Three Combinatorial Topics Arising from Phylogenetics

A dissertation presented by

ARIEL REGUA PANINGBATAN

to

Department of Applied Mathematics

in partial fulfilment of the requirement for the degree of

Doctor of Philosophy in Applied Mathematics

supervised by

Prof. Michael Fuchs

NATIONAL CHIAO TUNG UNIVERSITY

Hsinchu, Taiwan, R.O.C.

July 2020

Preface

Phylogenetics is the study of relationships between organisms. The inferred relationship between organisms is usually represented by a tree called phylogenetic tree.

The first part of this thesis establishes some tools that will be proven useful in the study of Phylogenetics. After this, we present and study three sets of parameters that arise from investigating the structure and applications of phylogenetic trees.

The first set of parameters that we will consider is concerned with the Shapley values. These values give a fair distribution of resources among species. Different definitions of Shapley values have appeared in recent studies, the rooted and unrooted Shaley values. In one of those studies, e.g., the rooted Shapley value was proven to be equal to the fair proportion index, a value that is a lot easier to compute than the rooted Shapley value. Moreover, numerical data suggested that the unrooted Shapley value is highly correlated to the fair proportion index. In this thesis, we will give a theoretical justification of the above claim. To be precise, we will show that the the correlation coefficient between unrooted Shapley value and fair proportion index tends to 1 under the β-splitting model for $\beta > -1$. We will also present data that the convergence slows down as β approaches -1.

The second set of parameters we will consider in this thesis involves the group formation process. We will consider the number of groups formed, groups of fixed size, and the size of the largest group. Here, we will derive moments and limit laws of these values under the uniform model. Our result show that there are only a finite number of groups. Moreover, we observed that there is only one large group and the other groups are small in size. We end this part by comparing our results to the existing results under the Yule-Harding model.

Finally, the ancestral configuration at a node will give the last set of parameters that we will study. The ancestral configuration at a node is the set of distinct gene lineages that pass through the node. In particular, we are interested in the number of ancestral configurations at the root R_n and the total number of ancestral configurations T_n. We studied the mean, variance and

limit laws of R_n and T_n under the uniform and Yule-Harding model. We observed that both R_n and T_n follow a lognormal distribution. Moreover, the moments of R_n and T_n have the same exponential growth factor.

Finally, we will conclude the thesis by comparing our results to the existing results about these parameters and give some outlook on possible future research on these parameters.

Contents

List of Figures

Chapter 1

Introduction

Phylogenetics concerns the study of evolutionary relationships between species or groups of organisms. Physiological properties such as physical appearance and behaviour were used in classical studies to determine the relationship of species. Advancement of technology offers modern phylogenetics more precise connections between species by extracting the information embedded in their DNA sequences. Researchers studying phylogenetics use evolutionary (phylogenetic) trees to represent the relationship of the species. This representation can be traced back to the "Transmutation of Species" by Charles Darwin dated back to 1837 (see Figure 1.1). Due to the complexity of reconstructing evolutionary histories and developing evolutionary models, phylogenetics is now a thriving area of research which includes mathematics, statistics, computer science, and biology.

Currently, phylogenetics is concerned with far more than the mere analysis of the structure of evolutionary trees. Applications of phylogenetics can be found in a lot of different areas such as conservation biology (see (Baker and Palumbi, 1994)), epidemiology (see (Bush et al., 1999)), forensics (see (Ou et al., 1992)), gene function prediction (see (Chang and Donoghue, 2000)), and drug development (see (Chang et al., 2002)).

The goal of this thesis is to solve some problems in phylogenetics. The thesis is divided into three problems. First, we study the phylogenetic diversity index of species in an evolutionary tree model which gives each species a "rank" to determine its importance in the group. Second, we study a model for animal grouping based on evolutionary trees. Finally, we study a structural property of evolutionary trees which reflects how different genetic trees arise from an underlying evolutionary trees.

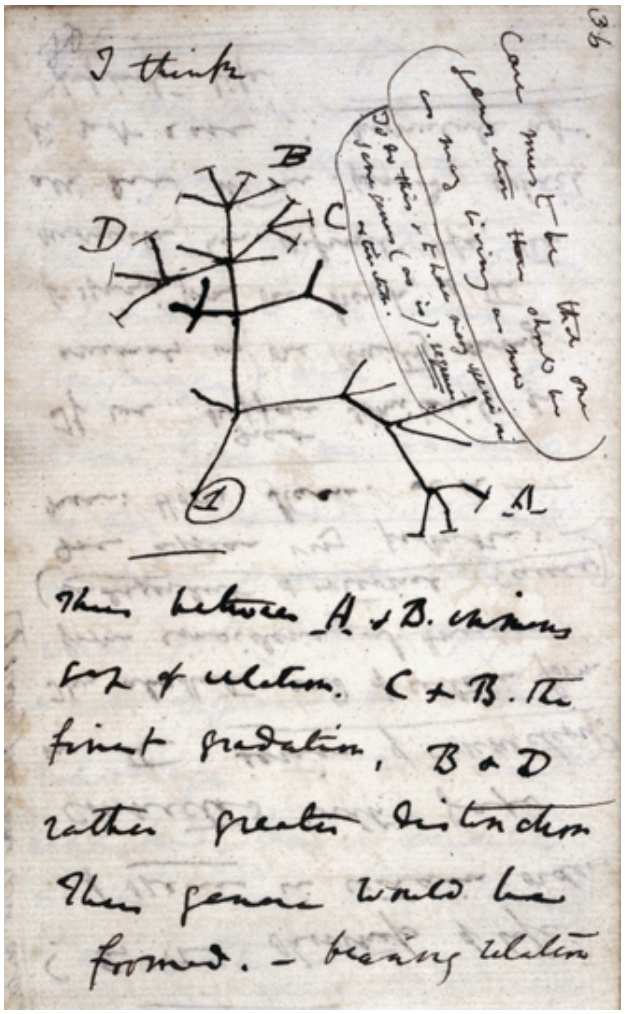


Figure 1.1: A page of "Transmutation of Species" by Charles Darwin which includes a diagram used to represent the relationship of species. The node labelled by 1 represents the common ancestor of the current species. Moreover, the leaves labelled by letters A, B, C, and D are the current existing species, whereas the leaves without labels are the extinct species.

1.1 Evolutionary Trees

This section presents different tree structures that will be used in the thesis. All trees discussed throughout the thesis possess a common underlying structure which is they are rooted bifurcating trees. Moreover, edges are often labelled where the labels represent evolutionary information (such as, e.g., time); where we usually choose all weights equal to 1 since we will mainly consider this case in the succeeding chapters. The only difference of these tree structures are whether the internal nodes and leaves are labelled or not and whether the tree is embedded into the plane or not. The labelling and embedding are used to interpret different phenomena

such as distinction of species and hierarchy in evolutionary history.

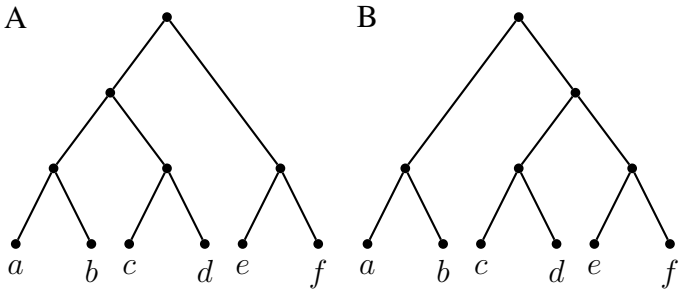


Figure 1.2: The trees A and B represent different phylogenetic trees with six taxa. However, the two trees are isomorphic since they have the same underlying topology.

Phylogenetic Tree. A *phylogenetic tree* is a rooted bifurcating tree with n labelled leaves (or taxa). A phylogenetic tree is also known as labelled topology. We denote the set of phylogenetic trees of size n by $\mathcal{T}_n$ and $\mathcal{T} = \bigcup_{n\geq 2} \mathcal{T}_n$. As a convention, we assume that the labels of the taxa are ordered linearly, that is, there is an ordering $\prec$ on the labels $\{a, b, c, d, e, \ldots\}$ such that $a \prec b \prec c \prec d \prec e \cdots$. We also say that two phylogenetic trees $\tau_1, \tau_2 \in \mathcal{T}_n$ are isomorphic, denoted by $\tau_1 \cong \tau_2$, if the trees are equal when the labelling is removed. In Figure 1.2, the phylogenetic tree in A and phylogenetic tree in B are two different phylogenetic trees but the two trees have the same underlying tree and hence they are isomorphic. Using the fact that every phylogenetic tree with n taxa produces $2n-1$ (which is the number of edges of a phylogenetic tree) distinct phylogenetic trees with $n+1$ taxa by inserting a leaf on one of its edges, the number of phylogenetic trees with n taxa is $|\mathcal{T}_n| = (2n-3)!! = 1 \times 3 \times \cdots \times (2n-3)$. This expression can be rewritten as

$$|\mathcal{T}_n| = \frac{(2n)!}{2^n(2n-1)n!} \tag{1.1}$$

with associated exponential generating function

$$T(z) = \sum_{n=1}^{\infty} \frac{|\mathcal{T}_n| z^n}{n!} = z + \frac{z}{2} + \frac{3z^2}{6} + \frac{15z^3}{24} + \cdots$$

given by

$$T(z) = 1 - \sqrt{1-2z}.$$

Plane Binary Tree. An *orientation* of a rooted bifurcating tree τ whose leaves are unlabelled is an embedding of τ into the plane. This gives τ a left-right orientation on the branches of the tree arising from an internal node. Such a rooted bifurcating tree with an orientation is called

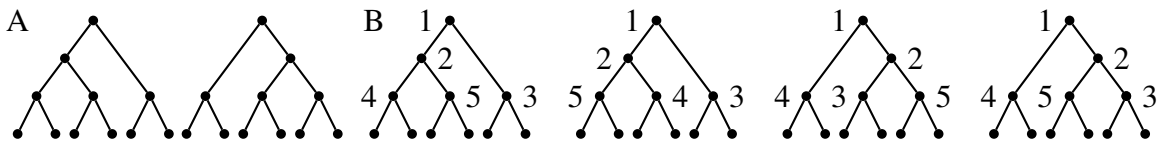


Figure 1.3: The trees in A show different orientations of a tree with six taxa. Meanwhile, the trees in B show different increasing labelling of its internal nodes.

plane binary tree. It is also called ordered unlabelled topology. Moreover, sometimes a plane binary tree is pruned which means that only the internal structure of the tree is considered. Here, we denote the set of plane binary trees with n taxa by $\mathcal{B}_n$ and $\mathcal{B} = \bigcup_{n\geq 2} \mathcal{B}_n$. The number of plane binary trees can be derived from (1.1) by adding first an orientation of the edges and then removing the labelling of its taxa. This gives us

$$|\mathcal{B}_n| = \frac{2^{n-1}}{n!} \cdot \frac{(2n)!}{2^n(2n-1)n!} = \frac{(2n-2)!}{n(n-1)!(n-1)!}. \tag{1.2}$$

Notice that this is the $n-1$-st Catalan number, denoted by C_{n-1}, where the n-th Catalan number is given by

$$C_n = \frac{1}{n+1}\binom{2n}{n} \tag{1.3}$$

with generating function

$$C(z) = \sum_{n\geq 0} C_n z^n = \frac{1-\sqrt{1-4z}}{2z}. \tag{1.4}$$

Ranked Plane Binary Tree. Let $\tau \in \mathcal{B}_n$ be a plane binary tree. For each internal node of τ assign a unique number in $\{1, 2, \ldots, n-1\}$ such that each ancestor has higher label than the descendants, that is, we assign a temporal ordering on the internal nodes of τ. A plane binary tree with such a temporal ordering on the internal nodes is called a *ranked plane binary tree*. It is also known as binary increasing tree or ordered unlabelled history. We denote the collection of ranked plane binary trees by $\mathcal{F}_n$ and $\mathcal{F} = \bigcup_{n\geq 2} \mathcal{F}_n$. Notice that every ranked plane binary tree with n taxa produces n unique ranked plane binary trees with $n+1$ taxa by attaching a cherry on a leaf of the tree. Thus, by induction, the number of ranked plane binary trees is given by

$$|\mathcal{F}_n| = (n-1)!. \tag{1.5}$$

Ranked Phylogenetic Tree. A phylogenetic tree with a temporal labelling is called a *ranked phylogenetic tree*. It is also known as labelled history or ranked dendogram. We denote the collection of ranked phylogenetic trees on n taxa by $\mathcal{H}_n$ and $\mathcal{H} = \bigcup_{n\geq 2} \mathcal{H}_n$. The number of

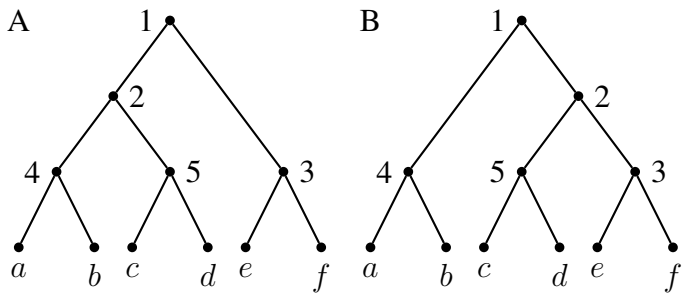


Figure 1.4: The trees A and B represent different ranked phylogenetic trees with six taxa.

ranked phylogenetic trees can be obtained from (1.5) by first labelling each taxon and then removing the orientation of the branches arising from internal nodes. Thus, we have

$$|\mathcal{H}_n| = \frac{n!}{2^{n-1}} \cdot (n-1)!. \tag{1.6}$$

1.2 Singularity Analysis

Consider the n-th Catalan number C_n and its generating function $C(z)$. Notice that for large values of n, computing the binomial coefficient is rather tedious. Moreover, just by looking at (1.3) we can not easily figure out the behaviour of C_n. Fortunately, analytic combinatorics gives us a systematic way to understand the behaviour of C_n through that of $C(z)$.

Actually, in the above simple example, we can apply Stirling's formula

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

to (1.3) to obtain

$$C_n \sim \frac{4^n}{\sqrt{\pi} n^{\frac{3}{2}}}. \tag{1.7}$$

Notice that the above expression is easier to evaluate than the binomial coefficient. Moreover, it gives an idea about how C_n behaves and in particular shows that it grows exponentially as fast as 4^n. The expression may not be precise but using more terms in the Stirling's formula, the error in (1.7) can be made as small as desired.

More generally, given a generating function $G(z)$ for a sequence g_n, we first need to extract the coefficient of $G(z)$ and then determine the asymptotic behaviour of the coefficient. However, most of the time (as we will see in the remainder of the thesis) despite having the generating function $G(z)$, extracting the exact form of the coefficient of $G(z)$ is almost impossible. Thus, a

natural question that arises is whether there is a method to obtain the asymptotics of g_n directly from the generating function $G(z)$?

Such a question can often be answered by the method of singularity analysis from analytic combinatorics. Notice that $C(z)$ has a removable singularity at $z = 0$. By redefining $C(0) = 1$, we can make $C(z)$ analytic at $z = 0$. Also, observe that $C(z)$ is not defined on the branch cut $\left[\frac{1}{4}, \infty\right)$. Unlike the removable singularity, $z = 1/4$ cannot be removed. Finally, note that C_n grows exponentially as fast as the reciprocal of $1/4$. Again, a natural question pops up: Is there a connection between the two? This will be also answered by singularity analysis. To begin with, we recall some definitions from complex analysis.

Let Ω be the interior of a simple closed curve γ. Consider a function $f(z)$ which is defined on Ω and z_0 be a point on the boundary curve γ. We say that z_0 is a *singular point* or *singularity* if $f(z)$ is not analytically continuable at z_0, that is, we cannot find an analytic function $f^*(z)$ and an open set Ω^* where $z_0 \in \Omega^*$ such that $f(z) = f^*(z)$ in $\Omega \cap \Omega^*$. For example, $z = 1/4$ is a singularity of (1.4).

Now, we need to find the locations of such singularities of $f(z)$. Recall that a power series is analytic inside its disc of convergence. Moreover, the function must not be analytic at at least one point on the boundary of this disc. This gives us the location of a dominant singularity of f which is a singularity with least modulus. The next theorem formalizes this.

Theorem 1. *Let $f(z)$ be a function which is analytic at the origin. Suppose that the power series expansion of $f(z)$ at the origin has finite radius of convergence R. Then $f(z)$ has a singularity on the boundary of the disc of convergence $\gamma = \{z \in \mathbb{C} \,:\, |z| = R\}$.*

Due to the fact that the problems in this thesis are arising from enumeration problems, the generating functions involved in our study will have non-negative coefficients. The next theorem is a refinement of Theorem 1 taking into consideration functions f whose power series has non-negative coefficients.

Theorem 2 (Pringsheim's Theorem)**.** *Let $f(z)$ be a function which is analytic at the origin. Suppose that the power series expansion of $f(z)$ at the origin has finite radius of convergence R and non-negative coefficients. Then $f(z)$ has a singularity at $z = R$.*

We have now a method for locating at least one of the dominant singularity of a function $f(z)$. Next, we need to establish the relationship between the coefficients of the series expansion of $f(z)$ and its dominant singularity. As observed earlier, we see that the coefficients seem to

grow as the reciprocal of the modulus of the dominant singularity. Before we get into that, we need to define the following.

A sequence $\{a_n\}_{n\geq 0}$ is said to be of *exponential order* K^n, denoted by the bowtie symbol $\bowtie$, if and only if $\limsup |a_n|^{1/n} = K$. We know that the limit supremum induces a natural upper bound and lower bound on a_n. That is, for any $\epsilon > 0$, we have

1. $|a_n| > (K-\epsilon)^n$ infinitely often, that is, $|a_n| > (K-\epsilon)^n$ for infinitely many values of n;
2. $|a_n| < (K+\epsilon)^n$ almost everywhere, that is, $|a_n| < (K+\epsilon)^n$ except for finite values of n.

With this we have a way of understanding the n-th coefficient $f_n := [z^n]f(z)$ of the series expansion of $f(z)$ at $z = 0$ in terms of its exponential behaviour. Notice that by the definition of the radius of convergence, for any $\epsilon > 0$, we have

1. $\lim_{n\to\infty} f_n(R-\epsilon)^n = 0$ since $\sum_{n\geq 0} f_n(R-\epsilon)^n$ converges, in particular, $f_n(R-\epsilon)^n < 1$ for sufficiently large n;
2. $f_n(R+\epsilon)^n$ is not bounded since $\sum_{n\geq 0} f_n(R+\epsilon)^n$ does not converge for all $\epsilon > 0$, in particular, $f_n(R+\epsilon)^n > 1$ for infinitely many n.

From the previous two remarks, we observe that

$$f_n \bowtie \frac{1}{R^n}.$$

The next theorem formalizes this observation.

Theorem 3. *Let $f(z)$ be a function which is analytic at the origin. Suppose that the dominant singularities are of modulus R. Then, the coefficients $f_n = [z^n]f(z)$ of its power series satisfy*

$$f_n \bowtie \frac{1}{R^n}.$$

From the previous theorem, we can now have the first principle of the coefficient asymptotic: The location of the dominant singularity of $f(z)$ gives the exponential growth behaviour of the coefficients of its power series.

From the above argument, if a function $f(z)$ has a dominant singularity of modulus R, then

$$f_n = R^{-n}\theta(n) \tag{1.8}$$

with $\limsup |\theta(n)|^{1/n} = 1$. In order to have a more precise form of f_n, we need to determine the behaviour of $\theta(n)$. The study of $\theta(n)$ leads us to the second principle of coefficient asymptotic: The sub-exponential factor $\theta(n)$ is determined by the nature of the dominant singularities of $f(z)$.

In order to explain this, first consider the function $(z-\beta)^{-r},\ r \in \mathbb{N}$ which has dominant singularity at $\beta \neq 0$. Notice that by extracting the coefficients of the power series of this function, we have

$$[z^n]\frac{1}{(z-\beta)^r} = \frac{(-1)^r}{\beta^r}[z^r]\frac{1}{(1-\frac{z}{\beta})^r} = \frac{(-1)^r}{\beta^r}\binom{n+r-1}{r-1}\beta^{-n}.$$

By definition, we know that the binomial coefficient above is a polynomial in n of degree $r-1$. Thus, the sub-exponential factor of $(z-\beta)^{-r}$ is a polynomial of degree $r-1$. In addition, notice that if β is the sole pole of a meromorphic function $f(z)$ of order r on $|z| \leq R$ we can write $f(z)$ locally around β as

$$f(z) = \sum_{j \geq -r} c_j (z-\beta)^j.$$

Let $h(z) = f(z) - \sum_{-r \leq j < 0} c_j (z-\beta)^j$. In (Flajolet and Sedgewick, 2009), the authors showed that

$$[z^n]h(z) = \mathcal{O}(R^{-n}).$$

By allowing more poles, we have the following theorem.

Theorem 4. *Let $f(z)$ be a function meromorphic on closed disc $D = \{z \in \mathbb{C} : |z| \leq R\}$ which is analytic at the origin and boundary of D. Let $\beta_1, \beta_2, \ldots, \beta_m$ be the distinct poles of $f(z)$ in D and $p_1, p_2, \ldots, p_m$ be the order of the poles, respectively. Then,*

$$f_n = \sum_{j=1}^{m} \Pi_j(n)\beta_j^{-n} + \mathcal{O}(R^{-n}). \tag{1.9}$$

for some polynomials $\Pi_j(x)$ of degree $p_j - 1$.

Notice that the asymptotic behaviour of meromorphic function relies strongly on the factor $(z-\beta)^{-r}$ as z tends to the singularity β. Also, it is sufficient to consider functions of the form $(1-z)^{-r}$ since

$$[z^n](z-\beta)^{-r} = (-1)^r \beta^{-n-r}[z^n](1-z)^{-r}.$$

The above situation motivates us to study more general functions $f(z)$ of the form $(1-z)^{-\alpha}$ where $\alpha \in \mathbb{C}$. Here, by using Cauchy integral formula

$$[z^n]f(z) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z^{n+1}}\,\mathrm{d}z$$

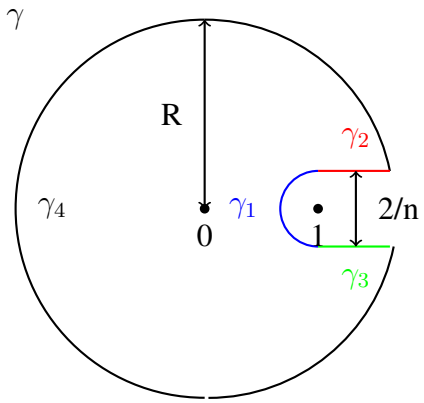


Figure 1.5: Contour γ consisting of the contours $\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4$.

and the contour $\gamma = \gamma_1 \cup \gamma_2 \cup \gamma_3 \cup \gamma_4$ where

$$\gamma_1 = \{z \in \mathbb{C} \ : \ z = 1 + e^{it}/n,\ t \in [-\pi/2, -3\pi/2]\},$$

$$\gamma_2 = \{z \in \mathbb{C} \ : \ z = t + i/n,\ t \in [1, \sqrt{R^2 - n^{-2}}]\},$$

$$-\gamma_3 = \{z \in \mathbb{C} \ : \ z = t - i/n,\ t \in [1, \sqrt{R^2 - n^{-2}}]\},$$

and

$$\gamma_4 = \{z \in \mathbb{C} \ : \ z = Re^{it},\ t \in [\arcsin((nR)^{-1}), 2\pi - \arcsin((nR)^{-1})]\}$$

(see Figure 1.5), we can analyse the asymptotic behaviour of the coefficient of $f(z)$. In particular, if we use the substitution

$$z = 1 + \frac{t}{n},$$

we have

$$\mathrm{d}z = \frac{\mathrm{d}t}{n} \ \text{ and } \ (1-z)^{-\alpha} = n^{\alpha}(-t)^{-\alpha}.$$

Moreover,

$$\lim_{n\to\infty} z^{n+1} = \lim_{n\to\infty} (1 + t/n)^{n+1} = e^t.$$

Thus,

$$[z^n](1-z)^{-\alpha} = \frac{1}{2i\pi}\int_{\gamma} \frac{f(z)}{z^{n+1}}\,\mathrm{d}z \sim \frac{n^{\alpha-1}}{2\pi i}\int_{\mathcal{H}} e^{-t}(-t)^{-\alpha}\,\mathrm{d}t,$$

where $\mathcal{H}$ is the Hankel contour. Recall that

$$\frac{1}{2\pi i}\int_{\mathcal{H}} e^{-t}(-t)^{-\alpha}\,\mathrm{d}t = \frac{1}{\Gamma(\alpha)}.$$

Consequently,

$$[z^n](1-z)^{-\alpha} \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}.$$

By refining this, we have the following result.

Theorem 5. *Let $\alpha \in \mathbb{C}\backslash\mathbb{Z}_{\le 0}$ and $f(z) = (1-z)^{-\alpha}$. Then, for sufficiently large n, the coefficient $f_n = [z^n]f(z)$ admits a complete asymptotic expansion*

$$f_n \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}\left(1 + \sum_{k=1}^{\infty}\frac{p_k}{n^k}\right),$$

where p_k is a polynomial in α of degree $2k$.

In particular, if we compute the first few terms of the asymptotic expansion, we have

$$[z^n]f(z) = \frac{n^{\alpha-1}}{\Gamma(\alpha)}\left(1 + \frac{\alpha(\alpha-1)}{2n} + \frac{\alpha(\alpha-1)(\alpha-2)(3\alpha-1)}{24n^2} + \mathcal{O}\left(\frac{1}{n^3}\right)\right).$$

For example, applying Theorem 5 to $C(z)$, we have

$$\begin{aligned}
[z^n]C(z) &= -\frac{1}{2}[z^{n+1}]\sqrt{1-4z} \\
&= -2\cdot 4^n[z^{n+1}]\sqrt{1-z} \\
&\sim \frac{-2\cdot 4^n n^{-3/2}}{\Gamma(-1/2)} = \frac{4^n}{\sqrt{\pi}n^{3/2}}.
\end{aligned}$$

Furthermore, we can also derive a more precise expansion for C_n, e.g.,

$$C_n = \frac{4^n}{\sqrt{\pi}n^{3/2}}\left(1 + \frac{3}{8n} + \frac{25}{128n^2} + \mathcal{O}\left(\frac{1}{n^3}\right)\right). \tag{1.10}$$

Next, we will study a more general form of $f(z)$ which is $(1-z)^{-\alpha}\left(\frac{1}{z}\log\frac{1}{1-z}\right)^{\beta}$. Here, using the same substitution for z, we have

$$\begin{aligned}
f(z) &= (1-z)^{-\alpha}\left(\frac{1}{z}\log\frac{1}{1-z}\right)^{\beta} \\
&\sim n^{\alpha}(-t)^{-\alpha}(\log n)^{\beta}\left(1 - \frac{\log(-t)}{\log n}\right)^{\beta}.
\end{aligned}$$

Thus, we have

$$f_n \sim \frac{n^{\alpha-1}(\log n)^{\beta}}{2\pi i}\int e^{-t}(-t)^{-\alpha}\left(1 - \frac{\log(-t)}{\log n}\right)^{\beta}\,\mathrm{d}t.$$

From this, we obtain a similar result for a larger class of functions.

Theorem 6. *Let $\alpha \in \mathbb{C}\backslash\mathbb{Z}_{\le 0}$, $\beta \in \mathbb{C}$, and $f(z) = (1-z)^{-\alpha}\left(\frac{1}{z}\log\frac{1}{1-z}\right)^{\beta}$. Then, for sufficiently large n, the coefficient $f_n = [z^n]f(z)$ admits a complete asymptotic expansion*

$$f_n \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}(\log n)^{\beta}\left(1 + \frac{c_1}{\log n} + \frac{c_2}{\log^2 n} + \cdots\right),$$

where $c_k = \binom{\beta}{k}\Gamma(\alpha)\left.\frac{\mathrm{d}^k}{\mathrm{d}s^k}\frac{1}{\Gamma(s)}\right|_{s=\alpha}$.

In fact, our functions usually do not have the form from Theorem 5 and Theorem 6. However, they do have this form if one gets close to singularities. Thus, we need the above results with the $\mathcal{O}$-notation and o-notation. Before we start, we give a definition.

Let $\phi, R \in \mathbb{R}$ with $R > 1$ and $0 < \phi < \frac{\pi}{2}$. For the parameters ϕ and R, define the open set $\Delta(\phi, R)$ by

$$\Delta(\phi, R) = \{z \mid |z| < R,\ z \neq 1,\ |arg(z-1)| > \phi\}.$$

An open set is called a Δ-*domain* if it is a $\Delta(\phi, R)$ for some R and ϕ. A function is Δ-*analytic* if it is analytic in some Δ-domain.

Theorem 7. *Let $\alpha, \beta \in \mathbb{R}$ and $f(z)$ be a function that is Δ-analytic.*

(i) Suppose $f(z)$ satisfies

$$f(z) = \mathcal{O}\left((1-z)^{-\alpha} \log\left(\frac{1}{1-z}\right)^{\beta}\right).$$

in the intersection of a neighbourhood of 1 and its Δ-domain. Then

$$[z^n]f(z) = \mathcal{O}(n^{\alpha-1}(\log n)^{\beta}).$$

(ii) Suppose $f(z)$ satisfies

$$f(z) = o\left((1-z)^{-\alpha} \log\left(\frac{1}{1-z}\right)^{\beta}\right).$$

in the intersection of a neighbourhood of 1 and its Δ-domain. Then

$$[z^n]f(z) = o(n^{\alpha-1}(\log n)^{\beta}).$$

Going back to $C(z)$, we can write the function in the form

$$C(z) = 2 - 2\sqrt{1-4z} + \mathcal{O}(\sqrt{1-4z})$$

as z tends to $1/4$. From this, we again find the asymptotics of the coefficients of $C(z)$ via Theorem 5 and Theorem 7.

With the above theorem, we cover every type of function that arises from problems involved in this thesis.

We now present a very useful tool in our computations. It shows that the singular expansion of a function is closed under derivatives and integration.

Proposition 1. *Let $f(z)$ be a Δ-analytic function with singular expansion*

$$f(z) = \sum_{j=1}^{k} c_j (1-z)^{\alpha_j} + \mathcal{O}\left((1-z)^{\beta}\right).$$

Then, for each positive integer r, $\frac{d^r}{dz^r} f(z)$ is Δ-analytic and admits a singular expansion

$$\frac{d^r}{dz^r} f(z) = (-1)^r \sum_{j=1}^{k} \frac{c_j \Gamma(\alpha_j + 1)}{\Gamma(\alpha_j + 1 - r)} (1-z)^{\alpha_j - r} + \mathcal{O}\left((1-z)^{\beta}\right).$$

With logarithmic functions, one has a similar result. Let

$$f(z) = \mathcal{O}\left((1-z)^{\alpha} \log^{\beta}(1-z)\right).$$

Then, we have

$$\frac{d^r}{dz^r} f(z) = \mathcal{O}\left((1-z)^{\alpha - r} \log^{\beta}(1-z)\right).$$

Proposition 2. *Let $f(z)$ be a Δ-analytic function with singular expansion*

$$f(z) = \sum_{j=1}^{k} c_j (1-z)^{\alpha_j} + \mathcal{O}\left((1-z)^{\beta}\right).$$

Then $\int_0^z f(t)\, dt$ is Δ-analytic. Moreover, assume that every α_j and β are not equal to 1.

1. *If $\beta < -1$, then $\int_0^z f(t)\, \mathrm{d}t$ has singular expansion*

$$\int_0^z f(t)\, \mathrm{d}t = -\sum_{j=1}^{k} \frac{c_j}{\alpha_j + 1} (1-z)^{\alpha_j + 1} + \mathcal{O}\left((1-z)^{\beta+1}\right).$$

2. *If $\beta > -1$, then $\int_0^z f(t)\, \mathrm{d}t$ has singular expansion*

$$\int_0^z f(t)\, \mathrm{d}t = -\sum_{j=1}^{k} \frac{c_j}{\alpha_j + 1} (1-z)^{\alpha_j + 1} + C + \mathcal{O}\left((1-z)^{\beta+1}\right)$$

where C is the integration constant given by

$$C = \sum_{\alpha_j < -1} \frac{c_j}{\alpha_j + 1} + \int_0^1 \left(f(t) - \sum_{\alpha_j < -1} c_j (1-z)^{\alpha_j} \right) \mathrm{d}t.$$

For the case that $\alpha = -1$ or $\beta = -1$, we have

$$\int_0^z (1-z)^{-1}\, \mathrm{d}z = \log z \qquad \text{and} \qquad \int_0^z \mathcal{O}\left((1-z)^{-1}\right)\, \mathrm{d}z = \mathcal{O}(\log z).$$

1.3 Generating Functions and Probability

Most of the enumeration we dealt with so far involves only one parameter, e.g., the number of leaves. In this section, we are interested in adding more restrictions or parameters in our enumeration problem. We start by defining the following as an analogy to the single parameter case.

Let $\mathcal{G}$ be a family of objects with associated double-indexed sequence $g_{n,k}$ which counts the number of objects in $\mathcal{G}$ with two parameters n and k. We define the *bivariate generating function* as

$$G(u,z) = \sum_{n\geq 0}\sum_{k\geq 0} g_{n,k}u^k z^n.$$

We say that $G(u,z)$ is the bivariate generating function of $\mathcal{G}$.

For example, we let $\mathcal{G}$ be the family of plane binary trees and $g_{n,k}$ denote the number of plane binary trees with size n and k cherries. Then, the bivariate generating function of $\mathcal{G}$ is given by

$$G(u,z) = \sum_{n\geq 0}\sum_{k\geq 0} g_{n,k}u^k z^n.$$

Observe that the number of cherries can be counted by adding the number of cherries in the left subtree and the number of cherries in the right subtree. Thus, we have

$$G(u,z) = z + uz^2 + G^2(u,z) - z^2.$$

Notice that $[z^n]G(u,z)$ is the generating function of the number of plane binary trees with k cherries in the class of plane binary trees with size n. This shows how plane binary trees with k cherries are distributed over the class of plane binary trees with size n. With this, we have an inkling of the relationship between bivariate generating function and probability. We make this now precise.

Let X be a discrete random variable defined over a probability space S which takes on only non-negative integers. The *probability generating function* of X is defined by

$$P(x) = \sum_{k\geq 0} \mathbb{P}_S(X = k)x^k.$$

Consider our previous example. If we let $u = 1$, then we get $[z^n]G(1,z)$ which is the generating function of the set of plane binary trees. If we define X_n as the number of cherries of a plane binary tree of size n chosen uniformly at random, then the probability generating

function of X_n is given by

$$\sum_{k=0} \mathbb{P}_{T_n}(X_n = k)u^k = \frac{[z^n]G(u,z)}{[z^n]G(1,z)}.$$

The above result is summarized in the following theorem.

Theorem 8. *Let $G(u,z)$ be the bivariate generating function of a class $\mathcal{G}$ associated with a parameters X and Y. Let $\mathcal{G}_n = \{g \in \mathcal{G} | Y(g) = n\}$. Then, the probability generating function of X with element from $\mathcal{G}_n$ picked uniformly at random is*

$$\sum_{k\geq 0} \mathbb{P}_{\mathcal{G}_n}(X = k)u^k = \frac{[z^n]G(u,z)}{[z^n]G(1,z)}.$$

Moments can be obtained from this as follows.

Corollary 1. *Let $G(u,z)$ be the bivariate generating function of a class $\mathcal{G}$ associated with a parameters X and Y. Let $\mathcal{G}_n = \{g \in \mathcal{G} | Y(g) = n\}$. Then*

$$\mathbb{E}_{\mathcal{G}_n}(X(X-1)\cdots(X-r+1)) = \frac{[z^n]\frac{\partial^r}{\partial u^r}G(u,z)|_{u=1}}{[z^n]G(1,z)}.$$

Apart from computing the expectation and higher moments of random variables, we are also interested in the limit laws of random variables. To study the limit laws, we define the following terms.

Let $D_n,\ D$ be a family of distribution functions. Then D_n is said to *converge weakly* to D if

$$\lim_{n\to\infty} D_n(x) = D(x)$$

for each $x \in \mathbb{R}$ where D is continuous. If X_n and X are the random variables associated with D_n and D, respectively, then we say that X_n converges in distribution or converges in law to X. We also say that X is the limit distribution of X_n.

The random variables that will be considered throughout this thesis will be discrete and non-negative since our problems arise from enumeration of a class of discrete objects. As for the limit laws in this thesis they will be either discrete or continuous. We first consider a discrete limit law and rephrase the above definition to this situation. The continuous case will be discussed in Section 1.5.

Let X_n be a sequence of non-negative discrete random variables. We say that X_n converges in distribution or converges in law to a discrete random variable X if

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq k) = \mathbb{P}(X \leq k)$$

for every $k \geq 0$. Moreover, we say that there exists a local limit law if

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$$

for every $k \geq 0$. Observe that both equations are equivalent by taking sums and differences of probabilities. Thus, to obtain the required limits law, it is necessary to study $\mathbb{P}(X_n \leq k)$ or $\mathbb{P}(X_n = k)$. Generating functions can be used to study $\mathbb{P}(X_n = k)$ since we have already established the relationship with $G(u, z)$ in Theorem 8.

The next theorem describes the conditions under which the convergence of probability generating functions implies a discrete limit law.

Theorem 9. *Let Ω be a set contained in the unit disc which has at least one accumulation point. Let X_n be a sequence of random variables with probability generating function $P_n(x) = \sum_{k\geq 0} P(X_n = k)x^k$. Suppose that there is a function $P(x) = \sum_{k\geq 0} p_k x^k$ satisfying*

$$\lim_{n\to\infty} P_n(x) = P(x)$$

for every $x \in \Omega$. Then, X_n converges in distribution and with all its moments to a discrete random variable X with probability generating function given by $P(x)$.

In this thesis we will be considering the composition schema

$$F(u, z) = g(uh(z)).$$

We assume that the coefficients of g and h are non-negative and the composition is well-defined. These conditions can be easily verified since our problem is enumeration of discrete objects. Next, let ρ_g and ρ_h be the radii of convergence of g and h, respectively. We also set

$$\tau_g = \lim_{x\to\rho_g^-} g(x) \quad \text{and} \quad \tau_h = \lim_{x\to\rho_h^-} h(x).$$

The limit exist or infinite since the coefficient of the functions are non-negative. Notice that the singularity of $F(u, z)$ on the positive real axis depends on the value of τ_h and ρ_g which divides the study in three different cases.

The composition schema $F(u,z) = g(uh(z))$ is said to be subcritical if $\tau_h < \rho_g$, critical if $\tau_h = \rho_g$, and supercritical if $\tau_h > \rho_g$.

In here, we are interested in the subcritical composition schema. The next theorem shows a limit law for a subcritical composition schema.

Theorem 10. *Consider the bivariate composition schema $F(u,z) = g(uh(z))$. Assume that $g(z)$ and $h(z)$ satisfy the subcritical condition. In addition, assume that $h(z)$ has a unique singularity at ρ_h on its disc of convergence and has an expansion*

$$h(z) = \tau_h - c\left(1 - \frac{z}{\rho_h}\right)^{\lambda} + o\left(\left((1 - \frac{z}{\rho_h}\right)^{\lambda}\right),$$

where $c \in \mathbb{R}^+$, $0 < \lambda < 1$ in a $\Delta-$domain. Then, for a sequence of random variables X_n defined by

$$P(X_n = k) = \frac{[u^k z^n]F(u,z)}{[z^n]F(1,z)},$$

we have a convergence in distribution and with all its moments to a discrete random variable X with probability generating function given by

$$P_X(u) = \frac{ug'(\tau_h u)}{g'(u)}.$$

Proof. First, we fix $u \in (0,1)$. The subcritical condition and the choice of u tells us that the dominant singularity of $F(u,z)$ is at ρ_h. Moreover, as $z \to \rho_h$, we have

$$F(u,z) = g(\tau_h u) - cug'(\tau_h u)(1 - z/\rho_h)^{\lambda}(1 + o(1)).$$

By Theorem 5 and Theorem 7, we have

$$\lim_{n\to\infty} \frac{[z^n]F(u,z)}{[z^n]F(1,z)} = \frac{ug'(\tau_h u)}{g'(\tau_h)}.$$

Finally, the result follows directly from Theorem 9. ∎

1.4 Random Models

There are different reasons as to why we are considering random models in studying structural properties and patterns in phylogenetic trees. Random models are used to reconstruct evolutionary processes and these reconstructed processes are then compared with the actual process or data. These models can be used to predict the outcome of some experiment or verify some hypothesis on evolutionary patterns. A wide array of statistical methods are now available which simplify seemingly complicated computations.

In this chapter, we will present two classical random models that generate random phylogenetic trees, namely the Uniform model and the Yule-Harding model. These random models where chosen in this study mainly because of their simplicity. In addition, many mathematical tools are now available to assist in the computation process. With these tools, results for different parameters of trees are widely been studied and can be used for further studies. At the end of this chapter, we will present a model proposed by Aldous which generalizes the two models.

Uniform Model. The first model we consider is the Uniform Model. This model is also known as the Proportional to Distinguishable Arrangement (PDA) model. In this model, every tree is assigned the same probability. The model can be similarly defined for different evolutionary tree structures. To avoid confusion when using different models simultaneously, we denote the probability of choosing τ under the uniform model to be

$$\mathbb{P}_{Unif_{\mathcal{M}}}(\tau) = \frac{1}{|\mathcal{M}|}$$

where $\mathcal{M} \in \{\mathcal{T}_n, \mathcal{B}_n, \mathcal{H}_n, \mathcal{F}_n\}$ and $\tau \in \mathcal{M}$. For example, the probabilities of a phylogenetic tree under the uniform model to have the shapes in Figure 1.6A and Figure 1.6B is given by $1/5$ and $4/5$, respectively.

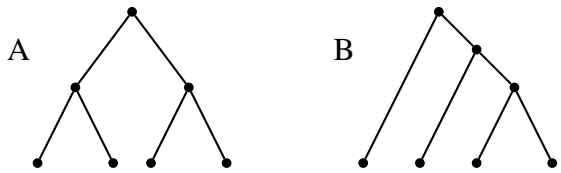


Figure 1.6: Figures A and B are the possible tree shapes for a tree with 4 taxa.

Notice that no matter how we label the tree in Figure 1.6A, for some parameters $\mathcal{X}$ on the tree, the parameters are independent of the labelling, e.g., the number of cherries (2), the height of the tree (2), etc. We say that such parameters depend only on the shape of the tree, that is, $\mathcal{X}(\tau_1) = \mathcal{X}(\tau_2)$ whenever the trees τ_1, τ_2 have the same underlying shape.

The next result shows a relationship between the uniform model of phylogenetic trees and plane binary trees.

Lemma 1. *Let $\mathcal{X}$ be a parameter that depends only on the shape of the tree. The distribution of $\mathcal{X}$ over random uniform phylogenetic trees of size n matches the distribution of $\mathcal{X}$ over random uniform plane binary trees of size n.*

Proof. We show that for every rooted bifurcating tree τ of size n, we have

$$\frac{or(\tau)}{|\mathcal{B}_n|} = \frac{lab(\tau)}{|\mathcal{T}_n|},$$

where $or(\tau)$ and $lab(\tau)$ are the orientations of τ and the number of labellings of the leaves of τ, respectively. To prove the claim, we observe that

$$or(\tau) = 2^{n-1-s(\tau)}$$

which is the number of ways to orient the children of the $n-1$ internal vertices of a rooted bifurcating tree τ and

$$lab(\tau) = \frac{n!}{2^{s(\tau)}}$$

which is the number of ways to label the leaves of τ. Here, $s(\tau)$ is the number of symmetric nodes of τ. The above equations, together with (1.1) and (1.2), justify the claim.

Note that $\frac{or(\tau)}{|\mathcal{B}_n|}$ and $\frac{lab(\tau)}{|\mathcal{T}_n|}$ are the probabilities of τ induced by the uniform distribution over the set of plane binary trees and phylogenetic trees of n taxa, respectively. Since $\mathcal{X}$ depends entirely on the shape τ, we have the desired result. ∎

Next, we have a similar result for ranked phylogenetic trees and ranked plane binary trees.

Lemma 2. *Let $\mathcal{X}$ be a parameter that depends only on the shape of the tree. The distribution of $\mathcal{X}$ over random uniform ranked phylogenetic trees of size n matches the distribution of $\mathcal{X}$ over random uniform ranked plane binary trees of size n.*

Proof. Similar to the proof of the previous theorem, we show that for every rooted bifurcating tree of size n with temporal labelling, we have

$$\frac{or(\tau)}{|\mathcal{F}_n|} = \frac{lab(\tau)}{|\mathcal{H}_n|}$$

where $or(\tau)$ and $lab(\tau)$ are the orientations of τ and the number of labellings of τ, respectively. Notice that

$$or(\tau) = 2^{n-1-c(\tau)}$$

which is the number of ways to orient the children of the $n-1$ internal vertices of a rooted bifurcating tree τ and

$$lab(\tau) = \frac{n!}{2^{c(\tau)}}$$

which is the number of ways to label the leaves of τ. Here, $c(\tau)$ is the number of cherries in τ. The above equation, together with (1.5) and (1.6), prove the claim.

Note that $\frac{or(\tau)}{|\mathcal{F}_n|}$ and $\frac{lab(\tau)}{|\mathcal{H}_n|}$ are the probabilities of τ induced by the uniform distribution over the set of ranked plane binary trees and ranked phylogenetic trees of n taxa, respectively. Since $\mathcal{X}$ depends entirely on the shape τ, we have the desired result. ∎

Due to the convenience of the left-right orientation of the children of a plane binary tree and with Lemma 1, the study of the uniform model on phylogenetic trees sometimes intermixes with the study of the uniform model on plane binary trees. The next result gives the distribution of the sizes of the left and right subtrees of a uniformly generated plane binary tree.

Lemma 3. *Let τ be a random uniform plane binary tree of size n. For $i = 1, 2, \ldots, n-1$, we let $\mathbb{P}_{Unif}(i)$ be the probability that the left subtree of τ is of size i. Then, we have*

$$\mathbb{P}_{Unif}(i) = \frac{C_{i-1}C_{n-i-1}}{C_{n-1}}$$

where C_n is the n-th Catalan number.

Proof. The proof directly follows from the number of plane binary trees of size n which is C_{n-1}, the $n-1$-st Catalan number. ∎

Yule-Harding Model. The second model we consider is the Yule-Harding model. In general, the random tree generated by the Yule-Harding model possesses the property that for each point of time, each taxa has equal chance to split. By taking note of the time split, notice that we generate a random ranked phylogenetic tree. Thus, we can say that the uniform model for $\mathcal{H}_n$ induces the Yule-Harding model for the set of phylogenetic trees $\mathcal{T}_n$.

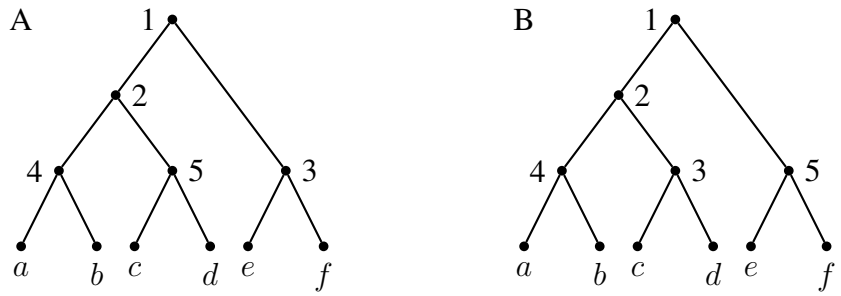


Figure 1.7: The trees A and B represent different ranked phylogenetic trees with six taxa with same underlying phylogenetic tree.

Different ranked phylogenetic trees may have the same underlying phylogenetic tree as shown in Figure 1.7. Thus, we need to determine the number of ranked phylogenetic trees that

share the same underlying phylogenetic tree. The correspond result can be found in Semple and Steel (2003), where they state that for each phylogenetic tree τ with n taxa, there exist exactly

$$\frac{(n-1)!}{\prod_{r=3}^{n}(r-1)^{d_r(\tau)}}$$

ranked phylogenetic trees with underlying shape τ. Here $d_r(\tau)$ is the number of internal vertices of τ with r leaves descending from the vertex. Taking all ranked phylogenetic trees with the same underlying phylogenetic tree, for each phylogenetic tree τ, we have

$$\mathbb{P}_{Yule}(\tau) = \frac{2^{n-1}}{n!\prod_{r=3}^{n}(r-1)^{d_r(\tau)}}. \tag{1.11}$$

Using this, the probabilities of a phylogenetic tree to have the shapes in Figure 1.6A and Figure 1.6B are $1/3$ or $2/3$, respectively. This gives us a big difference compared to the uniform model on phylogenetic trees. In particular, in contrast to the uniform model which gives each tree an equal probability, the Yule-Harding model gives higher probabily to more balanced trees since these trees have more rankings, compare with (1.11).

The next result relates the Yule-Harding model for phylogenetic trees and the uniform model for ranked plane binary trees.

Lemma 4. *Let $\mathcal{X}$ be a parameter that depends only on the shape of the tree. The distribution of $\mathcal{X}$ over phylogenetic trees of size n under the Yule distribution matches the distribution of $\mathcal{X}$ over random uniform ranked plane binary trees of size n.*

Proof. Since the Yule model for phylogenetic trees is induced by the uniform model for ranked phylogenetic trees, we see that the distributions of $\mathcal{X}$ is the same under the two models. By using Lemma 2, we have the desired result. ∎

The next result is an analog of Lemma 3 for uniform ranked plane binary trees.

Lemma 5. *Let τ be a random uniform ranked plane binary tree of size n. For $i = 1, 2, \ldots, n-1$, we let $\mathbb{P}_{Unif}(i)$ be the probability that the left subtree of τ is of size i. Then, we have*

$$\mathbb{P}_{Unif}(i) = \frac{1}{n-1}. \tag{1.12}$$

Proof. From equation (1.5), we know that the number of ranked plane binary trees of size n is $(n-1)!$. Moreover, the left subtree can be labelled by choosing $i-1$ labels from $\{2, 3, \ldots, n-$

$1\}$. Thus, we have

$$\begin{aligned}\mathbb{P}_{Unif}(i) &= \binom{n-2}{i-1}\frac{(i-1)!(n-i-1)!}{(n-1)!} \\ &= \frac{1}{n-1}.\end{aligned}$$

This proves the claim. ∎

The next model that we will introduce will be a generalization of the two models on phylogenetic trees above. It uses the same principle as the splitting process in the Yule model but generalizes the splitting probabilities.

Aldous β-splitting Model. We describe a probability distribution on phylogenetic trees using a probability distribution q_n which is symmetric, that is $q_n(i) = q_n(n-i)$ for $i \in \{1, 2, \cdots, n\}$. Figure 1.8 demonstrate the process. Consider a group of n nodes. Split the nodes into two group according to the probability distribution q_n with i nodes in the first group and $n-i$ in the second group. If one of the groups is empty, repeat this step until both groups are non-empty. Then, continue the process with each group until every group contains only one node. This defines a probability distribution on plane binary trees. By choosing random labels for the leaves and forgetting the order, it also gives a probability distribution on phylogenetic trees.

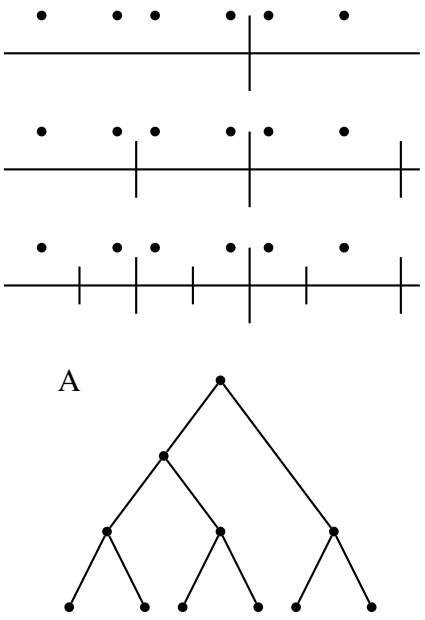


Figure 1.8: The process of constructing the phylogenetic tree A using the splitting process.

In (Aldous, 1996), Aldous suggested to choose the nodes in $[0, 1]$ and split according to a

density function $f(x)$ on $[0,1]$ which is symmetric. This gives

$$q_n(i) = a_n(\beta)^{-1}\binom{n}{i}\int_0^1 x^i(1-x)^{n-i}f(x)\,\mathrm{d}x \tag{1.13}$$

where $a_n(\beta)$ is a normalizing constant with

$$a_n(\beta) = \int_0^1 (1-x^n-(1-x)^n)f(x)\,\mathrm{d}x = 1-2\int_0^1 x^n f(x)\,\mathrm{d}x. \tag{1.14}$$

Finally, he suggested a symmetric density function which is parametrized by $-1<\beta<\infty$ and defined as

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}x^\beta(1-x)^\beta.$$

Combining the above equation with (1.13) and (1.14), we have

$$q_n(i) = \frac{1}{a_n(\beta)}\cdot\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)}$$

and

$$a_n(\beta) = \left(1-\frac{2\Gamma(2\beta+2)\Gamma(\beta+n+1)}{\Gamma(\beta+1)\Gamma(2\beta+n+2)}\right)\frac{\Gamma(\beta+1)^2\Gamma(2\beta+n+2)}{n!\Gamma(2\beta+2)}$$

for $1\le i\le n-1$ and $-1<\beta<\infty$. Notice that $q_n(i)$ is still well-defined for $-2<\beta\le -1$ (with a different normalizing factor $a_n(\beta)$). Thus, we can further extend the definition of $q_n(i)$ to $-2<\beta<\infty$.

In the sequel, we will need an asymptotic expansion of $q_n(i)$.

Lemma 6. *For $\beta>-1$, we have*

$$q_n(i) = \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)^2}n^{-2\beta-1}i^\beta(n-i)^\beta\left(1+\mathcal{O}\left(\frac{1}{i^\epsilon}+\frac{1}{(n-i)^\epsilon}\right)\right), \tag{1.15}$$

where $\epsilon>0$ is a sufficiently small constant.

Proof. For the proof, we use the well-known expansion

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\left(1+\mathcal{O}(1/z)\right), \qquad \text{as } z\longrightarrow\infty$$

which yields

$$\begin{aligned}\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{i!(n-i)!} &= \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)}\\ &= i^\beta(n-i)^\beta\left(1+\mathcal{O}\left(\frac{1}{i}+\frac{1}{n-i}\right)\right).\end{aligned}$$

Similarly,

$$c_n(\beta)^{-1} = \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)^2} n^{-2\beta-1} \left(1 + \mathcal{O}\left(\frac{1}{n^{\beta+1}}\right)\right)\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$
$$= \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)^2} n^{-2\beta-1} \left(1 + \mathcal{O}\left(\frac{1}{n^{\epsilon}}\right)\right),$$

where $\epsilon = \min\{0, \beta\} + 1$. Multiplying the last two formulas gives the claimed result. ∎

Since the β-splitting model induces both probability distributions on plane binary trees and phylogenetic trees, we have the following result.

Lemma 7. *Let $\mathcal{X}$ be a parameter that depends only on the shape of the tree. The distribution of $\mathcal{X}$ over phylogenetic trees of size n under the β-splitting model matches the distribution of $\mathcal{X}$ over plane binary trees of size n under the β-splitting model.*

Special Cases. The following are the specific models that arises from the β-splitting model.

- $\beta = -3/2$. We have
$$q_n(i) = \frac{C_{i-1}C_{n-i-1}}{C_{n-1}}.$$
Because of Lemma 3, this corresponds to the uniform model on plane binary trees and thus uniform model on phylogenetic trees.

- $\beta = 0$. We have
$$q_n(i) = \frac{1}{n-1}.$$
We will show that this induces the Yule-Harding model on phylogenetic trees. Therefore, let $\tilde{\tau}$ be a plane binary tree. Then, the splitting probability gives
$$\mathbb{P}_{\mathcal{B}_n}(\tilde{\tau}) = \frac{1}{\prod_{r=1}^{n}(r-1)^{d_r(\tilde{\tau})}}.$$
Thus, for phylogenetic trees τ we obtain
$$\mathbb{P}_{\mathcal{T}_n}(\tau) = \frac{2^{n-1}}{n!\prod_{r=1}^{n}(r-1)^{d_r(\tau)}}$$
which shows the claim; compare with (1.11).

- $\beta = -1$. We have
$$q_n(i) = \frac{n}{2H_{n-1}} \cdot \frac{1}{i(n-i)}$$
where H_n is the n-th harmonic number. This gives a particular interesting random model since it has been argued that this model gives the best fit for real-world models (see Blum and François (2006)). Unfortunately, tools used to study this model have not yet been developed. We will present some in the next Chapter.

1.5 Additive Tree Parameters

In this section we will be studying certain tree shape parameters $\mathcal{X}$ that follow a specific recursive pattern. Such patterns will be helpful in simplifying computations as will be seen in this section and proceeding chapters. Moreover, the majority of the parameters under consideration in this thesis will follow such a pattern.

Let τ be a rooted bifurcating tree. A parameter $\mathcal{X}$ of τ is said to be an *additive tree parameter* if

$$\mathcal{X}(\tau) = \mathcal{X}(\tau_1) + \mathcal{X}(\tau_2) + f(\tau) \tag{1.16}$$

where $\tau_0,\ \tau_1$ are the two root subtrees of τ. We call $f(\cdot)$ the *toll function*.

For example, the number of cherries $\mathcal{C}$ of a rooted bifurcating tree τ is an additive parameter since it follows

$$\mathcal{C}(\tau) = \mathcal{C}(\tau_1) + \mathcal{C}(\tau_2) + f(\tau)$$

where the toll function is give by

$$f(\tau) = \begin{cases} 1, & \text{if } |\tau| = 2; \\ 0, & \text{if } |\tau| \neq 2. \end{cases}$$

In Section 1.3, we have discussed that the limit law of the random variables that will be used in this thesis is either discrete or continuous. The discrete case was discussed previously and we are left with the continuous case. In the continuous case we will need the results in (Wagner, 2015). The theorems are stated as follows.

Proposition 3. *Let $\mathcal{X}$ be an additive parameter on plane binary trees with toll function f. Assume that the toll function satisfies*

$$\frac{\sum_{|\tau|=n} |f(\tau)|}{|\mathcal{B}_n|} = \mathcal{O}(c^n)$$

where $c \in (0,1)$ and the sum is taken over all plane binary trees of size n. Let τ_n be a random uniform plane binary tree of size n. Then, the mean $\mu_n = \mathbb{E}(\mathcal{X}(\tau_n))$ is given by

$$\mu_n = 2\mu n + \mathcal{O}(1)$$

where μ is given by

$$\mu = \sum_{\tau} f(\tau) 2^{1-2|\tau|}.$$

Moreover, the variance $\sigma_n^2 = \mathrm{Var}(\mathcal{X}(\tau_n))$ *is given by*

$$\sigma_n^2 = 2\sigma^2 n + \mathcal{O}(1)$$

where σ^2 *is given by*

$$\sigma^2 = \sum_{\tau} f(\tau)(2\mathcal{X}(\tau) - f(\tau))2^{1-2|\tau|} - 2\mu \sum_{\tau} f(\tau)(2|\tau| - 1)2^{1-2|\tau|}.$$

Finally, if in addition $\sigma \neq 0$*, then the random variable*

$$\frac{\mathcal{X}(\tau_n) - \mu_n}{\sigma_n}$$

converges weakly to a standard normal distribution $\mathcal{N}(0, 1)$.

We consider the parameter $\mathcal{C}$ to give an illustration of the theorem. Notice that the toll function f satisfies

$$\frac{\sum_{|\tau|=n} |f(\tau)|}{|\mathcal{B}_n|} = \begin{cases} 1, & \text{if } n = 2 \\ 0, & \text{if } n \neq 2 \end{cases}.$$

Thus, the bound is satisfied for any $c \in (0, 1)$. Applying the theorem above, we get

$$\mu = \frac{1}{8} \text{ and } \sigma^2 = \frac{1}{32}.$$

Therefore,

$$\mu_n = \frac{n}{4} + \mathcal{O}(1) \text{ and } \sigma_n^2 = \frac{n}{16} + \mathcal{O}(1).$$

Moreover,

$$\frac{\mathcal{C}(\tau_n) - \mu_n}{\sigma_n} \to \mathcal{N}(0, 1).$$

This coincides with the results of McKenzie and Steel (2000).

Proposition 4. *Let* $\mathcal{X}$ *be an additive parameter on ranked plane binary trees with toll function* f*. Assume that the toll function satisfies*

$$\frac{\sum_{|\tau|=n} |f(\tau)|}{|\mathcal{F}_n|} = \mathcal{O}(c^n)$$

where $c \in (0, 1)$ *and the sum is taken over all ranked plane binary trees of size* n*. Let* τ_n *be a random uniform ranked plane binary tree of size* n*. Then, the mean* $\mu_n = \mathbb{E}(\mathcal{X}(\tau_n))$ *is given by*

$$\mu_n = \mu n + \mathcal{O}(d^n)$$

for any $d \in (c, 1)$ *where* μ *is given by*

$$\mu = \sum_{\tau} \frac{2f(\tau)}{(|\tau|+1)!}.$$

Moreover, the variance $\sigma_n^2 = \mathrm{Var}(\mathcal{X}(\tau_n))$ *is given by*

$$\sigma_n^2 = \sigma^2 n + \mathcal{O}(d^n)$$

for any $d \in (c, 1)$ *and*

$$\sigma^2 = \sum_{\tau} \frac{2f(\tau)(2\mathcal{X}(\tau) - f(\tau))}{(|\tau|+1)!} - \mu^2 + \sum_{\tau_1} \sum_{\tau_2} \frac{4f(\tau_1)f(\tau_2)}{(|\tau_1|+1)!(|\tau_2|+1)!} \times$$
$$\left(\frac{(|\tau_1|-1)(|\tau_2|-1)}{|\tau_1|+|\tau_2|-1} - |\tau_1| - |\tau_2| + 2 + \frac{(|\tau_1|-1)(|\tau_2|-1)}{(|\tau_1|+|\tau_2|)(|\tau_1|+|\tau_2|+1)} \right.$$
$$\left. + \frac{(|\tau_1|-1)^2|(\tau_2|-1)^2}{(|\tau_1|+|\tau_2|-1)(|\tau_1|+|\tau_2|)(|\tau_1|+|\tau_2|+1)} \right).$$

Finally, if in addition $\sigma \neq 0$*, then the random variable*

$$\frac{\mathcal{X}(\tau_n) - \mu_n}{\sigma_n}$$

converges weakly to a standard normal distribution $\mathcal{N}(0, 1)$.

We consider the parameter $\mathcal{C}$ to give an illustration of the theorem. Notice that the toll function f satisfies

$$\frac{\sum_{|\tau|=n} |f(\tau)|}{|\mathcal{F}_n|} = \begin{cases} 1, & \text{if } n = 2 \\ 0, & \text{if } n \neq 2 \end{cases}.$$

Thus, the bound is satisfied for any $c \in (0, 1)$. Applying the theorem above, we get

$$\mu = \frac{1}{3} \text{ and } \sigma^2 = \frac{2}{45}.$$

Therefore, 1

$$\mu_n = \frac{n}{3} + \mathcal{O}(d^n) \text{ and } \sigma_n^2 = \frac{2n}{45} + \mathcal{O}(d^n)$$

for any $d \in (0, 1)$. Moreover,

$$\frac{\mathcal{C}(\tau_n) - \mu_n}{\sigma_n} \to \mathcal{N}(0, 1).$$

Since the number of cherries depends only on the shape of the tree, this again coincides with the results in (McKenzie and Steel, 2000) on the number of cherries of phylogenetic trees under Yule-Harding model.

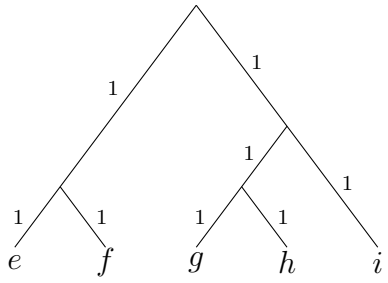
Chapter 2

Shapley Values and Fair Proportion Index

In recent year, conservation of biodiversity gains increased in popularity because of the declining population of several animal species in a particular ecosystem. Limited amount of resources prohibits groups to allocate optimal amount of resources to support every conservation projects. This led to the study on how resources should be allocated to maximize the worth of every supply available. This is also called the *Noah's Ark Problem* - optimal allocation of limited amount of resources to competing species, see (Weitzman, 1998). Along with this, a metric must be devised to measure the "importance" of species in the ecosystem.

In (Shapley, 1988), the author suggested a measure which gives a "fair" distribution of resources based on the performance of an individual in working with a group. It was originally used in cooperative game theory and subsequently refitted in the phylogeny setting. This measure is called the *Shapley value* which was named in honour of Lloyd Shapley. Several versions of this value appeared in biodiversity and the relationships between these values have been studied, see (Haake et al., 2008; Wicke and Fischer, 2017; Fuchs and Jin, 2015; Hartmann, 2013).

The Shapley value was naturally defined on phylogenetic trees since the trees represent



leaf	$\mathrm{SV}_\tau^{[r]}$	$\mathrm{SV}_\tau^{[u]}$
e	$3/2$	$26/15$
f	$3/2$	$26/15$
g	$11/6$	$91/60$
h	$11/6$	$91/60$
i	$4/3$	$27/20$

Figure 2.1: A phylogenetic tree with rooted and unrooted Shapley value for each leaf.

interaction between different species. So, we begin with a phylogenetic tree τ with edge weight 1, see Figure 2.1. For each subset S of leaves of τ, the *unrooted biodiversity* of S, denoted by $\mathrm{PD}_\tau^{[u]}(S)$, is defined to be the sum of all weights of the edges of the minimal spanning tree containing S (which is also called the *Steiner tree* of S in some combinatorics literature). For a leaf a of τ, we define the *unrooted Shapley value* of a by

$$\mathrm{SV}_\tau^{[u]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S|-1)!(n-|S|)! \left(\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\}) \right), \tag{2.1}$$

where n is the number of leaves of τ and the sum runs over all sets S of leaves containing a. Taking $a = e$ in Figure 2.1, we have $\mathrm{SV}_\tau^{[u]}(e) = 26/15$. The Shapley values of the other leaves can be seen in Figure 2.1.

We now explain the rationale behind the formula of the Shapley value. The unrooted biodiversity $\mathrm{PD}_\tau^{[u]}(S)$ describes the contribution of the group S and thus, the expression $\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\})$ describes the contribution of a in the group. Next, we take all the possible ordering of leaves of τ such that a is in the $|S|$-th position and the first $|S|$ positions are occupied by the leaves in S. Notice that regardless of how we reposition the first $|S|-1$ leaves and the final $n-|S|$ leaves, this always gives a $\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\})$ contribution. There are $(|S|-1)!(n-|S|)!$ such orderings. Taking into account all $n!$ possible ways to order the leaves of τ, we have the formula. This suggests that in Figure 2.1, the species e and f must have higher allocation of resources since they have higher contribution in the group.

The unrooted Shapley is simply called Shapley value in (Wicke and Fischer, 2017; Haake et al., 2008) but because of different definitions of the Shapley value, the prefix “unrooted" is added in this work. This is due to the fact that the root was disregarded in the definition of the unrooted Shapley value. This will be made clearer as we define a second Shapley value below.

In (Hartmann, 2013), the Shapley value was defined with the *rooted biodiversity* of S which is defined as the sum of all weights of the edges of the minimal spanning tree containing S and the root (which is also called the *ancestor tree* of S in combinatorics). Correspondingly, the *rooted Shapley value* is given by

$$\mathrm{SV}_\tau^{[r]}(a) = \frac{1}{n!} \sum_{S, a \in S} (|S|-1)!(n-|S|)! \left(\mathrm{PD}_T^{[r]}(S) - \mathrm{PD}_T^{[r]}(S \setminus \{a\}) \right). \tag{2.2}$$

The values for this Shapley value for the leaves of the tree from Figure 2.1 can also be found in the table in Figure 2.1. From the two definitions of Shapley values, the prefixes rooted and unrooted should be clear.

Even though the Shapley values give a very natural interpretation for the ordering of the leaves, both definition require a lot of computing time. Thus, many experts tried to define indices leading to easier ranking schema. In particular, Hartmann (2013) discussed another index for which he gives a heuristic and numeric data that explains that this index is closely related to the rooted Shapley value. This index is called the *fair proportion index* of a, denoted by $FP_\tau(a)$, which is given by

$$\mathrm{FP}_\tau(a) = \sum_e \frac{\lambda_e}{D_e},$$

where the sum runs over all edges on the unique path from a to the root, λ_e is the weight of edge e and D_e denotes the number of leaves below edge e. Notice that this index is much easier to compute, which is the main reason why it is frequently used in biodiversity conservation projects.

Fuchs and Jin (2015) tried to give a theoretical justification of the data in Hartmann's paper. However, they observed that

$$\mathrm{SV}_\tau^{[r]}(a) = \mathrm{FP}_\tau(a)$$

for all rooted phylogenetic trees τ and leaves a which deviates from the numerical data of Hartmann. Because of this, Fuchs and Jin (2015) believed that Hartmann used another version of the Shapley value. The authors then defined the *modified rooted Shapley value* which is given by

$$\widetilde{\mathrm{SV}}_\tau(a) = \frac{1}{n!} \sum_{|S|\geq 2, a\in S} (|S|-1)!(n-|S|)! \left(\mathrm{PD}_\tau^{[r]}(S) - \mathrm{PD}_\tau^{[r]}(S\setminus\{a\})\right).$$

Now, notice that the sum runs only over non-singletons which is quite natural since we are taking the performance of a leaf in a group. For this Shapley value, Fuchs and Jin (2015) then proved that for all weights equal to one and under the uniform and Yule-Harding model for τ,

$$\rho(\widetilde{\mathrm{SV}}_n, \mathrm{FP}_n) \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty,$$

where ρ denotes the correlation coefficient and $\widetilde{\mathrm{SV}}_n$ and FP_n are the random variables arising from the modified rooted Shapley value and fair proportion index for a random rooted phylogenetic tree with n leaves that is generated by the uniform resp. Yule-Harding model and the leaf is also chosen uniformly at random. The confusion about the data of Hartmann was resolved through an email by Mike Steel who pointed out that Hartmann used the unrooted Shapley value for his analysis.

To give now really a justification of the data of Hartmann, we tried to investigate the relationship between unrooted Shapley value and the fair proportion index. Our investigation led us

to the main result of this section which shows that the two index are indeed highly correlated. The result is stated as follows.

Theorem 11. *Assume that a random phylogenetic tree with n taxa is generated by the β-splitting model with $\beta > -1$ and choose a taxon uniformly at random. Then,*

$$\rho(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty,$$

where $\mathrm{SV}_n^{[u]}$ resp. FP_n denote the unrooted Shapley value resp. fair proportion index.

The outline of the proof is as follows. First, we observe that the main result follows directly by showing that the variance of the difference of the unrooted Shapley value and the fair proportion index tends to 0 (see Proposition 8). To obtain this convergence, moments of several parameters of phylogenetic trees under the β-splitting model are derived in Section 2.1. These parameters appear when the difference is studied in Section 2.2 and in the proof of the main theorem which is given in Section 2.3.

Unfortunately, the method of the proof does not cover $\beta \leq -1$ which includes the most important case $\beta = -1$ along with the uniform model $\beta = -3/2$. One of the reasons behind this is that the bounds rely on the asymptotic of $q_n(i)$ presented in Lemma 6. It is not clear whether our main result can be extended to the case of $\beta < -1$ as the numerical data which will be presented in Section 2.4 shows that the convergence slows down as β tends to -1.

2.1 Shape Parameters under the β-splitting model

As stated in the previous section, we will present different tree parameters that we found useful in proving Theorem 2.3. One common aspect about these parameters is the fact that they depend only on the shape of the tree. Thus, we can use the equivalence in Lemma 7 and may give a left-right orientation of the subtrees of the phylogenetic trees. Now, fix a random phylogenetic tree τ with all weights equal to 1 under the β-splitting model.

We start with the sum of all taxon-to-root distances which we denote by S_n. This parameter is important in phylogenetics where it is called the *Sackin index* and is used as a measure of imbalance of τ and also in computer science where it is called the *total path length* and used as a complexity measure for algorithms on τ. From the definition of the β-splitting model, it is clear that that S_n can be computed recursively as follows

$$S_n \stackrel{d}{=} S_{I_n} + S^*_{n-I_n} + n, \qquad (n \geq 2) \tag{2.3}$$

with $S_1 = 0$, where the equality holds in distribution and I_n is the size of the left subtree (with distribution $P(I_n = i) = q_n(i)$ for $1 \le i \le n-1$, see Section 1.4) and S_n^* is an independent copy of S_n. This follows from the fact that the Sackin index equals the sum of the Sackin indices of left and right subtree with n counting the edge from the root to left resp. right subtree for each leaf. In Figure 2.1, S_5 can be computed recursively as $S_2 + S_3 + 5 = 12$ where $S_2 = 2$ and $S_3 = 5$.

Another shape parameter, which we will also need is the distance of a random taxon to the root which we will denote by D_n. This parameter is called it the *depth* (this name is also borrowed from computer science). Similarly as S_n, it can be computed recursively:

$$D_n|(I_n = j) \stackrel{d}{=} \begin{cases} D_j + 1, & \text{with probability } j/n; \\ D_{n-j} + 1, & \text{with probability } (n-j)/n, \end{cases} \qquad (n \ge 2) \tag{2.4}$$

with $D_1 = 0$. The two cases in the bracket above correspond to the case where the chosen leaf is in the left or right subtree, respectively, and the 1's count the contribution of the left resp. right edge that connect the root to the left resp. right subtree.

Finally, note that the fair proportion index FP_n of a random taxon also allows a similar recursive description:

$$\mathrm{FP}_n|(I_n = j) \stackrel{d}{=} \begin{cases} \mathrm{FP}_j + 1/j, & \text{with probability } j/n; \\ \mathrm{FP}_{n-j} + 1/(n-j), & \text{with probability } (n-j)/n, \end{cases} \qquad (n \ge 2) \tag{2.5}$$

with $\mathrm{FP}_1 = 0$. The terms $1/j$ and $1/(n-j)$ are again coming from the left resp. right edge that connects the root with the left resp. right subtree.

As stated earlier, bounds for the moments for these parameters are needed to prove Theorem 11. Note that, from (2.3), we have

$$\mathbb{E}(S_n) = 2\sum_{j=1}^{n-1} q_n(j)\mathbb{E}(S_j) + n \qquad (n \ge 2)$$

with $\mathbb{E}(S_1) = 0$. In general, it turns out that the moments for S_n satisfy a recurrence

$$a_n = 2\sum_{j=1}^{n-1} q_n(j) a_j + b_n, \qquad (n \ge 2) \tag{2.6}$$

with $a_1 = 0$ and b_n is a given sequence. Similarly, for D_n and FP_n they will satisfy a recurrence

$$c_n = 2\sum_{j=1}^{n-1} q_n(j)\frac{j}{n} c_j + d_n, \qquad (n \ge 2)$$

with $c_1 = 0$ and d_n is again a given sequence. Setting $a_n := nc_n$ and $b_n := nd_n$ shows that in fact we only need to study the first recurrence.

The proof of the results in this section rely on the asymptotic of $q_n(j)$ in Lemma 6. Thus, throughout this section, all of the results will have an additional assumption that $\beta > -1$.

We will prove two general results about a sequence a_n which satisfies (2.6). The first result is a version of what computer scientists call a *master theorem* (see, e.g., Sokal and Rohlf (1962) for similar results).

Proposition 5. *Let a_n be sequence which satisfies (2.6) with*

$$b_n = \mathcal{O}(n^\gamma (\log n)^\delta)$$

for non-negative integers γ and δ. Then, we have

(i) if $\gamma = 1$, then $a_n = \mathcal{O}(n(\log n)^{\delta+1})$;

(ii) if $\gamma > 1$, then $a_n = \mathcal{O}(n^\gamma (\log n)^\delta)$.

Proof. We first prove part (i). By assumption, we have that $b_n \leq dn(\log n)^\delta$ for some $d > 0$. We will proceed by induction and therefore assume that $a_k \leq ck(\log k)^{\delta+1}$ for $1 \leq k < n$ with a suitable constant $c > 0$ (which can be chosen such that this holds up to some fixed n).

First, notice that by Lemma 6,

$$2\sum_{j=1}^{n-1} q_n(j) j (\log j)^{\delta+1} = \frac{2\Gamma(2\beta+2)}{\Gamma(\beta+1)^2} \frac{1}{n^{2\beta+1}} \sum_{j=1}^{n-1} j^{\beta+1}(n-j)^\beta (\log j)^{\delta+1} \left(1 + \mathcal{O}\left(\frac{1}{j^\epsilon} + \frac{1}{(n-j)^\epsilon}\right)\right). \tag{2.7}$$

Next, observe that

$$\begin{aligned}
&\frac{1}{n^{2\beta+1}} \sum_{j=1}^{n-1} j^{\beta+1}(n-j)^\beta (\log j)^{\delta+1} \\
&\quad = \frac{1}{n^{2\beta+1}} \sum_{j=1}^{n-1} j^{\beta+1}(n-j)^\beta (\log n + \log(j/n))^{\delta+1} \\
&\quad = (\log n)^{\delta+1} \sum_{j=1}^{n-1} \left(\frac{j}{n}\right)^{\beta+1} \left(1 - \frac{j}{n}\right)^\beta + (\delta+1)(\log n)^\delta \sum_{j=1}^{n-1} \left(\frac{j}{n}\right)^{\beta+1} \left(1 - \frac{j}{n}\right)^\beta \log(j/n) \\
&\qquad + \mathcal{O}\left((\log n)^{\delta-1} \sum_{j=1}^{n-1} \left(\frac{j}{n}\right)^{\beta+1} \left(1 - \frac{j}{n}\right)^\beta (\log(j/n))^2\right).
\end{aligned}$$

From a standard application of the Euler-Maclaurin summation formula

$$\sum_{j=1}^{n-1}\left(\frac{j}{n}\right)^{\beta+1}\left(1-\frac{j}{n}\right)^{\beta} = n\int_0^1 x^{\beta+1}(1-x)^{\beta}\mathrm{d}x + \mathcal{O}(n^{1-\epsilon})$$
$$= n\frac{\Gamma(\beta+2)\Gamma(\beta+1)}{\Gamma(2\beta+3)} + \mathcal{O}(n^{1-\epsilon})$$

and

$$\sum_{j=1}^{n-1}\left(\frac{j}{n}\right)^{\beta+1}\left(1-\frac{j}{n}\right)^{\beta}\log(j/n) = n\int_0^1 x^{\beta+1}(1-x)^{\beta}\log(x)\mathrm{d}x + \mathcal{O}(n^{1-\epsilon}),$$

where $\epsilon > 0$ is a suitable small constant. Moreover, by replacing the sum by an integral,

$$\sum_{j=1}^{n-1}\left(\frac{j}{n}\right)^{\beta+1}\left(1-\frac{j}{n}\right)^{\beta}(\log(j/n))^2 = \mathcal{O}(n).$$

The error terms in (2.7) can be treated in a similar way (where $\epsilon > 0$ has to be chosen small enough such that the integral which is used to upper bound the sum is convergent).

Overall, by combining everything, we obtain that

$$2\sum_{j=1}^{n-1} q_n(j)j(\log j)^{\delta+1} = \frac{2\Gamma(2\beta+2)}{\Gamma(\beta+1)^2}\cdot\frac{(\beta+1)\Gamma(\beta+1)^2}{(2\beta+2)\Gamma(2\beta+2)}n(\log n)^{\delta+1}$$
$$+ Kn(\log n)^{\delta} + o(n(\log n)^{\delta})$$
$$= n(\log n)^{\delta+1} + Kn(\log n)^{\delta} + o(n(\log n)^{\delta}), \tag{2.8}$$

where

$$K = (2\delta+2)\frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)^2}\int_0^1 x^{\beta+1}(1-x)^{\beta}\log(x)\mathrm{d}x.$$

In particular, note that $K < 0$.

Now, by plugging the assumption and the induction hypothesis into (2.6) and using (2.8),

$$a_n \leq 2c\sum_{j=1}^{n-1} q_n(j)j(\log j)^{\delta+1} + dn(\log n)^{\delta}$$
$$\leq cn(\log n)^{\delta+1} + cKn(\log n)^{\delta} + c\epsilon n(\log n)^{\delta} + dn(\log n)^{\delta}$$
$$\leq cn(\log n)^{\delta+1},$$

where $\epsilon > 0$ is a sufficiently small constant (coming from the o-term in (2.8)) and the last step holds by choosing c such that $c \geq -d/(K+\epsilon)$ (which is possible since $K < 0$).

For the proof of part (ii), we proceed similarly. First, similar as above,

$$2\sum_{j=1}^{n-1} q_n(j)j^{\gamma}(\log j)^{\delta} = Kn^{\gamma}(\log n)^{\delta} + o(n^{\gamma}(\log n)^{\delta})$$

with

$$K = \frac{2\Gamma(2\beta+2)\Gamma(\beta+\gamma+1)}{\Gamma(\beta+1)\Gamma(2\beta+\gamma+2)} = \prod_{\ell=2}^{\gamma} \frac{\beta+\ell}{2\beta+1+\ell},$$

where in the last step we used that $\Gamma(z+1) = z\Gamma(z)$. Since $\beta > -1$, the above product representation shows that $0 < K < 1$. Thus, again by the induction hypothesis,

$$\begin{aligned} a_n &\leq 2c \sum_{j=1}^{n-1} q_n(j) j^\gamma (\log j)^\delta + dn^\gamma (\log n)^\delta \\ &\leq cKn^\gamma (\log n)^\delta + c\epsilon n^\gamma (\log n)^\delta + dn^\gamma (\log n)^\delta \\ &\leq cn^\gamma (\log n)^\delta, \end{aligned}$$

where $\epsilon > 0$ is as in part (i) and the last step follows by choosing c such that $c \geq d/(1 - K - \epsilon)$ (which is possible since $K < 1$). This concludes the proof of part (ii) and thus also the proposition. ∎

As a consequence, we obtain the following bounds which will be needed in the proof of Theorem 11 (the bounds for the mean also follows from the results by Aldous (1996)).

Corollary 2. *For the Sackin index S_n under the β-splitting model, we have*

$$\mathbb{E}(S_n) = \mathcal{O}(n \log n) \qquad \textit{and} \qquad \mathbb{E}(S_n^2) = \mathcal{O}(n^2 (\log n)^2).$$

Proof. In order to prove this, we need to derive the recurrences satisfied by the mean and the second moment of S_n. To this end, we use the moment-generating function:

$$P_n(u) := \mathbb{E}\left(e^{S_n u}\right).$$

Then, from (2.3),

$$P_n(u) = \sum_{j=1}^{n-1} q_n(j) P_j(u) P_{n-j}(u) e^{nu}, \qquad (n \geq 2)$$

with $P_1(u) = 1$.

Taking the derivative with respect to u and setting $u = 0$, then gives

$$\mathbb{E}(S_n) = 2 \sum_{j=1}^{n-1} q_n(j) \mathbb{E}(S_j) + n, \qquad (n \geq 2)$$

with $\mathbb{E}(S_1) = 0$. This is (2.6) with $b_n = n$. Thus, from Proposition 5, we obtain that $\mathbb{E}(S_n) = \mathcal{O}(n \log n)$ as claimed.

For the second moment, we take the second derivative with respect to u and again set $u = 0$. Then, we see that the second moment also satisfies (2.6) with

$$\begin{aligned} b_n &= n^2 + 2\sum_{j=1}^{n-1} q_n(j)\mathbb{E}(S_j)\mathbb{E}(S_{n-j}) + 4n\sum_{j=1}^{n-1} q_n(j)\mathbb{E}(S_j) \\ &= \mathcal{O}(n^2(\log n)^2), \end{aligned}$$

where the last estimate follows from that for the mean. Thus, again by Proposition 5, we obtain that $\mathbb{E}(S_n^2) = \mathcal{O}(n^2(\log n)^2)$ which concludes the proof. ∎

Corollary 3. *For the depth D_n under the β-splitting model, we have*

$$\mathbb{E}(D_n) = \mathcal{O}(\log n) \qquad \textit{and} \qquad \mathbb{E}(D_n^2) = \mathcal{O}((\log n)^2).$$

Proof. First by the additivity of the expected value, we have

$$\mathbb{E}(S_n) = n\mathbb{E}(D_n)$$

and thus the claimed bound for the mean of the depth follows from that of the mean of the Sackin index which was obtained in the corollary above.

As for the second moment, we again use a moment-generating function:

$$P_n(u) := \mathbb{E}\left(e^{D_n u}\right).$$

Then, from (2.4), we obtain that

$$P_n(u) = 2e^u \sum_{j=1}^{n-1} \frac{j}{n} q_n(j) P_j(u), \qquad (n \geq 2)$$

with $P_1(u) = 1$. Taking the second derivative with respect to u and setting $u = 0$ shows that $n\mathbb{E}(D_n^2)$ satisfies (2.6) with

$$b_n = n + 4\sum_{j=1}^{n-1} q_n(j) j \mathbb{E}(D_j) = \mathcal{O}(n \log n).$$

Thus, Proposition 5 implies that $n\mathbb{E}(D_n^2) = \mathcal{O}(n(\log n)^2)$ from which the claimed result follows. ∎

We also need a lower bound result on a_n satisfying (2.6). The next proposition provides such a result by showing that if b_n is non-negative, then either $a_n \equiv 0$ for all n or otherwise it grows at least linearly.

Proposition 6. *Let a_n be a sequence which satisfies (2.6) with b_n non-negative and $b_{n_0} > 0$ for at least one n_0. Then,*

$$a_n = \Omega(n).$$

Proof. Let n_0 be the smallest positive integer such that $b_{n_0} > 0$. Define

$$\tilde{b}_n = \begin{cases} 0, & \text{if } 1 \leq n \leq n_0; \\ b_n + 2q_n(n_0)b_{n_0}, & \text{if } n \geq n_0 + 1 \end{cases}$$

and denote by $\tilde{a}_n$ the corresponding sequence which satisfies (2.6). Then, clearly $a_n \geq \tilde{a}_n$ and thus it suffices to show that the claim holds for $\tilde{a}_n$. Also, note that

$$\tilde{b}_n \geq 2q_n(n_0)b_{n_0} \geq \frac{K_1}{n^{\beta+1}}, \qquad (n \geq n_0 + 1) \tag{2.9}$$

for some suitable constant $K_1 > 0$ (this follows with similar arguments as in Lemma 6).

Now, we claim that

$$\tilde{a}_n \geq cn, \qquad (n \geq n_0 + 1)$$

with a suitable constant $c > 0$. We will prove this claim by induction where we can safely assume that it holds for sufficiently large n by a suitable choice of c. Plugging now the induction hypothesis into (2.6) gives for n sufficiently large:

$$\begin{aligned} \tilde{a}_n &\geq 2c\sum_{j=1}^{n-1} q_n(j)j - 2c\sum_{j=1}^{n_0} q_n(j)j + \frac{K_1}{n^{\beta+1}} \\ &\geq cn - \frac{cK_2}{n^{\beta+1}} + \frac{K_1}{n^{\beta+1}}, \end{aligned}$$

where we used:

$$2\sum_{j=1}^{n-1} q_n(j)j = n$$

and

$$2\sum_{j=1}^{n_0} q_n(j)j \leq \frac{K_2}{n^{\beta+1}}$$

which is proved as (2.9). Finally, choosing $0 < c \leq K_1/K_2$ gives $\tilde{a}_n \geq cn$ which shows the claim. ∎

From this we obtain the following lower bound for the variance of the fair proportion index which will also be needed in the proof of Theorem 11.

Corollary 4. *For the fair proportion index FP_n under the β-splitting model, we have*

$$\mathrm{Var}(\mathrm{FP}_n) = \Omega(1).$$

Proof. First observe that it was proved by Fuchs and Jin (2015) that

$$\mathbb{E}(\mathrm{FP}_n) = 2 - \frac{2}{n}$$

which follows immediately from the (deterministic) identity

$$\sum_a \mathrm{FP}_\tau(a) = 2n - 2,$$

where the sum runs over all leaves of a phylogenetic tree τ.

In order to find a recurrence for the variance of FP_n, we use the moment-generating function:

$$P_n(u) := \mathbb{E}\left(e^{(\mathrm{FP}_n - \mathbb{E}(\mathrm{FP}_n))u}\right)$$

which by (2.5) satisfies the recurrence

$$P_n(u) = 2\sum_{j=1}^{n-1} \frac{j}{n} q_n(j) P_j(u) e^{\Delta_{n,j}u}, \qquad (n \geq 2)$$

with $P_1(u) = 1$, where $\Delta_{n,j}$ is defined as

$$\Delta_{n,j} = \frac{1}{j} - \mathbb{E}(\mathrm{FP}_n) + \mathbb{E}(\mathrm{FP}_j) = \frac{2}{n} - \frac{1}{j}.$$

Taking the second derivative with respect to u and setting $u = 0$ gives

$$\mathrm{Var}(\mathrm{FP}_n) = 2\sum_{j=1}^{n-1} \frac{j}{n} q_n(j) \mathrm{Var}(\mathrm{FP}_j) + 2\sum_{j=1}^{n-1} \frac{j}{n} q_n(j) \Delta_{n,j}^2.$$

Thus, $n\mathrm{Var}(\mathrm{FP}_n)$ satisfies (2.6) with

$$b_n = 2\sum_{j=1}^{n-1} j q_n(j) \Delta_{n,j}^2$$

which satisfies the assumptions from the Proposition 18. Consequently, $n\mathrm{Var}(\mathrm{FP}_n) = \Omega(n)$ which is the claimed result. ∎

2.2 Difference between Unrooted and Rooted Shapley Value

As stated earlier, Theorem 11 will be a direct consequence of the variance of unrooted Shapley value and fair proportion index tending to 0. Because of the equality of the rooted Shapley value and the fair proportion index, and the two indices possess an almost similar "shape", we instead consider the difference between the unrooted and rooted Shaply values.

Before stating our result, we need to describe the following notations. First, we fix a phylogenetic tree τ. Let τ_ℓ and τ_r denote left and right subtree of the root of τ, respectively. Also, let a be a fixed leaf and assume w.l.o.g. that $a \in \tau_\ell$. For a set of leaves S, we consider the *least common ancestor of* S which is the least common ancestor of S under the order induced by τ where the root is the largest element. In Figure 1.4, the least common ancestor of $\{a, b, c\}$ is the node labelled 2. Define

$$X_\tau^{[i]} := \text{sum of all least common ancestor to root distances for all sets } S \text{ of leaves of size } i$$

and

$$Y_\tau^{[i]}(a) := \text{sum of all distances between the least common ancestors of } S \text{ and } S \cup \{a\} \text{ for all sets } S \text{ of leaves of size } i.$$

The parameter $X_\tau^{[i]}$ is related to the cophenetic value; see (Sokal and Rohlf, 1962).

Proposition 7. *For the difference between unrooted and rooted Shapley value, we have*

$$\begin{aligned}\mathrm{SV}_T^{[u]}(a) - \mathrm{SV}_\tau^{[r]}(a) = &-\frac{1}{n} D_\tau(a) + \frac{1}{n!} \sum_{i=1}^{|\tau_r|} i!(n-i-1)! \left(X_{\tau_r}^{[i]} + \binom{|\tau_r|}{i} \right) \\ &+ \frac{1}{n!} \sum_{i=1}^{|\tau_\ell|-1} i!(n-i-1)! Y_{\tau_\ell}^{[i]}(a), \end{aligned} \tag{2.10}$$

where $D_\tau(a)$ is the distance of a to the root.

Proof. In order to prove the result, we have to compare the difference of $\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\})$ and $\mathrm{PD}_\tau^{[r]}(S) - \mathrm{PD}_\tau^{[r]}(S \setminus \{a\})$ for all sets S of leaves with $a \in S$. Fix such a set S and assume that $S = \{a\} \cup S_\ell \cup S_r$ where S_ℓ are all the leaves of S except a from τ_ℓ and S_r are all the leaves of S from τ_r. We have to distinguish between four cases.

- Case 1: $S_\ell \neq \emptyset$ and $S_r \neq \emptyset$.

  In this case, we have

  $$\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\}) = \mathrm{PD}_\tau^{[r]}(S) - \mathrm{PD}_\tau^{[r]}(S \setminus \{a\})$$

  since the smallest spanning tree of S and $S \setminus \{a\}$ both contain the root. Thus, the contribution of this case to the difference of the two Shapley values is zero.

- Case 2: $S_\ell = S_r = \emptyset$.

  In this case, we have

  $$\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\}) = 0$$

  and

  $$\mathrm{PD}_\tau^{[r]}(S) - \mathrm{PD}_\tau^{[r]}(S \setminus \{a\}) = D_\tau(a)$$

  which gives the first term on the right hand side of (2.10).

- Case 3: $S_\ell = \emptyset$ and $S_r \neq \emptyset$.

  Assume that S_r has size i and least common ancestor c_r. Then, $\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\})$ and $\mathrm{PD}_\tau^{[r]}(S) - \mathrm{PD}_\tau^{[r]}(S \setminus \{a\})$ are explained in Figure 2.2. Moreover, the difference is explained on the left of Figure 2.4. Overall, by summing over all sets of size i, we obtain the second term on the right hand side of (2.10). Note that the term $\binom{|\tau_r|}{i}$ comes from the contribution of the edge from to the root to the right subtree of τ.

- Case 4: $S_\ell \neq \emptyset$ and $S_r = \emptyset$.

  Assume that S_ℓ has size i and least common ancestor c_ℓ. Moreover, denote the least common ancestor of $\{a\} \cup S_\ell$ by c (note that $c = c_\ell$ might be possible). Then, $\mathrm{PD}_\tau^{[u]}(S) - \mathrm{PD}_\tau^{[u]}(S \setminus \{a\})$ and $\mathrm{PD}_\tau^{[r]}(S) - \mathrm{PD}_\tau^{[r]}(S \setminus \{a\})$ are explained in Figure 2.3. Moreover, the difference is explained on the right of Figure 2.4. Overall, by summing over all sets of size i and noting that sets S_ℓ which contain a do not contribute to $Y_{\tau_\ell}^{[i]}(a)$, we obtain the final term on the right hand side of (2.10).

Combining all four cases concludes the proof. ∎

Our result for the difference has to be compared with that from the recent work by Stahn, see (Stahn, 2020), which was useful from a linear algebra poit of view. Our expression, on the other hand, is of combinatorial nature and useful from a computational point of view as will be explained next.

Note that similar to (2.4), we have the following recurrence for $D_\tau(a)$:

$$D_\tau(a) = \begin{cases} D_{\tau_\ell}(a) + 1, & \text{if } a \in \tau_\ell; \\ D_{\tau_r}(a) + 1, & \text{if } a \in \tau_r \end{cases}$$

Moreover, we have also similar recurrences for $X_\tau^{[i]}$ and $Y_\tau^{[i]}(a)$, namely,

$$X_\tau^{[i]} = X_{\tau_\ell}^{[i]} + X_{\tau_r}^{[i]} + \binom{|\tau_\ell|}{i} + \binom{|\tau_r|}{i}$$

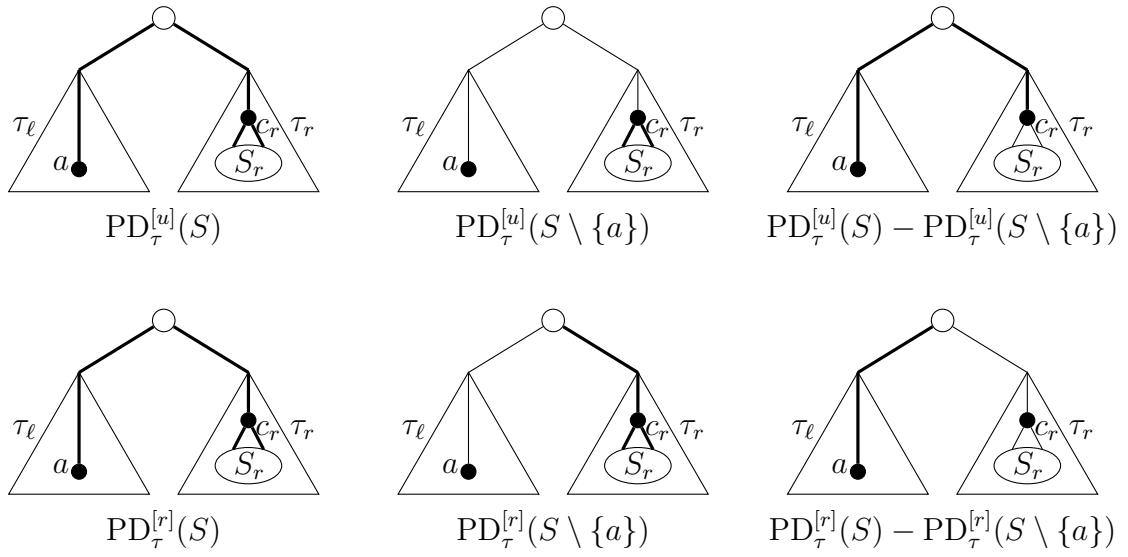


Figure 2.2: Explanation of $\mathrm{PD}_\tau^{[\star]}(S) - \mathrm{PD}_\tau^{[\star]}(S \setminus \{a\})$ for Case 3 in the proof of Proposition 7 where $\star = u$ (first row) or $\star = r$ (second row). The bold parts in the trees visualize the subtree of relevance for the computation of the quantity below the tree.

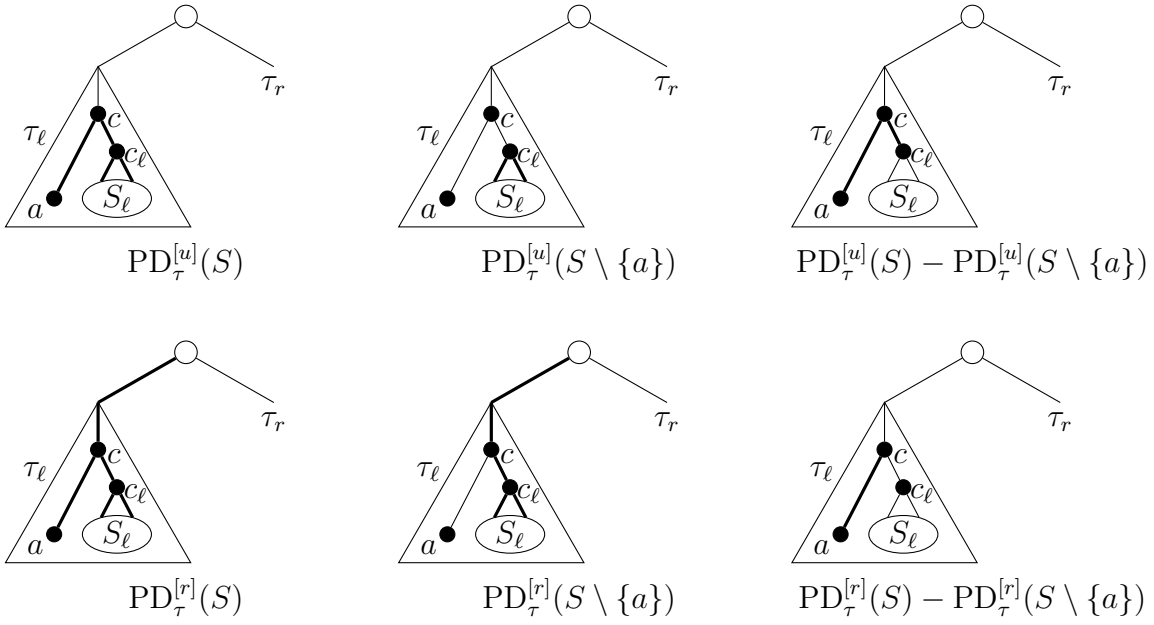


Figure 2.3: Explanation of $\mathrm{PD}_\tau^{[\star]}(S) - \mathrm{PD}_\tau^{[\star]}(S \setminus \{a\})$ for Case 4 in the proof of Proposition 7 where $\star = u$ (first row) or $\star = r$ (second row). The bold parts in the trees visualize the subtree of relevance for the computation of the quantity below the tree.

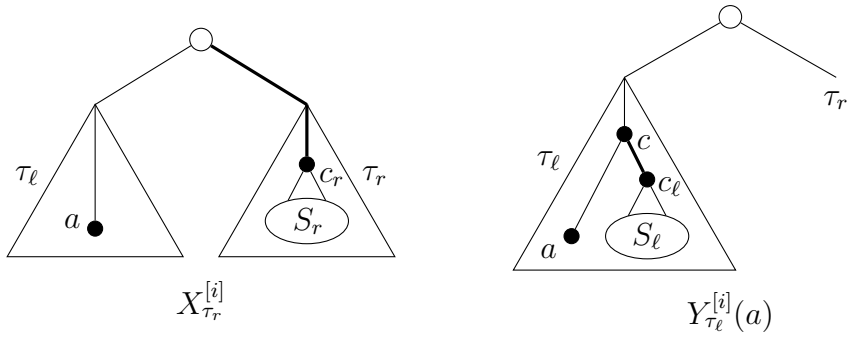


Figure 2.4: The contributions of Case 3 (left) and Case 4 (right) to the expression (2.10).

with $X^{[i]}_\tau = 0$ for all τ with $|\tau| \leq i$ and

$$Y^{[i]}_\tau(a) = \begin{cases} Y^{[i]}_{\tau_\ell}(a) + X^{[i]}_{\tau_r} + \binom{|\tau_r|}{i}, & \text{if } a \in \tau_\ell; \\ Y^{[i]}_{\tau_r}(a) + X^{[i]}_{\tau_\ell} + \binom{|\tau_\ell|}{i}, & \text{if } a \in \tau_r \end{cases}$$

with $Y^{[i]}_\tau(a) = 0$ for all τ with $|\tau| \leq i$. Here, note that in both cases, we only have to consider sets of leaves which are either completely contained in the left or right subtree because all other sets of leaves do not contribute to $X^{[i]}_\tau$ and $Y^{[i]}_\tau(a)$. Moreover, again in both cases, the binomial coefficients count the contribution of the edge connecting the root to the left and/or right right subtree.

These recurrences together with the above proposition lead to a (reasonable fast) recursive method of computing $\mathrm{SV}^{[u]}_\tau(a)$. In particular, this method is faster than doing the computation directly from the definition of $\mathrm{SV}^{[u]}_\tau(a)$. We will use it in order to produce some numerical results in Section 2.4.

2.3 Proof of Theorem 11

In this section, we will prove our main result, namely, that the correlation coefficient of unrooted Shapley value and fair proportion index tends to one. We will start by reducing this task to one which involves the difference between these two indices.

Proposition 8. *If*

$$\mathrm{Var}(\mathrm{SV}^{[u]}_n - \mathrm{FP}_n) = o(1),$$

then

$$\rho(\mathrm{SV}^{[u]}_n, \mathrm{FP}_n) \sim 1.$$

Proof. First consider

$$\begin{aligned}\mathrm{Cov}(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) &= \mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n + \mathrm{FP}_n, \mathrm{FP}_n) \\ &= \mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n, \mathrm{FP}_n) + \mathrm{Var}(\mathrm{FP}_n). \end{aligned} \tag{2.11}$$

Note that the first term in (2.11) can be bounded by Cauchy-Schwartz inequality as

$$\mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n, \mathrm{FP}_n) \leq \sqrt{\mathrm{Var}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n) \cdot \mathrm{Var}(\mathrm{FP}_n)}. \tag{2.12}$$

Also, recall that from Corollary 4,

$$\mathrm{Var}(\mathrm{FP}_n) \geq c,$$

where c is a suitable positive constant. Thus, from this, (2.11), (2.12) and the assumption, we obtain that

$$\mathrm{Cov}(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) = \mathrm{Var}(\mathrm{FP}_n)(1 + o(1)). \tag{2.13}$$

Next, consider

$$\begin{aligned}\mathrm{Var}(\mathrm{SV}_n^{[u]}) &= \mathrm{Cov}(\mathrm{SV}_n^{[u]}, \mathrm{SV}_n^{[u]}) \\ &= \mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n + \mathrm{FP}_n, \mathrm{SV}_n^{[u]} - \mathrm{FP}_n + \mathrm{FP}_n) \\ &= \mathrm{Var}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n) + 2\mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n, \mathrm{FP}_n) + \mathrm{Var}(\mathrm{FP}_n).\end{aligned}$$

Then, by a similar line of reasoning as above

$$\mathrm{Var}(\mathrm{SV}_n^{[u]}) = \mathrm{Var}(\mathrm{FP}_n)(1 + o(1)). \tag{2.14}$$

Finally, by combining (2.13) and (2.14), we have

$$\begin{aligned}\rho(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) &= \frac{\mathrm{Cov}(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n)}{\sqrt{\mathrm{Var}(\mathrm{SV}_n^{[u]})} \cdot \sqrt{\mathrm{Var}(\mathrm{FP}_n)}} \\ &= \frac{\mathrm{Var}(\mathrm{FP}_n)(1 + o(1))}{\sqrt{\mathrm{Var}(\mathrm{FP}_n)(1 + o(1))}\sqrt{\mathrm{Var}(\mathrm{FP}_n)}} = 1 + o(1)\end{aligned}$$

which is the claimed result. ∎

From the previous proposition, it is indeed sufficient to show that the variance of the difference goes to 0. In order to show this, we need the following simple inequality for conditional expectations.

Lemma 8. *Let U be a discrete random variable and $R_1, \ldots, R_k$ be random variables. Then,*

$$\mathbb{E}\left(\left(\sum_{i=1}^{k} R_i\right)^2 | U\right) \leq \left(\sum_{i=1}^{k} \sqrt{\mathbb{E}(R_i^2|U)}\right)^2.$$

Proof. This is an easy consequence of the Cauchy-Schwartz inequality for conditional expectations:

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{i=1}^{k} R_i\right)^2 |U\right) &= \sum_{i,j=1}^{k} \mathbb{E}(R_i R_j|U) \\ &\leq \sum_{i,j=1}^{k} \sqrt{\mathbb{E}(R_i^2|U)}\sqrt{\mathbb{E}(R_j^2|U)} \\ &= \left(\sum_{i=1}^{k} \sqrt{\mathbb{E}(R_i^2|U)}\right)^2.\end{aligned}$$

This proves the claim. ∎

Next, we will simplify the condition of Proposition 8 once more.

We first need some notations. Let the three terms on the right hand side of (2.10) for a random phylogenetic tree of size n under the β-splitting model with $\beta > -1$ and a random taxon a be $Z_n^{[1]}$, $Z_n^{[2]}$ and $Z_n^{[3]}$. Note that Proposition 7 actually gives conditional random variables:

$$\begin{aligned}Z_n^{[1]}|(\mathbf{Y}_n = (j, \text{left})) &= -\frac{D_j + 1}{n}; \\ Z_n^{[2]}|(\mathbf{Y}_n = (j, \text{left})) &= \frac{1}{n!}\sum_{i=1}^{n-j} i!(n-i-1)!\left(X_{n-j}^{[i]} + \binom{n-j}{i}\right); \\ Z_n^{[3]}|(\mathbf{Y}_n = (j, \text{left})) &= \frac{1}{n!}\sum_{i=1}^{j-1} i!(n-i-1)!Y_j^{[i]},\end{aligned}$$

where D_n is the depth (see Section 2.1) and $X_n^{[i]}$ and $Y_n^{[i]}$ denote the random versions of $X_\tau^{[i]}$ and $Y_\tau^{[i]}(a)$. Note that we have replaced D_n by $D_j + 1$ since the random taxon is in the left subtree which has size j; compare with (2.4).

The $\mathbf{Y}_n$ in the expressions above is a random vector with values (j, x) where $1 \leq j \leq n-1$ and $x \in \{\text{left}, \text{right}\}$. Here, j is the size of the left subtree and x gives the location of a. Note that

$$\mathbb{P}(\mathbf{Y}_n = (j, \text{left})) = q_n(j)\frac{j}{n} \tag{2.15}$$

and we have a similar expression if $x = \text{left}$ is replaced by $x = \text{right}$ (also for the $Z_n^{[\ell]}$'s above we have similar expressions in that case).

With these notations, we have the following proposition.

Proposition 9. *If*

$$\mathbb{E}(\mathbb{E}((Z_n^{[\ell]})^2|\mathbf{Y}_n)) = o(1) \text{ for } \ell = 1, 2, 3$$

then

$$\mathrm{Var}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n) = o(1).$$

Proof. First, note that

$$\begin{aligned}\mathrm{Var}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n) &\le \mathbb{E}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n)^2 \\ &= \mathbb{E}(Z_n^{[1]} + Z_n^{[2]} + Z_n^{[3]})^2 \\ &= \mathbb{E}(\mathbb{E}((Z_n^{[1]} + Z_n^{[2]} + Z_n^{[3]})^2 | \mathbf{Y}_n)).\end{aligned}$$

Applying Lemma 8 twice gives

$$\begin{aligned}\mathrm{Var}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n) &\le \mathbb{E}\left(\sum_{\ell=1}^{3} \sqrt{\mathbb{E}((Z_n^{[\ell]})^2 | \mathbf{Y}_n)}\right)^2 \\ &\le \left(\sum_{\ell=1}^{3} \sqrt{\mathbb{E}(\mathbb{E}((Z_n^{[\ell]})^2 | \mathbf{Y}_n))}\right)^2.\end{aligned}$$

From this the claim follows. ∎

The hypothesis for $\ell = 1$ in this proposition is easy to check.

Proposition 10. *The following bound holds:*

$$\mathbb{E}(\mathbb{E}((Z_n^{[1]})^2 | \mathbf{Y}_n)) = \mathcal{O}\left(\frac{(\log n)^2}{n^2}\right).$$

Proof. First,

$$\mathbb{E}((Z_n^{[1]})^2 | \mathbf{Y}_n = (j, \text{left})) = \frac{\mathbb{E}(D_j^2) + 2\mathbb{E}(D_j) + 1}{n^2}.$$

Next, by Corollary 3,

$$\mathbb{E}(D_n) = \mathcal{O}(\log n) \qquad \text{and} \qquad \mathbb{E}(D_n^2) = \mathcal{O}((\log n)^2).$$

Thus,

$$\mathbb{E}((Z_n^{[1]})^2 | \mathbf{Y}_n = (j, \text{left})) = \mathcal{O}\left(\frac{(\log j)^2 + 1}{n^2}\right).$$

A similar expression holds if $x = \text{left}$ is replaced by $x = \text{right}$. Finally, from (2.15),

$$\mathbb{E}(\mathbb{E}((Z_n^{[1]})^2 | \mathbf{Y}_n)) = \mathcal{O}\left(\sum_{j=1}^{n-1} j q_n(j) \frac{(\log j)^2 + 1}{n^3}\right) = \mathcal{O}\left(\frac{(\log n)^2}{n^2}\right)$$

as claimed. ∎

The other two conditions of Proposition 9 are slightly harder to prove. First, we need the following identity which is certainly well-known and can also be verified with, e.g., Maple. However, for the sake of simplicity, we give an easy, elementary, and self-contained proof.

Lemma 9. *The following identity holds:*

$$\sum_{i=1}^{n-j} \frac{\binom{n-j}{i}}{(i+1)\binom{n}{i+1}} = \frac{n-j}{nj}.$$

Proof. First, by the beta integral,

$$\frac{1}{\binom{n}{i+1}} = (n+1)\int_0^1 t^{i+1}(1-t)^{n-i-1}\mathrm{d}t.$$

Then, the above sum becomes

$$\sum_{i=1}^{n-j} \frac{\binom{n-j}{i}}{(i+1)\binom{n}{i+1}} = (n+1)\int_0^1 \sum_{i=1}^{n-j} \frac{\binom{n-j}{i}}{i+1} t^{i+1}(1-t)^{n-i-1}\mathrm{d}t.$$

The sum inside can be simplified as

$$\begin{aligned}
\sum_{i=1}^{n-j} \frac{\binom{n-j}{i}}{i+1} t^{i+1}(1-t)^{n-i-1} &= \int_0^t \sum_{i=1}^{n-j} \binom{n-j}{i} u^i (1-t)^{n-i-1}\mathrm{d}u \\
&= (1-t)^{j-1}\int_0^t \sum_{i=1}^{n-j} \binom{n-j}{i} u^i (1-t)^{n-j-i}\mathrm{d}u \\
&= (1-t)^{j-1}\int_0^t \left((u+1-t)^{n-j} - (1-t)^{n-j}\right)\mathrm{d}u \\
&= \frac{(1-t)^{j-1}}{n-j+1} - \frac{(1-t)^n}{n-j+1} - (1-t)^{n-1}t.
\end{aligned}$$

Plugging this into the integral above, we obtain that

$$\begin{aligned}
\sum_{i=1}^{n-j} \frac{\binom{n-j}{i}}{(i+1)\binom{n}{i+1}} &= (n+1)\int_0^1 \left(\frac{(1-t)^{j-1}}{n-j+1} - \frac{(1-t)^n}{n-j+1} - (1-t)^{n-1}t\right)\mathrm{d}t \\
&= (n+1)\left(\frac{1}{(n-j+1)j} - \frac{1}{(n-j+1)(n+1)} - \frac{1}{n} + \frac{1}{n+1}\right) \\
&= (n+1)\frac{n-j}{(n+1)nj} = \frac{n-j}{nj}.
\end{aligned}$$

This is the desired result. ∎

Now, we can verify the other two conditions from Proposition 9 which then completes the proof of Theorem 11.

Proposition 11. *The following bounds hold:*

$$\mathbb{E}(\mathbb{E}((Z_n^{[2]})^2|\mathbf{Y}_n)) = \begin{cases} \mathcal{O}\left(\frac{(\log n)^2}{n^{\min\{\beta+2,2\}}}\right), & \text{if } \beta \neq 0; \\ \mathcal{O}\left(\frac{(\log n)^3}{n^2}\right), & \text{if } \beta = 0 \end{cases}$$

and

$$\mathbb{E}(\mathbb{E}((Z_n^{[3]})^2|\mathbf{Y}_n)) = \begin{cases} \mathcal{O}\left(\frac{(\log n)^2}{n^{\min\{\beta+1,2\}}}\right), & \text{if } \beta \neq 1; \\ \mathcal{O}\left(\frac{(\log n)^3}{n^2}\right), & \text{if } \beta = 1. \end{cases}$$

Proof. We start with $Z_n^{[2]}$. First note that

$$\begin{aligned} \mathbb{E}((Z_n^{[2]})^2|\mathbf{Y}_n = (j, \text{left})) &= \frac{1}{n!^2}\mathbb{E}\left(\sum_{i=1}^{n-j} i!(n-i-1)!\left(X_{n-j}^{[i]} + \binom{n-j}{i}\right)\right)^2 \\ &\leq \frac{1}{n!^2}\left(\sum_{i=1}^{n-j} i!(n-i-1)!\sqrt{\mathbb{E}\left(X_{n-j}^{[i]} + \binom{n-j}{i}\right)^2}\right)^2, \end{aligned} \tag{2.16}$$

where in the last step we used Lemma 8. What is under the square-root can be written as

$$\mathbb{E}\left(X_{n-j}^{[i]} + \binom{n-j}{i}\right)^2 = \mathbb{E}(X_{n-j}^{[i]})^2 + 2\binom{n-j}{i}\mathbb{E}(X_{n-j}^{[i]}) + \binom{n-j}{i}^2. \tag{2.17}$$

In order to go on, we need the following bound for $X_n^{[i]}$:

$$X_n^{[i]} \leq \frac{1}{i}\binom{n-1}{i-1} S_n, \tag{2.18}$$

where S_n is the Sackin index from Section 2.1. This upper bound is explained as follows: every leaf is contained in $\binom{n-1}{i-1}$ subsets of leaves of size i. Thus,

$$\binom{n-1}{i-1} S_n \tag{2.19}$$

is the sum of taxon-to-root distances for all taxa in all subsets of taxa of size i. Since $X_n^{[i]}$ is the sum of the distance from the least common ancestor to the root of all subsets of taxa of size i, obviously $1/i$-th of (2.19) is at least as large as $X_n^{[i]}$.

Now, plugging (2.18) into (2.17) and using that

$$\mathbb{E}(S_n) = \mathcal{O}(n \log n) \qquad \text{and} \qquad \mathbb{E}(S_n^2) = \mathcal{O}(n^2(\log n)^2),$$

which was obtained in Corollary 2, we have

$$\mathbb{E}\left(X_{n-j}^{[i]} + \binom{n-j}{i}\right)^2 = \mathcal{O}\left(\binom{n-j}{i}^2 (\log n)^2\right).$$

Plugging this in turn into (2.16) gives

$$\mathbb{E}((Z_n^{[2]})^2|\mathbf{Y}_n=(j,\text{left}))=\mathcal{O}\left((\log n)^2\left(\sum_{i=1}^{n-j}\frac{\binom{n-j}{i}}{(i+1)\binom{n}{i+1}}\right)^2\right)=\mathcal{O}\left(\frac{(n-j)^2(\log n)^2}{n^2j^2}\right),$$

where in the last step we used Lemma 9. A similar expression holds if $x=$ left is replaced by $x=$ right.

Finally, using (2.15) gives

$$\begin{aligned}\mathbb{E}(\mathbb{E}((Z_n^{[2]})^2|\mathbf{Y}_n))&=\mathcal{O}\left((\log n)^2\sum_{j=1}^{n-1}q_n(j)\frac{(n-j)^2}{n^3j}\right)\\&=\mathcal{O}\left(n^{-2\beta-4}(\log n)^2\sum_{j=1}^{n-1}j^{\beta-1}(n-j)^{\beta+2}\right),\end{aligned}$$

where in the last step we used Lemma 6. From this the claimed result follows from the bounds

$$\sum_{j=1}^{n-1}j^{\beta-1}(n-j)^{\beta+2}=\begin{cases}\mathcal{O}(n^{\beta+2}), & \text{if } \beta<0;\\ \mathcal{O}(n^2\log n), & \text{if } \beta=0;\\ \mathcal{O}(n^{2\beta+2}), & \text{if } \beta>0.\end{cases}$$

Next, for $Z_n^{[3]}$, the same method as above can be used since it trivially holds that

$$Y_n^{[i]}\le X_n^{[i]}.$$

In particular, we obtain that

$$\mathbb{E}((Z_n^{[3]})^2|\mathbf{Y}_n=(j,\text{left}))=\mathcal{O}\left(\frac{j^2(\log n)^2}{n^2(n-j)^2}\right)$$

and a similar result holds if $x=$ left is replaced by $x=$ right.

Thus, again by (2.15), we have

$$\begin{aligned}\mathbb{E}(\mathbb{E}((Z_n^{[3]})^2|\mathbf{Y}_n))&=\mathcal{O}\left((\log n)^2\sum_{j=1}^{n-1}q_n(j)\frac{j^3}{n^3(n-j)^2}\right)\\&=\mathcal{O}\left(n^{-2\beta-4}(\log n)^2\sum_{j=1}^{n-1}j^{\beta+3}(n-j)^{\beta-2}\right)\end{aligned}$$

from which the claim follows by the bounds

$$\sum_{j=1}^{n-1}j^{\beta+3}(n-j)^{\beta-2}=\begin{cases}\mathcal{O}(n^{\beta+3}), & \text{if } \beta<1;\\ \mathcal{O}(n^4\log n), & \text{if } \beta=1;\\ \mathcal{O}(n^{2\beta+2}), & \text{if } \beta>1.\end{cases}$$

This concludes the proof. ∎

2.4 Numerical Data

In this section, we present some numerical data to illustrate Theorem 11. For this data, we used the splitting probabilities (1.15) to generate a random phylogenetic tree. Then, we picked a leaf uniformly at random from all taxa and computed the corresponding pair of unrooted Shapley value and fair proportion index. This was repeated five hundred times for each fixed choice of β and n. As for β, we chose $\beta = 0$ (Yule-Harding model), $\beta = -1/2$ and $\beta = -1$; for n, we chose $40, 80$ and 160.

The computation of the unrooted Shapley value and fair proportion index was done recursively as outlined in Section 2.2. The recursions for $\mathrm{FP}_T(a)$ and $D_T(a)$, $X_T^{[i]}$ and $Y_T^{[i]}(a)$ yield a sufficiently fast method for doing the computation. Our results can be found in Figure 2.5 and the code is available at

https://github.com/arpaningbatan/ShapleyValue.

Note that the case $\beta = -1$ is not covered by Theorem 11. Indeed, one sees that whereas the convergence of the data to the line $y = x$ is very good for $\beta = 0$, it seems to slow down if $\beta = -1/2$ and $\beta = -1$. In fact, this is in accordance with the proof method of Theorem 11 from the previous section which also gives a bound of the speed of convergence to 1. This bound is dominated by the second bound in Proposition 11 which gets worse as β approaches -1. However, our bound might be too conservative and it might be the case that the correlation also converges to 1 when $\beta = -1$. (Figure 2.5 seems to suggest that concentration on the line $y = x$ also takes place in this case.)

Also, note that points with small fair proportion index tend to be above the line $y = x$ whereas points with large fair proportion index tend to be below the line $y = x$. This is explained by (2.10) since the only negative term on the right hand side is also small in the former case and large in the latter case.

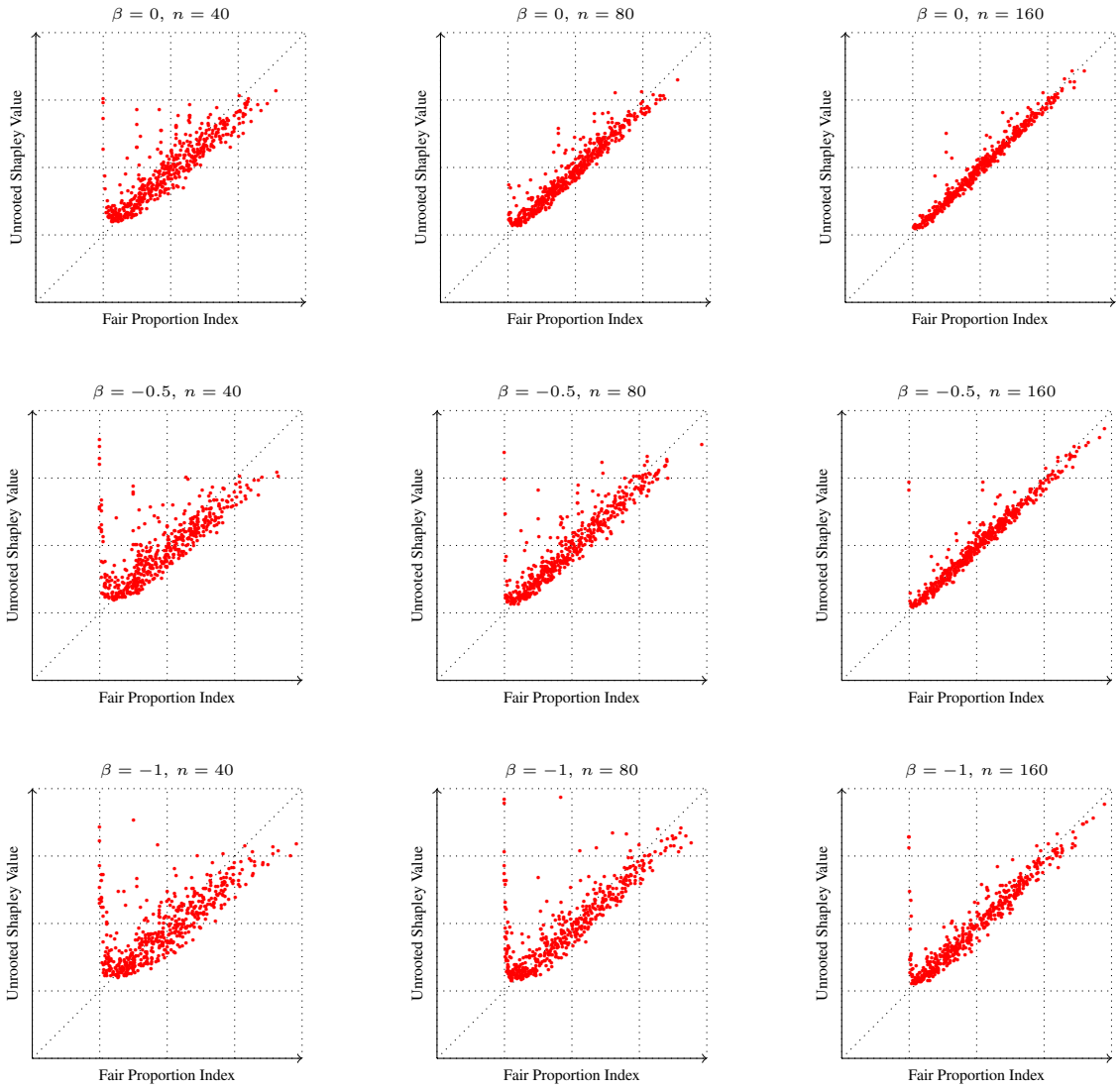


Figure 2.5: Numerical data for $\beta = 0$ (Yule-Harding model), $\beta = -1/2$ and $\beta = -1$.

Chapter 3

Animal Grouping

Group formation is a natural process in an ecosystem which can easily be seen in packs of social animals. Many studies suggest that kinship plays a vital role in the group formation process. Others suggest that organisms form groups to acquire benefits in the group such as hunting of predators, group protection from herds, preservation of hereditary traits, etc. Understanding the behaviours of animals and finding the reasons why the group formation process takes place are some of the problems being studied by biologist. Many models have been proposed to emulate group formation processes such as game theoretic models, aggregation and splitting models and kin-selection processes; see (Durand et al., 2007) for a detailed discussion.

In (Durand et al., 2007), the authors proposed a model in which group formation is based on the genetic relatedness of animals. This model is called the *neutral model* because of the use of neutral gene trees, which in our case are the phylogenetic trees from Section 1.1, to represent genetic relatedness of organisms. In this model, the individuals must satisfy the following conditions to form a group:

1. Each group must have more than one member.

2. Each individual must be grouped to its nearest kin.

The first condition is a natural condition since we are trying to form a group. Meanwhile, the second condition takes care of the genetic relatedness in the group formation process. The minimal groups satisfying the two conditions are called the *minimal clades* or the *clades* of the phylogenetic tree. In Figure 3.1A, $\{a, b\}$ can form a group but they are the nearest kin of c and cannot form a group by themselves. Therefore, $\{a, b, c\}$ forms a clade. Notice that the group $\{f, g, h, i\}$ satisfies both conditions but it is not minimal. Thus, it is not a clade. The clades in

Figure 3.1A are $\{a, b, c\}$, $\{f, g\}$ and $\{h, i\}$.

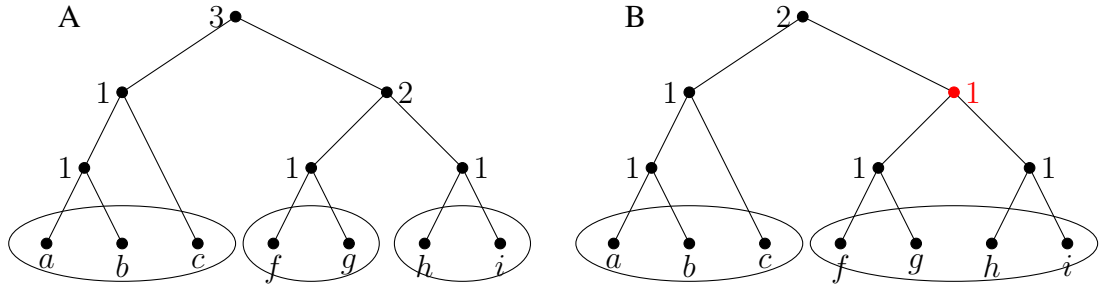


Figure 3.1: Recursive computation of the number of clades of a phylogenetic tree which represents the genetic relationship between 7 animals $\{a, b, c, f, g, h, i\}$. Extra-clustering event occurs at the red node in Figure B.

Due to the difficulty in constructing the accurate phylogenetic tree relationhip between animals, we may again consider the random tree models discussed in Section 1.4. Many parameters such as number of clades and clade spectrum under the Yule-Harding model have been studied; see (Durand and François, 2010; Drmota et al., 2014, 2016). The authors mainly used the distributional recurrence (see Section 2.1) of these parameters to derive the limit distributions. In contrast to their method, we will be using the combinatorial properties of the trees to derive their limit distributions. Thus, our study will focus on the uniform model on phylogenetic trees rather than the Yule-Harding model.

The following parameters will be of our interest in this section. First, we will again take another look at the number of clades and the number of clades of size m which we will denote by N_n and $N_n^{[m]}$, respectively. In Figure 3.1A, the values of these parameters are $N_7 = 3$, $N_7^{[2]} = 2$, $N_7^{[3]} = 1$, $N_7^{[4]} = N_7^{[5]} = N_7^{[6]} = N_7^{[7]} = 0$. In passing, we will observe that the number of groups is very small. This leads us to assume that there must be a clade with large size. Thus, we also consider the size of the largest clade which we denote by M_n. In Figure 3.1A, we have $M_7 = 3$. Notice that these parameters are dependent only on the shape of the tree and thus, by exploiting Lemma 1, we may consider the plane binary trees instead of the phylogenetic trees.

The neutral model relies only on the genetic relatedness of animals for group formation. This may be true for some classes of social animals but not for all, see (Durand et al., 2007) for real-world statistical data. In order to take into account other events that lead to the formation of groups, the authors in (Durand et al., 2007) introduced a parameter p, called the *clustering*

rate, which gives a probability of group formation in every splitting event in the plane binary tree; compare with the recursive way the uniform model arises from the β-splitting model of Section 1.4. Thus, in case that such an event occurs, all animals under that event form a group. This more general model is called the *extra-clustering model*. Note that the neutral model is a particular case of the extra-clustering model which can be obtained by taking $p = 0$. In Figure 3.1B, an extra-clustering event happens at the red node. Then, we have $N_7 = 2$, $N_7^{[2]} = N_7^{[5]} = N_7^{[6]} = N_7^{[7]} = 0$, $N_7^{[3]} = 1$, $N_7^{[4]} = 1$ and $M_7 = 4$. The case $p = 1$ will not be considered in this thesis since the results under this case are trivial. The only possibility in this case is that all animals are grouped together in a single clade.

The remainder of the chapter is organized as follows. In the next section, we will introduce the *cluster trees* and two important generating functions that are associated with it. The proofs of the results in this Chapter will heavily rely on these generating functions. Next, the limit distribution of N_n and $N_n^{[m]}$ will be derived in Section 3.2. Finally, in Section 3.3 we will derive for the moments and limit distribution of M_n.

3.1 Cluster Trees and Weights

First, note that the definition of the extra-clustering model can be broken into two probabilistic stages:

1. a plane binary tree of size n is picked uniformly at random and

2. the picked tree is traced (starting from the root and then recursively in the subtrees) and one stops if either a node is encountered whose left or right subtree is a leaf (parent of the leaves under the same clade) or an extra-clustering event has occurred.

In the second step, we replace the subtrees at the places where one has stopped by leaves. The resulting tree is called a *cluster tree* of the picked tree. Note that cluster trees are again plane binary trees and the leaves of the cluster tree corresponds to the clades in the extra-clustering process. Moreover, note that they are not unique but rather depend on the outcome of the probabilistic procedure in the extra clustering process, see Figure 3.2. Figure 3.2B happens when there is no extra clustering in the root and the tracing stops at the red internal nodes of the original tree with probability $pq^2 + 2p^2q + q^3$ (with probability pq^2 if extra-clustering events occured at both red internal nodes, twice p^2q if an extra-clustering event occured at exactly one

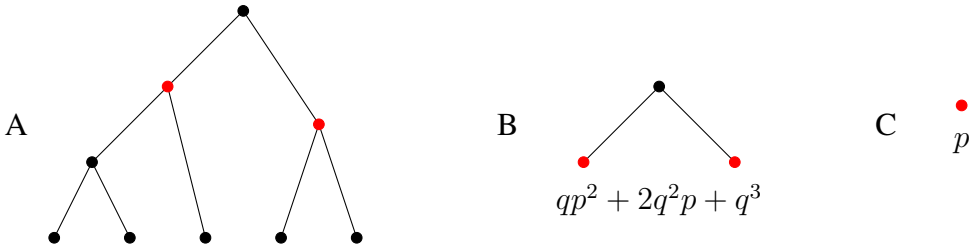


Figure 3.2: Plane binary tree of size 5 together with all its possible cluster trees in B and C with corresponding probabilities.

of the grey internal nodes, or q^3 if no extra-clustering event occured at red internal nodes). On the other hand, Figure 3.2C happens only when there is an extra-clustering event in the root with probability p.

Now, in order to keep track of the probabilities attached to cluster trees, we associate two generating functions with them. First, since no extra-clustering event has occurred at any internal node of a cluster tree, we attach the probability $q := 1 - p$ to these nodes, i.e., we consider

$$G(z) := \sum_{n \geq 1} q^{n-1} C_{n-1} z^n = zC(qz),$$

where $C(z)$ is the generating function of the Catalan numbers. Next, for the leaves of the cluster tree, they either resulted from one of the following:

1. an extra-clustering event occurred at the leaf,
2. exactly one of its children is a leaf, or
3. both of its children are leaves.

In the first case, we attach a weight p and note that there are C_{n-1} such possible trees. Next, in the second case, we attach a weight q and there are C_{n-2} possible trees because either the left or right subtree is a leaf. Finally, in the last case, we attach a weight q and there is only one possible tree. Thus, for each single leaf of the cluster tree, we consider the generating function

$$\begin{aligned} H(z) :&= (pC_1 + q)z^2 + \sum_{n \geq 3} (pC_{n-1} + 2qC_{n-2}) z^n \\ &= z^2 + pz(C(z) - 1 - z) + 2qz^2(C(z) - 1). \end{aligned}$$

Note that the composition of these two generating functions, namely $G(H(z))$, generates for any plane binary trees all its associated cluster trees with their corresponding probabilities.

In particular, since for each plane binary tree the probabilities of its cluster trees sum up to 1, we have

$$[z^n]G(H(z)) = C_{n-1}, \qquad (n \geq 2)$$

and $[z^1]G(H(z)) = 0$ because all plane binary trees have at least two leaves. We formulate this as a lemma.

Lemma 10. *For all* $0 \leq p < 1$*, we have* $G(H(z)) = z(C(z) - 1)$.

The case $p = 1$ is trivial since we have a clustering at the root

3.2 Number of Clades and Number of Fixed-size Clades

We now consider the parameter N_n. We first track the number of leaves of the cluster trees by marking each occurrence of $H(z)$ by a variable u. Then, we have a composition schema $G(uH(z))$. From the discussions in Section 1.3, the distribution of N_n is given by

$$\mathbb{P}(N_n = k) = \frac{[u^k z^n]G(uH(z))}{C_{n-1}}.$$

Using the notations from Theorem 10, note that

$$\rho_H = \frac{1}{4}, \rho_G = \frac{1}{4q} \qquad \text{and} \qquad \tau_H = \frac{3+p}{16}, \ \tau_G = \frac{1}{2q}.$$

Moreover, the subcritical condition is satisfied since

$$\tau_H = \frac{3+p}{16} < \frac{1}{4} \leq \frac{1}{4q} = \rho_G.$$

In addition, $H(z)$ has expansion about $\rho_H = 1/4$ which is given by

$$H(z) = \frac{3+p}{16} - \frac{1+p}{4}\sqrt{1-4z} + o(\sqrt{1-4z})$$

in a suitable Δ-domain. Thus, we can apply Theorem 10 from Section 1.3 to obtain the following result.

Theorem 12. *We have the limit distribution result*

$$N_n \xrightarrow{d} N$$

with convergence of all moments, where

$$N \stackrel{d}{=} \mathrm{NB}\left(\frac{1}{2}, \frac{3-2p-p^2}{4}\right) + 1.$$

Here, $\mathrm{NB}(r, p)$ *denotes the negative binomial distribution.*

Proof. By applying Theorem 10, we obtain the claimed result with the probability generating function of N given by

$$P_N(u) = u\sqrt{\frac{1-4q\tau_H}{1-4q\tau_H u}}.$$

Moreover, by the above form of the probability generating function of N, it is clear that N has the claimed distribution. ∎

Remark 1. $\mathrm{NB}(r,p)$ in the above theorem is more precisely the (standard) generalization of the negative binomial distribution to the case where the first parameter is allowed to be any positive real number. The mean of such distribution is given by

$$\mu = \frac{pr}{1-p}$$

and probability generating function

$$P_{NB(r,p)}(z) = \left(\frac{1-p}{1-pz}\right)^r.$$

As a direct consequence, we have the following corollary.

Corollary 5. *We have,*

$$\mathbb{E}(N_n) \sim \frac{5+2p+p^2}{2+4p+2p^2}.$$

Thus, on average, there are only a finite number of groups.

Now, we consider the number of clades with fixed-size. We fix $m \geq 2$ and consider the number of clades of size m which we denoted by $N_n^{[m]}$. We again use the two generating functions $G(z)$ and $H(z)$ to study this parameter. In this case, we only mark with the variable u those leaves of the cluster tree which corresponds to the groups of size m, that is, only the coefficient of z^m. Note that $[z^m]H(z)$ is given by

$$pC_{m-1} + (2-\delta_{2,m})qC_{m-2}$$

were $\delta_{2,m}$ is the Kronecker delta function. Thus, by marking the m-th coefficient of $H(z)$ by u, we have

$$G((pC_{m-1} + (2-\delta_{2,m})qC_{m-2})(u-1)z^m + H(z)).$$

Then,

$$\mathbb{P}(N_n^{[m]} = k) = \frac{[u^k z^n]G((pC_{m-1} + (2-\delta_{2,m})qC_{m-2})(u-1)z^m + H(z))}{C_{n-1}}.$$

Unfortunately, we can not apply Theorem 10 directly since the function is not of the form $G(uH(z))$. However, the method of proof of Theorem 10 can be applied and yields the following result.

Theorem 13. *We have the limit distribution result*

$$N_n^{[m]} \xrightarrow{d} N^{[m]}$$

with convergence of all moments, where

$$N^{[m]} \stackrel{d}{=} \mathrm{NB}\left(\frac{1}{2}, \frac{4^{2-m}(1-p)(pC_{m-1}+(1-p)(2-\delta_{2,m})C_{m-2})}{1+2p+p^2+4^{2-m}(1-p)(pC_{m-1}+(1-p)(2-\delta_{2,m})C_{m-2})}\right).$$

Proof. Let

$$H_m(u,z) = (pC_{m-1}+(2-\delta_{2,m})qC_{m-2})(u-1)z^m + H(z)$$

which has dominant singularity at $z = 1/4$. By a straightforward expansion, as $z \to 1/4$,

$$H_m(u,z) = c_m(u) - \frac{1+p}{4}\sqrt{1-4z} + o(\sqrt{1-4z}),$$

where

$$c_m(u) = (pC_{m-1}+(2-\delta_{2,m})qC_{m-2})(u-1)4^{-m} + \frac{3+p}{16}.$$

Note that for u close to 1, we have

$$|c_m(u)| < \frac{1}{4q}$$

and the upper bound is the dominant singularity of $G(z)$. Thus, $G(H_m(u,z))$ has also dominant singularity at $z = 1/4$. Moreover, as $z \to 1/4$,

$$G(H_m(u,z)) = \frac{1-\sqrt{1-4qc_m(u)}}{2q} - \frac{1+p}{4\sqrt{1-4qc_m(u)}}\sqrt{1-4z} + o(\sqrt{1-4z}).$$

Now, by singularity analysis,

$$[z^n]G(H_m(u,z)) \sim \frac{1+p}{8\sqrt{\pi}\sqrt{1-4qc_m(u)}} \cdot \frac{4^n}{n^{3/2}}$$

and by using the expansion of the Catalan numbers (see (1.10))

$$C_n = \frac{4^n}{\sqrt{\pi}n^{3/2}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right), \tag{3.1}$$

we obtain that

$$P_{N_n^{[m]}}(u) \sim \frac{1+p}{2\sqrt{1-4qc_m(u)}},$$

where $P_{N_n^{[m]}}(u)$ denotes the probability generating function of $N_n^{[m]}$. From this the claimed result follows from Theorem 9. ∎

As a consequence, we again obtain the asymptotics of the mean.

Corollary 6. *We have,*

$$\mathbb{E}(N_n^{[m]}) \sim 2\frac{4^{1-m}(1-p)(pC_{m-1}+(1-p)(2-\delta_{2,m})C_{m-2})}{1+2p+p^2}.$$

Corollary 5 and Corollary 6 now imply Proposition 12 below. But first, we prove a lemma that will help us to prove this proposition.

Lemma 11. *For $k \geq 3$, we have*

$$\sum_{m=3}^{k} 4^{-m}(pC_{m-1}+2qC_{m-2}) = \frac{p+2}{16} - \frac{(kp+k-p)(2k-2)!}{4^k(k-1)!k!}. \tag{3.2}$$

Proof. Let $S(k) := \dfrac{p+2}{16} - \dfrac{(kp+k-p)(2k-2)!}{4^k(k-1)!k!}$. Then we have

$$\begin{aligned}
S(k+1)-S(k) &= \frac{(kp+k-p)(2k-2)!}{4^k(k-1)!k!} - \frac{(kp+k+1)(2k)!}{4^{k+1}k!(k+1)!} \\
&= \frac{(2k-2)!}{4^k k!(k-1)!}\left(kp+k-p-\frac{(kp+k+1)(2k-1)}{2(k+1)}\right) \\
&= \frac{(2k-2)!}{4^k k!(k-1)!}\left(\frac{kp+k-2p+1}{2(k+1)}\right) \\
&= \frac{(2k-2)!}{4^k k!(k-1)!}\left(\frac{p(2k-1)}{2(k+1)} - \frac{p(2k-1)}{2(k+1)} + \frac{kp+k-2p+1}{2(k+1)}\right) \\
&= \frac{(2k-2)!}{4^k k!(k-1)!}\left(\frac{p(2k-1)}{2(k+1)} + \frac{1-p}{2}\right) \\
&= 4^{-k-1}(pC_k + 2qC_{k-1}).
\end{aligned}$$

The result immediately follows by summation. ∎

Now, we ready to state and prove the following proposition.

Proposition 12. *We have,*

$$\mathbb{E}(N) = 1 + \sum_{m\geq 2} \mathbb{E}(N^{[m]}).$$

Proof. Notice that if we take the limit of (3.2) as $k \to \infty$, we have

$$\sum_{m\geq 3} 4^{-m}(pC_{m-1}+2qC_{m-2}) = \frac{p+2}{16}.$$

Thus, we have

$$\begin{aligned}
1 + \sum_{m\geq 2} \mathbb{E}(N^{[m]}) &= 1 + \frac{1-p}{2(1+2p+p^2)} + \frac{(1-p)(p+2)}{2(1+2p+p^2)} \\
&= \frac{5+2p+p^2}{2+4p+2p^2}.
\end{aligned}$$

This proves the desired result. ∎

Corollary 5 already tells us that there are only few clades since

$$\frac{5+2p+p^2}{2+4p+2p^2} < \frac{5}{2}$$

for any $p \geq 0$. This suggest that there is a clade with large size. In Proposition 12, the left-hand side of the equation suggests that there is only one big clade (represented by the term 1) and all other clades are small (represented by the summation). This is indeed the case and will be proven in the next section.

3.3 Largest Clade Size

Finally, we study the last parameter we are interested in in this section which is the size of the largest clade which we denoted by M_n. Due to the observation from the last section that there should be one big group, we set $X_n := n - M_n$.

In order to find the distribution of X_n, we again make use of the two generating functions from Section 3.1 for the cluster tree. The main observation is that for $0 \leq k < n/2$, we have

$$\mathbb{P}(X_n = k) = \frac{[z^k]G'(H(z))[z^{n-k}]H(z)}{C_{n-1}}.$$

This is explained as follows. Since the largest clade size is equal to $n-k$, we have to replace one leaf of the cluster tree by a group of size $n-k$. This corresponds to the factor $[z^{n-k}]H(z)$. Then, all other leaves of the cluster tree are replaced by arbitrary groups which corresponds to the other factor $[z^k]G'(H(z))$. The restriction $0 \leq k < n/2$ is essential here, because it ensures that all other groups are indeed of size smaller than $n-k$. Moreover, the range $0 \leq k < n/2$ is expected to be sufficient for our purpose since we expect that the largest group size is close to n.

We start with the following lemma.

Lemma 12. *Uniformly for $0 \leq k < n/2$, we have*

$$\mathbb{P}(X_n = k) = \frac{1+p}{2} 4^{-k} [z^k]G'(H(z)) \left(1 - \frac{k}{n}\right)^{-3/2} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Proof. Note that

$$[z^{n-k}]H(z) = pC_{n-k-1} + 2qC_{n-k-2}.$$

The result follows directly by using the expansion of the Catalan numbers in (3.1). ∎

From the last lemma, we obtain the limit distribution of X_n.

Theorem 14. *We have the limit distribution result*

$$X_n \xrightarrow{d} X,$$

where X is a discrete random variable with probability generating function

$$P_X(u) = \sum_{k\geq 0} p_k u^k = \frac{1+p}{2F(u/4)}.$$

Here,

$$F(u) = \sqrt{1-2p+2p^2-4(1-2p)(1-p)u+4(1-p)^2u^2-2(1-p)(p-2(1-p)u)\sqrt{1-4u}}. \tag{3.3}$$

Proof. From Lemma 12, we have for fixed k

$$p_k := \lim_{n\to\infty} \mathbb{P}(X_n = k) = \frac{1+p}{2} 4^{-k}[z^k]G'(H(z)).$$

Thus,

$$P_X(u) = \sum_{k\geq 0} p_k u^k = \frac{1+p}{2} G'(H(u/4))$$

and the claimed form follows now by plugging into this the expressions for $G(z)$ and $H(z)$ and straightforward computation. ∎

Note that $F(u)$ has dominant singularity at $u = 1/4$. Moreover, as $u \to 1/4$,

$$P_X(u) = 1 - \frac{2(1-p)}{1+p}\sqrt{1-u} + o(\sqrt{1-u}).$$

From this and Theorems 5 and 7,

$$p_k = \frac{1-p}{(1+p)\sqrt{\pi}k^{3/2}}\left(1+\mathcal{O}\left(\frac{1}{k}\right)\right), \qquad (k\to\infty). \tag{3.4}$$

Note that all moments of X are infinite. Thus, we cannot apply Theorem 9 to have moment convergence in the above limit theorem for the largest clade size, in contrast to Theorem 12 and Theorem 13.

Due to the latter remark, it is interesting to compute moments of X_n (and consequently of M_n). We will do this next with the help of Lemma 12 and (3.4). To prove the next theorem, we first need the following crucial lemma.

Lemma 13. *We have,*

$$\sum_{0\leq k<n/2} \mathbb{P}(X_n = k) = 1 + o(n^{-1/2}) \tag{3.5}$$

and for $\ell \geq 1$

$$\sum_{0\leq k<n/2} k^{\ell}\mathbb{P}(X_n = k) \sim d_{\ell} n^{\ell-1/2} \tag{3.6}$$

where

$$d_{\ell} = \frac{1-p}{(1+p)\sqrt{\pi}} \int_0^{1/2} x^{\ell-3/2}(1-x)^{-3/2}\mathrm{d}x.$$

Proof. We will derive the asymptotics of the sum in (3.5) by splitting it into two parts:

$$\sum_{0\leq k<n/2} \mathbb{P}(X_n = k) = \sum_{0\leq k<n^{\rho}} \mathbb{P}(X_n = k) + \sum_{n^{\rho}\leq k<n/2} \mathbb{P}(X_n = k), \tag{3.7}$$

where $\rho > 0$ will be chosen as the proof proceeds.

For the first part, we have by Lemma 12,

$$\sum_{0\leq k<n^{\rho}} \mathbb{P}(X_n = k) = \sum_{0\leq k<n^{\rho}} p_k(1+\mathcal{O}(n^{\rho-1})) = \sum_{0\leq k<n^{\rho}} p_k(1+o(n^{-1/2})),$$

where p_k was defined in Theorem 14 and $\rho < 1/2$ so that the last equality holds. Note that

$$\sum_{0\leq k<n^{\rho}} p_k = 1 - \sum_{k\geq n^{\rho}} p_k = 1 - \frac{1-p}{(1+p)\sqrt{\pi}} \sum_{k\geq n^{\rho}} k^{-3/2}(1+\mathcal{O}(1/k)),$$

where we used (3.4) in the last step. Combining the two equations above, we get

$$\sum_{0\leq k<n^{\rho}} \mathbb{P}(X_n = k) = 1 - \frac{1-p}{(1+p)\sqrt{\pi}} \sum_{k\geq n^{\rho}} k^{-3/2}(1+\mathcal{O}(1/k)) + o(n^{-1/2}). \tag{3.8}$$

The asymptotic of the sum on the right-hand side of the equation can be derived by using the Euler-Maclaurin summation formula:

$$\sum_{k\geq n^{\rho}} k^{-3/2} = \int_{n^{\rho}}^{\infty} x^{-3/2}\mathrm{d}x + \mathcal{O}(n^{-3\rho/2}) = 2n^{-\rho/2} + o(n^{-1/2}),$$

where the last step holds whenever $\rho > 1/3$. The asymptotic of the $\mathcal{O}$-term in (3.8) can be derived in a similar manner and we have

$$\sum_{k\geq n^{\rho}} \mathcal{O}(k^{-5/2}) = \mathcal{O}(n^{-3\rho/2}) = o(n^{-1/2}),$$

where the last equality holds for $\rho > 1/3$. Thus, we obtain that

$$\sum_{0\leq k<n^{\rho}} \mathbb{P}(X_n = k) = 1 - \frac{2(1-p)}{(1+p)\sqrt{\pi}} n^{-\rho/2} + o(n^{-1/2}). \tag{3.9}$$

Now, we turn to the second part of the decomposition of (3.7) for which we use the expansions from Lemma 12 and (3.4):

$$\sum_{n^{\rho}\leq k<n/2} \mathbb{P}(X_n = k) = \frac{1-p}{(p+1)\sqrt{\pi}} \sum_{n^{\rho}\leq k<n/2} k^{-3/2}(1-k/n)^{-3/2}(1+\mathcal{O}(1/k)). \tag{3.10}$$

Using again Euler-Maclaurin summation formula,

$$\sum_{n^\rho \le k < n/2} k^{-3/2}(1-k/n)^{-3/2} = \int_{n^\rho}^{n/2} x^{-3/2}(1-x/n)^{-3/2}\mathrm{d}x + o(n^{-1/2}).$$

Note that

$$\int x^{-3/2}(1-x/n)^{-3/2}\mathrm{d}x = \frac{2(2x-n)}{\sqrt{nx(n-x)}}$$

and thus

$$\sum_{n^\rho \le k < n/2} k^{-3/2}(1-k/n)^{-3/2} = 2n^{-\rho/2} + o(n^{-1/2}).$$

Together with a similar treatment of the $\mathcal{O}$-term in (3.10), we obtain that

$$\sum_{n^\rho \le k < n/2} \mathbb{P}(X_n = k) = \frac{2(1-p)}{(1+p)\sqrt{\pi}} n^{-\rho/2} + o(n^{-1/2}). \tag{3.11}$$

Finally, substituting (3.9) and (3.11) into (3.7) gives the desired result.

Next, we proceed to the proof of (3.6). In a similar manner, we split the sum into

$$\sum_{0 \le k < n/2} k^\ell \mathbb{P}(X_n = k) = \sum_{0 \le k < n^\rho} k^\ell \mathbb{P}(X_n = k) + \sum_{n^\rho \le k < n/2} k^\ell \mathbb{P}(X_n = k), \tag{3.12}$$

where ρ is again chosen as the proof proceed.

For the first term on the right-hand side of (3.12):

$$\sum_{0 \le k < n^\rho} k^\ell \mathbb{P}(X_n = k) \le n^{\rho\ell} = o(n^{\ell - 1/2}),$$

where the last step holds when $\rho < 1/2$.

For the second term on the right-hand side of (3.12), we again apply the expansions in Lemma 12 and (3.4):

$$\sum_{n^\rho \le k < n/2} k^\ell \mathbb{P}(X_n = k) = \frac{1-p}{(1+p)\sqrt{\pi}} \sum_{n^\rho \le k < n/2} k^{\ell - 3/2}(1-k/n)^{-3/2}(1+\mathcal{O}(1/k)). \tag{3.13}$$

Using once more Euler-Maclaurin summation formula yields

$$\begin{aligned}
\sum_{n^\rho \le k < n/2} k^{\ell-3/2}(1-k/n)^{-3/2} &= \int_{n^\rho}^{n/2} x^{\ell-3/2}(1-x/n)^{-3/2}\mathrm{d}x + o(n^{\ell-1/2}) \\
&= \int_{0}^{n/2} x^{\ell-3/2}(1-x/n)^{-3/2}\mathrm{d}x + o(n^{\ell-1/2}) \\
&= \left(\int_0^{1/2} x^{\ell-3/2}(1-x)^{-3/2}\mathrm{d}x \right) n^{\ell-1/2} + o(n^{\ell-1/2}).
\end{aligned}$$

The $\mathcal{O}$-term in (3.13) is treated similarly.

Finally, substituting the above two equations into (3.12) gives the desired result. ∎

From this lemma, we obtain now the asymptotics of all moments of X_n.

Theorem 15. *For $\ell \geq 1$, we have*

$$\mathbb{E}(X_n^\ell) \sim d_\ell n^{\ell-1/2},$$

where d_ℓ is as in Lemma 13.

Proof. Since

$$\mathbb{E}(X_n^\ell) = \sum_{0\leq k\leq n} k^\ell \mathbb{P}(X_n = k) = \sum_{0\leq k<n/2} k^\ell \mathbb{P}(X_n = k) + \sum_{n/2\leq k\leq n} k^\ell \mathbb{P}(X_n = k)$$

we only need to show that the second term is $o(n^{\ell-1/2})$. This follows directly from

$$\sum_{n/2\leq k\leq n} k^\ell \mathbb{P}(X_n = k) \leq n^\ell \left(1 - \sum_{0\leq k<n/2} \mathbb{P}(X_n = k)\right) = o(n^{\ell-1/2}),$$

where (3.5) is used in the last estimate. ∎

Finally, we obtain the asymptotics of moments of the maximal clade size M_n which indeed shows that the groups have one big clade.

Corollary 7. *We have,*

$$\mathbb{E}(M_n) = n - \frac{2(1-p)}{(1+p)\sqrt{\pi}} n^{1/2} + o(n^{1/2})$$

and for $\ell \geq 2$

$$\mathbb{E}(M_n - \mathbb{E}(M_n))^\ell \sim (-1)^\ell d_\ell n^{\ell-1/2},$$

where d_ℓ is as in Lemma 13.

Chapter 4

Ancestral Configurations

In the previous chapters, we need that the evolutionary structures reflect the genetic relationship between different species. Due to incomplete lineage sorting, problems on inferences on the species structure via genetic data often occurs. Thus, strategies on more accurate inferences in species structure have been presented in different articles, see (Degnan and Salter, 2005; Slatkin and Pollack, 2006; Degnan and Rosenberg, 2005).

Another problem that arises in the study of the relationship between species tree and gene tree is computing gene tree probabilities. In (Degnan and Salter, 2005), the authors developed a non-recursive algorithm to compute such probabilities. Unfortunately, the method uses a lot of computing time and is thus not suitable for large computations. In (Wu, 2012), the author proposed a recursive way to compute these probabilities. The probabilities were computed via *ancestral configurations*.

Before we give the definition of ancestral configurations, we first establish the following notations and conventions. Here, we will consider a gene tree G with its matching species tree S which will be represented by a phylogenetic tree τ. Throughout this section, we will only consider matching gene trees G and species trees S, that is, $G = S = \tau$ for some phylogenetic tree τ. In addition, we will give an arbitrary labelling of the internal nodes of the phylogenetic tree, see Figure 4.1. Moreover, we will also label the edges of the tree by the label of the node immediately descendant of the edge. For example, in Figure 4.1, we label edge gj by g.

A realization R of a gene tree G is one of the evolutionary possibilities of G on its corresponding species tree S. In Figure 4.1, R_1 and R_2 are different realizations of G. In R_1, note that the gene j of G appears way back further than species j of S (it appears before the appearance of the species k of S). However, no matter which realization of S, j cannot appear after

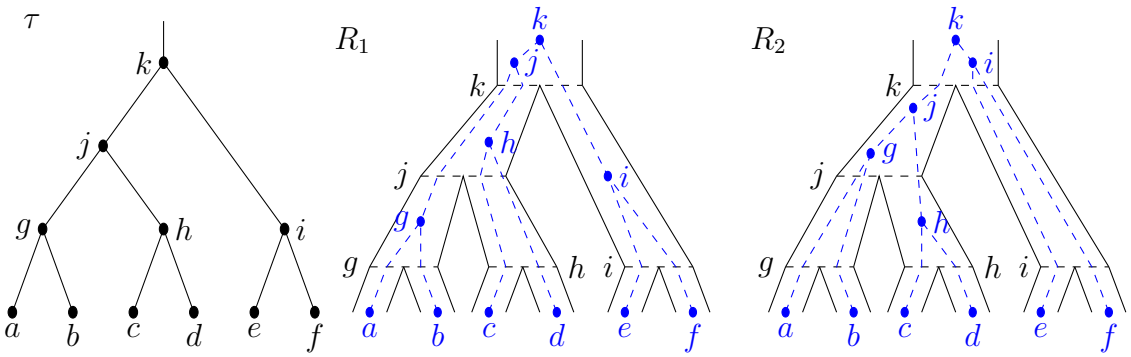


Figure 4.1: A gene tree G and species tree S with matching phylogenetic tree τ. R_1 and R_2 are different realization of the gene tree G (blue dashed line) embedded on its corresponding species tree S (black solid line).

the appearance of the species j of S.

Let R be a fixed realization of a gene tree G on its species tree S. For each internal node x of S, the *ancestral configuration* of R at x is the set of the gene lineages (edges of R) that are in S at the point just before the species x. We denote the ancestral configuration of R at x by $\mathcal{C}(x, R)$. For example, consider the realization R_1 in Figure 4.1. The gene lineages that are present at k of S are $g,\ h$, and i. Hence, $\mathcal{C}(k, R_1) = \{g, h, i\}$. The other ancestral configurations of G are $\mathcal{C}(j, R_1) = \{g, c, d\}$, $\mathcal{C}(g, R_1) = \{a, b\}$, $\mathcal{C}(h, R_1) = \{c, d\}$, and $\mathcal{C}(i, R_1) = \{e, f\}$.

By convention, we say that the leaves have no ancestral configuration. This is due to the fact that information of the gene lineages that appears in the leaves are still unavailable.

Next, denote by $\Re(\tau)$ the set of all realizations of the gene tree $G = \tau$ on its species tree $S = \tau$. For each internal node x, we denote the set of ancestral configurations of G at x by

$$\mathcal{C}(x) = \{\mathcal{C}(x, R) \ : \ R \in \Re(\tau)\}$$

and its corresponding cardinality by

$$\mathcal{C}_x = |\mathcal{C}(x)|.$$

Consider the phylogenetic tree τ in Figure 4.1. Then, we have

$$\mathcal{C}(g) = \{\{a, b\}\},\ \mathcal{C}(h) = \{\{c, d\}\},\ \mathcal{C}(i) = \{\{e, f\}\},$$

$$\mathcal{C}(j) = \{\{a, b, c, d\}, \{g, c, d\}, \{a, b, h\}, \{g, h\}\}$$

and

$$\mathcal{C}(k) = \{\{j, i\}, \{j, e, f\}, \{g, h, i\}, \{g, h, e, f\}, \{a, b, h, i\}, \{a, b, h, e, f\}, \{g, c, d, i\},$$

$$\{g, c, d, e, f\}, \{a, b, c, d, i\}, \{a, b, c, d, e, f\}\}. \tag{4.1}$$

Their corresponding cardinalities are given by $\mathcal{C}_g = \mathcal{C}_h = \mathcal{C}_i = 1$, $\mathcal{C}_j = 4$ and $\mathcal{C}_k = 10$.

Note that $\mathcal{C}_x$ counts the number of ways the gene lineages of G reaches x in S over all possible realization of G. Moreover, note that $\mathcal{C}_x$ depends only on the shape of τ for any internal nodes x. Thus, the following are well-defined. First, we define the *root configurations* of τ denoted by R_τ which is given by

$$R_\tau = \mathcal{C}_r$$

where r is the root of τ. We also define the *total configurations* of τ denoted by T_τ which is given by

$$T_\tau = \sum_x \mathcal{C}_x$$

where the sum runs over all internal nodes x of τ. Note that the two parameters can be computed recursively by

$$R_\tau = (R_{\tau_\ell} + 1)(R_{\tau_r} + 1) \tag{4.2}$$

and

$$T_\tau = T_{\tau_\ell} + T_{\tau_r} + R_\tau \tag{4.3}$$

where τ_ℓ (resp. τ_r) are the two subtrees of τ. The first equation can be explain as follows. A root configuration of τ can be obtained by either a union of a root configuration of τ_ℓ and τ_r, union of a root configuration of τ_ℓ and the root of τ_r, union of a root configuration of τ_r and the root of τ_ℓ, or the root of the tree. This gives us the desired result. The second equation follows directly from the definition of the total configurations of τ.

Applying the formula to the phylogenetic tree τ in Figure 4.1, we have $R_\tau = (R_{\tau_\ell}+1)(R_{\tau_r}+1) = (4+1)(1+1) = 10$ and $T_\tau = T_{\tau_\ell} + T_{\tau_r} + R_\tau = 4 + 1 + 10 = 15$.

In this chapter, we will be studying the limit law and the moments of these parameters under the uniform model and the Yule-Harding model. The rest of the section is as follows. Since the total configurations depends on the root configurations, we will start by studying the root configurations in Section 4.1. The limit law and the moments of the root configurations will be derived in this section. Finally, in Section 4.2, we will study the total configurations for which we use similar tools to derive the limit law and the moments. The limit laws were derived using the results in Section 1.5.

4.1 Root Configurations

We now consider the root configurations as a random parameter which we will denote by R_n. Note that this parameter depends only on the shape of the tree and thus we can consider τ to be embedded into the plane. Thus,

$$R_n \stackrel{d}{=} (R_{I_n} + 1)(R^*_{n-I_n} + 1) \tag{4.4}$$

with $R_1 = 0$, where equality holds in distribution and I_n is the size of the left subtree and R^*_n is an independent copy of R_n. This holds for both uniform and Yule-Harding model. This follows directly from (4.2).

Since we want to apply the results of Wagner (2015), we need to consider an additive parameter. By simple manipulation of (4.2), we have

$$R_\tau + 1 = (R_{\tau_\ell} + 1)(R_{\tau_r} + 1)\left(1 + \frac{1}{R_\tau}\right).$$

Thus, we have

$$\log(R_\tau + 1) = \log(R_{\tau_\ell} + 1) + \log(R^*_{\tau_r} + 1) + \log\left(1 + \frac{1}{R_\tau}\right).$$

Therefore, we have an additive parameter $\mathcal{X}(\tau) = \log(R_\tau + 1)$ with toll function $f(\tau) = \left(1 + \frac{1}{R_\tau}\right)$.

First, we recall some known result for root configurations. Disanto and Rosenberg (2017) showed that under the uniform model, we have

$$\mathbb{E}(R_n) \sim \sqrt{\frac{3}{2}}\left(\frac{4}{3}\right)^n$$

and

$$\mathrm{Var}(R_n) \sim \sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n.$$

Thus, we are left with finding the limit distributions under the uniform and Yule-Harding model and the mean and variance of R_n under the Yule-Harding model.

4.1.1 Limit Distributions of Root Configurations

First, we consider the root configurations R_n under the uniform model. Before we give our result, we first establish the relationship of antichains of a poset and root configurations.

Fix a plane binary tree τ of size n. Let $\tilde{\tau}$ be a plane tree obtained by removing the leaves of τ. The resulting tree $\tilde{\tau}$ is called a *pruned binary tree*, see Section 1.1. Figure 4.2 shows an

example of the pruning process. Now, consider the poset $(X, <)$ induced by $\tilde{\tau}$ by considering $\tilde{\tau}$ as its Hasse diagram with X be the set of the internal nodes of τ. Let $A(\tilde{\tau})$ be the number of non-empty anti-chains of $(X, <)$. In Figure 4.2, the non-empty antichains are

$$\{g\},\ \{h\},\ \{i\},\ \{j\},\ \{k\},\ \{g,h\},\ \{g,i\},\ \{h,i\},\ \{j,i\},\ \{k,i\},\ \text{and}\ \{g,h,i\}. \tag{4.5}$$

Thus, $A(\tilde{\tau}) = 10$.

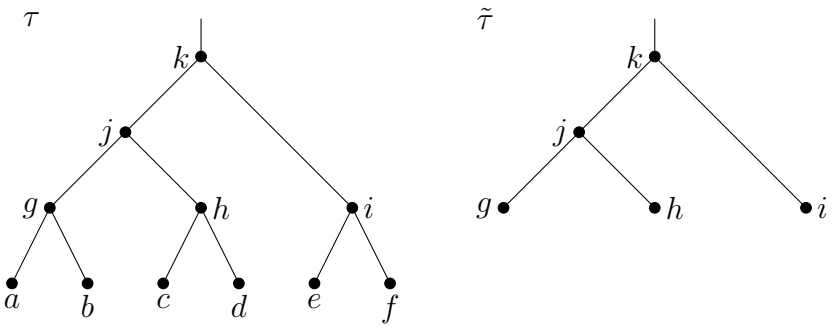


Figure 4.2: A bifurcating tree τ together with is corresponding pruned tree $\tilde{\tau}$.

Note that $A(\tilde{\tau})$ can be computed recursively by

$$A(\tilde{\tau}) = (1 + A(\tilde{\tau}_\ell))(1 + A(\tilde{\tau}_r))$$

where $\tilde{\tau}_\ell$ (resp. $\tilde{\tau}_r$) are the left (resp. right) subtree of $\tilde{\tau}$. This is explained as follows. The non-empty antichains in $A(\tilde{\tau})$ are either the root or the union of the antichains (possibly empty set) in $A(\tilde{\tau}_\ell)$ and $A(\tilde{\tau}_r)$. The empty set in the union is being compensated by the root. This suggests that the non-empty antichains of the pruned binary tree are related to the root configurations of phylogenetic trees. Actually, (4.5) can be obtained from (4.1) by removing all the leaves in the sets of (4.1) and replacing the empty set by the set containing the root.

Indeed, if we take τ to be a random uniform plane binary tree and A_n denotes the number of non-empty antichains of $\tilde{\tau}$, then

$$A_n \stackrel{d}{=} (1 + A_{I_n})(1 + A^*_{n-I_n})$$

with $A_1 = 0$ and I_n is the size of the left subtree of τ and A^*_n is an independent copy of A_n. This is the same as the distributional recurrence of R_n. Thus, their limit laws must coincide.

Wagner (2015) showed the limit law for $\log(A_n)$. The result is stated as follows.

Proposition 13. *Let A_n be the number of non-empty antichains of random uniform pruned binary trees with $n-1$ nodes. Let $\mu_n = \mathbb{E}(\log(A_n))$ and $\sigma_n^2 = \mathrm{Var}(\log(A_n))$. Then, we have*

$$\mu_n \sim \mu n \qquad \text{and} \qquad \sigma_n^2 \sim \sigma^2 n$$

where $\mu \approx 0.272$ and $\sigma^2 \approx 0.034$. Moreover,

$$\frac{\log(A_n) - \mu_n}{\sigma_n}$$

converges weakly to the standard normal distribution $\mathcal{N}(0,1)$.

Consequently, we have the following theorem.

Theorem 16. *Let R_n be the root configuration of random uniform phylogenetic trees of size n. Let $\mu_n = \mathbb{E}(\log(R_n))$ and $\sigma_n^2 = \mathrm{Var}(\log(R_n))$. Then, we have*

$$\mu_n \sim \mu n \qquad \text{and} \qquad \sigma_n^2 \sim \sigma^2 n$$

where $\mu \approx 0.272$ and $\sigma^2 \approx 0.034$. Moreover,

$$\frac{\log(R_n) - \mu_n}{\sigma_n}$$

converges weakly to the standard normal distribution $\mathcal{N}(0,1)$.

Proof. This follows directly from Proposition 13 and Lemma 1. ∎

We now consider the root configuration R_n under the Yule-Harding model. We will apply Proposition 4 directly. Thus, we need a bound for the toll function.

Lemma 14. *Let R_n be the root configurations of a random uniform ranked plane binary trees of size n. Then, we have*

$$\mathbb{E}\left(\log\left(1+\frac{1}{R_n}\right)\right) = \mathcal{O}((0.9)^n).$$

Proof. First, we fix a ranked plane binary tree τ. Note that $R_\tau \geq 2^{\mathrm{ch}(\tau)}$ where $\mathrm{ch}(\tau)$ is the number of cherries of τ. The inequality holds since every set set cherry nodes can form a configuration at the root. Since $\log(1+x) \leq x$ for $x \geq 0$, we have

$$\mathbb{E}\left(\log\left(1+\frac{1}{R_n}\right)\right) \leq \mathbb{E}(R_n^{-1}) \leq \mathbb{E}(2^{-\mathrm{ch}}).$$

Therefore, it is sufficient to show that $\mathbb{E}(2^{-\mathrm{ch}}) = \mathcal{O}((0.9)^n)$.

Disanto and Wiehe (2013) studied the generating function $F(x, z)$ which counts the number of ranked bifurcating trees τ of size n with a given number of cherries, where each ranked bifurcating tree τ is weighted by its probability $2^{n-1-\mathrm{ch}(\tau)}\}/(n-1)!$ under the Yule-Harding distribution:

$$F(x, z) = \sum_{\tau} \frac{2^{n-1-\mathrm{ch}(\tau)}}{(n-1)!} x^{\mathrm{ch}(\tau)} z^n.$$

The sum is taken over ranked bifurcating trees. The coefficient of $x^h z^n$ in $F(x, z)$ gives the probability of h cherries in ranked bifurcating trees of size n under the Yule-Harding model, or equivalently, the probability of h cherries in ranked plane binary trees of size n selected uniformly at random. Hence, the expression $\mathbb{E}_n[2^{-\mathrm{ch}}]$ can be obtained from the coefficient of z^n in $F(\frac{1}{2}, z)$. Disanto and Wiehe (2013) showed that

$$F\left(\frac{1}{2}, z\right) = f(z) = \frac{ze^{z\sqrt{2}} - z}{(\sqrt{2}-2)e^{z\sqrt{2}} + 2 + \sqrt{2}}.$$

We now apply singularity analysis to $f(z)$ to extract its coefficient. Note that the dominant singularity of $f(z)$ is α which is a zero of the denominator. Computing for the value of α, we have

$$\alpha = \frac{1}{\sqrt{2}} \log\left(\frac{2+\sqrt{2}}{2-\sqrt{2}}\right) = \frac{\sqrt{2}\log(3+2\sqrt{2})}{2},$$

Therefore, $\alpha^{-1} \approx 0.802$ and consequently, $\mathbb{E}_n[2^{-\mathrm{ch}}] = \mathcal{O}((0.9)^n)$ which completes the proof. ∎

Now, we are ready to apply Proposition 4. Note that

$$\frac{\sum_{\tau} \log\left(1 + \frac{1}{R_\tau}\right)}{|\mathcal{F}_n|} = \mathbb{E}\left(\log\left(1 + \frac{1}{R_n}\right)\right) = \mathcal{O}((0.9)^n)$$

which satisfies the condition of the proposition.

Theorem 17. *Let R_n be the root configuration of phylogenetic trees of size n selected at random under the Yule-Harding model. Let $\mu_n = \mathbb{E}(\log(R_n))$ and $\sigma_n^2 = \mathrm{Var}(\log(R_n))$. Then, we have*

$$\mu_n \sim \mu n \qquad \text{and} \qquad \sigma_n^2 \sim \sigma^2 n$$

where $\mu \approx 0.351$ and $\sigma^2 \approx 0.008$. Moreover,

$$\frac{\log(R_n) - \mu_n}{\sigma_n}$$

converges weakly to the standard normal distribution $\mathcal{N}(0, 1)$.

Proof. This follows directly from Proposition 4 and Lemma 4. ∎

Exact values of μ and σ can be obtained from Proposition 4. Unfortunately, we find it difficult to write a simple closed form for μ and σ and thus, we only gave their approximate values.

4.1.2 Mean and Variance of Root Configurations under the Yule-Harding Model

In this section, we derive the mean and variance of the number of root configurations under the Yule-Harding model. Our result is as follows.

Theorem 18. *Let R_n be the number of root configurations of a phylogenetic tree selected at random under the Yule-Harding model. Then, we have*

$$\mathbb{E}(R_n) \sim \frac{1}{(1-e^{-2\pi\sqrt{3}/9})^n} \qquad \textit{and} \qquad \mathrm{Var}(R_n) \sim c^n$$

where $c = 2.0449954971....$

The exact value of the constant c is difficult to obtain but using the method in the proof of Proposition 14, we can find an approximation for c with error as small as we want.

Proof. We begin with the proof for the mean.

Let $e_n = \mathbb{E}(R_n)$ and $E(z) = \sum_{n\geq 1} e_n z^n$ be its generating function. By (4.4) and Lemma 5, we have

$$e_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} e_j e_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} e_j, \tag{4.6}$$

with $e_1 = 0$. Multiplying both sides of the equation by $n-1$ and taking the necessary sum, we have

$$\sum_{n\geq 1}(n-1)e_n z^n = zE'(z) - E(z), \qquad \sum_{n\geq 1}(n-1)z^n = \frac{z^2}{(1-z)^2},$$

$$\sum_{n\geq 1}\left(\sum_{j=1}^{n-1} e_n e_{n-j}\right) z^n = E(z)^2 \qquad \text{and} \qquad \sum_{n\geq 1}\left(\sum_{j=1}^{n-1} e_j\right) z^n = \frac{z}{1-z}E(z).$$

Therefore, we have the Riccati differential equation

$$zE'(z) = E(z)^2 + \frac{1+z}{1-z}E(z) + \frac{z^2}{(1-z)^2},$$

with initial condition $E(0) = 0$. Solving this differential equation, we have

$$E(z) = \frac{2z \sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)}{(z-1)\left[\sqrt{3}\cos\left(\frac{\sqrt{3}}{2}\log(1-z)\right) + \sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)\right]}.$$

Note that the dominant singularity of $E(z)$ is $\beta = 1 - e^{-2\pi\sqrt{3}/9}$ which is a zero of

$$\sqrt{3}\cos\left(\frac{\sqrt{3}}{2}\log(1-z)\right) + \sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right).$$

As $z \to \beta$, we have

$$S(z) \sim \frac{1}{1-\frac{z}{\beta}}.$$

Therefore, by singularity analysis, we have $\mathbb{E}(R_n) = (1 - e^{-2\pi\sqrt{3}/9})^{-n}$.

Next, we take a look at the variance of R_n.

Let $s_n = \mathbb{E}(R_n^2)$ and $S(z) = \sum_{n\geq 1} s_n z^n$ be its generating function. Again, from (4.4) and Lemma 5, we have

$$s_n = 1 + \frac{1}{n-1}\sum_{j=1}^{n-1} s_j\, s_{n-j} + \frac{2}{n-1}\sum_{j=1}^{n-1} s_j + \frac{4}{n-1}\sum_{j=1}^{n-1} s_j\, e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j\, e_{n-j} + \frac{4}{n-1}\sum_{j=1}^{n-1} e_j,$$

with $s_1 = 0$. Using a similar method above, we obtain the Riccati differential equation

$$z\,S'(z) = S(z)^2 - S(z)\left[\frac{1+z}{z-1} - 4E(z)\right] + \frac{[z - 2(z-1)E(z)]^2}{(z-1)^2}$$

with initial condition $S(0) = 0$. Solving for the exact expression of $S(z)$ is quite difficult since $E(z)$ is already complicated. So, we find a roundabout way to analyse the coefficient of $S(z)$.

First, by setting

$$\Big(g_2(z), g_1(z), g_0(z)\Big) := \left(\frac{1}{z}, \left(4E(z) - \frac{1+z}{z-1}\right)\frac{1}{z}, \frac{[z-2(z-1)E(z)]^2}{z(z-1)^2}\right),$$

we have

$$S'(z) = g_2(z)\,S(z)^2 + g_1(z)\,S(z) + g_0(z).$$

Let $U(z) = \exp\left(-\int_0^z \frac{S(x)}{x}\,dx\right)$. Then, we have

$$S(z) = -\frac{zU'(z)}{U(z)}.$$

In addition, we have

$$U''(z) - \left(g_1(z) + \frac{g_2'(z)}{g_2(z)}\right)U'(z) + g_2(z)\,g_0(z)\,U(z) = 0 \tag{4.7}$$

with $U(0) = 1$ and $U'(0) = -S'(0)/2$. Notice that the coefficients in (4.7) are analytic for $|z| < 0.702$ (lower bound for the radius of convergence of $E(z)$), with removable singularity at $z = 0$. Thus, $U(z)$ must be analytic for $|z| < 0.702$. Therefore, $S(z)$ is a meromorphic function in $|z| < 0.702$. To analyse $S(z)$, we need to find the zeroes of $U(z)$ with least modulus. In fact, we will show later that $U(z)$ has a unique root $\alpha = 0.4889986317...$ on $\mathcal{B} = \{z \in \mathbb{C} : |z| \leq \frac{1}{2}\}$, see Proposition 14. Thus, the dominant singularity of $S(z)$ is α. In addition, notice that $U(z) = (z - \alpha)\tilde{U}(z)$ where $\tilde{U}(\alpha) \neq 0$. Moreover, $U'(\alpha) = \tilde{U}(\alpha) \neq 0$. Thus, as $z \to \alpha$, we have

$$S(z) = -\frac{\alpha[U'(\alpha) + U''(\alpha)(z-\alpha) + \cdots]}{U(\alpha) + U'(\alpha)(z-\alpha) + U''(\alpha)(z-\alpha)^2 + \cdots} \sim -\frac{\alpha U'(\alpha)}{U'(\alpha)(z-\alpha)} = \frac{1}{1 - \frac{z}{\alpha}}.$$

Hence, we have $s_n \sim \alpha^{-n}$.

Since $\mathrm{Var}(R_n) = s_n - en^2$ and $\alpha^{-1} > (1 - e^{-2\pi\sqrt{3}/9})^2$, we have

$$\mathrm{Var}(R_n) \sim \alpha^{-n}.$$

This proves the second claim of the theorem. ∎

To complete the proof, we still need to show that $U(z)$ has a unique root α on $\mathcal{B}$. Moreover, we need compute an approximate value of α^{-1}.

Before we state and prove the claim, we give an outline of the proof of this claim. First, since $U(z)$ is analytic on $|z| < 0.702$, it can be written in its power series expansion $U(z) = \sum_{n\geq 0} u_n z^n$. The recurrence u_n will be derived in Lemma 15. Then we will obtain a bound for $|u_n|$ in Lemma 17 in $\mathcal{B}$ with the help of the bound for e_n which we will obtain in Lemma 16. Next, we will split the terms of the power series into $U_1(z) = \sum_{n\geq 0}^{100} u_n z^n$ and $U_2(z) = \sum_{n\geq 101} u_n z^n$. In Lemma 18, we will obtain a bound for $|U_1(z)|$ in $\partial\mathcal{B}$ by using the bound of $|u_n|$. We will conclude in Proposition 14 by applying Rouché's theorem.

Lemma 15. *For $n \geq 2$, we have*

$$u_n = \frac{1}{n(n-1)} \sum_{k=0}^{n-1} (3n - k - 3)u_k$$

$$- \frac{4}{n(n-1)} \sum_{k=0}^{n-1} (n - 2k - 1)e_{n-k}u_k + \frac{4}{n(n-1)} \sum_{k=0}^{n-1} \left(\sum_{j=0}^{n-k-1} e_j \right) u_k, \tag{4.8}$$

with $u_0 = 1$ and $u_1 = 0$.

Proof. First notice that for $n \geq 0$, the coefficient of z^n in each term of (4.7) can be written as

$$[z^n]U''(z) = (n+2)(n+1)u_{n+2}$$

$$-[z^n]\left(g_1 + \frac{g_2'}{g_2}\right)U'(z) = -\sum_{k=0}^{n}(n-k+1)(4e_{k+1}+2)u_{n-k+1}$$

$$[z^n]g_2g_0U(z) = \sum_{k=0}^{n}\left[(k+1) + 4\sum_{j=0}^{k}e_{j+1} + 4\sum_{j=0}^{k+2}e_je_{k-j+2}\right]u_{n-k},$$

where for convenience we set $e_0 = 0$.

Making a substitution to the index of summation, we have

$$-4\sum_{k=0}^{n}(n-k+1)e_{k+1}u_{n-k+1} = -4\sum_{k=0}^{n+1}ke_{n-k+2}u_k.$$

Hence, the sum for $-[z^n](g_1 + g_2'/g_2)U'(z)$ can be simplified as

$$-[z^n]\left(g_1 + \frac{g_2'}{g_2}\right)U'(z) = -4\sum_{k=0}^{n+1}ke_{n-k+2}u_k - 2\sum_{k=0}^{n}(n-k+1)u_{n-k+1}.$$

The second sum in this equation together with the first sum $\sum_{k=0}^{n}(k+1)u_{n-k}$ of $[z^n]g_2g_0U(z)$ give

$$-2\sum_{k=0}^{n}(n-k+1)u_{n-k+1} + \sum_{k=0}^{n}(k+1)u_{n-k} = \sum_{k=0}^{n+1}(n-3k+1)u_k.$$

Furthermore, by setting $n = k+2$ in (4.6), the inner sums of $[z^n]g_2g_0U(z)$ can be rewritten as

$$4\sum_{j=0}^{k}e_{j+1} + 4\sum_{j=0}^{k+1}e_je_{k-j+2} = 4(k+1)e_{k+2} - 4(k+1) - 4\sum_{j=1}^{k+1}e_j.$$

Hence, the coefficient of z^n in (4.7) becomes

$$(n+2)(n+1)u_{n+2} - 4\sum_{k=0}^{n+1}ke_{n-k+2}u_k + \sum_{k=0}^{n+1}(n-3k+1)u_k$$

$$+\sum_{k=0}^{n}\left[4(k+1)e_{k+2} - 4(k+1) - 4\sum_{j=1}^{k+1}e_j\right]u_{n-k}.$$

In this expression, we make two substitutions:

$$\sum_{k=0}^{n}4(k+1)e_{k+2}u_{n-k} = \sum_{k=0}^{n+1}4(n-k+1)e_{n-k+2}u_k$$

$$\sum_{k=0}^{n+1}(n-3k+1)u_k - 4\sum_{k=0}^{n}(k+1)u_{n-k} = \sum_{k=0}^{n+1}(n-3k+1)u_k - 4\sum_{k=0}^{n}(n-k+1)u_k$$

$$= \sum_{k=0}^{n+1}(-3n+k-3)u_k,$$

obtaining

$$(n+2)(n+1)u_{n+2} - 4\sum_{k=0}^{n+1} ke_{n-k+2}u_k + \sum_{k=0}^{n+1} 4(n-k+1)e_{n-k+2}u_k + \sum_{k=0}^{n+1}(-3n+k-3)u_k$$
$$+\sum_{k=0}^{n}\Big(-4\sum_{j=1}^{k+1} e_j\Big)u_{n-k},$$

and thus

$$(n+2)(n+1)u_{n+2}+\sum_{k=0}^{n+1} 4(n-2k+1)e_{n-k+2}u_k+\sum_{k=0}^{n+1}(-3n+k-3)u_k+\sum_{k=0}^{n}\Big(-4\sum_{j=1}^{k+1} e_j\Big)u_{n-k}.$$

Finally, because $e_0 = 0$, in this expression we can substitute

$$\begin{aligned}\sum_{k=0}^{n}\Big(-4\sum_{j=1}^{k+1} e_j\Big)u_{n-k} &= \sum_{k=0}^{n}\Big(-4\sum_{j=0}^{k+1} e_j\Big)u_{n-k}\\ &= \sum_{k=0}^{n}\Big(-4\sum_{j=0}^{n-k+1} e_j\Big)u_k\\ &= \sum_{k=0}^{n+1}\Big(-4\sum_{j=0}^{n-k+1} e_j\Big)u_k,\end{aligned}$$

obtaining for $n \geq 0$

$$(n+2)(n+1)u_{n+2}+\sum_{k=0}^{n+1} 4(n-2k+1)e_{n-k+2}u_k-\sum_{k=0}^{n+1}(3n-k+3)u_k-4\sum_{k=0}^{n+1}\Big(\sum_{j=0}^{n-k+1} e_j\Big)u_k = 0,$$

which rescaled is recurrence (4.8). The starting conditions $u_0 = 1$ and $u_1 = 0$, follow from the fact that $U(0) = 1$ and $U'(0) = 0$ as $U(z) = \exp[\int_0^z S(x)/(-x)\,dx]$. ∎

In Lemma 17, we use the recurrence to find an upper bound for $|u_n|$. First, we need an upper bound for e_n.

Lemma 16. *For $n \geq 0$, we have*

$$e_n \leq \left(\frac{9}{10}\right)\left(\frac{3}{2}\right)^n.$$

Proof. Using (4.6), with the help of computing software we have shown that the inequality holds for $0 \leq n \leq 41$. We proceed by induction. Suppose the inequality holds for all $k < n$ with $n > 41$. By (4.6),

$$\begin{aligned}e_n &\leq 1 + \frac{81}{100(n-1)}\sum_{j=1}^{n-1}\left(\frac{3}{2}\right)^n + \frac{9}{5(n-1)}\sum_{j=1}^{n-1}\left(\frac{3}{2}\right)^j\\ &= 1 + \frac{81}{100}\left(\frac{3}{2}\right)^n + \frac{18}{5(n-1)}\left(\frac{3}{2}\right)^n - \frac{27}{5(n-1)}\\ &= \frac{9}{10}\left(\frac{3}{2}\right)^n - \frac{9}{10}\left(\frac{1}{10} - \frac{4}{n-1}\right)\left(\frac{3}{2}\right)^n - \frac{27}{5(n-1)} + 1.\end{aligned}$$

In the last step, we can see that a positive number is subtracted from $\frac{9}{10}(\frac{3}{2})^n$ for $n > 41$, as

$$\frac{9}{10}\left(\frac{1}{10}-\frac{4}{n-1}\right)\left(\frac{3}{2}\right)^n+\frac{27}{5(n-1)}-1>\frac{9}{10}\frac{1}{400}\left(\frac{3}{2}\right)^{42}-1>0.$$

Thus, the claim is proved. ∎

Lemma 17. *For $n \geq 0$, we have*

$$|u_n| \leq \left(\frac{9}{5}\right)^n.$$

Proof. Using (4.8), computing software verifies the inequality for $0 \leq n \leq 25$. We proceed by induction. Suppose that the inequality holds for all $k < n$ with $n > 25$. For simplicity of computation, instead of the bound in Lemma 16, we use the more conservative $(\frac{3}{2})^n$ as a bound for e_n. With (4.8), we get

$$\begin{aligned}
|u_n| &\leq \frac{3}{n}\sum_{k=0}^{n-1}\left(\frac{9}{5}\right)^k+\frac{4}{n}\sum_{k=0}^{n-1}\left(\frac{3}{2}\right)^{n-k}\left(\frac{9}{5}\right)^k+\frac{4}{n(n-1)}\sum_{k=0}^{n-1}\left(\sum_{j=0}^{n-k-1}\left(\frac{3}{2}\right)^j\right)\left(\frac{9}{5}\right)^k \\
&= \frac{15}{4n}\left(\frac{9}{5}\right)^n-\frac{15}{4n}+\frac{20}{n}\left(\frac{9}{5}\right)^n-\frac{20}{n}\left(\frac{3}{2}\right)^n+\frac{30}{n(n-1)}\left(\frac{9}{5}\right)^n \\
&\quad -\frac{40}{n(n-1)}\left(\frac{3}{2}\right)^n+\frac{10}{n(n-1)} \\
&= \frac{5(19n+5)}{4n(n-1)}\left(\frac{9}{5}\right)^n-\frac{20(n+1)}{n(n-1)}\left(\frac{3}{2}\right)^n-\frac{5(3n-11)}{4n(n-1)}.
\end{aligned}$$

In the last step, we have $|u_n| \leq (\frac{9}{5})^n$, as for $n > 25$, the following two inequalities hold:

$$\begin{aligned}
\frac{5(19n+5)}{4n(n-1)} &\leq 1 \\
-\frac{20(n+1)}{n(n-1)}\left(\frac{3}{2}\right)^n-\frac{5(3n-11)}{4n(n-1)} &\leq 0.
\end{aligned}$$

Thus, the claim is proved. ∎

We now consider the set $\mathcal{B} \equiv \{z \in \mathbb{C} \ : \ |z| \leq \frac{1}{2}\}$, and the partition $U(z) = \sum_{k=0}^{\infty} u_k z^k = U_1(z) + U_2(z)$, $U_1(z) \equiv \sum_{k=0}^{100} u_k z^k$ and $U_2(z) \equiv \sum_{k=101}^{\infty} u_k z^k$. Using the bound for $|u_n|$ from Lemma 17, for each $z \in \mathcal{B}$ we have

$$|U_2(z)| \leq \sum_{k=101}^{\infty}|u_k|\,|z|^k \leq \sum_{k=101}^{\infty}\left(\frac{9}{5}\right)^k\left(\frac{1}{2}\right)^k = 10\left(\frac{9}{10}\right)^{101} \approx 0.0002390525900. \tag{4.9}$$

Next, we need a lower bound for $|U_1(z)|$.

Lemma 18. *We have*

$$\min_{z\in\partial\mathcal{B}} |U_1(z)| \geq \frac{3}{1000}.$$

Proof. We obtain the result by considering a function

$$G(t) \equiv \left[\sum_{k=0}^{100} u_k \cos(kt) \left(\frac{1}{2}\right)^k\right]^2 + \left[\sum_{k=0}^{100} u_k \sin(kt) \left(\frac{1}{2}\right)^k\right]^2.$$

$G(t)$ has period 2π, with $G(\pi - t) = G(\pi + t)$, if $t \in [0, \pi]$. For $|z| \in \partial\mathcal{B}$ we can write $z = \frac{1}{2}[\cos t + i \sin t]$ for $t \in [0, 2\pi)$, and thus

$$\begin{aligned}
|U_1(z)| &= \left|\sum_{k=0}^{100} u_k \left[\left(\frac{1}{2}\right)[\cos t + i \sin t]\right]^k\right| \\
&= \left|\sum_{k=0}^{100} u_k \cos(kt) \left(\frac{1}{2}\right)^k + i \sum_{k=0}^{100} u_k \sin(kt) \left(\frac{1}{2}\right)^k\right| \\
&= \sqrt{G(t)}.
\end{aligned}$$

By using the bound in Lemma 17, we have the following inequality

$$\begin{aligned}
|G'(t)| = \Bigg| & 2\left[\sum_{k=0}^{100} u_k \cos(kt) \left(\frac{1}{2}\right)^k\right]\left[-\sum_{k=0}^{100} k u_k \sin(kt) \left(\frac{1}{2}\right)^k\right] \\
& +2\left[\sum_{k=0}^{100} u_k \sin(kt) \left(\frac{1}{2}\right)^k\right]\left[\sum_{k=0}^{100} k u_k \cos(kt) \left(\frac{1}{2}\right)^k\right]\Bigg| \\
\leq\ & 2\left|\sum_{k=0}^{100} u_k \cos(kt) \left(\frac{1}{2}\right)^k\right|\left|\sum_{k=0}^{100} k u_k \sin(kt) \left(\frac{1}{2}\right)^k\right| \\
& + 2\left|\sum_{k=0}^{100} u_k \sin(kt) \left(\frac{1}{2}\right)^k\right|\left|\sum_{k=0}^{100} k u_k \cos(kt) \left(\frac{1}{2}\right)^k\right| \\
\leq\ & 2\left[\sum_{k=0}^{100} |u_k||\cos(kt)| \left(\frac{1}{2}\right)^k\right]\left[\sum_{k=0}^{100} k|u_k||\sin(kt)| \left(\frac{1}{2}\right)^k\right] \\
& + 2\left[\sum_{k=0}^{100} |u_k||\sin(kt)| \left(\frac{1}{2}\right)^k\right]\left[\sum_{k=0}^{100} k|u_k||\cos(kt)| \left(\frac{1}{2}\right)^k\right] \\
\leq\ & 4\left[\sum_{k=0}^{100} \left(\frac{9}{10}\right)^k\right]\left[\sum_{k=0}^{100} k \left(\frac{9}{10}\right)^k\right] \approx 3598.862135. \qquad (4.10)
\end{aligned}$$

We set $\mathcal{I} = \{\frac{k\pi}{1000000} \ : \ k \in \mathbb{Z}, 0 \leq k \leq 1000000\}$. A numerical calculation shows that

$$\min_{t\in\mathcal{I}} G(t) = G(0) \approx 0.01949528529. \qquad (4.11)$$

With these preparations complete, we prove our claim by showing that

$$\min_{t\in[0,\pi]} G(t) \geq \frac{9}{1000000}. \qquad (4.12)$$

We prove (4.12) by contradiction. Suppose there exists $t_0 \in [0, \pi]$ such that $G(t_0) < \frac{9}{1000000}$. Then we can find $t_1 \in \mathcal{I}$ such that

$$|t_1 - t_0| \leq \frac{\pi}{2000000}. \tag{4.13}$$

By the Mean Value Theorem, we can find $c \in (t_0, t_1)$ such that $G(t_1) - G(t_0) = G'(c)(t_1 - t_0)$. From (4.10) and (4.13),

$$\frac{1800\pi}{1000000} \geq |G'(c)(t_1 - t_0)| = |G(t_1) - G(t_0)| \geq G(t_1) - G(t_0). \tag{4.14}$$

However, because $t_1 \in \mathcal{I}$, by (4.11), we have

$$G(t_1) - G(t_0) \geq G(0) - G(t_0) \geq \frac{1}{100} - \frac{9}{1000000} = \frac{9991}{1000000}.$$

This result contradicts the upper bound in (4.14). Thus, (4.12) holds and the claim has been proven. ∎

Lemma 19. *The polynomial $U_1(z)$ has a unique (simple) root β inside $\mathcal{B}$, with $\beta \approx 0.4889986317$.*

Proof. First, by the Intermediate Value Theorem, there exists a real root β with $0 < \beta < \frac{1}{2}$, as we can numerically compute $U_1(0)\, U_1(\frac{1}{2}) < 0$ for the polynomial $U_1(z)$. Thus, we must prove

$$\frac{U_1(z)}{z - \beta} = \frac{U_1(z) - U_1(\beta)}{z - \beta} = \sum_{k=0}^{100} u_k \frac{z^k - \beta^k}{z - \beta} = \sum_{k=0}^{100} u_k \sum_{\ell=0}^{k-1} \beta^{k-1-\ell} z^\ell = \sum_{\ell=0}^{99} \left(\sum_{k=\ell+1}^{100} u_k \beta^{k-1-\ell} \right) z^\ell$$

satisfies $|U_1(z)/(z - \beta)| > 0$ in $\mathcal{B}$.

To do so, we first use the bisection method for root-finding to numerically approximate β by

$$\tilde{\beta} = \frac{1101127027820569}{2251799813685248} \approx 0.4889986317,$$

with the approximation error

$$|\beta - \tilde{\beta}| \leq \frac{1}{2^{50}}. \tag{4.15}$$

Then, we define the polynomial

$$Q(z) \equiv \sum_{\ell=0}^{99} a_\ell z^\ell, \text{ with } a_\ell \equiv \sum_{k=\ell+1}^{100} u_k \tilde{\beta}^{k-1-\ell},$$

through which we can write

$$\begin{aligned} \frac{U_1(z)}{z - \beta} &= Q(z) + (\beta - \tilde{\beta}) R(z), \\ R(z) &\equiv \sum_{\ell=0}^{99} \left(\sum_{k=\ell+1}^{100} u_k \frac{\beta^{k-1-\ell} - \tilde{\beta}^{k-1-\ell}}{\beta - \tilde{\beta}} \right) z^\ell = \sum_{\ell=0}^{99} \left(\sum_{k=\ell+2}^{100} u_k \sum_{j=0}^{k-2-\ell} \beta^j \tilde{\beta}^{k-2-\ell-j} \right) z^\ell. \end{aligned}$$

Note that on $\mathcal{B}$,

$$|R(z)| \leq \sum_{\ell=0}^{99} \sum_{k=\ell+2}^{100} \sum_{j=0}^{k-2-\ell} |u_k||\beta|^j |\tilde{\beta}|^{k-2-\ell-j} |z|^\ell \leq \sum_{\ell=0}^{99} \sum_{k=\ell+2}^{100} \sum_{j=0}^{k-2-\ell} \left(\frac{9}{5}\right)^k \left(\frac{1}{2}\right)^{k-2} \approx 3234.224489, \tag{4.16}$$

where we used the bound for $|u_n|$ from Lemma 17 and the fact that $\beta, \tilde{\beta}, |z| \leq \frac{1}{2}$.

Next, let us consider the function

$$S(r,\theta) \equiv \sum_{\ell=0}^{99} a_\ell r^\ell \cos(\ell\theta)$$

defined over the rectangle $(r,\theta) \in [0,\frac{1}{2}] \times [0,\pi]$, where $S(r,\theta) = \Re(Q(z))$ if $z = r[\cos(\theta) \pm i\sin(\theta)] \in \mathcal{B}$. We need the following bound for the gradient of S:

$$\begin{aligned} |\nabla S| &= \left| \left(\sum_{\ell=0}^{99} \ell a_\ell r^{\ell-1} \cos(\ell\theta), \sum_{\ell=0}^{99} -\ell a_\ell r^\ell \sin(\ell\theta) \right) \right| = \left| \sum_{\ell=0}^{99} \left(\ell a_\ell r^{\ell-1} \cos(\ell\theta), -\ell a_\ell r^\ell \sin(\ell\theta) \right) \right| \\ &= \left| \sum_{\ell=0}^{99} \ell a_\ell r^{\ell-1} \left(\cos(\ell\theta), -r\sin(\ell\theta) \right) \right| \leq \sum_{\ell=0}^{99} \ell |a_\ell| |r|^{\ell-1} | \left(\cos(\ell\theta), -r\sin(\ell\theta) \right) | \\ &\leq \sum_{\ell=0}^{99} \ell |a_\ell| |r|^{\ell-1} \leq \sum_{\ell=0}^{99} \ell |a_\ell| \left(\frac{1}{2}\right)^{\ell-1} \approx 89.628949. \end{aligned} \tag{4.17}$$

Here, we have made use of $|r| < \frac{1}{2}$ and for $|r| < 1$, $\sqrt{\cos^2 x + r^2 \sin^2 x} \leq \sqrt{\cos^2 x + \sin^2 x} = 1$.

A numerical calculation shows that over the grid $\mathcal{I} \equiv \{(\frac{k}{2000}, \frac{j\pi}{1000}) \ : \ (k,j) \in \mathbb{Z}^2, 0 \leq k,j \leq 1000\}$, we have

$$\min_{(r,\theta)\in\mathcal{I}} |S(r,\theta)| = \left| S\left(\frac{1}{2}, \frac{502\pi}{1000}\right) \right| \approx 0.9518894218. \tag{4.18}$$

We now show—with a similar method to that used to prove Lemma 18—that

$$\min_{(r,\theta)\in[0,\frac{1}{2}]\times[0,\pi]} |S(r,\theta)| \geq \frac{3235}{2^{50}}. \tag{4.19}$$

Suppose for contradiction that there exists $z_0 = (r_0,\theta_0) \in [0,\frac{1}{2}]\times[0,\pi]$ such that $|S(r_0,\theta_0)| < 3235/2^{50}$. Then let us take $z_1 = (r_1,\theta_1) \in \mathcal{I}$ such that

$$|z_1 - z_0| < \sqrt{\frac{1}{16} + \frac{\pi^2}{4}} \left(\frac{1}{1000}\right) \leq \frac{1}{500}. \tag{4.20}$$

By the Mean Value Theorem, there exists a point (r,θ) on the line segment from (r_0,θ_0) to (r_1,θ_1) such that

$$\nabla S(r,\theta) \cdot (z_1 - z_0) = S(r_1,\theta_1) - S(r_0,\theta_0),$$

where $\cdot$ is the inner product of $\mathbb{R}^2$. By using the Cauchy-Schwarz inequality together with (4.17), (4.18) and (4.20), the assumption $|S(r_0,\theta_0)| < 3235/2^{50}$ would thus give

$$\begin{aligned}\frac{90}{500} &\geq |\nabla S(r,\theta)||z_1 - z_0| \geq |\nabla S(r,\theta)\cdot(z_1-z_0)| = |S(r_1,\theta_1) - S(r_0,\theta_0)| \\ &\geq |S(r_1,\theta_1)| - |S(r_0,\theta_0)| \geq \frac{9}{10} - \frac{3235}{2^{50}} > 0.89,\end{aligned}$$

which is a contradiction. Hence, (4.19) holds.

Finally, because for $z \in \mathcal{B}$ we have

$$|Q(z)| \geq |\Re(Q(z))| \geq \min_{(r,\theta)\in[0,\frac{1}{2}]\times[0,\pi]} |S(r,\theta)|,$$

by using (4.15), (4.16), and (4.19) it follows that in $\mathcal{B}$,

$$\begin{aligned}\left|\frac{U_1(z)}{z-\beta}\right| &= \left|Q(z) + (\beta-\tilde{\beta})R(z)\right| \geq \left||Q(z)| - |(\tilde{\beta}-\beta)R(z)|\right| \geq \frac{3235}{2^{50}} - |(\tilde{\beta}-\beta)||R(z)| \\ &\geq \frac{3235}{2^{50}} - \frac{|R(z)|}{2^{50}} > \frac{3235}{2^{50}} - \frac{3234.224489\ldots}{2^{50}} > 0.\end{aligned}$$

This concludes the proof. ∎

Combining Lemmas 18 and 19 with the inequality in (4.9), we obtain the following proposition.

Proposition 14. *The function $U(z)$ has a unique (simple) root α inside $\mathcal{B}$, where $\alpha \approx 0.4889986317$.*

Proof. For the decomposition $U(z) = U_1(z) + U_2(z)$, (4.9) together with Lemma 18 gives for $z \in \partial\mathcal{B}$

$$|U_1(z)| \geq \frac{3}{1000} > 0.00025 > |U_2(z)|.$$

Hence, from Rouché's theorem, inside $\mathcal{B}$ the function $U(z)$ has the same number of roots (considered with multiplicity) as polynomial $U_1(z)$. From Lemma 19, we know that $U_1(z)$ has one (simple) root inside $\mathcal{B}$.

The only remaining step is the numerical computation of α, whose first ten digits turn out to coincide with the constant β found in Lemma 19 as the root of $U_1(z)$ inside $\mathcal{B}$. We again decompose $U(z)$:

$$U(z) = \sum_{k=0}^{\infty} u_k z^k = \sum_{k=0}^{500} u_k z^k + \sum_{k=501}^{\infty} u_k z^k = \tilde{U}_1(z) + \tilde{U}_2(z).$$

Note that from our bound for $|u_k|$ (Lemma 17), for each $z \in \mathcal{B}$ we have

$$|\tilde{U}_2(z)| \leq \sum_{k=501}^{\infty} |u_k|\,|z|^k \leq \sum_{k=501}^{\infty} \left(\frac{9}{5}\right)^k \left(\frac{1}{2}\right)^k = 10\left(\frac{9}{10}\right)^{501} \leq 10^{-21}. \tag{4.21}$$

Let us now consider

$$\begin{aligned}\alpha' &= \frac{550563513910285}{1125899906842624} \approx 0.48899863172938484723\\ \alpha'' &= \frac{1101127027820571}{2251799813685248} \approx 0.48899863172938529132.\end{aligned}$$

These values were chosen using the bisection method such that

$$\tilde{U}_1(\alpha') = 2.708185805\ldots \cdot 10^{-16} \quad \text{and} \quad \tilde{U}_1(\alpha'') = -4.953373282\ldots \cdot 10^{-15}.$$

From the bound of $|\tilde{U}_2(z)|$ in (4.21), it is clear that $U(\alpha') > 0$ and $U(\alpha'') < 0$. Let α be the unique root of $U(z)$ in $\mathcal{B}$, which by the Intermediate Value Theorem must be a real root in (α', α''), and let $\epsilon \equiv \alpha - \alpha' \leq 10^{-14}$. Note that

$$\frac{1}{\alpha'} - \frac{1}{\alpha} = \frac{\epsilon}{\alpha'(\alpha' + \epsilon)} \leq \frac{\epsilon}{(\alpha')^2} \leq 5 \cdot 10^{-14}.$$

Thus, we can use

$$\begin{aligned}\alpha' &= 0.48899863172938484723\\ (\alpha')^{-1} &= 2.0449954971518340953\end{aligned}$$

to approximate α and α^{-1}, respectively. ∎

This completes the proof of Theorem 18.

4.2 Total Configurations

We now consider the total configurations as a random parameter which we will denote by T_n. Note that this parameter depends again only on the shape of the tree and thus we can consider τ to be embedded into the plane. In addition, in the Yule-Harding model, we will consider the uniform model for the ranked plane binary trees. Note that

$$T_n \stackrel{d}{=} T_{I_n} + T^*_{n-I_n} + R_n \tag{4.22}$$

with $T_1 = 0$, where equality holds in distribution and I_n is the size of the left subtree and T^*_n is an independent copy of T_n. This holds for both uniform and Yule-Harding model. This distributional recurrence follows directly from (4.3).

Notice that, for any phylogenetic tree τ, we have

$$R_\tau \leq T_\tau \leq nR_\tau. \tag{4.23}$$

The first part of the inequality is trivial since R_τ is one of the terms in the sum T_τ. The second part of the inequality holds since the root configuration is an upper bound to the number of ancestral configurations of any of the $n-1$ internal nodes of τ. Thus, we have

$$\mathbb{E}(T_n) \bowtie \mathbb{E}(R_n) \qquad \text{and} \qquad \mathrm{Var}(T_n) \bowtie \mathrm{Var}(R_n).$$

In particular, we will show in Sections 4.2.3 and 4.2.2 that

$$\mathbb{E}(T_n) \sim c_1 \mathbb{E}(R_n) \qquad \text{and} \qquad \mathrm{Var}(T_n) \sim c_2 \mathrm{Var}(R_n)$$

for some constant c_1, c_2.

4.2.1 Limit Distribution of Total Configurations

Observe that from (4.23) we have

$$\log(T_n) = \log(R_n) + \mathcal{O}(\log n).$$

Thus, we have

$$\mathbb{E}(\log(T_n)) \sim \mathbb{E}(\log(R_n)) \qquad \text{and} \qquad \mathrm{Var}(\log(T_n)) \sim \mathrm{Var}(\log(R_n)).$$

Also, notice that as $n \to \infty$, we have

$$\frac{\mathcal{O}(\log n)}{\sqrt{\mathrm{Var}(\log(R_n))}} \to 0.$$

This holds for both uniform and Yule-Harding model. Thus, we have the following result.

Theorem 19. *Let T_n be the total configurations of phylogenetic trees of size n selected at random under uniform (resp. Yule-Harding) model. Let $\mu_n = \mathbb{E}(\log(T_n))$ and $\sigma_n^2 = \mathrm{Var}(\log(T_n))$. Then, we have*

$$\mu_n \sim \mu n \qquad \textit{and} \qquad \sigma_n^2 \sim \sigma^2 n$$

where $\mu \approx 0.272$ and $\sigma^2 \approx 0.034$ (resp. $\mu \approx 0.351$ and $\sigma^2 \approx 0.008$). Moreover,

$$\frac{\log(T_n) - \mu_n}{\sigma_n}$$

converges weakly to the standard normal distribution $\mathcal{N}(0,1)$.

4.2.2 Mean and Variance of Total Configurations under the Yule-Harding Model

In this section, we derive the mean and variance of the number of total configurations under the Yule-Harding model. We are going to use the results at the end of Section 1.2. Our result is as follows.

Theorem 20. *Let R_n be the total configurations of a phylogenetic tree selected at random under Yule-Harding model. Then, we have*

$$\mathbb{E}(T_n) \sim \frac{1}{(1-e^{-2\pi\sqrt{3}/9})^n} \qquad \textit{and} \qquad \mathrm{Var}(T_n) \sim c^n$$

where $c = 2.0449954971....$ Moreover, we have

$$\rho(T_n, R_n) \sim 1.$$

Proof. Let us first derive the mean of T_n. By Lemma 5 and (4.22), we have

$$\mathbb{E}(T_n) = \frac{2}{n-1}\sum_{j=1}^{n-1}\mathbb{E}(T_j) + \mathbb{E}(R_n). \tag{4.24}$$

Consider the generating functions $R(z) = \sum_{n\geq 1}\mathbb{E}(R_n)z^n$ and $T(z) = \sum_{n\geq 1}\mathbb{E}(T_n)z^n$. From (4.24), we have

$$T'(z) + \frac{z+1}{z^2-z}T(z) = R'(z) - \frac{R(z)}{z}. \tag{4.25}$$

Set $M(z) = \frac{(z-1)^2}{z}$. Notice that $M'(z) = \frac{z+1}{z^2-z}M(z)$. Thus, we have

$$\begin{aligned}\left(\frac{(z-1)^2}{z}T(z)\right)' &= (T(z)M(z))' \\ &= T'(z)M(z) + \frac{z+1}{z^2-z}M(z)T(z) \\ &= \left(R'(z) - \frac{R(z)}{z}\right)M(z).\end{aligned}$$

Solving the differential equation, with initial condition $[T(z)(z-1)^2/z]|_{z=0} = 0$ since the first nonzero term in the expansion of $T(z)$ is the quadratic part, we have

$$T(z) = \frac{z}{(z-1)^2}\int_0^z \left(R'(x) - \frac{R(x)}{x}\right)M(x)\,dx.$$

From the proof of Theorem 18, as $z \to \beta$, we have

$$R(z) \sim \frac{1}{1-\frac{z}{\beta}} \qquad \text{and} \qquad R'(z) \sim \frac{1}{\alpha(1-\frac{z}{\beta})^2}$$

where $\beta = 1 - e^{-2\pi\sqrt{3}/9} \approx 0.702$. Consequently, as $z \to \beta$, we have

$$\left(R'(z) - \frac{R(z)}{z}\right) M(z) \sim \frac{(\beta-1)^2}{\beta} \left(\frac{1}{\beta(1-\frac{z}{\beta})^2}\right)$$

Therefore, as $z \to \beta$, we have

$$\begin{aligned} T(z) &\sim \frac{\alpha}{(\beta-1)^2} \int_0^z \frac{(\beta-1)^2}{\beta} \left(\frac{1}{\beta(1-\frac{x}{\beta})^2}\right) dx \\ &\sim \int_0^z \frac{1}{\beta(1-\frac{x}{\beta})^2}\, dx \\ &\sim \frac{1}{1-\frac{z}{\beta}}. \end{aligned}$$

Thus, $\mathbb{E}(T_n) = [z^n]T(z) \sim \beta^{-n}$.

Now, we find the variance of T_n. To compute for the variance, we need to consider $T_n R_n$ and T_n^2. Their distributional recurrences are given by

$$\begin{aligned} T_n R_n \stackrel{d}{=} T_{I_n} R_{I_n} R^*_{n-I_n} + T_{I_n} R_{I_n} + T_{I_n} R^*_{n-I_n} + T_{I_n} + T^*_{n-I_n} R_{I_n} R^*_{n-I_n} + T^*_{n-I_n} R_{I_n} \\ + T^*_{n-I_n} R^*_{n-I_n} + T^*_{n-I_n} + R_n^2, \end{aligned} \tag{4.26}$$

and

$$T_n^2 \stackrel{d}{=} T_{I_n}^2 + (T^*_{I_n})^2 + 2T_{I_n} T^*_{n-I_n} + 2T_n R_n - R_n^2. \tag{4.27}$$

which follows directly from (4.4) and (4.22). Set

$$S(z) = \sum_{n\geq 1} \mathbb{E}(R_n^2) z^n,$$

$$V(z) = \sum_{n\geq 1} \mathbb{E}(T_n R_n) z^n,$$

and

$$Q(z) = \sum_{n\geq 1} \mathbb{E}(T_n^2) z^n.$$

Thus, from (4.26) and (4.27),

$$V'(z) = V(z)\left(\frac{2R(z)}{z} + \frac{z+1}{z-z^2}\right) + \frac{2T(z)R(z) + \frac{2z}{1-z}T(z) + zS'(z) - S(z)}{z} \tag{4.28}$$

$$Q'(z) + \frac{z+1}{z^2-z} Q(z) = \frac{2T(z)^2 + 2zV'(z) - 2V(z) - zS'(z) + S(z)}{z} \tag{4.29}$$

We proceed in a similar manner as we did for $T(z)$.

We first consider $V(z)$. Define the functions

$$P(z) = \frac{2T(z)R(z) + \frac{2z}{1-z}T(z) + zS'(z) - S(z)}{z} \tag{4.30}$$

and

$$M_1(z) = \frac{(z-1)^2}{z} \exp\left(\int_0^z -\frac{R(x)}{x}\,\mathrm{d}x\right).$$

Then, we have

$$V(z) = \frac{1}{M_1(z)} \int_0^z P(x)M_1(x)\,\mathrm{d}x. \tag{4.31}$$

Notice that the dominant singularity of $V(z)$ is located at $\alpha = 0.488998631729...$ which is the dominant singularity of $S(z)$, see proof of Proposition 14. Here, $T(z)$ and $R(z)$ are both analytic in $|z| < 1/2$ which contains β. Thus, $T(z) \to T(\alpha)$ and $R(z) \to R(\alpha)$ as $z \to \alpha$. In addition, we have the following as $z \to \alpha$

$$M_1(z) \sim k, \qquad S(z) \sim \frac{1}{1-\frac{z}{\alpha}}, \qquad \text{and} \qquad S'(z) \sim \frac{1}{\alpha(1-\frac{z}{\alpha})^2}$$

for some nonzero constant k. Plugging these expressions to (4.30), as $z \to \alpha$, we get

$$\begin{aligned} P(z) &\sim \frac{2T(\alpha)R(\alpha) + \frac{2\alpha}{1-\alpha}T(\alpha)}{\alpha} + \frac{1}{\beta(1-\frac{z}{\alpha})^2} \\ &\sim \frac{1}{\alpha(1-\frac{z}{\alpha})^2} \end{aligned}$$

Finally, we have

$$\begin{aligned} V(z) &\sim \frac{1}{k}\int_0^z \frac{1}{\alpha(1-\frac{x}{\alpha})^2} k\,\mathrm{d}x \\ &\sim \frac{1}{1-\frac{z}{\alpha}} \end{aligned}$$

as $z \to \alpha$. Thus, $\mathbb{E}[T_n R_n] \sim \alpha^{-n}$.

Now, let us consider $Q(z)$. Set

$$G(z) = \frac{2T(z)^2 + 2zV'(z) - 2V(z) - zS'(z) + S(z)}{z}. \tag{4.32}$$

which yields to

$$Q(z) = \frac{1}{M(z)} \int_0^z G(x)M(x)\,\mathrm{d}x.$$

Now, as $z \to \alpha$, we have

$$G(z) \sim \frac{1}{\alpha(1-\frac{z}{\alpha})^2}.$$

Consequently, as $z \to \alpha$, we have

$$Q(z) \sim \frac{1}{1-\frac{z}{\alpha}}$$

since $M(z) \to k_1$ for some non-zero constant k_1. Therefore, $\mathbb{E}(T_n^2) \sim \alpha^{-n}$.

Computing for the variance of T_n, we have

$$\begin{aligned}\mathrm{Var}(T_n) &= \mathbb{E}(T_n)^2 - \mathbb{E}^2(T_n) \\ &\sim \mathbb{E}(T_n)^2.\end{aligned}$$

Hence, $\mathrm{Var}(T_n) \sim \alpha^{-n} = (2.0449954971...)^n$. See proof of Proposition 14 for more precise value of α^{-1}.

Finally, for the last part of the theorem. Notice that

$$\mathrm{Cov}(T_n, R_n) = \mathbb{E}(T_n R_n) - \mathbb{E}(T_n)\mathbb{E}(R_n) \sim \mathbb{E}(T_n R_n).$$

Thus,

$$\rho(T_n, R_n) = \frac{\mathrm{Cov}(T_n, R_n)}{\sqrt{\mathrm{Var}(T_n)}\sqrt{\mathrm{Var}(R_n)}} \sim 1.$$

This completes the proof. ∎

4.2.3 Mean and Variance of Total Configurations under the Uniform Model

In this section, we derive the mean and variance of the number of total configurations under the uniform model. Our result is as follows.

Theorem 21. *Let T_n be the total configurations of a random uniform phylogenetic tree of size n. Then, we have*

$$\mathbb{E}(T_n) \sim \sqrt{6}\left(\frac{4}{3}\right)^n \qquad \textit{and} \qquad \mathrm{Var}(T_n) \sim c\left(\frac{4}{7(8\sqrt{2}-11)}\right)^n$$

where $c = \frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}$. Moreover, we have

$$\rho(T_n, R_n) \sim \frac{1+\frac{\sqrt{2}}{2}}{\sqrt{\frac{2}{17}(15+11\sqrt{2})}} \approx 0.9003666874.$$

Proof. First, we compute for the mean of T_n. Observe that from (4.22) and Lemma 3, we have

$$\mathbb{E}(T_n) = 2\sum_{j=1}^{n-1} \frac{C_{j-1}C_{n-1-j}}{C_{n-1}}\mathbb{E}(T_j) + \mathbb{E}(R_n), \tag{4.33}$$

where $\mathbb{E}[T_1] = 0$. Similarly, by (4.4) and Lemma 3, we have

$$\mathbb{E}(R_n) = \sum_{j=1}^{n-1} \frac{C_{j-1}C_{n-1-j}}{C_{n-1}}\left(\mathbb{E}(R_j)\mathbb{E}(R_{n-j}) + \mathbb{E}(R_j) + \mathbb{E}(R_{n-j}) + 1\right). \tag{4.34}$$

with $\mathbb{E}(R_1) = 0$. Define the following generating function

$$T(z) = \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_n)z^n \qquad \text{and} \qquad R(z)\sum_{n\geq 1} C_{n-1}\mathbb{E}(R_n)z^n.$$

Using (4.33) and (4.34), we have

$$T(z) = 2zC(z)T(z) + R(z) \text{ and } R(z) = R^2(z) + 2zC(z)R(z) + z^2R(z)^2,$$

where $C(z)$ is the generating function of the Catalan numbers C_n. Solving for $T(z)$ from the two equations, we get

$$T(z) = \frac{R(z)}{1-2zC(z)} = \frac{\sqrt{1-4z} - \sqrt{2\sqrt{1-4z}-1}}{2\sqrt{1-4z}}.$$

Notice that the dominant singularity of $T(z)$ is $\alpha = 3/6$ which is the root of $2\sqrt{1-4z} - 1$. Thus, as $z \to \alpha$, we have the asymtotic expansion

$$T(z) \sim \frac{1}{2} - \sqrt{\frac{3}{2}}\sqrt{1-\frac{16z}{3}}.$$

By singularity analysis, we have

$$\mathbb{E}(T_n) = \frac{[z^n]T(z)}{C_{n-1}} \sim \frac{\sqrt{\frac{3}{2}}\frac{(16/3)^n}{2\sqrt{\pi n^3}}}{\frac{4^{n-1}}{\sqrt{\pi n^3}}} = \sqrt{6}\left(\frac{4}{3}\right)^n.$$

Now, let us compute for the variance of T_n. The method is similar but it involves more functions.

Let $\tilde{R}_n = R_n + 1$. The we have the following distributional recurrences

$$\begin{aligned}
\tilde{R}_n &\stackrel{d}{=} \tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1;\\
(\tilde{R}_n)^2 &\stackrel{d}{=} (\tilde{R}_{I_n})^2(\tilde{R}^*_{n-I_n})^2 + 2\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1;\\
T_n\tilde{R}_n &\stackrel{d}{=} T_{I_n}\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + T^*_{n-I_n}\tilde{R}^*_{n-I_n}\tilde{R}_{I_n} + T_{I_n} + T^*_{n-I_n} + (\tilde{R}_n)^2 - \tilde{R}_n;\\
(T_n)^2 &\stackrel{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2T_nR_n - (R_n)^2.
\end{aligned}$$

Now let us consider their corresponding generating functions

$$\begin{aligned}
\tilde{R}(z) &= \sum_{n\geq 1} C_{n-1}\mathbb{E}(\tilde{R}_n)z^n = \sum_{n\geq 1} C_{n-1}\mathbb{E}(R_n)z^n + \sum_{n\geq 1} C_{n-1}z^n = R(z) + zC(z)\\
\tilde{S}(z) &:= \sum_{n\geq 1} C_{n-1}\mathbb{E}(\tilde{R}_n^2)z^n\\
S(z) &:= \sum_{n\geq 1} C_{n-1}\mathbb{E}(R_n^2)z^n = \sum_{n\geq 1} C_{n-1}\mathbb{E}(\tilde{R}_n^2 - 2\tilde{R}_n + 1)z^n\\
&= \sum_{n\geq 1} C_{n-1}\mathbb{E}(\tilde{R}_n^2 - 2(R_n+1) + 1)z^n\\
&= \sum_{n\geq 1} C_{n-1}\mathbb{E}(\tilde{R}_n^2)z^n - 2\sum_{n\geq 1} C_{n-1}\mathbb{E}(R_n)z^n - \sum_{n\geq 1} C_{n-1}z^n\\
&= \tilde{S}(z) - 2R(z) - zC(z)\\
\tilde{V}(z) &:= \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_n\tilde{R}_n)z^n\\
V(z) &:= \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_nR_n)z^n = \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_n(\tilde{R}_n - 1))z^n\\
&= \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_n\tilde{R}_n)z^n - \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_n)z^n = \tilde{V}(z) - T(z)\\
U(z) &:= \sum_{n\geq 1} C_{n-1}\mathbb{E}(T_n^2)z^n.
\end{aligned}$$

Using the distributional recurrence above, we have

$$\begin{aligned}
\tilde{S}(z) - z &= \tilde{S}(z)^2 + 2\tilde{R}(z)^2 + z^2C(z)^2\\
\tilde{V}(z) &= 2\tilde{V}(z)\tilde{R}(z) + 2zC(z)T(z) + \tilde{S}(z) - \tilde{R}(z)\\
U(z) &= 2zC(z)U(z) + 2T(z)^2 + 2V(z) - S(z).
\end{aligned}$$

With the help of computing software, the above equations can be solved as follows. First, we know $R(z)$ and $C(z)$ and thus we get an expression for $\tilde{R}(z)$. With $\tilde{R}(z)$, we can solve for $\tilde{S}(z)$. Next, we can solve simultaneously the expressions for $S(z)$ and $\tilde{V}(z)$. Finally, we can compute for $U(z)$. Using this algorithm, we have

$$V(z) = \frac{-\sqrt{2r-1} + r\left(-r + \sqrt{2r-1} - \sqrt{-r + 4\sqrt{2r-1} - 1} + 3\right) - 1}{2r\sqrt{2r-1}};$$

$$U(z) = \frac{1}{2}\left(-\frac{1}{r^3} + \frac{-\frac{2\sqrt{-r+4\sqrt{2r-1}-1}}{\sqrt{2r-1}} + \sqrt{-2r + 4\sqrt{2r-1} - 1} + \frac{4}{\sqrt{2r-1}} + 3}{r} - \frac{6}{\sqrt{2r-1}} + 1\right),$$

where $r = \sqrt{1-4z}$. Both $U(z)$ and $V(z)$ have dominant singularity at $\beta = 7(8\sqrt{2} - 11)/16$ which is the root of $-r + 4\sqrt{2r-1} - 1$. Finally, using singularity analysis and with help of

computing software, we have

$$\mathbb{E}(T_n R_n) = \frac{[z^n]V(z)}{C_{n-1}} \sim \left(1 + \frac{\sqrt{2}}{2}\right) \sqrt{\frac{7(11-\sqrt{2})}{34}} \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n;$$

$$\mathbb{E}(T_n^2) = \frac{[z^n]U(z)}{C_{n-1}} \sim \frac{2}{17}(15+11\sqrt{2}) \sqrt{\frac{7(11-\sqrt{2})}{34}} \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n.$$

Therefore, $\mathrm{Var}(T_n) = \mathbb{E}(T_n^2) - \mathbb{E}^2(T_n) \sim \mathbb{E}(T_n^2)$.

For the last part of the theorem, we first compute for the covariance of T_n and R_n. Thus, we have

$$\mathrm{Cov}(T_n, R_n) = \mathbb{E}(T_n R_n) - \mathbb{E}(T_n)\mathbb{E}(R_n) \sim \mathbb{E}(T_n R_n).$$

Therefore,

$$\rho(T_n, R_n) = \frac{\mathrm{Cov}(T_n, R_n)}{\sqrt{\mathrm{Var}(T_n)}\sqrt{\mathrm{Var}(R_n)}} \sim \frac{1 + \frac{\sqrt{2}}{2}}{\sqrt{\frac{2}{17}(15+11\sqrt{2})}}$$

This completes the proof. ∎

Chapter 5

Conclusion and Outlook

In the last three chapters of this thesis, we have seen different applications of evolutionary structures such as resource allocation, group formation process and genetic analysis. We also have shown several results about these evolutionary structures which enhance the existing researches involving phylogenetic trees. Despite these recent progresses, scientists are still facing a lot of open problems which still makes phylogenetics an active field of study. Algorithms are still needed in order to help us solve some of the problems in phylogenetics. For example, how to deal with more general types of parameters (in this thesis we mainly considered parameters which are additive in nature or have a simple recursion structure) or how to deal with the β-splitting model (Chapter 2 does not cover $\beta \leq -1$ because of the lack of tools to solve this case). Apart from studying phylogenetic trees, an important recent trend are generalizations of phylogenetic trees. One such generalization are *phylogenetic networks*, see (Huson et al., 2010).

Now, will give some post-analysis on the topics of this thesis. We will follow the flow of the thesis in the discussion. So, we begin with the Shapley values.

In recent years, different versions of the Shapley value have appeared in different articles, see (Hartmann, 2013; Haake et al., 2008; Fuchs and Jin, 2015), including unrooted, rooted, and modified rooted Shapley values. Fuchs and Jin (2015) showed that under the uniform and Yule-Harding model the correlation coefficient between rooted and modified Shapley value is asymptotic to 1 which means that the two values are basically the same. One of the goals of this thesis was to determine the correlation coefficient between the unrooted and rooted Shapley value under the β-splitting model.

In this thesis, we have shown that the correlation coefficient of the unrooted and rooted

Shapley values under the β-splitting model tends to 1 for $\beta > -1$. Unfortunately, the most practical case which is $\beta = -1$ (see (Blum and François, 2006)) is not covered in this study. Also, the uniform model which is the case where $\beta = -3/2$ is also not covered. This makes further studies on this topic necessary.

We will briefly explain why the method in this thesis does not provide a result for $\beta \leq -1$. In particular, we try to consider $\beta = -1$. In the method of proof, we bound the parameters using the asymptotic behaviour of the splitting probability $q_n(j)$. Thus, looking at the splitting probability when $\beta = -1$, we have

$$q_n(j) = \frac{n}{H_{n-1}} \cdot \frac{1}{j(n-j)} \qquad \text{for } 1 \leq j \leq n-1,$$

where $H_n = \sum_{k=1}^{n} 1/k$ is the n-th Harmonic number. Using the same methods as in Section 2.1, we obtain the following bounds

$$\mathbb{E}(S_n) = \mathcal{O}(n(\log n)^2) \qquad \text{and} \qquad \mathcal{O}(n^2(\log n)^4)$$

Applying the methods in Section 2.3 to the third term of (2.10), we have

$$\mathbb{E}(\mathbb{E}((Z_n^{[3]})^2|\mathbf{Y}_n)) = \mathcal{O}\left(\frac{(\log n)^4}{n^2 H_{n-1}} \sum_{j=1}^{n-1} \frac{j^2}{(n-j)^3}\right) = \mathcal{O}((\log n)^3).$$

This explains why the result does not work since this term is required to tend to 0. Note that the above term is expected to be the largest of the terms in Proposition 2.10 since we pick a leaf uniformly at random which means that the leaf is most likely located at the larger subtree. Thus, the value of Y_τ is large while X_τ is small. With this, there is a need to develop a new method to deal with $\beta \leq -1$, especially $\beta = -1$.

Finally, note that the Shapley values are defined via the weights of the edges of the phylogenetic tree. In this thesis, we only considered the case when the weights are all equal to 1. So, one may look at other cases when the weights are not equal. But then, one need to develop a random model for the weights of the edges (and hope that this model has practical applications). So far only models for constructing trees with weights equal to 1 are available. It would be more interesting to have weights depending on the number of leaves present under the edge since we are looking at distributions of resources (more species need more resource) but how one would construct such a model is an open problem.

Now, let us discuss the group formation process.

In this thesis, we considered the number of clades N_n, clades with fixed size $N_n^{[m]}$, and size of the maximal clade M_n in the extra clustering model. We have used a combinatorial approach

to find the limit distribution of these parameters under the uniform model which covers all possible cases for the extra-clustering model. So, we will try to compare our results with the existing results under the Yule-Harding model, see (Durand and François, 2010; Drmota et al., 2014, 2016).

We start with N_n. In (Durand and François, 2010), the authors computed the mean for N_n which is given by

$$\mathbb{E}(N_n) = \begin{cases} \dfrac{c(p)}{\Gamma(2(1-p))} n^{1-2p}, & \text{if } 0 \leq p < 1/2; \\ \dfrac{\log n}{2}, & \text{if } p = 1/2; \\ \dfrac{p}{2p-1}, & \text{if } 1/2 < p < 1, \end{cases}$$

where

$$c(p) = \frac{1}{e^{2(1-p)}} \int_0^1 (1-t)^{-2p} e^{2(1-p)t} \left(1 - (1-p)t^2\right) \mathrm{d}t.$$

In contrast to the uniform case which has finite mean for all $0 \leq p < 1$, the Yule-Harding model has only finite mean when $1/2 < p < 1$. In addition to this, higher moments and limit laws were discussed in (Drmota et al., 2014, 2016). They showed that N_n has a continuous limit law when $p = 0$, for $0 < p < 1/2$ has a mixture of discrete and continuous limit law, and has always a discrete law when $p \geq 1/2$. On the other hand, in our situation, the limit law is always discrete for $0 \leq p < 1$. Moreover, we have convergence of all moment for all the cases while in the Yule-Harding model, moment convergence only holds when $0 < p < 1/2$ and $1/2 < p < 1$.

For $N_n^{[m]}$, in the Yule-Harding model only a result for the mean was proved in previous work. More precisely, Durand and François (2010) showed that the mean is of order n^{1-2p} for $0 \leq p < 1/2$. This leaves us with finding higher moments and the limit distribution for $N_n^{[m]}$. For this, the tools from (Drmota et al., 2014, 2016) should be helpful.

A result for M_n under the Yule-Harding is not available in the literature so far. This may be due to the fact that M_n is not additive and thus most of our methods do not apply for M_n. Thus, we need to develop new methods in order to investigate such parameters.

Finally, we discussed ancestral configurations. Here, we considered the number of root configurations R_n and the total number of configurations T_n. Disanto and Rosenberg (2017) derived the mean and variance of R_n under the uniform model. In this thesis, we filled all remaining gaps by deriving the limit laws, mean and variance for both R_n and T_n under the uniform and Yule-Harding model. However, recall that we only considered a matching gene tree and species tree, so one can work on the non-matching case. There are still recursive ways

to obtain results in such a case but the computations are getting more messy, see (Wu, 2012).

Bibliography

Aldous, D. (1996). Probability distributions on cladograms. In *Random Discrete Structures*, pages 1–18. Springer.

Baker, C. S. and Palumbi, S. (1994). Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science*, 265(5178):1538–1540.

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

Bush, R., Bender, C., Subbarao, K., Cox, N., and Fitch, W. (1999). Predicting the evolution of human influenza a. *Science*, 286(5446):1921–1925.

Chang, B. and Donoghue, M. (2000). Recreating ancestral proteins. *Trends in Ecology & Evolution*, 15(3):109–114.

Chang, C.-C., Yoon, K.-J., Zimmerman, J. J., Harmon, K. M., Dixon, P. M., Dvorak, C., and Murtaugh, M. (2002). Evolution of porcine reproductive and respiratory syndrome virus during sequential passages in pigs. *Journal of Virology*, 76(10):4750–4763.

Degnan, J. H. and Rosenberg, N. A. (2005). Discordance of species trees with their most likely gene trees. *PLOS Genetics*, 2:762–768.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37.

Disanto, F. and Rosenberg, N. A. (2017). Enumeration of ancestral configurations for matching gene trees and species trees. *J. Comput. Biol.*, 24(9):831–850.

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

Drmota, M., Fuchs, M., and Lee, Y.-W. (2014). Limit laws for the number of groups formed by social animals under the extra clustering model. In *Proceedings of the 25th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, Discrete Math. Theor. Comput. Sci. Proc., BA, pages 73–84. Assoc. Discrete Math. Theor. Comput. Sci., Nancy.

Drmota, M., Fuchs, M., and Lee, Y.-W. (2016). Stochastic analysis of the extra clustering model for animal grouping. *J. Math. Biol.*, 73(1):123–159.

Durand, E., Blum, M. G. B., and François, O. (2007). Prediction of group patterns in social mammals based on a coalescent model. *J. Theoret. Biol.*, 249(2):262–270.

Durand, E. and François, O. (2010). Probabilistic analysis of a genealogical model of animal group patterns. *J. Math. Biol.*, 60(3):451–468.

Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University Press, Cambridge.

Fuchs, M. and Jin, E. Y. (2015). Equality of Shapley value and fair proportion index in phylogenetic trees. *J. Math. Biol.*, 71(5):1133–1147.

Haake, C.-J., Kashiwada, A., and Su, F. E. (2008). The Shapley value of phylogenetic trees. *J. Math. Biol.*, 56(4):479–497.

Hartmann, K. (2013). The equivalence of two phylogenetic biodiversity measures: the Shapley value and fair proportion index. *J. Math. Biol.*, 67(5):1163–1170.

Huson, D. H., Rupp, R., and Scornavacca, C. (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press.

McKenzie, A. and Steel, M. (2000). Distributions of cherries for two models of trees. *Math. Biosci.*, 164(1):81–92.

Ou, C.-Y., Ciesielski, C., Myers, G., Bandea, C., Luo, C.-C., Korber, B., Mullins, J., Schochetman, G., Berkelman, R., Economou, A. N., Witte, J., Furman, L., Satten, G., Maclnnes, K., Curran, J., Jaffe, H., Group, L. I., and Group, E. I. (1992). Molecular epidemiology of hiv transmission in a dental practice. *Science*, 256(5060):1165–1171.

Semple, C. and Steel, M. (2003). *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford.

Shapley, L. S. (1988). A value for n-person games [from *Contributions to the Theory of Games, Vol. 2*, 307–317, Princeton Univ. Press, Princeton, NJ, 1953; MR **14**, 779]. In *The Shapley Value*, pages 31–40. Cambridge Univ. Press, Cambridge.

Slatkin, M. and Pollack, J. (2006). The concordance of gene trees and species trees at two linked loci. *Genetics*, 172(3):1979–1984.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40.

Stahn, H. (2020). Biodiversity, Shapley value and phylogenetic trees: some remarks. *J. Math. Biol.*, 80(3):717–741.

Wagner, S. (2015). Central limit theorems for additive tree parameters with small toll functions. *Combin. Probab. Comput.*, 24(1):329–353.

Weitzman, M. L. (1998). The Noah's ark problem. *Econometrica*, 66(6):1279–1298.

Wicke, K. and Fischer, M. (2017). Comparing the rankings obtained from two biodiversity indices: the fair proportion index and the Shapley value. *J. Theoret. Biol.*, 430:207–214.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *International Journal of Organic Evolution*, 66(3):763–775.