Noname manuscript No.
(will be inserted by the editor)

On joint subtree distributions under two evolutionary models

Taoyang Wu · Kwok Pui Choi

November 5, 2015

Abstract In population and evolutionary biology, hypotheses about micro-evolutionary and macro-evolutionary processes are commonly tested by comparing the shape indices of empirical evolutionary trees with those predicted by neutral models. A key ingredient in this approach is the ability to compute and quantify distributions of various tree shape indices under random models of interest. As a step to meet this challenge, in this paper we investigate the joint distribution of cherries and pitchforks (that is, subtrees with two and three leaves) under two widely used null models: the Yule-Harding-Kingman (YHK) model and the proportional to distinguishable arrangements (PDA) model. Based on two novel recursive formulae, we propose a dynamic approach to numerically compute the exact joint distribution (and hence the marginal distributions) for trees of any size. We also obtained insights into the statistical properties of trees generated under these two models, including a constant correlation between the cherry and the pitchfork distributions under the YHK model, and the log-concavity and unimodality of the cherry distributions under both models. In addition, we show that there exists a unique change point for the cherry distributions between these two models.

Keywords phylogenetic tree · subtree distribution · Yule-Harding-Kingman model · PDA model · tree indices · joint distribution

T. Wu (✉)
School of Computing Sciences, University of East Anglia, Norwich, United Kingdom
E-mail: taoyang.wu@uea.ac.uk, taoyang.wu@gmail.com

K.P. Choi
Department of Statistics and Applied Probability, and Department of Mathematics, National University of Singapore, Singapore 117546,
E-mail: stackp@nus.edu.sg

1 Introduction

Phylogenetic tree shapes have been utilised to test evolutionary processes (see, e.g. Mooers and Heard, 1997; Nordborg, 2001; Blum and François, 2006; Purvis et al, 2011; Stadler, 2013), and more recently, to resolve disease transmission patterns (see, e.g. Colijn and Gardy, 2014). One challenge in these approaches is the ability to compute the distributions of various tree shape indices under the models of interest, which is needed in statistical testing for calculating the p-value of the empirical shape statistics or constructing a confidential interval. Even for some relatively simple null models, this can still be a challenging task. Many current approaches are based on approximating techniques, such as Monte Carlo sampling (see, e.g. Blum and François, 2006) or Gaussian approximation (see, e.g. McKenzie and Steel, 2000), which could be computationally intensive or restricting the tests to the trees above a certain size. Therefore it is desirable to explore efficient ways of computing these distributions exactly.

Two widely used null models for generating random trees in population and evolutionary biology are the Yule-Harding-Kingman (YHK) model (Harding, 1971; Yule, 1925; Kingman, 1982) and the proportional to different arrangements (PDA) model (Aldous, 2001). Under the PDA model all rooted binary trees of the same size are chosen with the same probability (Aldous, 2001) whilst under the YHK model each tree is chosen with a probability proportional to the number of total orderings that can be assigned to its internal nodes so that the relative partial ordering derived from the tree topology is preserved.

In this paper, we are interested in the exact computation of the joint distribution for the number of subtrees under the YHK and PDA model. Here a subtree, also known as a fringe subtree in Aldous (1991), consists of a node and all its descendants. More specifically, we study the distributions of the number of cherries, subtrees with two leaves, and that of pitchforks, subtrees with three leaves. Note that this is equivalent to study the joint distributions of 2-pronged and 3-pronged nodes as defined in (Rosenberg, 2006), as well as the joint distributions of clades of size two and three as defined in (Zhu et al, 2011).

We now describe the contents of the rest of this paper. In the next section we gather some necessary notation and background. In particular, we present a random tree generating process for realising both the YHK and PDA models as described in McKenzie and Steel (2000). In contrast to the splitting model that were used in several previous studies concerning the asymptotical distributions of subtrees (see, e.g. Chang and Fuchs, 2010), the process used here is based on iteratively attaching leaves. We therefore also collect some observations on the change of the numbers of cherries and pitchforks in a tree when an additional leaf is attached.

In Sections 3 and 4 we study subtree distributions under the YHK and the PDA models, respectively. Our main results include two novel recursive formulae on the joint distributions of cherries and pitchforks; see Theorem 1 for the one under the YHK model and Theorem 4 for the one under the PDA model. These recursions enable us to develop a dynamic approach to numerically compute the joint distributions, and hence also their marginal distributions, for trees of any size.

Rewritten in functional forms, the recursions also provide a way to compute the covariance and correlation of the joint distributions under these two models. Somewhat surprisingly, we find that under the YHK model the correlation between the cherry and the pitchfork distributions is a constant $-\sqrt{14/69}$, which is independent

of the number of leaves (see, e.g., Corollary 3). In contrast to currently methods developed respectively for the two models (see, e.g. Rosenberg, 2006; Chang and Fuchs, 2010), the recursions also lead to an alternative and arguably more unified approach to compute the moments of the cherry and the pitchfork distributions, and we demonstrate this by reaffirming several results obtained in previous studies.

Using the recursions on the cherry distribution derived from the joint distribution, we obtain in Theorem 6 the exact formula for the cherry distribution under the PDA model, and derive some interesting properties for cherry distributions, including that they are log-concave and hence unimodal under both models (see Theorems 3 and 7).

In Section 5 we present a comparative study of cherry and pitchfork distributions under the YHK and PDA models. We first compare the mean and the variance of these two distributions under these two models. Then we show in Theorem 8 that there exists a unique change point when comparing cherry distributions, that is, there exists a critical value τ_n for each $n \geq 4$ such that the probability that a random tree with n leaves generated under the YHK model contains k cherries is lower than that under the PDA model if $1 < k < \tau_n$, and higher if $\tau_n < k \leq n/2$. Finally, we conclude in Section 6 with discussions and some open problems.

2 Preliminaries

For later use, we present in this section some basic notation and results concerning phylogenetic trees. Throughout this article, X denotes a finite set with $|X| = n \geq 2$.

Phylogenetic trees A *phylogenetic tree* $T = (V(T), E(T))$ on X is a rooted tree with leaf set $L(T) = X$ such that the root has one child whilst all other vertices have either zero or two children (see Fig. 1 for an example). Note that in this paper phylogenetic trees are rooted, with their edges directed away from the root. In addition, for technical simplicity we assume without loss of generality that the root has one child (also referred to as planted phylogenetic trees by Baroni et al (2005)). Let $E^*(T)$ be the set of pendant edges in T, i.e., those edges incident with a leaf. Then we have $|E(T)| = 2n - 1$ and $|E^*(T)| = n$.

Let e be an edge in a phylogenetic tree T. The tree consisting of e and all edges below e is called a *subtree* of T, and is denoted by $T(e)$. In particular, a *cherry* is a subtree with two leaves, and a *pitchfork* is a subtree with three leaves. The number of cherries and pitchforks contained in T are denoted by $C(T)$ and $A(T)$, respectively. Note first that we always have $1 \leq C(T) \leq n/2$ and $0 \leq A(T) \leq n/3$. Moreover, in our definition a cherry contains three edges and a pitchfork contains five edges. As an example, for the tree T depicted in Fig. 1 we have $C(T) = 2$ and $A(T) = 1$. In addition, $T(e_8)$ is a pitchfork with edge set $\{e_1, e_3, e_5, e_7, e_8\}$, and $T(e_7)$ is a cherry with edge set $\{e_1, e_5, e_7\}$. Finally, $C(T)$ and $A(T)$ are respectively equal to the number of 2-pronged nodes and 3-pronged nodes contained in T (see Rosenberg (2006) for the definitions of r-pronged nodes).

Given an edge e in a phylogenetic tree T and a taxon $x_0 \notin L(T)$, let $T[e; x_0]$ be the phylogenetic tree obtained from T by attaching a new leaf labelled with x_0 to the edge e. Formally, let $e = \{u, v\}$ and let w be a vertex not contained in $V(T)$, then $T[e; x_0]$ has vertex set $V(T) \cup \{x_0, w\}$ and edge set $\big(E(T) \setminus \{e\}\big) \cup$

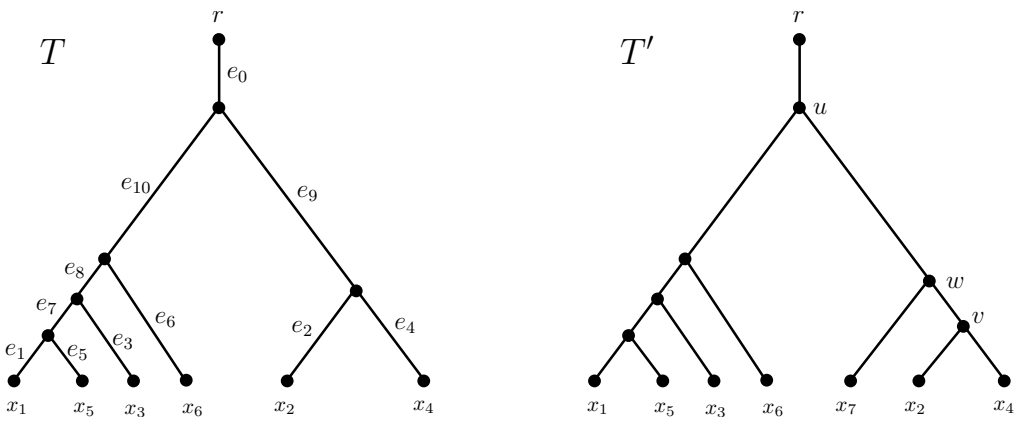

Fig. 1: Examples of phylogenetic trees. T is a phylogenetic tree on $X = \{x_1, \ldots, x_6\}$, and $T' = T[e_9; x_7]$ is a phylogenetic tree on $\{x_1, \ldots, x_7\}$ that is obtained from T by attaching the leaf labelled x_7 to edge e_9. Here the directions of all edges are directed away from the root r, and hence omitted for simplicity.

$\{(u, w), (v, w), (w, x_0)\}$ (see Fig. 1 for an illustration of this construction). When the labelling of the new leaf is clear from the context, $T[e; x_0]$ is abbreviated to $T[e]$.

The YHK and the PDA model In this subsection, we present a formal definition of the two null models investigated in this paper: the *proportional to distinguishable arrangements* (PDA) model and the *Yule–Harding–Kingman* (YHK) model. In contrast to the splitting process used by Aldous (2001) to accommodate the two models, the random process used here is based on iteratively attaching leaves.

Under the Yule–Harding model (Harding, 1971; Yule, 1925), a rooted phylogenetic tree on X is generated as follows. Beginning with the tree with two leaves, we "grow" it by repeatedly uniformly sampling a pendant edge e in the current tree T_{cur} and replace T_{cur} by $T_{cur}[e]$. This process continues until a binary tree with n leaves is obtained. Finally, we label each of its leaves with a label sampled randomly uniformly (without replacement) from $\{x_1, \ldots, x_n\}$. When branch lengths are ignored, the Yule–Harding model is shown by Aldous (1996) to be equivalent to the trees generated by the coalescent process, a backward tree generating process that is widely used in population genetics (Kingman, 1982), and so we call it the YHK model. The probability of generating a tree T under this model is denoted by $\mathbb{P}_y(T)$.

Let $\mathcal{T}_n$ be the set of phylogenetic trees with leaf set $\{x_1, \ldots, x_n\}$. It is well known that the number of trees contained in $\mathcal{T}_n$ is $\varphi(n) := (2n-3)!! = 1 \times 3 \times \cdots \times (2n-3)$ (see e.g. Semple and Steel, 2003). Here we adopt the convention that $\varphi(1) = 1$. Under the PDA model, each tree has the same probability, that is, $1/\varphi(n)$, to be generated. Alternatively, a tree can be generated under the PDA model using a Markov process similar to the one used in the YHK model; the only difference is that the edge e is uniformly sampled from $E(T)$, instead of $E^*(T)$ (see, e.g., McKenzie and Steel, 2000). We use $\mathbb{E}_y$, $\mathbb{V}_y$, Cov_y and ρ_y to denote respectively the expectation, variance, covariance and correlation taken with respect to the probability measure $\mathbb{P}_y$ under the YHK model. Similarly, $\mathbb{E}_u$, $\mathbb{V}_u$, Cov_u and ρ_u are defined with respect to the probability $\mathbb{P}_u$ under the PDA model.

For $n \geq 2$, let A_n (resp. C_n) be the random variable $A(T)$ (resp. $C(T)$) for a random tree T in $\mathcal{T}_n$. In this paper, we are interested in the joint distributions and the marginal properties of A_n and C_n under the YHK and the PDA models.

Subtree Pattern For later use, we present in this subsection several technical results concerning the change of the numbers of cherries and pitchforks when a new leaf is attached to a phylogenetic tree.

We begin with the following notation. Given a phylogenetic tree T, let $E_1(T)$ be the set of pendant edges that are contained in a pitchfork but not a cherry; $E_2(T)$ the set of edges in T that are contained in a cherry but not in a pitchfork (note that in our notation a cherry contains three leaves); $E_3(T)$ the set of pendant edges that are contained in neither a cherry nor a pitchfork; and $E_4(T) = E(T) \setminus (E_1(T) \cup E_2(T) \cup E_3(T))$. For instance, for the tree T depicted in Fig. 1, we have $E_1(T) = \{e_3\}$, $E_2(T) = \{e_2, e_4, e_9\}$, $E_3(T) = \{e_6\}$ and $E_4(T) = \{e_0, e_1, e_5, e_7, e_8, e_{10}\}$. In addition, $E(T)$ can be decomposed into the disjoint union of these four sets of edges. The following lemma, whose proof is straightforward and hence omitted here, shows this observation holds for all phylogenetic trees, where $\sqcup$ denotes disjoint union.

Lemma 1 *Suppose that T is a phylogenetic tree with n leaves. Then we have*

$$E(T) = E_1(T) \sqcup E_2(T) \sqcup E_3(T) \sqcup E_4(T). \tag{1}$$

In addition, we have $|E_1(T)| = A(T)$, $|E_2(T)| = 3(C(T) - A(T))$, $|E_3(T)| = n - A(T) - 2C(T)$, and $|E_4(T)| = n - 1 + 3A(T) - C(T)$.

The last lemma provides a decomposition for the set of edges in a phylogenetic tree, which is useful to the study of the PDA model. For the YHK model, we need an analogous decomposition for $E^*(T)$, the set of the pendant edges in T. To this end, note first that we have $E_1(T) \subseteq E^*(T)$ and $E_3(T) \subseteq E^*(T)$. In addition, let $E_i^*(T) := E_i(T) \cap E^*(T)$ be the set of pendant edges in $E_i(T)$ for $i = 2, 4$. Then we have the following lemma, whose proof is straightforward and hence omitted.

Lemma 2 *Suppose that T is a phylogenetic tree with n leaves. Then we have*

$$E^*(T) = E_1(T) \sqcup E_2^*(T) \sqcup E_3(T) \sqcup E_4^*(T). \tag{2}$$

In addition, we have $|E_2^*(T)| = 2(C(T) - A(T))$ and $|E_4^*(T)| = 2A(T)$.

We end this section with the following result relating the values $C(T[e]) - C(T)$ and $A(T[e]) - A(T)$ to the choice of e.

Proposition 1 *Suppose that e is an edge in a phylogenetic tree T and $T' = T[e]$. Then we have*

$$A(T') = \begin{cases} A(T) & \text{if } e \in E_3(T) \cup E_4(T), \\ A(T) - 1 & \text{if } e \in E_1(T), \\ A(T) + 1 & \text{if } e \in E_2(T); \end{cases} \quad \text{and} \quad C(T') = \begin{cases} C(T) & \text{if } e \in E_2(T) \cup E_4(T), \\ \\ C(T) + 1 & \text{if } e \in E_1(T) \cup E_3(T). \end{cases}$$

Proof Let $\{F_1, \ldots, F_k\}$ be the set of pitchforks contained in T, and let $\{H_1, \ldots, H_l\}$ $(l \geq k)$ be the set of cherries contained in T. Here we may assume that indices are chosen in the way so that H_1 is contained in F_1.

Suppose first that $e = (u, v) \in E_1(T)$. Swapping the labelling of F_i if necessary, we may assume that e is the pendant edge contained in the pitchfork F_1 but not in the cherry H_1. Let u_0 be the parent of u, and let u_1 be the child of u that is distinct from v. In addition, let w be the newly added interior vertex in T'. Now consider $e_0 = (u_0, u)$ and $e' = (u, w)$ in T'. Then $T'(e_0)$ is not a pitchfork as u_0 has four leaves as its descendants. On the other hand, $T'(e')$ is a cherry of T' that is not contained in T. Therefore, we have $A(T') = k - 1$ and $C(T') = l + 1$, as required.

By a similar argument, we can establish the proposition for the other three cases, i.e., $e \in E_i(T)$ for $2 \leq i \leq 4$. Since by Lemma 1 these four cases cover all possible choices of e, the proposition follows. □

One useful consequence of the last proposition is the following corollary, whose proof is straightforward and hence omitted.

Corollary 1 *Suppose that e is an edge in a phylogenetic tree T with $A(T) = a$ and $C(T) = b$. Then for the phylogenetic tree $T' = T[e]$, we have*

$$(A(T'), C(T')) \in \{(a-1, b+1), (a+1, b), (a, b+1), (a, b)\}$$

according to the index i $(1 \leq i \leq 4)$ with $e \in E_i(T)$.

3 Subtree Distributions under the YHK Model

In this section, we study the distributions of the random variables A_n (i.e., the number of pitchforks) and C_n (i.e., the number of cherries) under the YHK model. Our starting point is the following recursion on their joint distribution.

Theorem 1 *We have*

$$\begin{aligned}\mathbb{P}_y(A_{n+1} = a, C_{n+1} = b) = \frac{2a}{n}\mathbb{P}_y(A_n = a, C_n = b) + \frac{(a+1)}{n}\mathbb{P}_y(A_n = a+1, C_n = b-1) \\ + \frac{2(b-a+1)}{n}\mathbb{P}_y(A_n = a-1, C_n = b) + \frac{(n-a-2b+2)}{n}\mathbb{P}_y(A_n = a, C_n = b-1)\end{aligned} \tag{3}$$

for $n > 3$ and $1 < b < n$. Moreover, $\mathbb{P}_y(A_3 = a, C_3 = b)$ equals to 1 if $(a, b) = (1, 1)$, and 0 otherwise.

Proof Fix $n > 3$, and let $T_2, \ldots, T_n, T_{n+1}$ be a sequence of random trees generated by the YHK process, that is, T_2 contains two leaves and $T_{i+1} = T_i[e_i]$ for a uniformly chosen pendant edge e_i in T_i for $2 \leq i \leq n$. In particular, we have $|E^*(T_i)| = i$ for $2 \leq i \leq n+1$. Then we have

$$\begin{aligned}\mathbb{P}_y(A_{n+1} = a, C_{n+1} = b) &= \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b) \\ &= \sum_{p,q} \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \,|\, A(T_n) = p, C(T_n) = q)\mathbb{P}(A(T_n) = p, C(T_n) = q) \\ &= \sum_{p,q} \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \,|\, A(T_n) = p, C(T_n) = q)\mathbb{P}_y(A_n = p, C_n = q),\end{aligned} \tag{4}$$

where the first and second equalities follow from the law of total probability, and the definition of random variables A_n and C_n.

Let e_n be the pendant edge in T_n chosen in the above YHK process for generating T_{n+1}, that is, $T_{n+1} = T_n[e_n]$. Since Corollary 1 implies that

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = p, C(T_n) = q) = 0 \tag{5}$$

for $(p, q) \notin \{(a, b), (a + 1, b - 1), (a - 1, b), (a, b - 1)\}$, it suffices to consider the following four cases in the summation in (4): case (i): $p = a, q = b$; case (ii): $p = a + 1, q = b - 1$; case (iii): $p = a - 1, q = b$; and case (iv): $p = a, q = b - 1$.

Firstly, Proposition 1 implies that case (i) occurs if and only if $e_n \in E_4(T_n) \cap E^*(T_n) = E_4^*(T_n)$. Together with Lemma 2, we have

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = a, C(T_n) = b) = \frac{|E_4^*(T_n)|}{|E^*(T_n)|} = \frac{2A(T_n)}{n} = \frac{2a}{n}. \tag{6}$$

Similarly, Proposition 1 implies that case (ii) occurs if and only if $e_n \in E_1(T_n) \cap E^*(T_n) = E_1(T_n)$. Hence by Lemma 1 we have

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = a + 1, C(T_n) = b - 1) = \frac{|E_1(T_n)|}{|E^*(T_n)|} = \frac{a + 1}{n}. \tag{7}$$

Next, Proposition 1 implies case (iii) occurs if and only if $e_n \in E_2(T_n) \cap E^*(T_n) = E_2(T_n)$. Hence using Lemma 1 we have

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = a + 1, C(T_n) = b) = \frac{|E_2(T_n)|}{|E^*(T_n)|} = \frac{2(b - a - 1)}{n}. \tag{8}$$

Finally, by Proposition 1 case (iv) occurs if and only if e_n is contained in $E_3(T_n) \cap E^*(T_n) = E_3^*(T_n)$. Hence by Lemma 2 it follows that

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1} = b) \mid A(T_n) = a, C(T_n) = b - 1) = \frac{|E_3^*(T_n)|}{|E^*(T_n)|} = \frac{n - a - 2b + 2}{n}. \tag{9}$$

Now substituting Eq. (6)–(9) into Eq. (4) completes the proof of the theorem. □

The recursion in the last theorem can be used for a dynamic approach to numerically compute the joint distribution of A_n and C_n . More precisely, let M_m $(m \geq 3)$ be the $(m+1) \times (m+1)$ matrix whose (i, j)-entry is $\mathbb{P}_y(A_m = i - 1, C_m = j - 1)$. Then M_3 contains a unique non-zero entry, which is at position $(2, 2)$ and has a value of 1 . Next, starting with $m = 4$ and assuming that M_{m-1} is already constructed, each entry in M_m can be computed using time $O(1)$, and hence M_m can be constructed in time $O(m^2)$ with M_{m-1} given. In other words, M_n, which specifies the joint distribution of cherry and pitchfork under the YHK model, can be computed in $O(n^3)$ (see Fig. 2 for the contour plots with $n = 50$ and $n = 200$). Note that an alternative way of computing the joint distribution of A_n and C_n under the YHK model is proposed in Disanto and Wiehe (2013), which is based on integrating and differentiating generating functions.

For later use, we rewrite the recursion in Theorem 1 in the following functional form.

Theorem 2 *Let $\varphi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be an arbitrary function. Then, under the YHK model, we have*

$$\begin{aligned}\mathbb{E}_y\varphi(A_{n+1}, C_{n+1}) = {} & \frac{2}{n}\mathbb{E}_y[A_n\, \varphi(A_n, C_n)] + \frac{1}{n}\mathbb{E}_y[A_n\, \varphi(A_n - 1, C_n + 1)] \\ & + \frac{2}{n}\mathbb{E}_y[(C_n - A_n)\, \varphi(A_n + 1, C_n)] + \frac{1}{n}\mathbb{E}_y[(n - A_n - 2C_n)\, \varphi(A_n, C_n + 1)]\end{aligned}$$

for $n > 2$.

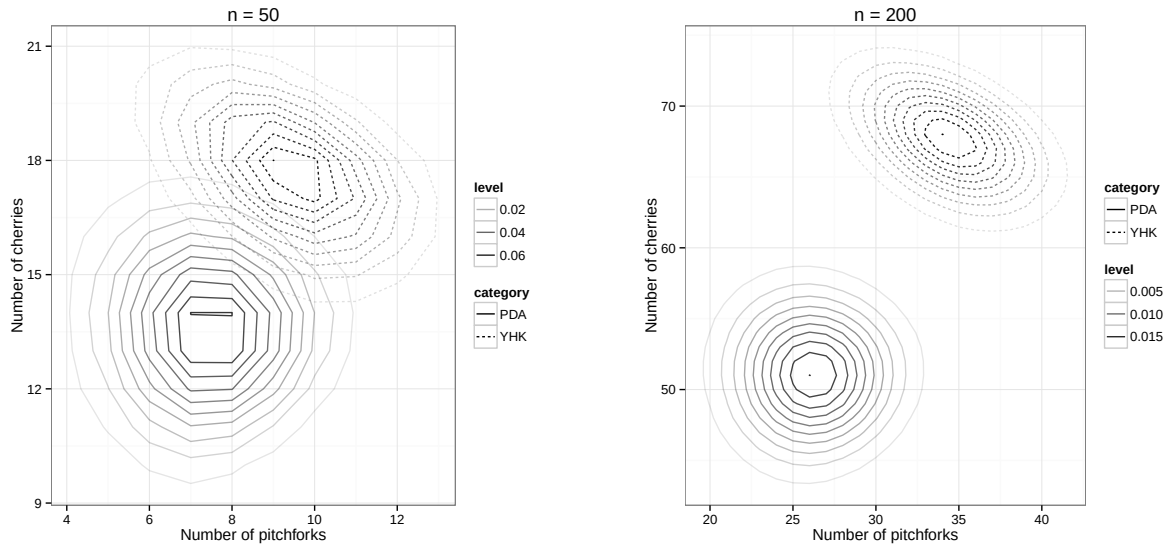

Fig. 2: Contour plots of the probability density functions for the joint distribution of cherries and pitchforks on phylogenetic trees with 50 and 200 leaves. The density functions are computed using a dynamic approach based on the recursions in Theorems 1 and 4. The polygonal contours arise because the joint distribution is defined only on integer lattice points.

Proof Consider the indicator function $I_{(a,b)}$ on $\mathbb{R}\times\mathbb{R}$ defined as

$$I_{(a,b)}(x,y)=\begin{cases}1 & \text{if } x=a \text{ and } y=b,\\ 0 & \text{otherwise.}\end{cases}$$

We multiply Eq. (3) in Theorem 1 by $\varphi(a,b)$ and rewrite them as follows

$$\begin{aligned}\mathbb{E}_y[\varphi(A_{n+1},C_{n+1})I_{(a,b)}(A_{n+1},C_{n+1})] &= \frac{1}{n}\big\{2\mathbb{E}_y[A_n\varphi(A_n,C_n)I_{(a,b)}(A_n,C_n)]\\ &+\mathbb{E}_y[A_n\varphi(A_n-1,C_n+1)I_{(a,b)}(A_n-1,C_n+1)]+2\mathbb{E}_y[(C_n-A_n)\varphi(A_n+1,C_n)I_{(a,b)}(A_n+1,C_n)]\\ &+\mathbb{E}_y[(n-A_n-2C_n)\varphi(A_n,C_n+1)I_{(a,b)}(A_n,C_n+1)]\big\}.\end{aligned}$$

Summing over all a and b completes the proof. □

In the remainder of this section we study cherry and pitchfork distributions using Theorem 2. We begin with a functional recursion on the cherry distribution C_n. This enables us to show that the cherry distribution is log-concave under the YHK model, and obtain an alternative approach to computing the central moments of cherry distribution.

Proposition 2 *Let $\psi:\mathbb{R}\to\mathbb{R}$ be an arbitrary function. Then we have*

$$\mathbb{E}_y\psi(C_{n+1})=\frac{1}{n}\mathbb{E}_y[2C_n\,\psi(C_n)+(n-2C_n)\,\psi(C_n+1)] \tag{10}$$

for $n>2$. In particular, we have $\mathbb{P}_y(C_2=1)=1$, $\mathbb{P}_y(C_2=k)=0$ for $k\neq 1$, and

$$\mathbb{P}_y(C_{n+1}=k)=\frac{2k}{n}\mathbb{P}_y(C_n=k)+\frac{n-2k+2}{n}\mathbb{P}_y(C_n=k-1) \tag{11}$$

for $n>2$ and $1<k<n$.

Proof For $\psi : \mathbb{R} \to \mathbb{R}$ given in the statement of the proposition, we define $\varphi^*(x, y) = \psi(y)$, a function on $\mathbb{R} \times \mathbb{R}$. Applying Theorem 2 to the function φ^* leads to Eq. (10). Eq. (11) follows from Eq. (10) by taking $\psi(x) = I_k(x)$, where $I_k(x)$ equals 1 if $x = k$, and 0 otherwise. □

Using the last proposition, the mean and the variance of cherry distribution can be obtained by substituting $\psi(x) = x$ and $\psi(x) = x^2$, respectively, in the recursive equation Eq. (10) in Proposition 2.

Corollary 2 *(Heard, 1992; McKenzie and Steel, 2000) We have* $\mathbb{E}_y(C_n) = n/3$ *for* $n > 2$ *and* $\mathbb{V}_y(C_n) = 2n/45$ *for* $n \geq 5$.

Recall that a sequence of numbers, $\{y_1, \ldots, y_m\}$, is said to be *positive* if each number in the sequence is greater than zero. It is called *log-concave* if $y_{k-1}y_{k+1} \leq y_k^2$ holds for $2 \leq k \leq m-1$. Clearly, a positive sequence $\{y_k\}_{1\leq k\leq m}$ is log-concave if and only if the sequence $\{y_k/y_{k+1}\}_{1\leq k\leq m-1}$ is increasing. Therefore, a log-concave sequence is necessarily *unimodal*, that is, there exists an index $1 \leq k \leq m$ such that

$$y_1 \leq y_2 \leq \cdots \leq y_k \geq y_{k+1} \geq \cdots \geq y_m \tag{12}$$

holds. Finally, a non-negative integer valued random variable Y with probability mass function $\{p_k : k \geq 0\}$ is log-concave if $\{p_k\}_{k\geq 0}$ is a log-concave sequence.

To show that the probability density function of C_n is log-concave, we need the following lemma.

Lemma 3 *Let* z_1, z_2, z_3, z_4 *be four positive numbers with* $z_2^2 \geq z_1z_3$ *and* $z_3^2 \geq z_2z_4$. *Then we have*

$$z_2z_3 \geq z_1z_4 \qquad \textit{and} \qquad z_1z_3 + z_2z_4 \geq 2z_1z_4.$$

Proof Since z_i are positive for $1 \leq i \leq 4$, from $z_2^2 \geq z_1z_3$ and $z_3^2 \geq z_2z_4$ if follows that

$$\frac{z_2}{z_1} \geq \frac{z_3}{z_2} \geq \frac{z_4}{z_3}.$$

Hence we have

$$z_2z_3 \geq z_1z_4, \tag{13}$$

which completes the proof of the first inequality in the lemma.

To prove the second inequality in the lemma, we consider the following two cases.

Case 1: $z_1 \geq z_2$. Together with $z_2^2 \geq z_1z_3$, this implies $z_2 \geq z_3$, and hence $z_3 \geq z_4$ in view of $z_3^2 \geq z_2z_4$. Therefore, we have

$$(z_1 - z_2)(z_3 - z_4) \geq 0.$$

This leads to $z_1z_3 + z_2z_4 \geq z_1z_4 + z_2z_3 \geq 2z_1z_4$, where the last inequality follows from Eq. (13).

Case 2: $z_1 < z_2$. If $z_3 \leq z_4$, then we have $(z_1 - z_2)(z_3 - z_4) \geq 0$, and hence $z_1z_3 + z_2z_4 \geq z_1z_4 + z_2z_3 \geq 2z_1z_4$, as required. Therefore, we may assume that $z_3 > z_4$. This implies $z_1z_3 \geq z_1z_4$ and $z_2z_4 \geq z_1z_4$, and hence $z_1z_3 + z_2z_4 \geq 2z_1z_4$, as required. □

Using the last lemma, we present the following theorem concerning the log-concavity of the cherry distribution under the YHK model.

Theorem 3 *Under the YHK model, we have*

$$\mathbb{P}_y(C_n = k)^2 \geq \mathbb{P}_y(C_n = k+1)\mathbb{P}_y(C_n = k-1) \tag{14}$$

for $n > 2$ *and* $1 < k < n$.

Proof For simplicity, we put $a_{n,k} := \mathbb{P}_y(C_n = k)$. We prove this theorem by induction; the basic case $n = 3$ is straight-forward. Now assuming that $n \geq 3$ and Eq. (14) holds for all $1 < k < n$, it suffices to show that

$$a_{n+1,k}^2 \geq a_{n+1,k-1}a_{n+1,k+1} \tag{15}$$

for all $1 < k \leq n$. Using the recursion described in Eq. (11), we have

$$a_{n+1,k}^2 = 4k^2a_{n,k}^2 + (n+2-2k)^2a_{n,k-1}^2 + 4k(n+2-2k)a_{n,k}a_{n,k-1}$$

and $a_{n+1,k-1}a_{n+1,k+1}$ is equal to

$$\begin{aligned}(2k+2)(2k-2)a_{n,k-1}a_{n,k+1} + (2k+2)(n-2k+4)a_{n,k-2}a_{n,k+1}\\ + (n-2k)(2k-2)a_{n,k}a_{n,k-1} + (n-2k)(n-2k+4)a_{n,k}a_{n,k-2}.\end{aligned}$$

Therefore, by the inductive assumption and Lemma 3 we have

$$\begin{aligned}&a_{n+1,k}^2 - a_{n+1,k-1}a_{n+1,k+1}\\ &= 2[k(n-2k)+(n+2k)](a_{n,k}a_{n,k-1} - a_{n,k-2}a_{n,k+1}) + 4k^2(a_{n,k}^2 - a_{n,k-1}a_{n,k+1})\\ &\quad + 4a_{n,k-1}(a_{n,k-1} - a_{n,k-2}) + 4(n+2-2k)^2(a_{n,k+1}^2 - a_{n,k}a_{n,k-2}) + 4a_{n,k-2}(a_{n,k} - a_{n,k+1})\\ &\geq 4a_{n,k+1}(a_{n,k-1} - a_{n,k-2}) + 4a_{n,k-2}(a_{n,k} - a_{n,k+1})\\ &= 4(a_{n,k-1}a_{n,k+1} + a_{n,k}a_{n,k-2} - 2a_{n,k-2}a_{n,k+1}) \geq 0,\end{aligned}$$

from which Eq. (15) follows, as required. □

In the next result we compute the mean and the variance of pitchfork distribution A_n under the YHK model, and calculate the covariance and correlation of A_n and C_n. Note that the mean and the variance of A_n was also obtained by Rosenberg (2006, Theorem 4.4). Since the proof is similar to that of Corollary 2, we only outline the main step here.

Proposition 3 *For* $n \geq 7$ *we have*

$$\mathbb{E}_y(A_n) = \frac{n}{6}, \quad Cov_y(A_n, C_n) = -\frac{n}{45}, \quad \textit{and} \quad \mathbb{V}_y(A_n) = \frac{23n}{420}. \tag{16}$$

Proof Applying Theorem 2 to $\varphi(x, y) = x$ and using Corollary 2, it follows that

$$\begin{aligned}\mathbb{E}_y(A_{n+1}) &= \frac{1}{n}\mathbb{E}_y[2A_n^2 + A_n(A_n-1) + 2(C_n - A_n)(A_n+1) + (n - A_n - 2C_n)A_n]\\ &= \frac{2}{3} + \frac{n-3}{n}\mathbb{E}_y(A_n)\end{aligned}$$

holds for $n > 2$. Together with $\mathbb{E}_y(A_3) = 1$, we have $\mathbb{E}_y(A_n) = n/6$ for $n \geq 4$, as required.

Next, applying Theorem 2 to the function $\varphi(x,y)=xy$ shows that

$$\mathbb{E}_y(A_{n+1}C_{n+1})=\frac{n-5}{n}\mathbb{E}_y(A_nC_n)+\frac{n-1}{n}\mathbb{E}_y(A_n)+\frac{2}{n}\mathbb{E}_y(C_n^2)$$

holds for $n>2$. By Corollary 2 and $\mathbb{E}_y(A_n)=n/6$ it follows that

$$\begin{aligned}Cov_y(A_{n+1},C_{n+1})&=\mathbb{E}_y(A_{n+1}C_{n+1})-\frac{(n+1)^2}{18}\\&=\frac{n-5}{n}\mathbb{E}_y(A_nC_n)+\frac{n-1}{6}+\frac{2(5n+2)}{45}-\frac{(n+1)^2}{18}\\&=\frac{n-5}{n}Cov_y(A_n,C_n)-\frac{2}{15}\end{aligned}$$

holds for $n\geq 5$. Solving the last recursion equation, we obtain $Cov_y(A_n,C_n)=-n/45$ for $n\geq 6$, as required.

Now the formula on $\mathbb{V}_y(A_n)$ can be established by an argument similar to that for $Cov_y(A_n,C_n)$ by applying Theorem 2 to the function $\varphi(x,y)=x^2$. □

Interestingly, the last proposition implies that the correlation coefficient between the cherry and pitchfork distribution under the YHK model is a negative constant for $n\geq 7$. Note that negative correlation is to be expected as the more cherries are found in a tree, the more likely that there are fewer pitchforks in that tree.

Corollary 3 *Under the YHK model, the correlation coefficient $\rho_y(A_n,C_n)$ between A_n and C_n is $-\sqrt{14/69}$, which is independent of n for $n\geq 7$.*

Proof The proposition follows directly from Corollary 2 and Proposition 3. □

4 Subtree Distributions under the PDA model

In this section, we shall investigate the cherry and pitchfork distributions under the PDA model. Similar to the study on the YHK model in Section 3, our starting point is the following recursion relating the joint distribution of cherries and pitchforks.

Theorem 4 *We have*

$$\begin{aligned}\mathbb{P}_u(A_{n+1}=a,C_{n+1}=b)&=\frac{n+3a-b-1}{2n-1}\mathbb{P}_u(A_n=a,C_n=b)+\frac{a+1}{2n-1}\mathbb{P}_u(A_n=a+1,C_n=b-1)\\&+\frac{3(b-a+1)}{2n-1}\mathbb{P}_u(A_n=a-1,C_n=b)+\frac{n-a-2b+2}{2n-1}\mathbb{P}_u(A_n=a,C_n=b-1)\end{aligned}$$

for $n>3$ and $1<b<n$. Moreover, $\mathbb{P}_u(A_3=a,C_3=b)$ equals to 1 if $(a,b)=(1,1)$ and 0 otherwise.

Proof We give a sketch of the proof as it is similar to the proof of Theorem 1.

The only modifications needed are the conditional probabilities in the four cases there. For case (i), by Proposition 1 this case occurs if and only if $e_n\in E_4(T_n)$, and hence the conditional probability is $|E_4(T_n)|/|E(T_n)|=(n+3a-b-1)/(2n-1)$ by Lemma 1. Using similar arguments, for case (ii), the conditional probability is $|E_1(T_n)|/|E(T_n)|=(a+1)/(2n-1)$. For case (iii), the conditional probability is $|E_2(T_n)|/|E(T_n)|=3(b-a+1)(2n-1)$. Finally, for case (iv), the conditional probability is $|E_3(T_n)/|E(T_n)|=(n-a-2b+2)/(2n-1)$. The rest of the proof proceeds as in the proof of Theorem 1. □

Using an approach similar to the remark after Theorem 1, the last theorem leads to a dynamic programming approach to compute the joint distribution of cherry and pitchfork (see Fig. 2 for the contour plots with $n = 50$ and $n = 200$). In addition, we present the following result which will enable us to study the moments of A_n and C_n, whose proof is similar to that of Theorem 2 and hence omitted.

Theorem 5 *Let $\varphi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be an arbitrary function. For $n > 3$ we have*

$$\begin{aligned}
&\mathbb{E}_u\varphi(A_{n+1}, C_{n+1}) \\
=&\frac{1}{2n-1}\mathbb{E}_u[(n + 3A_n - C_n - 1)\,\varphi(A_n, C_n)] + \frac{1}{2n-1}\mathbb{E}_u[A_n\,\varphi(A_n - 1, C_n + 1)] \\
&+ \frac{3}{2n-1}\mathbb{E}_u[(C_n - A_n)\,\varphi(A_n + 1, C_n)] + \frac{1}{2n-1}\mathbb{E}_u[(n - A_n - 2C_n)\,\varphi(A_n, C_n + 1)].
\end{aligned}$$

In the remainder of this section we shall apply Theorem 5 to study cherry and pitchfork distributions under the PDA model. To begin with, we present the following functional recursion between cherry distributions, which will enable us to obtain the exact formula for cherry distributions and show that cherry distribution is log-concave under this model.

Proposition 4 *Let $\psi : \mathbb{R} \to \mathbb{R}$ be an arbitrary function. Then for $n > 2$ we have*

$$\mathbb{E}_u\psi(C_{n+1}) = \frac{1}{2n-1}\mathbb{E}_u[(n + 2C_n - 1)\,\psi(C_n)] + \frac{1}{2n-1}\mathbb{E}_u[(n - 2C_n)\,\psi(C_n + 1)] \tag{17}$$

and

$$\mathbb{P}_u(C_{n+1} = k) = \frac{n + 2k - 1}{2n-1}\mathbb{P}_u(C_n = k) + \frac{n - 2k + 2}{2n-1}\mathbb{P}_u(C_n = k - 1), \quad 1 \le k < n. \tag{18}$$

Proof Let $\psi : \mathbb{R} \to \mathbb{R}$ be an arbitrary function as in the statement of the proposition. Then $\varphi^*(x, y) = \psi(y)$ is an function on $\mathbb{R} \times \mathbb{R}$. Now applying Theorem 5 to the function φ^* leads to Eq. (17). Finally, Eq. (18) follows from Eq. (17) by taking $\psi(x) = I_k(x)$, where $I_k(x)$ equals 1 if $x = k$, and 0 otherwise. □

Note that the recursion presented in the last proposition enables us to study the moments of cherry distribution under the PDA model. As an example, we present below an alternative computation for the mean and the variance of C_n. Since the techniques used to solve difference equations under this model is rather different from that used under the YHK model (i.e., Corollary 2), a complete proof is included here. Note that in the proof we will use the following well-known Faulhaber's formulae (also known as Bernoulli's formulae) concerning the sum of powers of integers (see e.g. Conway and Guy, 1996).

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}, \qquad \sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4},$$
$$\sum_{i=1}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}, \qquad \sum_{i=1}^{n} i^5 = \frac{n^2(n+1)^2(2n^2+2n-1)}{12}.$$

Corollary 4 *(Chang and Fuchs, 2010, Proposition 5) For $n \ge 2$ we have*

$$\mathbb{E}_u(C_n) = \frac{n(n-1)}{2(2n-3)} \sim \frac{n}{4} \quad \textit{and} \quad \mathbb{V}_u(C_n) = \frac{n(n-1)(n-2)(n-3)}{2(2n-3)^2(2n-5)} \sim \frac{n}{16}.$$

Proof We may assume that $n \geq 3$ in the remainder of the proof as the case $n = 2$ clearly holds. Substituting $\psi(x) = x$ in the recursive equation Eq. (17) in Proposition 4 leads to that

$$\mathbb{E}_u(C_{n+1}) = \frac{1}{2n-1}\mathbb{E}_u\big[(n + 2C_n - 1)C_n + (n - 2C_n)(C_n + 1)\big] = \frac{n}{2n-1} + \frac{2n-3}{2n-1}\mathbb{E}_u(C_n)$$

holds for $n > 2$. Together with the initial condition $\mathbb{E}_u(C_2) = 1$, multiplying the both sides of the last difference equation on $\mathbb{E}_u(C_n)$ by $2n - 1$ and solving it leads to

$$(2n-3)\mathbb{E}_u(C_n) = 1 + \cdots + (n-1) = \frac{n(n-1)}{2}$$

for $n > 2$, as required.

For simplicity, let $f(n) = (2n-3)(2n-5)$ and $g(n) = (n-1)(n^2-2n-1)$. Then

$$\sum_{k=1}^{n} g(k) = \sum_{k=1}^{n}(k^3 - 3k^2 + k + 1) = \frac{n(n-1)(n^2-n-4)}{4}.$$

Next, applying Proposition 4 to the function $\psi(x) = x^2$ implies that

$$\begin{aligned}\mathbb{E}_u(C_{n+1}^2) &= \frac{1}{2n-1}\mathbb{E}_u\left[(n + 2C_n - 1)C_n^2 + (n - 2C_n)(C_n + 1)^2\right] \\ &= \frac{n}{2n-1} + \frac{2n-2}{2n-1}\mathbb{E}_u(C_n) + \frac{2n-5}{2n-1}\mathbb{E}_u(C_n^2) \\ &= \frac{g(n+1)}{(2n-1)(2n-3)} + \frac{2n-5}{2n-1}\mathbb{E}_u(C_n^2)\end{aligned}$$

holds for $n > 2$. Now multiplying $f(n+1)$ on both sides of the above recursion leads to

$$f(n)\mathbb{E}_u(C_n^2) - f(n-1)\mathbb{E}_u(C_{n-1}^2) = g(n)$$

for $n \geq 3$. Since $\mathbb{E}_u(C_2^2) = 1 = -g(2)$ and $g(1) = 0$, we have

$$(2n-3)(2n-5)\mathbb{E}_u(C_n^2) = f(n)\mathbb{E}_u(C_n^2) = \sum_{k=1}^{n} g(k) = \frac{n(n-1)(n^2-n-4)}{4}$$

for $n \geq 3$, from which we have

$$\mathbb{E}_u(C_n^2) = \frac{n(n-1)(n^2-n-4)}{4(2n-3)(2n-5)} \tag{19}$$

and hence $\mathbb{V}_y(C_n)$ follows. □

Another consequence of the recursion in Proposition 4 is the following exact formula on the cherry distribution for the PDA model, whose proof is a straightforward application of induction and hence omitted here.

Theorem 6 *For $n \geq 2$ and $1 \leq k \leq n/2$ we have*

$$\mathbb{P}_u(C_n = k) = \frac{n!(n-1)!(n-2)!2^{n-2k}}{(n-2k)!(2n-2)!k!(k-1)!}\ . \tag{20}$$

Interestingly, a similar formula for unrooted trees was obtained by Hendy and Penny (1982), that is, the probability that a random tree generated by the PDA model contains exactly k cherries is

$$\frac{n!(n-2)!(n-4)!2^{n-2k}}{(n-2k)!(2n-4)!k!(k-2)!} \tag{21}$$

for $2 \leq k \leq n/2$ (see, also McKenzie and Steel, 2000, Theorem 4). A direct consequence of Theorem 6 is that the cherry distribution under the PDA model is log-concave, and hence also unimodal.

Theorem 7 *For $n \geq 2$ and $1 < k < n$ we have*

$$\mathbb{P}_u(C_n = k)^2 \geq \mathbb{P}_u(C_n = k+1)\mathbb{P}_u(C_n = k-1). \tag{22}$$

Moreover, let $\Delta(n) = \frac{(n+1)(n+2)}{2(2n+1)}$. Then

$$\mathbb{P}_u(C_n = k-1) < \mathbb{P}_u(C_n = k) \text{ for } 1 < k < \Delta(n), \text{ and } \mathbb{P}_u(C_n = k) > \mathbb{P}_u(C_n = k+1) \text{ for } \Delta(n) \leq k < n/2.$$

Proof Since $\mathbb{P}_u(C_n = k) = 0$ for $k > n/2$, the theorem clearly holds for $k \geq n/2 - 1$. Hence in the remainder of the proof we may assume $k < (n-2)/2$. Now by Theorem 6 we have

$$\frac{\mathbb{P}_u(C_n = k-1)}{\mathbb{P}_u(C_n = k)} = \frac{4k(k-1)}{(n-2k+1)(n-2k+2)} := g(k,n). \tag{23}$$

Considering the function $g(k,n)$ defined in Eq. (23), then $g(k+1,n) > g(k,n)$ holds for $1 < k < (n-2)/2$. This, together with Eq. (23), completes the proof of Eq. (22).

The second part of the theorem follows from the observation that $g(k,n) > 1$ if and only if $k \geq \Delta(n)$. □

Now we apply Theorem 5 to study pitchfork distribution, and the joint distribution between pitchforks and cherries under the PDA model. Note that the mean and the variance of pitchfork distributions under this model were also derived by Chang and Fuchs (2010, Proposition 5). Since the proof is similar to that in Corollary 4, we only outline the main steps used here.

Proposition 5 *For $n \geq 3$ we have*

$$\mathbb{E}_u(A_n) = \frac{n(n-1)(n-2)}{2(2n-3)(2n-5)} \sim \frac{n}{8}, \tag{24}$$

$$Cov_u(A_n, C_n) = \frac{-n(n-1)(n-2)(n-3)}{2(2n-3)^2(2n-5)(2n-7)} \sim -\frac{n}{32}, \tag{25}$$

$$\mathbb{V}_u(A_n) = \frac{3n(n-1)(n-2)(n-3)(4n^3-40n^2+123n-110)}{4(2n-3)^2(2n-5)^2(2n-7)(2n-9)} \sim \frac{3n}{64}. \tag{26}$$

Proof Applying Theorem 5 to the function $\varphi(x,y) = x$ and using Corollary 4, we have

$$\mathbb{E}_u(A_{n+1}) = \frac{3n(n-1)}{2(2n-1)(2n-3)} + \frac{2n-5}{2n-1}\mathbb{E}_u(A_n)$$

for $n > 2$. Now Eq. (24) follows by solving the last recursion with an approach similar to that in Corollary 4.

To this end, applying Theorem 5 to the function $\varphi(x,y) = xy$ implies that

$$\mathbb{E}_u(A_{n+1}C_{n+1}) = \frac{2n-7}{2n-1}\mathbb{E}_u(A_nC_n) + \frac{n(n-1)(5n^2-9n-8)}{4(2n-1)(2n-3)(2n-5)}$$

holds for $n > 2$. Solving this recursion we have

$$\mathbb{E}_u(A_nC_n) = \frac{n(n-1)(n-2)(n^2-3n-2)}{4(2n-3)(2n-5)(2n-7)}, \tag{27}$$

from which Eq. (25) follows.

Finally, applying Theorem 5 to the function $\varphi(x,y) = x^2$ shows that

$$\mathbb{E}_u(A_{n+1}^2) = \frac{2n-9}{2n-1}\mathbb{E}_u(A_n^2) + \frac{g(n+1)}{4(2n-1)(2n-3)(2n-5)(2n-7)}$$

holds for $n > 2$. Solving the above recursion leads to

$$\mathbb{E}_u A_n^2 = \frac{n(n-1)(n-2)(n^3-4n^2-17n+66)}{4(2n-3)(2n-5)(2n-7)(2n-9)},$$

from which Eq. (26) follows. □

We end this section with the following correlation result for the PDA model.

Corollary 5 *For $n \geq 4$ we have*

$$\rho_u(A_n, C_n) = -\sqrt{\frac{2(2n-5)(2n-9)}{3(2n-7)(4n^3-40n^2+123n-110)}} \sim -\frac{1}{\sqrt{3}\,n}. \tag{28}$$

In addition, $\{|\rho_u(A_n, C_n)|\}_{n\geq 4}$ is a decreasing sequence converging to 0.

Proof Note first that Eq. (28) follows from Corollary 4 and Proposition 5. Since the sequence $\{|\rho_u(A_n, C_n)|\}_{n\geq 4}$ clearly approaches 0, it remains to show that this sequence is decreasing. To this end, it suffices to show that the ratio

$$R(n) = \frac{\rho_u(A_n, C_n)^2}{\rho_u(A_{n+1}, C_{n+1})^2}$$

is greater than 1 for $n \geq 4$. Using Eq. (28), we have

$$R(n) = \frac{(2n-5)^2(2n-9)(4(n+1)^3-40(n+1)^2+123(n+1)-110}{(2n-7)^2(2n-3)(4n^3-40n^2+123n-110)}.$$

By numerical computation, we can check that $R(n) > 1$ for $4 \leq n \leq 15$, therefore we may assume that $n > 15$ in the remainder of the proof. Now denoting the numerator and denominator of $R(n)$ by $R_1(n)$ and $R_2(n)$, respectively, then we have

$$\begin{aligned} R_1(n) - R_2(n) &= 64n^5 - 944n^4 + 5408n^3 - 15048n^2 + 20436n - 10995 \\ &> 64n^4(n-15) + 5408n^3(n-15) + 20436(n-15) > 0 \end{aligned}$$

for $n > 15$. This implies $R(n) = R_1(n)/R_2(n) > 1$ for $n > 15$, as required. □

5 A comparative study of two models

In this section, we compare and contrast the distributional properties of the number of cherries and the pitchforks in random trees generated under the YHK and the PDA models.

To begin with, note that the recursions in Theorems 1 and 4 provide us exact computation of the joint distribution, and hence also the marginal distributions, of A_n and C_n under the two models. For example, the joint distributions with $n = 50$ and $n = 200$ for the two models are depicted in Fig. 2. They suggest that on average, trees of a given size generated by the YHK model contain more cherries and more pitchforks than those by the PDA model. This is confirmed by the following result.

Proposition 6 *For $n > 3$, we have*

$$\mathbb{E}_u(C_n) < \mathbb{E}_y(C_n) < \frac{4}{3}\mathbb{E}_u(C_n) \tag{29}$$

and

$$\mathbb{E}_u(A_n) < \mathbb{E}_y(A_n) < \frac{4}{3}\mathbb{E}_u(A_n). \tag{30}$$

Proof By Corollaries 2 and 4 we have

$$\mathbb{E}_y(C_n) = \left[1 + \frac{n-3}{3(n-1)}\right] \mathbb{E}_u(C_n),$$

from which Eq. (29) follows. Similarly, by Propositions 3 and 5 we have

$$\mathbb{E}_y(A_n) = \left[1 + \frac{n^2-7n+9}{3(n-1)(n-2)}\right] \mathbb{E}_u(A_n),$$

from which Eq. (30) follows. □

Next, we study the variances of cherry and pitchfork distributions under the two models.

Proposition 7 *For $n > 5$, we have*

$$\frac{32}{45}\mathbb{V}_y(C_n) < \mathbb{V}_u(C_n) < \frac{49}{54}\mathbb{V}_y(C_n).$$

Proof Let $R_n = \mathbb{V}_y(C_n)/\mathbb{V}_u(C_n)$. Then by Corollaries 2 and 4 we have

$$R_n = \frac{4(2n-3)^2(2n-5)}{45(n-1)(n-2)(n-3)}.$$

This implies

$$\frac{R_n}{R_{n+1}} = \frac{n(2n-3)(2n-5)}{(n-3)(2n-1)^2} = 1 + \frac{2n+3}{(n-3)(2n-1)^2} > 1,$$

and hence that R_n is decreasing in n. Noting that $\lim_{n\to\infty} R_n = \frac{32}{45}$, we have

$$\frac{32}{45} < R_n < R_5 = \frac{49}{54} < 1,$$

from which the proposition follows. □

Proposition 8 *For $n \geq 7$, we have*

$$1.168\,\mathbb{V}_u(A_n) < \frac{368}{315}\mathbb{V}_u(A_n) < \mathbb{V}_y(A_n) < \frac{8349}{6520}\mathbb{V}_u(A_n) < 1.281\,\mathbb{V}_u(A_n).$$

Proof Let $R_n = \mathbb{V}_y(A_n)/\mathbb{V}_u(A_n)$ for $n \geq 7$. Then by Proposition 3 and 5 we have

$$R_n = \frac{23(2n-3)^2(2n-5)^2(2n-7)(2n-9)}{315(n-1)(n-2)(n-3)(4n^3-40n^2+123n-110)},$$

and hence

$$\begin{aligned}\frac{R_n}{R_{n+1}} &= \frac{n(2n-5)(2n-9)(4n^3-28n^2+55n-23)}{(n-3)(2n-1)^2(4n^3-40n^2+123n-110)} \\ &= 1 + \frac{2(24n^3-180n^2+382n-165)}{(n-3)(2n-1)^2(4n^3-40n^2+123n-110)} > 1.\end{aligned}$$

Therefore, R_n is strictly decreasing in n and we have

$$\frac{\mathbb{V}_y(A_n)}{\mathbb{V}_u(A_n)} = R_n > \lim_{m\to\infty} R_m = \frac{23\cdot 4\cdot 4\cdot 4}{315\cdot 4} = \frac{368}{315} > 1.168\,.$$

This, together with $R_7 = \frac{8349}{6520} < 1.281$, completes the proof. □

Proposition 6 shows that trees generated by the YHK model have smaller variation in the number of cherries than trees of the same size generated by the PDA model. On the contrary, Proposition 8 shows that YHK model generates trees with larger variation in the number of pitchforks than the PDA model does. This is not unexpected as the covariances of cherries and pitchforks are found to be negative by Propositions 3 and 5.

Now we present a result concerning the correlation coefficients between the cherry and the pitchfork distributions under the two models. Since these two distributions are negatively correlated, we will focus on their absolute values.

Proposition 9 *For $n \geq 7$, we have $|\rho_y(A_n, C_n)| \geq |\rho_u(A_n, C_n)|$. Moreover, $\left\{\frac{\rho_u(A_n,C_n)}{\rho_y(A_n,C_n)}\right\}_{n\geq 7}$ is a monotonically decreasing sequence with limit 0.*

Proof This follows from Corollary 3 and 5 and the observation that

$$|\rho_u(A_7, C_7)| = \sqrt{\frac{30}{7 \times 163}} \leq |\rho_y(A_7, C_7)| = \sqrt{\frac{14}{69}}.$$

□

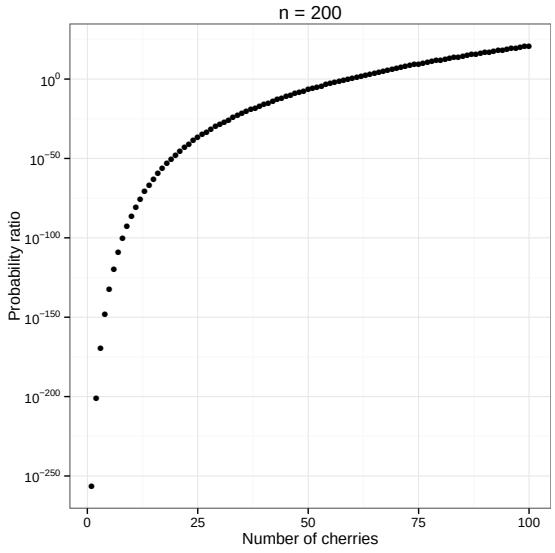

Fig. 3: Plot of the ratio $\mathbb{P}_y(C_n = k)/\mathbb{P}_u(C_n = k)$ for $n = 200$ and $1 \leq k \leq 100$. The probabilities are computed using Eq. (11) and Eq. (18).

Proposition 6 states that the mean of C_n is greater under the YHK model than the PDA model. In the remainder of this section, we shall present a more detailed study on C_n. Intuitively, it is easy to see that the number of cherries contained in a random tree generated by the YHK model is likely to be greater than that by the PDA model: firstly, by Proposition 1 we know that the number of cherries in $T[e]$, the phylogenetic tree obtained from T by attaching a new leaf to edge e in T, is strictly greater than that in T precisely

when e is a pendant edge of T; secondly, in the YHK process the edge to which the new leaf is attached is sampled only from the pendant edges while in the PDA model that edge is sampled from all possible edges. Indeed, this intuition can also be corroborated by numerical results. As an example, considering the ratio of $\mathbb{P}_y(C_n = k)/\mathbb{P}_u(C_n = k)$ with $n = 200$ as depicted in Fig. 3 using a logarithmic scale, then it is clear that the ratio is strictly increasing and is greater than 1 when k is greater than a certain value. The following theorem establishes the existence of a unique change point between the two models for $n \geq 4$. Note that a similar phenomenon is shown to hold for clade sizes by Zhu et al (2015, Theorem 5).

Theorem 8 *Suppose $n \geq 3$. The ratio $\mathbb{P}_y(C_n = k)/\mathbb{P}_u(C_n = k)$ is strictly increasing for $1 \leq k \leq n/2$. In particular, there exists a number τ_n with $1 \leq \tau_n \leq n/2$ such that*

$$\mathbb{P}_y(C_n = k) < \mathbb{P}_u(C_n = k) \quad \textit{for } 1 \leq k < \tau_n, \textit{ and } \ \mathbb{P}_y(C_n = k) > \mathbb{P}_u(C_n = k) \quad \textit{for } \tau_n < k \leq n/2.$$

Proof For simplicity, put $a_n^k = \mathbb{P}_y(C_n = k)$. By Eq. (23), it suffices to show that

$$f(k,n) =: \frac{a_n^{k-1}}{a_n^k} \leq \frac{\mathbb{P}_u(C_n = k-1)}{\mathbb{P}_u(C_n = k)} = \frac{4k(k-1)}{(n-2k+1)(n-2k+2)} := g(k,n) \tag{31}$$

holds for $1 < k \leq n/2$. To this end, we shall use induction on n. The base case $n = 3$ is clear because $\mathbb{P}_y(C_3 = 1) = \mathbb{P}_u(C_3 = 1) = 1$. For induction step, assuming that $f(k,m) < g(k,m)$ holds for a given $m > 3$ and all $1 \leq k \leq m/2$, it remains to show that $f(k,m+1) \leq g(k,m+1)$ for all $1 < k \leq (m+1)/2$.

Note first that $f(1,m+1) = g(1,m+1) = 0$. Now fix $2 \leq k \leq (m+1)/2$, then $a_m^{k-1} > 0$. Since Proposition 2 implies

$$a_{m+1}^{k-1} = \frac{2k-2}{m} a_m^{k-1} + \frac{m-2k+4}{m} a_m^{k-2} \ \text{ and } \ a_{m+1}^k = \frac{2k}{m} a_m^k + \frac{m-2k+2}{m} a_m^{k-1},$$

it follows that Eq. (31) is equivalent to

$$(2k-2)a_m^{k-1} + (m-2k+4)a_m^{k-2} < \left(2ka_m^k + (m-2k+2)a_m^{k-1}\right) g(k,m+1).$$

Since $a_m^{k-1} > 0$, dividing both sides of the last inequality by a_m^{k-1} leads to

$$(2k-2) + (m-2k+4)f(k-1,m) < 2k\frac{g(k,m+1)}{f(k,m)} + (m-2k+2)g(k,m+1). \tag{32}$$

By induction assumption we have $f(k-1,m) \leq g(k-1,m)$ and $0 < f(k,m) < g(k,m)$, hence it remains to show that

$$(2k-2) + (m-2k+4)g(k-1,m) < 2k\frac{g(k,m+1)}{g(k,m)} + (m-2k+2)g(k,m+1). \tag{33}$$

To this end, denoting the left and the right side of Inequality (33) by $L(k,m)$ and $R(k,m)$, respectively, then we need to show that $R(k,m) - L(k,m) > 0$. Note first that

$$L(k,m) = (2k-2) + \frac{4(k-1)(k-2)}{m-2k+3}.$$

On the other hand, since

$$\frac{g(k,m+1)}{g(k,m)} = \frac{m-2k+1}{m-2k+3} = 1 - \frac{2}{m-2k+3},$$

we have

$$R(k,m) = 2k - \frac{4k}{m-2k+3} + \frac{4k(k-1)}{m-2k+3}.$$

Therefore, we have

$$R(k,m) - L(k,m) = 2 + \frac{4k(k-1) - 4k - 4(k-1)(k-2)}{m - 2k + 3}$$
$$= 2 + \frac{4(k-2)}{m - 2k + 3} > 0,$$

as required. Here the last inequality follows from $m > 3$ and $2 \leq k \leq (m+1)/2$. □

6 Discussion and Conclusion

Tree shape indices are summary statistics of some aspect of the shape of a phylogenetic tree, particularly the 'balance' of a tree. Since the introduction of the first tree shape index by Sackin (1972), many such indices have been proposed (see Mooers and Heard (1997) for an excellent review and Mir et al (2013) for some recent development).

In this paper we present several results concerning the distributions of cherries and pitchforks under the YHK and PDA models. Our main results include two novel recursive formulae on the joint distributions of cherries and pitchforks under these two models, which enable us to numerically compute their joint probability density functions (and hence also the marginal distributions) for trees of any size numerically. This is relevant because one of the main applications of tree indices is their use as test statistics to discriminate stochastic models of evolution. For example, statistics based on the number of cherries and on that of pitchforks are utilised by Blum and François (2005) to test the goodness-of-fit of the YHK model to an HIV-1 dataset. However, cherries and pitchforks are used alone in those statistics while the contour plots in Figure 1 suggest that the joint distributions of the cherries and pitchforks might be better than the marginal distributions to discriminate the two models. Therefore, developing powerful statistical tests based on the joint distributions and a thorough data analysis of phylogenetic trees will be explored elsewhere.

Our numerical results (e.g. Fig. 4) indicate that the limiting joint distributions of cherries and pitchforks can be well approximated by bivariate normal distributions. For the YHK model, this was recently confirmed by Holmgren and Janson (2015) and it remains open to establish the analogous result for the PDA model. In addition, several asymptotic results on cherries and pitchforks have been established by Plazzotta and Colijn (preprint) for the general Crump-Mode-Jagers branching process, and it would be interesting to study the joint distribution of cherries and pitchforks under this process.

In this paper we concentrate on rooted trees, but it is of interest to investigate to what extent the results obtained in this paper for rooted trees can be carried over to unrooted trees. For example, using Eq. (21) and an argument similar to the proof in Theorem 7, it follows that the cherry distribution of unrooted trees under the PDA model is also log-concave, and hence unimodal. However, whether the same property holds for the cherry distribution of unrooted trees under the YHK model remains open. One challenge is to derive the recursions for unrooted trees as in Theorem 1 or an exact formula as in Theorem 6.

To end this article, we mention several additional questions. For instance, cherries and pitchforks are special instances of a rooted caterpillar (i.e., a rooted tree containing precisely one cherry), and hence it is of interest to see how the recursion formulae in Theorems 1 and 4 might be extended to rooted caterpillars in general. Next, is pitchfork distribution, or other subtree distributions, log-concave? Our numerical calculation suggests

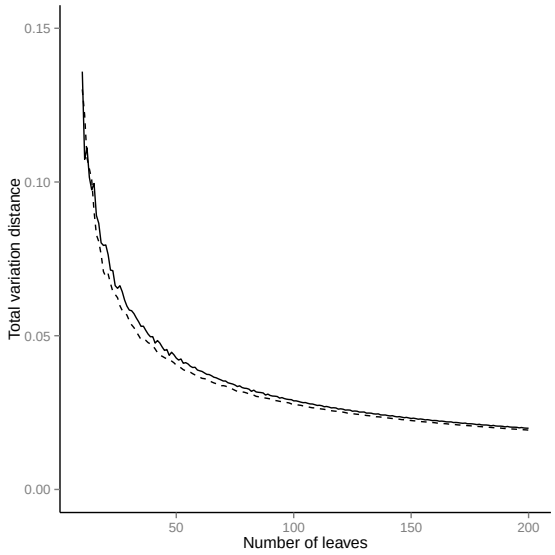

Fig. 4: Plot of the total variation distance between the joint distributions of cherry and pitchfork and discretised bivariate normal distributions for $10 \leq n \leq 200$ under the YHK model (solid line) and PDA model (dotted line). The mean vectors and covariance matrices of the normal distributions are derived from Corollary 2 and Proposition 3 for the YHK model, and Corollary 4 and Proposition 5 for the PDA model. The bivariate normal distributions are discretised by assigning to each point (x, y) of the two-dimensional integer lattice with the probability that a point randomly generated according to the given normal distribution is contained in the unit square centred at (x, y).

the pitchfork distribution is log-concave. A related question is whether there also exists a unique change point for other subtree distributions. Finally, cherry pattern and pitchfork pattern are closely related to instances of recursive shape index (in the sense of Matsen (2007)), therefore it would also be of interest to see whether some of the properties obtained here can be carried over to some other tree indices as well.

Acknowledgements K.P. Choi acknowledges the support of Singapore Ministry of Education Academic Research Fund R-155-000-147-112. We thank the Institute for Mathematical Sciences (Singapore) where part of the work presented here was done during its 'Networks in Biological Sciences' program in July 2015. We would also like to thank Prof. Noah Rosenberg and two anonymous referees for their constructive suggestions on a previous version of this article.

References

Aldous D (1991) Asymptotic fringe distributions for general families of random trees. The Annals of Applied Probability 1:228–266

Aldous D (1996) Probability distributions on cladograms. In: Aldous D, Pemantle R (eds) Random Discrete Structures, The IMA Volumes in Mathematics and its Applications, vol 76, Springer-Verlag, pp 1–18

Aldous D (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statistical Science 16(1):23–34

Baroni M, Grünewald S, Moulton V, Semple C (2005) Bounding the number of hybridisation events for a consistent evolutionary history. Journal of Mathematical Biology 51(2):171–182

Blum MGB, François O (2005) On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. Mathematical Biosciences 195:141–153

Blum MGB, François O (2006) Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Systematic Biology 55(4):685–691

Chang H, Fuchs M (2010) Limit theorems for patterns in phylogenetic trees. Journal of Mathematical Biology 60(4):481–512

Conway JH, Guy R (1996) The Book of Numbers. Springer

Colijn C, Gardy J (2014) Phylogenetic tree shapes resolve disease transmission patterns. Evolution, Medicine, and Public Health 1:96-108.

Disanto F, Wiehe T (2013) Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. Mathematical Biosciences 242(2):195–200.

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Advances in Applied Probability 3(1):44–77

Heard SB (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46(6):1818–1826

Hendy M, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. Mathematical Biosciences 59(2):277–290

Holmgren C, Janson S (2015) Limit laws for functions of fringe trees for binary search trees and recursive trees. Electronic Journal of Probability 20:1–51

Kingman J (1982) On the genealogy of large populations. Journal of Applied Probability 19:27–43

Matsen F (2007) Optimization over a class of tree shape statistics. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 4(3):506–512

McKenzie A, Steel MA (2000) Distributions of cherries for two models of trees. Mathematical Biosciences 164:81–92

Mir A, Rosselló F, Rotger L (2013) A new balance index for phylogenetic trees. Mathematical Biosciences 241(1):125–136

Mooers AO, Heard SB (1997) Evolutionary process from phylogenetic tree shape. Quarterly Review of Biology 72:31–54

Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of Statistical Genetics, Wiley, Chichester, UK, chap 7, pp 179–212

Plazzotta G, Colijn C (2015) Asymptotic frequency of shapes in supercritical branching trees. Preprint, arXiv:1507.02699.

Purvis A, Fritz SA, Rodríguez J, Harvey PH, Grenyer R (2011) The shape of mammalian phylogeny: patterns, processes and scales. Philosophical Transactions of the Royal Society B: Biological Sciences 366(1577):2462–

2477

Rosenberg NA (2006) The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Annals of Combinatorics 10:129–146

Sackin MJ (1972) "Good" and "bad" phenograms. Systematic Zoology 21(2):225–226

Semple C, Steel MA (2003) Phylogenetics. Oxford University Press, Oxford, UK

Stadler T (2013) Recovering speciation and extinction dynamics based on phylogenies Journal of Evolutionary Biology 26(6):1203-1219.

Yule GU (1925) A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis, F.R.S. In: Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, vol 213, The Royal Society, pp 21–87

Zhu S, Degnan JH, Steel MA (2011) Clades, clans and reciprocal monophyly under neutral evolutionary models. Theoretical Population Biology 79:220–227

Zhu S, Than C, Wu T (2015) Clades and clans: A comparison study of two evolutionary models. Journal of Mathematical Biology 71:99–124