest to smallest, at $r = 3, 2, 4,$ and 5;* (ii) *for $3 \leq r \leq n-2$, $V_{cat}(n, r) > V_{cat}(n, r+1)$.*

Proof. The cases of $n = 6, 7, 9, 10$ can be verified from Corollary 5.2 (ii) by direct computation. For $n \geq 11$ the proof follows from the following six inequalities: (a), (b), and (e) are easily proven from the corresponding statements in the proof of Theorem 5.4, and (c), (d) and (f) are straightforward using elementary methods.

(a) For integers n, r with $n \geq 3$ and $2 \leq r \leq n-1$, $V_{2cat}(n, r) > V_{1cat}(n, r) > V_{3cat}(n, r)$,
(b) For integers n, r with $n \geq 3$ and $3 \leq r \leq n-1$, $V_{1cat}(n, r) > V_{1cat}(n, r+1)$,
(c) For positive n, $V_{1cat}(n, 3) > V_{1cat}(n, 2) > V_{1cat}(n, 4) > V_{1cat}(n, 5)$,
(d) For even integers $n \geq 12$, $V_{1cat}(n, 5) > V_{2cat}(n, n/2)$,

(e) For integers n, r with $n \geq 10$ and $n/2 < r \leq n-1$, $V_{3cat}(n,r) > V_{3cat}(n,r+1)$,
(f) For even integers $n \geq 8$, $V_{2cat}(n,n/2) < V_{1cat}(n,n/2-1)$.

The proof of (i) uses (a), (b), (c), and (d), and follows that of Theorem 4.8 (i), except that the positions of $V_{1cat}(2,r)$ and $V_{1cat}(4,r)$ are reversed. The proof of (ii) uses (a), (b), (e) and (f) and follows that of Theorem 4.8 (ii), except that the direction of (f) guarantees $V_{cat}(n,r) > V_{cat}(n,r+1)$ at $r+1 = n/2$ for even n. ∎

Remark 5.5. For large r relative to n, a genealogy is likely to contain at most one r-caterpillar, and V_{cat} approaches M_{cat}. For large n, with $r \geq 2$, $V_{cat}(n,r)/n$ follows the sequence $2/45, 23/420, 67/1575, 364/19305, 28466/4729725, 823/526500\ldots$. At small n ($n \leq 8$), however, as in the case of r-pronged nodes, the behavior of the variance can differ from that specified by the theorem (Figure 7).

It is interesting to compare the mean and variance of the number of r-pronged nodes with those of the number of r-caterpillars. For $r = 2$ and $r = 3$, an r-pronged node is necessarily a caterpillar, and the numbers of r-pronged nodes and r-caterpillars in a genealogy are equal. For $r \geq 4$, an r-pronged node need not be an r-caterpillar, and consequently, the mean number of r-caterpillars is strictly less than the mean number of r-pronged nodes. For $r \geq 4$, as can be verified by elementary comparison of the formulas in Theorem 4.4 and Corollary 5.2, the variance of the number of r-caterpillars is strictly less than the variance of the number of r-pronged nodes, with two exceptions: for $(n,r) = (6,4)$, both variances equal $6/25$, and for $(n,r) = (7,4)$, the variance of the number of caterpillars equals $56/225$, while the variance of the number of r-pronged nodes equals only $21/100$.

Summing over all r, each internal node is r-pronged for one value of r, and the mean total number of r-pronged nodes (not counting the root) is $\sum_{r=2}^{n-1} 2n/[r(r+1)] = n-2$, as it should. The mean number of caterpillars — internal nodes (not counting the root) that are r-caterpillars for some r — is $\sum_{r=2}^{n-1} 2^{r-1} n/(r+1)!$. As a fraction of n, this sum has a large-n limit of $(e^2-5)/4 \approx 0.59726$; thus, for large Yule-generated genealogies, on average $\sim 60\%$ of subgenealogies are pectinate.

Acknowledgments. I thank the referees for careful readings of the manuscript. This work was supported by a National Science Foundation Postdoctoral Fellowship in Biological Informatics and by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences.

References

1. D. Aldous, Probability distributions on cladograms, In: Discrete Random Structures, D. Aldous and R. Pemantle, Eds., Springer-Verlag, New York, (1996) pp. 1–18.
2. D.J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Statist. Sci. **16** (2001) 23–34.
3. M.G.B. Blum and O. François, On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited, Math. Biosci. **195** (2005) 141–153.
4. J.K.M. Brown, Probabilities of evolutionary trees, Syst. Biol. **43** (1994) 78–91.
5. J.H. Degnan and L.A. Salter, Gene tree distributions under the coalescent process, Evolution **59** (2005) 24–37.

6. L. Devroye, Limit laws for local counters in random binary search trees, Random Structures Algorithms **2** (1991) 303–315.
7. L. Devroye, Limit laws for sums of functions of subtrees of random binary search trees, SIAM J. Comput. **32** (2003) 152–171.
8. A.W.F. Edwards, Estimation of the branch points of a branching diffusion process, J. Roy. Statist. Soc. Ser. B **32** (1970) 155–174.
9. J. Felsenstein. Inferring Phylogenies, Sinauer, Sunderland, MA, 2004.
10. Y.X. Fu, Statistical properties of segregating sites, Theor. Pop. Biol. **48** (1995) 172–197.
11. H.W. Gould, Combinatorial Identities, Gould Publications, Morgantown, WV, 1972.
12. E.F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. Appl. Probab. **3** (1971) 44–77.
13. S.B. Heard, Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees, Evolution **46** (1992) 1818–1826.
14. R.R. Hudson, Gene genealogies and the coalescent process, Oxf. Surv. Evol. Biol. **7** (1990) 1–44.
15. J.F.C. Kingman, On the genealogy of large populations, J. Appl. Probab. **19A** (1982) 27–43.
16. W.P. Maddison and M. Slatkin, Null models for the number of evolutionary steps in a character on a phylogenetic tree, Evolution **45** (1991) 1184–1197.
17. H.M. Mahmoud, Evolution of Random Search Trees, Wiley, New York, 1992.
18. A. McKenzie and M. Steel, Distributions of cherries for two models of trees, Math. Biosci. **164** (2000) 81–92.
19. A.O. Mooers and S.B. Heard, Evolutionary process from phylogenetic tree shape, Quart. Rev. Biol. **72** (1997) 31–54.
20. A.O. Mooers and S.B. Heard, Using tree shape, Syst. Biol. **51** (2002) 833–834.
21. F. Murtagh, Counting dendrograms: a survey, Discrete Appl. Math. **7** (1984) 191–199.
22. J.E. Neigel and J.C. Avise, Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation, In: Evolutionary Processes and Theory, S. Karlin and E. Nevo, Eds., Academic Press, New York, (1986) pp. 515–534.
23. M. Nordborg, On the probability of Neanderthal ancestry, Amer. J. Hum. Genetics **63** (1998) 1237–1240.
24. M. Nordborg, Coalescent theory, In: Handbook of Statistical Genetics, Chapter 7, D.J. Balding, M. Bishop, and C. Cannings, Eds., Wiley, Chichester, UK, (2001) pp. 179–212.
25. R.D.M. Page, Random dendrograms and null hypotheses in cladistic biogeography, Syst. Zool. **40** (1991) 54–62.
26. P. Pamilo and M. Nei, Relationships between gene trees and species trees, Mol. Biol. Evol. **5** (1988) 568–583.
27. M. Petkovšek, H.S. Wilf, and D. Zeilberger, $A = B$, Peters, Wellesley, MA, 1996.
28. J.A. Rice, Mathematical Statistics and Data Analysis, 2nd edition, Duxbury Press, Belmont, CA, 1995.
29. N.A. Rosenberg, The probability of topological concordance of gene trees and species trees, Theoret. Popul. Biol. **61** (2002) 225–247.
30. N.A. Rosenberg, The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model, Evolution **57** (2003) 1465–1477.
31. N.A. Rosenberg, Gene genealogies, In: Evolutionary Genetics: Concepts and Case Studies, C.W. Fox and J.B. Wolf, Eds., Oxford University Press, Oxford, 2006.

32. C. Semple and M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.
33. J.B. Slowinski, Probabilities of *n*-trees under two models: a demonstration that asymmetrical interior nodes are not improbable, Syst. Zool. **39** (1990) 89–94.
34. J.B. Slowinski and C. Guyer, Testing the stochasticity of patterns of organismal diversity: an improved null model, Amer. Naturalist **134** (1989) 907–921.
35. M. Steel and A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, Math. Biosci. **170** (2001) 91–112.
36. J. Stone and J. Repka, Using a nonrecursive formula to determine cladogram probabilities, Syst. Biol. **47** (1998) 617–624.
37. J.R. Stone, Probabilities for completely pectinate and symmetric cladograms, Cladistics **19** (2003) 565–566.
38. F. Tajima, Evolutionary relationship of DNA sequences in finite populations, Genetics **105** (1983) 437–460.
39. 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.