Theoretical Population Biology 149 (2023) 27–38

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Distributions of cherries and pitchforks for the Ford model

Gursharn Kaur [a], Kwok Pui Choi [b], Taoyang Wu [c,*]

[a] *Biocomplexity Institute, University of Virginia, Charlottesville, 22911, USA*
[b] *Department of Statistics and Data Science, and the Department of Mathematics, National University of Singapore, 117546, Singapore*
[c] *School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK*

ARTICLE INFO

Article history:
Received 2 May 2022
Available online 22 December 2022

Keywords:
Tree shape
Ford's alpha model
Polya urn model
Random trees
Subtree statistics

ABSTRACT

Distributional properties of tree shape statistics under random phylogenetic tree models play an important role in investigating the evolutionary forces underlying the observed phylogenies. In this paper, we study two subtree counting statistics, the number of cherries and that of pitchforks for the Ford model, the alpha model introduced by Daniel Ford. It is a one-parameter family of random phylogenetic tree models which includes the proportional to distinguishable arrangement (PDA) and the Yule models, two tree models commonly used in phylogenetics. Based on a non-uniform version of the extended Pólya urn models in which negative entries are permitted for their replacement matrices, we obtain the strong law of large numbers and the central limit theorem for the joint distribution of these two statistics for the Ford model. Furthermore, we derive a recursive formula for computing the exact joint distribution of these two statistics. This leads to exact formulas for their means and higher order asymptotic expansions of their second moments, which allows us to identify a critical parameter value for the correlation between these two statistics. That is, when the number of tree leaves is sufficiently large, they are negatively correlated for $0 \leq \alpha \leq 1/2$ and positively correlated for $1/2 < \alpha < 1$.

1. Introduction

In many branches of biology, it is important to elucidate the evolutionary events and forces leading to the current biological systems, such as a group of species or strains of a virus. To this end, the evolutionary relationships among the biological system under investigation are typically represented by a phylogenetic tree, that is, a binary tree whose leaves are labelled by the taxon units in the system. As these events and forces, such as the rates of speciation and expansion, are often not directly observable (Mooers et al., 2007; Heath et al., 2008), one popular approach is to compare the empirical shape indices computed from the trees inferred from real datasets with those predicted by a null tree growth model (Blum and François, 2006; Hagen et al., 2015). Furthermore, some tree shape indices are also closely related to several fundamental statistics in population genetics (Ferretti et al., 2017; Arbisser et al., 2018), and to certain important parameters in the dynamics of virus evolution and propagation (Colijn and Gardy, 2014; Colijn and Plazzotta, 2017).

In phylogenetic analysis, two commonly used random tree growth models are the Yule model and the proportional to distinguishable arrangements (PDA) model (Aldous, 1996). The PDA model is a uniform model on phylogenies in which each phylogenetic tree with the same set of leaf labels has the same probability to be sampled. On the other hand, the Yule model, also known as the Yule–Harding–Kingman (YHK) model, can be realized by a uniform distribution on ranked phylogenies and is closely related to a number of important random processes in theoretical biology, such as the Yule process of speciation and the coalescent process (see, e.g., Steel, 2016). However, for phylogenetic trees inferred from real datasets, it is often observed that the Yule or the PDA model may not always provide a good fit (Aldous, 1996; Blum and François, 2006). Partially motivated by this, several general classes of random trees have been proposed for modelling and analysing the observed data, two well-known ones being the beta model by Aldous (1996) and the alpha model by Ford (2006).

The alpha model, which will be referred to as the Ford model in this paper, is a family of random binary tree growth processes with some desirable properties. First, indexed by a parameter α ranging from 0 to 1, the Ford model interpolates continuously between the Yule model ($\alpha = 0$), the PDA model ($\alpha = 1/2$), and the Comb model ($\alpha = 1$) which generates only most unbalanced tress (i.e., those with precisely one cherry). Second, this model is sampling consistent (Ford, 2006), i.e., the probability that a given tree with n leaves is sampled under the Ford model with a given parameter α is the same as randomly deleting a leaf from a random tree with $n+1$ leaves sampled by the Ford model with the same parameter α. Furthermore, the Ford model has been further

* Corresponding author.
E-mail address: taoyang.wu@uea.ac.uk (T. Wu).

https://doi.org/10.1016/j.tpb.2022.12.002
0040-5809/

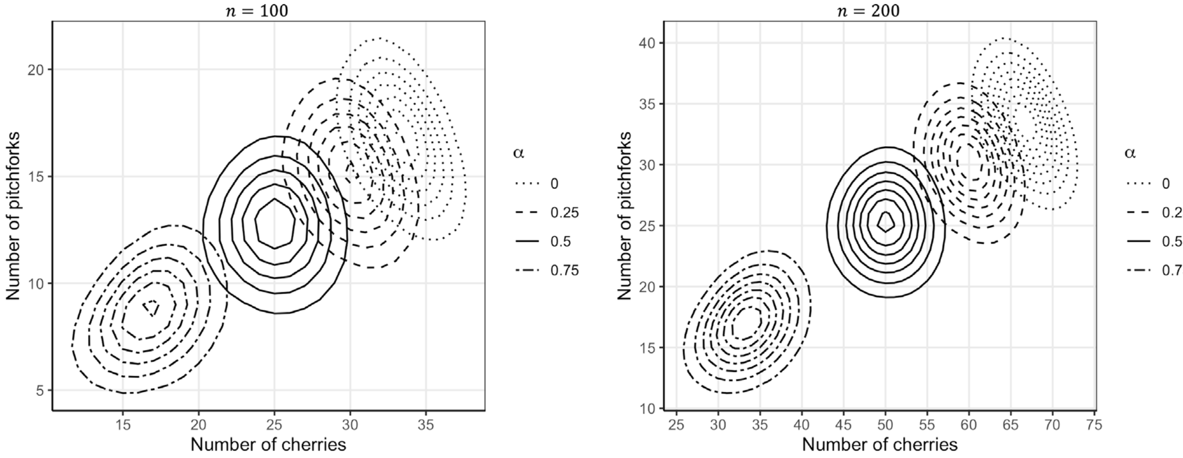

Fig. 1. Contour plots of the probability density functions for the joint distribution of the numbers of cherries and pitchforks under the Ford model with 100 leaves (left) and 200 leaves (right). The polygonal contours arise because the joint distribution is defined only on integer lattice points.

extended by Chen et al. (2009) to a two-parameter family of tree growth process called the alpha–gamma model for sampling trees that are not necessary binary. As such, the Ford model and its variants have been increasingly appreciated and studied in the past decade, including its ability to generate trees for applications on real datasets (e.g., Pompei et al., 2012; Coronado et al., 2018, 2019; Wirtz and Wiehe, 2019).

One commonly used family of tree shape statistics or indices are the number of subtrees. More precisely, in this paper we are interested in the number of cherries and the number of pitchforks. A cherry is a fringe subtree (i.e., a subtree consisting of an edge (u, v) and all the descendants of v) with precisely two leaves. A pitchfork, which was introduced in Rosenberg (2006), is a fringe subtree with three leaves. The study of the number of fringe subtrees of a random tree can be traced back to a paper by Aldous (1991) and has since been extended to various random tree models (see, e.g., Holmgren and Janson, 2017). In phylogenetics, asymptotic properties of the number of cherries were first studied by McKenzie and Steel (2000), who showed that the number of cherries is asymptotically normal for the Yule and the PDA models as the number of leaves tends to infinity. Later, similar properties of the number of cherries are extended to the alpha model by Ford (2006, Theorem 57) and to the Crump–Mode–Jagers branching process by Plazzotta and Colijn (2016). For the number of pitchforks, Rosenberg (2006) obtained its mean and variance, and Chang and Fuchs (2010) proved that the number of pitchforks is also asymptotically normal for the Yule and the PDA models. For the joint distributions, that is, the likelihood that a random tree has a given number of cherries and a given number of pitchforks, Holmgren and Janson (2015) showed that the joint distribution is asymptotically normal for the Yule model, using a correspondence between the Yule model and a classical tree model in computer science known as random binary search trees. This was recently extended by Choi et al. (2021) to the PDA model based on a version of the extended urn models in which negative entries are permitted for their replacement matrices.

In this paper, we establish the strong law of large numbers and the central limit theorems for the joint distribution of cherries and pitchforks under the Ford model (Theorem 3.2) by considering an associated non-uniform urn model (Theorem 3.1). These results are presented in Section 3, following Section 2 in which we collect background information concerning the Ford model and limiting theorems on uniform urn models. Furthermore, we derive a recurrence formula for computing the exact joint distribution under the Ford model (Theorem 4.1) in Section 4, generalizing the results by Wu and Choi (2016) for the Yule and the PDA models. This leads to an efficient way to compute the joint distributions under the Ford model, see Fig. 1 for an illustration. Furthermore, it enables us to obtain recurrence expressions for the moments of these two statistics. As an application, we obtain the exact formula for the mean of the number of cherries (first reported by Ford, 2006) and that of pitchforks under the Ford model in Theorem 4.5. We also obtain higher order expansions of the second moments of their marginal and joint distributions (Theorem 4.6), which allows us to identify a critical parameter value for the correlation between these two statistics: when n is sufficiently large, these two statistics are negatively correlated for $0 \leq \alpha \leq 1/2$ and positively correlated for $1/2 < \alpha < 1$. The proofs of Theorems 4.5 and 4.6 are presented in the appendix. We end the paper with a discussion of open problems in Section 5.

2. Ford model and urn model

In this section, we first introduce the Ford model, which is a one-parameter family of random phylogenetic tree models. Next we present a non-uniform version of the extended urn models associated with the Ford tree model. Finally, we recall certain conditions on the related uniform version of the extended urn model under which the strong law of large numbers and the central limit theorem are obtained.

2.1. Ford model

A rooted binary tree is a finite connected simple graph without cycles that contains a unique vertex of degree 1 designated as the root and all the remaining vertices are of degrees 3 (interior vertices) or 1 (leaves). A phylogenetic tree T with n leaves is a rooted binary tree whose leaves are bijectively labelled with the elements in $\{1, \ldots, n\}$. Note that all the edges in T are directed away from the root, and edges incident with leaves are referred to as pendant edges. Furthermore, for technical simplicity we assume that the root has one child, which is also referred to as planted phylogenetic trees in the literature (e.g. Wu and Choi, 2016). A fringe subtree in T consists of an edge (u, v) and all the edges that are included in the paths from v to all its descendants. A cherry (resp. pitchfork) is a fringe subtree with two (resp. three) leaves. A cherry that is not contained in a pitchfork will be referred to as an essential cherry. Finally, we let $A(T)$ and $C(T)$ denote the number of pitchforks and the number of cherries in the tree T, respectively.

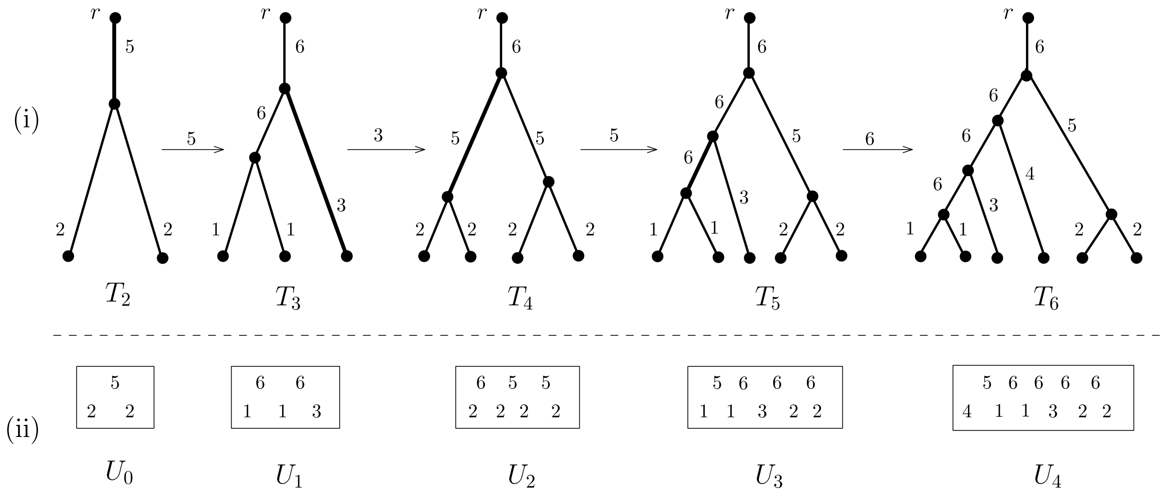

Fig. 2. A sample path of the Ford model and the associated trajectory under the urn model. (i) A sample path of the Ford model evolving from T_2 with two leaves to T_6 with six leaves. All edges are directed away from the root r towards the leaves. The labels of the leaves are omitted for simplicity. The type of an edge is indicated by the number next to it. For $2 \leq i \leq 5$, the edge e_i selected in T_i to generate $T_{i+1} = T_i[e_i]$ is highlighted in bold and the associated edge type of e_i is indicated in the number above the arrow. (ii) The associated urn model with six colours, derived from the types of edges in the trees. In vector form, we have $U_0 = (0, 2, 0, 0, 1, 0)$, $U_1 = (2, 0, 1, 0, 0, 2)$, $U_2 = (0, 4, 0, 0, 2, 1)$, $U_3 = (2, 2, 1, 0, 1, 3)$, and $U_4 = (2, 2, 1, 1, 1, 4)$.

Under the Ford model with parameter $0 \leq \alpha \leq 1$, a random phylogenetic tree T_n with n leaves is constructed recursively by adding one leaf at a time as follows (see Fig. 2 for an illustration.). Fix a random permutation $(x_1, \dots, x_n)$ of $\{1, \dots, n\}$. The initial tree T_2 contains precisely two leaves (e.g. one cherry) which are labelled as x_1 and x_2. For the recursive step, given a tree T_m with m leaves constructed so far, choose a random edge e in T_m according to weight $1 - \alpha$ for each pendant edge and weight α for each of the other edges. That is, an edge e of T_m is sampled with probability $\alpha/(m - \alpha)$ if e is an interior edge, and with probability $(1 - \alpha)/(m - \alpha)$ if e is a pendant edge. The tree $T_{m+1} := T_m[e]$ is obtained by subdividing the selected edge e with a new node v_e and attaching a new leaf labelled x_{m+1} to v_e. That is, every single addition of a leaf to the tree results in an increase of the number of edges by two. Finally, we let $A_n = A(T_n)$ and $C_n = C(T_n)$ denote the numbers of pitchforks and cherries in the tree T_n, respectively.

2.2. An urn model associated with trees

Consider an urn containing balls of d different colours where the colours are denoted by integers $\{1, 2, \dots, d\}$. Let $U_n = (U_{n,1}, \dots, U_{n,d})$ be the configuration vector of length d such that the ith element of U_n is the number of balls of colour i at time n. Let U_0 be the initial vector of colour configuration. At time $n \geq 1$, a ball is selected uniformly at random from the urn, and if the colour of the selected ball is i then the ball is replaced along with $R_{i,j}$ many balls of colour j, for every $1 \leq j \leq d$. The dynamics of the urn configuration depends on its initial configuration U_0 and the $d \times d$ replacement matrix $R = (R_{i,j})_{1 \leq i,j \leq d}$.

We study the limiting properties of the numbers of cherries and pitchforks via an equivalent urn process. Towards this, we use six different colours and assign one colour to each type of edges of a tree T in the following scheme introduced by Choi et al. (2021): colour 1 for all pendant edges of a cherry in a pitchfork; colour 2 for pendant edges of an essential cherry (i.e., cherry not contained in any pitchforks); colour 3 for pendant edges in a pitchfork but not in any cherry; colour 4 for pendant edges in neither a cherry nor a pitchfork; colour 5 for the internal edge of an essential cherry (i.e., those adjacent to colour 2 edges), and colour 6 for all other (necessarily internal) edges. See Fig. 2 for an illustration of the scheme. For $1 \leq i \leq 6$, let $E_i(T)$ be the set of edges of colour i in T.

Let $U_n = (U_{n,1}, \dots, U_{n,6})$ denote the colour configuration of the urn at time n, where $U_{n,i}$ denotes the number of edges of colour i in the tree at time n. When generating a tree under the Ford model, it will have precisely $n+2$ leaves. At the initial time step ($n = 0$), the tree T_2 is an essential cherry, which has two pendant edges and one interior edge (see T_2 in Fig. 2), so $U_0 = (0, 2, 0, 0, 1, 0)$. Based on the colouring scheme of the edges, at any time $n \geq 0$, we have

$$(A_{n+2}, C_{n+2}) = \frac{1}{2}\left(U_{n,1}, U_{n,1} + U_{n,2}\right), \tag{1}$$

where A_{n+2} and C_{n+2} are the numbers of pitchforks and cherries in T_{n+2}, respectively. Under the alpha tree model, the dynamics of the corresponding urn process evolve according to the following replacement matrix

$$R = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 2 & -2 & 1 & 0 & -1 & 2 \\ -2 & 4 & -1 & 0 & 2 & -1 \\ 0 & 2 & 0 & -1 & 1 & 0 \\ 2 & -2 & 1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Let $\mathbf{e}_i$, $1 \leq i \leq 6$, denote a 6-vector in which the ith component is 1 and 0 elsewhere; and χ_n, the random vector taking value $\mathbf{e}_i$ if speciation happens at an edge with type i at time n. Thus, we have the following recursion

$$U_n = U_{n-1} + \chi_n R, \qquad n \geq 1,$$

where

$$P(\chi_n = \mathbf{e}_i | \mathcal{F}_{n-1}) \propto \begin{cases} (1-\alpha) U_{n-1,i}, & \text{for } i \in \{1, 2, 3, 4\}, \\ \alpha\, U_{n-1,i}, & \text{for } i \in \{5, 6\}. \end{cases} \tag{2}$$

Observe that the process $(U_n)_{n \geq 0}$, which describes the dynamics of the numbers of cherries and pitchforks, is a *non-uniform urn model* since the balls are not selected uniformly at random from the urn, which is different from the classical *uniform* urn models in which the balls are selected uniformly at random from the urn (see, e.g., Hofri and Mahmoud, 2019, Chapter 7).

We end this subsection with the following observation regarding the edge colour scheme with the number of pitchforks and that of cherries, which follows directly from the replacement matrix R (see also Wu and Choi, 2016, Section 2).

Lemma 2.1. *Suppose that T is a phylogenetic tree with $n \geq 2$ leaves. Then, $|E_3(T)| = A(T)$, $|E_2(T) \cup E_5(T)| = 3(C(T) - A(T))$, $|E_4(T)| = n - A(T) - 2C(T)$, and $|E_1(T) \cup E_6(T)| = n - 1 + 3A(T) - C(T)$. Furthermore, suppose that e is an edge in T and $T' = T[e]$. Then,*

$$A(T') = \begin{cases} A(T) - 1, & \text{if } e \in E_3(T), \\ A(T) + 1, & \text{if } e \in E_2(T) \cup E_5(T), \text{ and} \\ A(T), & \text{otherwise}; \end{cases}$$

$$C(T') = \begin{cases} C(T) + 1, & \text{if } e \in E_3(T) \cup E_4(T), \\ \\ C(T), & \text{otherwise}. \end{cases}$$

2.3. Limiting theorems on uniform urn models

In this subsection, we review the strong law of large numbers and the central limit theorem on a version of uniform urn models developed by Choi et al. (2021), which will be applied to the non-uniform urn process in Section 2.2 using the urn coupling idea in Bandyopadhyay and Kaur (2018).

For the classical uniform urn models, Bai and Hu (2005) showed that the random process U_n/n, when properly adjusted by a scalar factor, converges almost surely to the left eigenvector of R corresponding to the maximal eigenvalue, and is asymptotic normal with a known limiting variance matrix under certain assumptions on R. Standard assumptions made in the urn model theory are that the replacement matrix is irreducible with a constant row sum and all the off-diagonal elements are non-negative (see, e.g., Mahmoud, 2009). Choi et al. (2021) extended this to the case when off-diagonal elements of a replacement matrix can be negative satisfying the following set of assumptions **(A1)–(A4)**. Let $\mathrm{diag}(a_1, \ldots, a_d)$ denote the diagonal matrix whose diagonal elements are $a_1, \ldots, a_d$.

(A1) *Tenable:* It is always possible to draw balls and follow the replacement rule.

(A2) *Small:* All eigenvalues of R are real. The maximal eigenvalue λ_1, called the *principal eigenvalue*, is positive with $\lambda_1 > 2\lambda$ for all other eigenvalues λ of R.

(A3) *Strictly balanced:* The column vector $\mathbf{u}_1 = (1, 1, \ldots, 1)^\top$ is a right eigenvector of R corresponding to λ_1, and R has a unique *principal left eigenvector* $\mathbf{v}_1$ that is both a left eigenvector corresponding to λ_1 and a probability vector.

(A4) *Diagonalizable:* There exists an invertible matrix V with real entries such that its first row equals to $\mathbf{v}_1$, the first column of V^{-1} is $\mathbf{u}_1$, and

$$VRV^{-1} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) =: \Lambda, \tag{3}$$

where $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$ are eigenvalues of R.

Let $\mathcal{N}(\mathbf{0}, \Sigma)$ be the multivariate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and covariance matrix Σ. Then, we have the following result from Choi et al. (2021, Theorems 1 & 2), which also follows from Janson (2004, Theorems 3.21 and 3.22 and Remark 4.2).

Theorem 2.2. *Under assumptions* **(A1)–(A4)**, *we have*

$$(n\lambda_1)^{-1}U_n \xrightarrow{a.s.} \mathbf{v}_1 \quad \text{and} \quad n^{-1/2}(U_n - n\lambda_1\mathbf{v}_1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{4}$$

where λ_1 is the principal eigenvalue, $\mathbf{v}_1$ is the principal left eigenvector of R, and

$$\Sigma = \sum_{i,j=2}^{d} \frac{\lambda_1\lambda_i\lambda_j\mathbf{u}_i^\top \mathit{diag}(\mathbf{v}_1)\mathbf{u}_j}{\lambda_1 - \lambda_i - \lambda_j}\mathbf{v}_i^\top\mathbf{v}_j, \tag{5}$$

where $\mathbf{v}_j$ is the jth row of V, and $\mathbf{u}_j$ the jth column of V^{-1} for $2 \leq j \leq d$.

3. Limit theorems for the joint distribution

In this section, we present the strong law of large numbers and the central limit theorems for the joint distribution of the number of cherries and the number of pitchforks under the Ford model.

3.1. Main convergence results

We introduce the following six polynomials in α for later use:

$$\begin{aligned} \phi_1 &= 8\alpha^3 - 32\alpha^2 + 45\alpha - 23, & \phi_4 &= 8\alpha^3 - 40\alpha^2 + 37\alpha + 13, \\ \phi_2 &= 40\alpha^3 - 164\alpha^2 + 221\alpha - 97, & \phi_5 &= 40\alpha^3 - 112\alpha^2 - 31\alpha + 181, \\ \phi_3 &= 56\alpha^3 - 248\alpha^2 + 367\alpha - 181, & \phi_6 &= 8\alpha^3 + 4\alpha^2 - 71\alpha + 71. \end{aligned} \tag{6}$$

For simplicity of notation, we do not indicate ϕ_i as a function of α. It can be verified directly that $\phi_1, \phi_2, \phi_3 < 0$ and $\phi_4, \phi_5, \phi_6 > 0$ for $\alpha \in (0, 1)$. Then, we have the following asymptotic results of the urn model process associated with the Ford model.

Theorem 3.1. *Suppose $(U_n)_{n \geq 0}$ is the urn process associated with the Ford model with parameter $\alpha \in (0, 1)$. Then,*

$$\frac{U_n}{n} \xrightarrow{a.s.} \mathbf{v} \quad \text{and} \quad \frac{U_n - n\mathbf{v}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{7}$$

as $n \to \infty$, where

$$\mathbf{v} = \frac{1}{2(3 - 2\alpha)}(2(1 - \alpha),\ 2(1 - \alpha),\ 1 - \alpha,\ 1 + \alpha,\ 1 - \alpha,\ 5 - 3\alpha), \tag{8}$$

and with the polynomials $\phi_1, \ldots, \phi_6$ defined in (6),

$$\Sigma = \frac{1 - \alpha}{4(3 - 2\alpha)^2(5 - 4\alpha)(7 - 4\alpha)} \times \begin{bmatrix} -12\phi_1 & 4\phi_2 & -6\phi_1 & -2\phi_4 & 2\phi_2 & -2\phi_2 \\ 4\phi_2 & -4\phi_3 & 2\phi_2 & -2\phi_6 & -2\phi_3 & 2\phi_3 \\ -6\phi_1 & 2\phi_2 & -3\phi_1 & -\phi_4 & \phi_2 & -\phi_2 \\ -2\phi_4 & -2\phi_6 & -\phi_4 & \phi_5 & -\phi_6 & \phi_6 \\ 2\phi_2 & -2\phi_3 & \phi_2 & -\phi_6 & -\phi_3 & \phi_3 \\ -2\phi_2 & 2\phi_3 & -\phi_2 & \phi_6 & \phi_3 & -\phi_3 \end{bmatrix}. \tag{9}$$

The proof of Theorem 3.1 is given at the end of this section.

Remark 1. Theorem 3.1 provides the limiting results on the urn model using a scaling factor relating to the time n (which is motivated by noting that the number of leaves in the tree at time n is $n + 2$). However, the results can be readily rephrased using the vector $U_n/(3 + 2n)$ of proportions of colour balls in the urn process, in which $\sum_{i=1}^{6} U_{n,i} = 3 + 2n$ since two balls are added into the urn at every time point.

Remark 2. Using the approach outlined by Choi et al. (2021), Theorem 3.1 continues to hold for the unrooted Ford model.

With Theorem 3.1, we are ready to present one of our main results in this paper concerning limit theorems on the joint distribution of the number of cherries C_n and the number of pitchforks A_n under the Ford model.

Theorem 3.2. *Under the Ford model with parameter $\alpha \in [0, 1]$, we have*

$$\frac{1}{n}(A_n, C_n) \xrightarrow{a.s.} (\nu, \mu) := \frac{1 - \alpha}{2(3 - 2\alpha)}(1, 2), \tag{10}$$

and

$$\frac{(A_n, C_n) - n(\nu, \mu)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\big((0, 0), S\big),$$

where

$$S=\begin{bmatrix}\tau^2 & \rho\\ \rho & \sigma^2\end{bmatrix}=\frac{1-\alpha}{(3-2\alpha)^2(5-4\alpha)} \times\begin{bmatrix}\frac{-24\alpha^3+96\alpha^2-135\alpha+69}{4(7-4\alpha)} & \frac{-(2-\alpha)(1-2\alpha)}{2}\\ \frac{-(2-\alpha)(1-2\alpha)}{2} & 2-\alpha\end{bmatrix}. \tag{11}$$

Remark 3. We consider special cases of the Ford model, which are commonly studied in phylogenetics. The first two have been established by Choi et al. (2021).

1. The PDA model corresponds to $\alpha = 1/2$, where all edges, internal or leaf, are selected with equal weight and the limit results hold with
$$(\nu,\mu)=\frac{1}{8}(1,2)\quad\text{and}\quad\begin{bmatrix}\tau^2 & \rho\\ \rho & \sigma^2\end{bmatrix}=\frac{1}{64}\begin{bmatrix}3 & 0\\ 0 & 4\end{bmatrix}.$$
2. The Yule model corresponds to $\alpha = 0$, where only leaf edges are selected with equal weight and the limit results hold with
$$(\nu,\mu)=\frac{1}{6}(1,2)\quad\text{and}\quad\begin{bmatrix}\tau^2 & \rho\\ \rho & \sigma^2\end{bmatrix}=\frac{1}{45}\begin{bmatrix}69/28 & -1\\ -1 & 2\end{bmatrix}.$$
3. The Comb model corresponds to $\alpha = 1$, a degenerate case. It is easy to see that $(\nu,\mu)=(0,0)$ and $\tau^2=\rho=\sigma^2=0$.

Proof of Theorem 3.2. First note that the case $\alpha = 1$ reduces to the degenerate case, Comb model, and therefore we only consider $\alpha \in [0, 1)$. The limiting results for the case $\alpha = 0$ has been obtained by Choi et al. (2021), which agree with the above results when $\alpha = 0$. Thus, it is enough to prove the result for $\alpha \in (0, 1)$.

By (1), we have $(A_{n+2}, C_{n+2}) = U_n Q$ with

$$Q^\top=\frac{1}{2}\begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\\ 1 & 1 & 0 & 0 & 0 & 0\end{bmatrix}. \tag{12}$$

Since

$$\frac{U_n}{n}\xrightarrow{\text{a.s.}}\mathbf{v}=\frac{1}{2(3-2\alpha)}\big(2(1-\alpha),2(1-\alpha), 1-\alpha,1+\alpha,1-\alpha,5-3\alpha\big), \tag{13}$$

using the relation from Eq. (1) we get

$$\frac{1}{n+2}(A_{n+2},C_{n+2})=\frac{n}{n+2}\left(\frac{U_n}{n}\right)Q\xrightarrow{\text{a.s.}}\mathbf{v}Q=\frac{1-\alpha}{2(3-2\alpha)}(1,2).$$

This concludes the proof of the strong law of large numbers.

We now prove the central limit theorem, and obtain the expression for the limiting variance matrix. To this end, denoting the (i, j)-entry in the covariance matrix Σ of (9) by $\sigma_{i,j}$ for $1 \le i, j \le 6$, we consider the matrix

$$\begin{aligned}S=Q^\top\Sigma Q&=\frac{1}{4}\begin{bmatrix}\sigma_{1,1} & \sigma_{1,1}+\sigma_{1,2}\\ \sigma_{1,1}+\sigma_{2,1} & \sigma_{1,1}+\sigma_{2,1}+\sigma_{1,2}+\sigma_{2,2}\end{bmatrix}\\ &=\frac{1-\alpha}{16(3-2\alpha)^2(5-4\alpha)(7-4\alpha)}\\ &\quad\times\begin{bmatrix}-12\phi_1 & -12\phi_1+4\phi_2\\ -12\phi_1+4\phi_2 & -12\phi_1+8\phi_2-4\phi_3\end{bmatrix}\\ &=\frac{1-\alpha}{(3-2\alpha)^2(5-4\alpha)}\begin{bmatrix}\frac{-24\alpha^3+96\alpha^2-135\alpha+69}{4(7-4\alpha)} & \frac{-(2-\alpha)(1-2\alpha)}{2}\\ \frac{-(2-\alpha)(1-2\alpha)}{2} & 2-\alpha\end{bmatrix}\end{aligned}$$

Since $(A_{n+2}, C_{n+2}) = U_n Q$, where Q is as defined in (12), we get

$$\begin{aligned}\frac{(A_{n+2},C_{n+2})-n(\nu,\mu)}{\sqrt{n}}&=\frac{1}{\sqrt{n}}(U_n-n\mathbf{v})\,Q\xrightarrow{d}\mathcal{N}\left(\mathbf{0},Q^\top\Sigma Q\right)\\ &=\mathcal{N}(\mathbf{0},S)\,.\end{aligned}$$

Since $n^{-1/2}\{(A_{n+2}, C_{n+2}) - n(\nu, \mu)\} - n^{-1/2}\{(A_n, C_n) - n(\nu, \mu)\} \xrightarrow{d} 0$, this completes the proof. □

We end this subsection with the following results on the behaviour of the first and second moments of the limiting joint distribution of cherries and pitchforks in the parameter region, as indicated by their plots in Fig. 3.

Corollary 3.3.

(i) *For $0 \le \alpha < 1$, $A_n/C_n \xrightarrow{a.s.} 1/2$ as $n \to \infty$. That is, the number of pitchforks is asymptotically equal to the number of essential cherries.*

(ii) *$A_n/n \xrightarrow{a.s.} (1-\alpha)/(6-4\alpha)$, this limit decreases strictly from 1/6 to 0, as α increases from 0 to 1.*

(iii) *The limiting variance τ^2 of $A_n/\sqrt{n}$ decreases strictly from 23/420 to 0, as α increases from 0 to 1.*

(iv) *The limiting variance σ^2 of $C_n/\sqrt{n}$ increases strictly from 2/45 to 0.0695 over $(0, a_0)$ and decreases from 0.0695 to 0 over $(a_0, 1)$, where $a_0 = 0.7339$, the unique root of $19 - 48\alpha + 36\alpha^2 - 8\alpha^3 = 0$ in $(0, 1)$.*

(v) *The limiting covariance ρ of $A_n/\sqrt{n}$ and $C_n/\sqrt{n}$ changes sign from negative to positive at $\alpha = 1/2$. Specifically, it increases from $-1/45$ to 0.0225 over $(0, a_1)$ and decreases from 0.0225 over $(a_1, 1)$, where $a_1 = 0.8688$, the unique root of $-24\alpha^4 + 160\alpha^3 - 370\alpha^2 + 358\alpha - 123 = 0$ in $(0, 1)$.*

3.2. A uniform urn model derived from U_n

For $\alpha \in (0, 1)$, consider the diagonal 6 × 6 matrix

$$T_\alpha=\mathrm{diag}(1-\alpha,1-\alpha,1-\alpha,1-\alpha,\alpha,\alpha) \tag{14}$$

and

$$\widetilde{U}_n:=U_nT_\alpha=\big((1-\alpha)U_{n,1},\ldots,(1-\alpha)U_{n,4},\alpha U_{n,5},\alpha U_{n,6}\big)\,. \tag{15}$$

Clearly, there is a one to one correspondence between U_n and $\widetilde{U}_n = U_n T_\alpha$ for $\alpha \in (0, 1)$ and therefore it is sufficient to obtain the limiting results for the urn process $\widetilde{U}_n$. Note that the off-diagonal elements of the replacement matrix $R_\alpha = RT_\alpha$ of $\widetilde{U}_n$ are not all non-negative, therefore we will use the limit results from Choi et al. (2021) to obtain the convergence results for the urn process $\widetilde{U}_n$.

Theorem 3.4. *Suppose $\alpha \in (0, 1)$. Then $(\widetilde{U}_n)_{n\ge 0}$ is a uniform urn process with replacement matrix $R_\alpha = RT_\alpha$ and*

$$\frac{\widetilde{U}_n}{n}\xrightarrow{a.s.}\widetilde{\mathbf{v}}_1, \tag{16}$$

where

$$\widetilde{\mathbf{v}}_1=\frac{1}{2(3-2\alpha)}\big(2(1-\alpha)^2,2(1-\alpha)^2, (1-\alpha)^2,1-\alpha^2,\alpha(1-\alpha),\alpha(5-3\alpha)\big) \tag{17}$$

is the normalized left eigenvector of R_α corresponding to the largest eigenvalue $\lambda_1 = 1$. Furthermore,

$$\frac{\widetilde{U}_n-n\widetilde{\mathbf{v}}_1}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(\mathbf{0},\widetilde{\Sigma}), \tag{18}$$

with the polynomials $\phi_1, \ldots, \phi_6$ defined in (6) and $\beta = 1 - \alpha$,

$$\widetilde{\Sigma}=\frac{\beta}{4(3-2\alpha)^2(5-4\alpha)(7-4\alpha)}$$

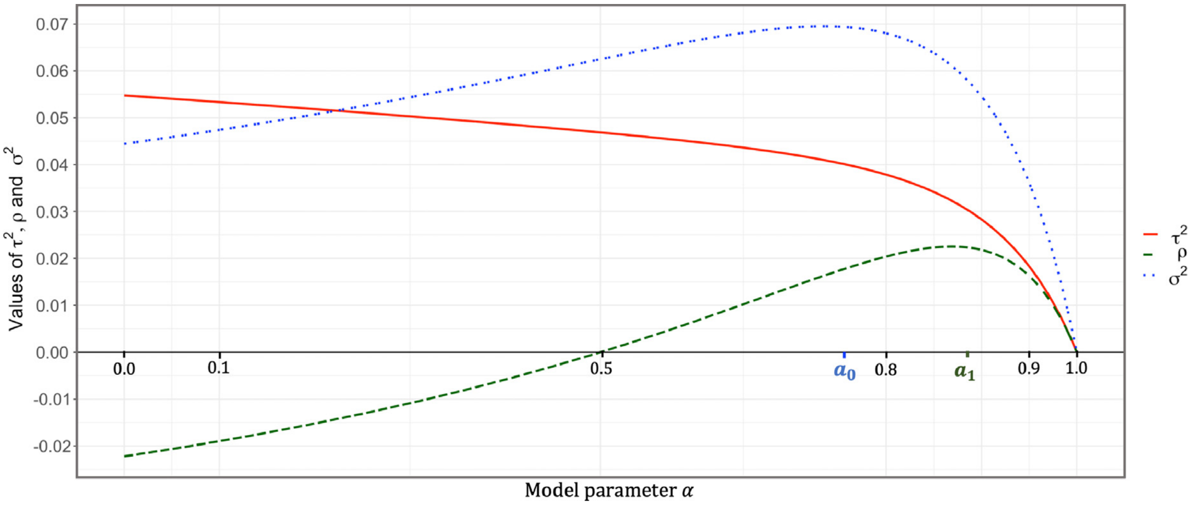

Fig. 3. Plot of the variances and covariance of the limiting normalized joint distribution of cherries and pitchforks with respect to parameter α under the Ford model. (i) The limiting variance τ^2 of the normalized pitchfork counting statistic $A_n/\sqrt{n}$ decreases as the parameter α increases. (ii) The limiting variance σ^2 of the normalized cherry counting statistic $C_n/\sqrt{n}$ increases over $[0, \alpha_0)$ and then decreases over $(\alpha_0, 1]$. (iii) The limiting covariance ρ of $A_n/\sqrt{n}$ and $C_n/\sqrt{n}$ increases over $[0, \alpha_1)$ and decreases over $(\alpha_1, 1]$. The formulas for τ^2, σ^2, and ρ are presented in (11) of Theorem 3.2; and the exact values of α_0 and α_1 are given in Corollary 3.3.

$$\times \begin{bmatrix} -12\beta^2\phi_1 & 4\beta^2\phi_2 & -6\beta^2\phi_1 & -2\beta^2\phi_4 & 2\alpha\beta\phi_2 & -2\alpha\beta\phi_2 \\ 4\beta^2\phi_2 & -4\beta^2\phi_3 & 2\beta^2\phi_2 & -2\beta^2\phi_6 & -2\alpha\beta\phi_3 & 2\alpha\beta\phi_3 \\ -6\beta^2\phi_1 & 2\beta^2\phi_2 & -3\beta^2\phi_1 & -\beta^2\phi_4 & \alpha\beta\phi_2 & -\alpha\beta\phi_2 \\ -2\beta^2\phi_4 & -2\beta^2\phi_6 & -\beta^2\phi_4 & \beta^2\phi_5 & -\alpha\beta\phi_6 & \alpha\beta\phi_6 \\ 2\alpha\beta\phi_2 & -2\alpha\beta\phi_3 & \alpha\beta\phi_2 & -\alpha\beta\phi_6 & -\alpha^2\phi_3 & \alpha^2\phi_3 \\ -2\alpha\beta\phi_2 & 2\alpha\beta\phi_3 & -\alpha\beta\phi_2 & \alpha\beta\phi_6 & \alpha^2\phi_3 & -\alpha^2\phi_3 \end{bmatrix}. \tag{19}$$

Proof of Theorem 3.4. First, observe that at any time n, there are $n+2$ pendant edges and $n+1$ internal edges in a rooted tree. That is,

$$U_{n,1} + U_{n,2} + U_{n,3} + U_{n,4} = n + 2 \quad \text{and} \quad U_{n,5} + U_{n,6} = n + 1.$$

This gives

$$\begin{aligned} \|\widetilde{U}_n\|_1 &= (1-\alpha)\sum_{j=1}^{4} U_{n,j} + \alpha \sum_{j=5}^{6} U_{n,j} \\ &= (1-\alpha)(n+2) + \alpha(n+1) = n + 2 - \alpha. \end{aligned}$$

Therefore, from (2) we get,

$$\mathbb{E}[\chi_n|\mathcal{F}_{n-1}] = \frac{U_{n-1}T_\alpha}{\|U_{n-1}T_\alpha\|_1} = \frac{U_{n-1}T_\alpha}{n+1-\alpha},$$

and

$$\mathbb{E}[U_n|\mathcal{F}_{n-1}] = U_{n-1} + \mathbb{E}[\chi_n|\mathcal{F}_{n-1}]R = U_{n-1} + \frac{1}{n+1-\alpha}U_{n-1}T_\alpha R.$$

Multiplying both sides by T_α, we get

$$\mathbb{E}[\widetilde{U}_n|\mathcal{F}_{n-1}] = \widetilde{U}_{n-1} + \left(\frac{1}{\|\widetilde{U}_{n-1}\|_1}\widetilde{U}_{n-1}\right) RT_\alpha.$$

Hence, $(\widetilde{U}_n)_{n\geq 0}$ is a classical uniform urn model with replacement matrix $R_\alpha = RT_\alpha$.

Note that **(A1)** holds because the general Ford's dynamics on a rooted tree is well defined at every time n, thus the corresponding urn model satisfies the assumption of tenability. That is, it is always possible to draw balls without getting stuck with the replacement rule. Note that R_α is diagonalizable as

$$VR_\alpha V^{-1} = \Lambda$$

holds with $\Lambda = \text{diag}\big(1, 0, 0, 0, -2(1-\alpha), -(3-2\alpha)\big)$,

$$V^{-1} = \begin{bmatrix} 1 & \frac{1}{\beta} & 0 & 0 & 1 & 1-\alpha \\ 1 & 0 & \frac{1}{\beta} & 0 & 1 & 3-\alpha \\ 1 & \frac{-2}{\beta} & 0 & \frac{3}{\beta} & \frac{-(2-\alpha)}{\beta} & -5+\alpha \\ 1 & 0 & 0 & \frac{1}{\beta} & \frac{-(2-\alpha)}{\beta} & -3+\alpha \\ 1 & 0 & \frac{-2}{\alpha} & \frac{1}{\alpha} & 1 & 3-\alpha \\ 1 & 0 & 0 & \frac{-1}{\alpha} & 1 & 1-\alpha \end{bmatrix}, \tag{20}$$

and V in (21) that is given in Box I. Therefore, R_α satisfies condition **(A4)**. Next, **(A2)** holds because R_α has eigenvalues

$$1, \quad 0, \quad 0, \quad 0, \quad -2(1-\alpha), \quad -(3-2\alpha),$$

which are all real. The maximal eigenvalue $\lambda_1 = 1$ is positive with $\lambda_1 > 2\lambda$ holds for all other eigenvalues λ of R_α. Furthermore, put $\widetilde{\mathbf{u}}_i = V^{-1}\mathbf{e}_i^\top$ and $\widetilde{\mathbf{v}}_i = \mathbf{e}_i V$ for $1 \leq i \leq 6$. Then **(A3)** follows by noting that $\widetilde{\mathbf{u}}_1 = (1, 1, 1, 1, 1, 1)^\top$ is a right eigenvector, and

$$\begin{aligned} \widetilde{\mathbf{v}}_1 = \frac{1}{2(3-2\alpha)}\big(&2(1-\alpha)^2, 2(1-\alpha)^2, \\ &(1-\alpha)^2, 1-\alpha^2, \alpha(1-\alpha), \alpha(5-3\alpha)\big) \end{aligned}$$

is the principal left eigenvector.

Since all the assumptions **(A1)–(A4)** are satisfied by the replacement matrix R_α, by Theorem 2.2, (16) holds. Furthermore, since

$$\widetilde{\Sigma} = \sum_{i,j=2}^{6} \frac{\lambda_i\lambda_j \widetilde{\mathbf{u}}_i^\top \text{diag}(\widetilde{\mathbf{v}}_1)\widetilde{\mathbf{u}}_j}{1-\lambda_i-\lambda_j}\widetilde{\mathbf{v}}_i^\top\widetilde{\mathbf{v}}_j, \tag{22}$$

by (16), it follows that (18) holds. □

3.3. *Proof of Theorem 3.1*

Proof. Since $\alpha \in (0, 1)$, it follows that the matrix T_α defined in (14) is invertible and its inverse is

$$T_\alpha^{-1} = \frac{1}{\alpha(1-\alpha)}\text{diag}(\alpha, \alpha, \alpha, \alpha, 1-\alpha, 1-\alpha),$$

which is also a diagonal matrix, and so $(T_\alpha^{-1})^\top = T_\alpha^{-1}$. Note that we have $U_n = \widetilde{U}_n T_\alpha^{-1}$. Furthermore, consider the vector $\widetilde{\mathbf{v}}_1$ in (17)

$$V = \frac{1}{2(3-2\alpha)} \begin{bmatrix} 2\beta^2 & 2\beta^2 & \beta^2 & (1+\alpha)\beta & \alpha\beta & \alpha(5-3\alpha) \\ 2\beta(1+\alpha-\alpha^2) & 2\beta^3 & -(2-\alpha)\beta^2 & (2-\alpha)\beta^2 & -\alpha\beta^2 & -\alpha\beta(5-3\alpha) \\ 2\alpha\beta^2 & 2\alpha(2-\alpha)\beta & \alpha\beta^2 & -\alpha\beta^2 & -\alpha(3-\alpha)\beta & -3\alpha\beta^2 \\ 2\alpha(2-\alpha)\beta & 2\alpha\beta^2 & \alpha(2-\alpha)\beta & -\alpha(2-\alpha)\beta & \alpha^2\beta & -3\alpha(2-\alpha)\beta \\ 2(2-\alpha)\beta & -2\beta^2 & (2-\alpha)\beta & -(4-\alpha)\beta & -\alpha\beta & \alpha\beta \\ -2\beta & 2\beta & -\beta & \beta & \alpha & -\alpha \end{bmatrix}. \tag{21}$$

Box I.

and let

$$\mathbf{v} = \widetilde{\mathbf{v}}_1 T_\alpha^{-1} = \frac{1}{2(3-2\alpha)}\big(2(1-\alpha), 2(1-\alpha), 1-\alpha, 1+\alpha, 1-\alpha, 5-3\alpha\big).$$

Since $\widetilde{U}_n/n \xrightarrow{\text{a.s.}} \widetilde{\mathbf{v}}_1$ holds in view of (16) in Theorem 3.4,

$$\frac{U_n}{n} \xrightarrow{\text{a.s.}} \mathbf{v}, \tag{23}$$

which concludes the proof of the almost sure convergence in (7).

Consider the covariance matrix $\widetilde{\Sigma}$ for $\widetilde{U}_n$ as stated in (19), then by straightforward calculation we have

$$\Sigma = (T_\alpha^{-1})^\top \widetilde{\Sigma} T_\alpha^{-1} = T_\alpha^{-1} \widetilde{\Sigma} T_\alpha^{-1}.$$

Since

$$\frac{\widetilde{U}_n - n\widetilde{\mathbf{v}}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \widetilde{\Sigma}),$$

in view of Theorem 3.4, we get

$$\frac{U_n - n\mathbf{v}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, (T_\alpha^{-1})^\top \widetilde{\Sigma}\, T_\alpha^{-1}\big) = \mathcal{N}(\mathbf{0}, \Sigma).$$

This completes the proof. □

4. Exact distributions

In this section, we present recursion formulas for computing the joint distributions of cherries and pitchforks, their means, variances and covariance for fixed n under the Ford model.

The following result on the exact computation of the joint probability mass function (pmf) of A_n and C_n can be regarded as a generalization of the existing results on the Yule model (e.g., when $\alpha = 0$ by Wu and Choi, 2016, Theorem 1) and the PDA model (e.g., $\alpha = 1/2$ by Wu and Choi, 2016, Theorem 4). A related result for unrooted trees is presented in Choi et al. (2020).

Theorem 4.1. *Consider $n \geq 3$, $0 \leq a \leq n/3$ and $1 \leq b \leq n/2$. Under the Ford model with parameter $\alpha \in [0, 1]$, we have*

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

Proof of Theorem 4.1. Fix $n \geq 3$, and let $T_2, \ldots, T_n, T_{n+1}$ be a sequence of random trees generated by the Ford process, that is, T_2 contains two leaves and $T_{i+1} = T_i[e_i]$ for a random edge e_i in T_i chosen according to the Ford model for $2 \leq i \leq n$. Then, we have

$$\begin{aligned} \mathbb{P}(A_{n+1} = a, C_{n+1} = b) &= \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b) \\ &= \sum_{p,q} \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = p, C(T_n) = q) \\ &\quad \times \mathbb{P}(A(T_n) = p, C(T_n) = q) \\ &= \sum_{p,q} \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = p, C(T_n) = q) \\ &\quad \times \mathbb{P}(A_n = p, C_n = q), \end{aligned} \tag{24}$$

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

Let e_n be the edge in T_n chosen in the above Ford process for generating T_{n+1}, that is, $T_{n+1} = T_n[e_n]$. Since Lemma 2.1 implies that

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = p, C(T_n) = q) = 0 \tag{25}$$

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 (24): case (i): $p = a, q = b$; case (ii): $p = a+1, q = b-1$; case (iii): $p = a-1, q = b$; and case (iv): $p = a, q = b-1$.

First, Lemma 2.1 implies that case (i) occurs if and only if $e_n \in E_1(T_n) \cup E_6(T_n)$, and that $E_1(T_n) \cup E_6(T_n)$ contains precisely $2A(T_n)$ pendent edges and $(n-1) + A(T_n) - C(T_n)$ interior edges. Therefore, we have

$$\begin{aligned} \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = a, C(T_n) = b) &= \frac{2A(T_n)(1-\alpha) + \alpha(n-1+A(T_n)-C(T_n))}{n-\alpha} \\ &= \frac{2a + \alpha(n-a-b-1)}{n-\alpha}. \end{aligned} \tag{26}$$

Similarly, case (ii) occurs if and only if $e_n \in E_3(T_n)$, which contains $A(T_n)$ pendent edges and no interior edges. Therefore we have

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = a+1, C(T_n) = b-1) = \frac{(a+1)(1-\alpha)}{n-\alpha}. \tag{27}$$

Next, case (iii) occurs precisely when $e_n \in E_2(T_n) \cup E_5(T_n)$, which contains $2(C(T_n) - A(T_n))$ pendent edges and $C(T_n) - A(T_n)$ interior edges. Thus

$$\begin{aligned} \mathbb{P}(A(T_{n+1}) = a, C(T_{n+1}) = b \mid A(T_n) = a-1, C(T_n) = b) \\ = \frac{2(1-\alpha)(b-a+1) + \alpha(b-a+1)}{n-\alpha} = \frac{(2-\alpha)(b-a+1)}{n-\alpha}. \end{aligned} \tag{28}$$

Finally, case (iv) occurs if and only if e_n is in $E_4(T_n)$, which contains precisely $n - A(T_n) - 2C(T_n)$ pendent edges and no interior edges. Hence,

$$\mathbb{P}(A(T_{n+1}) = a, C(T_{n+1} = b) \mid A(T_n) = a, C(T_n) = b-1)$$

$$= \frac{(1-\alpha)(n-a-2b+2)}{n-\alpha}. \tag{29}$$

Substituting Eqs. (26)–(29) into Eq. (24) completes the proof of the theorem. □

To study the moments of A_n and C_n, we present below a functional recursion form of Theorem 4.1, whose proof is similar to that in Wu and Choi (2016, Theorem 2) by using the recursion in Theorem 4.1 and a bivariate indicator function, and hence omitted here.

Proposition 4.2. *Let $\varphi : \mathbb{N}\times\mathbb{N} \to \mathbb{R}$ be an arbitrary function. For $n \geq 3$, under the Ford model with parameter $\alpha \in [0, 1]$, we have*

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

For a fix integer k, consider the indicator function $I_k(y)$ that equals to 1 if $y = k$, and 0 otherwise. Then, applying Proposition 4.2 with $\varphi(x, y) = I_k(x)$ leads to the following result on the distribution of cherries.

Corollary 4.3. *For integers $n \geq 3$ and $0 \leq k \leq n/2$, under the Ford model with parameter $\alpha \in [0, 1]$ we have*

$$\begin{aligned}&(n-\alpha)\mathbb{P}(C_{n+1} = k) = [(n-1)\alpha + 2(1-\alpha)k]\mathbb{P}(C_n = k)\\ &\quad + (1-\alpha)(n-2k+2)\mathbb{P}(C_n = k-1).\end{aligned}$$

Similarly, applying Proposition 4.2 with appropriate functions φ leads to the following recurrence relation on the moments of the joint distributions. The proofs are similar to those in Wu and Choi (2016, Corollary 4 & Proposition 5) and hence omitted here.

Corollary 4.4. *For $n \geq 3$, under the Ford model with parameter $\alpha \in [0, 1]$, we have*

$$(n-\alpha)\mathbb{E}[C_{n+1}] - (n-2+\alpha)\mathbb{E}[C_n] = n(1-\alpha), \tag{30}$$

$$(n-\alpha)\mathbb{E}[A_{n+1}] - (n-3+\alpha)\mathbb{E}[A_n] = (2-\alpha)\mathbb{E}[C_n], \tag{31}$$

$$\begin{aligned}&(n-\alpha)\mathbb{E}[C_{n+1}^2] - (n-4+3\alpha)\mathbb{E}[C_n^2]\\ &\quad = 2(n-1)(1-\alpha)\mathbb{E}[C_n] + n(1-\alpha),\end{aligned} \tag{32}$$

$$\begin{aligned}&(n-\alpha)\mathbb{E}[A_{n+1}C_{n+1}] - (n-5+3\alpha)\mathbb{E}[A_nC_n]\\ &\quad = (n-1)(1-\alpha)\mathbb{E}[A_n] + (2-\alpha)\mathbb{E}[C_n^2],\end{aligned} \tag{33}$$

$$\begin{aligned}&(n-\alpha)\mathbb{E}[A_{n+1}^2] - (n-6+3\alpha)\mathbb{E}[A_n^2] = 2(2-\alpha)\mathbb{E}[A_nC_n]\\ &\quad + (2-\alpha)\mathbb{E}[C_n] - \mathbb{E}[A_n],\end{aligned} \tag{34}$$

with initial conditions $\mathbb{E}[A_3] = \mathbb{E}[C_3] = \mathbb{E}[A_3^2] = \mathbb{E}[C_3^2] = \mathbb{E}[A_3C_3] = 1$.

Remark 4. Let $\mu_n = \mathbb{E}[C_n]$ and $\sigma_n^2 = \mathrm{var}(C_n)$. Substituting $\mathbb{E}[C_n^2] = \sigma_n^2 + \mu_n^2$ into (32) and applying (30), we obtain below a recurrence relation of the σ_n^2, which was also obtained by Ford (2006, Theorem 60):

$$\begin{aligned}&(n-\alpha)\sigma_{n+1}^2 - (n-4+3\alpha)\sigma_n^2 = -\frac{4(1-\alpha)^2}{n-\alpha}\mu_n^2\\ &\quad + \frac{2(1-\alpha)[(1-2\alpha)n+\alpha]}{n-\alpha}\mu_n + \frac{\alpha(1-\alpha)n(n-1)}{n-\alpha}.\end{aligned}$$

As an application of Corollary 4.4, we obtain the formulas for the mean of A_n and that of C_n under the Ford model in the next theorem. This theorem extends existing results on the Yule and the PDA models (see, e.g., Wu and Choi, 2016 and the references therein). Note that the mean of C_n as stated in Theorem 4.5(i) was first obtained by Ford (2006) and is included here for completeness.

Theorem 4.5. *Under the Ford model with parameter $\alpha \in [0, 1]$, for $n \geq 3$ we have*

(i) $\mathbb{E}[C_n] = \dfrac{1-\alpha}{3-2\alpha}\,n + \dfrac{\alpha}{2(3-2\alpha)} + x_n$, *where* $x_3 = \frac{\alpha}{2(3-2\alpha)}$ *and for $n \geq 4$,*

$$\begin{aligned}x_n &= \frac{\alpha}{2(3-2\alpha)}\prod_{i=3}^{n-1}\frac{i-2+\alpha}{i-\alpha}\\ &= \frac{\alpha\Gamma(3-\alpha)}{2(3-2\alpha)\Gamma(1+\alpha)}n^{-2(1-\alpha)}\left(1+o(1)\right);\end{aligned} \tag{35}$$

(ii) $\mathbb{E}[A_n] = \dfrac{1-\alpha}{2(3-2\alpha)}\,n + \dfrac{\alpha}{2(3-2\alpha)} + y_n$, *where* $y_3 = \frac{1}{2}$, $y_4 = \frac{\alpha(5-3\alpha)}{2(3-\alpha)(3-2\alpha)}$, *and for $n \geq 5$,*

$$\begin{aligned}y_n &= \frac{\alpha(2n-3+\alpha-n\alpha)}{2(3-2\alpha)(3-\alpha)}\prod_{j=4}^{n-1}\frac{j-3+\alpha}{j-\alpha}\\ &= \frac{\alpha(2-\alpha)}{2(3-2\alpha)}\frac{\Gamma(3-\alpha)}{\Gamma(1+\alpha)}n^{-2(1-\alpha)}\left(1+o(1)\right).\end{aligned} \tag{36}$$

The proof of Theorem 4.5 and that of Theorem 4.6, which concerns higher order expansions of the second moments, are presented in the Appendix.

Theorem 4.6. *Under the Ford model with parameter $\alpha \in [0, 1]$, we have*

(i) $\mathrm{var}(C_n) = \frac{(1-\alpha)(2-\alpha)}{(3-2\alpha)^2(5-4\alpha)}\,n - \frac{\alpha(1-\alpha)(2-\alpha)}{(3-2\alpha)^2(5-4\alpha)} + \mathcal{O}(n^{-2(1-\alpha)})$,

(ii) $\mathrm{cov}(A_n, C_n) = \frac{-(1-\alpha)(2-\alpha)(1-2\alpha)}{2(3-2\alpha)^2(5-4\alpha)}\,n - \frac{\alpha(1-\alpha)(2-\alpha)}{(3-2\alpha)^2(5-4\alpha)} + \mathcal{O}(n^{-2(1-\alpha)})$, *and*

(iii) $\mathrm{var}(A_n) = \frac{(1-\alpha)(69-135\alpha+96\alpha^2-24\alpha^3)}{4(3-2\alpha)^2(5-4\alpha)(7-4\alpha)}\,n + \frac{3\alpha(1-\alpha)(1-2\alpha)(5-3\alpha)}{4(3-2\alpha)^2(5-4\alpha)(7-4\alpha)} + \mathcal{O}(n^{-2(1-\alpha)})$.

Let $\rho_\alpha(A_n, C_n)$ be the correlation of A_n and C_n under the Ford model with parameter $\alpha \in [0, 1)$, which is not defined for $\alpha = 1$ because A_n and C_n are both degenerate random variables in this case. It is shown by Wu and Choi (2016, Corollaries 3 & 5) that for the Yule model $\rho_0(A_n, C_n) = -\sqrt{14/69}$ holds for $n \geq 7$, and for the PDA model $\{\rho_{1/2}(A_n, C_n)\}_{n\geq 4}$ is an increasing sequence converging to 0. Together with Theorem 4.6(ii), this leads directly to the following result which shows that $\alpha = 1/2$ is a critical value for $\rho_\alpha(A_n, C_n)$: when n is large, A_n and C_n are negatively correlated for $\alpha \in [0, 1/2]$, which is expected; and positively correlated for $\alpha \in (1/2, 1)$, which is less expected.

Corollary 4.7. *Under the Ford model, for each $0 \leq \alpha \leq 1/2$, there exists a constant $n_0(\alpha)$ such that $\rho_\alpha(A_n, C_n) < 0$ for all $n > n_0(\alpha)$. Furthermore, for each $1/2 < \alpha < 1$, there exists a constant $n_0(\alpha)$ such that $\rho_\alpha(A_n, C_n) > 0$ for all $n > n_0(\alpha)$.*

5. Discussion

Motivated by developing a unified approach to subtree statistics for the Yule and the PDA models as outlined in Wu and Choi (2016) and Choi et al. (2020, 2021), we study the joint distribution of the number of cherries and that of pitchforks for the Ford model in this paper. Our results include formulas for computing the exact joint distribution (Theorem 4.1), their means and higher order asymptotic expansions of their second moments (Theorems 4.5 and 4.6), the strong law of large numbers and the central limit theorems (Theorem 3.2). As a consequence, in

Corollary 4.7, we show that $1/2$ is a critical model parameter value for their correlation, that is, for sufficiently large n, they are negatively correlated for $0 \leq \alpha \leq 1/2$ and positively correlated for $1/2 < \alpha < 1$.

The results obtained in this paper also naturally lead to several broad questions for future work. First, suppose that we observe a rooted binary tree of n leaves with a pitchforks and b cherries. A natural question arises as to which α under the Ford model best fits the observed tree. We can employ the maximum likelihood principle to estimate α: that is $\widehat{\alpha}_{\mathrm{MLE}} = \mathrm{argmax}_{0\leq\alpha\leq 1}\mathbb{P}_\alpha(A_n = a, C_n = b)$, where $\mathbb{P}_\alpha(A_n = a, C_n = b)$ denotes the joint probability mass function at (a, b) parametrized by α. We can apply Theorem 4.1 to compute $\mathbb{P}_\alpha(A_n = a, C_n = b)$ for α belongs to a grid of the interval between 0 and 1, from which we can approximate $\widehat{\alpha}_{\mathrm{MLE}}$. Whether there exists a computationally more effective algorithm than this brute force grid search will be left for future work. One can also possibly use a statistic in the form of a function of a convex combination of A_n and C_n (i.e., $\lambda A_n + (1-\lambda)C_n$ where $0 \leq \lambda \leq 1$) for estimating the parameter α in the Ford model. A systematic study of which λ to use and which functional form of this convex combination, in our opinion, is an interesting question to address.

Second, in this paper we focused on rooted trees under the Ford model, and it would be interesting to extend them to unrooted trees as well. For instance, subtree structures under the PDA model are utilized recently in Pouryahya and Sankoff (2022) to study the evolution of polyploids, and it would be interesting to see whether the unrooted Ford model could provide a more powerful statistical model. Furthermore, Choi et al. (2020) proved that the mean number of cherries and that of pitchforks for unrooted trees converge respectively to those for rooted trees when $\alpha = 0$ (i.e., under the Yule model) while there exists a limiting gap of $1/4$ for $\alpha = 1/2$ (i.e., under the PDA model). Clearly, for the case $\alpha = 1$, this gap is 1, but other cases remain open.

Next, in addition to cherry and pitchfork statistics, one may study other subtree statistics such as k-pronged nodes and k-caterpillars, for which some analytical results are obtained by Rosenberg (2006) for the Yule model and by Chang and Fuchs (2010) for the PDA model. Other than subtree counts, a number of indices such as Colless' index, Sackin's index (see, e.g., Fischer et al., 2021 and the references therein) are used to measure tree balance or the lack of it. It would be interesting to investigate their properties under the Ford model.

Finally, it would be interesting to investigate shape statistics for several recently proposed graphical structures in evolutionary biology, which include the distribution of branch lengths in Ferretti et al. (2017) and Arbisser et al. (2018), relatively ranked tree shapes by Kim et al. (2020), and also shape statistics in phylogenetic networks, in which events such as lateral gene transfer and viral recombinations are better accommodated (see, e.g., Bienvenu et al., 2022 and Fuchs et al., 2022 for some recent results).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The work of Gursharn Kaur was supported by NUS Research Grant R-155-000-198-114, and that of Kwok Pui Choi by Singapore Ministry of Education Academic Research Fund R-155-000-188-114 and R-155-000-222-114. We thank Chris Greenman and Ariadne Thompson for stimulating discussions on the Ford model, and Weichang Yu for suggestion to improve our presentation. We would like to thank the editor and two anonymous referees for their valuable suggestions. All authors approved the version of the manuscript to be published.

Appendix. Proofs of Theorems 4.5 and 4.6

In the appendix we present the proofs of Theorems 4.5 and 4.6. To this end, we start with the two lemmas below.

Lemma 5.1. *Let a, b and c be three positive real numbers with $a > b - 1$. Given an integer $n_0 \geq 2$, suppose that $\{X_n\}_{n\geq n_0}$ is a sequence of real numbers satisfying the recursion*

$$X_{n+1} = f_n X_n + g_n, \qquad n \geq n_0,$$

where $\{f_n\}_{n\geq n_0}$ and $\{g_n\}_{n\geq n_0}$ are two sequences with $\prod_{i=\ell}^{n-1}|f_i| \leq c(n/\ell)^{-a}$ and $|g_\ell| \leq c\ell^{-b}$ for every $\ell \geq n_0$. Then, there exists a positive number C such that $|X_n| \leq Cn^{1-b}$ for all $n \geq n_0$.

Proof of Lemma 5.1. Since the solution to the given recursion is given by

$$X_n = X_{n_0}\prod_{i=n_0}^{n-1} f_i + \sum_{i=n_0}^{n-1} g_i \prod_{j=i+1}^{n-1} f_j$$

for $n > n_0$, we have

$$|X_n| \leq |X_{n_0}| \prod_{i=n_0}^{n-1} |f_i| + \sum_{i=n_0}^{n-1} |g_i| \prod_{j=i+1}^{n-1} |f_j| .$$

Considering $C = 2\max\{c(n_0)^a|X_{n_0}|, (c^2 2^a)/(a-b+1)\}$, then the lemma follows by noticing that

$$|X_{n_0}| \prod_{i=n_0}^{n-1} |f_i| \leq c{n_0}^a|X_{n_0}|n^{-a} \leq Cn^{-a}/2 \leq Cn^{1-b}/2,$$

and

$$\sum_{i=n_0}^{n-1} |g_i| \prod_{j=i+1}^{n-1} |f_j| \leq c \sum_{i=n_0}^{n-1} |g_i| \left(\frac{i+1}{n}\right)^a$$

$$\leq \frac{c^2 2^a}{n^a} \sum_{i=n_0}^{n-1} i^{a-b} \leq \frac{c^2 2^a n^{a-b+1}}{n^a(a-b+1)} \leq \frac{C}{2} n^{1-b}. \quad \square$$

Lemma 5.2. *For $\alpha \in [0, 1]$ and three finite non-negative integers ℓ, k, m such that $\ell \geq k$ and $m \geq 1$, there exists a positive constant $K = K(\alpha, m)$ such that*

$$\prod_{i=\ell}^{n-1} \left| \frac{i-k+m\alpha}{i-\alpha} \right| \leq K \left(\frac{n}{\ell}\right)^{-k+(m+1)\alpha} \quad \text{for all } 1 \leq \ell \leq n-1. \tag{37}$$

Furthermore, as $n \to \infty$ we have

$$\prod_{i=\ell}^{n-1} \frac{i-k+m\alpha}{i-\alpha} = \frac{\Gamma(\ell-\alpha)}{\Gamma(\ell-k+m\alpha)} n^{-k+(m+1)\alpha} \left(1+o(1)\right). \tag{38}$$

Proof of Lemma 5.2. First, (37) follows from Choi et al. (2021, Lemma 2). To prove (38), note that

$$\begin{aligned}\prod_{i=\ell}^{n-1}\frac{i-k+m\alpha}{i-\alpha} &= \prod_{i=\ell}^{n-1}\frac{\Gamma(i+1-k+m\alpha)\Gamma(i-\alpha)}{\Gamma(i-k+m\alpha)\Gamma(i+1-\alpha)}\\ &= \frac{\Gamma(n-k+m\alpha)}{\Gamma(\ell-k+m\alpha)}\frac{\Gamma(\ell-\alpha)}{\Gamma(n-\alpha)}\\ &= \frac{\Gamma(\ell-\alpha)}{\Gamma(\ell-k+m\alpha)}\frac{\Gamma(n+m\alpha)}{\Gamma(n-\alpha)}\prod_{j=1}^{k}\frac{1}{n-j+m\alpha}.\end{aligned} \tag{39}$$

By Stirling's approximation formula, $\Gamma(x) = \sqrt{2\pi}\;x^{x-1/2}e^{-x}(1+o(1))$, we have the following well-known approximation:

$$\frac{\Gamma(n+m\alpha)}{\Gamma(n-\alpha)} = n^{(m+1)\alpha}\,(1+o(1))\,. \tag{40}$$

Combining (40) and

$$\prod_{j=1}^{k}\frac{1}{n-j+m\alpha} = n^{-k}\,(1+o(1))\,,$$

we get (38). □

Next, we present the proof of Theorem 4.5. Intuitively, from our asymptotic result, we notice that for a fixed parameter α, the value $\mathbb{E}[C_n]$ can be decomposed into the sum of a linear term $(1-\alpha)n/(3-2\alpha)$ (where the coefficient $(1-\alpha)/(3-2\alpha)$ is given by (10)) and a remaining term for which a recurrence relation can be derived from the recurrence of $\mathbb{E}[C_n]$ in Corollary 4.4. A similar method is utilized for computing $\mathbb{E}[A_n]$ in the proof below.

Proof of Theorem 4.5. To prove part (i), we consider

$$x_n = \mathbb{E}[C_n] - \frac{1-\alpha}{3-2\alpha}\,n - \frac{\alpha}{2(3-2\alpha)}, \quad n \geq 3. \tag{41}$$

Since $\mathbb{E}[C_3] = 1$, we get $x_3 = \alpha/(6-4\alpha)$. Furthermore, substituting (41) into (30) leads to

$$(n-\alpha)x_{n+1} - (n-2+\alpha)x_n = 0, \quad n \geq 3,$$

and hence

$$x_n = x_3\prod_{i=3}^{n-1}\frac{i-2+\alpha}{i-\alpha} = x_3\frac{\Gamma(3-\alpha)}{\Gamma(1+\alpha)}\frac{\Gamma(n-2+\alpha)}{\Gamma(n-\alpha)}, \quad n \geq 4.$$

Together with Lemma 5.2, this establishes (35), and hence completes the proof of part (i).

To prove part (ii), we consider

$$y_n = \mathbb{E}[A_n] - \frac{1-\alpha}{2(3-2\alpha)}\,n - \frac{\alpha}{2(3-2\alpha)} \tag{42}$$

for $n \geq 3$. Then, $y_3 = 1/2$. Furthermore, substituting (42) and (41) into (31) leads to

$$y_{n+1} = \frac{n-3+\alpha}{n-\alpha}y_n + \frac{2-\alpha}{n-\alpha}x_n, \quad n \geq 3.$$

Solving this recurrence relation gives us $y_4 = \alpha(5-3\alpha)/[2(3-\alpha)(3-2\alpha)]$ and for $n \geq 5$,

$$\begin{aligned} y_n &= y_3\prod_{i=3}^{n-1}\frac{i-3+\alpha}{i-\alpha} + \sum_{i=3}^{n-1}\frac{2-\alpha}{i-\alpha}x_i\prod_{j=i+1}^{n-1}\frac{j-3+\alpha}{j-\alpha}\\ &= \frac{1}{2}\prod_{i=3}^{n-1}\frac{i-3+\alpha}{i-\alpha} + \frac{(2-\alpha)\alpha}{2(3-2\alpha)}\sum_{i=3}^{n-1}\prod_{j=i+1}^{n-1}\frac{j-3+\alpha}{j-\alpha}\\ &\quad\times\frac{1}{i-\alpha}\times\prod_{j=3}^{i-1}\frac{j-2+\alpha}{j-\alpha}\\ &= \frac{1}{2}\prod_{i=3}^{n-1}\frac{i-3+\alpha}{i-\alpha} + \frac{(2-\alpha)\alpha}{2(3-2\alpha)}\sum_{i=3}^{n-1}\frac{1}{3-\alpha}\prod_{j=4}^{n-1}\frac{j-3+\alpha}{j-\alpha}\\ &= \frac{1}{2}\prod_{i=3}^{n-1}\frac{i-3+\alpha}{i-\alpha} + \frac{(2-\alpha)\alpha}{2(3-2\alpha)}\frac{(n-3)}{(3-\alpha)}\prod_{j=4}^{n-1}\frac{j-3+\alpha}{j-\alpha}\\ &= \frac{\alpha(2n-3+\alpha-n\alpha)}{2(3-2\alpha)(3-\alpha)}\prod_{j=4}^{n-1}\frac{j-3+\alpha}{j-\alpha}.\end{aligned}$$

By Lemma 5.2,

$$\begin{aligned} y_n &= \frac{\alpha(2n-3+\alpha-n\alpha)}{2(3-2\alpha)(3-\alpha)}\frac{\Gamma(4-\alpha)}{\Gamma(1+\alpha)}n^{-3+2\alpha}\,(1+o(1))\\ &= \frac{\alpha(2-\alpha)}{2(3-2\alpha)}\frac{\Gamma(3-\alpha)}{\Gamma(1+\alpha)}n^{-2+2\alpha}\,(1+o(1))\,,\end{aligned}$$

as $n \to \infty$. This completes the proof of part (ii) and hence the theorem. □

In the remainder of this section we present the proof of Theorem 4.6. Theorem 3.2 indicates that for a fixed parameter α, we have $\mathrm{var}(C_n) = \mathbb{E}[C_n^2] - (\mathbb{E}[C_n])^2 = \mathcal{O}(n)$. Therefore, $\mathbb{E}[C_n^2]$ can be decomposed into the sum of a quadratic term $(1-\alpha)^2n^2/(3-2\alpha)^2$ whose coefficient is derived from Theorem 4.5, a linear term derived from (41) and Theorem 3.2 and a remaining term for which a recurrence relation can be derived from Corollary 4.4. A similar method is utilized for computing $\mathbb{E}[A_nC_n]$ and $\mathbb{E}[A_n^2]$ in the proof below.

Proof of Theorem 4.6.

Since the theorem clearly holds for $\alpha = 1$, we shall assume that $\alpha \in [0, 1)$ in the remainder of the proof. Furthermore, we will use the same x_n and y_n as defined in Theorem 4.5, and the fact that $x_n = \mathcal{O}(n^{-2(1-\alpha)})$ and $y_n = \mathcal{O}(n^{-2(1-\alpha)})$ as $n \to \infty$. We start with the proof of part (i). To this end, we let

$$\begin{aligned} z_n = \mathbb{E}[C_n^2] &- \frac{(1-\alpha)^2}{(3-2\alpha)^2}\,n^2 - \frac{2(1-\alpha)(1+2\alpha-2\alpha^2)}{(5-4\alpha)(3-2\alpha)^2}\,n\\ &+ \frac{\alpha(8-17\alpha+8\alpha^2)}{4(5-4\alpha)(3-2\alpha)^2}, \quad n \geq 3.\end{aligned} \tag{43}$$

Since $\mathbb{E}[C_3^2] = 1$, we get

$$z_3 = \frac{88\alpha^3-213\alpha^2+152\alpha-24}{4(3-2\alpha)^2(5-4\alpha)}.$$

Next, substituting (43) into (32) leads to

$$(n-\alpha)z_{n+1} - (n-4+3\alpha)z_n = 2(1-\alpha)(n-1)x_n, \quad n \geq 3.$$

Furthermore, using Theorem 4.5(i) we have

$$\begin{aligned}\mathrm{var}(C_n) = \mathbb{E}[C_n^2] - (\mathbb{E}[C_n])^2 &= \frac{(1-\alpha)(2-\alpha)}{(5-4\alpha)(3-2\alpha)^2}\,n\\ &- \frac{\alpha(1-\alpha)(2-\alpha)}{(5-4\alpha)(3-2\alpha)^2} + \upsilon_n - x_n^2,\end{aligned} \tag{44}$$

where

$$\upsilon_n = z_n - \frac{2(1-\alpha)}{3-2\alpha}nx_n - \frac{\alpha}{3-2\alpha}x_n = z_n - \frac{[2(1-\alpha)n+\alpha]}{3-2\alpha}x_n.$$

Then, for $n \geq 3$, we have

$$\begin{aligned}(n-\alpha)\upsilon_{n+1} &= (n-\alpha)z_{n+1} - \frac{[2(1-\alpha)(n+1)+\alpha]}{3-2\alpha}(n-\alpha)x_{n+1}\\ &= (n-4+3\alpha)\upsilon_n - \frac{2(1-\alpha)}{3-2\alpha}x_n\end{aligned}$$

and hence also

$$v_{n+1} = \frac{n-4+3\alpha}{n-\alpha} v_n - \frac{2(1-\alpha)}{(3-2\alpha)} \frac{x_n}{(n-\alpha)}. \tag{45}$$

Consider

$$f_n = \frac{n-4+3\alpha}{n-\alpha} \quad \text{and} \quad g_n = -\frac{2(1-\alpha)x_n}{(3-2\alpha)(n-\alpha)}$$

for $n \geq 3$, and let $a = 4-4\alpha$ and $b = 3-2\alpha$. Then, by Lemma 5.2, it follows that there exists a constant K_1 such that

$$\prod_{i=\ell}^{n-1} |f_i| = \prod_{i=\ell}^{n-1} \frac{i-4+3\alpha}{i-\alpha} \leq K_1 \left(\frac{n}{\ell}\right)^{-4+4\alpha} = K_1 \left(\frac{n}{\ell}\right)^{-a}$$
$$\text{for all} \quad 3 \leq \ell \leq n-1,$$

and by Theorem 4.5 there exists a constant K_2 such that

$$|g_n| = \frac{2(1-\alpha)x_n}{(3-2\alpha)(n-\alpha)} < \frac{4(1-\alpha)}{(3-2\alpha)} \frac{x_n}{n} \leq K_2\, n^{-3+2\alpha} = K_2 n^{-b}$$
$$\text{for} \quad n \geq 3.$$

Since $a - b + 1 = 2(1-\alpha) > 0$ for $\alpha \in [0,1)$, an application of Lemma 5.1 on the recursion (45) with the above f_n, g_n, a, b, and $c = \max\{K_1, K_2\}$ leads to $v_n = \mathcal{O}(n^{-2+2\alpha})$, and hence $v_n - x_n^2 = \mathcal{O}(n^{-2+2\alpha})$. This, together with (44), completes the proof of (i).

To prove part (ii), we consider

$$t_n = \mathbb{E}[A_n C_n] - \frac{(1-\alpha)^2}{2(3-2\alpha)^2} n^2 + \frac{(1-\alpha)(4-25\alpha+16\alpha^2)}{4(5-4\alpha)(3-2\alpha)^2} n + \frac{\alpha(8-17\alpha+8\alpha^2)}{4(5-4\alpha)(3-2\alpha)^2} \tag{46}$$

for $n \geq 3$. Combining (33) and (46) leads to

$$(n-\alpha)t_{n+1} - (n-5+3\alpha)t_n = (2-\alpha)z_n + (1-\alpha)(n-1)y_n, \quad n \geq 3.$$

Since $\operatorname{cov}(A_n, C_n) = \mathbb{E}[A_n C_n] - \mathbb{E}[A_n]\mathbb{E}[C_n]$, by (41), (42) and (46), we have

$$\operatorname{cov}(A_n, C_n) = \frac{-(1-\alpha)(2-\alpha)(1-2\alpha)}{2(5-4\alpha)(3-2\alpha)^2} n - \frac{\alpha(1-\alpha)(2-\alpha)}{(5-4\alpha)(3-2\alpha)^2} + w_n - x_n y_n, \tag{47}$$

where

$$w_n = t_n - \frac{[(1-\alpha)n+\alpha]x_n + [2(1-\alpha)n+\alpha]y_n}{2(3-2\alpha)}.$$

Using straightforward but tedious algebraic simplification steps, we can show that

$$(n-\alpha)w_{n+1} - (n-5+3\alpha)w_n = (2-\alpha)v_n - \frac{1-\alpha}{3-2\alpha} x_n$$

holds for $n \geq 3$, and hence

$$w_{n+1} = \frac{n-5+3\alpha}{n-\alpha} w_n + (2-\alpha)\frac{v_n}{n-\alpha} - \frac{(1-\alpha)}{(3-2\alpha)} \frac{x_n}{(n-\alpha)}. \tag{48}$$

Similarly to the proof of part (i), applying Lemma 5.1 to the recursion (48) with $a = 5-4\alpha$, $b = 3-2\alpha$,

$$f_n = \frac{n-5+3\alpha}{n-\alpha} \quad \text{and} \quad g_n = (2-\alpha)\frac{v_n}{n-\alpha} - \frac{(1-\alpha)}{(3-2\alpha)} \frac{x_n}{(n-\alpha)},$$

we get $w_n = \mathcal{O}(n^{-2+2\alpha})$, and hence $w_n - x_n y_n = \mathcal{O}(n^{-2+2\alpha})$. This proves part (ii) in view of (47).

To prove part (iii), we consider

$$s_n = \mathbb{E}[A_n^2] - \frac{(1-\alpha)^2}{4(3-2\alpha)^2} n^2 - \frac{2(1-\alpha)(1+2\alpha+2\alpha^2)}{(5-4\alpha)(3-2\alpha)^2} n - \frac{\alpha(5-3\alpha+\alpha^2)}{4(3-2\alpha)(5-4\alpha)(7-4\alpha)}$$

for $n \geq 3$. Let

$$u_n = s_n - \frac{[(1-\alpha)n+\alpha]y_n}{3-2\alpha}.$$

Then by straightforward simplification steps, we have

$$u_{n+1} = \frac{n-6+3\alpha}{n-\alpha} u_n + \frac{\alpha(2-\alpha)}{(3-2\alpha)} \frac{y_n}{(n-\alpha)} + \frac{(2-\alpha)^2}{(3-2\alpha)} \frac{x_n}{(n-\alpha)}, \quad n \geq 3. \tag{49}$$

Furthermore, by $\operatorname{var}(A_n) = \mathbb{E}[A_n^2] - (\mathbb{E}[A_n])^2$ and Theorem 4.5(ii), we have

$$\operatorname{var}(A_n) = \frac{(1-\alpha)(69-135\alpha+96\alpha^2-24\alpha^3)}{4(3-2\alpha)^2(5-4\alpha)(7-4\alpha)} n + \frac{3\alpha(1-\alpha)(1-2\alpha)(5-3\alpha)}{4(3-2\alpha)^2(5-4\alpha)(7-4\alpha)} + u_n - y_n^2.$$

Similarly to the proof of part (i), applying Lemma 5.1 on the recursion (49) with $a = 6-4\alpha$, $b = 3-2\alpha$,

$$f_n = \frac{n-6+3\alpha}{n-\alpha}, \quad \text{and}$$
$$g_n = \left[\frac{\alpha(2-\alpha)}{(3-2\alpha)} \frac{y_n}{(n-\alpha)} + \frac{(2-\alpha)^2}{(3-2\alpha)} \frac{x_n}{(n-\alpha)}\right],$$

we get $u_n = \mathcal{O}(n^{-2+2\alpha})$ and hence also $u_n - y_n^2 = \mathcal{O}(n^{-2+2\alpha})$, which completes the proof of (iii) and hence the theorem. □

References

Aldous, D., 1991. Asymptotic fringe distributions for general families of random trees. Ann. Appl. Probab. 1, 228–266.

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

Arbisser, I.M., Jewett, E.M., Rosenberg, N.A., 2018. On the joint distribution of tree height and tree length under the coalescent. Theor. Popul. Biol. 122, 46–56.

Bai, Z.D., Hu, F., 2005. Asymptotics in randomized urn models. Ann. Appl. Probab. 15, 914–940.

Bandyopadhyay, A., Kaur, G., 2018. Linear de-preferential urn models. Adv. Appl. Probab. 50 (4), 1176–1192.

Bienvenu, F., Lambert, A., Steel, M., 2022. Combinatorial and stochastic properties of ranked tree-child networks. Random Struct. Algorithms 60, 653–689.

Blum, M.G.B., François, O., 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55 (4), 685–691.

Chang, H., Fuchs, M., 2010. Limit theorems for patterns in phylogenetic trees. J. Math. Biol. 60 (4), 481–512.

Chen, B., Ford, D., Winkel, M., 2009. A new family of Markov branching trees: the alpha-gamma model. Electron. J. Probab. 14, 400–430.

Choi, K.P., Kaur, G., Wu, T., 2021. On asymptotic joint distributions of cherries and pitchforks for random phylogenetic trees. J. Math. Biol. 83 (4), 1–34.

Choi, K.P., Thompson, A., Wu, T., 2020. On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees. Theor. Popul. Biol. 132, 92–104.

Colijn, C., Gardy, J., 2014. Phylogenetic tree shapes resolve disease transmission patterns. Evol. Medi. Public Health 2014 (1), 96–108.

Colijn, C., Plazzotta, G., 2017. A metric on phylogenetic tree shapes. Syst. Biol. 67 (1), 113–126.

Coronado, T.M., Mir, A., Rosselló, F., 2018. The probabilities of trees and cladograms under Ford's α-model. Sci. World J. 2018, 1916094.

Coronado, T.M., Mir, A., Rosselló, F., Valiente, G., 2019. A balance index for phylogenetic trees based on rooted quartets. J. Math. Biol. 79 (3), 1105–1148.

Ferretti, L., Ledda, A., Wiehe, T., Achaz, G., Ramos-Onsins, S.E., 2017. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics 207 (1), 229–240.

Fischer, M., Herbst, L., Kersting, S., Kühn, L., Wicke, K., 2021. Tree balance indices: a comprehensive survey. arXiv preprint arXiv:2109.12281.

Ford, D.J., 2006. Probabilities on Cladograms: Introduction To the Alpha Model (Ph.D. thesis). ProQuest LLC, Ann Arbor, MI, Stanford University, p. 241.

Fuchs, M., Liu, H., Yu, T.C., 2022. Limit theorems for patterns in ranked tree-child networks. arXiv preprint arXiv:2204.07676.

Hagen, O., Hartmann, K., Steel, M., Stadler, T., 2015. Age-dependent speciation can explain the shape of empirical phylogenies. Syst. Biol. 64 (3), 432–440.

Heath, T.A., Zwickl, D.J., Kim, J., Hillis, D.M., 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst. Biol. 57 (1), 160–166.

Hofri, M., Mahmoud, H., 2019. Algorithmics of Nonuniformity: Tools and Paradigms. CRC Press.

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

Holmgren, C., Janson, S., 2017. Fringe trees, Crump–Mode–Jagers branching processes and m-ary search trees. Probab. Surv. 14, 53–154.

Janson, S., 2004. Functional limit theorems for multitype branching processes and generalized Pólya urns. Stochastic Process. Appl. 110 (2), 177–245.

Kim, J., Rosenberg, N.A., Palacios, J.A., 2020. Distance metrics for ranked evolutionary trees. Proc. Natl. Acad. Sci. 117 (46), 28876–28886.

Mahmoud, H.M., 2009. Pólya Urn Models. In: Texts in Statistical Science Series, CRC Press, Boca Raton, FL, p. xii+290.

McKenzie, A., Steel, M.A., 2000. Distributions of cherries for two models of trees. Math. Biosci. 164, 81–92.

Mooers, A., Harmon, L.J., Blum, M.G., Wong, D.H., Heard, S.B., 2007. Some models of phylogenetic tree shape. In: Gascuel, O., Steel, M. (Eds.), Reconstructing Evolution: New Mathematical and Computational Advances. Oxford University Press, Oxford, pp. 149–170.

Plazzotta, G., Colijn, C., 2016. Asymptotic frequency of shapes in supercritical branching trees. J. Appl. Probab. 53 (4), 1143–1155.

Pompei, S., Loreto, V., Tria, F., 2012. Phylogenetic properties of RNA viruses. PLoS One 7 (9), e44849.

Pouryahya, F., Sankoff, D., 2022. Peripheral structures in unlabelled trees and the accumulation of subgenomes in the evolution of polyploids. J. Theoret. Biol. 532, 110924.

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

Steel, M., 2016. Phylogeny: Discrete and Random Processes in Evolution. SIAM.

Wirtz, J., Wiehe, T., 2019. The evolving Moran genealogy. Theor. Popul. Biol. 130, 94–105.

Wu, T., Choi, K.P., 2016. On joint subtree distributions under two evolutionary models. Theor. Popul. Biol. 108, 13–23.