Subtree Moments in Random Phylogenetic Trees

Ariadne István Thompson

Doctor of Philosophy

University of East Anglia

School of Computing Sciences

July 2023

This research was supported by the UKRI Biotechnology and Biological Sciences Research Council Norwich Research Park Biosciences Doctoral Training Partnership, UK [Grant number BB/M011216/1].

Declaration

I certify that the work contained in the thesis submitted by me for the degree of Doctor of Philosophy is my original work except where due reference is made to other authors, and has not been previously submitted by me for a degree at this or any other university. Some results in Chapter 3 have previously been published in the paper: Kwok Pui Choi, Ariadne Thompson, and Taoyang Wu, On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees, Theoretical Population Biology, 2020. In addition, some results from Chapter 4 have been published in the paper: Kwok Pui Choi, Gursharn Kaur, Ariadne Thompson, and Taoyang Wu, Distributions of 4-subtree patterns for uniform random unrooted phylogenetic trees, Journal of Theoretical Biology, 2024. Furthermore, manuscripts based on the research in Chapters 4 and 5 are currently in development with the intention of submission to an academic journal.

Abstract

Null models for randomly generating trees are useful for detecting and understanding evolutionary processes in real phylogenetic trees. The two most commonly used null models are the Yule-Harding-Kingman (YHK) and Proportional to Distinguishable Arrangements (PDA) models. We investigate the trees generated under these two models through the lens of subtrees, which are small recurring structures inside trees. Through a recursive approach, we obtain exact results for the joint and marginal statistical distributions of subtrees in trees generated under both models, in both rooted and unrooted trees, and for subtrees with up to four leaves (cherries, pitchforks, 4-caterpillars, and crabs). Furthermore, we present more limited marginal results for subtrees with arbitrarily many leaves.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to sincerely thank Taoyang Wu and Kwok Pui Choi for their guidance and expertise without which this would not have been possible. In addition, I would like to thank Alex Madigan, Andrew Copeman, Avril Randle, Klára Morvai, Stuart Thompson, and Penelope for their love and support.

1 Introduction

Trees are an important subclass of graphs, which can be randomly generated. In this thesis, we seek to study one property of random trees, that is, if you randomly generate a tree, statistically how many of what subtrees can you expect to find in it? To this end, in this chapter we present some general background context for this problem, and summarise existing research in the field.

Trees have been used for applications as diverse as efficient data compression [Huffman, 1952, Van Leeuwen, 1976], parsing grammar [Carnie, 2021], machine learning [Suthaharan and Suthaharan, 2016], business and economics [Magee, 1964, Gepp et al., 2010], organic chemistry [Balaban, 1985], robotics [Colledanchise and Ögren, 2018, Marzinotto et al., 2014] and even video games [Nicolau et al., 2016].

Here, we focus on their use in modelling evolutionary histories. This use can be traced back to Charles Darwin, see Figure 1.1 for an example. This includes both phylogenetic trees (in which each leaf is a species) and genealogical trees (in which each leaf is an individual in a population). See Fig.1.2 for an example of a tree in which each leaf is a different COVID-19 sample.

Even if an evolutionary tree is constructed accurately, evolutionary forces such as speciation, extinction, mutation, and natural selection can be difficult to isolate or infer. One way to accomplish this is through a null model, which is neutral to these evolutionary forces. We can then compare real empirically inferred trees against what would be statistically likely under a null model to establish evidence for other forces.

Trees can be studied statistically using topological measures such as

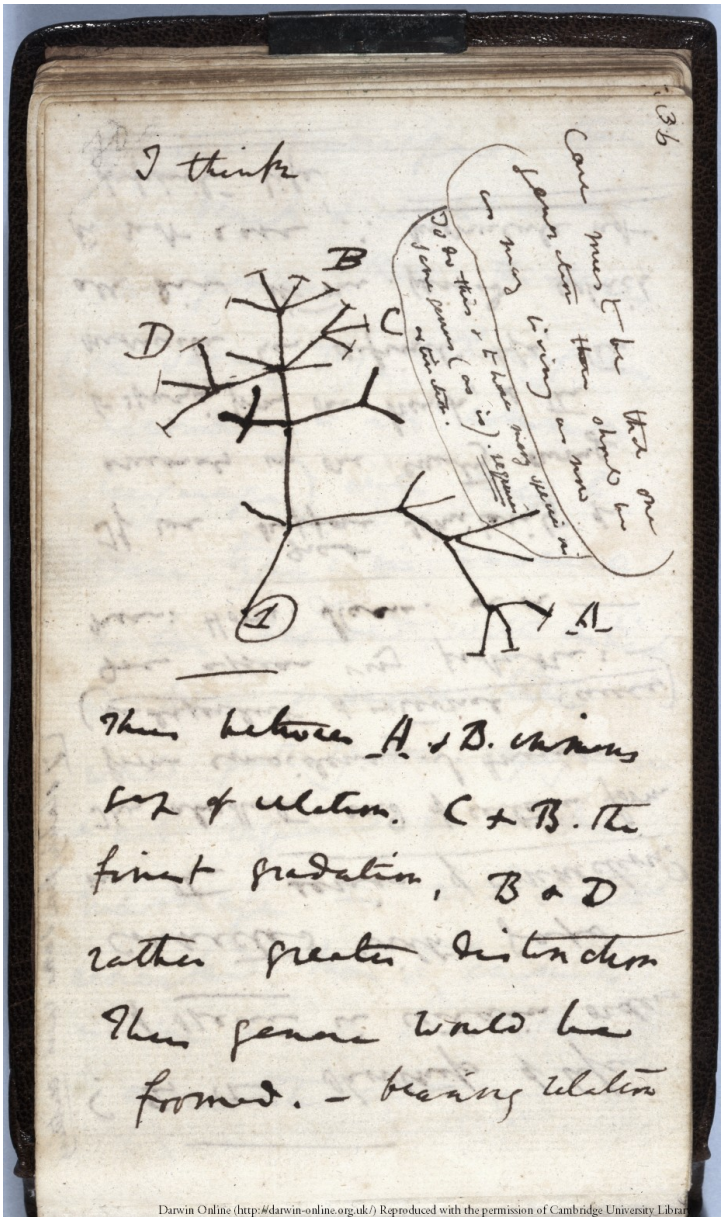

Figure 1.1: The first known phylogenetic tree, from Charles Darwin's notebook [Van Wyhe, 2002].

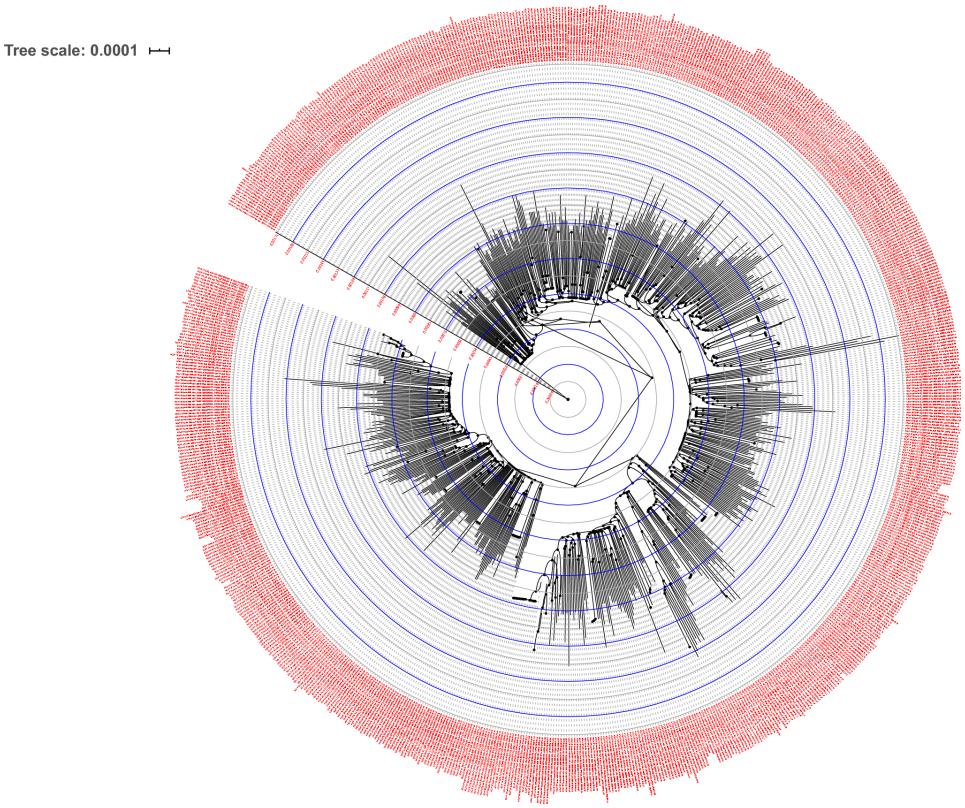

Figure 1.2: A phylogenetic tree showing the relationships between 775 different samples of COVID-19 taken in India [Kumar et al., 2021].

Sackin's index [Sackin, 1972, M. Coronado et al., 2020] (using the depth of each leaf to estimate how balanced a tree is) or branch length (where each edge in the tree is given a "length", usually representing timespan, which is useful for considering mutation). In this thesis, however, we will focus on trees' topology, and more specifically on subtrees, which are smaller topological structures arising within trees.

McKenzie and Steel [2000] found exact results for means and variances of the numbers of cherries (subtrees with two leaves) in rooted trees under the YHK and PDA models (Yule-Harding-Kingman and Proportional to Distinguishable Arrangements; see Section 2.3), as well as making some conjectures regarding corresponding results for unrooted trees, and giving an example of how these could be applied to empirical phylogenetic trees. These results were later extended to pitchforks and cherries [Wu and Choi, 2016] and then from there to unrooted trees [Choi et al., 2020].

Rosenberg [2006] considered the YHK model on random rooted trees. By first considering the probability of generating a given genealogy, he was able to compute the means and variances of the numbers of k-subtrees and k-caterpillars generated. While these results are highly notable, unfortunately the methods used do not generalise well to unrooted trees or trees generated under other models, and require more steps than the recursive approaches detailed in Chapters 3 onwards.

Ford [2006] introduced the alpha model, a random tree model generalising the YHK and PDA models, and found some limiting results for the mean and variance of the number of cherries. This was extended further to multifurcating trees with the alpha-gamma model by Chen et al. [2009].

The remainder of the thesis is organised as follows: in Chapter 2 we obtain some preliminary results on small trees. In Chapter 3 we find the means and variances of cherries and pitchforks in unrooted trees, and in Chapter 4 we extend this to 4-subtrees, in both rooted and unrooted trees. In Chapter 5 we obtain some results for general k-subtrees. Finally, in Chapter 6 we discuss the results obtained and future avenues for research. The appendix contains code used to verify the initial conditions in Section 2.6, and a summary of the results from Chapters 3 and 4.

2 Random Trees

In this chapter, we formally define the concepts and methods we will use throughout the rest of this thesis, beginning with trees and subtrees. We also present some exact results for small trees. While not significant on their own, these will be required to prove results in chapters 3, 4, and 5.

2.1 Trees

A *network* is defined as a set of points (known as *nodes* or *vertices*), and a set of pairs of points that are linked to one another (known as *edges*). A *tree* is a particular type of network which is *connected* and *acyclic*: all vertices can reach all other vertices via some sequence of edges, and the edges do not form any closed loops; or, equivalently, there is exactly one possible path between any two given vertices, following the edges.

The *degree* of a vertex is the number of edges attached to it. In a *binary* tree, all vertices have degree one or three. A *rooted* tree has a particular vertex of degree one designated as the *root*, and an *unrooted* tree does not. A vertex that has degree one, and is not the root, is known as a *leaf*. Edges attached to leaves are known as *pendant* edges, and all other edges are known as *internal* edges. A tree is *labelled* if the leaves are in some way distinguishable from one another; otherwise, it is *unlabelled* [Semple et al., 2003][Deo, 2017].

In this text, unless explicitly stated otherwise, we consider only binary, unlabelled trees, also known as *tree shapes*.

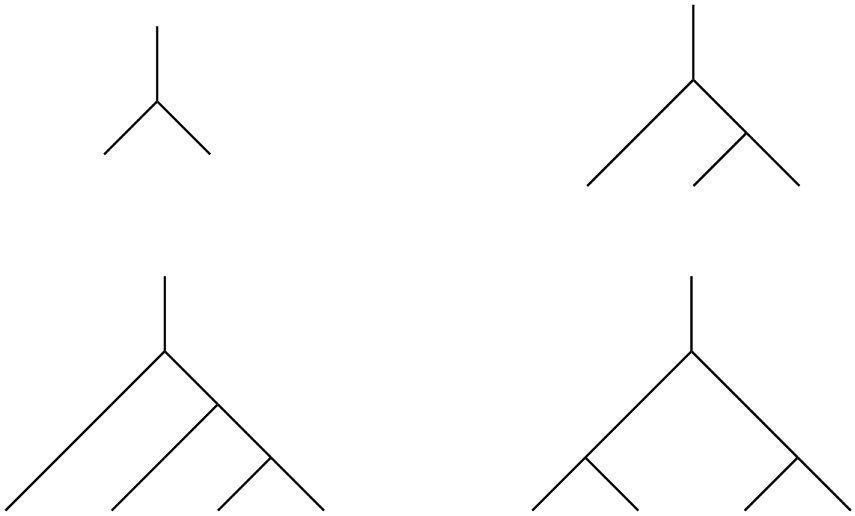

Figure 2.1: Examples of small subtrees. Top left: a cherry. Top right: a pitchfork. Bottom left: a 4-caterpillar. Bottom right: a crab.

2.2 Subtrees

A *subtree* is a single edge, and all other vertices and edges that would be included if that edge were taken as the root of a rooted tree. This is sometimes referred to specifically as a *pendant subtree* [Allen, 1998]. Note that subtrees are always rooted, even if the tree they are a part of is unrooted.

A *cherry* is a subtree comprising two pendant edges joined at a single vertex, and the root edge above. A *pitchfork* is a subtree comprising a cherry joined to another pendant edge, and the root edge above. Additionally, we define a *crab* as two cherries joined at a single vertex, and the root above. A *k-subtree* is any subtree containing k leaves, regardless of its topology, and a *k-caterpillar* is a k-subtree in which every internal vertex is joined to a pendant edge.

Let A_n, B_n, C_n, D_n be the random variables denoting the number of pitchforks, cherries, 4-caterpillars, and crabs, respectively, in a randomly generated tree of n leaves. Similarly, let $C_{k,n}$ and $S_{k,n}$ be the number of k-caterpillars and k-subtrees, respectively. Finally, we adopt the convention that for a random variable X_n denoting the number of a given subtree, let

X_n^* and X_n° denote the rooted and unrooted cases, respectively.

2.3 Random Trees

We will focus on two methods of randomly generating trees: the Yule-Harding-Kingman (YHK) model [Yule, 1925, Harding, 1971, Kingman, 1982], and the Proportional to Distinguishable Arrangements (PDA) model [Aldous, 2001]. Other more general models for generating random trees exist, such as the Ford alpha model [Ford, 2006] and beta splitting model [Aldous, 1996].

The YHK model was developed as a simple model for evolution and the growth of real phylogenetic trees. The process can be understood as follows:

1. Begin with a given small tree (usually the unique tree with two leaves in the rooted case, or with four leaves in the unrooted case).

2. Randomly choose a single edge from the pendant edge set. The edges are sampled uniformly, so every pendant edge is equally likely to be chosen.

3. "Split" the edge by inserting a node in the middle of it, connected to a new pendant edge with a new leaf.

4. Repeat steps 2 and 3 until the tree has the desired number of leaves.

The splitting process can be understood intuitively as a speciation event in the evolutionary history of a phylogenetic tree. Speciation can only occur on currently extant lineages, hence why only pendant edges can be split.

The PDA model was developed as a null model, under which every unique labelled tree has an equal probability of being generated. Its process can be understood as being identical to the YHK process, except at step 2 the edge is sampled from the entire edge set, not just the pendant edge set.

Despite the YHK model being much closer to the expected reality of how phylogenetic trees are generated in nature, real life phylogenetic trees are consistently more imbalanced than predicted by the YHK model [Aldous,

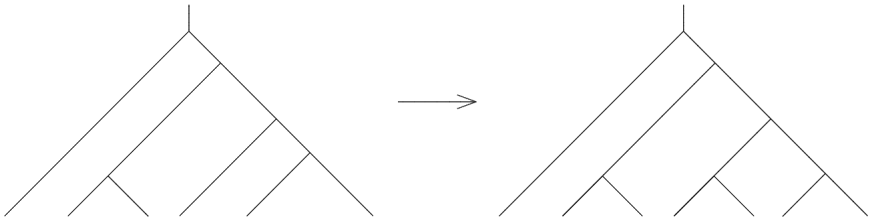

Figure 2.2: An example of the result of splitting an edge in a rooted tree. In this case, the third pendant edge from the right has been split. As a result, the number of leaves has increased from six to seven, the number of pitchforks has reduced from one to zero, the numbers of cherries has increased from two to three, and the number of crabs has increased from zero to one.

2001, Blum and François, 2006]. However, the PDA model is not a better predictor for real phylogenetic trees [Jones, 2011].

To distinguish between the two models, we let $\mathbb{E}_y$, $\mathbb{V}_y$, Cov_y, and ρ_y be the expectation, variance, covariance, and correlation coefficient respectively under the YHK model, and $\mathbb{E}_u$, $\mathbb{V}_u$, Cov_u, and ρ_u be the expectation, variance, covariance, and correlation coefficient respectively under the PDA model.

2.4 Recursion

A *recursion* is a set of equations defining the behaviour of a function of a discrete variable or variables. Typically they consist of:

1. One equation relating the value of the function at one point to the value of the function at another point, for example $f(n) = 2f(n-1)$, and

2. One equation giving the precise value of the function at a given fixed point, for example $f(0) = 1$.

Recursions can be understood as a discrete analogue of differential equations. Similarly to differential equations, they can be solved to yield a closed

form of the function described, for example $f(n) = 2^n$ is a solution to the above equations.

There are many ways to solve recursions, but we will principally be using the summation factor method, detailed in Graham et al. [1989], (eq. 2.9–2.11)

2.5 Falling Factorials

We define the falling factorial as

$$n^{\underline{k}} = \prod_{i=0}^{k-1}(n-i) = n(n-1)(n-2)\dots(n-k+1),$$

and the double falling factorial as

$$n^{\underline{\underline{k}}} = \prod_{i=0}^{k-1}(n-2i) = n(n-2)(n-4)\dots(n-2k+2).$$

We note the following identity (see Graham et al. [1989], equation 2.53)

$$\sum_{a \le k < b} k^{\underline{m}} = \frac{b^{\underline{m+1}} - a^{\underline{m+1}}}{m+1}, \quad m \neq -1. \tag{2.1}$$

Furthermore, we note (ibid., equation 6.10)

$$k^n = \sum_{i=0}^{n} \left\{ {n \atop i} \right\} k^{\underline{i}}, \quad n \geq 1 \tag{2.2}$$

by which any monomial can be converted to a sum of falling factorials. Here $\left\{ {n \atop i} \right\}$ denotes Stirling numbers of the second kind. As they are used frequently

throughout, the first few examples are given below:

$$
\begin{aligned}
k^1 &= k^{\underline{1}}, \\
k^2 &= k^{\underline{2}} + k^{\underline{1}}, \\
k^3 &= k^{\underline{3}} + 3k^{\underline{2}} + k^{\underline{1}}, \\
k^4 &= k^{\underline{4}} + 6k^{\underline{3}} + 7k^{\underline{2}} + k^{\underline{1}}, \\
k^5 &= k^{\underline{5}} + 10k^{\underline{4}} + 25k^{\underline{3}} + 15k^{\underline{2}} + k^{\underline{1}}, \\
k^6 &= k^{\underline{6}} + 15k^{\underline{5}} + 65k^{\underline{4}} + 90k^{\underline{3}} + 31k^{\underline{2}} + k^{\underline{1}}.
\end{aligned}
$$

2.6 Exact Probabilities and Expectations of Subtrees

In rooted trees, there is only one unique tree with three leaves, comprising a pitchfork. Similarly, in unrooted trees, there is only one unique tree with four leaves (comprising two cherries joined by a single edge) and one unique tree with five leaves (comprising a cherry joined to a pitchfork).

Given the procedure for recursively growing trees described in Section 2.3, it is possible to calculate the exact probabilities of particular trees being generated under the YHK or PDA models, in both rooted and unrooted trees (See Figures 2.3, 2.4, 2.5, and 2.6).

From this, it is also possible to calculate the expectations of particular subtrees. For example, consider $\mathbb{E}_y(A_6^\circ B_6^\circ)$. From Figure 2.5, there are two unique unrooted trees with six leaves. The first, on the left, contains two cherries and two pitchforks and has a probability of $\frac{4}{5}$. The second, on the right, contains three cherries and zero pitchforks, and has a probability of $\frac{1}{5}$. We can then easily see that $\mathbb{E}_y(A_6^\circ B_6^\circ) = \frac{16}{5}$.

As the number of trees to consider grows rapidly with the number of leaves, it quickly becomes intractable to calculate probabilities or expectations by this method for larger trees, but it is useful for calculating initial values for recursions, which we will use extensively in chapters 3 and 4.

We note that $\mathbb{E}_y(B_4^\circ) = 2$, $\mathbb{E}_y(B_4^{\circ 2}) = 4$, $\mathbb{E}_u(B_4^\circ) = 2$, and $\mathbb{E}_u(B_4^{\circ 2}) = 4$ all follow from the fact that the unique four-leaved unrooted tree has two cherries. Most other initial conditions can be calculated directly from

Figures 2.3, 2.4, 2.5, and 2.6. The remainder can easily be calculated by manually extending the same method to larger trees, or using the code in Appendix 7.1.

For convenience, we list here all initial conditions used in the remainder of the thesis. For rooted trees generated under the YHK model, from Figure 2.3 we have

$$\begin{aligned}
\mathbb{E}_y(C_5^*) &= \frac{1}{3} \\
\mathbb{E}_y(B_7^* C_7^*) &= \frac{8}{9} \\
\mathbb{E}_y(A_8^* C_8^*) &= \frac{86}{105} \\
\mathbb{E}_y(C_9^{*2}) &= \frac{26}{35} \\
\mathbb{E}_y(D_5^*) &= \frac{1}{6} \\
\mathbb{E}_y(B_7^* D_7^*) &= \frac{61}{90} \\
\mathbb{E}_y(A_8^* D_8^*) &= \frac{1}{7} \\
\mathbb{E}_y(D_9^{*2}) &= \frac{47}{140} \\
\mathbb{E}_y(C_9^* D_9^*) &= \frac{1}{14}
\end{aligned}$$

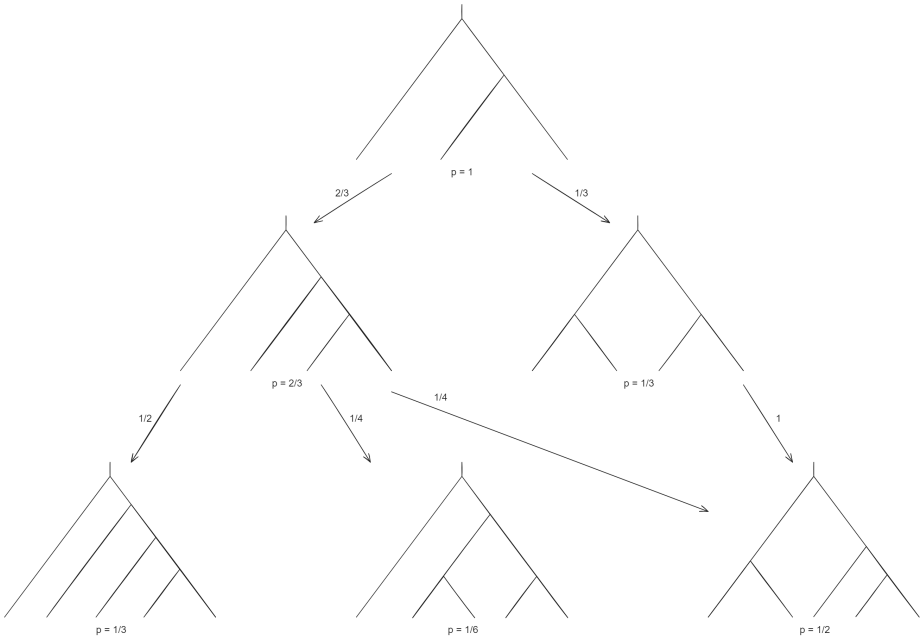

Figure 2.3: The exact probabilities of small rooted trees being generated under the YHK model. Numbers on arrows are the probability of a split in the source tree resulting in the destination tree.

Next, for rooted trees generated under the PDA model, from Fig.2.4 we have

$$\begin{aligned}
\mathbb{E}_u(C_4^*) &= \frac{4}{5} \\
\mathbb{E}_u(B_4^* C_4^*) &= \frac{4}{5} \\
\mathbb{E}_u(A_4^* C_4^*) &= \frac{4}{5} \\
\mathbb{E}_u(C_4^{*2}) &= \frac{4}{5} \\
\mathbb{E}_u(D_4^*) &= \frac{1}{5} \\
\mathbb{E}_u(B_4^* D_4^*) &= \frac{2}{5} \\
\mathbb{E}_u(A_4^* D_4^*) &= 0 \\
\mathbb{E}_u(D_4^{*2}) &= \frac{1}{5} \\
\mathbb{E}_u(C_4^* D_4^*) &= 0
\end{aligned}$$

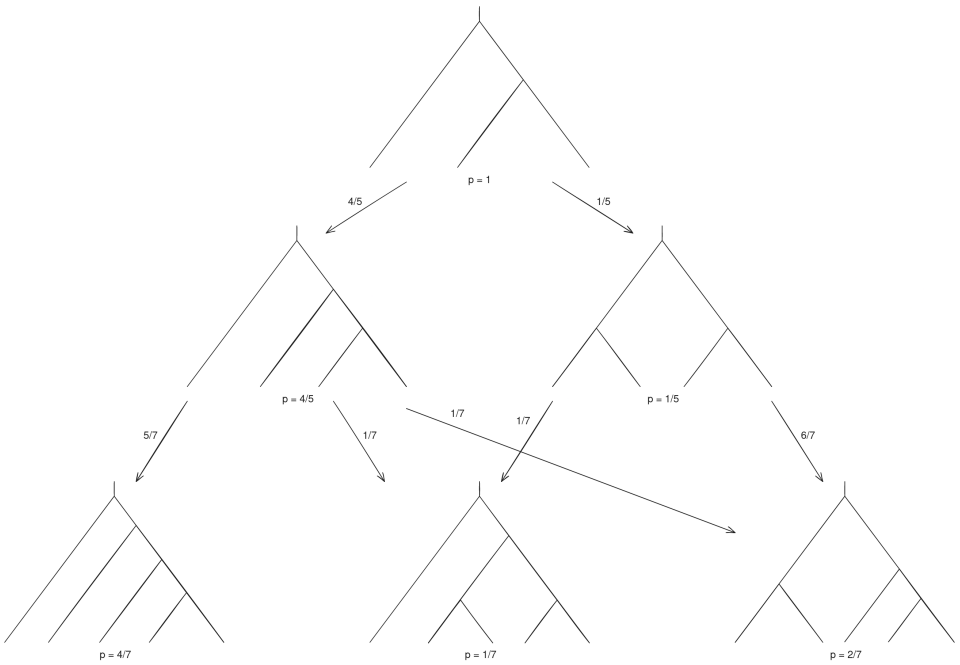

Figure 2.4: The exact probabilities of small rooted trees being generated under the PDA model. Numbers on arrows are the probability of a split in the source tree resulting in the destination tree.

Similarly, for unrooted trees generated under the YHK model, from Figure 2.5 we have

$$\begin{aligned}
\mathbb{E}_y(B_4^\circ) &= 2 \\
\mathbb{E}_y(B_4^{\circ 2}) &= 4 \\
\mathbb{E}_y(A_6^\circ) &= \frac{8}{5} \\
\mathbb{E}_y(A_6^\circ B_6^\circ) &= \frac{16}{5} \\
\mathbb{E}_y(A_6^{\circ 2}) &= \frac{16}{5} \\
\mathbb{E}_y(C_8^\circ) &= \frac{94}{105} \\
\mathbb{E}_y(B_9^\circ C_9^\circ) &= \frac{467}{210} \\
\mathbb{E}_y(A_9^\circ C_9^\circ) &= \frac{323}{210} \\
\mathbb{E}_y(C_9^{\circ 2}) &= \frac{241}{210} \\
\mathbb{E}_y(D_8^\circ) &= \frac{44}{105} \\
\mathbb{E}_y(B_9^\circ D_9^\circ) &= \frac{311}{210} \\
\mathbb{E}_y(A_8^\circ D_8^\circ) &= \frac{2}{7} \\
\mathbb{E}_y(D_9^{\circ 2}) &= \frac{67}{140} \\
\mathbb{E}_y(C_9^\circ D_9^\circ) &= \frac{31}{210}
\end{aligned}$$

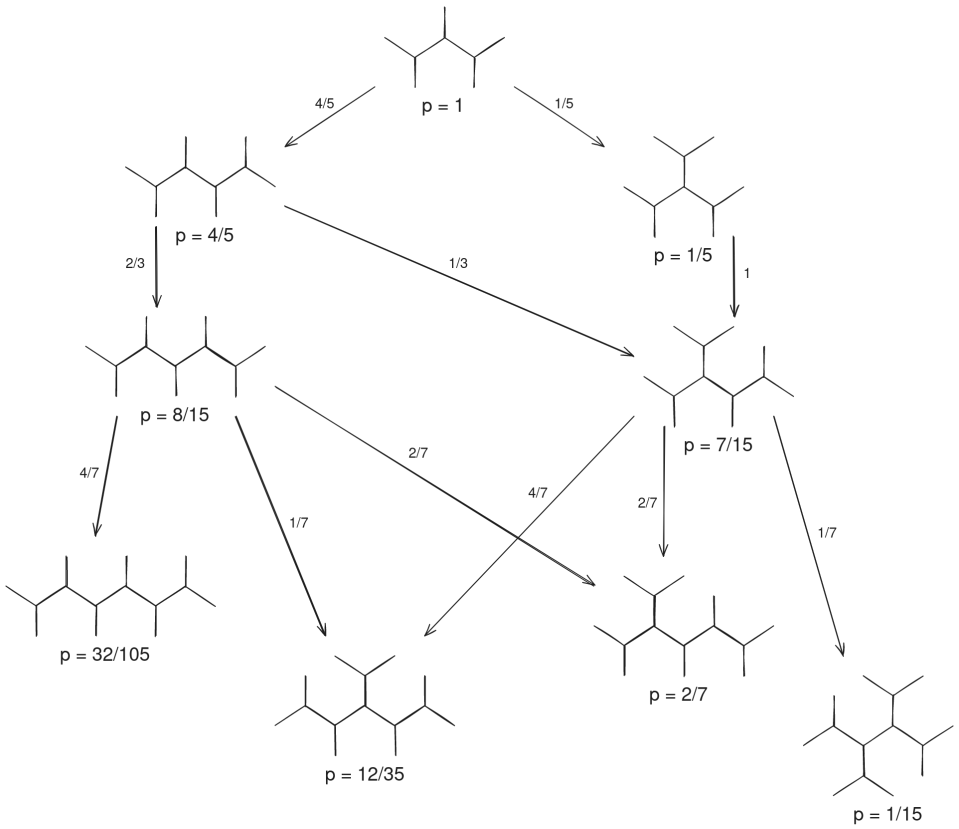

Figure 2.5: The exact probabilities of small unrooted trees being generated under the YHK model. Numbers on arrows are the probability of a split in the source tree resulting in the destination tree.

Finally, for unrooted trees generated under the PDA model, from Figure 2.6 we have

$$\begin{aligned}
\mathbb{E}_u(B_4^\circ) &= 2 \\
\mathbb{E}_u(B_4^{\circ 2}) &= 4 \\
\mathbb{E}_u(A_6^\circ) &= \frac{12}{7} \\
\mathbb{E}_u(A_6^\circ B_6^\circ) &= \frac{24}{7} \\
\mathbb{E}_u(A_6^{\circ 2}) &= \frac{24}{7} \\
\mathbb{E}_u(C_8^\circ) &= \frac{10}{33} \\
\mathbb{E}_u(B_8^\circ C_8^\circ) &= \frac{8}{3} \\
\mathbb{E}_u(A_8^\circ C_8^\circ) &= \frac{24}{11} \\
\mathbb{E}_u(C_8^{\circ 2}) &= \frac{72}{33} \\
\mathbb{E}_u(D_8^\circ) &= \frac{10}{33} \\
\mathbb{E}_u(B_8^\circ D_8^\circ) &= \frac{32}{33} \\
\mathbb{E}_u(A_8^\circ D_8^\circ) &= \frac{8}{33} \\
\mathbb{E}_u(D_8^{\circ 2}) &= \frac{12}{33} \\
\mathbb{E}_u(C_8^\circ D_8^\circ) &= \frac{8}{33}
\end{aligned}$$

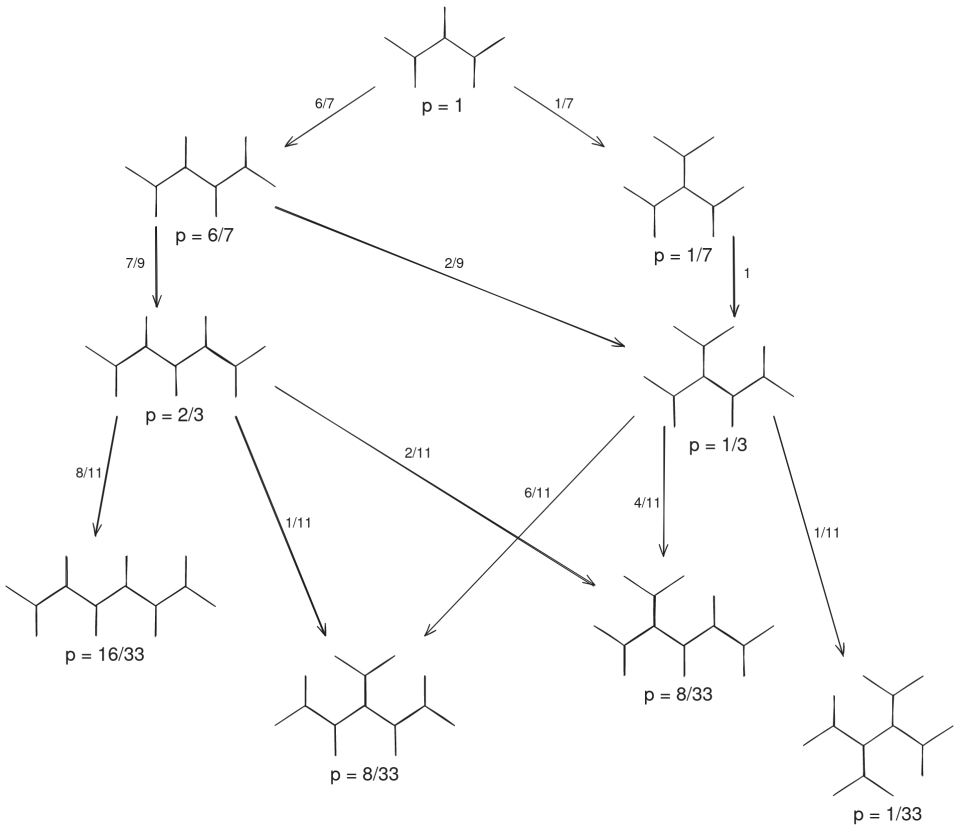

Figure 2.6: The exact probabilities of small unrooted trees being generated under the PDA model. Numbers on arrows are the probability of a split in the source tree resulting in the destination tree.

3 Moments of Unrooted Cherries and Pitchforks

In this chapter, we consider the means, variances, and covariances of two-leaved and three-leaved subtrees (cherries and pitchforks). Exact cherry and pitchfork moments are already known for rooted trees under both the YHK and PDA models [Wu and Choi, 2016]. We extend these results to unrooted trees [Choi et al., 2020].

3.1 Moments of Unrooted Cherries and Pitchforks under the YHK model

Theorem 3.1.1. *For $n \geq 6$, we have*

$$\begin{aligned}\mathbb{P}_y(A^\circ_{n+1} = a, B^\circ_{n+1} = b) = &\frac{2a}{n}\mathbb{P}_y(A^\circ_n = a, B^\circ_n = b) \\ &+ \frac{a+1}{n}\mathbb{P}_y(A^\circ_n = a+1, B^\circ_n = b-1) \\ &+ \frac{2(b-a+1)}{n}\mathbb{P}_y(A^\circ_n = a-1, B^\circ_n = b) \\ &+ \frac{n-a-2b+2}{n}\mathbb{P}_y(A^\circ_n = a, B^\circ_n = b-1).\end{aligned}$$

Proof. We begin by observing that the four edge sets E_1, E_2, E_3, E_4 as defined in Table 3.1 comprise a partition of the set of pendant edges in a tree of $n \geq 6$ leaves generated under the YHK model. If we let e be the edge that is split to grow the tree from n to $n+1$ leaves, then it follows that e must be a member of exactly one of the edge sets E_1, E_2, E_3, E_4. Thus, we

Edge set	Definition	Subtree numbers after being split	Size (YHK)	Size (PDA)
E_1	Any edge in both a cherry and a pitchfork, plus any internal edge not contained in the other three edge sets	A_n, B_n	$2A_n$	$n-3+3A_n-B_n$
E_2	Any pendant edge in a pitchfork but not in a cherry	A_n-1, B_n+1	A_n	
E_3	Any edge in a cherry but not in a pitchfork	A_n+1, B_n	$2(B_n-A_n)$	$3(B_n-A_n)$
E_4	Any pendant edge not in a cherry or pitchfork	A_n, B_n+1	$n-A_n-2B_n$	

Table 3.1: Edge sets defined under the YHK and PDA models. Note that under the YHK model we consider only pendant edges. See Figure 3.1 for an example.

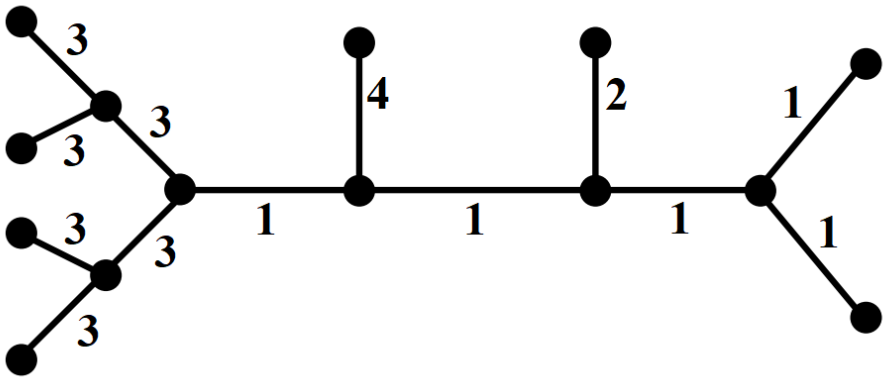

Figure 3.1: An illustration of the edge sets from Table 3.1. Edges in E_1 are labelled 1, edges in E_2 are labelled 2, and so on.

have

$$\begin{aligned}\mathbb{P}_y(A^\circ_{n+1} = a, B^\circ_{n+1} = b) = {} & \mathbb{P}_y(e \in E_1)\mathbb{P}_y(A^\circ_n = a, B^\circ_n = b) \\ & + \mathbb{P}_y(e \in E_2)\mathbb{P}_y(A^\circ_n = a+1, B^\circ_n = b-1) \\ & + \mathbb{P}_y(e \in E_3)\mathbb{P}_y(A^\circ_n = a-1, B^\circ_n = b) \\ & + \mathbb{P}_y(e \in E_4)\mathbb{P}_y(A^\circ_n = a, B^\circ_n = b-1).\end{aligned}$$

Theorem 3.1.1 then follows from the sizes of the edge sets and the fact that any pendant edge is equally likely to be split. □

Theorem 3.1.2. *Let $n \geq 6$, and let $\varphi : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ be an arbitrary function. Then we have*

$$\begin{aligned}\mathbb{E}_y\Big(\varphi(A^\circ_{n+1}, B^\circ_{n+1})\Big) = {} & \frac{2}{n}\mathbb{E}_y\Big(A^\circ_n \varphi(A^\circ_n, B^\circ_n)\Big) \\ & + \frac{1}{n}\mathbb{E}_y\Big(A^\circ_n \varphi(A^\circ_n - 1, B^\circ_n + 1)\Big) \\ & + \frac{2}{n}\mathbb{E}_y\Big((B^\circ_n - A^\circ_n)\varphi(A^\circ_n + 1, B^\circ_n)\Big) \\ & + \frac{1}{n}\mathbb{E}_y\Big((n - A^\circ_n - 2B_n)\varphi(A^\circ_n, B^\circ_n + 1)\Big).\end{aligned}$$

Proof. We define the indicator function $I_{a,b} : \mathbb{Z} \times \mathbb{Z} \to \{0, 1\}$ as

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

Multiplying both sides of Theorem 3.1.1 by our arbitrary function φ, we can rewrite it in the form

$$\begin{aligned}
&\mathbb{E}_y\Big(\varphi(A^\circ_{n+1}, B^\circ_{n+1}) I_{a,b}(A^\circ_{n+1}, B^\circ_{n+1})\Big) \\
&\quad = \frac{2}{n}\mathbb{E}_y\Big(A^\circ_n \varphi(A^\circ_n, B^\circ_n) I_{a,b}(A^\circ_n, B^\circ_n)\Big) \\
&\quad + \frac{1}{n}\mathbb{E}_y\Big(A^\circ_n \varphi(A^\circ_n - 1, B^\circ_n + 1) I_{a,b}(A^\circ_n - 1, B^\circ_n + 1)\Big) \\
&\quad + \frac{2}{n}\mathbb{E}_y\Big((B^\circ_n - A^\circ_n) \varphi(A^\circ_n + 1, B^\circ_n) I_{a,b}(A^\circ_n + 1, B^\circ_n)\Big) \\
&\quad + \frac{1}{n}\mathbb{E}_y\Big((n - A^\circ_n - 2B_n) \varphi(A^\circ_n, B^\circ_n + 1) I_{a,b}(A^\circ_n, B^\circ_n + 1)\Big).
\end{aligned}$$

If we then sum over all possible a and b we are left only with the values for which the indicator function is equal to 1, completing the proof. □

Note that Theorems 3.1.1 and 3.1.2 are identical to the corresponding results for the rooted case (see Wu and Choi [2016], Theorems 1 and 2) because the partitioning of the pendant edge set remains the same, and any pendant edge is equally likely to be split for both rooted and unrooted trees. However, the particular solutions to these equations for the unrooted cases will differ due to the initial conditions being different.

Proposition 3.1.1. *We have*

$$\begin{aligned}
\mathbb{E}_y(B^\circ_n) &= \frac{n}{3} + \frac{4}{(n-1)(n-2)}, \quad n \geq 4; \\
\mathbb{V}_y(B^\circ_n) &= \frac{2n}{45} - \frac{4(n^2 - 3n + 14)}{3(n-1)^2(n-2)^2}, \quad n \geq 5.
\end{aligned}$$

Proof. Letting $\varphi(x, y) = y$ in Theorem 3.1.2, we obtain

$$\begin{aligned}\mathbb{E}_y(B^\circ_{n+1}) =& \frac{2}{n}\mathbb{E}_y(A^\circ_n B^\circ_n) + \frac{1}{n}\mathbb{E}_y(A^\circ_n B^\circ_n + A^\circ_n) + \frac{2}{n}\mathbb{E}_y(B^{\circ 2}_n - A^\circ_n B^\circ_n)\\ &+ \frac{1}{n}\mathbb{E}_y(nB^\circ_n + n - A^\circ_n B^\circ_n - A^\circ_n - 2B^{\circ 2} - 2B^\circ_n)\\ =& \frac{1}{n}\mathbb{E}_y\Big((n-2)B^\circ_n + n\Big)\\ =& \frac{n-2}{n}\mathbb{E}_y(B^\circ_n) + 1.\end{aligned}$$

If we then multiply throughout by the summation factor $n^{\underline{2}}$ and let $\alpha(n) = (n-1)^{\underline{2}}\mathbb{E}_y(B^\circ_n)$, we obtain

$$\alpha(n) = \alpha(n-1) + (n-1)^{\underline{2}}.$$

The unique unrooted tree with four leaves contains two cherries, that is, $\mathbb{E}_y(B^\circ_4) = 2$, and therefore $\alpha(4) = 12$. We can then rearrange the above equation to give

$$\begin{aligned}\alpha(n) &= \sum_{k=5}^{n}\Big(\alpha(k) - \alpha(k-1)\Big) + \alpha(4)\\ &= \sum_{k=5}^{n}(k-1)^{\underline{2}} + 12\\ &= \frac{n^{\underline{3}} - 4^{\underline{3}}}{3} + 12\\ &= \frac{n^{\underline{3}}}{3} + 4,\end{aligned}$$

where the third equality follows from equation (2.1). Dividing throughout by $(n-1)^{\underline{2}}$ we are left with the formula for $\mathbb{E}_y(B^\circ_n)$.

For the variance, we first let $\varphi(x, y) = y^2$ in Theorem 3.1.2, yielding

$$\begin{aligned}\mathbb{E}_y(B^{\circ 2}_{n+1}) &= \frac{n-4}{n}\mathbb{E}_y(B^{\circ 2}_n) + 2\frac{n-1}{n}\mathbb{E}(B^\circ_n) + 1\\ &= \frac{n-4}{n}\mathbb{E}_y(B^{\circ 2}_n) + \frac{2n+1}{3} + \frac{8}{n(n-2)}.\end{aligned}$$

This time we multiply through by the summation factor $n^{\underline{4}}$. Letting $\beta(n) = (n-1)^{\underline{4}}\mathbb{E}_y(B_n^{\circ 2})$, we obtain

$$\begin{aligned}\beta(n) &= \beta(n-1) + \frac{(2n-1)(n-1)^{\underline{4}}}{3} + 8(n-2)(n-4) \\ &= \frac{1}{3}\sum_{k=5}^{n}(2k-1)(k-1)^{\underline{4}} + 8\sum_{k=5}^{n}(k-2)(k-4) + \beta(4).\end{aligned}$$

Note again that the unique four leaved unrooted tree has two cherries, and hence $\mathbb{E}_y(B_4^{\circ 2}) = 4$. We then have $\beta(4) = 0$, and using equation (2.2), we can express the above as

$$= \frac{(5n+2)n^{\underline{5}}}{45} + \frac{8(n-1)^{\underline{3}}}{3} - 4(n-1)^{\underline{2}} + 8.$$

Dividing through by our summation factor $n^{\underline{4}}$ we can then recover

$$\mathbb{E}_y(B_n^{\circ 2}) = \frac{(5n+2)n}{45} + \frac{4(2n-1)}{3(n-1)^{\underline{2}}},$$

from which the variance easily follows. Though Theorem 3.1.2 is only valid for $n \geq 6$, we can in fact manually verify that our formula for the mean is valid for $n \geq 4$ and our formula for the variance is valid for $n \geq 5$. □

We remark that the proof for $\mathbb{E}_y(B_n^{\circ})$ has notable similarities to that of the formula for the tetrahedral numbers (see, for instance, Baumann [2019] section 3).

Theorem 3.1.3. *We have*

$$\mathbb{E}_y(A_n^{\circ}) = \frac{n}{6} + \frac{4(2n-3)}{(n-1)^{\underline{3}}}, \quad n \geq 6;$$

and

$$Cov_y(A_n^{\circ}, B_n^{\circ}) = -\frac{n}{45} - \frac{4(n^3 - 6n^2 + 35n - 42)}{3(n-1)^{\underline{3}}(n-1)^{\underline{2}}}, \quad n \geq 6.$$

Furthermore,

$$\mathbb{V}_y(A_n^{\circ}) = \frac{23n}{420} - \frac{16(2n-3)^2}{\left((n-1)^{\underline{3}}\right)^2}, \quad n \geq 7.$$

Proof. Letting $\varphi(x, y) = x$ in theorem 3.1.2, we obtain

$$\begin{aligned}\mathbb{E}_y(A^\circ_{n+1}) &= \frac{1}{n}\mathbb{E}_y\Big((n-3)A^\circ_n + 2B^\circ_n\Big) \\ &= \frac{n-3}{n}\mathbb{E}_y(A^\circ_n) + \frac{2}{3} + \frac{8}{n^{\underline{3}}}.\end{aligned}$$

We can then multiply through by the summation factor $n^{\underline{3}}$ and define $\gamma(n) = (n-1)^{\underline{3}}\mathbb{E}_y(A^\circ_n)$ to give

$$\gamma(n+1) = \gamma(n) + \frac{2n^{\underline{3}}}{3} + 8.$$

Noting that $\mathbb{E}_y(A^\circ_6) = \frac{8}{5}$ (see Section 2.6) and hence $\gamma(6) = 192$, we then obtain the sum

$$\gamma(n) = \gamma(6) + \sum_{k=7}^{n-1} \frac{2k^{\underline{3}}}{3} + 8.$$

This can easily be solved to give the expectation above.

Next, taking $\varphi(x, y) = xy$ in Theorem 3.1.2, we obtain

$$\begin{aligned}\mathbb{E}_y(A^\circ_{n+1}B^\circ_{n+1}) &= \frac{n-5}{n}\mathbb{E}_y(A^\circ_n B^\circ_n) + \frac{n-1}{n}\mathbb{E}_y(A^\circ_n) + \frac{2}{n}\mathbb{E}_y(B^{\circ 2}_n) \\ &= \frac{n-5}{n}\mathbb{E}_y(A^\circ_n B^\circ_n) + \frac{35n-7}{90} + \frac{4(10n^2-29n+15)}{3n^{\underline{4}}}.\end{aligned}$$

Taking the summation factor $n^{\underline{5}}$ and letting $\delta(n) = (n-1)^{\underline{5}}\mathbb{E}_y(A^\circ_n B^\circ_n)$, this simplifies to

$$\delta(n+1) = \delta(n) + \frac{35n-7}{90}n^{\underline{5}} + \frac{4}{3}(n-4)(10n^2-29n+15).$$

Combined with the initial condition $\mathbb{E}_y(A^\circ_6 B^\circ_6) = \frac{16}{5}$ and hence $\delta(6) = 384$ (see Section 2.6), this can be solved by similar methods to give

$$\mathbb{E}_y(A^\circ_n B^\circ_n) = \frac{5n^4 - 27n^3 + 40n^2 + 288n - 360}{90(n-2)^{\underline{2}}},$$

from which the covariance follows.

Finally, taking $\varphi(x, y) = x^2$ in Theorem 3.1.2 gives us

$$\begin{aligned}\mathbb{E}_y(A^{\circ 2}_{n+1}) =& \frac{n-6}{n}\mathbb{E}_y(A^\circ_n) + \frac{4}{n}\mathbb{E}_y(A^\circ_n B^\circ_n) - \frac{1}{n}\mathbb{E}_y(A^\circ_n) + \frac{2}{n}\mathbb{E}_y(B^\circ_n) \\ =& \frac{n-6}{n}\mathbb{E}_y(A^\circ_n) + \frac{20n^5 - 83n^4 - 2n^3 + 1487n^2 - 2862n + 360}{90n^{\underline{4}}}\end{aligned}$$

Using the summation factor $n^{\underline{6}}$ and initial condition $\mathbb{E}_y(A^{\circ 2}_6) = \frac{16}{5}$ (see Section 2.6), this can be solved to give

$$\mathbb{E}_y(A^{\circ 2}_n) = \frac{35n^5 - 141n^4 - 29n^3 + 3909n^2 - 5454n}{1260(n-1)^{\underline{3}}},$$

from which the variance follows. □

3.2 Moments of Unrooted Cherries and Pitchforks under the PDA model

Theorem 3.2.1. *For $n \geq 6$, we have*

$$\begin{aligned}\mathbb{P}_u(A^\circ_{n+1} = a, B^\circ_{n+1} = b) =& \frac{n + 3a - b - 3}{2n-3}\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b) \\ &+ \frac{a+1}{2n-3}\mathbb{P}_u(A^\circ_n = a+1, B^\circ_n = b-1) \\ &+ \frac{3(b-a+1)}{2n-3}\mathbb{P}_u(A^\circ_n = a-1, B^\circ_n = b) \\ &+ \frac{n - a - 2b + 2}{2n-3}\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b-1).\end{aligned}$$

Proof. Similarly to the proof of Theorem 3.1.1, we begin by observing that the four edge sets E_1, E_2, E_3, E_4 as defined in Table 3.1 comprise a partition of the set of edges in a tree of $n \geq 6$ leaves generated under the PDA model. If we let e be the edge that is split to grow the tree from n to $n+1$ leaves, then it follows that e must be a member of exactly one of the edge sets

E_1, E_2, E_3, E_4. Thus, we have

$$\begin{aligned}\mathbb{P}_u(A^\circ_{n+1} = a, B^\circ_{n+1} = b) &= \mathbb{P}_u(e \in E_1)\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b)\\ &\quad + \mathbb{P}_u(e \in E_2)\mathbb{P}_u(A^\circ_n = a+1, B^\circ_n = b-1)\\ &\quad + \mathbb{P}_u(e \in E_3)\mathbb{P}_u(A^\circ_n = a-1, B^\circ_n = b)\\ &\quad + \mathbb{P}_u(e \in E_4)\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b-1).\end{aligned}$$

Theorem 3.2.1 then follows from the sizes of the edge sets and the fact that any edge is equally likely to be split.

□

Theorem 3.2.2. *Let $n \geq 6$, and let $\varphi : \mathbb{Z}\times\mathbb{Z} \to \mathbb{Z}$ be an arbitrary function. Then we have*

$$\begin{aligned}\mathbb{E}_u\Big(\varphi(A^\circ_{n+1}, B^\circ_{n+1})\Big) &= \frac{1}{2n-3}\mathbb{E}_u\Big((n + 3A^\circ_n - B^\circ_n - 3)\varphi(A^\circ_n, B^\circ_n)\Big)\\ &\quad + \frac{1}{2n-3}\mathbb{E}_u\Big(A^\circ_n\varphi(A^\circ_n - 1, B^\circ_n + 1)\Big)\\ &\quad + \frac{3}{2n-3}\mathbb{E}_u\Big((B^\circ_n - A^\circ_n)\varphi(A^\circ_n + 1, B^\circ_n)\Big)\\ &\quad + \frac{1}{2n-3}\mathbb{E}_u\Big((n - A^\circ_n - 2B_n)\varphi(A^\circ_n, B^\circ_n + 1)\Big).\end{aligned}$$

Proof. Using the same methodof proof as in Theorem 3.1.2, we multiply both sides of Theorem 3.2.1 by our arbitrary function φ, then rearrange in terms of the indicator function $I_{a,b}$. Summing over all possible a and b completes the proof. □

Theorem 3.2.3. *For $n \geq 4$, we have*

$$\mathbb{E}_u(B^\circ_n) = \frac{n^{\underline{2}}}{2(2n-5)};$$

and

$$\mathbb{V}_u(B^\circ_n) = \frac{n^{\underline{2}}(n-4)^{\underline{2}}}{2(2n-5)(2n-5)^{\underline{2}}}.$$

Proof. Substituting $\varphi(x, y) = y$ into Theorem 3.2.2 gives us

$$\mathbb{E}_u(B^\circ_{n+1}) = \frac{2n-5}{2n-3}\mathbb{E}_u(B^\circ_n) + \frac{n}{2n-3}.$$

The unique four leaved unrooted tree has two cherries, giving us the initial condition $\mathbb{E}_u(B^\circ_4) = 2$. Using this and multiplying through by the summation factor $2n-3$, we can easily solve the above recursion to give the mean.

Similarly, substituting $\varphi(x, y) = y^2$ into Theorem 3.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(B^{\circ 2}_{n+1}) &= \frac{2n-7}{2n-3}\mathbb{E}_u(B^{\circ 2}_n) + \frac{2n-2}{2n-3}\mathbb{E}_u(B^\circ_n) + \frac{n}{2n-3} \\ &= \frac{2n-7}{2n-3}\mathbb{E}_u(B^{\circ 2}_n) + \frac{n(n-1)^2}{(2n-3)^{\underline{2}}} + \frac{n}{2n-3}.\end{aligned}$$

By the same argument as above, we have $\mathbb{E}_u(B^{\circ 2}_4) = 4$. Using this initial condition and the summation factor $(2n-3)^{\underline{2}}$ we can solve the recursion to give

$$\mathbb{E}_u(B^{\circ 2}_n) = \frac{n^{\underline{2}}(n^2-n-8)}{4(2n-5)^{\underline{2}}},$$

from which the variance follows. □

Theorem 3.2.4. *For $n \geq 6$, we have*

$$\mathbb{E}_u(A^\circ_n) = \frac{n^{\underline{3}}}{2(2n-5)^{\underline{2}}},$$

$$Cov_u(A^\circ_n, B^\circ_n) = -\frac{3n^{\underline{3}}(n-5)}{2(2n-5)(2n-5)^{\underline{3}}},$$

and

$$\mathbb{V}_u(A^\circ_n) = \frac{3n^{\underline{3}}(4n^4-76n^3+527n^2-1555n+1610)}{4(2n-5)^{\underline{2}}(2n-5)^{\underline{4}}}.$$

Proof. Substituting $\varphi(x, y) = x$ into Theorem 3.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(A^\circ_{n+1}) &= \frac{2n-7}{2n-3}\mathbb{E}_u(A^\circ_n) + \frac{3}{2n-3}\mathbb{E}_u(B^\circ_n) \\ &= \frac{2n-7}{2n-3}\mathbb{E}_u(A^\circ_n) + \frac{3n^{\underline{2}}}{2(2n-3)^{\underline{2}}}.\end{aligned}$$

This recursion can be solved using the summation factor $(2n-3)^{\underline{2}}$ and initial condition $\mathbb{E}_u(A^\circ_6) = \frac{12}{7}$ (see Section 2.6) to give the mean above.

Next, substituting $\varphi(x, y) = xy$ into Theorem 3.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(A^\circ_{n+1}B^\circ_{n+1}) &= \frac{2n-9}{2n-3}\mathbb{E}_u(A^\circ_nB^\circ_n) + \frac{n-1}{2n-3}\mathbb{E}_u(A^\circ_n) + \frac{3n}{2n-3}\mathbb{E}_u(B^{\circ 2}_n) \\ &= \frac{2n-9}{2n-3}\mathbb{E}_u(A^\circ_nB^\circ_n) + \frac{n^{\underline{2}}(5n^2-9n-20)}{4(2n-3)^{\underline{3}}},\end{aligned}$$

which can be solved with the summation factor $(2n-3)^{\underline{3}}$ and initial condition $\mathbb{E}_u(A^\circ_6B^\circ_6) = \frac{24}{7}$ (see Section 2.6) to give

$$\mathbb{E}_u(A^\circ_nB^\circ_n) = \frac{n^5-6n^4+5n^3+12n^2-12n}{4(2n-3)^{\underline{3}}},$$

from which the covariance follows.

Finally, substituting $\varphi(x, y) = x^2$ into Theorem 3.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(A^{\circ 2}_{n+1}) = \frac{2n-11}{2n-3}\mathbb{E}_u(A^{\circ 2}_n) + \frac{6}{2n-3}\mathbb{E}_u(A^\circ_nB^\circ_n) - \frac{2}{2n-3}\mathbb{E}_u(A^\circ_n) \\ + \frac{3}{2n-3}\mathbb{E}_u(B^\circ_n),\end{aligned}$$

which can be solved with the summation factor $(2n-3)^{\underline{4}}$ and initial condition $\mathbb{E}_u(A^{\circ 2}_6) = \frac{24}{7}$ (see Section 2.6) to give

$$\mathbb{E}_u(A^{\circ 2}_n) = \frac{n^6-7n^5-19n^4+229n^3-480n^2+276n}{4(2n-3)^{\underline{4}}},$$

from which the variance follows. □

4 Moments of 4-subtrees

In this chapter we extend the results in the previous chapter to subtrees with up to four leaves. To this end, recall that A_n, B_n, C_n, D_n are the random variables denoting the number of pitchforks, cherries, 4-caterpillars, and crabs, respectively, in a randomly generated tree of n leaves.

4.1 Moments of 4-Subtrees Under the YHK Model

Theorem 4.1.1. *For both rooted and unrooted trees under the YHK model, we have*

$$\begin{aligned}
&\mathbb{P}_y(A_{n+1} = a, B_{n+1} = b, C_{n+1} = c, D_{n+1} = d) \\
&\quad = \frac{4(d+1)}{n}\mathbb{P}_y(A_n = a-1, B_n = b, C_n = c, D_n = d+1) \\
&\quad + \frac{2c}{n}\mathbb{P}_y(A_n = a, B_n = b, C_n = c, D_n = d) \\
&\quad + \frac{c+1}{n}\mathbb{P}_y(A_n = a+1, B_n = b-1, C_n = c+1, D_n = d-1) \\
&\quad + \frac{c+1}{n}\mathbb{P}_y(A_n = a, B_n = b-1, C_n = c+1, D_n = d) \\
&\quad + \frac{2(a-c+1)}{n}\mathbb{P}_y(A_n = a, B_n = b, C_n = c-1, D_n = d) \\
&\quad + \frac{a-c+1}{n}\mathbb{P}_y(A_n = a+1, B_n = b-1, C_n = c, D_n = d-1) \\
&\quad + \frac{2(b-2d-a+1)}{n}\mathbb{P}_y(A_n = a-1, B_n = b, C_n = c, D_n = d) \\
&\quad + \frac{n-2b-a-c+2}{n}\mathbb{P}_y(A_n = a, B_n = b-1, C_n = c, D_n = d).
\end{aligned}$$

Proof. Similarly to Theorems 3.1.1 and 3.2.1, we can partition the pendant

Edge set	Definition	Subtree numbers after being split	Size (YHK)	Size (rooted PDA)	Size (unrooted PDA)
E_1	Any edge in a crab, other than the root edge	$A_n + 1, B_n, C_n, D_n - 1$	$4D_n$	$6D_n$	
E_2	Any edge in both a cherry and a 4-caterpillar, plus any internal edge not contained in the other seven edge sets	A_n, B_n, C_n, D_n	$2C_n$	$n - 1 + 4C_n - A_n - B_n$	$n - 3 + 4C_n - A_n - B_n$
E_3	Any pendant edge in both a pitchfork and a 4-caterpillar, but not in a cherry	$A_n - 1, B_n + 1, C_n - 1, D_n + 1$	C_n		
E_4	Any pendant edge in a 4-caterpillar, but not in a pitchfork	$A_n, B_n + 1, C_n - 1, D_n$	C_n		

Edge set	Definition	Subtree numbers after being split	Size (YHK)	Size (rooted PDA)	Size (unrooted PDA)
E_5	Any internal edge in a pitchfork, plus any edge in both a cherry and a pitchfork	$A_n, B_n, C_n + 1, D_n$	$2(A_n - C_n)$	$4(A_n - C_n)$	
E_6	Any pendant edge in a pitchfork that is not also in a cherry	$A_n - 1, B_n + 1, C_n, D_n + 1$	$A_n - C_n$		
E_7	Any edge in a cherry that is not also in a pitchfork or crab	$A_n + 1, B_n, C_n, D_n$	$2(B_n - A_n - 2D_n)$	$3(B_n - A_n - 2D_n)$	
E_8	Any pendant edge that is not in a cherry or 4-caterpillar	$A_n, B_n + 1, C_n, D_n$	$n - 2B_n - A_n - C_n$		

Table 4.1: 4-subtree edge sets defined under the YHK and PDA models. Note that under the YHK model we consider only pendant edges. See Figure 4.1 for an example.

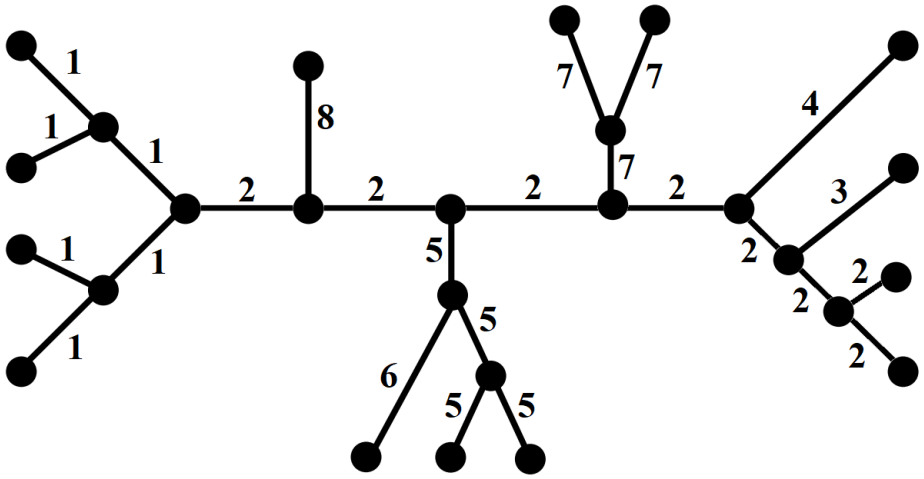

Figure 4.1: An illustration of the edge sets from Table 4.1. Edges in E_1 are labelled 1, edges in E_2 are labelled 2, and so on.

edge set into the edge sets given in Table 4.1. Theorem 4.1.1 then follows from the sizes of each pendant edge set, combined with the fact that every pendant edge is equally likely to be split. □

Theorem 4.1.2. *For both rooted and unrooted trees under the YHK model,*

we have

$$\begin{aligned}
&\mathbb{E}_y\big(\varphi(A_{n+1}, B_{n+1}, C_{n+1}, D_{n+1})\big) \\
&\quad = \frac{4}{n}\mathbb{E}_y\big(D_n\varphi(A_n+1, B_n, C_n, D_n-1)\big) \\
&\quad + \frac{2}{n}\mathbb{E}_y\big(C_n\varphi(A_n, B_n, C_n, D_n)\big) \\
&\quad + \frac{1}{n}\mathbb{E}_y\big(C_n\varphi(A_n-1, B_n+1, C_n-1, D_n+1)\big) \\
&\quad + \frac{1}{n}\mathbb{E}_y\big(C_n\varphi(A_n, B_n+1, C_n-1, D_n)\big) \\
&\quad + \frac{2}{n}\mathbb{E}_y\big((A_n-C_n)\varphi(A_n, B_n, C_n+1, D_n)\big) \\
&\quad = \frac{1}{n}\mathbb{E}_y\big((A_n-C_n)\varphi(A_n-1, B_n+1, C_n, D_n+1)\big) \\
&\quad + \frac{2}{n}\mathbb{E}_y\big((B_n-A_n-2D_n)\varphi(A_n+1, B_n, C_n, D_n)\big) \\
&\quad + \frac{1}{n}\mathbb{E}_y\big((n-2B_n-A_n-C_n)\varphi(A_n, B_n+1, C_n, D_n)\big).
\end{aligned}$$

Proof. Using the same method as Theorems 3.1.2 and 3.2.2, we multiply both sides of Theorem 4.1.1 by our arbitrary function φ, then rearrange in terms of the indicator function $I_{a,b}$. Summing over all possible a and b completes the proof. □

4.1.1 In Rooted Trees

Theorem 4.1.3. *We have*

$$\mathbb{E}_y(C_n^*) = \frac{n}{15}, \quad n \geq 5;$$

$$Cov_y(B_n^*, C_n^*) = -\frac{n}{35}, \quad n \geq 7;$$

$$Cov_y(A_n^*, C_n^*) = \frac{17n}{1260}, \quad n \geq 8;$$

and

$$\mathbb{V}_y(C_n^*) = \frac{67n}{1575}, \quad n \geq 9.$$

Proof. Substituting $\varphi(a,b,c,d) = c$ into Theorem 4.1.2 gives us

$$\begin{aligned}\mathbb{E}_y(C^*_{n+1}) &= \frac{n-4}{n}\mathbb{E}_y(C^*_n) + \frac{2}{n}\mathbb{E}_y(A^*_n)\\ &= \frac{n-4}{n}\mathbb{E}_y(C^*_n) + \frac{1}{3},\end{aligned}$$

which, combined with the summation factor $n^{\underline{4}}$ and initial condition $\mathbb{E}_y(C^*_5) = \frac{1}{3}$ (see Section 2.6) gives us $\mathbb{E}_y(C^*_n) = \frac{n}{15}$.

Next, substituting $\varphi(a,b,c,d) = bc$ into Theorem 4.1.2 gives us

$$\begin{aligned}\mathbb{E}_y(B^*_{n+1}C^*_{n+1}) &= \frac{n-6}{n}\mathbb{E}_y(B^*_nC^*_n) + \frac{n-2}{n}\mathbb{E}_y(C^*_n) + \frac{2}{n}\mathbb{E}_y(A^*_nB^*_n)\\ &= \frac{n-6}{n}\mathbb{E}_y(B^*_nC^*_n) + \frac{8(n-1)}{45}.\end{aligned}$$

Combined with the summation factor $n^{\underline{6}}$ and initial condition $\mathbb{E}_y(B^*_7C^*_7) = \frac{8}{9}$ (see Section 2.6) this gives us $\mathbb{E}_y(B^*_nC^*_n) = \frac{n^2}{45} - \frac{n}{35}$, from which the covariance follows.

Substituting $\varphi(a,b,c,d) = ac$ into Theorem 4.1.2 gives

$$\begin{aligned}\mathbb{E}_y(A^*_{n+1}C^*_{n+1}) &= \frac{n-7}{n}\mathbb{E}_y(A^*_nC^*_n) + \frac{1}{n}\mathbb{E}_y(C^*_n) + \frac{2}{n}\mathbb{E}_y(A^{*2}_n) + \frac{2}{n}\mathbb{E}_y(B^*_nC^*_n)\\ &= \frac{n-7}{n}\mathbb{E}_y(A^*_nC^*_n) + \frac{n}{10} + \frac{5}{42}.\end{aligned}$$

Combined with the summation factor $n^{\underline{7}}$ and initial condition $\mathbb{E}_y(A^*_8C^*_8) = \frac{86}{105}$ (see Section 2.6) this gives us $\mathbb{E}_y(A^*_nC^*_n) = \frac{n^2}{90} - \frac{17n}{1260}$, from which the covariance follows.

Finally, substituting $\varphi(a,b,c,d) = c^2$ into Theorem 4.1.2 gives

$$\begin{aligned}\mathbb{E}_y(C^{*2}_{n+1}) &= \frac{n-8}{n}\mathbb{E}_y(C^{*2}_n) + \frac{4}{n}\mathbb{E}_y(A^*_nC^*_n) + \frac{2}{n}\mathbb{E}_y(A^*_n)\\ &= \frac{n-8}{n}\mathbb{E}_y(C^{*2}_n) + \frac{2n}{45} + \frac{122}{315}.\end{aligned}$$

Combined with the summation factor $n^{\underline{8}}$ and initial condition $\mathbb{E}_y(C^{*2}_9) = \frac{26}{35}$ (see Section 2.6) this gives us $\mathbb{E}_y(A^*_nC^*_n) = \frac{n^2}{225} - \frac{67n}{1575}$, from which the variance follows. □

Theorem 4.1.4. *We have*

$$\mathbb{E}_y(D_n^*) = \frac{n}{30}, \quad n \geq 5;$$

$$Cov_y(B_n^*, D_n^*) = \frac{2n}{105}, \quad n \geq 7;$$

$$Cov_y(A_n^*, D_n^*) = -\frac{67n}{2520}, \quad n \geq 8;$$

$$\mathbb{V}_y(D_n^*) = \frac{43n}{1575}, \quad n \geq 9;$$

and

$$Cov_y(C_n^*, D_n^*) = -\frac{19n}{1575}, \quad n \geq 9.$$

Proof. Substituting $\varphi(a, b, c, d) = d$ into Theorem 4.1.2 gives us

$$\begin{aligned}\mathbb{E}_y(D_{n+1}^*) &= \frac{n-4}{n}\mathbb{E}_y(D_n^*) + \frac{1}{n}\mathbb{E}_y(A_n^*) \\ &= \frac{n-4}{n}\mathbb{E}_y(D_n^*) + \frac{1}{6}\end{aligned}$$

which, combined with the summation factor $n^{\underline{4}}$ and initial condition $\mathbb{E}_y(D_5^*) = \frac{1}{6}$ (see Section 2.6) gives us $\mathbb{E}_y(C_n^*) = \frac{n}{30}$.

Next, substituting $\varphi(a, b, c, d) = bd$ into Theorem 4.1.2 gives us

$$\begin{aligned}\mathbb{E}_y(B_{n+1}^* D_{n+1}^*) &= \frac{n-6}{n}\mathbb{E}_y(B_n^* D_n^*) + \frac{1}{n}\mathbb{E}_y(A_n^* B_n^*) + \frac{1}{n}(A_n^*) + \mathbb{E}_y(D_n^*) \\ &= \frac{n-6}{n}\mathbb{E}_y(B_n^* C_n^*) + \frac{4n}{45} + \frac{13}{90}.\end{aligned}$$

Combined with the summation factor $n^{\underline{6}}$ and initial condition $\mathbb{E}_y(B_7^* D_7^*) = \frac{61}{90}$ (see Section 2.6) this gives us $\mathbb{E}_y(B_n^* D_n^*) = \frac{n^2}{90} + \frac{2n}{105}$, from which the covariance follows.

Substituting $\varphi(a,b,c,d) = ad$ into Theorem 4.1.2 gives

$$\begin{aligned}\mathbb{E}_y(A^*_{n+1}D^*_{n+1}) =& \frac{n-7}{n}\mathbb{E}_y(A^*_nD^*_n) + \frac{2}{n}\mathbb{E}_y(B^*_nD^*_n) - \frac{4}{n}\mathbb{E}_y(D^*_n) + \frac{1}{n}\mathbb{E}_y(A^{*2}_n)\\ &- \frac{1}{n}\mathbb{E}_y(A^*_n)\\ =& \frac{n-7}{n}\mathbb{E}_y(A^*_nD^*_n) + \frac{n}{20} + \frac{29}{140}.\end{aligned}$$

Combined with the summation factor $n^{\underline{7}}$ and initial condition $\mathbb{E}_y(A^*_8D^*_8) = \frac{1}{7}$ (see Section 2.6) this gives us $\mathbb{E}_y(A^*_nD^*_n) = \frac{n^2}{180} - \frac{67n}{2520}$, from which the covariance follows.

Substituting $\varphi(a,b,c,d) = d^2$ into Theorem 4.1.2 gives

$$\begin{aligned}\mathbb{E}_y(D^{*2}_{n+1}) &= \frac{n-8}{n}\mathbb{E}_y(D^{*2}_n) + \frac{2}{n}\mathbb{E}_y(A^*_nD^*_n) + \frac{4}{n}\mathbb{E}_y(D^*_n) + \frac{1}{n}\mathbb{E}_y(A^*_n)\\ &= \frac{n-8}{n}\mathbb{E}_y(C^{*2}_n) + \frac{n}{90} + \frac{311}{1260}.\end{aligned}$$

Combined with the summation factor $n^{\underline{8}}$ and initial condition $\mathbb{E}_y(D^{*2}_9) = \frac{47}{140}$ (see Section 2.6) this gives us $\mathbb{E}_y(D^{*2}_n) = \frac{n^2}{900} + \frac{43n}{1575}$, from which the variance follows.

Finally, substituting $\varphi(a,b,c,d) = cd$ into Theorem 4.1.2 gives

$$\begin{aligned}\mathbb{E}_y(C^*_{n+1}D^*_{n+1}) =& \frac{n-8}{n}\mathbb{E}_y(C^*_nD^*_n) + \frac{1}{n}\mathbb{E}_y(A^*_nC^*_n) + \frac{2}{n}\mathbb{E}_y(A^*_nD^*_n) - \frac{1}{n}\mathbb{E}_y(C^*_n)\\ =& \frac{n-8}{n}\mathbb{E}_y(C^*_nD^*_n) + \frac{n}{45} - \frac{67}{630}.\end{aligned}$$

Combined with the summation factor $n^{\underline{8}}$ and initial condition $\mathbb{E}_y(C^*_9D^*_9) = \frac{1}{14}$ (see Section 2.6) this gives us $\mathbb{E}_y(C^*_nD^*_n) = \frac{n^2}{450} - \frac{19n}{1575}$, from which the covariance follows. □

Corollary 4.1.1. *We have*

$$\rho_y(B^*_n, C^*_n) = -\frac{9\sqrt{4690}}{938}, \quad n \geq 7;$$

$$\rho_y(B^*_n, D^*_n) = -\frac{3\sqrt{3010}}{301}, \quad n \geq 7;$$

$$\rho_y(A_n^*, C_n^*) = \frac{17\sqrt{23115}}{9246}, \quad n \geq 8;$$

$$\rho_y(A_n^*, D_n^*) = -\frac{67\sqrt{14835}}{11868}, \quad n \geq 8;$$

and

$$\rho_y(C_n^*, D_n^*) = -\frac{19}{\sqrt{2881}}, \quad n \geq 9.$$

Proof. These results follow easily from Theorems 4.1.3 and 4.1.4. □

4.1.2 In Unrooted Trees

Theorem 4.1.5. *We have*

$$\mathbb{E}_y(C_n^\circ) = \frac{n}{15} + \frac{8(n^2 - 4n + 6)}{(n-1)^{\underline{4}}}, \quad n \geq 8;$$

$$Cov_y(B_n^\circ, C_n^\circ) = -\frac{n}{35} - \frac{y_{bc}}{45(n-1)^{\underline{4}}(n-1)^{\underline{2}}}, \quad n \geq 9$$

where $y_{bc} = 37n^5 - 4n^4 - 895n^3 + 4606n^2 - 10080n + 10656$;

$$Cov_y(A_n^\circ, C_n^\circ) = \frac{17n}{1260} + \frac{4y_{ac}}{15(n-1)^{\underline{4}}(n-1)^{\underline{3}}}, \quad n \geq 9$$

where $y_{ac} = 4n^5 - 31n^4 - 130n^3 + 1075n^2 - 2574n + 2016$*; and*

$$\mathbb{V}_y(C_n^\circ) = \frac{67n}{1575} + \frac{y_{cc}}{1575\big((n-1)^{\underline{4}}\big)^2}, \quad n \geq 9$$

where $y_{cc} = 7n^{11} - 80n^{10} - 77n^9 + 7620n^8 - 62799n^7 + 239400n^6 - 441631n^5 + 78980n^4 + 1543892n^3 - 4048320n^2 + 5404608n - 3628800$.

Proof. Exactly the same recursions and summation factors apply as in Theorem 4.1.3, but different initial conditions (see Section 2.6).

Taking $\mathbb{E}_y(C_8^\circ) = \frac{94}{105}$ gives us

$$\mathbb{E}_y(C_n^\circ) = \frac{n}{15} + \frac{8(n^2 - 4n + 6)}{(n-1)^{\underline{4}}}.$$

Similarly, $\mathbb{E}_y(B_9^\circ C_9^\circ) = \frac{467}{210}$ gives us

$$\mathbb{E}_y(B_n^\circ C_n^\circ) = \frac{7n^6 - 79n^5 + 335n^4 + 259n^2 - 4338n^2 + 10368n - 7056}{315(n-1)^{\underline{4}}}$$

from which the covariance follows. Next, $\mathbb{E}_y(A_9^\circ C_9^\circ) = \frac{323}{210}$ gives us

$$\mathbb{E}_y(A_n^\circ C_n^\circ) = \frac{14n^6 - 123n^5 + 320n^4 + 2247n^3 - 9586n^2 + 12168n + 8064}{1260(n-1)^{\underline{4}}}.$$

from which the covariance follows. Finally, $\mathbb{E}_y(C_9^{\circ 2}) = \frac{241}{210}$ gives us

$$\mathbb{E}_y(C_n^{\circ 2}) = \frac{7n^6 - 3n^5 - 425n^4 + 3675n^3 - 4022n^2 - 11832n + 35280}{1575(n-1)^{\underline{4}}}$$

which leads to the variance, completing the proof. □

Theorem 4.1.6. *We have*

$$\mathbb{E}_y(D_n^\circ) = \frac{n}{30} + \frac{4n}{(n-1)^{\underline{3}}}, \quad n \geq 8;$$

$$Cov_y(B_n^\circ, D_n^\circ) = \frac{2n}{105} + \frac{4(3n^3 - 8n^2 - 17n + 2)}{5(n-1)^{\underline{3}}(n-1)^{\underline{2}}}, \quad n \geq 9;$$

$$Cov_y(A_n^\circ, D_n^\circ) = \frac{67n}{2520} - \frac{y_{ad}}{1260\big((n-1)^{\underline{3}}\big)^2}, \quad n \geq 9$$

where $y_{ad} = 67n^7 - 804n^6 + 3886n^5 - 5280n^4 - 14789n^3 + 88596n^2 - 100908n + 9072$*;*

$$\mathbb{V}_y(D_n^\circ) = \frac{43n}{1575} - \frac{y_{dd}}{5670\big((n-1)^{\underline{3}}\big)^2}, \quad n \geq 9$$

where $y_{dd} = 67n^7 - 804n^6 + 3886n^5 - 26280n^4 + 112723n^3 - 101076n^2 + 102204n$*; and*

$$Cov_y(C_n^\circ, D_n^\circ) = -\frac{19n}{1575} - \frac{y_{cd}}{11340(n-1)^{\underline{4}}(n-1)^{\underline{3}}}, \quad n \geq 9$$

where $y_{cd} = 151n^8 - 2416n^7 + 16006n^6 - 32584n^5 - 125801n^4 + 1145672n^3 - 3011412n^2 + 3534480n - 435456$.

Proof. Exactly the same recursions and summation factors apply as in Theorem 4.1.4, but different initial conditions (see Section 2.6).

Taking $\mathbb{E}_y(D_8^\circ) = \frac{44}{105}$ gives us

$$\mathbb{E}_y(D_n^\circ) = \frac{n}{30} + \frac{4n}{(n-1)^{\underline{3}}}.$$

Similarly, $\mathbb{E}_y(B_9^\circ D_9^\circ) = \frac{311}{210}$ gives us

$$\mathbb{E}_y(B_n^\circ D_n^\circ) = \frac{7n^5 - 30n^4 + 5n^3 + 1014n^2 + 1188n + 504}{630(n-1)^{\underline{3}}}$$

from which the covariance follows.

Next, $\mathbb{E}_y(A_8^\circ D_8^\circ) = \frac{2}{7}$ gives us

$$\mathbb{E}_y(A_n^\circ D_n^\circ) = \frac{14n^5 - 151n^4 + 556n^3 + 1531n^2 - 9342n + 3024}{2520(n-1)^{\underline{3}}}$$

from which the covariance follows.

Taking $\mathbb{E}_y(D_9^{\circ 2}) = \frac{67}{140}$ gives us

$$\mathbb{E}_y(D_n^{\circ 2}) = \frac{7n^5 + 130n^4 - 955n^3 + 3530n^2 + 17448n}{6300(n-1)^{\underline{3}}}$$

from which the variance follows, and finally $\mathbb{E}_y(C_9^\circ D_9^\circ) = \frac{31}{210}$ gives us

$$\mathbb{E}_y(C_n^\circ D_n^\circ) = \frac{7n^3 - 52n^2 + 76n + 1680}{3150(n-2)}$$

from which the covariance follows. □

4.2 Moments of 4-Subtrees Under the PDA Model

4.2.1 In Rooted Trees

Theorem 4.2.1. *For rooted trees under the PDA model, we have*

$$\begin{aligned}
&\mathbb{P}_u(A^*_{n+1} = a, B^*_{n+1} = b, C^*_{n+1} = c, D^*_{n+1} = d) \\
&\quad = \frac{6d+6}{2n-1}\mathbb{P}_u(A^*_n = a-1, B^*_n = b, C^*_n = c, D^*_n = d+1) \\
&\quad + \frac{n-1+4c-a-b}{2n-1}\mathbb{P}_u(A^*_n = a, B^*_n = b, C^*_n = c, D^*_n = d) \\
&\quad + \frac{c+1}{2n-1}\mathbb{P}_u(A^*_n = a+1, B^*_n = b-1, C^*_n = c+1, D^*_n = d-1) \\
&\quad + \frac{c+1}{2n-1}\mathbb{P}_u(A^*_n = a, B^*_n = b-1, C^*_n = c+1, D^*_n = d) \\
&\quad + \frac{4(a-c+1)}{2n-1}\mathbb{P}_u(A^*_n = a, B^*_n = b, C^*_n = c-1, D^*_n = d) \\
&\quad + \frac{a-c+1}{2n-1}\mathbb{P}_u(A^*_n = a+1, B^*_n = b-1, C^*_n = c, D^*_n = d-1) \\
&\quad + \frac{3(b-a-2d+1)}{2n-1}\mathbb{P}_u(A^*_n = a-1, B^*_n = b, C^*_n = c, D^*_n = d) \\
&\quad + \frac{n-2b-a-c+2}{2n-1}\mathbb{P}_u(A^*_n = a, B^*_n = b-1, C^*_n = c, D^*_n = d).
\end{aligned}$$

Proof. Similarly to Theorems 3.1.1, 3.2.1, and 4.1.1, we can partition the edge set into the edge sets given in Table 4.1. Theorem 4.2.1 then follows from the sizes of each edge set, combined with the fact that every edge is equally likely to be split. □

Theorem 4.2.2. *For rooted trees under the PDA model, we have*

$$\begin{aligned}
&\mathbb{E}_u\big(\varphi(A^*_{n+1}, B^*_{n+1}, C^*_{n+1}, D^*_{n+1})\big)\\
&\quad= \frac{6}{2n-1}\mathbb{E}_u\big(D^*_n\varphi(A^*_n+1, B^*_n, C^*_n, D^*_n-1)\big)\\
&\quad+ \frac{1}{2n-1}\mathbb{E}_u\big((n-1+4C^*_n-A^*_n-B^*_n)\varphi(A^*_n, B^*_n, C^*_n, D^*_n)\big)\\
&\quad+ \frac{1}{2n-1}\mathbb{E}_u\big(C^*_n\varphi(A^*_n-1, B^*_n+1, C^*_n-1, D^*_n+1)\big)\\
&\quad+ \frac{1}{2n-1}\mathbb{E}_u\big(C^*_n\varphi(A^*_n, B^*_n+1, C^*_n-1, D^*_n)\big)\\
&\quad+ \frac{4}{2n-1}\mathbb{E}_u\big((A^*_n-C^*_n)\varphi(A^*_n, B^*_n, C^*_n+1, D^*_n)\big)\\
&\quad+ \frac{1}{2n-1}\mathbb{E}_u\big((A^*_n-C^*_n)\varphi(A^*_n-1, B^*_n+1, C^*_n, D^*_n+1)\big)\\
&\quad+ \frac{3}{2n-1}\mathbb{E}_u\big((B^*_n-A^*_n-2D^*_n)\varphi(A^*_n+1, B^*_n, C^*_n, D^*_n)\big)\\
&\quad+ \frac{1}{2n-1}\mathbb{E}_u\big((n-2B^*_n-A^*_n-C^*_n)\varphi(A^*_n, B^*_n+1, C^*_n, D^*_n)\big).
\end{aligned}$$

Proof. Using the same method as Theorems 3.1.2, 3.2.2, and 4.1.2, we multiply both sides of Theorem 4.2.1 by our arbitrary function φ, then rearrange in terms of the indicator function $I_{a,b}$. Summing over all possible a and b completes the proof. □

Theorem 4.2.3. *For $n \geq 4$, we have*

$$\mathbb{E}_u(C^*_n) = \frac{n^{\underline{4}}}{2(2n-3)^{\underline{3}}};$$

$$Cov_u(B^*_n, C^*_n) = -\frac{n^{\underline{4}}(n^2-5n+3)}{2(2n-3)(2n-3)^{\underline{4}}};$$

$$Cov_u(A^*_n, C^*_n) = \frac{n^{\underline{4}}(4n^4-62n^3+340n^2-768n+585)}{2(2n-3)^{\underline{2}}(2n-3)^{\underline{5}}};$$

and

$$\mathbb{V}_u(C^*_n) = \frac{3n^{\underline{4}}u_{cc}}{2(2n-3)^{\underline{3}}(2n-3)^{\underline{6}}}$$

where $u_{cc} = 12n^6 - 296n^5 + 2937n^4 - 14934n^3 + 40971n^2 - 56290n + 30345$.

Proof. For the initial conditions used in this proof, see Section 2.6. Substituting $\varphi(a, b, c, d) = c$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(C^*_{n+1}) &= \frac{2n-7}{2n-1}\mathbb{E}_u(C^*_n) + \frac{4}{2n-1}\mathbb{E}_u(A^*_n) \\ &= \frac{2n-7}{2n-1}\mathbb{E}_u(C^*_n) + \frac{2n^{\underline{3}}}{(2n-1)^{\underline{3}}}.\end{aligned}$$

Using the summation factor $(2n-1)^{\underline{3}}$ and the initial condition $\mathbb{E}_u(C^*_4) = \frac{4}{5}$ this can easily be solved to give $\mathbb{E}_u(C^*_n) = \frac{n^{\underline{4}}}{2(2n-3)^{\underline{3}}}$.

Similarly, substituting $\varphi(a, b, c, d) = bc$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(B^*_{n+1}C^*_{n+1}) &= \frac{2n-9}{2n-1}\mathbb{E}_u(B^*_nC^*_n) + \frac{n-2}{2n-1}\mathbb{E}_u(C^*_n) + \frac{4}{2n-1}\mathbb{E}_u(A^*_nB^*_n) \\ &= \frac{2n-9}{2n-1}\mathbb{E}_u(B^*_nC^*_n) + \frac{n^{\underline{3}}(3n^2-11n+2)}{2(2n-1)^{\underline{4}}}\end{aligned}$$

which can be solved with the summation factor $(2n-1)^{\underline{4}}$ and the initial condition $\mathbb{E}_u(B^*_4C^*_4) = \frac{4}{5}$ to give

$$\mathbb{E}_u(B^*_nC^*_n) = \frac{n^{\underline{4}}(n^2-5n+2)}{4(2n-3)^{\underline{4}}}$$

from which the covariance follows.

Next, substituting $\varphi(a, b, c, d) = ac$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(A^*_{n+1}C^*_{n+1}) =& \frac{2n-11}{2n-1}\mathbb{E}_u(A^*_nC^*_n) + \frac{1}{2n-1}\mathbb{E}_u(C^*_n) + \frac{4}{2n-1}\mathbb{E}_u(A^{*2}_n) \\ &+ \frac{3}{2n-1}\mathbb{E}_u(B^*_nC^*_n) \\ =& \frac{2n-11}{2n-1}\mathbb{E}_u(A^*_nC^*_n) + \frac{n^{\underline{3}}(7n^3-36n^2-47n+300)}{4(2n-1)^{\underline{5}}}\end{aligned}$$

Using the summation factor $(2n-1)^{\underline{5}}$ and initial condition $\mathbb{E}_u(A^*_4C^*_4) = \frac{4}{5}$ this can be solved to give

$$\mathbb{E}_u(A_n^* C_n^*) = \frac{n^{\underline{4}}(n^3 - 7n^2 - 6n + 78)}{4(2n-3)^{\underline{5}}}$$

from which the covariance follows.

Finally, substituting $\varphi(a, b, c, d) = c^2$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(C_{n+1}^{*2}) =& \frac{2n-13}{2n-1}\mathbb{E}_u(C_n^{*2}) + \frac{8}{2n-1}\mathbb{E}_u(A_n^* C_n^*) - \frac{2}{2n-1}\mathbb{E}_u(C_n^*) \\ &+ \frac{4}{2n-1}\mathbb{E}_u(A_n^*) \\ =& \frac{2n-13}{2n-1}\mathbb{E}_u(C_n^{*2}) + \frac{n^{\underline{3}}(2n^4 - 8n^3 - 134n^2 + 929n - 1557)}{(2n-1)^{\underline{6}}}.\end{aligned}$$

Using the summation factor $(2n-1)^{\underline{6}}$ and initial condition $\mathbb{E}_u(C_4^{*2}) = \frac{4}{5}$ this can be solved to give

$$\mathbb{E}_u(C_n^{*2}) = \frac{n^{\underline{4}}(n^4 - 6n^3 - 85n^2 + 798n - 1734)}{4(2n-3)^{\underline{6}}}$$

from which the variance follows. □

Theorem 4.2.4. *For $n \geq 4$, we have*

$$\mathbb{E}_u(D_n^*) = \frac{n^{\underline{4}}}{8(2n-3)^{\underline{3}}};$$

$$Cov_u(B_n^*, D_n^*) = \frac{n^{\underline{4}}(2n^3 - 35n + 48)}{16(2n-3)(2n-3)^{\underline{4}}};$$

$$Cov_u(A_n^*, D_n^*) = -\frac{3n^{\underline{4}}(2n^4 - 27n^3 + 126n^2 - 236n + 150)}{4(2n-3)^{\underline{2}}(2n-3)^{\underline{5}}};$$

$$\mathbb{V}_u(D_n^*) = \frac{3n^{\underline{4}}u_{dd}}{32(2n-3)^{\underline{3}}(2n-3)^{\underline{6}}}$$

where $u_{dd} = 76n^6 - 1832n^5 + 17737n^4 - 87894n^3 + 233827n^2 - 314434n + 165480$*; and*

$$Cov_u(C_n^*, D_n^*) = -\frac{n^{\underline{4}} u_{cd}}{8(2n-3)^{\underline{3}}(2n-3)^{\underline{6}}},$$

where $u_{cd} = 28n^6 - 648n^5 + 5989n^4 - 28158n^3 + 70663n^2 - 89274n + 44100$.

Proof. For the initial conditions used in this proof, see Section 2.6. Substituting $\varphi(a,b,c,d) = d$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(D_{n+1}^*) &= \frac{2n-7}{2n-1}\mathbb{E}_u(D_n^*) + \frac{1}{2n-1}\mathbb{E}_u(A_n^*) \\ &= \frac{2n-7}{2n-1}\mathbb{E}_u(D_n^*) + \frac{n^{\underline{3}}}{2(2n-1)^{\underline{3}}}.\end{aligned}$$

Using the summation factor $(2n-1)^{\underline{3}}$ and initial condition $\mathbb{E}_u(D_4^*) = \frac{1}{5}$ this can be solved to give $\mathbb{E}_u(D_n^*) = \frac{n^{\underline{4}}}{8(2n-3)^{\underline{3}}}$.

Next, substituting $\varphi(a,b,c,d) = bd$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(B_{n+1}^* D_{n+1}^*) =& \frac{2n-9}{2n-1}\mathbb{E}_u(B_n^* D_n^*) + \frac{n}{2n-1}\mathbb{E}_u(D_n^*) + \frac{1}{2n-1}\mathbb{E}_u(A_n^*) \\ &+ \frac{1}{2n-1}\mathbb{E}_u(A_n^* B_n^*) \\ =& \frac{2n-9}{2n-1}\mathbb{E}_u(B_n^* D_n^*) + \frac{n^{\underline{3}}(3n^2 - n - 32)}{8(2n-1)^{\underline{4}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-1)^{\underline{4}}$ and initial condition $\mathbb{E}_u(B_4^* D_4^*) = \frac{2}{5}$ to give

$$\mathbb{E}_u(B_n^* D_n^*) = \frac{n^{\underline{4}}(n^2 - n - 16)}{16(2n-3)^{\underline{4}}}$$

from which the covariance follows.

Substituting $\varphi(a,b,c,d) = ad$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(A_{n+1}^* D_{n+1}^*) =& \frac{2n-11}{2n-1}\mathbb{E}_u(A_n^* D_n^*) - \frac{6}{2n-1}\mathbb{E}_u(D_n^*) + \frac{1}{2n-1}\mathbb{E}_u(A_n^{*2}) \\ &- \frac{1}{2n-1}\mathbb{E}_u(A_n^*) + \frac{3}{2n-1}\mathbb{E}_u(B_n^* D_n^*) \\ =& \frac{2n-11}{2n-1}\mathbb{E}_u(A_n^* D_n^*) + \frac{7n^{\underline{6}}}{16(2n-1)^{\underline{5}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-1)^{\underline{5}}$ and initial condition $\mathbb{E}_u(A_4^*D_4^*) = 0$ to give

$$\mathbb{E}_u(A_n^*D_n^*) = \frac{n^{\underline{7}}}{16(2n-3)^{\underline{5}}}.$$

Now substituting $\varphi(a,b,c,d) = d^2$ into Theorem 4.2.2 gives us

$$\begin{aligned}\mathbb{E}_u(D_{n+1}^{*2}) =& \frac{2n-13}{2n-1}\mathbb{E}_u(D_n^{*2}) + \frac{2}{2n-1}\mathbb{E}_u(A_n^*D_n^*) + \frac{6}{2n-1}\mathbb{E}_u(D_n^*) \\ &+ \frac{1}{2n-1}\mathbb{E}_u(A_n^*) \\ =& \frac{2n-13}{2n-1}\mathbb{E}_u(D_n^{*2}) + \frac{n^{\underline{3}}(n^4+38n^3-625n^2+2884n-4194)}{8(2n-1)^{\underline{6}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-1)^{\underline{6}}$ and initial condition $\mathbb{E}_u(D_4^{*2}) = \frac{1}{5}$ to give

$$\mathbb{E}_u(D_n^{*2}) = \frac{n^{\underline{4}}(n^4+42n^3-877n^2+5106n-9456)}{64(2n-3)^{\underline{6}}}$$

from which the variance follows.

Finally, substituting $\varphi(a,b,c,d) = cd$ into Theorem 4.2.2 gives

$$\begin{aligned}\mathbb{E}_u(C_{n+1}^*D_{n+1}^*) =& \frac{2n-13}{2n-1}\mathbb{E}_u(C_n^*D_n^*) + \frac{1}{2n-1}\mathbb{E}_u(A_n^*C_n^*) \\ &+ \frac{4}{2n-1}\mathbb{E}_u(A_n^*D_n^*) - \frac{1}{2n-1}\mathbb{E}_u(C_n^*) \\ =& \frac{2n-13}{2n-1}\mathbb{E}_u(C_n^*D_n^*) + \frac{n^{\underline{7}}}{2(2n-1)^{\underline{6}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-1)^{\underline{6}}$ and initial condition $\mathbb{E}_u(C_4^*D_4^*) = 0$ to give

$$\mathbb{E}_u(C_n^*D_n^*) = \frac{n^{\underline{8}}}{16(2n-3)^{\underline{6}}}$$

from which the covariance follows. □

4.2.2 In Unrooted Trees

Theorem 4.2.5. *For unrooted trees under the PDA model, we have*

$$\begin{aligned}
&\mathbb{P}_u(A^\circ_{n+1} = a, B^\circ_{n+1} = b, C^\circ_{n+1} = c, D^\circ_{n+1} = d) \\
&\quad = \frac{6d+6}{2n-3}\mathbb{P}_u(A^\circ_n = a-1, B^\circ_n = b, C^\circ_n = c, D^\circ_n = d+1) \\
&\quad + \frac{n-3+4c-a-b}{2n-3}\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b, C^\circ_n = c, D^\circ_n = d) \\
&\quad + \frac{c+1}{2n-3}\mathbb{P}_u(A^\circ_n = a+1, B^\circ_n = b-1, C^\circ_n = c+1, D^\circ_n = d-1) \\
&\quad + \frac{c+1}{2n-3}\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b-1, C^\circ_n = c+1, D^\circ_n = d) \\
&\quad + \frac{4(a-c+1)}{2n-3}\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b, C^\circ_n = c-1, D^\circ_n = d) \\
&\quad + \frac{a-c+1}{2n-3}\mathbb{P}_u(A^\circ_n = a+1, B^\circ_n = b-1, C^\circ_n = c, D^\circ_n = d-1) \\
&\quad + \frac{3(b-a-2d+1)}{2n-3}\mathbb{P}_u(A^\circ_n = a-1, B^\circ_n = b, C^\circ_n = c, D^\circ_n = d) \\
&\quad + \frac{n-2b-a-c+2}{2n-3}\mathbb{P}_u(A^\circ_n = a, B^\circ_n = b-1, C^\circ_n = c, D^\circ_n = d).
\end{aligned}$$

Proof. Similarly to Theorems 3.1.1, 3.2.1, 4.1.1, and 4.2.1, we can partition the edge set into the edge sets given in Table 4.1. Theorem 4.2.5 then follows from the sizes of each edge set, combined with the fact that every edge is equally likely to be split. □

Theorem 4.2.6. *For unrooted trees under the PDA model, we have*

$$\begin{aligned}
&\mathbb{E}_u\big(\varphi(A^\circ_{n+1}, B^\circ_{n+1}, C^\circ_{n+1}, D^\circ_{n+1})\big)\\
&\quad= \frac{6}{2n-3}\mathbb{E}_u\big(D^\circ_n\varphi(A^\circ_n+1, B^\circ_n, C^\circ_n, D^\circ_n-1)\big)\\
&\quad+ \frac{1}{2n-3}\mathbb{E}_u\big((n-3+4C^\circ_n-A^\circ_n-B^\circ_n)\varphi(A^\circ_n, B^\circ_n, C^\circ_n, D^\circ_n)\big)\\
&\quad+ \frac{1}{2n-3}\mathbb{E}_u\big(C^\circ_n\varphi(A^\circ_n-1, B^\circ_n+1, C^\circ_n-1, D^\circ_n+1)\big)\\
&\quad+ \frac{1}{2n-3}\mathbb{E}_u\big(C^\circ_n\varphi(A^\circ_n, B^\circ_n+1, C^\circ_n-1, D^\circ_n)\big)\\
&\quad+ \frac{4}{2n-3}\mathbb{E}_u\big((A^\circ_n-C^\circ_n)\varphi(A^\circ_n, B^\circ_n, C^\circ_n+1, D^\circ_n)\big)\\
&\quad+ \frac{1}{2n-3}\mathbb{E}_u\big((A^\circ_n-C^\circ_n)\varphi(A^\circ_n-1, B^\circ_n+1, C^\circ_n, D^\circ_n+1)\big)\\
&\quad+ \frac{3}{2n-3}\mathbb{E}_u\big((B^\circ_n-A^\circ_n-2D^\circ_n)\varphi(A^\circ_n+1, B^\circ_n, C^\circ_n, D^\circ_n)\big)\\
&\quad+ \frac{1}{2n-3}\mathbb{E}_u\big((n-2B^\circ_n-A^\circ_n-C^\circ_n)\varphi(A^\circ_n, B^\circ_n+1, C^\circ_n, D^\circ_n)\big).
\end{aligned}$$

Proof. Using the same method as Theorems 3.1.2, 3.2.2, 4.1.2, and 4.2.2, we multiply both sides of Theorem 4.2.5 by our arbitrary function φ, then rearrange in terms of the indicator function $I_{a,b}$. Summing over all possible a and b completes the proof. □

Theorem 4.2.7. *For $n \geq 8$ we have*

$$\mathbb{E}_u(C^\circ_n) = \frac{n^{\underline{4}}}{2(2n-5)^{\underline{3}}};$$

$$Cov_u(B^\circ_n C^\circ_n) = -\frac{n^{\underline{4}}(n^2-5n-5)}{2(2n-5)(2n-5)^{\underline{4}}};$$

$$Cov_u(A^\circ_n C^\circ_n) = \frac{n^{\underline{4}}(4n^4-90n^3+736n^2-2520n+2905)}{2(2n-5)^{\underline{2}}(2n-5)^{\underline{5}}};$$

and

$$\mathbb{V}_u(C^\circ_n) = \frac{3n^{\underline{4}}u_{cc}}{2(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}}$$

where $u_{cc} = 12n^6 - 384n^5 + 5013n^4 - 34006n^3 + 125715n^2 - 238730n + 181125$.

Proof. For the initial conditions used in this proof, see Section 2.6. Substituting $\varphi(a, b, c, d) = c$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(C^\circ_{n+1}) &= \frac{2n-9}{2n-3}\mathbb{E}_u(C^\circ_n) + \frac{4}{2n-3}\mathbb{E}_u(A^\circ_n) \\ &= \frac{2n-9}{2n-3}\mathbb{E}_u(C^\circ_n) + \frac{2n^{\underline{3}}}{(2n-3)^{\underline{3}}}.\end{aligned}$$

Using the summation factor $(2n-3)^{\underline{3}}$ and the initial condition $\mathbb{E}_u(C^\circ_8) = \frac{10}{33}$ this can easily be solved to give $\mathbb{E}_u(C^\circ_n) = \frac{n^{\underline{4}}}{2(2n-5)^{\underline{3}}}$.

Similarly, substituting $\varphi(a, b, c, d) = bc$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(B^\circ_{n+1}C^\circ_{n+1}) &= \frac{2n-11}{2n-3}\mathbb{E}_u(B^\circ_n C^\circ_n) + \frac{n-2}{2n-3}\mathbb{E}_u(C^\circ_n) + \frac{4}{2n-3}\mathbb{E}_u(A^\circ_n B^\circ_n) \\ &= \frac{2n-11}{2n-3}\mathbb{E}_u(B^\circ_n C^\circ_n) + \frac{n^{\underline{3}}(3n^2-11n-6)}{2(2n-3)^{\underline{4}}}\end{aligned}$$

which can be solved with the summation factor $(2n-3)^{\underline{4}}$ and the initial condition $\mathbb{E}_u(B^\circ_8 C^\circ_8) = \frac{8}{3}$ to give

$$\mathbb{E}_u(B^\circ_n C^\circ_n) = \frac{n^{\underline{4}}(n^2-5n-2)}{4(2n-5)^{\underline{4}}}$$

from which the covariance follows. Next, substituting $\varphi(a, b, c, d) = ac$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(A^\circ_{n+1}C^\circ_{n+1}) =& \frac{2n-13}{2n-3}\mathbb{E}_u(A^\circ_n C^\circ_n) + \frac{1}{2n-3}\mathbb{E}_u(C^\circ_n) + \frac{4}{2n-3}\mathbb{E}_u(A^{\circ 2}_n) \\ &+ \frac{3}{2n-3}\mathbb{E}_u(B^\circ_n C^\circ_n) \\ =& \frac{2n-13}{2n-3}\mathbb{E}_u(A^\circ_n C^\circ_n) + \frac{n^{\underline{3}}(n-4)(7n^2-8n)}{4(2n-3)^{\underline{5}}}.\end{aligned}$$

Using the summation factor $(2n-3)^{\underline{5}}$ and initial condition $\mathbb{E}_u(A^\circ_8 C^\circ_8) = \frac{24}{11}$ this can be solved to give

$$\mathbb{E}_u(A_n^\circ C_n^\circ) = \frac{n^{\underline{4}}(n^3 - 7n^2 - 22n + 166)}{4(2n-5)^{\underline{5}}}$$

from which the covariance follows.

Finally, substituting $\varphi(a, b, c, d) = c^2$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(C_{n+1}^{\circ 2}) =& \frac{2n-15}{2n-3}\mathbb{E}_u(C_n^{\circ 2}) + \frac{8}{2n-3}\mathbb{E}_u(A_n^\circ C_n^\circ) - \frac{2}{2n-3}\mathbb{E}_u(C_n^\circ) \\ &+ \frac{4}{2n-3}\mathbb{E}_u(A_n^\circ) \\ =& \frac{2n-15}{2n-3}\mathbb{E}_u(C_n^{\circ 2}) + \frac{n^{\underline{3}}(2n^4 - 8n^3 - 206n^2 + 1613n - 3141)}{(2n-3)^{\underline{6}}}.\end{aligned}$$

Using the summation factor $(2n-3)^{\underline{6}}$ and initial condition $\mathbb{E}_u(C_8^{\circ 2}) = \frac{72}{33}$ this can be solved to give

$$\mathbb{E}_u(C_n^{\circ 2}) = \frac{n^{\underline{4}}(n^4 - 6n^3 - 133n^2 + 1374n - 3450)}{4(2n-5)^{\underline{6}}}$$

from which the variance follows. □

Theorem 4.2.8. *For $n \geq 8$, we have*

$$\mathbb{E}_u(D_n^\circ) = \frac{n^{\underline{4}}}{8(2n-5)^{\underline{3}}};$$

$$Cov_u(B_n^\circ D_n^\circ) = \frac{3n^{\underline{6}}}{8(2n-5)(2n-5)^{\underline{4}}};$$

$$Cov_u(A_n^\circ D_n^\circ) = -\frac{3n^{\underline{4}}(2n^4 - 33n^3 + 188n^2 - 432n + 350)}{4(2n-5)^{\underline{2}}(2n-5)^{\underline{5}}};$$

$$\mathbb{V}_u(D_n^\circ) = \frac{3n^{\underline{4}}u_{dd}}{32(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}},$$

where $u_{dd} = 76n^6 - 2304n^5 + 28453n^4 - 182806n^3 + 642751n^2 - 1169090n + 856800$; and

$$Cov_u(C_n^\circ D_n^\circ) = \frac{n^{\underline{4}} u_{cd}}{8(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}},$$

where $u_{cd} = 28n^6 - 768n^5 + 8401n^4 - 46782n^3 + 139891n^2 - 214170n + 132300$.

Proof. For the initial conditions used in this proof, see Section 2.6. Substituting $\varphi(a,b,c,d) = d$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(D_{n+1}^\circ) &= \frac{2n-9}{2n-3}\mathbb{E}_u(D_n^\circ) + \frac{1}{2n-3}\mathbb{E}_u(A_n^\circ) \\ &= \frac{2n-9}{2n-3}\mathbb{E}_u(D_n^\circ) + \frac{n^{\underline{3}}}{2(2n-3)^{\underline{3}}}.\end{aligned}$$

Using the summation factor $(2n-3)^{\underline{3}}$ and initial condition $\mathbb{E}_u(D_8^\circ) = \frac{10}{33}$ this can be solved to give $\mathbb{E}_u(D_n^\circ) = \frac{n^{\underline{4}}}{8(2n-5)^{\underline{3}}}$.

Next, substituting $\varphi(a,b,c,d) = bd$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(B_{n+1}^\circ D_{n+1}^\circ) =& \frac{2n-11}{2n-3}\mathbb{E}_u(B_n^\circ D_n^\circ) + \frac{n}{2n-3}\mathbb{E}_u(D_n^\circ) + \frac{1}{2n-3}\mathbb{E}_u(A_n^\circ) \\ &+ \frac{1}{2n-3}\mathbb{E}_u(A_n^\circ B_n^\circ) \\ =& \frac{2n-11}{2n-3}\mathbb{E}_u(B_n^\circ D_n^\circ) + \frac{n^{\underline{3}}(3n^2-n-48)}{8(2n-3)^{\underline{4}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-3)^{\underline{4}}$ and initial condition $\mathbb{E}_u(B_8^\circ D_8^\circ) = \frac{32}{33}$ to give

$$\mathbb{E}_u(B_n^\circ D_n^\circ) = \frac{n^{\underline{4}}(n^2-n-24)}{16(2n-5)^{\underline{4}}}$$

from which the covariance follows.

Substituting $\varphi(a,b,c,d) = ad$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(A_{n+1}^\circ D_{n+1}^\circ) =& \frac{2n-13}{2n-3}\mathbb{E}_u(A_n^\circ D_n^\circ) - \frac{6}{2n-3}\mathbb{E}_u(D_n^\circ) + \frac{1}{2n-3}\mathbb{E}_u(A_n^{\circ 2}) \\ &- \frac{1}{2n-3}\mathbb{E}_u(A_n^\circ) + \frac{3}{2n-3}\mathbb{E}_u(B_n^\circ D_n^\circ) \\ =& \frac{2n-13}{2n-3}\mathbb{E}_u(A_n^\circ D_n^\circ) + \frac{7n^{\underline{6}}}{16(2n-3)^{\underline{5}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-3)^{\underline{5}}$ and initial condition $\mathbb{E}_u(A_8^\circ D_8^\circ) = \frac{8}{33}$ to give

$$\mathbb{E}_u(A_n^\circ D_n^\circ) = \frac{n^{\underline{7}}}{16(2n-5)^{\underline{5}}}.$$

Now substituting $\varphi(a,b,c,d) = d^2$ into Theorem 4.2.6 gives us

$$\begin{aligned}\mathbb{E}_u(D_{n+1}^{\circ 2}) =& \frac{2n-15}{2n-3}\mathbb{E}_u(D_n^{\circ 2}) + \frac{2}{2n-3}\mathbb{E}_u(A_n^\circ D_n^\circ) + \frac{6}{2n-3}\mathbb{E}_u(D_n^\circ) \\ &+ \frac{1}{2n-3}\mathbb{E}_u(A_n^\circ) \\ =& \frac{2n-15}{2n-3}\mathbb{E}_u(D_n^{\circ 2}) + \frac{n^{\underline{3}}(n^4+38n^3-769n^2+4252n-7362)}{8(2n-3)^{\underline{6}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-3)^{\underline{6}}$ and initial condition $\mathbb{E}_u(D_8^{\circ 2}) = \frac{12}{33}$ to give

$$\mathbb{E}_u(D_n^{\circ 2}) = \frac{n^{\underline{4}}(n^4+42n^3-1069n^2+7410n-16320)}{64(2n-5)^{\underline{6}}}$$

from which the variance follows.

Finally, substituting $\varphi(a,b,c,d) = cd$ into Theorem 4.2.6 gives

$$\begin{aligned}\mathbb{E}_u(C_{n+1}^\circ D_{n+1}^\circ) =& \frac{2n-15}{2n-3}\mathbb{E}_u(C_n^\circ D_n^\circ) + \frac{1}{2n-3}\mathbb{E}_u(A_n^\circ C_n^\circ) \\ &+ \frac{4}{2n-3}\mathbb{E}_u(A_n^\circ D_n^\circ) - \frac{1}{2n-3}\mathbb{E}_u(C_n^\circ) \\ =& \frac{2n-15}{2n-3}\mathbb{E}_u(C_n^\circ D_n^\circ) + \frac{n^{\underline{7}}}{2(2n-3)^{\underline{6}}}.\end{aligned}$$

This can be solved using the summation factor $(2n-3)^{\underline{6}}$ and initial condition $\mathbb{E}_u(C_8^\circ D_8^\circ) = \frac{8}{33}$ to give

$$\mathbb{E}_u(C_n^\circ D_n^\circ) = \frac{n^{\underline{8}}}{16(2n-5)^{\underline{6}}}$$

from which the covariance follows. □

5 Moments of General Subtrees

In this chapter, we consider moments of subtrees of arbitrary size. First we consider the number of k-leaved caterpillars, $C_{k,n}$, then the number of k-leaved subtrees, $S_{k,n}$. Additionally, we introduce the rootedness parameter, r, defined

$$r = \begin{cases} 1, & \text{if the tree is rooted,} \\ 0, & \text{otherwise.} \end{cases}$$

5.1 Moments of k-Caterpillars

5.1.1 Moments of k-Caterpillars under the YHK Model

Theorem 5.1.1. *Under the YHK model, for $k \geq 2$ and $n > 2k$, the conditional probability that the number of k-leaved caterpillars in a tree of n leaves is equal to some constant c is given by*

$$\mathbb{P}_y(C_{k,n+1} = c \mid C_{k,n} = a, C_{k-1,n} = b) = \begin{cases} \frac{2(b-a)}{n}, & \textit{when } a = c-1, \\ \frac{(k-2)a}{n}, & \textit{when } a = c+1, \\ \frac{(n-2)b+(4-k)a}{n}, & \textit{when } a = c, \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. Let T_n be a random n-leaved tree containing a and b k-leaved caterpillars and $(k-1)$-leaved caterpillars respectively. Then, suppose we apply one step of the YHK process, splitting a random pendant edge e to give T_{n+1}. Partitioning the pendant edge set in T_n as in Table 5.1, we obtain the

Edge set	Definition	$C_{k,n+1}$ after being split	Size (YHK)	Size (PDA)
E_1	Any edge that is either internal or part of a cherry, and is inside a $k-1$-legged caterpillar but not in a k-legged caterpillar	$C_{k,n}+1$	$2(C_{k-1,n}-C_{k,n})$	$k(C_{k-1,n}-C_{k,n})$
E_2	Any pendant edge that is in a k-legged caterpillar and not also in a cherry	$C_{k,n}-1$	$(k-2)C_{k,n}$	
E_3	All other edges	$C_{k,n}$	$n-2C_{k-1,n}+(4-k)C_{k,n}$	$2n-3+2r-kC_{k-1,n}+2C_{k,n}$

Table 5.1: Caterpillar related edge sets defined under the YHK and PDA models. Note that under the YHK model we consider only pendant edges. See Figure 5.1 for an example.

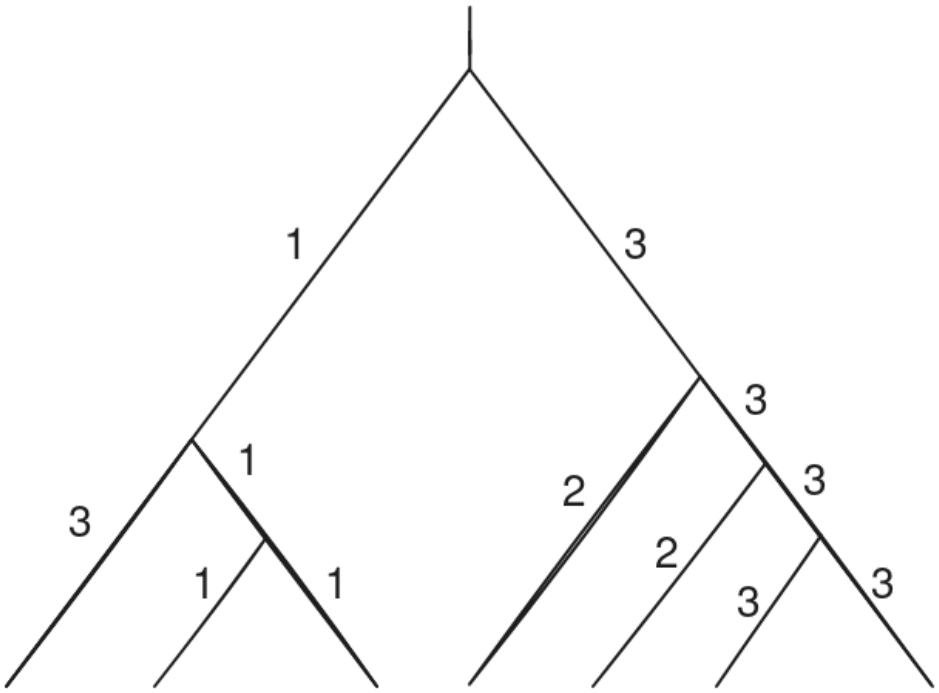

Figure 5.1: An illustration of the edge sets from Table 5.1, for the case $k = 4$. Edges in E_1 are labelled 1, edges in E_2 are labelled 2, and so on.

following cases:

1. $e \in E_1$: the split occurs in a $(k-1)$-caterpillar such that it induces a new k-caterpillar, hence $a = c - 1$. This occurs with conditional probability $\frac{2(b-a)}{n}$.

2. $e \in E_2$: the split occurs in a k-caterpillar such that it destroys an existing k-caterpillar, hence $a = c + 1$. This occurs with conditional probability $\frac{(k-2)a}{n}$.

3. $e \in E_3$: the split occurs elsewhere and does not alter the number of k-caterpillars. This occurs with conditional probability $\frac{(n-2)b+(4-k)a}{n}$.

□

Theorem 5.1.2. *Let $\psi : \mathbb{Z} \to \mathbb{Z}$ be an arbitrary function. Then, for $k \geq 3$ and $n > k$, we have*

$$\begin{aligned}\mathbb{E}_y\big(\psi(C_{k,n+1})\big) = \frac{2}{n}\mathbb{E}_y\big((C_{k-1,n} - C_{k,n})\psi(1 + C_{k,n})\big) \\ + \frac{k-2}{n}\mathbb{E}_y\big(C_{k,n}\psi(-1 + C_{k,n})\big) \\ + \frac{1}{n}\mathbb{E}_y\big((n - 2C_{k-1,n} + (4-k)C_{k,n})\psi(C_{k,n})\big).\end{aligned}$$

Proof. As before, the expectation can be found by multiplying our arbitrary function ψ by the probability, then summing over all possible cases.

$$\mathbb{E}_y\big(\psi(C_{k,n+1})\big) = \sum_c \psi(c)\mathbb{P}_y(C_{k,n+1} = c).$$

By definition of conditional probability, this is equal to

$$= \sum_c \sum_a \sum_b \psi(c)\mathbb{P}_y(C_{k,n+1} = c \mid C_{k,n} = a, C_{k-1,n} = b)\mathbb{P}_y(C_{k,n} = a, C_{k-1,n} = b).$$

As a, b, and c vary independently of one another, we can rearrange the order

of summation.

$$= \sum_a \sum_b \sum_c \psi(c) \mathbb{P}_y(C_{k,n+1} = c \mid C_{k,n} = a, C_{k-1,n} = b) \mathbb{P}_y(C_{k,n} = a, C_{k-1,n} = b)$$

$$= \sum_a \sum_b \mathbb{P}_y(C_{k,n} = a, C_{k-1,n} = b) \left(\sum_c \psi(c) \mathbb{P}_y(C_{k,n+1} = c \mid C_{k,n} = a, C_{k-1,n} = b) \right).$$

Substituting in Theorem 5.1.1, we obtain

$$\begin{aligned}
&= \sum_a \sum_b \mathbb{P}_y(C_{k,n} = a, C_{k-1,n} = b) \frac{1}{n} \Big(2(b-a)\psi(a+1) + (k-2)a\psi(a-1) \\
&\quad + \big(n - 2b + (4-k)a\big)\psi(a) \Big) \\
&= \frac{2}{n} \mathbb{E}_y\big((C_{k-1,n} - C_{k,n})\psi(1 + C_{k,n})\big) + \frac{k-2}{n} \mathbb{E}_y\big(C_{k,n}\psi(-1 + C_{k,n})\big) \\
&\quad + \frac{1}{n} \mathbb{E}_y\big((n - 2C_{k-1,n} + (4-k)C_{k,n})\psi(C_{k,n})\big)
\end{aligned}$$

which completes the proof. □

Theorem 5.1.3. *Rosenberg [2006, Corollary 5.2] The expected number of k-caterpillars in an unrooted tree generated under the YHK model is given by*

$$\mathbb{E}_y(C^*_{k,n}) = \frac{2^{k-1} n}{(k+1)!}.$$

for $n \geq 3$ and $2 \leq k < n$.

Proof. We present an alternative proof. First, we take $\psi(x) = x$ in theorem 5.1.2, giving us

$$\mathbb{E}_y(C^*_{k,n+1}) = \frac{n-k}{n} \mathbb{E}_y(C^*_{k,n}) + \frac{2}{n} \mathbb{E}_y(C^*_{k-1,n}).$$

As this is a recursion in multiple variables, it resists most standard solution methods, including the summation factor method we have used elsewhere. Inspired by the separation of variables method for solving partial differential equations (see, for example, Polyanin and Nazaikinskii [2015] Section 0.4), we assume a solution of the form

$$\mathbb{E}_y(C^*_{k,n}) = K(k)N(n),$$

where K and N are unknown functions. This gives us

$$\begin{aligned} nK(k)N(n+1) &= (n-k)K(k)N(n) + 2K(k-1)N(n); \\ n\frac{N(n+1)}{N(n)} - n &= 2\frac{K(k-1)}{K(k)} - k. \end{aligned}$$

Because the left hand side of the equation depends on n only and the right hand side depends on k only, for them to be equal to one another, they must also be equal to a constant, which we denote c. We first consider the left hand side, which can be rewritten as

$$N(n+1) = \frac{n+c}{n}N(n).$$

From Wu and Choi [2016] Corollary 2 and Proposition 3, and Theorem 4.1.3, we have

$$\begin{aligned} \mathbb{E}_y(C^*_{2,n}) = K(2)N(n) &= \frac{n}{3}, \\ \mathbb{E}_y(C^*_{3,n}) = K(3)N(n) &= \frac{n}{6}, \\ \mathbb{E}_y(C^*_{4,n}) = K(4)N(n) &= \frac{n}{15}, \end{aligned}$$

which suggests a solution of the form $N(n) = an$. We can take $N(n) = n$ without loss of generality, giving us $c = 1$. The right hand side then becomes

$$K(k) = \frac{2}{k+1}K(k-1).$$

Combined with the initial condition that $K(2) = \frac{1}{3}$, this can easily be solved to give

$$K(k) = \frac{2^{k-1}}{(k+1)!}$$

from which Theorem 5.1.3 follows. □

Lemma 5.1.1. *The probability that a rooted YHK tree or subtree with n leaves is a caterpillar is given by*

$$\mathbb{P}_y(T_n^* = C_{n,n}^*) = \frac{2^{n-2}}{(n-1)!}$$

for $n \geq 2$.

Proof. A tree with three leaves is always a pitchfork, and hence is a caterpillar. If a tree is already a caterpillar, for it to remain a caterpillar after being split, it must have been split at one of the two pendant edges in the cherry. Hence, we have

$$\mathbb{P}_y(T_3^* = C_{3,3}^*) = 1$$

and

$$\mathbb{P}_y(T_n^* = C_{n,n}^*) = \frac{2}{n-1}\mathbb{P}_y(T_{n-1}^* = C_{n-1,n-1}^*).$$

Together, these give us

$$\mathbb{P}_y(T_n^* = C_{n,n}^*) = \prod_{i=4}^{n} \frac{2}{i-1}$$

which can easily be solved to give Lemma 5.1.1. □

Theorem 5.1.4. *The expected number of k-caterpillars in an unrooted YHK tree is given by*

$$\mathbb{E}_y(C_{k,n}^\circ) = \frac{2^{k-1}n}{(k+1)!} + \frac{2^k}{(n-1)^{\underline{k}}} \sum_{j=1}^{k-1} \frac{2n-j-3}{2n-4} \binom{n-2}{j-1}$$

for $k \geq 2$ and $n > 2k$. The sum on the right hand side is a partial sum of binomial coefficients, which has no known closed form solution [Graham et al. [1989], p165]. However, approximations do exist [Worsch, 1994].

Proof. Results on unrooted YHK trees can be understood intuitively as the result of an unrooting process applied to a rooted YHK tree. Thus, the mean

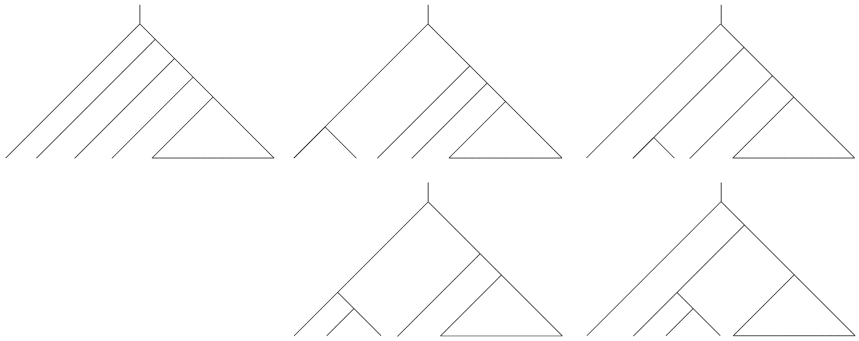

Figure 5.2: Examples of trees that will induce a 4-caterpillar when unrooted. The triangles represent the $(n-4)$-subtree comprising the remainder of the tree. Left: case 1. Middle: case 2, with a 2-caterpillar (above) or 3-caterpillar (below). Right: Case 3, again with a 2-caterpillar (above) or 3-caterpillar (below).

number of k-caterpillars in an unrooted tree is equal to the mean number of k-caterpillars already present in a rooted tree, plus the mean number of caterpillars induced by the unrooting process itself.

A new k-caterpillar can be induced in only three ways:

1. The k splits closest to the root each have exactly one pendant edge.

2. As above, but the split closest to the root instead has a j-caterpillar, where $j < k$.

3. As above, but the j-caterpillar is at the split second closest to the root.

See Figure 5.2 for an example. Any other arrangement cannot induce a k-caterpillar when unrooted, and there is no way to induce multiple k-caterpillars by unrooting.

By Rosenberg [2006] Theorem 3.7, the probability that an n leaved tree or subtree is split into a k-subtree and an $(n-k)$-subtree is equal to $\frac{1}{n-1}$ if $k = \frac{n}{2}$ and n is even, or $\frac{2}{n-1}$ if $k \neq \frac{n}{2}$ and $k \in \{1, 2, \ldots, n-1\}$. For the remainder of the proof we assume that $k < \frac{n}{2}$.

In case 1, the tree is split into a pendant edge and an $(n-1)$-subtree, which has probability $\frac{2}{n-1}$. The $(n-1)$-subtree is then split into a pendant

edge and an $(n-2)$-subtree, and so on. The probability of k splits like this is equal to

$$\prod_{i=1}^{k} \frac{2}{n-i} = \frac{2^k}{(n-1)^{\underline{k}}}.$$

For cases 2 and 3, it is simplest to consider them as variations of case 1, with two differences. First, j of the splits do not occur, and second we must multiply by the probability that the j-subtree is a caterpillar, which is provided by Lemma 5.1.1. We must also sum over all possible values of j. Summing all three cases together, we obtain

$$\begin{aligned}
\mathbb{E}_y(C^{\circ}_{k,n}) &= \mathbb{E}_y(C^{*}_{k,n}) + \frac{2^k}{(n-1)^{\underline{k}}}\left(1 + \sum_{j=2}^{k-1} \frac{2^{j-2}}{(j-1)!} \frac{(n-2)^{\underline{j-1}} + (n-3)^{\underline{j-1}}}{2^{j-1}}\right) \\
&= \mathbb{E}_y(C^{*}_{k,n}) + \frac{2^k}{(n-1)^{\underline{k}}}\left(1 + \sum_{j=2}^{k-1} \frac{(n-2)^{\underline{j-1}} + (n-3)^{\underline{j-1}}}{2(j-1)!}\right)
\end{aligned}$$

By the definition of the binomial operator, then, we have

$$= \mathbb{E}_y(C^{*}_{k,n}) + \frac{2^k}{(n-1)^{\underline{k}}}\left(1 + \frac{1}{2}\sum_{j=2}^{k-1} \binom{n-2}{j-1} + \binom{n-3}{j-1}\right)$$

We note that the 1 representing case 1. can be absorbed into the sum by changing the lower limit to $j = 1$.

$$= \mathbb{E}_y(C^{*}_{k,n}) + \frac{2^k}{(n-1)^{\underline{k}}}\,\frac{1}{2}\sum_{j=1}^{k-1} \binom{n-2}{j-1} + \binom{n-3}{j-1}.$$

By careful manipulation of the second binomial, we can obtain

$$\begin{aligned}
&= \mathbb{E}_y(C^{*}_{k,n}) + \frac{2^k}{(n-1)^{\underline{k}}}\,\frac{1}{2}\sum_{j=1}^{k-1} \left(1 + \frac{n-j-1}{n-2}\right)\binom{n-2}{j-1} \\
&= \mathbb{E}_y(C^{*}_{k,n}) + \frac{2^k}{(n-1)^{\underline{k}}}\sum_{j=1}^{k-1} \frac{2n-j-3}{2n-4}\binom{n-2}{j-1}
\end{aligned}$$

completing the proof. □

5.1.2 Moments of k-Caterpillars under the PDA Model

Theorem 5.1.5. *Under the PDA model, the conditional probability that the number of k-leaved caterpillars in a tree of n leaves is equal to some constant c is given by*

$$\mathbb{P}_u(C_{k,n+1} = c \mid C_{k,n} = a, C_{k-1,n} = b) = \begin{cases} \frac{k(b-a)}{2n-3+2r}, & \text{when } a = c-1; \\ \frac{(k-2)a}{2n-3+2r}, & \text{when } a = c+1; \\ 1 + \frac{2a-kb}{2n-3+2r}, & \text{when } a = c; \\ 0, & \text{otherwise} \end{cases}$$

for $k \geq 2$ and $n > 2k$.

Proof. The proof follows similarly to Theorem 5.1.1. Let T_n again be a random n-leaved tree containing a and b k-leaved and $(k-1)$-leaved caterpillars respectively. Suppose we then apply one step of the PDA process, splitting a random edge e to give T_{n+1}. Partitioning the edge set in T_n as in Table 5.1, we obtain the following cases:

1. $e \in E_1$: the split occurs in a $(k-1)$-caterpillar such that it induces a new k-caterpillar, hence $a = c - 1$. This occurs with conditional probability $\frac{k(b-a)}{2n-3+2r}$.

2. $e \in E_2$: the split occurs in a k-caterpillar such that it destroys an existing k-caterpillar, hence $a = c + 1$. This occurs with conditional probability $\frac{(k-2)a}{2n-3+2r}$.

3. $e \in E_3$: the split occurs elsewhere and does not alter the number of k-caterpillars. This occurs with conditional probability $1 + \frac{2a-kb}{2n-3+2r}$.

□

Theorem 5.1.6. *The number of k-caterpillars in a rooted or unrooted tree generated under the PDA model obeys the recursion*

$$\begin{aligned}\mathbb{E}_u\big(\psi(C_{k,n+1})\big) &= \frac{k}{2n-3+2r}\mathbb{E}_u\big((C_{k-1,n}-C_{k,n})\psi(C_{k,n}+1)\big)\\ &+ \frac{k-2}{2n-3+2r}\mathbb{E}_u\big(C_{k,n}\psi(C_{k,n}-1)\big)\\ &+ \frac{1}{2n-3+2r}\mathbb{E}_u\big((2n-3+2r+2C_{k,n}-kC_{k-1,n})\psi(C_{k,n})\big)\end{aligned}$$

for $k \geq 2$ and $n > 2k$.

Proof. The proof follows similarly to that of Theorem 5.1.2. By exactly the same arguments, we have

$$\begin{aligned}\mathbb{E}_u\big(\psi(C_{k,n})\big) =& \sum_a \sum_b \mathbb{P}_u(C_{k,n}=a, C_{k-1,n}=b)\\ &\times\left(\sum_c \psi(c)\mathbb{P}_u(C_{k,n+1}=c \mid C_{k,n}=a, C_{k-1,n}=b)\right).\end{aligned}$$

Substituting in Theorem 5.1.5, this is then equal to

$$\begin{aligned}=&\sum_a\sum_b \mathbb{P}_u(C_{k,n}=a, C_{k-1,n}=b)\frac{1}{2n-3+2r}\Big(k(b-a)\psi(a+1)\\ &+(k-2)a\psi(a-1)+\big(2n-3+2r+2a-kb\big)\psi(a)\Big)\\ =&\frac{2}{2n-3+2r}\mathbb{E}_u\big((C_{k-1,n}-C_{k,n})\psi(1+C_{k,n})\big)\\ &+\frac{k-2}{2n-3+2r}\mathbb{E}_u\big(C_{k,n}\psi(-1+C_{k,n})\big)\\ &+\frac{1}{2n-3+2r}\mathbb{E}_u\big((2n-3+2r+2C_{k,n}-kC_{k-1,n})\psi(C_{k,n})\big)\end{aligned}$$

which completes the proof. □

Theorem 5.1.7. *The mean number of k-caterpillars in an n-leaved tree generated under the PDA model is given by*

$$\mathbb{E}_u(C_{k,n}) = \frac{n^{\underline{k}}}{2(2n-5+2r)^{\underline{\underline{k-1}}}}$$

for $k \geq 2$ and $n > 2k$.

Proof. Substituting $\psi(x) = x$ into Theorem 5.1.6 gives us

$$\mathbb{E}_u(C_{k,n+1}) = \frac{k}{2n-3+2r}\mathbb{E}_u(C_{k-1,n}) + \frac{2n-2k-1+2r}{2n-3+2r}\mathbb{E}_u(C_{k,n})$$

We begin by summarising known values of $\mathbb{E}_u(C_{k,n})$ for small k. From Corollary 4 and Proposition 5 of Wu and Choi [2016], and Theorem 4.2.3, we have

$$\mathbb{E}_u(C^*_{2,n}) = \mathbb{E}_u(B^*_n) = \frac{n^{\underline{2}}}{2(2n-3)}$$
$$\mathbb{E}_u(C^*_{3,n}) = \mathbb{E}_u(A^*_n) = \frac{n^{\underline{3}}}{2(2n-3)^{\underline{2}}}$$
$$\mathbb{E}_u(C^*_{4,n}) = \mathbb{E}_u(C^*_n) = \frac{n^{\underline{4}}}{2(2n-3)^{\underline{3}}}$$

Then, from Theorems 3.2.3, 3.2.4, and 4.2.7, we have

$$\mathbb{E}_u(C^\circ_{2,n}) = \mathbb{E}_u(B^\circ_n) = \frac{n^{\underline{2}}}{2(2n-5)}$$
$$\mathbb{E}_u(C^\circ_{3,n}) = \mathbb{E}_u(A^\circ_n) = \frac{n^{\underline{3}}}{2(2n-5)^{\underline{2}}}$$
$$\mathbb{E}_u(C^\circ_{4,n}) = \mathbb{E}_u(C^\circ_n) = \frac{n^{\underline{4}}}{2(2n-5)^{\underline{3}}}$$

Given these forms, it is natural to take the ansatz

$$\mathbb{E}_u(C_{k,n}) = \frac{n^{\underline{k}}}{2(2n-5+2r)^{\underline{\underline{k-1}}}}$$

which can be confirmed by substitution to be a solution to the above recursion.

□

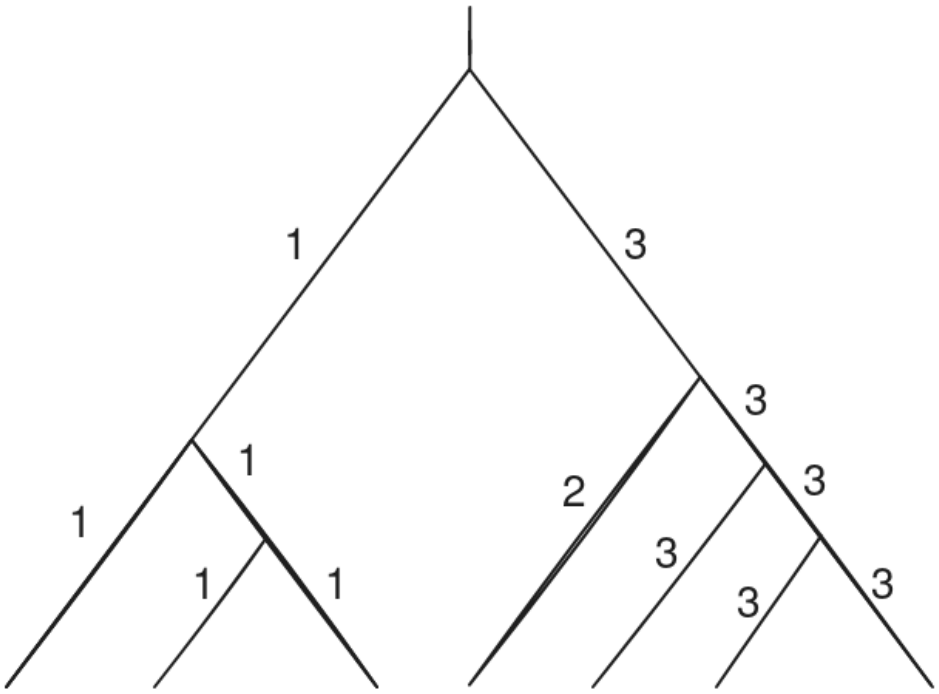

Figure 5.3: An illustration of the edge sets from Table 5.2. Edges in E_1 are labelled 1, edges in E_2 are labelled 2, and so on.

5.2 Moments of k-Subtrees

Recall that $S_{k,n}$ is the number of k-subtrees in a random tree with n leaves. Furthermore, let $\tilde{S}_{k,n}$ be the number of *independent* k-subtrees in a random tree with n leaves. Here a k-subtree is defined as independent if it is not contained in a $(k+1)$-subtrees.

5.2.1 Moments of k-Subtrees under the YHK Model

Theorem 5.2.1. *Under the YHK model, for $k \geq 3$, the conditional probability that the number of k-subtrees in a tree of n leaves is equal to some*

Edge set	Definition	$S_{k,n+1}$ after being split	Size (YHK)	Size (PDA)
E_1	Any edge that is in a $(k-1)$-subtree (including its root edge) but not in a k-subtree	$S_{k,n}+1$	$(k-1)S_{k-1,n}-2S_{k,n}$	$(2k-3)S_{k-1,n}-kS_{k,n}$
E_2	Any edge that is in a k-subtree (not including its root edge) but not in a $(k-1)$-subtree	$S_{k,n}-1$	$(k-2)S_{k,n}$	
E_3	All other edges	$S_{k,n}$	$n-(k-1)S_{k-1,n}-(k-4)S_{k,n}$	$2n-3+2r+2S_{k,n}-(2k-3)S_{k-1,n}$

Table 5.2: k-subtree related edge sets defined under the YHK and PDA models. Note that under the YHK model we consider only pendant edges. See Figure 5.3 for an example.

given s is given by

$$\mathbb{P}_y(S_{k,n+1} = s \mid S_{k,n} = a, S_{k-1,n} = b, \tilde{S}_{k-1,n} = c)$$
$$= \begin{cases} \frac{(k-1)c}{n}, & \text{when } a = s-1; \\ \frac{ka-(k-1)(b-c)}{n}, & \text{when } a = s+1; \\ \frac{n-ka+(k-1)b-2(k-1)c}{n}, & \text{when } a = s; \\ 0, & \text{otherwise} \end{cases}$$

for $k \geq 2$ and $n > 2k$.

Proof. Let T_n be a random n-leaved tree containing a k-leaved subtrees, b total $(k-1)$-trees (both dependent and independent), and c independent $(k-1)$-subtrees. Then, suppose we apply one step of the YHK process, splitting a random pendant edge e to give T_{n+1}. Partitioning the pendant edge set in T_n as in Table 5.2, we obtain the following cases:

1. $e \in E_1$: the split occurs in an independent $(k-1)$-subtree such that it induces a new k-subtree, hence $a = s-1$. This occurs with conditional probability $\frac{(k-1)c}{n}$.

2. $e \in E_2$: the split occurs in a k-subtree such that it destroys an existing k-subtree, hence $a = s+1$. This occurs with conditional probability $\frac{ka-(k-1)(b-c)}{n}$.

3. $e \in E_3$: the split occurs elsewhere and does not alter the number of k-subtrees, hence $a = s$. This occurs with conditional probability $\frac{n-ka+(k-1)b-2(k-1)c}{n}$.

□

Theorem 5.2.2. *The expectation of the number of k-subtrees in a random*

rooted or unrooted tree generated under the YHK model obeys the recursion

$$\begin{aligned}\mathbb{E}_y\Big(\psi(S_{k,n+1})\Big) &= \frac{1}{n}\mathbb{E}_y\Big((k-1)\tilde{S}_{k-1,n}\psi(S_{k,n}+1)\Big) \\ &\quad + \frac{1}{n}\mathbb{E}_y\Big(\big(kS_{k,n}-(k-1)(S_{k-1,n}-\tilde{S}_{k-1,n})\big)\psi(S_{k,n}-1)\Big) \\ &\quad + \frac{1}{n}\mathbb{E}_y\Big(\big(n-kS_{k,n}+(k-1)S_{k-1,n}-2(k-1)\tilde{S}_{k-1,n}\big)\psi(S_{k,n})\Big)\end{aligned}$$

for $k \geq 2$ and $n > 2k$.

Proof. The proof follows similarly to that of Theorem 5.1.2. We begin by noting

$$\begin{aligned}&\mathbb{E}_y\big(\psi(S_{k,n+1})\big) = \sum_s \psi(s)\mathbb{P}_y(S_{k,n+1}) \\ =&\sum_s\sum_a\sum_b\sum_c \psi(s)\mathbb{P}_y(S_{k,n+1}=s \mid S_{k,n}=a, S_{k-1,n}=b, \tilde{S}_{k-1,n}=c) \\ &\times \mathbb{P}_y(S_{k,n}=a, S_{k-1,n}=b, \tilde{S}_{k-1,n}=c).\end{aligned}$$

As a, b, c, and s all vary independently of one another, we can rearrange the order of summation.

$$\begin{aligned}=&\sum_a\sum_b\sum_c \mathbb{P}_y(S_{k,n}=a, S_{k-1,n}=b, \tilde{S}_{k-1,n}=c) \\ &\times \sum_s \psi(s)\mathbb{P}_y(S_{k,n+1}=s \mid S_{k,n}=a, S_{k-1,n}=b, \tilde{S}_{k-1,n}=c).\end{aligned}$$

From Theorem 5.2.1, then, we have

$$
\begin{aligned}
=&\sum_a \sum_b \sum_c \mathbb{P}_y(S_{k,n} = a, S_{k-1,n} = b, \tilde{S}_{k-1,n} = c) \\
&\times \left(\frac{(k-1)c\psi(a+1)}{n} + \frac{\big(ka - (k-1)(b-c)\big)\psi(a-1))}{n} \right. \\
&\left. + \frac{\big(n - ka + (k-1)b - 2(k-1)c\big)\psi(a)}{n} \right) \\
=&\frac{1}{n}\mathbb{E}_y\Big((k-1)\tilde{S}_{k-1,n}\psi(S_{k,n}+1)\Big) \\
&+ \frac{1}{n}\mathbb{E}_y\Big(\big(kS_{k,n} - (k-1)(S_{k-1,n} - \tilde{S}_{k-1,n})\big)\psi(S_{k,n}-1)\Big) \\
&+ \frac{1}{n}\mathbb{E}_y\Big(\big(n - kS_{k,n} + (k-1)S_{k-1,n} - 2(k-1)\tilde{S}_{k-1,n}\big)\psi(S_{k,n})\Big)
\end{aligned}
$$

which completes the proof. □

Theorem 5.2.3. *Rosenberg [2006] Theorem 4.4 (i) states that*

$$\mathbb{E}_y(S^*_{k,n}) = \frac{2n}{k(k+1)}$$

for $k \geq 2$ and $n > 2k$.

Proof. We present an alternative proof, similar to that of Theorem 5.1.3. Substituting $\psi(x) = x$ into Theorem 5.2.2 gives us

$$\mathbb{E}_y(S^*_{k,n+1}) = \frac{n-k}{n}\mathbb{E}_y(S^*_{k,n}) + \frac{k-1}{n}\mathbb{E}_y(S^*_{k-1,n}).$$

Again, we assume that the solution takes the form

$$\mathbb{E}_y(S^*_{k,n}) = K(k)N(n).$$

Substituting in, this gives us

$$
\begin{aligned}
nK(k)N(n+1) &= (n-k)K(k)N(n) + (k-1)K(k-1)N(n) \\
n\frac{N(n+1)}{N(n)} - n &= (k-1)\frac{K(k-1)}{K(k)} - k.
\end{aligned}
$$

As the left hand side depends only on n and the right hand side depends only on k, for them to be equal to one another, both must be equal to a constant, which we denote c. By the same argument as in Theorem 5.1.3, we have $N(n) = n$ and $c = 1$. This leaves us with

$$K(k) = \frac{k-1}{k+1} K(k-1).$$

Combined with the initial condition $K(2) = \frac{1}{3}$, this can easily be solved to give

$$K(k) = \frac{2}{k(k+1)}$$

from which Theorem 5.2.3 follows. □

Theorem 5.2.4. *Let $\psi(k) = P(\{2, \ldots, k\}) \setminus \{\}$, where $P(S)$ is the power set of S. Then, for $k \geq 2$ and $n > 2k$, we have*

$$\mathbb{E}_y(S^{\circ}_{k,n}) = \frac{2n}{k(k+1)} + \frac{2}{n-1} \sum_{S \in \psi(k)} \left(\prod_{j \in S} \frac{2}{n-j} \right).$$

Proof. Similar to the proof of Theorem 5.1.4, we consider this problem in terms of unrooting. Hence, the mean number of k-subtrees in an unrooted tree will be equal to the number of k-subtrees in a rooted tree, plus the probability that a k-subtree is induced by the unrooting process.

In order to induce a k-subtree, an $(n-k)$-subtree must exist, and the remaining k leaves must not comprise a k-subtree already. For this to occur, on the path between the root and the $(n-k)$-subtree, there must be $2 \leq i \leq k$ nodes, each of which must have a daughter subtree of size $1 \leq j_i \leq k-1$, such that the sum of all j is equal to k.

The first node below the root will separate the tree into a j_1-subtree and an $(n-j_1)$-subtree, with probability $\frac{2}{n-1}$. The second node will separate into a j_2-subtree and an $(n-j_1-j_2)$-subtree, with probability $\frac{2}{n-j_1-1}$ and so on. The probability of any particular topology will then be equal to

$$\prod_i \frac{2}{n-j_i}.$$

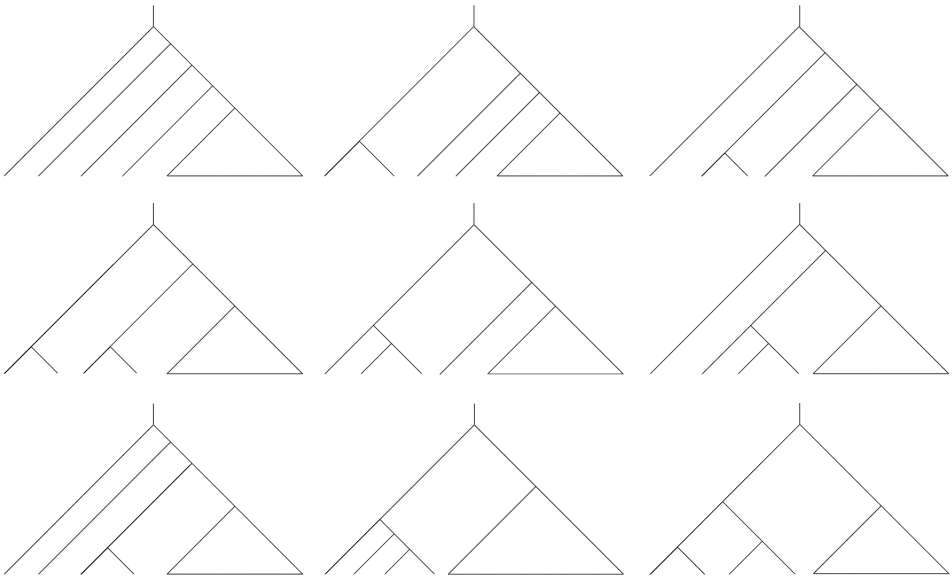

Figure 5.4: An example for $k = 4$. The triangles represent the subtrees containing the remaining $n - 4$ leaves. The first seven trees will induce a 4-subtree when unrooted, whereas the last two (corresponding to the empty set of splits considered) will not.

Any given $\frac{2}{n-j}$ term may exist or not, based on the size of the previous j-subtree, with two exceptions: the first $\frac{2}{n-1}$ term must always be present, and there must be at least one other term. The set of all combinations of j being present or not is equivalent to the power set of $\{2, \ldots, k\}$, with the exception of the empty set, which in this case corresponds to a k-subtree already being present. Summing over all members of the modified power set completes the proof. □

Corollary 5.2.1. *Again for $k \geq 2$, $n > 2k$, we have*

$$\frac{2n}{k(k+1)} + \frac{2^k(2^{k-1}-1)}{(n-1)^{\underline{k}}} < \mathbb{E}_y(S^{\circ}_{k,n}) < \frac{2n}{k(k+1)} + \frac{4(2^{k-1}-1)}{(n-1)(n-k)}.$$

Proof. If a given set has cardinality n then its power set has cardinality 2^n (see, for example, Halmos [1960] p20). Hence, we have

$$|\psi(k)| = 2^{k-1} - 1$$

Next, given that $k \geq 2$ and $n > 2k$, we have $n - k > \frac{5}{2}$, and hence

$$0 < \frac{2}{n-j} < 1$$

for all possible j. It follows that the largest possible term in the sum seen in Theorem 5.2.4 is equal to

$$\frac{2}{n-k}$$

and the smallest possible term is equal to

$$\frac{2^{k-1}}{(n-2)^{\underline{k-1}}}.$$

If we assume that all $2^{k-1} - 1$ terms in the sum are equal to the largest possible value, we obtain an upper bound,

$$\mathbb{E}_y(S^{\circ}_{k,n}) < \frac{2}{n-1}(2^{k-1}-1)\frac{2}{n-k} = \frac{4(2^{k-1}-1)}{(n-1)(n-k)}$$

and conversely if all terms are equal to the smallest value we obtain a lower bound

$$\mathbb{E}_y(S^{\circ}_{k,n}) > \frac{2}{n-1}(2^{k-1}-1)\frac{2^{k-1}}{(n-2)^{\underline{k-1}}} = \frac{2^k(2^{k-1}-1)}{(n-1)^{\underline{k}}}.$$

□

5.2.2 Moments of k-Subtrees under the PDA Model

Theorem 5.2.5. *Under the PDA model, for $k \geq 3$, the conditional probability that the number of k-subtrees in a tree of n leaves is equal to some given s is given by*

$$\mathbb{P}_u(S_{k,n+1} = s \mid S_{k,n} = a, S_{k-1,n} = b, \tilde{S}_{k-1,n} = c)$$
$$= \begin{cases} \frac{(2k-3)c}{2n-3+2r}, & \text{when } a = s-1; \\ \frac{(2k-2)a-(2k-3)(b-c)}{2n-3+2r}, & \text{when } a = s+1; \\ \frac{2n-3+2r-(2k-2)a+(2k-3)b-2(2k-3)c}{2n-3+2r}, & \text{when } a = s; \\ 0, & \text{otherwise} \end{cases}$$

for $k \geq 2$ and $n > 2k$.

Proof. Let T_n be a random n-leaved tree containing a k-leaved subtrees, b $(k-1)$-trees, and c independent $(k-1)$-subtrees. Then, suppose we apply one step of the PDA process, splitting a random edge e to give T_{n+1}. Partitioning the edge set in T_n as in Table 5.2, we obtain the following cases:

1. $e \in E_1$: the split occurs in an independent $(k-1)$-subtree such that it induces a new k-subtree, hence $a = s-1$. This occurs with probability $\frac{(2k-3)c}{2n-3+2r}$ as it is conditioned in that T_n contains c copies of independent $(k-1)$-subtrees.

2. $e \in E_2$: the split occurs in a k-subtree such that it destroys an existing k-subtree, hence $a = s+1$. This occurs with conditional probability $\frac{(2k-2)a-(2k-3)(b-c)}{2n-3+2r}$.

3. $e \in E_3$: the split occurs elsewhere and does not alter the number of k-subtrees, hence $a = s$. This occurs with conditional probability $\frac{2n-3+2r-(2k-2)a+(2k-3)b-2(2k-3)c}{2n-3+2r}$.

□

Theorem 5.2.6. *The expectation of the number of k-subtrees in a random rooted or unrooted tree generated under teh PDA model obeys the recursion*

$$\begin{aligned}\mathbb{E}_u\big(\psi(S_{k,n+1})\big) =& \frac{1}{2n-3+2r}\mathbb{E}_u\Big((2k-3)\tilde{S}_{k-1,n}\psi(S_{k,n}+1)\Big)\\ &+\frac{1}{2n-3+2r}\mathbb{E}_u\Big((2k-2)S_{k,n}\\ &-(2k-3)(S_{k-1,n}-\tilde{S}_{k-1,n})\psi(S_{k,n}-1)\Big)\\ &+\frac{1}{2n-3+2r}\mathbb{E}_u\Big(2n-3+2r-(2k-2)S_{k,n}\\ &+(2k-3)S_{k-1,n}-2(2k-3)\tilde{S}_{k-1,n}\psi(S_{k,n})\Big)\end{aligned}$$

for $k \geq 2$ and $n > 2k$.

The proof follows similarly to that of Theorem 5.2.2 and hence is omitted here.

Theorem 5.2.7. *The expected number of k-subtrees in an n-leaved tree generated under the PDA model is given by*

$$\mathbb{E}_u(S_{k,n}) = \frac{(2k-3)!!}{k!}\frac{n^{\underline{k}}}{(2n-5+2r)^{\underline{\underline{k-1}}}}$$

for $k \geq 2$ and $n > 2k$.

Proof. Letting $\psi(x) = x$ in Theorem 5.2.6 and assuming the solution takes the form $\mathbb{E}_u(S_{k,n}) = f(k)\mathbb{E}_u(C_{k,n})$, we obtain the straightforward recursion

$$f(k) = \frac{2k-3}{k}f(k-1).$$

Combined with the initial condition $f(2) = \frac{1}{2}$, this can be easily solved to give

$$f(k) = \frac{(2k-3)!!}{k!}$$

from which the expectation follows. □

6 Discussion

In this final chapter we present a discussion of the main results obtained in this thesis, as well as a section on applications and directions for future research

6.1 Summary of Results

In Chapter 3, we showed that existing results based on subtree moments in rooted trees can be directly extended to unrooted trees. In some cases, even the recursions are the same (compare, for example, Theorem 3.1.2 and Theorem 2 of Wu and Choi [2016]) however the differences in initial conditions can lead to very different forms for the joint and marginal distributions. This shows that caution must be taken when directly comparing rooted and unrooted trees, especially smaller trees where the presence or absence of the root has a proportionally larger effect.

Chapter 4 shows that results can be obtained for larger subtrees, including directly comparing different subtrees of the same number of leaves, which opened the way towards separately considering k-caterpillars and k-subtrees in Chapter 5. In Chapter 5 we showed that limited but nonetheless useful results can be obtained for subtrees of arbitrary size and topology.

Tables 6.1, 6.2, 6.3 and 6.4 provide an easy comparison of all the moments calculated in Chapters 3, 4, and 5. It is immediately apparent that the mean numbers of cherries, pitchforks, 4-caterpillars, and crabs are each greater in unrooted trees than rooted trees, under both the YHK and PDA models. Indeed, comparing the rooted and unrooted results in Table 6.4 shows this is true for all k-subtrees and caterpillars.

The variances and covariances do not follow such an obvious pattern, with some being greater in the rooted case and others greater in the unrooted case. However, we observe that as $n \to \infty$, the difference between the rooted and unrooted cases approaches zero. This makes intuitive sense, as the larger a tree is, the greater the number of non-root edges in the tree, and hence the smaller an effect the presence or absence of the root would be expected to make. For a more detailed summary of results, see Appendix 7.2.

6.2 Applications and future direction

Understanding tree generation processes is crucial for analysing phylogenetic trees inferred from empirical datasets, and can help recognise events in the evolutionary history such as adaptive radiation or mass extinctions [Mooers and Heard, 1997]. It also has applications to conservation and maintaining biodiversity [Gernhard et al., 2008].

Pouryahya and Sankoff [2022] studied the polyploidisation history of the genome of a widely cultivated variety of sugarcane, *Saccharum officinarum*. They used a "one-branch-at-a-time" model, equivalent to the PDA model studied here, as a null model to explain the observed polyploidisation of subgenomes. Through experimentally calculating the expected numbers of cherries ("terminal pairs"), pitchforks ("triples") and 4-subtrees ("quadruples"), they were able to reject the null hypothesis. The results in Theorems 3.2.3, 3.2.4, 4.2.7, and 4.2.8 of this thesis confirm Conjectures 2.2, 2.3, 3.2, 3.3, 4.2, 4.3, 4.5 and 4.6 in their paper. Together with Theorems 5.1.7 and 5.2.7, we open the way to applying this method to equivalent problems even on much larger genomes [Choi et al., 2024].

An obvious immediate avenue for future research is finding the variances and covariances for k-subtrees and k-caterpillars. We also conjecture that it is possible to obtain forms of Theorems 5.1.1, 5.1.5, 5.2.1, and 5.2.5 that are not conditional on the independence of smaller subtrees. It may also be possible to find other statistical measures of subtree distributions, such as log-concavity or total variation distance between distributions of different subtrees.

	YHK	PDA
$\mathbb{E}(B_n^*)$	$\frac{n}{3}$	$\frac{n(n-1)}{2(2n-3)}$
$\mathbb{E}(B_n^\circ)$	$\frac{n}{3} + \frac{4}{(n-1)(n-2)}$	$\frac{n(n-1)}{2(2n-5)}$
$\mathbb{V}(B_n^*)$	$\frac{2n}{45}$	$\frac{n^{\underline{4}}}{2(2n-3)^2(2n-5)}$
$\mathbb{V}(B_n^\circ)$	$\frac{2n}{45} - \frac{4(n^2-3n+14)}{3(n-1)^2(n-2)^2}$	$\frac{n^{\underline{2}}(n-4)^{\underline{2}}}{2(2n-5)(2n-5)^{\underline{2}}}$
$\mathbb{E}(A_n^*)$	$\frac{n}{6}$	$\frac{n^{\underline{3}}}{2(2n-3)^{\underline{2}}}$
$\mathbb{E}(A_n^\circ)$	$\frac{n}{6} + \frac{4(2n-3)}{(n-1)^{\underline{3}}}$	$\frac{n^{\underline{3}}}{2(2n-5)^{\underline{2}}}$
$\mathbb{V}(A_n^*)$	$\frac{23n}{420}$	$\frac{3n^{\underline{4}}(4n^3-40n^2+123n-110)}{4(2n-3)^{\underline{2}}(2n-3)^{\underline{4}}}$
$\mathbb{V}(A_n^\circ)$	$\frac{23n}{420} - \frac{16(2n-3)^2}{\left((n-1)^{\underline{3}}\right)^2}$	$\frac{3n^{\underline{3}}(4n^4-76n^3+527n^2-1555n+1610)}{4(2n-5)^{\underline{2}}(2n-5)^{\underline{4}}}$
$Cov(A_n^*, B_n^*)$	$-\frac{n}{45}$	$-\frac{3n^{\underline{4}}}{2(2n-3)(2n-3)^{\underline{3}}}$
$Cov(A_n^\circ, B_n^\circ)$	$-\frac{n}{45} - \frac{4(n^3-6n^2+35n-42)}{3(n-1)^{\underline{3}}(n-1)^{\underline{2}}}$	$-\frac{3n^{\underline{3}}(n-5)}{2(2n-5)(2n-5)^{\underline{3}}}$

Table 6.1: A summary of the results from Chapter 3, alongside corresponding results on rooted trees from Wu and Choi [2016].

	YHK	PDA
$\mathbb{E}(C_n^*)$	$\frac{n}{15}$	$\frac{n^{\underline{4}}}{2(2n-3)^{\underline{3}}}$
$\mathbb{E}(C_n^\circ)$	$\frac{n}{15} + \frac{8(n^2-4n+6)}{(n-1)^{\underline{4}}}$	$\frac{n^{\underline{4}}}{2(2n-5)^{\underline{3}}}$
$Cov(B_n^*, C_n^*)$	$-\frac{n}{35}$	$-\frac{n^{\underline{4}}(n^2-5n+3)}{2(2n-3)(2n-3)^{\underline{4}}}$
$Cov(B_n^\circ, C_n^\circ)$	$-\frac{n}{35} - \frac{y_{bc}}{45(n-1)^{\underline{4}}(n-1)^{\underline{2}}}$	$-\frac{n^{\underline{4}}(n^2-5n-5)}{2(2n-5)(2n-5)^{\underline{4}}}$
$Cov(A_n^*, C_n^*)$	$\frac{17n}{1260}$	$\frac{n^{\underline{4}}(4n^4-62n^3+340n^2-768n+585)}{2(2n-3)^{\underline{2}}(2n-3)^{\underline{5}}}$
$Cov(A_n^\circ, C_n^\circ)$	$\frac{17n}{1260} + \frac{4y_{ac}}{15(n-1)^{\underline{4}}(n-1)^{\underline{3}}}$	$\frac{n^{\underline{4}}(4n^4-90n^3+736n^2-2520n+2905)}{2(2n-5)^{\underline{2}}(2n-5)^{\underline{5}}}$
$\mathbb{V}(C_n^*)$	$\frac{67n}{1575}$	$\frac{3n^{\underline{4}}u_{cc}}{2(2n-3)^{\underline{3}}(2n-3)^{\underline{6}}}$
$\mathbb{V}(C_n^\circ)$	$\frac{67n}{1575} + \frac{y_{cc}}{1575\left((n-1)^{\underline{4}}\right)^2}$	$\frac{3n^{\underline{4}}u_{cc}}{2(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}}$

Table 6.2: A summary of the results from Chapter 4. Some results are too large to fit in the table; for the full forms, see Theorems 4.1.5, 4.2.3, and 4.2.7. Continued in Table 6.3.

	YHK	PDA
$\mathbb{E}(D_n^*)$	$\frac{n}{30}$	$\frac{n^{\underline{4}}}{8(2n-3)^{\underline{\underline{3}}}}$
$\mathbb{E}(D_n^\circ)$	$\frac{n}{30} + \frac{4n}{(n-1)^{\underline{3}}}$	$\frac{n^{\underline{4}}}{8(2n-5)^{\underline{\underline{3}}}}$
$Cov(B_n^*, D_n^*)$	$\frac{2n}{105}$	$\frac{n^{\underline{4}}(2n^3-35n+48)}{16(2n-3)(2n-3)^{\underline{\underline{4}}}}$
$Cov(B_n^\circ, D_n^\circ)$	$\frac{2n}{105} + \frac{4(3n^3-8n^2-17n+2)}{5(n-1)^{\underline{3}}(n-1)^{\underline{2}}}$	$\frac{3n^{\underline{6}}}{8(2n-5)(2n-5)^{\underline{\underline{4}}}}$
$Cov(A_n^*, D_n^*)$	$-\frac{67n}{2520}$	$-\frac{3n^{\underline{4}}(2n^4-27n^3+126n^2-236n+150)}{4(2n-3)^{\underline{\underline{2}}}(2n-3)^{\underline{\underline{5}}}}$
$Cov(A_n^\circ, D_n^\circ)$	$-\frac{67n}{2520} - \frac{y_{ad}}{1260\left((n-1)^{\underline{3}}\right)^2}$	$-\frac{3n^{\underline{4}}(2n^4-33n^3+188n^2-432n+350)}{4(2n-5)^{\underline{\underline{2}}}(2n-5)^{\underline{\underline{5}}}}$
$\mathbb{V}(D_n^*)$	$\frac{43n}{1575}$	$\frac{3n^{\underline{4}}u_{dd}}{32(2n-3)^{\underline{\underline{3}}}(2n-3)^{\underline{\underline{6}}}}$
$\mathbb{V}(D_n^\circ)$	$\frac{43n}{1575} - \frac{y_{dd}}{5670\left((n-1)^{\underline{3}}\right)^2}$	$\frac{3n^{\underline{4}}u_{dd}}{32(2n-5)^{\underline{\underline{3}}}(2n-5)^{\underline{\underline{6}}}}$
$Cov(C_n^*, D_n^*)$	$-\frac{19n}{1575}$	$-\frac{n^{\underline{4}}u_{cd}}{8(2n-3)^{\underline{\underline{3}}}(2n-3)^{\underline{\underline{6}}}}$
$Cov(C_n^\circ, D_n^\circ)$	$-\frac{19n}{1575} - \frac{y_{cd}}{11340(n-1)^{\underline{4}}(n-1)^{\underline{3}}}$	$\frac{n^{\underline{4}}u_{cd}}{8(2n-5)^{\underline{\underline{3}}}(2n-5)^{\underline{\underline{6}}}}$

Table 6.3: A summary of the results from Chapter 4. Some results are too large to fit in the table; for the full forms, see Theorems 4.1.6, 4.2.4, and 4.2.8. Continued from Table 6.2.

	YHK	PDA
$\mathbb{E}(C^*_{k,n})$	$\frac{2^{k-1}n}{(k+1)!}$	$\frac{n^{\underline{k}}}{2(2n-3)^{\overline{\overline{k-1}}}}$
$\mathbb{E}(C^\circ_{k,n})$	$\frac{2^{k-1}n}{(k+1)!} + \frac{2^k}{(n-1)^{\underline{k}}} \sum_{j=1}^{k-1} \frac{2n-j-3}{2n-4} \binom{n-2}{j-1}$	$\frac{n^{\underline{k}}}{2(2n-5)^{\overline{\overline{k-1}}}}$
$\mathbb{E}(S^*_{k,n})$	$\frac{2n}{k(k+1)}$	$\frac{(2k-3)!!}{k!} \frac{n^{\underline{k}}}{(2n-3)^{\overline{\overline{k-1}}}}$
$\mathbb{E}(S^\circ_{k,n})$	$\frac{2n}{k(k+1)} + \frac{2}{n-1} \sum_{S \in \psi(k)} \left(\prod_{j \in S} \frac{2}{n-j} \right)$	$\frac{(2k-3)!!}{k!} \frac{n^{\underline{k}}}{(2n-5)^{\overline{\overline{k-1}}}}$

Table 6.4: A summary of the results from Chapter 5.

Another avenue would be to extend these results to the more general Ford alpha model, which encompasses both the YHK and PDA models. This has already been accomplished for cherries and pitchforks [Kaur et al., 2023] but not yet for larger subtrees.

Extended Pólya urn models have been used to obtain further statistical results for subtrees under the YHK, PDA, and alpha models, most notably proving that the numbers of cherries and pitchforks are asymptotically normally distributed [McKenzie and Steel, 2000, Choi et al., 2021, Kaur et al., 2023]. However, it remains open to extend these results to larger subtrees.

While less straightforward than the recursive approach taken in this thesis, generating functions have been used in corresponding problems on ranked trees [Disanto and Wiehe, 2013], and could prove fruitful on unranked trees as well.

More ambitiously, it may be possible to apply similar methods to studying subtrees (or subgraphs) in random reticulated phylogenetic networks (or random graphs). See, for example, Bienvenu et al. [2022], Stufler [2022] for recent progress.

7 Appendices

The appendices include two sections. Section 7.1 comprises the code used to verify the initial conditions from Section 2.6. Section 7.2 comprises a detailed summary of results, alongside the initial conditions and summation factors used to calculate them.

7.1 Computed Probabilities

```
# -*- coding: utf-8 -*-
"""
Functions to calculate the probabilities of finding a given number of subtrees
in a tree with n leaves.
"""

from math import factorial as fact


Ry_known_values = {}

def Ry(n, k):
    """
    Probability that a rooted tree under the YHK model with n leaves contains k cherries.
    """
    global Ry_known_values

    # Check for impossible values.
    if n < 3:
        raise ValueError

    if k > n / 2:
        return 0

    # Check for initial conditions.
    if n == 3:
        if k == 1:
            return 1
        else:
            return 0
```

```
    # Check if it's already been calculated.
    if (n, k) in Ry_known_values:
        return Ry_known_values[(n, k)]

    # Otherwise, calculate it recursively.
    else:
        current_value = ( 2 * k * Ry(n-1, k) / (n-1) ) + ( (n - 2*k + 1) * Ry(n-1, k-1) / (n - 1) )
        Ry_known_values[(n, k)] = current_value
        return current_value


Uy_known_values = {}

def Uy(n, k):
    """
    Probability that an unrooted tree under the YHK model with n leaves contains k cherries.
    """
    global Uy_known_values

    if n < 4:
        raise ValueError

    if k > n / 2:
        return 0

    if n == 4:
        if k == 2:
            return 1
        else:
            return 0

    if (n, k) in Uy_known_values:
        return Uy_known_values[(n, k)]

    else:
        current_value = ( 2 * k * Uy(n-1, k) / (n-1) ) + ( (n - 2*k + 1) * Uy(n-1, k-1) / (n - 1) )
        Uy_known_values[(n, k)] = current_value
        return current_value


def Ru(n, k):
    """
    Probability that a rooted tree under the PDA model with n leaves contains k cherries.
    """
    return ( fact(n) * fact(n - 1) * fact(n - 2) * 2 ** (n - 2*k) ) / ( fact(n - 2*k) * fact(2*n - 2) * fact(k)


def Uu(n, k):
    """
    Probability that an unrooted tree under the PDA model with n leaves contains k cherries.
    """
    if k == 1:
        return 0

    return ( fact(n) * fact(n - 2) * fact(n - 4) * 2 ** (n - 2*k) ) / ( fact(n - 2*k) * fact(2*n - 4) * fact(k)


Py_known_values = {}
```

```
def Py(n, a, b):
    """
    Probability that an unrooted tree under the YHK model with n leaves contains a pitchforks and b cherries.
    """
    global Py_known_values

    # Initial known value.
    if n == 7:
        if a == 1 and b == 3:
            return 7/15
        elif a == 2 and b == 2:
            return 8/15
        else:
            return 0

    # Value has already been calculated.
    elif (n,a,b) in Py_known_values:
        return Py_known_values[(n,a,b)]

    # Have to actually calculate the value recursively.
    else:
        current_value = (2*a / (n - 1)) * Py(n-1, a, b) + ((a+1) / (n-1)) * Py(n-1, a+1, b-1) + ((2*(b - a + 1)
        Py_known_values[(n,a,b)] = current_value
        return current_value



Pu_known_values = {}

def Pu(n, a, b):
    """
    Probability that an unrooted tree under the PDA model with n leaves contains a pitchforks and b cherries.
    """
    global Pu_known_values

    # Initial known value.
    if n == 7:
        if a == 1 and b == 3:
            return 1/3
        elif a == 2 and b == 2:
            return 2/3
        else:
            return 0

    # Value has already been calculated.
    elif (n,a,b) in Pu_known_values:
        return Pu_known_values[(n,a,b)]

    # Have to actually calculate the value recursively.
    else:
        current_value = ((n + 3*a - b - 4) / (2*n - 5)) * Pu(n-1, a, b) + ((a + 1) / (2*n - 5)) * Pu(n-1, a+1,
        Pu_known_values[(n,a,b)] = current_value
        return current_value



P4y_known_values = {}

def P4y(n, a, b, c, d, r):
    """
    Probability that a tree under the YHK model with n leaves contains exactly
    a pitchforks, b cherries, c 4-caterpillars, and d 4-forks. r is True if the
    tree is rooted, or False if the tree is unrooted.
```

```
    """
    global P4y_known_values

    if r:
        if n < 5:
            raise ValueError("Recursion not valid for n < 5.")

        if n == 5:
            if (a, b, c, d) == (1, 2, 0, 0):
                return 1/2
            elif (a, b, c, d) == (0, 2, 0, 1):
                return 1/6
            elif (a, b, c, d) == (1, 1, 1, 0):
                return 1/3
            else:
                return 0

    else:
        if n < 8:
            raise ValueError("Recursion not valid for n < 8.")

        if n == 8:
            if (a, b, c, d) == (2, 2, 2, 0):
                return 32/105
            elif (a, b, c, d) == (2, 3, 0, 0):
                return 12/35
            elif (a, b, c, d) == (1, 3, 1, 1):
                return 2/7
            elif (a, b, c, d) == (0, 4, 0, 2):
                return 1/15
            else:
                return 0

    if (n, a, b, c, d, r) in P4y_known_values:
        return P4y_known_values[(n, a, b, c, d, r)]

    else:
        current_value = (1/(n - 1)) * (
                  (4*d + 4) * P4y(n-1, a-1, b, c, d+1, r)
                + (2*c) * P4y(n-1, a, b, c, d, r)
                + (c + 1) * P4y(n-1, a+1, b-1, c+1, d-1, r)
                + (c + 1) * P4y(n-1, a, b-1, c+1, d, r)
                + 2 * (a - c + 1) * P4y(n-1, a, b, c-1, d, r)
                + (a - c + 1) * P4y(n-1, a+1, b-1, c, d-1, r)
                + 2 * (b - 2*d - a + 1) * P4y(n-1, a-1, b, c, d, r)
                + (n - 2*b - a - c + 1) * P4y(n-1, a, b-1, c, d, r)
                )
        P4y_known_values[(n, a, b, c, d, r)] = current_value
        return current_value


P4u_known_values = {}

def P4u(n, a, b, c, d, r):
    """
    Probability that a tree under the PDA model with n leaves contains exactly
    a pitchforks, b cherries, c 4-caterpillars, and d 4-forks. r is True if the
    tree is rooted, or False if the tree is unrooted.
    """
    global P4u_known_values

    if r:
```

```
        if n < 5:
            raise ValueError("Recursion not valid for n < 5.")

        if n == 5:
            if (a, b, c, d) == (1, 2, 0, 0):
                return 2/7
            elif (a, b, c, d) == (0, 2, 0, 1):
                return 1/7
            elif (a, b, c, d) == (1, 1, 1, 0):
                return 4/7
            else:
                return 0

    else:
        if n < 8:
            raise ValueError("Recursion not valid for n < 8.")

        if n == 8:
            if (a, b, c, d) == (2, 2, 2, 0):
                return 16/33
            elif (a, b, c, d) == (2, 3, 0, 0):
                return 8/33
            elif (a, b, c, d) == (1, 3, 1, 1):
                return 8/33
            elif (a, b, c, d) == (0, 4, 0, 2):
                return 1/33
            else:
                return 0

    if (n, a, b, c, d, r) in P4u_known_values:
        return P4u_known_values[(n, a, b, c, d, r)]

    else:
        current_value = (1/(2*n - 5 + 2*r)) * (
                  (6*d + 6) * P4u(n-1, a-1, b, c, d+1, r)
                + (n - 4 + 2*r + 4*c - a - b) * P4u(n-1, a, b, c, d, r)
                + (c + 1) * P4u(n-1, a+1, b-1, c+1, d-1, r)
                + (c + 1) * P4u(n-1, a, b-1, c+1, d, r)
                + 4 * (a - c + 1) * P4u(n-1, a, b, c-1, d, r)
                + (a - c + 1) * P4u(n-1, a+1, b-1, c, d-1, r)
                + 3 * (b - a - 2*d + 1) * P4u(n-1, a-1, b, c, d, r)
                + (n + 1 - 2*b - a - c) * P4u(n-1, a, b-1, c, d, r)
                  )
        P4u_known_values[(n, a, b, c, d, r)] = current_value
        return current_value



def Ey(N, A, B, C, D, r):
    """
    Expected value of the random variables A, B, C, D, in a tree with N leaves, under the YHK model.
    Arguments A, B, C, D are the respective powers of the random variables.
    e.g. Ey(B_8 C_8) = E(8, 0, 1, 1, 0)
    """
    E = 0

    for a in range(N//3 +1):
        for b in range(1, N//2 + 1):
            for c in range(N//4 + 1):
                for d in range(N//4 + 1):
                    E += (a ** A) * (b ** B) * (c ** C) * (d ** D) * P4y(N, a, b, c, d, r)

    return E
```

```
def Eu(N, A, B, C, D, r):
    """
    Expected value of the random variables A, B, C, D, in a tree with N leaves, under the PDA model.
    Arguments A, B, C, D are the respective powers of the random variables.
    e.g. Eu(B_8 C_8) = E(8, 0, 1, 1, 0)
    """
    E = 0

    for a in range(N//3 +1):
        for b in range(1, N//2 + 1):
            for c in range(N//4 + 1):
                for d in range(N//4 + 1):
                    E += (a ** A) * (b ** B) * (c ** C) * (d ** D) * P4u(N, a, b, c, d, r)

    return E
```

7.2 Summary of Results

Expectation	Recursion	Summation Factor	Initial Condition	Result
$E_y(B_n^*)$	$E_y(B_{n+1}^*) = \frac{n-2}{n} E_y(B_n^*) + 1$	$n^{\underline{2}}$	$E_y(B_3^*) = 1$	$\frac{n}{3}$
$E_y(A_n^*)$	$E_y(A_{n+1}^*) = \frac{n-3}{n} E_y(A_n^*) + \frac{2}{n} E_y(B_n^*)$	$n^{\underline{3}}$	$E_y(A_4^*) = \frac{2}{3}$	$\frac{n}{6}$
$E_y(C_n^*)$	$E_y(C_{n+1}^*) = \frac{n-4}{n} E_y(C_n^*) + \frac{2}{n} E_y(A_n^*)$	$n^{\underline{4}}$	$E_y(C_5^*) = \frac{1}{3}$	$\frac{n}{15}$
$E_y(D_n^*)$	$E_y(D_{n+1}^*) = \frac{n-4}{n} E_y(C_n^*) + \frac{1}{n} E_y(A_n^*)$	$n^{\underline{4}}$	$E_y(D_5^*) = \frac{1}{6}$	$\frac{n}{30}$

Expectation	Recursion	Summation Factor	Initial Condition	Result
$E_y(B_n^{*2})$	$E_y(B_{n+1}^{*2}) = \frac{n-4}{n}E_y(B_n^{*2}) + \frac{2n+1}{3}$	$n^{\underline{4}}$	$E_y(B_5^{*2}) = 3$	$\frac{2n}{45} + \frac{n^2}{9}$
$E_y(A_n^{*2})$	$E_y(A_{n+1}^{*2}) = \frac{n-6}{n}E_y(A_n^{*2}) + \frac{2}{n}E_y(B_n^*) + \frac{4}{n}E_y(B_n^*A_n^*) - \frac{1}{n}E_y(A_n^*)$	$n^{\underline{6}}$	$E_y(A_7^{*2}) = \frac{157}{90}$	$\frac{23n}{420} + \frac{n^2}{36}$
$E_y(C_n^{*2})$	$E_y(C_{n+1}^{*2}) = \frac{n-8}{n}E_y(C_n^{*2}) + \frac{2}{n}E_y(A_n^*) + \frac{4}{n}E_y(A_n^*C_n^*)$	$n^{\underline{8}}$	$E_y(C_9^{*2}) = \frac{26}{35}$	$\frac{67n}{1575} + \frac{n^2}{225}$
$E_y(D_n^{*2})$	$E_y(D_n^{*2} + 1) = \frac{n-8}{n}E_y(D_n^{*2}) + \frac{4}{n}E_y(D_n^*) + \frac{2}{n}E_y(A_n^*D_n^*) + \frac{1}{n}E_y(A_n^*)$	$n^{\underline{8}}$	$E_y(D_9^*) = \frac{47}{140}$	$\frac{43n}{1575} + \frac{n^2}{900}$

Expectation	Recursion	Summation Factor	Initial Condition	Result
$E_y(B_n^*A_n^*)$	$E_y(B_{n+1}^*A_{n+1}^*) = \frac{n-5}{n}E_y(B_n^*A_n^*) + \frac{n-1}{n}E_y(A_n^*) + \frac{2}{n}E_y(B_n^{*2})$	$n^{\underline{5}}$	$E_y(B_6^*A_6^*) = \frac{28}{15}$	$\frac{n^2}{18} - \frac{n}{45}$
$E_y(B_n^*C_n^*)$	$E_y(B_{n+1}^*C_{n+1}^*) = \frac{n-6}{n}E_y(B_n^*C_n^*) + \frac{n-2}{n}E_y(C_n^*) + \frac{2}{n}E_y(B_n^*A_n^*)$	$n^{\underline{6}}$	$E_y(B_7^*C_7^*) = \frac{8}{9}$	$\frac{n^2}{45} - \frac{n}{35}$
$E_y(A_n^*C_n^*)$	$E_y(A_{n+1}^*C_{n+1}^*) = \frac{n-7}{n}E_y(A_n^*C_n^*) + \frac{1}{n}E_y(C_n^*) + \frac{2}{n}E_y(A_n^{*2}) + \frac{2}{n}E_y(B_n^*C_n^*)$	$n^{\underline{7}}$	$E_y(B_8^*C_8^*) = \frac{86}{105}$	$\frac{n^2}{90} + \frac{17n}{1260}$
$E_y(B_n^*D_n^*)$	$E_y(B_{n+1}^*D_{n+1}^*) = \frac{n-6}{n}E_y(B_n^*D_n^*) + \frac{1}{n}E_y(B_n^*A_n^*) + \frac{1}{n}E_y(A_n^*) + E_y(D_n^*)$	$n^{\underline{6}}$	$E_y(B_7^*D_7^*) = \frac{61}{90}$	$\frac{n^2}{90} + \frac{2n}{105}$
$E_y(A_n^*D_n^*)$	$E_y(A_{n+1}^*D_{n+1}^*) = \frac{n-7}{n}E_y(A_n^*D_n^*) + \frac{2}{n}E_y(B_n^*D_n^*) - \frac{4}{n}E_y(D_n^*) + \frac{1}{n}E_y(A_n^{*2}) - \frac{1}{n}E_y(A_n^*)$	$n^{\underline{7}}$	$E_y(A_8^*D_8^*) = \frac{1}{7}$	$\frac{n^2}{180} - \frac{67n}{2520}$
$E_y(C_n^*D_n^*)$	$E_y(C_{n+1}^*D_{n+1}^*) = \frac{n-8}{n}E_y(C_n^*D_n^*) - \frac{1}{n}E_y(C_n^*) + \frac{2}{n}E_y(A_n^*D_n^*) + \frac{1}{n}E_y(A_n^*C_n^*)$	$n^{\underline{8}}$	$E_y(C_9^*D_9^*) = \frac{1}{14}$	$\frac{n^2}{450} - \frac{19n}{1575}$

Expectation	Recursion	Summation Factor	Initial Condition	Result
$E_u(B_n^*)$	$E_u(B_{n+1}^*) = \frac{2n-3}{2n-1}E_u(B_n^*)+\frac{n}{2n-1}$	$2n-1$	$E_u(B_2^*) = 1$	$\frac{n^{\underline{2}}}{2(2n-3)}$
$E_u(A_n^*)$	$E_u(A_{n+1}^*) = \frac{2n-5}{2n-1}E_u(A_n^*) + \frac{3}{2n-1}E_u(B_n^*)$	$(2n-1)^{\underline{2}}$	$E_u(A_3^*) = 1$	$\frac{n^{\underline{3}}}{2(2n-3)^{\underline{2}}}$
$E_u(C_n^*)$	$E_u(C_{n+1}^*) = \frac{2n-7}{2n-1}E_u(C_n^*) + \frac{4}{2n-1}E_u(A_n^*)$	$(2n-1)^{\underline{3}}$	$E_u(C_4^*) = \frac{4}{5}$	$\frac{n^{\underline{4}}}{2(2n-3)^{\underline{3}}}$
$E_u(D_n^*)$	$E_u(D_{n+1}^*) = \frac{2n-7}{2n-1}E_u(D_n^*) + \frac{1}{2n-1}E_u(A_n^*)$	$(2n-1)^{\underline{3}}$	$E_u(D_4^*) = \frac{1}{5}$	$\frac{n^{\underline{4}}}{8(2n-3)^{\underline{3}}}$

Expectation	Recursion	Summation Factor	Initial Condition	Result
$E_u(B_n^{*2})$	$E_u(B_{n+1}^{*2}) = \frac{2n-5}{2n-1}E_u(B_n^{*2}) + \frac{2n-2}{2n-1}E_u(B_n^*) + \frac{n}{2n-1}$	$(2n-1)^{\underline{2}}$	$E_u(B_2^{*2}) = 1$	$\frac{n^{\underline{2}}(n^2-n-4)}{4(2n-3)^{\underline{2}}}$
$E_u(A_n^{*2})$	$E_u(A_{n+1}^{*2}) = \frac{2n-9}{2n-1}E_u(A_n^{*2}) + \frac{6}{2n-1}E_u(B_n^*A_n^*) - \frac{2}{2n-1}E(A_n^*) + \frac{3}{2n-1}E(B_n^*)$	$(2n-1)^{\underline{4}}$	$E_u(A_3^{*2}) = 1$	$\frac{n^{\underline{3}}(n^3-4n^2-17n+66)}{4(2n-3)^{\underline{4}}}$
$E_u(C_n^{*2})$	$E_u(C_{n+1}^{*2}) = \frac{2n-13}{2n-1}E_u(C_n^{*2}) + \frac{8}{2n-1}E_u(A_n^*C_n^*) - \frac{2}{2n-1}E(C_n^*) + \frac{4}{2n-1}E(A_n^*)$	$(2n-1)^{\underline{6}}$	$E_u(C_4^{*2}) =$	$\frac{n^{\underline{4}}(n^4-6n^3-85n^2+798n-1734)}{4(2n-3)^{\underline{6}}}$
$E_u(D_n^{*2})$	$E_u(D_{n+1}^{*2}) = \frac{2n-13}{2n-1}E_u(D_n^{*2}) + \frac{2}{2n-1}E_u(A_n^*D_n^*) + \frac{6}{2n-1}E(D_n^*) + \frac{1}{2n-1}E(A_n^*)$	$(2n-1)^{\underline{6}}$	$E_u(D_4^{*2}) =$	$\frac{n^{\underline{4}}(n^4+42n^3-877n^2+5106n-9456)}{64(2n-3)^{\underline{6}}}$

Expectation	Recursion	Summation Factor	Initial Condition	Result
$E_u(B_n^* A_n^*)$	$E_u(B_{n+1}^* A_{n+1}^*) = \frac{2n-7}{2n-1} E_u(B_n^* A_n^*) + \frac{3}{2n-1} E_u(B_n^{*2}) + \frac{n-1}{2n-1} E_u(A_n^*)$	$(2n-1)^{\underline{3}}$	$E_u(B_3^* A_3^*) =$	$\frac{n^{\underline{3}}(n^2-3n-2)}{4(2n-3)^{\underline{3}}}$
$E_u(B_n^* C_n^*)$	$E_u(B_{n+1}^* C_{n+1}^*) = \frac{2n-9}{2n-1} E_u(B_n^* C_n^*) + \frac{4}{2n-1} E_u(B_n^* A_n^*) + \frac{n-2}{2n-1} E_u(C_n^*)$	$(2n-1)^{\underline{4}}$	$E_u(B_4^* C_4^*) =$	$\frac{n^{\underline{4}}(n^2-5n+2)}{4(2n-3)^{\underline{4}}}$
$E_u(A_n^* C_n^*)$	$E_u(A_{n+1}^* C_{n+1}^*) = \frac{2n-11}{2n-1} E_u(A_n^* C_n^*) + \frac{3}{2n-1} E_u(B_n^* C_n^*) + \frac{4}{2n-1} E_u(A_n^{*2}) + \frac{1}{2n-1} E_u(C_n^*)$	$(2n-1)^{\underline{5}}$	$E_u(A_4^* C_4^*) =$	$\frac{n^{\underline{4}}(n^3-7n^2-6n+78)}{4(2n-3)^{\underline{5}}}$
$E_u(B_n^* D_n^*)$	$E_u(B_{n+1}^* D_{n+1}^*) = \frac{2n-9}{2n-1} E_u(B_n^* D_n^*) + \frac{1}{2n-1} E_u(B_n^* A_n^*) + \frac{n}{2n-1} E_u(D_n^*) + \frac{1}{2n-1} E_u(A_n^*)$	$(2n-1)^{\underline{4}}$	$E_u(B_4^* D_4^*) =$	$\frac{n^{\underline{4}}(n^2-n-16)}{16(2n-3)^{\underline{4}}}$
$E_u(A_n^* D_n^*)$	$E_u(A_{n+1}^* D_{n+1}^*) = \frac{2n-11}{2n-1} E_u(A_n^* D_n^*) + \frac{3}{2n-1} E_u(B_n^* D_n^*) + \frac{1}{2n-1} E_u(A_n^{*2}) - \frac{6}{2n-1} E_u(D_n^*) - \frac{1}{2n-1} E_u(A_n^*)$	$(2n-1)^{\underline{5}}$	$E_u(A_4^* D_4^*) =$	$\frac{n^{\underline{7}}}{16(2n-3)^{\underline{5}}}$
$E_u(C_n^* D_n^*)$	$E_u(C_{n+1}^* D_{n+1}^*) = \frac{2n-13}{2n-1} E_u(C_n^* D_n^*) + \frac{1}{2n-1} E_u(A_n^* C_n^*) + \frac{4}{2n-1} E_u(A_n^* D_n^*) - \frac{1}{2n-1} E_u(C_n^*)$	$(2n-1)^{\underline{6}}$	$E_u(C_4^* D_4^*) =$	$\frac{n^{\underline{8}}}{16(2n-3)^{\underline{6}}}$

Bibliography

David Aldous. Probability distributions on cladograms. In Random Discrete Structures, pages 1–18. Springer, 1996.

David J Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statistical Science, pages 23–34, 2001.

Benjamin Allen. Subtree Transfer Operations and their Induced Metrics on Evolutionary Trees. PhD thesis, University of Canterbury, 1998.

Alexandru T Balaban. Applications of graph theory in chemistry. Journal of chemical information and computer sciences, 25(3):334–343, 1985.

Michael Heinrich Baumann. Die k-dimensionale champagnerpyramide. Mathematische Semesterberichte, 66(1):89–100, 2019.

François Bienvenu, Amaury Lambert, and Mike Steel. Combinatorial and stochastic properties of ranked tree-child networks. Random Structures & Algorithms, 60(4):653–689, 2022.

Michael GB Blum and Olivier François. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. Systematic Biology, 55(4):685–691, 2006.

Andrew Carnie. Syntax: A generative introduction. John Wiley & Sons, 2021.

Bo Chen, Daniel Ford, and Matthias Winkel. A new family of markov branching trees: the alpha-gamma model. Electronic Journal of Probability, 14:400, 2009.

Kwok Pui Choi, Ariadne Thompson, and Taoyang Wu. On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees. Theoretical Population Biology, 132:92, 2020.

Kwok Pui Choi, Gursharn Kaur, and Taoyang Wu. On asymptotic joint distributions of cherries and pitchforks for random phylogenetic trees. Journal of Mathematical Biology, 83(4):40, 2021.

Kwok Pui Choi, Gursharn Kaur, Ariadne Thompson, and Taoyang Wu. Distributions of 4-subtree patterns for uniform random unrooted phylogenetic trees. Journal of Theoretical Biology, 584, 2024.

Michele Colledanchise and Petter Ögren. Behavior trees in robotics and AI: An introduction. CRC Press, 2018.

Narsingh Deo. Graph Theory with Applications to Engineering and Computer Science. Courier Dover Publications, 2017.

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

Daniel J Ford. Probabilities on cladograms: introduction to the alpha model. Stanford University, 2006.

Adrian Gepp, Kuldeep Kumar, and Sukanto Bhattacharya. Business failure prediction using decision trees. Journal of Forecasting, 29(6):536–555, 2010.

Tanja Gernhard, Klaas Hartmann, and Mike Steel. Stochastic properties of generalised yule models, with biodiversity applications. Journal of Mathematical Biology, 57:713–735, 2008.

Ronald L Graham, Donald E Knuth, Oren Patashnik, and Stanley Liu. Concrete mathematics: a foundation for computer science. Computers in Physics, 3(5):106–107, 1989.

Paul Richard Halmos. Naive set theory. van Nostrand, 1960.

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

David A. Huffman. A method for the construction of minimum-redundancy codes. Proceedings of the IRE, 40(9):1098–1101, 1952. doi: 10.1109/JRPROC.1952.273898.

Graham R Jones. Tree models for macroevolution and phylogenetic analysis. Systematic Biology, 60(6):735–746, 2011.

Gursharn Kaur, Kwok Pui Choi, and Taoyang Wu. Distributions of cherries and pitchforks for the ford model. Theoretical Population Biology, 149: 27–38, 2023.

John FC Kingman. On the genealogy of large populations. Journal of Applied Probability, 19(A):27–43, 1982.

S Udhaya Kumar, N Madhana Priya, SR Nithya, Priyanka Kannan, Nikita Jain, D Thirumal Kumar, R Magesh, Salma Younes, Hatem Zayed, and C George Priya Doss. A review of novel coronavirus disease (covid-19): based on genomic structure, phylogeny, current shreds of evidence, candidate vaccines, and drug repurposing. 3 Biotech, 11:1–22, 2021.

Tomás M. Coronado, Arnau Mir, Francesc Rosselló, and Lucía Rotger. On sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index. BMC Bioinformatics, 21(1):154, 2020.

John F Magee. Decision trees for decision making. Harvard Business Review Brighton, MA, USA, 1964.

Alejandro Marzinotto, Michele Colledanchise, Christian Smith, and Petter Ögren. Towards a unified behavior trees framework for robot control. In 2014 IEEE International Conference on Robotics and Automation (ICRA), pages 5420–5427. IEEE, 2014.

Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. Mathematical Biosciences, 164(1):81–92, 2000.

Arne O Mooers and Stephen B Heard. Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology, 72(1):31–54, 1997.

Miguel Nicolau, Diego Perez-Liebana, Michael O'Neill, and Anthony Brabazon. Evolutionary behavior tree approaches for navigating platform games. IEEE Transactions on Computational Intelligence and AI in Games, 9(3):227–238, 2016.

Andrei D Polyanin and Vladimir E Nazaikinskii. Handbook of linear partial differential equations for engineers and scientists. CRC press, 2015.

Fatemeh Pouryahya and David Sankoff. Peripheral structures in unlabelled trees and the accumulation of subgenomes in the evolution of polyploids. Journal of Theoretical Biology, 532:110924, 2022.

Noah A Rosenberg. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. Annals of Combinatorics, 10(1):129–146, 2006.

Michael J Sackin. "good" and "bad" phenograms. Systematic Biology, 21 (2):225–226, 1972.

Charles Semple, Mike Steel, et al. Phylogenetics, volume 24. Oxford University Press on Demand, 2003.

Benedikt Stufler. A branching process approach to level-k phylogenetic networks. Random Structures & Algorithms, 61(2):397–421, 2022.

Shan Suthaharan and Shan Suthaharan. Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, pages 237–269, 2016.

Jan Van Leeuwen. On the construction of huffman trees. In ICALP, pages 382–410, 1976.

John Van Wyhe. The complete work of charles darwin online. http://darwin-online.org.uk/content/frameset?viewtype=side&itemID=CUL-DAR121.-&pageseq=38, 2002. Accessed: 29-02-2024.

Thomas Worsch. Lower and upper bounds for (sums of) binomial coefficients. 1994. Karlsruhe 1994. (Interner Bericht. Fakultät für Informatik, Universität Karlsruhe. 1994,31.).

Taoyang Wu and Kwok Pui Choi. On joint subtree distributions under two evolutionary models. Theoretical Population Biology, 108:13–23, 2016.

George Udny Yule. A mathematical theory of evolution, based on the conclusions of dr. jc willis, fr s. Philosophical Transactions of the Royal Society of London. Series B, containing papers of a biological character, 213(402-410):21–87, 1925.