Equality of Shapley Value and Fair Proportion Index in Phylogenetic Trees

Michael Fuchs
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, 300
Taiwan

Emma Yu Jin*
Department of Computer Science
University of Kaiserslautern
67663, Germany

December 2, 2014

Abstract

The Shapley value and the fair proportion index of phylogenetic trees have been introduced recently for the purpose of making conservation decisions in genetics. Moreover, also very recently, Hartmann (2013) has presented data which shows that there is a strong correlation between a slightly modified version of the Shapley value (which we call the modified Shapley value) and the fair proportion index. He gave an explanation of this correlation by showing that the contribution of both indices to an edge of the tree becomes identical as the number of taxa tends to infinity. In this note, we show that the Shapley value and the fair proportion index are in fact the same. Moreover, we also consider the modified Shapley value and show that its covariance with the fair proportion index in random phylogenetic trees under the Yule-Harding model and uniform model is indeed close to one.

1 Introduction and Definitions

The *Shapley value*, a parameter long studied in cooperative game theory, has recently been suggested as a prioritization tool for taxa in phylogenetics; see Haake, Kashiwada and Su [3]. However, this parameter has the drawback that its definition is relatively complicated making its computation complicated as well. Therefore, other (more simple) measures have been introduced. One of them is the *fair proportion index* for which a high correlation with the Shapley value was observed by Redding, Hartmann, Mimoto, Bokal, DeVos and Mooers [7].

This correlation was further investigated by Hartmann [4] who produced a lot of data which showed that the correlation index between a slightly modified version of the Shapley value (which we call the modified Shapley value) and the fair proportion index approaches one as the number of taxa tends to infinity. He also gave an (heuristic) explanation of this phenomenon by showing that the contribution of both parameters to an edge becomes equal when the number of taxa becomes large.

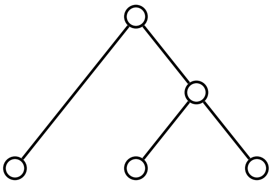
*Corresponding author: Email: jin@cs.uni-kl.de; Phone: 49-631-2053428; Fax: 49-631-2052573.

Key words: Phylogenetic trees, Shapley value, fair proportion index, moments, correlation.

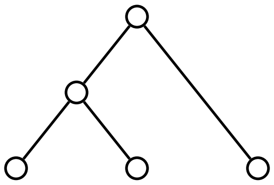
2010 Mathematics Subject Classification: 92B99, 05C05, 60C05.

In this note, we will further investigate the relationship between the Shapley value and the fair proportion index. We will show that they in fact coincide. Moreover, we will also show that Hartmann's modified Shapley value is indeed highly correlated to the fair proportion index under both the Yule-Harding model and the uniform model.

Before we give more details, we fix some notations. First, throughout this note, a *phylogenetic tree* T is a *rooted binary tree*, which is a rooted tree whose nodes have either two or zero children. Moreover, we assume that the tree is a plane tree which means that an ordering is specified for the children of each node. Accordingly,



and



are different trees. This definition differs from the standard definition of phylogenetic trees, where one normally assumes that the tree is non-plane and that the leaves are labeled; see for instance Semple and Steel [8]. However, this difference is irrelevant for the scope of the current study, as is frequently the case for probabilistic studies of phylogenetic trees; e.g., see Chang and Fuchs [1].

We continue introducing notations needed below. The nodes of a phylogenetic tree having zero (resp. two) children are called external (resp. internal) nodes. The taxa of a phylogenetic tree are the external nodes (sometimes also called leaves) of the tree. The left (resp. right) subtree of a tree is the tree rooted at the left (resp. right) child of the root.

The size of the tree will be the number of taxa and we will assign a weight to every edge e which will be denoted by λ_e, $\lambda_e \in \mathbb{R}$. In the sequel, we will also consider random phylogenetic trees under the Yule-Harding model and uniform model. These models will be briefly reviewed in Section 3; for more details see [1].

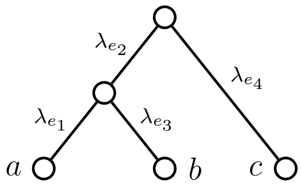
Now, we are going to give definitions of the two parameters above. Fix a phylogenetic tree T of size n, i.e. with n leaves. Moreover, fix a taxon a of T. Then, the *fair proportion index* is defined as

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

where the sum runs over all edges e on the path from the root to a, D_e denotes the number of taxa that are descendants of edge e and λ_e is the weight on edge e.

For the reader's convenience, we give an example. Consider

T_3 :



which is a phylogenetic tree of size 3, where λ_{e_i} is the weight on edge e_i. Consider a which is the leftmost taxon of this tree. Then there are two edges e_1, e_2 on the path from the root to a and $D_{e_1} = 1$, $D_{e_2} = 2$. Thus, $\mathrm{FP}_{T_3}(a) = \lambda_{e_1} + \lambda_{e_2}/2$.

Next, we define the *Shapley value* as

$$\mathrm{SV}_T(a) = \frac{1}{n!} \sum_{S, a \in S} (|S| - 1)!(n - |S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})), \tag{1}$$

where the sum runs over all sets of taxa containing a and $\mathrm{PD}_T(S)$ is the sum of weights of the edges of the ancestor tree of S (the smallest spanning tree containing S and the root). Here, PD is short for phylogenetic diversity, which is a common measure of biodiversity; see [4]. If S is empty, we set $\mathrm{PD}_T(S) = 0$. Note that the singleton set $S = \{a\}$ is included in the definition of the Shapley value (i.e., the sum runs over all S with $a \in S$ and $|S| \geq 1$).

We continue to use T_3 as an example to show the computation of $\mathrm{SV}_{T_3}(a)$. There are four sets of taxa in T_3 that contain a which are $S_1 = \{a\}$, $S_2 = \{a, b\}$, $S_3 = \{a, c\}$ and $S_4 = \{a, b, c\}$. It follows that

$$\begin{aligned}
\mathrm{PD}_{T_3}(S_1) - \mathrm{PD}_{T_3}(S_1 \setminus \{a\}) &= \lambda_{e_1} + \lambda_{e_2}, \\
\mathrm{PD}_{T_3}(S_2) - \mathrm{PD}_{T_3}(S_2 \setminus \{a\}) &= \lambda_{e_1}, \\
\mathrm{PD}_{T_3}(S_3) - \mathrm{PD}_{T_3}(S_3 \setminus \{a\}) &= \lambda_{e_1} + \lambda_{e_2}, \\
\mathrm{PD}_{T_3}(S_4) - \mathrm{PD}_{T_3}(S_4 \setminus \{a\}) &= \lambda_{e_1}
\end{aligned}$$

and therefore by definition $\mathrm{SV}_{T_3}(a) = \lambda_{e_1} + \lambda_{e_2}/2 = \mathrm{FP}_{T_3}(a)$. In Theorem 1, we will prove that the latter equality is not a coincidence, i.e., $\mathrm{SV}_T(a) = \mathrm{FP}_T(a)$ holds for every phylogenetic tree T of size n and every taxon a of T, namely the Shapley value is equal to the fair proportion index.

We will also consider a modified Shapley value, denoted by $\widetilde{\mathrm{SV}}_T(a)$, which is defined as

$$\widetilde{\mathrm{SV}}_T(a) = \frac{1}{n!} \sum_{\substack{S, a \in S \\ |S| \geq 2}} (|S| - 1)!(n - |S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})). \tag{2}$$

Comparing with (1), we have

$$\mathrm{SV}_T(a) = \widetilde{\mathrm{SV}}_T(a) + \frac{\mathrm{PD}_T(a)}{n}.$$

For the particular case $\lambda_e = 1$ for every edge e in T, $\mathrm{PD}_T(a)$ is the depth of taxon a in the tree T, i.e., the length of the path from a to the root in the tree T. The modified Shapley value is the one which apparently was used in [4] (the definition in [4] does not make it very clear).

We conclude the introduction with a brief sketch of the paper. In the next section, we show that the fair proportion index and the Shapley value are identical. In Section 3, we consider the fair proportion index of random phylogenetic trees under the Yule-Harding model and uniform model and derive asymptotic expansions for mean and variance. In Section 4, we use these results together with results for the depth to show that the correlation coefficient of the fair proportion index and the modified Shapley value tends to one as the number of taxa tends to infinity. We will finish the paper with a conclusion.

2 Fair Proportion Index = Shapley Value.

In this section, we will show that the fair proportion index and the Shapley value are actually the same. As in the previous section, we will fix a phylogenetic tree T of size n. Let T_l and T_r denote the left and the right subtree of the tree T. We will assume throughout this section that the size of T_l is j (and consequently, the size of T_r is $n - j$). Finally, we assume that a is a taxon in T_l and denote the left edge of the root by e; see Figure 1.

Clearly, the fair proportion index can be computed recursively by computing it first for the edge e and then computing it in the left subtree. This yields the following proposition.

Proposition 1. *We have,*

$$\mathrm{FP}_T(a) = \frac{\lambda_e}{j} + \mathrm{FP}_{T_l}(a).$$

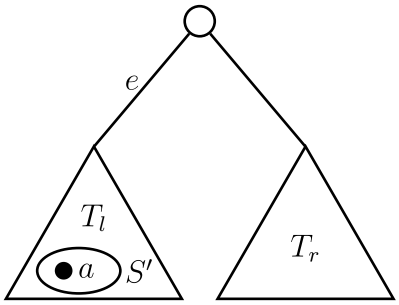


Figure 1: A phylogenetic tree T with subtrees T_l and T_r of size j and $n-j$, respectively. The node a is a taxon in the left subtree T_l.

We now show that the Shapley value satisfies the same recurrence.

Proposition 2. *We have,*

$$\mathrm{SV}_T(a) = \frac{\lambda_e}{j} + \mathrm{SV}_{T_l}(a).$$

Before proving this result, we state a lemma which is needed in the proof.

Lemma 1. *We have,*

$$\sum_{k=0}^{n-j} \binom{m-1+k}{k}\binom{n-m-k}{n-j-k} = \binom{n}{j} \tag{3}$$

for any $1 \le m \le j$.

Proof. This identity can be proved by using the combinatorial interpretation of the binomial coefficients. Let $\mathcal{M}_{n,j}$ be the family of sequences $a_1a_2\cdots a_n$ such that $a_i = 0$ or $a_i = 1$ for every i and there are j 1's contained in the sequence $a_1a_2\cdots a_n$. It is well-known that $|\mathcal{M}_{n,j}| = \binom{n}{j}$.

Now, for any sequence $s_1s_2\cdots s_n \in \mathcal{M}_{n,j}$ and m with $1 \le m \le j$, we consider the position of the m-th 1 in the sequence $s_1s_2\cdots s_n$. Suppose that $s_{m+k} = 1$ is the m-th 1 in the sequence $s_1s_2\cdots s_n$ for some k, $0 \le k \le n-j$. Then, we can decompose the sequence $s_1s_2\cdots s_n$ into two subsequences $s_1s_2\cdots s_{m+k-1}1$ and $s_{m+k+1}\cdots s_n$ such that the subsequence $s_1s_2\cdots s_{m+k-1}1$ contains k 0's and m 1's and the subsequence $s_{m+k+1}\cdots s_n$ contains $(n-j-k)$ 0's and $(j-m)$ 1's. So, the number of sequences $s_1s_2\cdots s_n$ in $\mathcal{M}_{n,j}$ such that $s_{m+k} = 1$ is the m-th 1 is counted by $\binom{m+k-1}{k}\binom{n-m-k}{n-j-k}$. By summing over all k, (3) follows. ∎

Proof of Proposition 2. For any set S_l of taxa contained in the left-subtree T_l and $a \in S_l$, we consider

$$\mathcal{A}_{S_l} = \{S : S = S_l \cup S_r, S_r \text{ is any set of taxa contained in the right subtree } T_r \text{ or } S_r = \emptyset\ \}.$$

It follows that for any $S \in \mathcal{A}_{S_l}$, we have

$$\begin{aligned}\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\}) &= \mathrm{PD}_T(S_l \cup S_r) - \mathrm{PD}_T((S_l \cup S_r) \setminus \{a\}) \\ &= \mathrm{PD}_T(S_l) + \mathrm{PD}_T(S_r) - \mathrm{PD}_T(S_l \setminus \{a\}) - \mathrm{PD}_T(S_r) \\ &= \mathrm{PD}_T(S_l) - \mathrm{PD}_T(S_l \setminus \{a\}). \end{aligned} \tag{4}$$

On the other hand, for any set S that contains a, there is a unique way to write $S = S_l \cup S_r$ where S_l (resp. S_r) is the set of taxa contained in the left (resp. right) subtree T_l (resp. T_r) and $a \in S_l$. This implies that $S \in \mathcal{A}_{S_l}$ and hence

$$\bigcup_{S_l} \mathcal{A}_{S_l} = \{S : a \in S, S \text{ is a set of taxa in the tree } T\},$$

where the (disjoint) union runs over all the subsets S_l of taxa contained in the left subtree T_l and $a \in S_l$.

The last property and (4) imply that

$$\begin{aligned}
\mathrm{SV}_T(a) &= \frac{1}{n!} \sum_{S, a \in S} (|S|-1)!(n-|S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})) \\
&= \frac{1}{n!} \sum_{S_l, a \in S_l} \sum_{S \in \mathcal{A}_{S_l}} (|S|-1)!(n-|S|)!(\mathrm{PD}_T(S) - \mathrm{PD}_T(S \setminus \{a\})) \\
&= \sum_{S_l, a \in S_l} \left(\frac{1}{n!} \sum_{S \in \mathcal{A}_{S_l}} (|S|-1)!(n-|S|)! \right) (\mathrm{PD}_T(S_l) - \mathrm{PD}_T(S_l \setminus \{a\})), \qquad (5)
\end{aligned}$$

where the first sum (from left to right) runs over all the sets S_l of taxa from the left subtree T_l with $a \in S_l$.

We now simplify the expression inside the bracket as follows

$$\frac{1}{n!} \sum_{S \in \mathcal{A}_{S_l}} (|S|-1)!(n-|S|)! = \frac{1}{n!} \sum_{\substack{S_r \\ S = S_l \cup S_r}} (|S_l| + |S_r| - 1)!(n - |S_l| - |S_r|)!$$

where the second sum (from left to right) runs over all the sets S_r of taxa from the right subtree T_r. Since the right subtree T_r has $n-j$ taxa, we have $0 \le |S_r| \le n-j$ for any S_r and therefore

$$\begin{aligned}
&\frac{1}{n!} \sum_{\substack{S_r \\ S = S_l \cup S_r}} (|S_l| + |S_r| - 1)!(n - |S_l| - |S_r|)! \\
&\qquad = \frac{1}{n!} \sum_{k=0}^{n-j} \binom{n-j}{k} (|S_l| + k - 1)!(n - |S_l| - k)! \\
&\qquad = \frac{(n-j)!}{n!} \sum_{k=0}^{n-j} \frac{(|S_l| + k - 1)!(n - |S_l| - k)!}{k!(n-j-k)!} \\
&\qquad = \frac{(n-j)!}{n!} (|S_l| - 1)!(j - |S_l|)! \sum_{k=0}^{n-j} \binom{|S_l| - 1 + k}{k} \binom{n - |S_l| - k}{n-j-k} \\
&\qquad = \frac{(|S_l| - 1)!(j - |S_l|)!}{j!},
\end{aligned}$$

where we used Lemma 1 in the last step.

Plugging this into (5) yields

$$\mathrm{SV}_T(a) = \frac{1}{j!} \sum_{S_l, a \in S_l} (|S_l| - 1)!(j - |S_l|)!(\mathrm{PD}_T(S_l) - \mathrm{PD}_T(S_l \setminus \{a\})), \qquad (6)$$

where the sum runs over all the sets S_l of taxa from the left subtree T_l with $a \in S_l$.

Now, note that for $S_l \neq \{a\}$, we have

$$\mathrm{PD}_T(S_l) - \mathrm{PD}_T(S_l \setminus \{a\}) = \mathrm{PD}_{T_l}(S_l) - \mathrm{PD}_{T_l}(S_l \setminus \{a\}) \tag{7}$$

since the edge e in Figure 1 is counted in both terms on the left-hand side. Moreover,

$$\mathrm{PD}_T(\{a\}) = \lambda_e + \mathrm{PD}_{T_l}(\{a\}), \tag{8}$$

where λ_e comes from the edge e in Figure 1. Plugging (7) and (8) into (6) yields

$$\mathrm{SV}_T(a) = \frac{\lambda_e}{j} + \frac{1}{j!} \sum_{S_l, a \in S_l} (|S_l| - 1)!(j - |S_l|)!(\mathrm{PD}_{T_l}(S_l) - \mathrm{PD}_{T_l}(S_l \setminus \{a\})).$$

This proves the claim. ∎

We can now state the main result of this paper.

Theorem 1. *The fair proportion index and the Shapley value are identical, i.e.,*

$$\mathrm{FP}_T(a) = \mathrm{SV}_T(a).$$

Proof. This follows from the above two propositions, together with the initial condition $\mathrm{FP}_{T_2}(a) = \mathrm{SV}_{T_2}(a) = \lambda_e$, where T_2 is the tree of size 2. ∎

3 Fair Proportion Index of Random Phylogenetic Trees

In what follows, we assume that $\lambda_e = 1$ for all edges e. Moreover, from now on, we will consider random phylogenetic trees of size n and denote by FP_n the fair proportion index of a random taxon (where a taxon is picked uniformly from the set of all taxa).

We first consider the Yule-Harding model [1] which is defined by a tree evolution process: the tree grows by choosing at random one of the leaves and replacing it by a cherry (an internal node with two children). We stop when a tree with n external nodes is constructed. This is the top-down construction of a phylogenetic tree of size n under the Yule-Harding model. Alternatively, a bottom-up construction can be used: start with n external nodes and successively choose a random pair and coalesce the two nodes; stop when only one node (the root) is left. It is clear that the random models arising from these two constructions are the same.

We next recall some basic properties of this model; for details see [1]. First, the size of the left subtree I_n is uniformly distributed on $\{1, \dots, n-1\}$. Moreover, if I_n is fixed, then both left and right subtrees are again random phylogenetic trees under the Yule-Harding model which are independent.

Using these properties and the same recursive procedure for computing the fair proportion index as in the last section, we obtain that, for $n \geq 2$ and $1 \leq j \leq n-1$,

$$\mathrm{FP}_n | (I_n = j) = \begin{cases} \dfrac{1}{j} + \mathrm{FP}_j, & \text{with probability } j/n; \\ \dfrac{1}{n-j} + \mathrm{FP}_{n-j}, & \text{with probability } (n-j)/n, \end{cases} \tag{9}$$

where $\mathrm{FP}_1 = 0$. From this recurrence, we will deduce the following result.

Theorem 2. *For the fair proportion index of a random phylogenetic tree under the Yule-Harding model, we have*

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

and

$$\mathrm{Var}(\mathrm{FP}_n) = 10 - 6H_{n-1}^{(2)} - \frac{6}{n} - \frac{4}{n^2}.$$

Consequently, as $n \to \infty$,

$$\mathbb{E}(\mathrm{FP}_n) \sim 2 \qquad \textit{and} \qquad \mathrm{Var}(X_n) \sim 10 - \pi^2 = 0.130395\cdots.$$

Proof. We first prove the result for the mean of FP_n. Therefore, observe that for any phylogenetic tree T of size n, we have

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

where the sum runs over all the taxa a of T. This identity is easily explained: for every edge e that is the ancestor of j taxa, we have $D_e = j$ and the edge e contributes exactly $j \cdot 1/j = 1$ to the summation $\sum_a \mathrm{FP}_T(a)$. Thus, $\sum_a \mathrm{FP}_T(a)$ is equal to the number of edges in the tree T of size n, which is $2n-2$. The above identity now yields

$$\mathbb{E}(\mathrm{FP}_n) = \frac{1}{n}\sum_a \mathrm{FP}_T(a) = 2 - \frac{2}{n}$$

since we consider a random taxon which is uniformly chosen from the set of all taxa.

For the second moment, we use (9) which yields, for $n \geq 2$,

$$\mathbb{E}(\mathrm{FP}_n^2) = \frac{1}{n-1}\sum_{j=1}^{n-1}\left(\frac{j}{n}\mathbb{E}\left(\frac{1}{j} + \mathrm{FP}_j\right)^2 + \frac{n-j}{n}\mathbb{E}\left(\frac{1}{n-j} + \mathrm{FP}_{n-j}\right)^2\right),$$

where $\mathbb{E}(\mathrm{FP}_1^2) = 0$. This recurrence can be simplified into

$$n(n-1)\mathbb{E}(\mathrm{FP}_n^2) = 8(n-1) - 6H_{n-1} + 2\sum_{j=1}^{n-1} j\mathbb{E}(\mathrm{FP}_j^2),$$

where $H_n = \sum_{j=1}^{n} 1/j$ denotes the n-th harmonic number. By computing the difference of successive terms of the sequence $\{n(n-1)\mathbb{E}(\mathrm{FP}_n^2)\}_{n\geq 1}$, we obtain, for $n \geq 1$,

$$\mathbb{E}(\mathrm{FP}_{n+1}^2) - \mathbb{E}(\mathrm{FP}_n^2) = \frac{8}{n(n+1)} - \frac{6}{n^2(n+1)}.$$

Iterating it yields

$$\mathbb{E}(\mathrm{FP}_n^2) = \sum_{j=1}^{n-1}\left(\frac{8}{j(j+1)} - \frac{6}{j^2(j+1)}\right) = 8 - \frac{8}{n} - 6\sum_{j=1}^{n-1}\frac{1}{j^2(j+1)} = 14 - 6H_{n-1}^{(2)} - \frac{14}{n} \tag{10}$$

$$\sim 14 - \pi^2, \tag{11}$$

where $H_n^{(2)} = \sum_{j=1}^{n} 1/j^2$ and we used $\lim_{n\to\infty} H_n^{(2)} = \pi^2/6$. From this, our result for the variance follows by subtracting the square of the mean. ∎

For our considerations below, we need the moments for the depth D_n of a random taxon in a tree of size n, where the depth is defined as the distance from the randomly chosen taxon to the root. Note that similar to the fair proportion index, we have the following recurrence, for $n \geq 2$ and $1 \leq j \leq n-1$,

$$D_n|(I_n = j) = \begin{cases} 1 + D_j, & \text{with probability } j/n; \\ 1 + D_{n-j}, & \text{with probability } (n-j)/n, \end{cases} \tag{12}$$

where $D_1 = 0$. Then, with exactly the same method as above, we obtain the following result.

Proposition 3. *For the depth in a random phylogenetic tree under the Yule-Harding model, we have*

$$\mathbb{E}(D_n) = 2H_n - 2$$

and

$$\mathrm{Var}(D_n) = 2H_n - 4H_n^{(2)} + 2.$$

Consequently, as $n \to \infty$,

$$\mathbb{E}(D_n) \sim \mathrm{Var}(D_n) \sim 2 \log n.$$

Remark 1. For the above asymptotics, we used the elementary fact $H_n \sim \log n$.

Remark 2. The above result can also be obtained from the well-known results for the unsuccessful search in random binary search trees (which are known to be equivalent to random phylogenetic trees under the Yule-Harding model); see Section 2.4 in Mahmoud [5].

Now, we turn to the uniform model which assigns the same probability to every phylogenetic tree of size n. It is well-known that the number of phylogenetic trees of size n is the $(n-1)$-th Catalan number C_{n-1}. This implies that under the uniform model the size I_n of the left subtree has the distribution

$$P(I_n = j) = \frac{C_{j-1}C_{n-1-j}}{C_{n-1}},$$

where $1 \leq j \leq n-1$. Moreover, again once the size of the left subtree is fixed, both left and right subtrees are independent random phylogenetic trees under the uniform model. As a consequence, the fair proportion index again satisfies (9) with the only difference that I_n is replaced by the distribution above.

Before starting with our analysis, we recall that

$$C(z) = \sum_{n \geq 0} C_n z^n = \frac{1 - \sqrt{1-4z}}{2z} \qquad \text{and} \qquad C_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n^3}}, \tag{13}$$

where the above asymptotics follows either from Stirling's formula or, alternatively, from $C(z)$ by singularity analysis; for the latter standard tool from analytic combinatorics (which will be also used below) see Chapter VI of Flajolet and Sedgewick [2].

We will now prove the following theorem.

Theorem 3. *For the fair proportion index of a random phylogenetic tree under the uniform model, we have, as $n \to \infty$,*

$$\mathbb{E}(\mathrm{FP}_n) = 2 - \frac{2}{n} \sim 2 \qquad \textit{and} \qquad \mathrm{Var}(\mathrm{FP}_n) \sim 12 \ln 2 - 8 = 0.317766 \cdots.$$

Proof. First, observe that the mean is the same as for the Yule-Harding model

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

since under both models, we consider a random taxon which is uniformly chosen from the n taxa.

Next, for the second moment, we obtain from (9) the following recurrence, for $n \geq 2$,

$$\mathbb{E}(\mathrm{FP}_n^2) = \sum_{j=1}^{n-1} \frac{C_{j-1}C_{n-1-j}}{C_{n-1}} \left(\frac{j}{n}\mathbb{E}\left(\frac{1}{j} + \mathrm{FP}_j\right)^2 + \frac{n-j}{n}\mathbb{E}\left(\frac{1}{n-j} + \mathrm{FP}_{n-j}\right)^2 \right),$$

where $\mathrm{FP}_1 = 0$. Set $X_n = 1/n + \mathrm{FP}_n$. Then, for $n \geq 2$,

$$nC_{n-1}\left(\mathbb{E}(X_n^2) + \frac{3-4n}{n^2}\right) = 2\sum_{j=1}^{n-1} C_{j-1}C_{n-1-j}j\mathbb{E}(X_j^2), \tag{14}$$

where $X_1 = 1$. In order to (asymptotically) solve this recurrence, we use generating functions. Therefore, set

$$F(z) = \sum_{n\geq 1} nC_{n-1}\mathbb{E}(X_n^2)z^{n-1}.$$

Our aim is to translate (14) into an identity for the generating functions $C(z)$ and $F(z)$ and then employ the singularity analysis method to estimate $\mathbb{E}(X_n^2)$. First, observe that

$$\sum_{n\geq 2} nC_{n-1}\mathbb{E}(X_n^2)z^{n-2} = \frac{F(z)-1}{z}$$

and

$$\sum_{n\geq 2} C_{n-1}\frac{3-4n}{n}z^{n-2} = \frac{1}{z} + \frac{1}{z^2} + \frac{3\ln z}{2z^2} - \frac{\sqrt{1-4z}}{z^2} + \frac{3}{2z^2}\ln\left(\frac{1+\sqrt{1-4z}}{1-\sqrt{1-4z}}\right).$$

This is the generating function of the left-hand side of (14). Next, note that the generating function of the right-hand side equals $2C(z)F(z)$. Thus, by equating and solving for $F(z)$, we obtain that

$$F(z) = -\frac{1}{z\sqrt{1-4z}} - \frac{3\ln z}{2z\sqrt{1-4z}} + \frac{1}{z} - \frac{3}{2z\sqrt{1-4z}}\ln\left(\frac{1+\sqrt{1-4z}}{1-\sqrt{1-4z}}\right).$$

From this, by singularity analysis and (13)

$$\mathbb{E}(X_n^2) = \frac{[z^{n-1}]F(z)}{nC_{n-1}} \sim 12\ln 2 - 4.$$

Consequently,

$$\mathbb{E}(\mathrm{FP}_n^2) = \mathbb{E}(X_n^2) + \frac{3-4n}{n^2} \sim 12\ln 2 - 4.$$

The claimed result for the variance follows from this and the result for the mean. ∎

Again, with the same method, a similar result for the depth (which again satisfies the same recurrence (12) as that under the Yule-Harding model) can be obtained.

Theorem 4. *For the depth in a random phylogenetic tree under the uniform model, we have, as $n \to \infty$,*

$$\mathbb{E}(D_n) = \frac{4^{n-1}}{nC_{n-1}} - 1 \sim \sqrt{\pi n} \qquad \textit{and} \qquad \mathrm{Var}(D_n) \sim (4-\pi)n$$

Remark 3. The depth for internal nodes (which correspond to ancestors) was also considered; see, e.g., Meir and Moon [6] for the mean.

4 Correlation between fair proportion index and modified Shapley value

In this section, we consider the modified Shapley value of a random taxon in a random phylogenetic tree of size n which will be denoted by $\widetilde{\mathrm{SV}}_n$. Similarly, the Shapley value will be denoted by SV_n. Recall that since $\lambda_e = 1$ for every edge e in the random phylogenetic tree, we have

$$\mathrm{FP}_n = \mathrm{SV}_n \qquad \text{and} \qquad \mathrm{SV}_n = \widetilde{\mathrm{SV}}_n + \frac{D_n}{n}, \tag{15}$$

where D_n is the depth from the previous section.

We again first consider the Yule-Harding model for which we have the following result for the correlation of FP_n and $\widetilde{\mathrm{SV}}_n$.

Theorem 5. *For the correlation coefficient of the fair proportion index and the modified Shapley value of a random phylogenetic tree under the Yule-Harding model, we have, as $n \to \infty$,*

$$\rho(\mathrm{FP}_n, \widetilde{\mathrm{SV}}_n) \sim 1.$$

Proof. From (15),

$$\mathrm{Cov}(\mathrm{FP}_n, \widetilde{\mathrm{SV}}_n) = \mathrm{Var}(\mathrm{FP}_n) - \frac{\mathrm{Cov}(\mathrm{FP}_n, D_n)}{n}.$$

Using the results from the previous section, we have

$$\mathrm{Cov}(\mathrm{FP}_n, D_n) \leq (\mathrm{Var}(\mathrm{FP}_n)\mathrm{Var}(D_n))^{1/2} = \mathcal{O}(\sqrt{\log n}). \tag{16}$$

From this, we obtain that, as $n \to \infty$,

$$\mathrm{Cov}(\mathrm{FP}_n, \widetilde{\mathrm{SV}}_n) \sim \mathrm{Var}(\mathrm{FP}_n).$$

Similarly,

$$\mathrm{Var}(\widetilde{\mathrm{SV}}_n) = \mathrm{Var}(\mathrm{FP}_n) + \frac{\mathrm{Var}(D_n)}{n^2} - 2\frac{\mathrm{Cov}(\mathrm{FP}_n, D_n)}{n} \sim \mathrm{Var}(\mathrm{FP}_n).$$

Thus, the claimed result follows. ▮

Remark 4. In fact, $\mathrm{Cov}(\mathrm{FP}_n, D_n)$ under the Yule-Harding model can be computed explicitly by solving its recursion. First, by definition

$$\mathrm{Cov}(\mathrm{FP}_n, D_n) = \mathbb{E}(\mathrm{FP}_n D_n) - \mathbb{E}(\mathrm{FP}_n)\mathbb{E}(D_n).$$

From Section 3, we know the expectations $\mathbb{E}(\mathrm{FP}_n)$ and $\mathbb{E}(D_n)$. It remains to compute the expectation $\mathbb{E}(\mathrm{FP}_n D_n)$ which satisfies, for $n \geq 2$,

$$\mathbb{E}(\mathrm{FP}_n D_n) = \frac{1}{n-1}\sum_{j=1}^{n-1}\left(\frac{j}{n}\mathbb{E}\left((1+D_j)\left(\frac{1}{j} + \mathrm{FP}_j\right)\right) + \frac{n-j}{n}\mathbb{E}\left((1+D_{n-j})\left(\frac{1}{n-j} + \mathrm{FP}_{n-j}\right)\right)\right)$$

which can be simplified as

$$\binom{n}{2}\mathbb{E}(\mathrm{FP}_n D_n) = \sum_{j=1}^{n-1}(2H_j + 2j - 3 + j\mathbb{E}(\mathrm{FP}_j D_j)).$$

Again by computing the successive terms of the sequence $\{\binom{n}{2}\mathbb{E}(\mathrm{FP}_nD_n)\}_{n\geq 1}$, we obtain that

$$\mathbb{E}(\mathrm{FP}_nD_n)=\sum_{i=1}^{n-1}\frac{2(2H_i+2i-3)}{i(i+1)},$$

which leads to

$$\begin{aligned}\mathrm{Cov}(\mathrm{FP}_n,D_n)&=\sum_{i=1}^{n-1}\frac{2(2H_i+2i-3)}{i(i+1)}-(2H_n-2)\left(2-\frac{2}{n}\right)\\&=\sum_{i=1}^{n-1}\frac{4H_i}{i(i+1)}+\frac{4H_n}{n}-6+\frac{2}{n}\\&=4H_{n-1}^{(2)}-6+\frac{2}{n}+\frac{4}{n^2}\sim\frac{2\pi^2}{3}-6=0.579736\cdots.\end{aligned}$$

Our final result of the paper is that the correlation coefficient of fair proportion index and modified Shapley value under the uniform model tends to one, too.

Theorem 6. *For the correlation coefficient of the fair proportion index and the modified Shapley value of a random phylogenetic tree under the uniform model, we have, as $n\to\infty$,*

$$\rho(\mathrm{FP}_n,\widetilde{\mathrm{SV}}_n)\sim 1.$$

Proof. This follows with the same proof as for the Yule-Harding model. ▮

Remark 5. Again, one can derive a more precise result for $\mathrm{Cov}(\mathrm{FP}_n,D_n)$. Therefore, observe that $\mathbb{E}(\mathrm{FP}_nD_n)$ satisfies, for $n\geq 2$,

$$\begin{aligned}\mathbb{E}(\mathrm{FP}_nD_n)=\sum_{j=1}^{n-1}\frac{C_{j-1}C_{n-j-1}}{C_{n-1}}&\left(\frac{j}{n}\mathbb{E}\left((1+D_j)\left(\frac{1}{j}+\mathrm{FP}_j\right)\right)\right.\\&\left.+\frac{n-j}{n}\mathbb{E}\left((1+D_{n-j})\left(\frac{1}{n-j}+\mathrm{FP}_{n-j}\right)\right)\right).\end{aligned}$$

By setting $X_n=1/n+\mathrm{FP}_n$ and using $\mathbb{E}(D_n)=\frac{4^{n-1}}{nC_{n-1}}-1$, $\mathbb{E}(\mathrm{FP}_n)=2-\frac{2}{n}$, we have, for $n\geq 2$,

$$\frac{nC_{n-1}}{2}\mathbb{E}(X_nD_n)=\sum_{j=1}^{n-1}(2j-1+j\mathbb{E}(X_jD_j))C_{j-1}C_{n-j-1}+\frac{2^{2n-3}}{n}-\frac{C_{n-1}}{2}.\tag{17}$$

Next, consider the generating function

$$G(z)=\sum_{n\geq 2}nC_{n-1}\mathbb{E}(X_nD_n)z^{n-1}.$$

Then, (17) can be translated into

$$\left(\frac{1}{2}-\frac{1-\sqrt{1-4z}}{2}\right)G(z)=-\frac{1}{8z}\log(1-4z)-\frac{1}{4z}(1-\sqrt{1-4z})+\frac{(1-\sqrt{1-4z})^2}{4z\sqrt{1-4z}}.$$

By performing singularity analysis on $G(z)$, we finally get

$$\mathbb{E}(\mathrm{FP}_nD_n)=\frac{[z^{n-1}]G(z)}{nC_{n-1}}\sim\log\, n+2\sqrt{\pi\, n}.$$

Therefore,

$$\mathrm{Cov}(\mathrm{FP}_n,D_n)=\mathbb{E}(\mathrm{FP}_nD_n)-\mathbb{E}(\mathrm{FP}_n)\mathbb{E}(D_n)\sim\log\, n.$$

5 Conclusion

In a recent paper, Hartmann [4] observed that the modified Shapley value and the fair proportion index are highly correlated and gave an explanation for this phenomenon. In this note, we proved that the Shapley value coincides with the fair proportion index. Moreover, we also considered Hartmann's modified Shapley value for which we (asymptotically) computed the correlation coefficient with the fair proportion index for random phylogenetic trees under the Yule-Harding model and the uniform model. For both models, the correlation coefficient tends to one as the tree grows which explains the observations made in [4].

Acknowledgments

The authors are grateful to the anonymous referees for many good suggestions. This research was done while both authors visited the Institut für Diskrete Mathematik und Geometrie, Technical University of Vienna. The authors thank the department for hospitality. The first author was partially supported by NSC grants NSC-102-2918-I-009-012 and MOST-103-2115-M-009-007-MY2; the second author was supported by the grant from German Research Foundation (DFG), JI 207/1-1.

References

[1] H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481-512.

[2] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*, Cambridge University Press, 2009.

[3] C. Haake, A. Kashiwada, F. E. Su (2008). The Shapley value of phylogenetic trees, *J. Math. Biol.*, **56**, 479–497.

[4] K. Hartmann (2013). The equivalence of two phylogenetic biodiversity measures: the Shapley value and Fair Proportion index, *J. Math. Biol.*, **67**, 1163–1170.

[5] H. M. Mahmoud. *Evolution of Random Search Trees*, Wiley, New York, 1992.

[6] A. Meir and J. W. Moon (1978). On the altitude of nodes in random trees, *Canad. J. Math.*, **30:5**, 997–1015.

[7] D. Redding, K. Hartmann, A. Mimotot, D. Bokal, M. DeVos, A. O. Mooers (2008). The most "original species" often capture more phylogenetic diversity than expected, *J. Theor. Biol.*, **251**, 606–615.

[8] C. Semple and M. Steel. *Phylogenetics*, Oxford University Press, 2003.