Correlation between Shapley Values of Rooted Phylogenetic Trees under the Beta-Splitting Model

Michael Fuchs*
Department of Mathematical Sciences
National Chengchi University
Taipei, 116
Taiwan

Ariel R. Paningbatan†
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, 300
Taiwan

September 8, 2019

Abstract

In recent years, several different versions of the Shapley value have been introduced in phylogenetics for the purpose of ranking biodiversity data in order to decide whether to preserve the data or not. Two of these Shapley values are the *rooted* and *unrooted Shapley value* which have been compared with the fair proportion index since this index is easier to compute. In particular, it was proved for the former that it is identical with the fair proportion index and numerical data was presented by several authors that the latter is strongly correlated with the fair proportion index. In this paper, we will prove a theoretical result which supports this observation. More precisely, we will prove that in random phylogenetic trees under the β-splitting model, the correlation coefficient between the unrooted Shapley value and the fair proportion index indeed tends to one for all β with $\beta > -1$. We also present data which suggests that the convergence worsens as β is approaching -1.

1 Introduction and Main Result

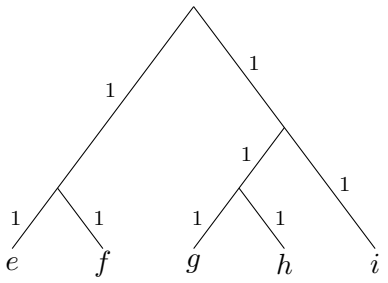
Conservation decisions in biodiversity are usually based on indices which measure the biodiversity value of the data. One such index is the *Shapley value* which was originally proposed as a measure of importance of each invidual player in a cooperative game. Recently, several versions of this value have been introduced in biodiversity and the interconnection between these different versions has been investigated; see Wicke and Fischer (2017); Fuchs and Jin (2015); Haake et al. (2008) and Hartmann (2013). Before recalling these results, we give some basic definitions.

First, a *rooted phylogenetic tree* is a rooted tree which is binary, non-plane and leaf-labeled with each leaf representing a taxon; see Figure 1 for one possible embedding of such a tree into the plane. Moreover, edges of a rooted phylogenetic tree are often labeled where the labels represent evolutionary information (such as, e.g., time); again see the tree in Figure 1 where we choose all weights equal to 1 since we will mainly consider this case in the sequel. Rooted phylogenetic trees have long been used to visualize, understand and analyze evolutionary data; see, e.g, Steel (2016) and references therein.

Let now T be a fixed rooted phylogenetic tree and a one of its taxa. Then, the *rooted Shapley value* is defined as follows

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

*Email address: mfuchs@nccu.edu.tw
†On leave from the Institute of Mathematics, University of the Philippines, Diliman, Quezon City 1101, Philippines.



leaf	$\mathrm{SV}_T^{[r]}$	$\mathrm{SV}_T^{[u]}$
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 1: A rooted phylogenetic tree with rooted and unrooted Shapley value for each leaf.

where n is the number of taxa of T and the sum runs over all sets S of taxa containing a. Moreover, $\mathrm{PD}_T^{[r]}(S)$ is 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 (this tree is usually called the *ancestor tree of* S in the combinatorial literature). For instance, if we choose $a = e$ in the tree from Figure 1, then we obtain $\mathrm{SV}_T^{[r]}(e) = 3/2$. (The values for the other leaves are listed in the table in Figure 1.)

The above Shapley value has appeared under different names (see, e.g., Wicke and Fischer (2017)), however, we prefer to add the prefix *rooted* because the above definition of the phylogenetic diversity includes the root. In fact, there is a closely related notion of phylogenetic diversity which does not include the root and this notion leads to what we will call *unrooted Shapley value* in this paper:

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

where $\mathrm{PD}_T^{[u]}(S)$ is the *unrooted biodiversity of* S which is defined as the sum of all weights of the edges of the minimal spanning tree containing S (again there is a standard name in combinatorics for this tree, namely, *Steiner tree of* S). Note that this Shapley value is also obtained by unrooting the tree T and using the definition of the Shapley value for unrooted trees by Haake et al. (2008). Again, different names have been used for it; see, e.g., Wicke and Fischer (2017). The values for this Shapley value for the leaves of the tree from Figure 1 can also be found in the table in Figure 1.

Hartmann (2013) gave the definition (1), but (confusingly) used the unrooted Shapley value for his results and numerical data. More precisely, Hartmann provided a heuristic and numeric explanation for the fact that the Shapley value is closely related to the *fair proportion index* which is defined as

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

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 taxa below edge e. Note that this index is a priori much easier to compute than the (rooted or unrooted) Shapley values, which is the main reason why it is used in biodiversity conservation projects.

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

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

for all rooted phylogenetic trees T and taxa a. Believing that Hartmann used a rooted version of the Shapley value, Fuchs and Jin (2015) then defined a *modified rooted Shapley value* as follows:

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

where the only difference to the definition of the rooted Shapley value is that the sum now runs over all non-singleton sets of taxa. 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 (which are two standard random models on rooted phylogenetic trees; see the next section)

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

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 taxa that is generated by the uniform resp. Yule-Harding model and the taxon is also chosen uniformly at random. Fuchs and Jin (2015) considered this result to be a theoretical justification of the results in Hartmann's paper. However, as mentioned above, Hartmann used the unrooted Shapley value as Mike Steel pointed out to us in an email communication (see also Wicke and Fischer (2017)).

The main purpose of this work is to investigate the relationship between the unrooted Shapley value and the fair proportion index and give a theoretical result which shows that they are indeed highly correlated. More precisely, the main result of the paper is the following theorem (for the definition of the β-splitting model see the next section).

Theorem 1. *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,*

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

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

Note that this result contains the Yule-Harding model ($\beta = 0$), but not the uniform model ($\beta = -3/2$). Moreover, the important case of $\beta = -1$ is also not covered (the importance of this case for phylogenetic applications was discussed, e.g., by Blum and François (2006)). Whether or not our result extends to these cases (especially $\beta = -3/2$) is not clear and also the data which we are going to present in Section 5 does not really give a clue since the convergence slows down as β tends to -1.

We conclude the introduction with an outline of the proof of our main result and with a brief sketch of the paper. Our main observation for the proof is that it will be sufficient to show that the variance of the difference between the unrooted Shapley value and the fair proportion index tends to 0 (see Proposition 4 below). Thus, we need a suitable expression for this difference and also suitable bounds for it. The bounds will be in terms of moments of certain recursive parameters of rooted phylogenetic trees under the β-splitting model. Therefore, we have to study these moments and this we will do first in Section 2 (the results in this section might be of independent interest because they are quite general). Next, we will consider the difference between the unrooted Shapley value and the fair proportion index and derive an expression for it. This is done in Section 3 and again the results of this sections are of independent interest. In particular, they should be compared with the recent work by Stahn (2019) where a similar expression for the difference was proved. However, whereas the expression given by Stahn is useful from a linear algebra point of view, our expression is of combinatorial nature and will be useful from a computational point of view. This expression will then be used in Section 4 together with the results from Section 2 to prove our main theorem and in Section 5 to produce numerical results. We will finish the paper with a conclusion in Section 6.

2 β-splitting Model and Shape Parameters

In this section, we recall the β-splitting model which was proposed by Aldous (1996). Moreover, we study properties of certain shape parameters that are defined on random trees generated by this model.

First, the random model by Aldous (1996) is defined as follows. Assume that the n taxa are points which are chosen uniformly at random from the unit interval. Then, break the unit interval into two pieces at a point x which is chosen randomly with a density function $f(x)$ that satisfies the symmetry condition $f(x) = f(1-x)$ for $x \in (0,1)$. If either one of the two intervals $(0,x)$ and $(x,1)$ does not contain a taxon, repeat this step by independently choosing another point x with density $f(x)$ until both intervals contain at least one taxon. The resulting two intervals can be seen as the two subtrees of the root and these subtrees will contain the taxa from the two respective intervals. Finally, repeat this procedure independently with each subinterval scaled to the unit interval. Stop with intervals which only contain one taxon.

Note that this procedure gives a random distribution on the set of rooted, binary, plane trees. However, if one forgets the embedding into the plane and labels the leaves with a random permutation of $\{1,\dots,n\}$, one also gets a random distribution on rooted phylogenetic trees (where now and in the sequel we set all edge weights equal to one). Nevertheless, since we are interested in parameters on trees which are only related to the shape and neither to the specific embedding of the tree nor to the labeling, this final step just amounts to a "re-scaling" of the probability model. Thus, we can safely drop the last step from all our subsequent discussions.

As for the choice of the density $f(x)$, Aldous (1996) suggested to use a (symmetric) β-distribution

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)^2} x^\beta (1-x)^\beta, \qquad x \in (0,1), \qquad -1 < \beta < \infty.$$

The resulting model is called the *β-splitting model*. For this model, Aldous showed that the number of taxa which fall into the interval $(0,x)$ generated by the first split equals j with probability

$$p_{n,j} = \frac{1}{c_n(\beta)} \frac{\Gamma(\beta+j+1)\Gamma(\beta+n-j+1)}{j!(n-j)!}, \qquad 1 \le j < n, \tag{2}$$

where

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

Note that the *splitting probabilities* $p_{n,j}$ completely describe the random process which generates the random tree under the β-splitting model (and also note that by (2), one can actually extend the range of β to $\beta > -2$; however, we will not consider this extended range of β in this paper).

In the sequel, we will need an asymptotic expansion of $p_{n,j}$.

Lemma 1. *We have,*

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

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

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

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. ∎

We fix now a random tree T of size n which was generated by the β-splitting model (recall that the tree is rooted, binary and plane). Then, we have the following two important properties:

- The distribution of the number I_n of taxa in the left subtree is given by the splitting probabilities, i.e.,
$$\mathbb{P}(I_n = j) = p_{n,j}, \qquad \text{for } 1 \le j \le n-1.$$
- Given that $I_n = j$, left and right subtree of T are random trees of size j and $n-j$ which are also generated by the β-splitting model; moreover, these two random trees are independent.

We next discuss certain shape parameters of T which can be computed recursively in a sense which will become clear soon.

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 T and also in computer science where it is called the *total path length* and used as a complexity measure for algorithms on T. From the above properties of the β-splitting model, it is immediately clear that S_n can be computed recursively as follows

$$S_n \stackrel{d}{=} S_{I_n} + S^*_{n-I_n} + n, \qquad (n \ge 2) \tag{3}$$

with $S_1 = 0$, where the equality holds in distribution 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 resp. right subtree plus the left edge resp. right edge connecting the root with the left resp. right subtree for each taxa of the tree (the latter gives the factor n).

Another shape parameter, which we will also need in the sequel and which is closely related to S_n, is the distance of a random taxon to the root. We will denote this parameter by D_n and call it the *depth* (this name is also borrowed from computer science). Again it can be computed recursively:

$$D_n|(I_n = j) \stackrel{d}{=} \begin{cases} D_j + 1, & \text{with probability } j/n; \\ D_{n-j} + 1, & \text{with probability } (n-j)/n, \end{cases} \qquad (n \ge 2) \tag{4}$$

with $D_1 = 0$. The two cases in the bracket above correspond to the case where the chosen taxon is in the left or right subtree, respectively, and the 1s counts 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:

$$\mathrm{FP}_n|(I_n = j) \stackrel{d}{=} \begin{cases} \mathrm{FP}_j + 1/j, & \text{with probability } j/n; \\ \mathrm{FP}_{n-j} + 1/(n-j), & \text{with probability } (n-j)/n, \end{cases} \qquad (n \ge 2) \tag{5}$$

with $\mathrm{FP}_1 = 0$. (Note that 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.)

We need in the sequel properties of the first two moments of these random variables. To obtain such properties, it will turn out that these moments for S_n satisfy a recurrence

$$a_n = 2\sum_{j=1}^{n-1} p_{n,j} a_j + b_n, \qquad (n \geq 2) \tag{6}$$

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

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

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.

We will prove two general results about a sequence a_n which satisfies (6). The first result is a version of what computer scientists call a *master theorem* (see, e.g., Roura (2001) for similar results).

Proposition 1. *Let a_n be sequence which satisfies (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 1,

$$2\sum_{j=1}^{n-1} p_{n,j} j(\log j)^{\delta+1} = \frac{2\Gamma(2\beta+2)}{\Gamma(\beta+1)^2} \frac{1}{n^{2\beta+1}} \sum_{j=1}^{n-1} j^{\beta+1}(n-j)^\beta (\log j)^{\delta+1} \left(1 + \mathcal{O}\left(\frac{1}{j^\epsilon} + \frac{1}{(n-j)^\epsilon}\right)\right). \tag{7}$$

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

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

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 (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

$$\begin{aligned} 2\sum_{j=1}^{n-1} p_{n,j} j(\log j)^{\delta+1} &= \frac{2\Gamma(2\beta+2)}{\Gamma(\beta+1)^2}\cdot\frac{(\beta+1)\Gamma(\beta+1)^2}{(2\beta+2)\Gamma(2\beta+2)} n(\log n)^{\delta+1} + Kn(\log n)^{\delta} + o(n(\log n)^{\delta}) \\ &= n(\log n)^{\delta+1} + Kn(\log n)^{\delta} + o(n(\log n)^{\delta}), \end{aligned} \tag{8}$$

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 (6) and using (8),

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

where $\epsilon > 0$ is a sufficiently small constant (coming from the o-term in (8)) and the last step holds by choosing c such that $c \ge -d/(K+\epsilon)$ (which is possible since $K < 0$).

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

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

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,

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

where $\epsilon > 0$ is as in part (i) and the last step follows by choosing c such that $c \ge d/(1-K-\epsilon)$ (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 1 (the bounds for the mean also follows from the results by Aldous (1996)).

Corollary 1. *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 (3),

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

with $P_1(u) = 1$.

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

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

with $\mathbb{E}(S_1) = 0$. This is (6) with $b_n = n$. Thus, from Proposition 1, 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 (6) with

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

where the last estimate follows from that for the mean. Thus, again by Proposition 1, we obtain that $\mathbb{E}(S_n^2) = \mathcal{O}(n^2 (\log n)^2)$ which concludes the proof. ∎

Corollary 2. *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 (4), we obtain that

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

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 (6) with

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

Thus, Proposition 1 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 (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 2. *Let a_n be a sequence which satisfies (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

$$\tilde{b}_n = \begin{cases} 0, & \text{if } 1 \le n \le n_0; \\ b_n + 2p_{n,n_0} b_{n_0}, & \text{if } n \ge n_0 + 1 \end{cases}$$

and denote by $\tilde{a}_n$ the corresponding sequence which satisfies (6). Then, clearly $a_n \ge \tilde{a}_n$ and thus it suffices to show that the claim holds for $\tilde{a}_n$. Also, note that

$$\tilde{b}_n \ge 2p_{n,n_0} b_{n_0} \ge \frac{K_1}{n^{\beta+1}}, \qquad (n \ge n_0 + 1) \tag{9}$$

for some suitable constant $K_1 > 0$ (this follows with similar arguments as in Lemma 1).

Now, we claim that

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

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 (6) gives for n sufficiently large:

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

where we used:

$$2\sum_{j=1}^{n-1} p_{n,j} j = n$$

and

$$2\sum_{j=1}^{n_0} p_{n,j} j \le \frac{K_2}{n^{\beta+1}}$$

which is proved as (9). Finally, choosing $0 < c \le K_1/K_2$ gives $\tilde{a}_n \ge cn$ 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 1.

Corollary 3. *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

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

where the sum runs over all taxa of T.

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 (5) satisfies the recurrence

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

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

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

Thus, $n\mathrm{Var}(\mathrm{FP}_n)$ satisfies (6) with

$$b_n = 2\sum_{j=1}^{n-1} j p_{n,j} \Delta_{n,j}^2$$

which satisfies the assumptions from the Proposition 2. Consequently, $n\mathrm{Var}(\mathrm{FP}_n) = \Omega(n)$ which is the claimed result. ∎

Remark 1. In the special case of $\beta = 0$, we have $p_{n,j} = 1/(n-1)$. This is the so-called Yule-Harding model; see, e.g., Steel (2016). Under this model, more precise results can be obtained either by elementary means (see, e.g., Fuchs and Jin (2015)) or by using analytic combinatorics (see Flajolet and Sedgewick (2009)).

3 Difference between Unrooted and Rooted Shapley Value

In this section, we consider the difference between unrooted and rooted Shapley value in a fixed rooted phylogenetic tree T. Let T_ℓ and T_r denote left and right subtree of the root of T, respectively. Also, let a be a fixed taxon and assume w.l.o.g. that $a \in T_\ell$.

Before stating our result, we need some further notations. First, for a set of taxa S, we consider the *least common ancestor of* S which is the smallest common ancestor of S under the order induced by T where the root is the largest element. Define

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

and

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

(Note that $X_T^{[i]}$ is related to the cophenetic value; see Sokal and Rohlf (1962).) Then, our main result in this section is as follows.

Proposition 3. *For the difference between unrooted and rooted Shapley value, we have*

$$\begin{aligned}\mathrm{SV}_T^{[u]}(a) - \mathrm{SV}_T^{[r]}(a) = &-\frac{1}{n}D_T(a) + \frac{1}{n!}\sum_{i=1}^{|T_r|} i!(n-i-1)!\left(X_{T_r}^{[i]} + \binom{|T_r|}{i}\right) \\ &+ \frac{1}{n!}\sum_{i=1}^{|T_\ell|-1} i!(n-i-1)!Y_{T_\ell^{[i]}}(a), \end{aligned} \tag{10}$$

where $D_T(a)$ is the distance of a to the root and $|T|$ denotes the number of taxa of T.

Proof. In order to prove the result, we have to compare the difference of $\mathrm{PD}_T^{[u]}(S) - \mathrm{PD}_T^{[u]}(S \setminus \{a\})$ and $\mathrm{PD}_T^{[r]}(S) - \mathrm{PD}_T^{[r]}(S \setminus \{a\})$ for all sets S of taxa 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 taxa of S except a from T_ℓ and S_r are all the taxa of S from T_r. We have to distinguish between four cases.

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

  In this case, we have

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

  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

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

  and

  $$\mathrm{PD}_T^{[r]}(S) - \mathrm{PD}_T^{[r]}(S \setminus \{a\}) = D_T(a)$$

  which gives the first term on the right hand side of (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, $\mathrm{PD}_T^{[u]}(S) - \mathrm{PD}_T^{[u]}(S \setminus \{a\})$ and $\mathrm{PD}_T^{[r]}(S) - \mathrm{PD}_T^{[r]}(S \setminus \{a\})$ are explained in Figure 2. Moreover, the difference is explained on the left of Figure 4. Overall, by summing over all sets of size i, we obtain the second term on the right hand side of (10). (Note that the term $\binom{|T_r|}{i}$ comes from the contribution of the edge from to the root to the right subtree of T.)

- 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, $\mathrm{PD}_T^{[u]}(S) - \mathrm{PD}_T^{[u]}(S \setminus \{a\})$ and $\mathrm{PD}_T^{[r]}(S) - \mathrm{PD}_T^{[r]}(S \setminus \{a\})$ are explained in Figure 3. Moreover, the difference is explained on the right of Figure 4. Overall, by summing over all sets of size i and noting that sets S_ℓ which contain a do not contribute to $Y_{T_\ell}^{[i]}(a)$, we obtain the final term on the right hand side of (10).

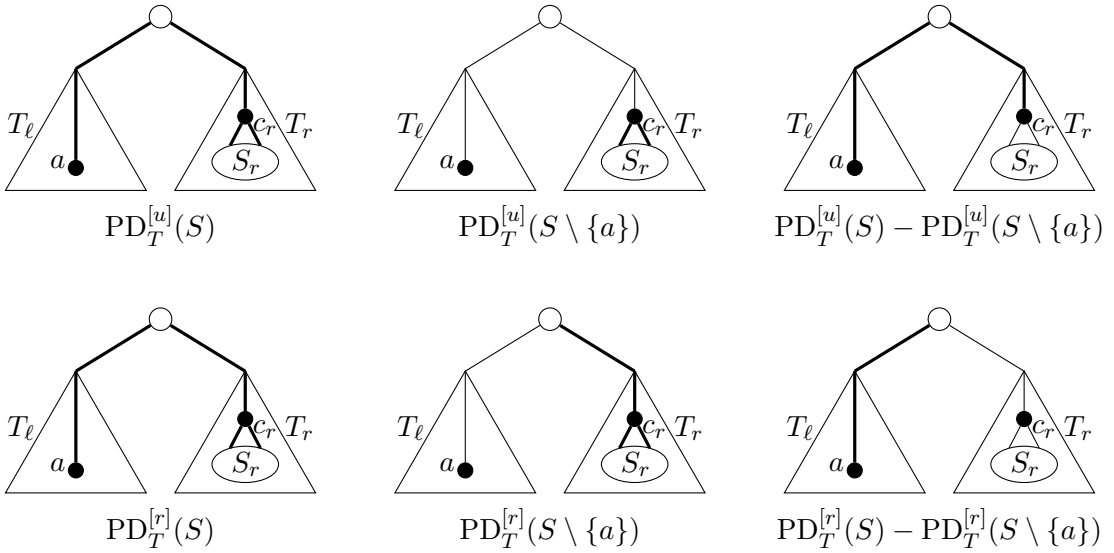


Figure 2: Explanation of $\mathrm{PD}_T^{[\star]}(S) - \mathrm{PD}_T^{[\star]}(S \setminus \{a\})$ for Case 3 in the proof of Proposition 3 where $\star = u$ (first row) or $\star = r$ (second row). The bold parts in the trees visualize the subtree of relevance for the computation of the quantity below the tree.

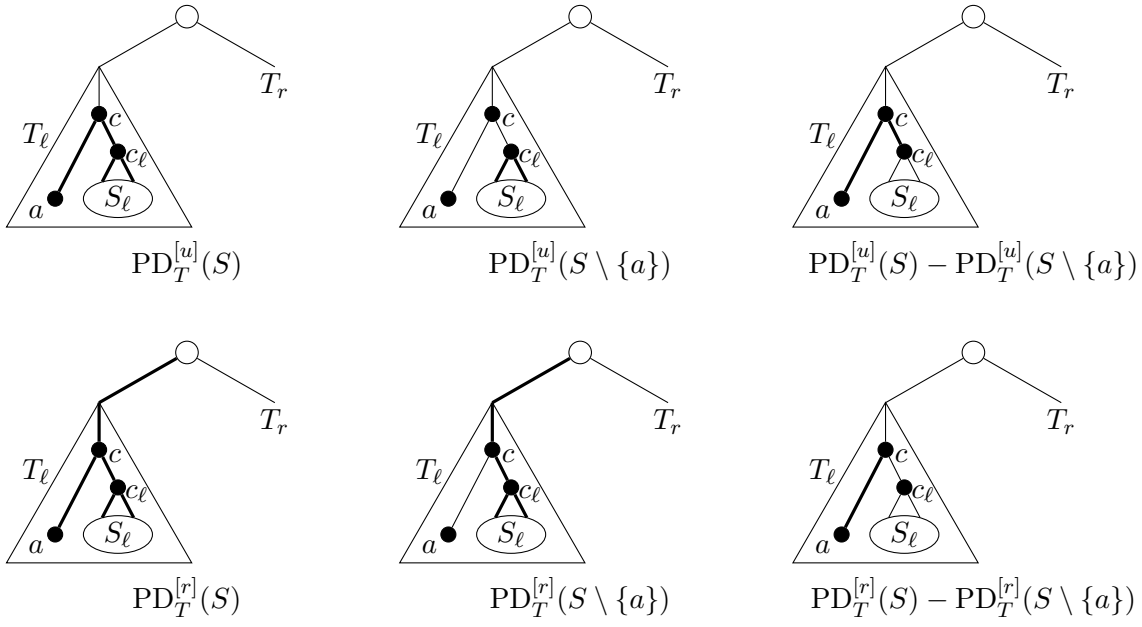


Figure 3: Explanation of $\mathrm{PD}_T^{[\star]}(S) - \mathrm{PD}_T^{[\star]}(S \setminus \{a\})$ for Case 4 in the proof of Proposition 3 where $\star = u$ (first row) or $\star = r$ (second row). The bold parts in the trees visualize the subtree of relevance for the computation of the quantity below the tree.

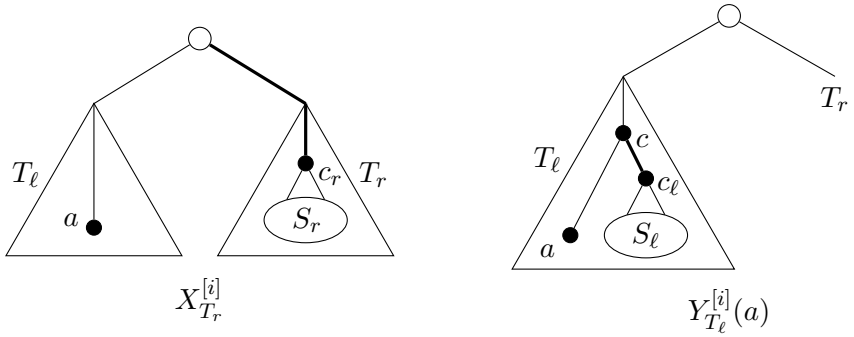


Figure 4: The contributions of Case 3 (left) and Case 4 (right) to the expression (10).

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 at the end of the introduction) which was useful from a linear algebra point of view. Our expression, on the other hand, is of combinatorial nature and useful from a computational point of view.

To see this, recall that by Fuchs and Jin (2015), we have that $\mathrm{SV}^{[r]}_T(a) = \mathrm{FP}_T(a)$ and the fair proportion index can be computed recursively:

$$\mathrm{FP}_T(a) = \begin{cases} \mathrm{FP}_{T_\ell}(a) + 1/|T_\ell|, & \text{if } a \in T_\ell; \\ \mathrm{FP}_{T_r}(a) + 1/|T_r|, & \text{if } a \in T_r \end{cases}$$

with $\mathrm{FP}_T(a) = 0$ when $|T| = 1$; compare with (5). Likewise, we have for $D_T(a)$:

$$D_T(a) = \begin{cases} D_{T_\ell}(a) + 1, & \text{if } a \in T_\ell; \\ D_{T_r}(a) + 1, & \text{if } a \in T_r \end{cases}$$

with $D_T(a) = 0$ when $|T| = 1$; see (4).

Moreover, we have also similar recurrences for $X^{[i]}_T$ and $Y^{[i]}_T(a)$, namely,

$$X^{[i]}_T = X^{[i]}_{T_\ell} + X^{[i]}_{T_r} + \binom{|T_\ell|}{i} + \binom{|T_r|}{i}$$

with $X^{[i]}_T = 0$ for all T with $|T| \leq i$ and

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

with $Y^{[i]}_T(a) = 0$ for all T with $|T| \leq i$. Here, note that in both cases, we only have to consider sets of taxa which are either completely contained in the left or right subtree because all other sets of taxa do not contribute to $X^{[i]}_T$ and $Y^{[i]}_T(a)$. 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 $\mathrm{SV}^{[u]}_T(a)$. In particular, this method is faster than doing the computation directly from the definition of $\mathrm{SV}^{[u]}_T(a)$. We will use it in order to produce some numerical results in Section 5.

4 Proof of Theorem 1

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 4. *If*

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

then

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

Proof. First consider

$$\begin{aligned}\mathrm{Cov}(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) &= \mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n + \mathrm{FP}_n, \mathrm{FP}_n) \\ &= \mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n, \mathrm{FP}_n) + \mathrm{Var}(\mathrm{FP}_n). \end{aligned} \tag{11}$$

Note that the first term in (11) can be bounded by Cauchy-Schwartz inequality as

$$\mathrm{Cov}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n, \mathrm{FP}_n) \leq \sqrt{\mathrm{Var}(\mathrm{SV}_n^{[u]} - \mathrm{FP}_n) \cdot \mathrm{Var}(\mathrm{FP}_n)}. \tag{12}$$

Also, recall that from Corollary 3,

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

where c is a suitable positive constant. Thus, from this, (12) and the assumption, we obtain that

$$\mathrm{Cov}(\mathrm{SV}_n^{[u]}, \mathrm{FP}_n) = \mathrm{Var}(\mathrm{FP}_n)(1 + o(1)). \tag{13}$$

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

$$\mathrm{Var}(\mathrm{SV}_n^{[u]}) = \mathrm{Var}(\mathrm{FP}_n)(1 + o(1)). \tag{14}$$

Finally, by combining (13) and (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. ∎

Before we go on, we need the following simple inequality for conditional expectations.

Lemma 2. *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 4 once more.

We first need some notations. Let the three terms on the right hand side of (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 3 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) and $X_n^{[i]}$ and $Y_n^{[i]}$ denote the random versions of $X_T^{[i]}$ and $Y_T^{[i]}(a)$. (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 (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

$$\mathbb{P}(\mathbf{Y}_n = (j, \text{left})) = p_{n,j}\frac{j}{n} \tag{15}$$

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 5. *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

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

Applying Lemma 2 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 6. *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 2,

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

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 (15),

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

as claimed. ∎

The other two conditions of Proposition 5 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 3. *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 5 which then completes the proof of Theorem 1.

Proposition 7. *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

$$\begin{aligned}\mathbb{E}((Z_n^{[2]})^2|\mathbf{Y}_n=(j,\text{left})) &= \frac{1}{n!^2}\mathbb{E}\left(\sum_{i=1}^{n-j} i!(n-i-1)!\left(X_{n-j}^{[i]}+\binom{n-j}{i}\right)\right)^2\\ &\leq \frac{1}{n!^2}\left(\sum_{i=1}^{n-j} i!(n-i-1)!\sqrt{\mathbb{E}\left(X_{n-j}^{[i]}+\binom{n-j}{i}\right)^2}\right)^2, \qquad (16)\end{aligned}$$

where in the last step we used Lemma 2. What is under the square-root can be written as

$$\mathbb{E}\left(X_{n-j}^{[i]}+\binom{n-j}{i}\right)^2 = \mathbb{E}(X_{n-j}^{[i]})^2+2\binom{n-j}{i}\mathbb{E}(X_{n-j}^{[i]})+\binom{n-j}{i}^2. \qquad (17)$$

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

$$X_n^{[i]} \leq \frac{1}{i}\binom{n-1}{i-1} S_n, \tag{18}$$

where S_n is the Sackin index from Section 2. This upper bound is explained as follows: every taxon is contained in $\binom{n-1}{i-1}$ subsets of taxa of size i. Thus,

$$\binom{n-1}{i-1} S_n \tag{19}$$

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 (19) is at least as large as $X_n^{[i]}$.

Now, plugging (18) into (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 1, 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 (16) gives

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

where in the last step we used Lemma 3. A similar expression holds if $x = \text{left}$ is replaced by $x = \text{right}$.

Finally, using (15) gives

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

where in the last step we used Lemma 1. 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

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

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 (15), we have

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

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. ∎

5 Numerical Data

In this section, we present some numerical data to illustrate Theorem 1. For this data, we used the splitting probabilities (2) to generate a random phylogenetic tree. Then, we picked a taxon 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 3. 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 5 and the code is available at

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

Note that the case $\beta = -1$ is not covered by Theorem 1. 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 1 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 7 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 5 seems to suggest that concentration on the line $y = x$ also takes place in this case.)

6 Conclusion

In recent years, several versions of the Shapley value have been introduced in biodiversity and there has been some confusion about which author used which value. The first author of this paper also added to this confusion by defining in a joint paper with Jin the *modified rooted Shapley value* whose correlation coefficient with the fair proportion index tends to 1 for a random phylogenetic tree under the Yule-Harding model and uniform model where the taxon is also chosen uniformly at random. Fuchs and Jin (2015) claimed that this explains the data presented by Hartmann (2013). However, it was pointed out recently that Hartman used in fact the *unrooted Shapley value*. Consequently, we studied the correlation of the unrooted Shapley value and fair proportion index in this paper.

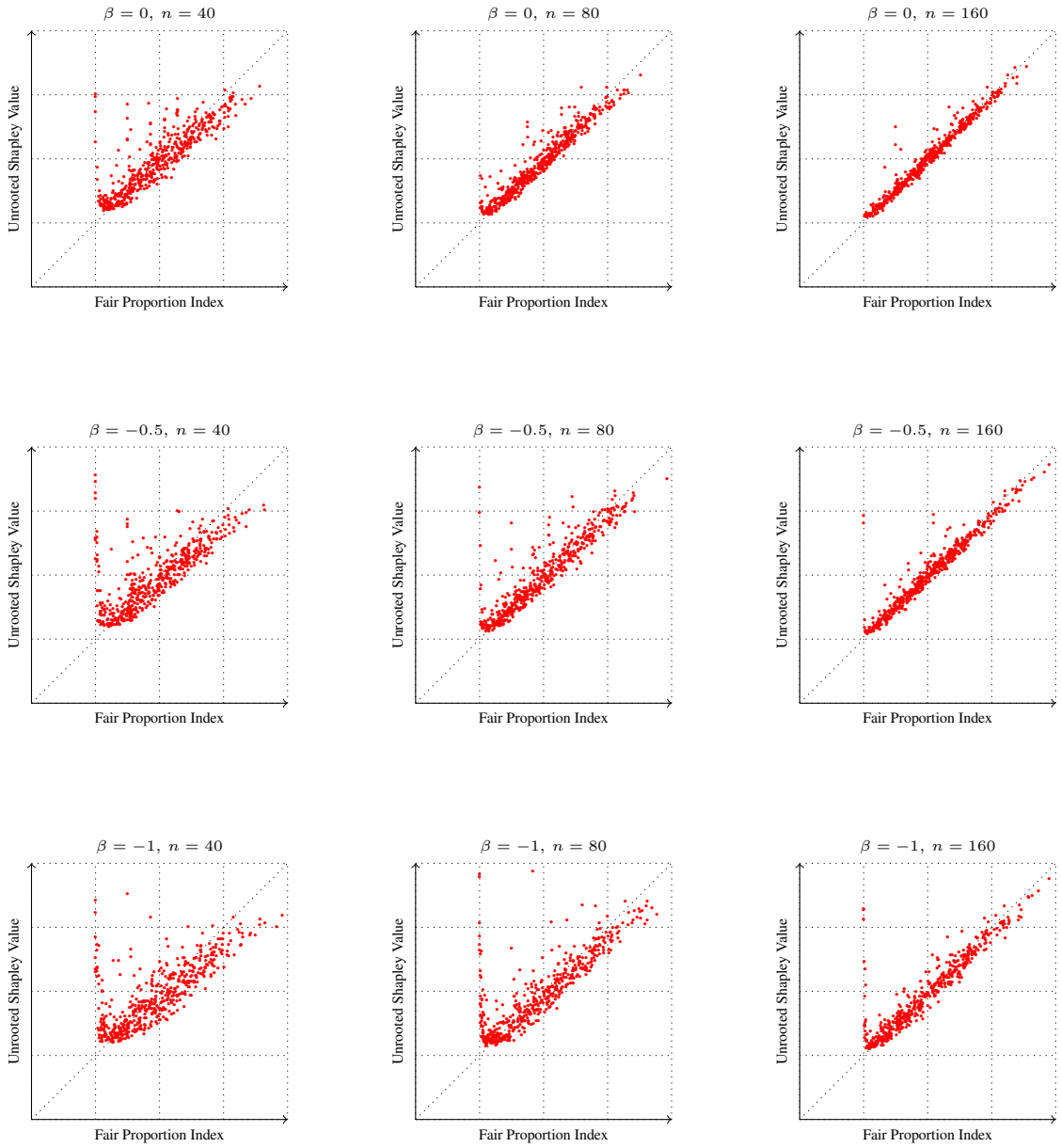


Figure 5: Numerical data for $\beta = 0$ (Yule-Harding model), $\beta = -1/2$ and $\beta = -1$. For each choice of β and n (indicated at the top of the plots), five hundred random trees were generated and for each of them a taxon was picked uniformly at random from the set of all taxa. For these random taxa, the corresponding pair of unrooted Shapley value and fair proportion index was computed by the recursive method from Section 3 and then plotted. 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 (10) since the only negative term on the right hand side is also small in the former case and large in the latter case.

Our main result states that the correlation coefficient of these two indices also tends to one as the number of taxa tends to infinity. We proved this result for a random phylogenetic tree generated by the β-splitting model with $\beta > -1$. The latter restriction is necessary for our method to work and (unfortunately) excludes the important case $\beta = -1$; see Blum and François (2006). (Also, the case $\beta = -3/2$ is not covered which corresponds to the uniform model; see Aldous (1996) and Steel (2016).)

We briefly outline why our method cannot cover the case $\beta = -1$. First, in Section 2, we studied the fundamental recurrence (6) only for $\beta > -1$. However, one could extend our analysis also to $\beta = -1$ where the splitting probabilities (2) take the form:

$$p_{n,j} = \frac{n}{H_{n-1}} \cdot \frac{1}{j(n-j)}, \qquad \text{for } 1 \le j \le n-1$$

with $H_n = \sum_{j=1}^{n} 1/j$ the n-th Harmonic number. Then, using similar methods as in Section 2, one obtains, e.g., that

$$\mathbb{E}(S_n) = \mathcal{O}(n(\log n)^2) \qquad \text{and} \qquad \mathbb{E}(S_n^2) = \mathcal{O}(n^2(\log n)^4).$$

From this, by using the same method as in Section 4, one obtains for the second bound in Proposition 7:

$$\mathbb{E}(\mathbb{E}((Z_n^{[3]})^2|\mathbf{Y}_n)) = \mathcal{O}\left(\frac{(\log n)^4}{n^2 H_{n-1}} \sum_{j=1}^{n-1} \frac{j^2}{(n-j)^3}\right) = \mathcal{O}((\log n)^3).$$

This bound is, however, too large and in particular does not tend to 0 anymore. (Note that the main contribution clearly comes from the third term in (10) since a is likely to be chosen from the larger subtree of the root and thus $X_{T_r}^{[i]}$ is expected to be small whereas $Y_{T_\ell}^{[i]}(a)$ is not.)

Overall, a new method of proof is necessary if one wants to show that the conclusion of Theorem 1 also holds for $\beta = -1$. Indeed, proving such a result would be an important task since the β-splitting model with $\beta = -1$ has the best fit with many real-world applications (see Blum and François (2006)). Thus, such a result would really give strong support to the widely used practice in biodiversity to use the fair proportion index instead of the Shapley value.

We conclude by briefly discussing another natural question arising from our study: what happens if one allows weights which are not necessarily all equal to 1? However, in order to answer this question, one first has to come up with a useful model of assigning weights to the edges of our random trees. In fact, this point is crucial since we are not aware of any model in phylogenetics for this (all the standard models produce random trees without weights). For instance, if one assign random weights independent of the random process that generates the trees and independent of the number of taxa, then many of the ideas of this paper should work as well. However, we suspect that weights which do depend on the number of taxa are more interesting, but then again, it is not clear to us what a useful random model for producing such weights should be.

Acknowledgments

We thank both reviewers for a careful reading and many insightful comments which led to an improvement of the paper. We also acknowledge support by the Ministry of Science, Taiwan under the grants MOST-104-2923-M-009-006-MY3 and MOST-107-2115-M-009-010-MY2.

References

Aldous, D. (1996). Probability distributions on cladograms. *Random discrete structures (Minneapolis, MN, 1993)*, 76:1–18.

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

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.

Roura, S. (2001). Improved master theorems for divide-and-conquer recurrences. *J. ACM*, 48(2):170–205.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40.

Stahn, H. (2019). Biodiversity, shapley value and phylogenetic trees: Some remarks. *J. Math. Biol.*, to appear.

Steel, M. (2016). *Phylogeny-discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

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.