The expected value of the squared cophenetic metric under the Yule and the uniform models

Gabriel Cardona, Arnau Mir, Francesc Rosselló*, Lucía Rotger

Department of Mathematics and Computer Science, University of the Balearic Islands, E-07122 Palma de Mallorca, Spain

Abstract

The cophenetic metrics $d_{\varphi,p}$, for $p \in \{0\} \cup [1, \infty[$, are a recent addition to the kit of available distances for the comparison of phylogenetic trees. Based on a fifty years old idea of Sokal and Rohlf, these metrics compare phylogenetic trees on a same set of taxa by encoding them by means of their vectors of cophenetic values of pairs of taxa and depths of single taxa, and then computing the L^p norm of the difference of the corresponding vectors. In this paper we compute the expected value of the square of $d_{\varphi,2}$ on the space of fully resolved rooted phylogenetic trees with n leaves, under the Yule and the uniform probability distributions.

Key words: Phylogenetic tree; Cophenetic metric; Uniform model; Yule model; Sackin index; Total cophenetic index

1. Introduction

The definition and study of metrics for the comparison of rooted phylogenetic trees on the same set of taxa is a classical problem in phylogenetics [11, Ch. 30], and many metrics have been introduced so far with this purpose. A recent addition to the set of metrics available in this context are the *cophenetic metrics* $d_{\varphi,p}$ introduced in [8]; see also [18] for a related metric. Based on a fifty years old idea of Sokal and Rohlf, the cophenetic metrics compare phylogenetic trees on a same set of taxa by first encoding the trees by means of their vectors of cophenetic values of pairs of taxa and depths of single taxa, and then computing the L^p norm of the difference of the corresponding vectors.

Once the disimilarity between two phylogenetic trees has been computed through a given metric, it is convenient in many situations to assess its significance. One possibility is to compare the value obtained with its expected, or

*Corresponding author

Email addresses: gabriel.cardona@uib.es (Gabriel Cardona), arnau.mir@uib.es (Arnau Mir), cesc.rossello@uib.es (Francesc Rosselló), lucia.rotger@uib.es (Lucía Rotger)

mean, value: is it much larger, much smaller, similar? [28] This makes it necessary to study the distribution of the metric, or, at least, to have a formula for the expected value of the metric for any number n of leaves. The distribution of several metrics has been studied so far: see, for instance, [5, 6, 10, 17, 19, 28].

The expected value of a distance depends on the probability distribution on the space of phylogenetic trees under consideration. The most popular distribution on the space $\mathcal{T}_n$ of binary phylogenetic trees with n leaves is the uniform distribution, under which all trees in $\mathcal{T}_n$ are equiprobable. But phylogeneticists consider also other probability distributions on $\mathcal{T}_n$, defined through stochastic models of evolution [11, Ch. 33]. The most popular is the so-called Yule model [15, 29], defined by an evolutionary process where, at each step, each currently extant species can give rise, with the same probability, to two new species. Under this model, different phylogenetic trees with the same number of leaves may have different probabilities, which depend on their shape.

In this paper we provide explicit formulas for the expected values of the square of the cophenetic metric $d_{\varphi,2}$ under the uniform and the Yule models. The proofs of these formulas are based on long and tedious algebraic computations and thus, to ease the task of the reader interested only in the formulas and the path leading to them, but not in the details, we have moved these computations to an Appendix at the end of the paper.

The spread of $d_{\varphi,2}^2$ around its expected value can be quantified by means of its variance. Unfortunately, we have not been able so far to derive an exact formula for this variance under any model. So, in §4 we provide instead an accurate estimation of its order, both under the uniform and the Yule models, based on simulations.

Besides the aforementioned application of this value in the assessment of tree comparisons, the knowledge of formulas for the expected value of $d_{\varphi,2}^2$ under different models may allow the use of $d_{\varphi,2}$ to test stochastic models of tree growth, a popular line of research in the last years which so far has been mostly based on shape indices; see, for instance, [3, 21]. As a proof of concept, in §5 we report on a basic, preliminary such test performed on the binary phylogenetic trees contained in the TreeBASE database [22].

2. Preliminaries

In this paper, by a *phylogenetic tree* on a set S of taxa we mean a fully resolved, or binary, rooted tree with its leaves bijectively labeled in S. We understand such a rooted tree as a directed graph, with its arcs pointing away from the root. To simplify the language, we shall always identify a leaf of a phylogenetic tree with its label. We shall also use the term *phylogenetic tree with n leaves* to refer to a phylogenetic tree on the set $\{1,\ldots,n\}$. We shall denote by $\mathcal{T}(S)$ the space of all phylogenetic trees on S and by $\mathcal{T}_n$ the space of all phylogenetic trees with n leaves.

Let T be a phylogenetic tree. If there exists a directed path from u to v in T, we shall say that v is a *descendant* of u and also that u is an *ancestor* of

v. The *lowest common ancestor* $\mathrm{LCA}_T(u,v)$ of a pair of nodes u,v in T is the unique common ancestor of them that is a descendant of every other common ancestor of them. The *depth* $\delta_T(v)$ of a node v in T is the distance (in number of arcs) from the root of T to v. The *cophenetic value* $\varphi_T(i,j)$ of a pair of leaves i,j in T is the depth of their LCA. To simplify the notations, we shall often write $\varphi_T(i,i)$ to denote the depth $\delta_T(i)$ of a leaf i.

Given two phylogenetic trees T,T' on disjoint sets of taxa S,S', respectively, we shall denote by $T\,\widehat{}\,T'$ the phylogenetic tree on $S\cup S'$ obtained by connecting the roots of T and T' to a (new) common root. Every phylogenetic tree $T\in\mathcal{T}_n$ is obtained as $T_k\,\widehat{}\,T'_{n-k}$, for some $1\leqslant k\leqslant n-1$, some subset $S_k\subseteq\{1,\ldots,n\}$ with k elements, some tree T_k on S_k and some tree T'_{n-k} on $S_k^c=\{1,\ldots,n\}\setminus S_k$. Actually, every phylogenetic tree in $\mathcal{T}_n$ is obtained in this way twice.

The *Yule*, or *Equal-Rate Markov*, model of evolution [15, 29] is a stochastic model of phylogenetic trees' growth. It starts with a node, and at every step a leaf is chosen randomly and uniformly and it is splitted into two leaves. Finally, the labels are assigned randomly and uniformly to the leaves once the desired number of leaves is reached. This corresponds to a model of evolution where, at each step, each currently extant species can give rise, with the same probability, to two new species. Under this stochastic model, if $T\in\mathcal{T}_n$ is a phylogenetic tree with set of internal nodes $V_{int}(T)$, and if for every $v\in V_{int}(T)$ we denote by $\ell_T(v)$ the number of its descendant leaves, then the probability of T is [4, 27]

$$P_Y(T)=\frac{2^{n-1}}{n!}\prod_{v\in V_{int}(T)}\frac{1}{\ell_T(v)-1}.$$

The *uniform*, or *Proportional to Distinguishable Arrangements*, model [24] is another stochastic model of phylogenetic trees' growth. Unlike the Yule model, its main feature is that all phylogenetic trees $T\in\mathcal{T}_n$ have the same probability:

$$P_U(T)=\frac{1}{(2n-3)!!},\text{ where } (2n-3)!!=(2n-3)(2n-5)\cdots 3\cdot 1.$$

From the point of view of tree growth, this model is described as the process that starts with a node labeled 1 and then, at the k-th step, a new pendant arc, ending in the leaf labeled $k+1$, is added either to a new root (whose other child will be, then, the original root) or to some edge, with all possible locations of this new pendant arc being equiprobable [9, 26]. Although this is not an explicit model of evolution, only of tree growth, several interpretations of it in terms of evolutionary processes have been given in the literature: see [3, p. 686] and the references therein.

3. Main results

Let $T\in\mathcal{T}_n$ be a phylogenetic tree with n leaves. The *cophenetic vector* $\varphi(T)$ of T is the vector consisting of, for each leaf, its depth, and for each pair of different leaves, the depth of their LCA. Formally,

$$\varphi(T)=\big(\varphi_T(i,j)\big)_{1\leqslant i\leqslant j\leqslant n}\in\mathbb{R}^{n(n+1)/2},$$

with its elements lexicographically ordered in (i, j). It turns out [8] that the mapping

$$\varphi : \mathcal{T}_n \to \mathbb{R}^{n(n+1)/2}$$

that sends each $T \in \mathcal{T}_n$ to its cophenetic vector $\varphi(T)$, is injective up to isomorphism. As it is well known, this allows to induce metrics on $\mathcal{T}_n$ from metrics defined on powers of $\mathbb{R}$. In particular, in this paper we consider the *cophenetic metric* $d_{\varphi,2}$ on $\mathcal{T}_n$ induced by the euclidean distance:

$$d_{\varphi,2}(T_1, T_2) = \sqrt{\sum_{1\leqslant i\leqslant j\leqslant n} (\varphi_{T_1}(i, j) - \varphi_{T_2}(i, j))^2}.$$

Example 1. Consider the phylogenetic trees $T, T' \in \mathcal{T}_4$ depicted in Fig. 1. Their total cophenetic vectors are

$$\begin{aligned} \varphi(T) \; &= (2, 1, 0, 0, 2, 0, 0, 2, 1, 2) \\ \varphi(T') &= (1, 0, 0, 0, 2, 1, 1, 3, 2, 3) \end{aligned}$$

and therefore $d_{\varphi,2}(T, T') = \sqrt{7}$.

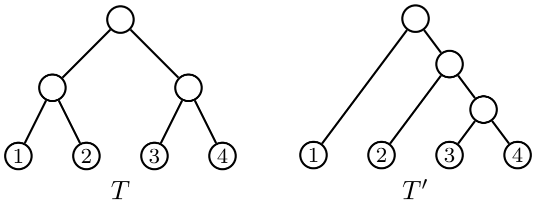

Figure 1: Two phylogenetic trees with 4 leaves.

Let D^2_n the random variable that chooses a pair of trees $T, T' \in \mathcal{T}_n$ and computes $d_{\varphi,2}(T, T')^2$. Its expected values under the Yule and the uniform models are given by the following two theorems. Recall that the n-th *harmonic number* H_n is defined as $H_n = \sum_{i=1}^{n} 1/i$.

Theorem 2. *For every $n \geqslant 2$, the expected value of D^2_n under the Yule model is*

$$E_Y(D^2_n) = \frac{2n}{n-1}\big(3n^2 - 10n - 1 + 8(n+1)H_n - 4(n+1)H^2_n\big).$$

Theorem 3. *For every $n \geqslant 2$, the expected value of D^2_n under the uniform model is*

$$E_U(D^2_n) = \frac{1}{3}(4n^3 + 18n^2 - 10n) - \frac{n(n+3)}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} - \frac{n(n+7)}{4}\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2$$

Since $H_n \sim \ln(n)$ and $(2n-2)!!/(2n-3)!! \sim \sqrt{\pi n}$, these formulas imply that

$$E_Y(D_n^2) \sim 6n^2, \quad E_U(D_n^2) \sim \Big(\frac{4}{3} - \frac{\pi}{4}\Big)n^3.$$

We prove the formulas in Theorems 2 and 3 by reducing the computation of the expected value of D_n^2 to that of the following random variables:

- S_n, the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes its Sackin index [25]
$$S(T) = \sum_{i=1}^{n} \delta_T(i)$$
- Φ_n, the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes its total cophenetic index [20], defined by
$$\Phi(T) = \sum_{1 \leqslant i < j \leqslant n} \varphi_T(i,j).$$
- $\overline{\Phi}_n^{(2)}$, the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes
$$\overline{\Phi}^{(2)}(T) = \sum_{1 \leqslant i \leqslant j \leqslant n} \varphi_T(i,j)^2$$

For the models under consideration, the expected values of these variables are related to that of D_n^2 by the next proposition. In it and henceforth, we shall denote by $E(X)$ the expected value of a random variable X on $\mathcal{T}_n$ under a generic probability distribution $p : \mathcal{T}_n \to [0,1]$ on $\mathcal{T}_n$ invariant under relabelings. The probability distributions p_Y and p_U defined by the Yule and the uniform models, respectively, are invariant under relabelings, and therefore the expected values under these specific models, which will be denoted by E_Y and E_U, respectively, are special cases of E.

Proposition 4. $E(D_n^2) = 2E(\overline{\Phi}_n^{(2)}) - 2 \cdot \dfrac{E(S_n)^2}{n} - 4 \cdot \dfrac{E(\Phi_n)^2}{n(n-1)}.$

Proof. To simplify the notations, let

- φ_n be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $\varphi_T(1,2)$.
- δ_n be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $\delta_T(1)$.

Let us compute now $E(D_n^2)$ from its very definition:

$$\begin{aligned}
E(D_n^2) &= \sum_{(T,T')\in\mathcal{T}_n^2} d_{\varphi,2}(T,T')^2 p(T)p(T')\\
&= \sum_{(T,T')\in\mathcal{T}_n^2}\Big(\sum_{1\leqslant i\leqslant j\leqslant n}(\varphi_T(i,j)-\varphi_{T'}(i,j))^2\Big)p(T)p(T')\\
&= \sum_{1\leqslant i\leqslant j\leqslant n}\ \sum_{(T,T')\in\mathcal{T}_n^2}(\varphi_T(i,j)^2+\varphi_{T'}(i,j)^2-2\varphi_T(i,j)\varphi_{T'}(i,j))p(T)p(T')\\
&= \sum_{1\leqslant i\leqslant j\leqslant n}\Big(\sum_{(T,T')\in\mathcal{T}_n^2}\varphi_T(i,j)^2p(T)p(T')+\sum_{(T,T')\in\mathcal{T}_n^2}\varphi_{T'}(i,j)^2p(T)p(T')\\
&\qquad -2\sum_{(T,T')\in\mathcal{T}_n^2}\varphi_T(i,j)\varphi_{T'}(i,j)p(T)p(T')\Big)\\
&= \sum_{1\leqslant i\leqslant j\leqslant n}\Big(\sum_{T\in\mathcal{T}_n}\varphi_T(i,j)^2p(T)+\sum_{T'\in\mathcal{T}_n}\varphi_{T'}(i,j)^2p(T')\\
&\qquad -2\Big(\sum_{T\in\mathcal{T}_n}\varphi_T(i,j)p(T)\Big)\Big(\sum_{T'\in\mathcal{T}_n}\varphi_{T'}(i,j)p(T')\Big)\Big)\\
&= \sum_{1\leqslant i\leqslant j\leqslant n}\Big(2\sum_{T\in\mathcal{T}_n}\varphi_T(i,j)^2p(T)-2\Big(\sum_{T\in\mathcal{T}_n}\varphi_T(i,j)p(T)\Big)^2\Big)\\
&= 2\sum_{T\in\mathcal{T}_n}\Big(\sum_{1\leqslant i\leqslant j\leqslant n}\varphi_T(i,j)^2\Big)p(T)-2\sum_{1\leqslant i<j\leqslant n}\Big(\sum_{T\in\mathcal{T}_n}\varphi_T(i,j)p(T)\Big)^2\\
&\qquad -2\sum_{1\leqslant i\leqslant n}\Big(\sum_{T\in\mathcal{T}_n}\varphi_T(i,i)p(T)\Big)^2\\
&= 2\sum_{T\in\mathcal{T}_n}\overline{\Phi}^{(2)}(T)p(T)-2\binom{n}{2}\Big(\sum_{T\in\mathcal{T}_n}\varphi_T(1,2)p(T)\Big)^2\\
&\qquad -2n\Big(\sum_{T\in\mathcal{T}_n}\delta_T(1)p(T)\Big)^2\\
&= 2E(\overline{\Phi}_n^{(2)})-n(n-1)E(\varphi_n)^2-2nE(\delta_n)^2
\end{aligned}$$

Now, the values of $E(\delta_n)$ and $E(\varphi_n)$ can be easily obtained from $E(S_n)$ and $E(\Phi_n)$, respectively, using the invariance under relabelings of the probability distribution under which we compute the expected values E:

$$E(\delta_n)=E(S_n)/n, \qquad E(\varphi_n)=E(\Phi_n)/\tbinom{n}{2}$$

The formula in the statement is then obtained by replacing $E(\delta_n)$ and $E(\varphi_n)$ by these values. □

The expected values of S_n and Φ_n under the Yule and the uniform models are known:

$$\begin{aligned}
E_Y(S_n) &= 2n(H_n-1) & E_Y(\Phi_n) &= n(n-1)-2n(H_n-1)\\
E_U(S_n) &= n\Big(\frac{(2n-2)!!}{(2n-3)!!}-1\Big) & E_U(\Phi_n) &= \frac{1}{2}\binom{n}{2}\Big(\frac{(2n-2)!!}{(2n-3)!!}-2\Big)
\end{aligned}$$

The formula for $E_Y(S_n)$ was proved in [16] and the other three, in [20].

To obtain the expected values of D_n^2, it remains to compute the expected values of $\overline{\Phi}_n^{(2)}$. They are given by the following result.

Proposition 5. *For every $n \geqslant 2$,*

(a) $E_Y(\overline{\Phi}_n^{(2)}) = 5n(n-1) - 8n(H_n - 1)$

(b) $E_U(\overline{\Phi}_n^{(2)}) = \dfrac{1}{6}n(4n^2+21n-7) - \dfrac{3}{4}n(n+3)\dfrac{(2n-2)!!}{(2n-3)!!}$

This proposition is proved in the Appendix at the end of this paper. Finally, the identities given in Theorems 2 and 3 are obtained by replacing, in the identity given in Proposition 4, $E(S_n)$, $E(\Phi_n)$, and $E(\overline{\Phi}_n^{(2)})$ by their values. We leave the last details to the reader.

4. On the variance of D_n^2

In order to assess the spread of the random variable D_n^2 around its expected value, it is useful to know its variance. Since $\mathrm{Var}(D_n^2) = E(D_n^4) - E(D_n^2)^2$, the computation of this variance involves the computation of the expected value of D_n^4. Developing this expected value as in Proposition 4, one can obtain an expression for $E(D_n^4)$ in the same spirit as the one given for $E(D_n^2)$ therein, but with 24 different terms instead of only 3, and so far we have not been able to convert it, either for the Yule or the uniform model, into a closed formula depending only on n.

Therefore, in order to be able to, at least, estimate the asymptotic order of $E(D_n^4)$, we have taken the Monte Carlo path. More specifically, both for the Yule and the uniform models, and for every $n = 3, \ldots, 100$, we have randomly generated $N = 10000$ pairs of binary trees $(T, T') \in \mathcal{T}_n \times \mathcal{T}_n$, we have computed the value of $d_{\varphi,2}(T,T')^4$ for each such pair (T,T'), and we have computed the arithmetic mean $\overline{D_n^4}$ of these N values. These arithmetic means are estimations of the value of $E(D_n^4)$ under the corresponding model.

Finally, we have computed the slope α of the regression line of $\log(\overline{D_n^4})$ as a function of $\log(n)$ using the values for $n = 50, \ldots, 100$. We have only considered the largest values of n because if smaller values were also included in the regression, the regression coefficient was considerably smaller, due to the fact that, for small n, the dominant term is not large enough to significantly stand out from terms of smaller degree. The results obtained are given in the following table:

Model	α	Regression coefficient R^2
Yule	4.439682	0.9999
Uniform	6.226358	0.9999

The intermediate results of all these computations, as well as the Python and R scripts used to compute them, are available in the Supplementary Material web page `http://bioinfo.uib.es/~recerca/phylotrees/expectedcophdist/`.

From this table, we estimate then that $E_Y(D_n^4) \approx \Theta(n^{4.44})$ and $E_U(D_n^4) \approx \Theta(n^{6.23})$, and the regression coefficients tell us that these orders explain quite well the estimated expected values up to $n = 100$. Since, by Theorems 2 and 3, $E_Y(D_n^2)^2 = \Theta(n^4)$ and $E_U(D_n^2)^2 = \Theta(n^6)$, and hence these asymptotic orders are smaller than those estimated for $E_Y(D_n^4)$ and $E_U(D_n^4)$, we estimate that the asymptotic orders of $\mathrm{Var}(D_n^2)$ under the Yule and the uniform models are the same as those of $E(D_n^4)$.

5. An experiment on TreeBASE

In this section we report on a very simple experiment to show how $d_{\varphi,2}$ can be used to test evolutionary hypotheses. In this experiment, we have compared the expected value of $d_{\varphi,2}^2$ on $\mathcal{T}_n$ under the uniform and the Yule models with its average value on the set $\mathrm{TreeBASE}_{bin,n}$ of binary phylogenetic trees with n leaves contained in TreeBASE [22] (downloaded on December 15, 2015).

To perform this experiment, we had to take some decisions. First, since there are only very few values $n > 50$ such that $|\mathrm{TreeBASE}_{bin,n}| > 10$, we have decided to consider only those binary trees contained in TreeBASE with $n \leqslant 50$ leaves. On the other hand, even for those n such that $\mathrm{TreeBASE}_{bin,n}$ is relatively large, in most cases it does not contain many pairs of trees with the same taxa. So, instead of computing the average value of $d_{\varphi,2}^2$ on $\mathrm{TreeBASE}_{bin,n}$ by averaging the values $d_{\varphi,2}^2(T, T')$ for pairs T, T' with exactly the same n taxa, we have made use of the formula given in Proposition 4, as if $\mathrm{TreeBASE}_{bin,n}$ was closed under relabelings: that is, we have taken only into account the shapes of the trees contained in it. This is consistent with the fact that our final goal is to test models of evolution that produce tree shapes.

So, we have computed the average values of $\overline{\Phi}^{(2)}$, of the Sackin index S, and of the total cophenetic index Φ on $\mathrm{TreeBASE}_{bin,n}$, and we have taken as average value of $d_{\varphi,2}^2$ on this set the result of appying the formula in Proposition 4. Let's call $E_{TrB}(D_n^2)$ the resulting value.

On the other hand, for each $n \leqslant 50$ we have used our estimations of $E_Y(D_n^4)$ and $E_U(D_n^4)$ and the exact values of $E_Y(D_n^2)$ and $E_U(D_n^2)$ to give estimations $sd_Y(D_n^2)$ and $sd_Y(D_n^2)$ of the values of the standard deviations of D_n^2 under both models. Finally, for every $n \leqslant 50$, we have taken the intervals $E_Y(D_n^2) \pm 2 \cdot sd_Y(D_n^2)$ and $E_U(D_n^2) \pm 2 \cdot sd_U(D_n^2)$ as reference intervals for D_n^2 under the Yule and the uniform model. The detailed results of these computations, as well as the Python and R scripts used to compute and analyze them, are also available in the Supplementary Material web page `http://bioinfo.uib.es/~recerca/phylotrees/expectedcophdist/`.

Fig. 2 plots $\log(E_{TrB}(D_n^2))$ as a function of $\log(n)$ (middle, continuous curve). We have added the curves of $\log(E_Y(D_n^2))$ (lower, dotted curve) and $\log(E_U(D_n^2))$ (upper, dashed curve), again as functions of $\log(n)$, and the logarithms of the corresponding reference intervals for D_n^2 (vertical segments). The graphic shows that the expected value of $d_{\varphi,2}^2$ on (the shapes of) the phylogenetic trees contained in TreeBASE is better explained by the uniform model than by the Yule model. This agrees with the results of similar experiments using other measures (see, for instance, [3, 20]).

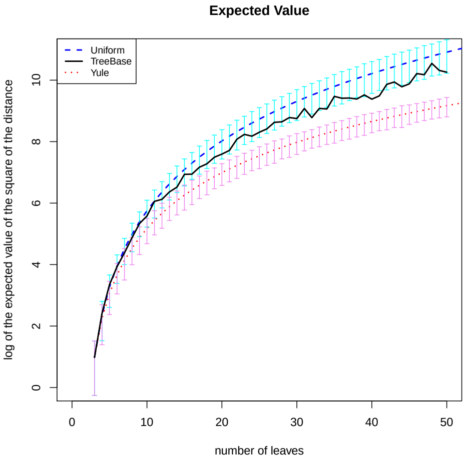

Figure 2: Log plots of the mean of D_n^2 for the binary trees in TreeBASE with a fixed number n of leaves, of $E_Y(D_n^2)$ (red curve) and $E_U(D_n^2)$ (blue curve).

6. Conclusions and discussion

In this paper we have obtained formulas for the expected values under the Yule and the uniform models of the square of the cophenetic metric $d_{\varphi,2}$ induced by the euclidean distance between cophenetic vectors. These formulas are explicit and hold on spaces $\mathcal{T}_n$ of fully resolved phylogenetic trees with any number n of leaves.

These formulas have been obtained through long algebraic manipulations of sums of sequences. To double-check our results, we have computed the exact value of $E_Y(D_n^2)$ and $E_U(D_n^2)$ for $n = 3, \ldots, 7$, by generating all trees with up to 7 leaves. Moreover, we have computed numerical approximations to these values for $n = 10, 20, \ldots, 100$, by generating pairs of random trees until the numerical method stabilizes. These numerical experiments confirm that our formulas give the right figures. Table 1 gives the exact values for $n = 3, \ldots, 7$. The Python scripts used in these computations, as well as a full account of the results, are also available in the Supplementary Material web page `http://bioinfo.uib.es/~recerca/phylotrees/expectedcophdist/`.

The formulas for $E_Y(D_n^2)$ and $E_U(D_n^2)$ grow in different orders: $E_Y(D_n^2)$ is in $\Theta(n^2)$, while $E_U(D_n^2)$ is in $\Theta(n^3)$. Therefore, D_n^2 can be used to test the Yule and the uniform models as null stochastic models of evolution for collections of phylogenetic trees reconstructed by different methods. We have reported on a first experiment of this type, which reinforces the conclusion that “real world” phylogenetic trees (that is, those contained in TreeBASE) are not consistent with the Yule model of evolution. We plan to report in a future paper on more extensive tests on stochastic models of evolutionary processes, including Ford's α-model [12] and Aldous' β-model [2].

	3	4	5	6	7
$E_Y(D_n^2)$	2.66667	9.40741	21.1833	38.712	62.5562
$E_U(D_n^2)$	2.66667	10.56	26.2367	52.3023	91.4086

Table 1: Values of $E_Y(D_n^2)$ and $E_U(D_n^2)$ for $n = 3, \dots, 7$. They agree with those given by our formulas.

We would like to close this paper with a conjecture. As we have seen in §4, from our simulations we have obtained that $E_Y(D_n^4) \approx \Theta(n^{4.44})$ and $E_U(D_n^4) \approx \Theta(n^{6.23})$. Now, the formulas in n usually obtained in the explicit computation of terms like those appearing in the formal development of $E(D_n^4)$ in the spirit of Proposition 4, are linear combinations of square, fourth, and, in general, 2^m-th roots of polynomials in n. Therefore, looking at the exponents in the aforementioned estimated asymptotic orders of $E_Y(D_n^4)$ and $E_U(D_n^4)$, we conjecture that their actual asymptotic orders are $E_Y(D_n^4) = \Theta(n^{4.5})$ and $E_U(D_n^4) = \Theta(n^{6.25})$.

Acknowledgements

The research reported in this paper has been partially supported by the Spanish government and the UE FEDER program, through project DPI2015-67082-P.

References

[1] M. Abramowitz, I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover (1964).

[2] D. Aldous. Probability distributions on cladograms. In: Random discrete structures, IMA Vol. Math. Appl. 76 (Springer,1996), 1–18.

[3] M. G. B. Blum, O. François. Which random processes describe the Tree of Life? A large-scale study of phylogenetic tree imbalance. Sys. Biol. 55 (2006), 685–691.

[4] J. Brown, Probabilities of evolutionary trees. Syst. Biol. 43 (1994), 78–91.

[5] D. Bryant, M. Steel, Computing the distribution of a tree metric. IEEE/ACM Transactions in Computational Biology and Bioinformatics 16 (2009), 420–426.

[6] G. Cardona, A. Mir, F. Rosselló, The expected value under the Yule model of the squared path-difference distance. Applied Mathematics Letters 25 (2012), pp. 2031–2036.

[7] G. Cardona, A. Mir, F. Rosselló, Exact formulas for the variance of several balance indices under the Yule model. Journal of Mathematical Biology, 67 (2013), 1833–1846.

[8] G. Cardona, A. Mir, L. Rotger, F. Rosselló, D. Sánchez, Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. BMC Bioinformatics (2013) 14:3

[9] L. L. Cavalli-Sforza, A. Edwards, Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet., 19 (1967), 233–257.

[10] D. Critchlow, D. Pearl, C. Qian, The triples distance for rooted bifurcating phylogenetic trees. Systematic Biology, 45 (1996), 323–334.

[11] J. Felsenstein, Inferring Phylogenies. Sinauer Associates Inc., 2004.

[12] D. Ford. Probabilities on cladograms: Introduction to the alpha model. arXiv:math/0511246 [math.PR] (2005).

[13] http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric3F2/03/07/02.

[14] http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric2F1/03/03/01.

[15] E. Harding, The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3 (1971), 44–77.

[16] S. B. Heard, Patterns in Tree Balance among Cladistic, Phenetic, and Randomly Generated Phylogenetic Trees. Evolution 46 (1992), 1818–1826.

[17] M. Hendy, C. Little, D. Penny, Comparing Trees with Pendant Vertices Labelled. SIAM J. Applied Mathematics 44 (1984), 1054–1065.

[18] M. Kendall, C. Colijn, Mapping phylogenetic trees to reveal distinct patterns of evolution. Molecular Biology and Evolution 33 (2016), 2735–2743.

[19] A. Mir, F. Rosselló, The mean value of the squared path-difference distance for rooted phylogenetic trees. Journal of Mathematical Analysis and Applications 371 (2010), 168–176.

[20] A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees. Mathematical Biosciences 241 (2013), 125–136.

[21] A. Mooers, S. B. Heard, Inferring evolutionary process from phylogenetic tree shape. Quart. Rev. Biol. 72 (1997), 31–54.

[22] V. Morell, TreeBASE: the roots of phylogeny. Science 273 (1996), 569–560. `http://www.treebase.org`.

[23] M. Petkovsek, H. Wilf, D. Zeilberger, $A = B$. AK Peters Ltd. (1996). Available online at `http://www.math.upenn.edu/~wilf/AeqB.html`.

[24] D. E. Rosen, Vicariant Patterns and Historical Explanation in Biogeography. Syst. Biol. 27 (1978), 159–188.

[25] M. J. Sackin, "Good" and "bad" phenograms. Sys. Zool, 21 (1972), 225–226.

[26] M. Steel, A. McKenzie, Distributions of cherries for two models of trees. Math. Biosc. 164 (2000), 81–92.

[27] M. Steel, A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosc. 170 (2001), 91–112.

[28] M. A. Steel, D. Penny, Distributions of tree comparison metrics—some new results, Syst. Biol. 42 (2) (1993) 126–141.

[29] G. U. Yule, A mathematical theory of evolution based on the conclusions of Dr J. C. Willis. Phil. Trans. Royal Soc. (London) Series B 213 (1924), 21–87.

Appendix: Proof of Proposition 5

Proof of Proposition 5.(a)

For every $T \in \mathcal{T}_n$, let

$$\overline{\Phi}(T) = S(T) + \Phi(T) = \sum_{1\leqslant i\leqslant j\leqslant n} \varphi_T(i,j),$$

and let $\overline{\Phi}_n$ be the random variable that chooses a tree $T \in \mathcal{T}_n$ and computes $\overline{\Phi}(T)$. We have that

$$E_Y(\overline{\Phi}_n) = E_Y(S_n) + E_Y(\Phi_n) = n(n-1).$$

To compute $E_Y(\overline{\Phi}_n^{(2)})$, we shall use an argument similar to the one used in the proof of [6, Prop. 3]. Notice that

$$\begin{aligned} E_Y(\overline{\Phi}_n^{(2)}) &= \sum_{T\in\mathcal{T}_n} \overline{\Phi}^{(2)}(T)\cdot p_Y(T) \\ &= \frac{1}{2}\sum_{k=1}^{n-1} \sum_{\substack{S_k\subsetneq\{1,\ldots,n\}\\ |S_k|=k}} \sum_{T_k\in\mathcal{T}(S_k)} \sum_{T'_{n-k}\in\mathcal{T}(S_k^c)} \overline{\Phi}^{(2)}(T_k\widehat{\ }T'_{n-k})\cdot p_Y(T_k\widehat{\ }T'_{n-k}) \end{aligned}$$

Now, on the one hand, we have the following easy lemma on $P_Y(T\widehat{\ }T')$: see [7, Lem. 1].

Lemma 6. *Let* $\emptyset \neq S_k \subsetneq \{1,\ldots,n\}$ *with* $|S_k| = k$*, let* $T_k \in \mathcal{T}(S_k)$ *and* $T'_{n-k} \in \mathcal{T}(S_k^c)$*. Then,*

$$P_Y(T_k\widehat{\ }T'_{n-k}) = \frac{2}{(n-1)\binom{n}{k}} P(T_k)P(T'_{n-k}).$$

On the other hand, we have the following recursive expression for $\overline{\Phi}^{(2)}(T\widehat{\ }T')$.

Lemma 7. *Let* $\emptyset \neq S_k \subsetneq \{1,\ldots,n\}$ *with* $|S_k| = k$*, let* $T_k \in \mathcal{T}(S_k)$ *and* $T'_{n-k} \in \mathcal{T}(S_k^c)$*. Then*

$$\overline{\Phi}^{(2)}(T_k\widehat{\ }T'_{n-k}) = \overline{\Phi}^{(2)}(T_k)+\overline{\Phi}^{(2)}(T'_{n-k})+2\overline{\Phi}(T_k)+2\overline{\Phi}(T'_{n-k})+\binom{k+1}{2}+\binom{n-k+1}{2}.$$

Proof. Let us assume, without any loss of generality, that $S = \{1,\ldots,m\}$ and $S' = \{m+1,\ldots,n\}$. Then

$$\varphi_{T_k\widehat{\ }T'_{n-k}}(i,j) = \begin{cases} \varphi_{T_k}(i,j)+1 & \text{if } 1 \leqslant i,j \leqslant k \\ \varphi_{T'_{n-k}}(i,j)+1 & \text{if } k+1 \leqslant i,j \leqslant n \\ 0 & \text{otherwise} \end{cases}$$

and therefore

$$\begin{aligned} \overline{\Phi}^{(2)}(T_k\widehat{\ }T'_{n-k}) &= \sum_{1\leqslant i\leqslant j\leqslant n} \varphi_{T_k\widehat{\ }T'_{n-k}}(i,j)^2 \\ &= \sum_{1\leqslant i\leqslant j\leqslant k} (\varphi_{T_k}(i,j)+1)^2 + \sum_{k+1\leqslant i\leqslant j\leqslant n} (\varphi_{T'_{n-k}}(i,j)+1)^2 \\ &= \sum_{1\leqslant i\leqslant j\leqslant k} (\varphi_{T_k}(i,j)^2 + 2\varphi_{T_k}(i,j)+1) + \sum_{k+1\leqslant i\leqslant j\leqslant n} (\varphi_{T'_{n-k}}(i,j)^2 + 2\varphi_{T'_{n-k}}(i,j)+1) \\ &= \overline{\Phi}^{(2)}(T_k) + 2\overline{\Phi}(T_k) + \binom{k+1}{2} + \overline{\Phi}^{(2)}(T'_{n-k}) + 2\overline{\Phi}(T'_{n-k}) + \binom{n-k+1}{2}. \end{aligned}$$

□

So, if we set

$$f(a,b) = \binom{a+1}{2} + \binom{b+1}{2},$$

we have that

$$
\begin{aligned}
&E_Y(\overline{\Phi}_n^{(2)})\\
&\quad= \frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}\sum_{T_k\in\mathcal{T}_k}\sum_{T'_{n-k}\in\mathcal{T}_{n-k}}\Big[\overline{\Phi}^{(2)}(T_k)+\overline{\Phi}^{(2)}(T'_{n-k})+2(\overline{\Phi}(T_k)+\overline{\Phi}(T'_{n-k}))\\
&\qquad+f(k,n-k)\Big]\frac{2}{(n-1)\binom{n}{k}}P_Y(T_k)P_Y(T'_{n-k})\\
&\quad= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big[\sum_{T_k}\sum_{T'_{n-k}}\overline{\Phi}^{(2)}(T_k)P_Y(T_k)P_Y(T'_{n-k})\\
&\qquad+\sum_{T_k}\sum_{T'_{n-k}}\overline{\Phi}^{(2)}(T'_{n-k})P_Y(T_k)P_Y(T'_{n-k})\\
&\qquad+2\sum_{T_k}\sum_{T'_{n-k}}\overline{\Phi}(T_k)P_Y(T_k)P_Y(T'_{n-k})\\
&\qquad+2\sum_{T_k}\sum_{T'_{n-k}}\overline{\Phi}(T'_{n-k})P_Y(T_k)P_Y(T'_{n-k})\\
&\qquad+\sum_{T_k}\sum_{T'_{n-k}}f(k,n-k)P_Y(T_k)P_Y(T'_{n-k})\Big]\\
&\quad= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big[\sum_{T_k}\overline{\Phi}^{(2)}(T_k)P_Y(T_k)+\sum_{T'_{n-k}}\overline{\Phi}^{(2)}(T'_{n-k})P_Y(T'_{n-k})\\
&\qquad\quad+2\sum_{T_k}\overline{\Phi}(T_k)P_Y(T_k)+2\sum_{T'_{n-k}}\overline{\Phi}(T'_{n-k})P_Y(T'_{n-k})+f(k,n-k)\Big]\\
&\qquad= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big[E_Y(\overline{\Phi}_k^{(2)})+E_Y(\overline{\Phi}_{n-k}^{(2)})+2E_Y(\overline{\Phi}_k)+2E_Y(\overline{\Phi}_{n-k})\\
&\qquad\quad+\binom{k+1}{2}+\binom{n-k+1}{2}\Big]\\
&\qquad= \frac{2}{n-1}\sum_{k=1}^{n-1}E_Y(\overline{\Phi}_k^{(2)})+\frac{4}{n-1}\sum_{k=1}^{n-1}E_Y(\overline{\Phi}_k)+\frac{1}{3}n(n+1).
\end{aligned}
$$

In particular

$$E_Y(\overline{\Phi}_{n-1}^{2}) = \frac{2}{n-2}\sum_{k=1}^{n-2}E_Y(\overline{\Phi}_k^{(2)})+\frac{4}{n-2}\sum_{k=1}^{n-2}E_Y(\overline{\Phi}_k)+\frac{1}{3}n(n-1).$$

and therefore

$$\begin{aligned}
E_Y(\overline{\Phi}^{(2)}_n) &= \frac{n-2}{n-1}\cdot\frac{2}{n-2}\sum_{k=1}^{n-2}E_Y(\overline{\Phi}^{(2)}_k)+\frac{2}{n-1}E_Y(\overline{\Phi}^{(2)}_{n-1})\\
&\quad+\frac{n-2}{n-1}\cdot\frac{4}{n-2}\sum_{k=1}^{n-2}E_Y(\overline{\Phi}_k)+\frac{4}{n-1}E_Y(\overline{\Phi}_{n-1})\\
&\quad+\frac{n-2}{n-1}\cdot\frac{1}{3}n(n-1)+n\\
&=\frac{n-2}{n-1}E_Y(\overline{\Phi}^{(2)}_{n-1})+\frac{2}{n-1}E_Y(\overline{\Phi}^{(2)}_{n-1})+\frac{4}{n-1}E_Y(\overline{\Phi}_{n-1})+n\\
&=\frac{n}{n-1}E_Y(\overline{\Phi}^{(2)}_{n-1})+5n-8.
\end{aligned}$$

Setting $x_n = E_Y(\overline{\Phi}^{(2)}_n)/n$, this recurrence becomes

$$x_n = x_{n-1}+5-\frac{8}{n}$$

and the solution of this recursive equation with $x_1 = E_Y(\overline{\Phi}^{(2)}_1) = 0$ is

$$x_n = \sum_{k=2}^{n}\Big(5-\frac{8}{k}\Big) = 5(n-1)-8(H_n-1) = 5n+3-8H_n$$

from where we deduce that $E_Y(\overline{\Phi}^{(2)}_n) = 5n^2+3n-8nH_n$, as we claimed.

Proof of Proposition 5.(b)

To compute $E_U(\overline{\Phi}^{(2)}_n)$, we shall use an argument similar to the one used in [19]. For every $k = 1,\ldots,n-1$, let

$$\begin{aligned}
f_{k,n} &= |\{T\in\mathcal{T}_n \mid \varphi_T(1,2)=k\}|\\
&= |\{T\in\mathcal{T}_n \mid \varphi_T(i,j)=k\}| \text{ for every } 1\leqslant i<j\leqslant n\\
d_{k,n} &= |\{T\in\mathcal{T}_n \mid \delta_T(1)=k\}|\\
&= |\{T\in\mathcal{T}_n \mid \delta_T(i)=k\}| \text{ for every } 1\leqslant i\leqslant n
\end{aligned}$$

(where $|X|$ denotes the cardinal of the set X).

Lemma 8. *For every $n\geqslant 2$,*

$$E_U(\overline{\Phi}^{(2)}_n) = \frac{1}{(2n-3)!!}\Big(n\sum_{k=1}^{n-1}k^2\cdot d_{k,n}+\binom{n}{2}\sum_{k=1}^{n-2}k^2\cdot f_{k,n}\Big)$$

Proof. Under the uniform model,

$$E_U(\overline{\Phi}^{(2)}_n) = \frac{\sum_{T\in\mathcal{T}_n}\overline{\Phi}^{(2)}(T)}{(2n-3)!!},$$

where

$$
\begin{aligned}
\sum_{T\in\mathcal{T}_n} \overline{\Phi}^{(2)}(T) &= \sum_{T\in\mathcal{T}_n}\sum_{1\leqslant i\leqslant j\leqslant n} \varphi_T(i,j)^2 = \sum_{1\leqslant i\leqslant j\leqslant n}\sum_{T\in\mathcal{T}_n} \varphi_T(i,j)^2\\
&= \sum_{1\leqslant i\leqslant n}\sum_{T\in\mathcal{T}_n} \delta_T(i)^2 + \sum_{1\leqslant i<j\leqslant n}\sum_{T\in\mathcal{T}_n} \varphi_T(i,j)^2\\
&= \sum_{1\leqslant i\leqslant n}\sum_{k=1}^{n-1} k^2\cdot|\{T\in\mathcal{T}_n \mid \delta_T(i)=k\}|\\
&\qquad + \sum_{1\leqslant i<j\leqslant n}\sum_{k=1}^{n-2} k^2\cdot|\{T\in\mathcal{T}_n \mid \varphi_T(i,j)=k\}|\\
&= \sum_{1\leqslant i\leqslant n}\sum_{k=1}^{n-1} k^2\cdot d_{k,n} + \sum_{1\leqslant i<j\leqslant n}\sum_{k=1}^{n-2} k^2\cdot f_{k,n}\\
&= n\sum_{k=1}^{n-1} k^2\cdot d_{k,n} + \binom{n}{2}\sum_{k=1}^{n-2} k^2\cdot f_{k,n}.
\end{aligned}
$$

□

A formula for $d_{k,n}$ was obtained in the proof of [20, Lem. 21]:

$$d_{k,n} = \frac{(2n-k-3)!\cdot k}{(n-k-1)!2^{n-k-1}}. \tag{1}$$

As far as $f_{k,n}$ goes, we have the following result. In it, and henceforth, ${}_pF_q$ denotes the (*generalized*) *hypergeometric function* defined by

$${}_pF_q\left(\begin{matrix} a_1, & \dots, & a_p \\ b_1, & \dots, & b_q\end{matrix} ; z\right) = \sum_{k\geqslant 0} \frac{(a_1)_k\cdots(a_p)_k}{(b_1)_k\cdots(b_q)_k}\cdot\frac{z^k}{k!},$$

where $(a)_0 = 1$ and $(a)_k := a\cdot(a+1)\cdots(a+k-1)$ for $k\geqslant 1$.

Lemma 9. *For every $n\geqslant 2$, $f_{0,n} = (2n-4)!!$ and*

$$f_{k,n} = \frac{(2n-k-5)!k}{(2n-2k-4)!!}\cdot {}_3F_2\left(\begin{matrix} 1, & 2-n, & k+2-n \\ \frac{k+5}{2}-n, & \frac{k}{2}-n+3 & \end{matrix} ; 1\right)$$

for every $k = 1,\dots,n-2$.

Proof. Let us start by proving $f_{0,n} = (2n-4)!!$ by induction on n. It is clear that $f_{0,2} = 1 = (2\cdot 2-4)!!$. Assume now that $f_{0,n-1} = (2(n-1)-4)!!$. Every phylogenetic tree T with n leaves such that $\varphi_T(1,2) = 0$, that is, where $LCA_T(1,2)$ is the root, is obtained by taking a phylogenetic tree T' with $n-1$ leaves such that $\varphi_{T'}(1,2) = 0$ and adding a new pendant edge, ending in the leaf n, to any edge in T'. Then, since there are $f_{0,n-1} = (2n-6)!!$ trees $T'\in\mathcal{T}_{n-1}$ such that $\varphi_{T'}(1,2) = 0$, and each one of them has $2(n-1)-2$ edges where we can add the new edge, we obtain

$$f_{0,n} = (2n-4)(2n-6)!! = (2n-4)!!.$$

Now, to compute $f_{k,n}$ for $k \geqslant 1$, we shall study the structure of a tree $T \in \mathcal{T}_n$ such that $\varphi_T(1,2) = k$; to simplify the notations, let us denote by x the node $LCA_T(1,2)$, which has depth k, and by T_0 the subtree of T rooted at x.

Then, on the one hand, T_0 is a phylogenetic tree on a subset $S_0 \subseteq \{1, \dots, n\}$ containing $1, 2$, and since its root x is the LCA of 1 and 2 in T, we have that $\varphi_{T_0}(1,2) = 0$. On the other hand, there is a path $(r = v_1, v_2, v_3, \dots, v_{k+1} = x)$ in T from r to x. For every $j = 1, \dots, k$, let T_j be the subtree rooted at the child of v_j other than v_{j+1}; see Fig. 3.

So, the tree T is determined by:

- A number $0 \leqslant m \leqslant n - k - 2$, so that $m + 2$ will be the number of leaves of the phylogenetic tree T_0 rooted at $LCA_T(1,2)$
- A subset $\{i_1, \dots, i_m\}$ of $\{3, \dots, n\}$. There are $\binom{n-2}{m}$ such subsets.
- A phylogenetic tree T_0 on $\{1, 2, i_1, \dots, i_m\}$ such that $\varphi_{T_0}(1,2) = 0$. There are $f_{0,m+2} = (2m)!!$ such trees.
- An *ordered k-forest*, that is, an ordered sequence of phylogenetic trees $(T_1, T_1, \dots, T_k)$ such that $\bigcup_{i=1}^{k} L(T_i) = \{1, \dots, n\} - \{1, 2, i_1, \dots, i_m\}$. The number of such ordered k-forests is (see, for instance, [19, Lem. 1])

$$\frac{(2n - 2m - k - 5)!k}{(n - m - k - 2)!2^{n-m-k-2}}.$$

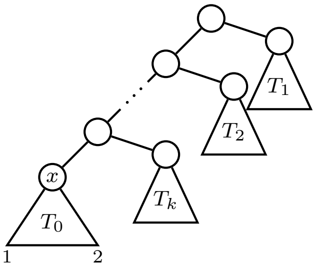

Figure 3: The structure of a tree T with $\varphi_T(1,2) = k$.

This shows that $f_{k,n}$ can be computed as

$$\begin{aligned}
f_{k,n} &= \sum_{m=0}^{n-k-2} (\text{number of ways of choosing } \{i_1,\ldots,i_m\}) \\
&\qquad \cdot(\text{number of trees in } \mathcal{T}_{m+2} \text{ with } \varphi_T(1,2)=0) \\
&\qquad \cdot(\text{number of ordered } k\text{-forests on } n-m-2 \text{ leaves}) \\
&= \sum_{m=0}^{n-k-2} \binom{n-2}{m} \cdot (2m)!! \cdot \frac{(2n-2m-k-5)!k}{(n-m-k-2)!2^{n-m-k-2}} \\
&= k \sum_{m=0}^{n-k-2} \frac{(n-2)!m!2^m(2n-2m-k-5)!}{m!(n-m-2)!(n-m-k-2)!2^{n-m-k-2}} \\
&= \frac{(n-2)!k}{2^{n-k-2}} \sum_{m=0}^{n-k-2} \frac{4^m(2n-2m-k-5)!}{(n-m-2)!(n-m-k-2)!}
\end{aligned}$$

Now, taking into account that

$$\begin{aligned}
&(1)_m = m! \\
&(2-n)_m = (-1)^m \frac{(n-2)!}{(n-m-2)!} \\
&(k+2-n)_m = (-1)^m \frac{(n-k-2)!}{(n-k-m-2)!} \\
&\left(\frac{k+5}{2}-n\right)_m = \frac{(-1)^m(2n-k-5)!!}{2^m(2n-k-2m-5)!!}, \\
&\left(\frac{k}{2}-n+3\right)_m = \frac{(-1)^m(2n-k-6)!!}{2^m(2n-k-2m-6)!!}
\end{aligned}$$

we have that

$$\begin{aligned}
&{}_3F_2\left(\begin{matrix} 1,\ 2-n,\ k+2-n \\ \frac{k+5}{2}-n,\ \frac{k}{2}-n+3 \end{matrix} ;1\right) = \sum_{m\geqslant 0} \frac{(1)_m \cdot (2-n)_m \cdot (k+2-n)_m}{(\frac{k+5}{2}-n)_m \cdot (\frac{k}{2}-n+3)_m} \cdot \frac{1}{m!} \\
&= \sum_{m\geqslant 0} \frac{m!(n-2)!(n-k-2)!2^m(2n-k-2m-5)!!2^m(2n-k-2m-6)!!}{(n-m-2)!(n-k-m-2)!(2n-k-5)!!(2n-k-6)!!m!} \\
&= \sum_{m=0}^{n-k-2} \frac{(n-2)!(n-k-2)!(2n-k-2m-5)!2^{2m}}{(n-m-2)!(n-k-m-2)!(2n-k-5)!} \\
&= \frac{(n-2)!(n-k-2)!}{(2n-k-5)!} \sum_{m=0}^{n-k-2} \frac{(2n-k-2m-5)!4^m}{(n-m-2)!(n-k-m-2)!}
\end{aligned}$$

from where we deduce that

$$\begin{aligned}
&\sum_{m=0}^{n-k-2} \frac{(2n-k-2m-5)!4^m}{(n-m-2)!(n-k-m-2)!} \\
&\qquad = \frac{(2n-k-5)!}{(n-2)!(n-k-2)!} {}_3F_2\left(\begin{matrix} 1,\ 2-n,\ k+2-n \\ \frac{k+5}{2}-n,\ \frac{k}{2}-n+3 \end{matrix} ;1\right)
\end{aligned}$$

and hence

$$\begin{aligned} f_{k,n} &= \frac{(n-2)!k}{2^{n-k-2}} \sum_{m=0}^{n-k-2} \frac{4^m (2n-2m-k-5)!}{(n-m-2)!(n-m-k-2)!} \\ &= \frac{(n-2)!k}{2^{n-k-2}} \cdot \frac{(2n-k-5)!}{(n-2)!(n-k-2)!} {}_3F_2\left(\begin{matrix} 1,\ 2-n,\ k+2-n \\ \frac{k+5}{2}-n,\ \frac{k}{2}-n+3 \end{matrix} ;1 \right) \\ &= \frac{(2n-k-5)!k}{(2n-2k-4)!!} \cdot {}_3F_2\left(\begin{matrix} 1,\ 2-n,\ k+2-n \\ \frac{k+5}{2}-n,\ \frac{k}{2}-n+3 \end{matrix} ;1 \right) \end{aligned}$$

as we claimed. □

We must compute now the sums

$$\sum_{k=1}^{n-1} k^2 \cdot d_{k,n}, \quad \sum_{k=1}^{n-2} k^2 \cdot f_{k,n}.$$

To do that, we shall use the following auxiliary lemma.

Lemma 10. *For every $n \geqslant 2$ and $m \geqslant 1$, let*

$$U_{n,m} = \sum_{k=0}^{n-2} \frac{k^m (n+k-2)!}{k! 2^k}.$$

Then,

$$\begin{aligned} U_{n,0} &= (2n-4)!! \\ U_{n,1} &= (n-1)(2n-4)!! - (2n-3)!! \\ U_{n,2} &= (n^2-1)(2n-4)!! - (2n-1)(2n-3)!! \\ U_{n,3} &= (n^3+3n^2-3n-1)(2n-4)!! - (3n^2+n-1)(2n-3)!! \end{aligned}$$

Proof. The proof of these identities is standard, using well known equalities for hypergeometric functions and the *lookup algorithm* given in [23, p. 36]. We shall prove in detail the identity for $m = 2$, and we leave the details of the rest to the reader.

Notice that

$$\begin{aligned} U_{n,2} = \sum_{k=0}^{n-2} \frac{k^2(n+k-2)!}{k!2^k} &= \sum_{k=1}^{n-2} \frac{k^2(n+k-2)!}{k!2^k} = \sum_{k=0}^{n-3} \frac{(k+1)^2(n+k-1)!}{(k+1)!2^{k+1}} \\ = \sum_{k=0}^{\infty} \frac{(k+1)^2(n+k-1)!}{(k+1)!2^{k+1}} &- \sum_{k=n-2}^{\infty} \frac{(k+1)^2(n+k-1)!}{(k+1)!2^{k+1}} \end{aligned}$$

Set

$$X_n = \sum_{k=0}^{\infty} \frac{(k+1)^2(n+k-1)!}{(k+1)!2^{k+1}}, \qquad Y_n = \sum_{k=n-2}^{\infty} \frac{(k+1)^2(n+k-1)!}{(k+1)!2^{k+1}}$$

We compute now these two summands.

As to X_n,

$$X_n = \frac{(n-1)!}{2} \sum_{k=0}^{\infty} \frac{(k+1)^2(n+k-1)!}{(n-1)!(k+1)!2^k}$$

If we set

$$t_k = \frac{(k+1)^2(n+k-1)!}{(n-1)!(k+1)!2^k},$$

we have that

$$\frac{t_{k+1}}{t_k} = \frac{(k+2)(k+n)}{(k+1)^2} \cdot \frac{1}{2}$$

and therefore, by the *lookup algorithm* [23, p. 36], we have that

$$\begin{aligned} X_n &= \frac{(n-1)!}{2} \cdot {}_2F_1\left(\begin{matrix} 2,\ n \\ 1 \end{matrix} ; \frac{1}{2} \right) \\ &= \frac{(n-1)!}{2} \cdot 2^n \cdot {}_2F_1\left(\begin{matrix} n,\ -1 \\ 1 \end{matrix} ; -1 \right) \quad \text{(using (15.3.4) in [1, p. 559])} \\ &= (n-1)!2^{n-1} \sum_{k \geqslant 0} \frac{(n)_k(-1)_k}{(1)_k} \cdot \frac{(-1)^k}{k!} \\ &= (n-1)!2^{n-1} \Big(\frac{(n)_0(-1)_0}{(1)_0} \cdot \frac{(-1)^0}{0!} + \frac{(n)_1(-1)_1}{(1)_1} \cdot \frac{(-1)^1}{1!} \Big) \\ &= (n-1)!2^{n-1}(n+1) \end{aligned}$$

As to Y_n,

$$\begin{aligned} Y_n &= \sum_{k=0}^{\infty} \frac{(k+n-1)^2(2n+k-3)!}{(k+n-1)!2^{k+n-1}} \\ &= \frac{(n-1)^2(2n-3)!}{(n-1)!2^{n-1}} \cdot \sum_{k=0}^{\infty} \frac{(k+n-1)^2(2n+k-3)!}{(k+n-1)!2^k \cdot \frac{(n-1)^2(2n-3)!}{(n-1)!}} \end{aligned}$$

If we take now

$$t_k = \frac{(k+n-1)^2(2n+k-3)!}{(k+n-1)!2^k \cdot \frac{(n-1)^2(2n-3)!}{(n-1)!}}$$

we have that

$$\frac{t_{k+1}}{t_k} = \frac{(n+k)(2n+k-2)}{(k+n-1)^2} \cdot \frac{1}{2}$$

and therefore, again by the *lookup algorithm* [23, p. 36], we have that

$$\begin{aligned} Y_n &= \frac{(n-1)^2(2n-3)!}{(n-1)!2^{n-1}} \cdot {}_3F_2\left(\begin{matrix} 1,\ n,\ 2n-2 \\ n-1,\ n-1 \end{matrix} ; \frac{1}{2} \right) \\ &= \frac{(n-1)^2(2n-3)!}{(n-1)!2^{n-1}} \Big[{}_2F_1\left(\begin{matrix} 2n-2,\ 1 \\ n-1 \end{matrix} ; \frac{1}{2} \right) + \frac{1}{n-1} \cdot {}_2F_1\left(\begin{matrix} 2n-1,\ 2 \\ n \end{matrix} ; \frac{1}{2} \right) \Big] \\ &\qquad \text{(using [13]).} \end{aligned}$$

Now

$$\begin{aligned}
{}_2F_1\left(\begin{matrix}2n-2,\ 1\\ n-1\end{matrix};\frac{1}{2}\right) &= 2\cdot {}_2F_1\left(\begin{matrix}1-n,\ 1\\ n-1\end{matrix};-1\right) \quad \text{(using (15.3.4) in [1, p. 559])}\\
&= 2\cdot\frac{2^{2(n-2)}\Gamma(n-1)}{\Gamma(2n-2)}\Big[\frac{\Gamma(n-1)}{\Gamma(0)}+\frac{\Gamma(n)}{\Gamma(1)}+\frac{2\Gamma(n-\frac{1}{2})}{\Gamma(\frac{1}{2})}\Big] \quad \text{(using [14])}\\
&= 2\cdot\frac{2^{2(n-2)}(n-2)!}{(2n-3)!}\Big[(n-1)!+2\cdot\frac{(2n-3)!!}{2^{n-1}}\Big]\\
&= \frac{2^{n-1}(n-1)!}{(2n-3)!!}+2\\
{}_2F_1\left(\begin{matrix}2n-1,\ 2\\ n\end{matrix};\frac{1}{2}\right) &= 2^2\cdot {}_2F_1\left(\begin{matrix}2,\ 1-n\\ n\end{matrix};-1\right) \quad \text{(using (15.3.4) in [1, p. 559])}\\
&= 4\cdot\frac{\Gamma(n)}{2^{2(2-n)}\Gamma(2n-1)}\Big(\frac{\Gamma(n-\frac{1}{2})}{\Gamma(\frac{1}{2})}+\frac{\Gamma(n+\frac{1}{2})}{\Gamma(\frac{3}{2})}+2\Gamma(n)\Big) \quad \text{(using [14])}\\
&= \frac{2^{2n-2}(n-1)!}{(2n-2)!}\Big(\frac{(2n-3)!!}{2^{n-1}}+\frac{(2n-1)!!}{2^{n-1}}+2\cdot(n-1)!\Big)\\
&= \frac{2^{n-1}(n-1)!}{(2n-2)!}((2n-3)!!+(2n-1)!!+2^n\cdot(n-1)!)
\end{aligned}$$

Therefore,

$$\begin{aligned}
Y_n = \frac{(n-1)^2(2n-3)!}{(n-1)!2^{n-1}}&\Big[\frac{2^{n-1}(n-1)!}{(2n-3)!!}+2\\
&+\frac{1}{n-1}\cdot\frac{2^{n-1}(n-1)!}{(2n-2)!}((2n-3)!!+(2n-1)!!+2^n\cdot(n-1)!)\Big]\\
= 2^{n-2}(n+1)(n-1)!&+(2n-1)!!
\end{aligned}$$

and finally

$$\begin{aligned}
U_{n,2} &= X_n - Y_n = 2^{n-2}(n+1)(n-1)! - (2n-1)!!\\
&= (n^2-1)(2n-4)!! - (2n-1)(2n-3)!!
\end{aligned}$$

as we claimed. □

Lemma 11. *For every* $n \geqslant 2$,

$$\sum_{k=1}^{n-1} k^2 d_{k,n} = (4n-1)(2n-3)!! - 3(2n-2)!!.$$

Proof. By equation (1),

$$\begin{aligned}
\sum_{k=1}^{n-1} k^2 d_{k,n} &= \sum_{k=1}^{n-1}\frac{k^3(2n-k-3)!}{(n-k-1)!2^{n-k-1}} = \sum_{k=0}^{n-2}\frac{(n-k-1)^3(n+k-2)!}{k!2^k}\\
&= (n-1)^3U_{n,0} - 3(n-1)^2U_{n,1} + 3(n-1)U_{n,2} - U_{n,3}\\
&= (n-1)^3(2n-4)!! - 3(n-1)^2\big((n-1)(2n-4)!! - (2n-3)!!\big)\\
&\quad +3(n-1)\big((n^2-1)(2n-4)!! - (2n-1)(2n-3)!!\big)\\
&\quad -\big((n^3+3n^2-3n-1)(2n-4)!! - (3n^2+n-1)(2n-3)!!\big)\\
&= (4n-1)(2n-3)!! - 3(2n-2)(2n-4)!!.
\end{aligned}$$

□

Lemma 12. *For every $n \geqslant 2$,*

$$\sum_{k=1}^{n-2} k^2 f_{k,n} = \frac{1}{3}(4n+1)(2n-3)!! - \frac{3}{2}(2n-2)!!.$$

Proof. To simplify the notations, set $S_n = \sum\limits_{k=1}^{n-2} k^2 f_{k,n}$. As we have seen in the proof of Lemma 9,

$$f_{k,n} = \frac{(n-2)!k}{2^{n-k-2}} \sum_{m=0}^{n-k-2} \frac{4^m(2n-2m-k-5)!}{(n-m-2)!(n-m-k-2)!}$$

and therefore

$$\begin{aligned}
S_n &= \frac{(n-2)!}{2^{n-2}} \sum_{k=1}^{n-2} 2^k k^3 \sum_{m=0}^{n-k-2} \frac{4^m(2n-k-2m-5)!}{(n-k-2)!(n-k-m-2)!} \\
&= \frac{(n-2)!}{2^{n-2}} \sum_{k=1}^{n-2} 2^k k^3 \sum_{m=0}^{n-k-2} \frac{4^{n-k-2-m}(k+2m-1)!}{(k+m)!m!} \\
&= (n-2)!2^{n-2} \sum_{k=1}^{n-2} \frac{k^3}{2^k} \left(\frac{1}{k} + \sum_{m=1}^{n-k-2} \frac{1}{4^m m} \binom{k+2m-1}{k+m} \right) \\
&= (n-2)!2^{n-2} \left(6 - \frac{n^2+2}{2^{n-2}} + \sum_{k=1}^{n-2} \frac{k^3}{2^k} \sum_{m=1}^{n-k-2} \frac{1}{4^m m} \binom{k+2m-1}{k+m} \right)
\end{aligned}$$

Set now

$$S'_n = \sum_{k=1}^{n-2} \frac{k^3}{2^k} \sum_{m=1}^{n-k-2} \frac{1}{4^m m} \binom{k+2m-1}{k+m} = \sum_{k=1}^{n-3} \frac{k^3}{2^k} \sum_{m=1}^{n-k-2} \frac{1}{4^m m} \binom{k+2m-1}{k+m}$$

Since $S'_3 = 0$, we have that

$$S'_n = \sum_{p=3}^{n-1} (S'_{p+1} - S'_p)$$

and

$$\begin{aligned}
S'_{p+1}-S'_p &= \frac{(p-2)^3}{2^p}+\sum_{k=1}^{p-3}\frac{k^3}{2^k(p-k-1)4^{p-k-1}}\binom{2p-k-3}{p-1}\\
&= \frac{(p-2)^3}{2^p}+\frac{1}{2^{2p-2}}\sum_{k=1}^{p-3}\frac{k^3(2p-k-3)!}{2^{-k}(p-k-1)(p-1)!(p-k-2)!}\\
&= \frac{(p-2)^3}{2^p}+\frac{1}{2^{2p-2}(p-1)!}\sum_{k=1}^{p-3}\frac{k^3(2p-k-3)!}{2^{-k}(p-k-1)!}\\
&= \frac{(p-2)^3}{2^p}+\frac{1}{2^{2p-2}(p-1)!}\sum_{k=1}^{p-3}\frac{(p-k-2)^3(p+k-1)!}{2^{k-p+2}(k+1)!}\\
&= \frac{(p-2)^3}{2^p}+\frac{1}{2^{p-1}(p-1)!}\sum_{k=2}^{p-2}\frac{(p-k-1)^3(p+k-2)!}{2^k k!}\\
&= \frac{(p-2)^3}{2^p}+\frac{1}{2^{p-1}(p-1)!}\Big[\sum_{k=0}^{p-2}\frac{(p-k-1)^3(p+k-2)!}{2^k k!}\\
&\qquad -(p-1)^3(p-2)!-\frac{1}{2}(p-2)^3(p-1)!\Big]\\
&= -\frac{(p-1)^2}{2^{p-1}}+\frac{1}{2^{p-1}(p-1)!}\sum_{k=0}^{p-2}\frac{(p-k-1)^3(p+k-2)!}{2^k k!}\\
&= -\frac{(p-1)^2}{2^{p-1}}+\frac{1}{(2p-2)!!}\big((4p-1)(2p-3)!!-3(2p-2)!!\big)\quad\text{(by Lemma 11)}\\
&= -\frac{(p-1)^2}{2^{p-1}}+(4p-1)\frac{(2p-3)!!}{(2p-2)!!}-3
\end{aligned}$$

Therefore

$$S'_n=\sum_{p=3}^{n-1}\Big((4p-1)\frac{(2p-3)!!}{(2p-2)!!}-\frac{(p-1)^2}{2^{p-1}}-3\Big)$$

Now, applying *Gosper's algorithm* [23, p. 77] we have that

$$\sum_{p=3}^{n-1}(4p-1)\frac{(2p-3)!!}{(2p-2)!!}=\frac{1}{3\cdot 2^{2n+1}}\Big(32(4n^2-3n-1)\binom{2n-3}{n-1}-39\cdot 2^{2n}\Big)$$

and then

$$\begin{aligned}
S'_n &= \frac{1}{3\cdot 2^{2n+1}}\Big(32(4n^2-3n-1)\binom{2n-3}{n-1}-39\cdot 2^{2n}\Big)\\
&\qquad -\frac{11\cdot 2^n-8(n^2+2)}{2^{n+1}}-3(n-3)\\
&= \frac{n^2+2}{2^{n-2}}-3(n+1)+\frac{(4n+1)(2n-3)!!}{3(2n-4)!!}.
\end{aligned}$$

Finally,

$$\begin{aligned} S_n &= (n-2)!2^{n-2}\left(6-\frac{n^2+2}{2^{n-2}}+S'_n\right)\\ &= -3(n-1)!2^{n-2}+\frac{(4n+1)(2n-3)!!}{3}\\ &= \frac{1}{3}(4n+1)(2n-3)!!-\frac{3}{2}(2n-2)!!. \end{aligned}$$

□

Finally, by Lemmas 8, 11, and 12, we have that

$$\begin{aligned} E_U(\overline{\Phi}_n^{(2)}) &= \frac{1}{(2n-3)!!}\Big(n\sum_{k=1}^{n-1}k^2\cdot d_{k,n}+\binom{n}{2}\sum_{k=1}^{n-2}k^2\cdot f_{k,n}\Big)\\ &= \frac{1}{(2n-3)!!}\Big[n((4n-1)(2n-3)!!-3(2n-2)!!)\\ &\qquad +\binom{n}{2}\Big(\frac{1}{3}(4n+1)(2n-3)!!-\frac{3}{2}(2n-2)!!\Big)\Big]\\ &= \frac{1}{6}n(4n^2+21n-7)-\frac{3n(n+3)}{4}\cdot\frac{(2n-2)!!}{(2n-3)!!} \end{aligned}$$

as we claimed.