



Universitat
de les Illes Balears

DOCTORAL THESIS
2019

**NEW BALANCE INDICES AND METRICS FOR
PHYLOGENETIC TREES**

Lucía Rotger García



Universitat
de les Illes Balears

DOCTORAL THESIS
2019

**Doctoral Programme of Information and
Communications Technology**

**NEW BALANCE INDICES AND METRICS FOR
PHYLOGENETIC TREES**

Lucía Rotger García

Thesis supervisor: Arnau Mir Torres

Thesis supervisor: Francesc Rosselló Llompart

Thesis tutor: Arnau Mir Torres

Doctor by the Universitat de les Illes Balears

Abstract

The belief that the shape of a phylogenetic tree reflects the properties of the evolutionary processes underlying it has motivated the study of indices quantifying the graph-theoretical properties of phylogenetic trees and of metrics allowing for their comparison. The main contribution of this PhD Thesis is then the addition to the set of available techniques for the analysis and comparison of phylogenetic trees of the *total cophenetic* balance index, the family of *Colless-like* balance indices, and the family of *cophenetic metrics*.

The total cophenetic index turns out to be a good alternative to other popular balance indices like Sackin's and Colless' indices. This index is defined for multifurcating trees and it achieves its maximum value exactly at the combs and its minimum value among the multifurcating trees exactly at the star trees and among the bifurcating trees at the maximally balanced trees, being the first balance index published in the literature satisfying this last property. We have computed closed formulas for its expected value under the Yule and the uniform models of bifurcating phylogenetic tree growth and a simple recurrence for its variance under the uniform model. As a by-product of this study, we have obtained a closed formula for the expected value of the Sackin index under the uniform model, a problem that remained open so far.

The *Colless-like indices* provide the first sound extension to multifurcating trees of the Colless index for bifurcating trees, in the sense that, when restricted to bifurcating trees, they give the classical Colless index up to a constant factor, and, for any given number of leaves, the only multifurcating trees that yield their minimum value are exactly the fully symmetric. These Colless-like indices depend on the choice of a dissimilarity function and of a *size* of rooted trees, and we show that this choice may affect how they measure the balance of a tree. In connection with these indices, we introduce in this Thesis our R package "CollessLike", available on the CRAN, that allows to perform goodness of fit tests of a phylogenetic tree with null model any α - γ -model.

Finally, we have defined the family of *cophenetic metrics* $d_{\varphi,p}$, with $p \in \{0\} \cup [1, \infty[$, for phylogenetic trees with possibly nested taxa and weights on the arcs. On different types of spaces of non-weighted trees, we have computed their least non-zero value, the order of their diameter, and the neighborhood of any given tree. Moreover, we have obtained closed formulas for the expected value under the Yule and the uniform models of the square of the metric $d_{\varphi,2}$.

Resumen

La creencia que la forma de un árbol filogenético es un reflejo de las propiedades de los procesos evolutivos subyacentes ha motivado el estudio de índices que cuantifiquen las propiedades gráficas de un árbol filogenético y de las métricas que permitan la comparación de árboles filogenéticos. La principal contribución de esta tesis doctoral es entonces la incorporación al conjunto de técnicas disponibles para el análisis y la comparación de árboles filogenéticos del *índice de balance cofenético total*, la familia de índices de balance *Colless-like* y la familia de *métricas cofenéticas*.

El índice cofenético total resulta ser una buena alternativa a otros índices populares de balance como los índices de Sackin y Colless. Este índice está definido para árboles no binarios, y alcanza su valor máximo exactamente en los árboles de tipo peine y su valor mínimo entre los árboles arbitrarios exactamente en los árboles estrella y entre los árboles binarios en los máximo balanceados, siendo el primer índice de balance publicado que satisface esta última propiedad. Hemos calculado fórmulas explícitas para su valor esperado bajo los modelos de Yule y uniforme de crecimiento de árboles filogenéticos binarios y una recurrencia simple para su varianza bajo el modelo uniforme. En el decurso de este estudio, hemos obtenido una fórmula explícita para el valor esperado del índice de Sackin bajo el model uniforme, un problema que aún permanecía abierto.

La familia de los índices *Colless-like* proporciona la primera extensión sólida a árboles filogenéticos arbitrarios del índice de Colless clásico para árboles binarios, en el sentido de que cuando se restringen a árboles binarios coinciden con el índice de Colless clásico salvo un factor constante y, para cualquier número de hojas, los únicos árboles que alcanzan su valor mínimo son exactamente los totalmente simétricos. Estos índices dependen de la elección de una función de disimilitud y de un *tamaño* de árboles, y mostramos que esta elección puede afectar la forma en que miden el balance del árbol. En relación con estos índices, presentamos en esta tesis nuestro paquete de R “CollessLike”, disponible en la CRAN, que permite realizar pruebas de bondad de ajuste de un árbol filogenético con cualquier modelo α - γ para árboles no binarios como modelo nulo.

Finalmente, hemos definido la familia de las *métricas cofenéticas* $d_{\varphi,p}$, con $p \in \{0\} \cup [1, \infty[$, para árboles filogenéticos con, posiblemente, nodos

interiores etiquetados y pesos en las aristas. Para diferentes tipos de espacios de árboles filogenéticos sin pesos en las aristas, hemos calculado el valor mínimo estrictamente positivo de estas métricas, el orden de magnitud de su diámetro y los entornos de los árboles. Además, hemos obtenido fórmulas explícitas para el valor esperado bajo los modelos de Yule y uniforme del cuadrado de la métrica $d_{\varphi,2}$.

Resum

La creença que la forma d'un arbre filogenètic és un reflex de les propietats dels processos evolutius que hi ha al darrere ha motivat l'estudi d'índexs que quantifiquin les propietats gràfiques dels arbres filogenètics i de mètriques que permetin la seva comparació. La contribució principal d'aquesta tesi doctoral és aleshores la incorporació al conjunt de tècniques disponibles per a l'anàlisi i la comparació d'arbres filogenètics de l'*índex de balanç cofenètic total*, la família d'índexs *Colless-like* i la família de *mètriques cofenètiques*.

L'índex cofenètic total és una bona alternativa a altres índexs de balanç populars com ara els de Sackin i de Colless. Aquest índex està definit per a arbres no binaris, i assoleix el seu valor màxim exactament als arbres de tipus pinta i el seu valor mínim entre els arbres arbitraris exactament als arbres estrella (no binaris) i entre els arbres binaris exactament als arbres màxim balancejats, sent el primer índex de balanç publicat que satisfà aquesta darrera propietat. Hem calculat fórmules explícites per al seu valor esperat sota els models de creixement d'arbres filogenètics binaris de Yule i uniforme i una recurrència simple per a la seva variància sota el model uniforme. Com a part d'aquest estudi, hem obtingut una fórmula explícita per al valor esperat de l'índex de Sackin sota el model uniforme, un problema que romanía obert.

Els índexs *Colless-like* són la primera extensió sòlida publicada per a arbres no binaris de l'índex de Colless, en el sentit que quan es restringeixen a arbres binaris coincideixen amb l'índex de Colless clàssic llevat d'un factor constant i, per a cada nombre de fulles, els arbres que assoleixen el seu valor mínim són exactament els totalment simètrics. Aquests índexs depenen de l'elecció d'una funció de dissimilitud i d'una *mida* d'arbres, i mostrem que aquesta tria pot afectar la forma com mesuren el balanç. En relació amb aquests índexs, presentem en aquesta tesi el nostre paquet de R "CollessLike", disponible a la CRAN, que permet realitzar proves de bondat d'ajust d'un arbre filogenètic amb qualsevol model α - γ per a arbres no binaris com a model nul.

Finalment, hem definit les *mètriques cofenètiques* $d_{\varphi,p}$, amb $p \in \{0\} \cup [1, \infty[$, per a arbres filogenètics amb, potser, nodes interiors etiquetats i pesos a les arestes. Per a alguns tipus d'espais d'arbres filogenètics sense pesos a les arestes, hem calculat el valor mínim no nul d'aquestes mètriques, l'ordre de magnitud del seu diàmetre i els entorns dels arbres. A més, donem fórmules explícites per a l'esperança sota els models de Yule i uniforme del quadrat de $d_{\varphi,2}$.

Agradecimientos

A mis directores de tesis, Cesc y Arnau, porque sin ellos esto no hubiera sido posible. Muchas gracias por todo, he aprendido muchas cosas durante estos años gracias a vosotros y espero poder seguir aprendiendo a vuestro lado.

A Mercè, por todo lo que se ha preocupado por mí, por cuidarme y darme buenos consejos. A Biel Cardona, por preocuparse, aconsejarme y guiarme aunque no fuera fácil. A Ricardo, por ser buen jefe y preocuparse por todos. A Adrià, por empezar la aventura juntos aunque nuestros caminos fueran divergentes. A Tomás, por ser un gran compañero filosófico de despacho. Al resto de integrantes del grupo BIOCOT: Jairo, Joan Carles, Pere, Irene, Biel R., Gabriel V., etc. por compartir el camino. Al Departamento de Ciencias Matemáticas e Informática y a la Universitat de les Illes Balears por acogerme durante todos estos años.

A Krzysztof Bartoszek, por su ayuda y consejo en la mejora de los test y pruebas estadísticas.

A Clara y Lorenzo, por adoptarnos en Logroño. A Pilar por preocuparse siempre de que estemos bien. A Miguel por amenizar las jornadas en el despacho. Al Departamento de Matemáticas y Computación por acogerme como una más. A la Universidad de La Rioja por darme la oportunidad de estar allí.

A Pedro, Henrik y Maribel, por ser más que amigos y compañeros de carrera, por compartir risas, momentos y frikadas durante tantos años y espero que muchos más. También a Marga, Cristina y demás compañeros.

A Eva, Ana y Erik por entender y compartir experiencias sobre el doctorado. Al resto del grupo valenciano que siempre están ahí.

A los amigos desperdigados por España, siempre es un placer volver a veros.

A Luisa, por su apoyo incondicional y sinceras palabras.

A mi familia, a Mari Sales que también ha entrado a formar parte de ella. A Rosalía y Manuel por ser como mis segundos padres. A mis tíos y tías. A Enrique, Noelia y demás primos. A mis abuelas Juana y Amalia, por valorar mi trabajo sin llegar a entenderlo, en recuerdo especial a mi abuela Amalia por no haber podido ver el final de este camino. A mis padres, por todo. Sin vosotros no hubiera llegado hasta aquí. Os quiero.

A Juanmi, por ayudarme, por preocuparse, por entenderme, por cuidarme, por todo. T'estime.

Esta investigación ha sido parcialmente financiada por el Ministerio de Ciencia, Innovación y Universidades y el Fondo Europeo de Desarrollo Regional (FEDER) a través de los proyectos *Grafos en biología computacional* (MTM2009-07165), *Aplicaciones bioinformáticas en filogenética, metagenómica, biología de sistemas y genómica del cáncer* (DPI2015-67082- P), *Desarrollo de estrategias -Ómicas para desvelar pangenomas, coevolución vírica y adaptación a los extremos de concentración salina- SP4 - MICROMATES* (PGC2018-096956-B-C43), *Creación de una red temática en computación biomolecular y celular* (TIN2008-04487-E/TIN) y *Renovación y nuevas actividades de la red temática en computación biomelecular y bioceular* (TIN2011-15874-E), también por la *Obra Social La Caixa* a través del “Programa Pont La Caixa per a grups de recerca de la UIB”.

Contents

Introduction	1
1 Preliminaries	9
1.1 Phylogenetic trees	9
1.2 Balance indices	16
1.3 Probabilistic models for phylogenetic trees	20
1.4 Hypergeometric series	26
2 The total cophenetic index	37
2.1 Main definitions	37
2.2 Trees with maximum and minimum Φ	41
2.3 Expected value of Φ under the Yule model	50
2.4 Expected value of Φ under the uniform model	54
2.5 On the variance of Φ under the uniform model	61
2.6 On the variance of S under the uniform model	74
2.7 On the covariance of Φ and S under the uniform model	82
2.8 Numerical experiments	93
2.8.1 The discriminative power of Φ	93
2.8.2 A test on TreeBASE	93
3 The cophenetic metrics	97
3.1 The cophenetic vectors	97
3.2 The definition of the cophenetic metrics	101
3.3 Minimum values	103
3.4 Diameters	123
3.5 Expected values in the Euclidean case	131
3.5.1 Expected value of D_n^2 under the Yule model	133
3.5.2 Expected value of D_n^2 under the uniform model	137
3.6 On the variance of D_n^2	146

4	The Colless-like indices	149
4.1	The Colless index	149
4.2	The Colless-like indices	151
4.3	Sound Colless-like indices	156
4.4	Maximally unbalanced trees	161
4.4.1	The case of $f(n) = \ln(n + e)$	161
4.4.2	The case of $f(n) = e^n$	182
4.5	The R package <i>CollessLike</i>	193
4.5.1	A real example	195
4.6	Experimental results on TreeBASE	197
4.6.1	Mean and variance as a function of the number of leaves of the trees	198
4.6.2	Numbers of ties	200
4.6.3	Spearman's rank correlation	201
4.6.4	Does TreeBASE fit the uniform model or the α - γ -model?	201
4.7	Tables of values for some examples of $\mathfrak{C}_{D,f}$	205
	Conclusions and future work	209
	Bibliography	213
	Appendix	223
A	Scripts	223
A.1	Packages required	223
A.2	List of all binary trees	223
A.3	General functions	224
A.4	Scripts from Chapter 2	225
A.4.1	Computation of $E_Y(\Phi_n)$	225
A.4.2	Computation of $E_U(\Phi_n)$	226
A.4.3	Computation of $\sigma_U^2(\Phi_n)$	227
A.4.4	Computation of $\sigma_U^2(S_n)$	230
A.4.5	Computation of $Cov_U(S_n, \Phi_n)$	233
A.4.6	Computation of the estimated probability of a tie	237
A.4.7	Testing Φ_n on TreeBASE	239
A.5	Scripts from Chapter 3	241
A.5.1	Computation of $E(D_n^2)$	241
A.5.2	Computation of $\sigma^2(D_n^2)$	243
A.6	Scripts from Chapter 4	247

A.6.1	The R package <i>CollessLike</i>	247
A.6.2	A real example	263
A.6.3	Computation of the mean and variance	265
A.6.4	Computation of the number of ties	274
A.6.5	Computation of Spearman's rank correlation	274
A.6.6	A test on the distribution of TreeBASE	277

Introduction

Almost two centuries ago, Charles Darwin observed common traits on different species. His studies led to his well-known book *On The Origin of Species* [35] and the idea of how evolution works by natural selection. The only figure in that book is the famous Fig. 1 below, an abstract representation of the descent through “modifications” (mutations) of a theoretical group of species from an initial common ancestor. This is considered to be the first published *phylogenetic tree*: a rooted tree representing the evolutionary history of a set of contemporary species —located at the leaves of the tree— from an unknown common ancestor —located at the root of the tree— through sets of mutations —represented by the arcs of the tree— involving intermediate ancestors that are also unknown —represented by the internal nodes of the tree.

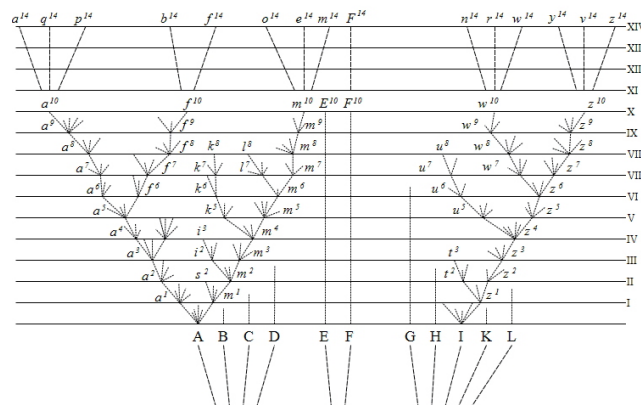


Figure 1: Darwin’s abstract phylogenetic tree as depicted in Chapter IV of *On The Origin of Species* (1859). Downloaded from http://commons.wikimedia.org/wiki/File:Origin_of_Species.svg.

Darwin’s choice of a line diagram as a metaphor of a evolutionary history was probably inspired by their previous use in genealogical trees and animal pedigrees [7]. So, to mention some examples from the scientific literature, Buffon [19] used a graph to represent the evolutionary origin of domesticated dog breeds in 1755 and Duchesne [39] represented in a similar way the evolutionary relationships between strawberry cultivars in 1766. These diagrams were not

trees, because they include hybridizations represented by nodes that are ends of pairs of branches coming from two different “parent” nodes. In a directed tree, however, every internal node is the end of a single arc, making it suitable to represent that a species is the direct descendant of a single species through a set of successful mutations. Also, Lamarck published 50 years before *On The Origin of Species* a tree representing a set of hypothetical evolutionary relationships among several major groups of animals [66]. It is not considered an example of phylogenetic tree with its current meaning of the term, because it illustrated the Lamarckian “evolution by adaptation” where a species can descend from another contemporary species through adaptation to new conditions. As a consequence, the extant groups of animals considered by Lamarck appeared both as leaves and as internal nodes in his tree, whereas, in a phylogenetic tree, extant species correspond only to leaves, and when a specific species is assigned to an internal node, it represents an extinct species that is known to be an ancestor of contemporary ones. For a more detailed account on the history of phylogenetic trees, see [85]; see also the “History” entries of *The genealogical world of phylogenetic networks* blog (<http://phylonetworks.blogspot.com>).

Most global biodiversity studies estimate that there are around 11 million different species of living organisms in the Earth [67]. The *Tree of Life*, the gigantic phylogenetic tree explaining the evolutionary history of all these current species, is considered one of the most important organizing principles in biology [38]. The goal of assembling such a global phylogenetic tree has been pushed by the new genomic sequencing techniques and the exponential increase of computers’ computational power that allows the alignment and comparison of large sets of genomic sequences [61]. Despite these advances, the Tree of Life is far from completed. In fact, there is an open and collaborative project to build an exhaustive and interactive version of it, called the *Open Tree of Life*. The first draft included more than 2 million species and it was released in 2015 [59]: the current version can be accessed at the url <https://tree.opentreeoflife.org>. Nevertheless, most phylogenetic studies focus on relatively small sets of species and therefore any technique dealing with phylogenetic trees must be able to handle any number of species, not only the whole set of species on Earth.

Since its first appearances, the use of phylogenetic trees as a representation diagram to describe the evolutionary relationships among species have become a fundamental tool in biology [10, 11]. Current phylogenetic trees usually include information about evolutive distances or a measure of the time elapsed between pairs of consecutive species by assigning lengths to their branches. Even so, it is a common belief that the *shape* of a phylogenetic tree, i.e., its raw branching structure, already reflects some of the properties of the evolutionary processes that have produced it, like for instance the differences in the rates of speciation or extinction among the different lineages gathered in the tree [106]. Moreover, different stochastic models of evolution give rise to different probability distributions of phylogenetic tree shapes, and this fact can be used to test these models using published phylogenetic trees [76] or to reveal

deviations in a tree from the speciation or extinction rates predicted by the model [54].

This motivates the interest in the study of *unweighted phylogenetic trees*, also called *cladograms*, that represent only the dependences in evolutionary histories, and they are the main topic of our Thesis. In other words, our phylogenetic trees represent a hierarchical classification of a set of species (or other *OTU*, from Operational Taxonomic Unit) into *clades*: subsets consisting of all descendant species of some internal node in the tree. In this way, the inclusion of one clade into another corresponds to the fact that the last common ancestor of the species in the former clade is a descendant of the last common ancestor of the latter. So, throughout most of this Thesis by a *phylogenetic tree* we simply mean a rooted directed tree with its leaves labeled with some set of labels in such a way that different leaves have different labels. Only in parts of Chapter 3 we shall allow also *nested taxa*, i.e., labeled internal nodes representing known extinct common ancestors of current species, and weights on the arcs, which can represent any measure of evolutive divergence between a species and a descendant.

The focus of this Thesis is the quantification of the *balance* of a phylogenetic tree, defined as the tendency of the direct descendants (the *children*) of any given node to have the same number of descendant leaves. This is one of the most studied properties of phylogenetic trees, because the *imbalance* of a phylogenetic tree reflects the propensity of evolutionary events to occur along specific lineages [78, 98], although sometimes it may also be due simply to a bias in the method or the data used to build it [99]. For instance, it has been observed that the phylogenetic trees arising in paleontology tend to be very unbalanced and this imbalance is usually a construct due to the incompleteness of data [54].

The balance of a phylogenetic tree is usually measured by means of *balance indices*, and many such indices have been defined so far in the phylogenetics literature. These balance indices are also used to quantify the informal notion of *symmetry* of a tree: a high degree of symmetry correlates positively with a high balance. Think for instance on the *fully symmetric* bifurcating trees on a number of leaves that is a power of 2, where the subtrees rooted at the children of each internal node are isomorphic. The two most popular and classic such indices are the *Colless index* [29] of a bifurcating tree, defined as the sum over all internal nodes of the absolute value of the difference between the number of descendant leaves of its pair of children; and the *Sackin index* [63, 98, 99] of a multifurcating tree, defined as the sum of the *depths* of its leaves (i.e., of their distances to the root). Other balance indices for bifurcating trees previous to our work include the variance of the depths of the leaves [63, 98], the sum of the reciprocals of the *heights* of the internal nodes (i.e., their maximum distances to descendant leaves) [99], and the number of *cherries* (pairs of leaves with a common parent) [70]. As for balance indices for multifurcating trees, one recent addition is the rooted quartets index [33]. More examples can be found

in [56, 58] and in the section “Measures of overall asymmetry” of Felsenstein’s book [43] (pp. 562–563). This abundance of balance indices is partly motivated by Shao and Sokal’s advice on using more than one such index to quantify tree balance, due to the fact that each index may measure different aspects of the balance of the tree [99].

Besides their natural application in the description of phylogenetic trees, pioneered more than 35 years ago by Sokal [101], the main application of balance indices has been as tools to test stochastic models of evolution [99, 76]. Other properties of the shapes of phylogenetic trees used in this connection include the distribution of clades’ sizes [118, 119] and the joint distribution of the numbers of rooted subtrees of different types [113]. The main idea in this connection is to obtain information of the probabilistic behaviour of a balance index under a given stochastic model of phylogenetic tree growth (for instance, its expected value and variance for any given number of leaves), be it by means of theorems or through Monte Carlo methods, and then to test whether the distribution of the balance indices of “real-life” phylogenetic trees gathered in phylogenetic databases is compatible with this information. Some useful databases with accessible phylogenetic data are TreeBASE [83, 109], PhyloFacts [2, 36, 65], TreeFam [68] and even the Open Tree of Life [59]. In this Thesis, we have used systematically the TreeBASE database in our experiments, because it is easy to consult it with R. Two of the most popular stochastic models of evolutionary tree growth are the Yule model [55, 116] and the uniform model [26, 70, 94] for bifurcating phylogenetic trees, and they are the most frequently used null models in this type of tests. A detailed description of them, as well as of two parametric generalizations, can be found in Section 1.3. Several properties of the distributions and moments of different balance indices have been established in the literature under these models [15, 14, 56, 63, 77, 90, 91, 92, 104], and this Thesis also contributes to this line of research.

In this PhD Thesis, we wanted to answer the necessity of having available balance indices satisfying several nice properties. One of the desirable properties of a balance index is that it must classify as most balanced and most unbalanced trees exactly those usually accepted as so in the phylogenetics literature, which would be a sign that this index measures balance in a proper way. Moreover, the index should have a low rate of ties among pairs of trees with the same number of leaves, in order to be useful to compare pairs of tree shapes. For instance, both the Sackin index and the Colless index classify as “maximally balanced” not only the so-called *maximally balanced trees*, but also other trees which are clearly less balanced than the maximally balanced ones. Another desirable property is that the index must be defined for arbitrary rooted trees. For instance, the Colless index can only be used on bifurcating trees and the attempts to generalize it to multifurcating trees previous to this Thesis have led to meaningless measures [99]. Also, the balance index must be easy to compute and to understand and with a low computational cost. Finally, it is convenient to be able to obtain information about its probability distribution

at least under the Yule and the uniform models (in the bifurcating case). For this purpose, in this Thesis we introduce two new balance indices: in Chapter 2, a new index that satisfies all the desirable properties listed above, and in Chapter 4, a meaningful generalization of the Colless index to multifurcating trees.

In addition to balance indices, another useful tool in the study of evolution through phylogenetic trees are the tree comparison metrics [105]. Given two phylogenetic trees, a similar value of a balance index could indicate that they have a similar structure, but quantifying “how similar they are” provides more relevant information. Phylogenetic tree metrics are applied, for instance, to compare different phylogenetic trees obtained from the same data by means of different algorithms [102, 93] or to compare phylogenetic trees obtained through numerical algorithms with other types of hierarchical classifications [102]. They are also used to assess the stability of reconstruction methods [112], in the comparative analysis of hierarchical cluster structures [53, 87], and in the construction of consensus supertrees [8]. Many phylogenetic tree comparison metrics have been proposed so far [43, Chapter 30]. Some of them are edit distances that count how many transformations of a given type are necessary to transform one tree into the other. For example, the nearest-neighbor interchange metric [110] and the subtree prune-and-regrafting distance [5] are metrics of this kind. Other metrics compare a pair of phylogenetic trees through some consensus subtree. This is the case for instance of the MAST distances defined in [44, 48, 117]. Finally, other metrics are based on the comparison of suitable encodings of the phylogenetic trees, like the Robinson-Foulds metric [88, 89], the triples metric [34], the nodal metrics for bifurcating trees [112, 105, 40, 41, 82], and the splitted nodal metric for multifurcating phylogenetic trees [21]. This last kind of metrics have the advantage that, unlike the edit and the consensus distances, they are usually computed in low polynomial time. In Chapter 3 we define a new metric for phylogenetic trees of the third aforementioned type that allows nested taxa and weighted arcs in the phylogenetic trees and that is inspired by the balance index introduced in the previous chapter.

Before detailing the contents of this Thesis, we want to emphasize that, as we have commented in the first paragraphs of this Introduction, phylogenetic trees only represent evolutionary descentance through speciation by mutation. This is a quite restrictive view of evolution, because it does not take into account other evolutive processes like genetic recombinations, which are a common mechanism in sexually reproducing species, hybridizations, a very common speciation mechanism in plants, and lateral gene transfers, a very common mechanism for the exchange of genetic material among bacteria [37]. All these processes cannot be suitably represented within the branching pattern of a tree, and require graphical representations where a node may be the end of more than one arc. These more general representations of evolutive histories are generically called *phylogenetic networks* [62]. For instance, the evolutionary diagrams by Buffon and Duchesne mentioned above are phylogenetic networks, and the Tree

of Life is actually not a tree but a *Network of Life* [80, 86]. But, as many other mathematical models, although phylogenetic trees do not model evolution perfectly, they are useful enough to be still the most common paradigm to describe evolution [71]. This Thesis deals only with phylogenetic trees, although it would be interesting the generalization of the concepts considered herein to the more general framework of networks.

Organization of this Thesis

This Introduction is followed by a chapter where we gather a series of definitions and basic results needed to develop the remaining chapters. Then, in Chapter 2 we define a new balance index for phylogenetic trees, the *total cophenetic index*, inspired by the philosophy of the classical Sackin index but, instead of adding up the leaves's depths, we add up the *cophenetic values* of all pairs of different leaves, i.e., the depths of their last common ancestors. This cophenetic index actually measures the *imbalance* of the tree, in the sense that it tends to grow with the imbalance. We characterize the phylogenetic trees with minimum and maximum cophenetic indices for every number of leaves, showing that these extremal values are achieved exactly at the trees commonly considered to be the most balanced and the most unbalanced trees, respectively. Moreover, we study the expected value and the variance of this index under the Yule and the uniform models: in some cases, we provide exact formulas for these moments and in the remaining cases we obtain recurrences. As a by-product, we obtain a closed formula for the expected value of the Sackin index under the uniform model, for which only an asymptotic formula was known previous to our work [14]. Finally, we perform some numerical experiments to test the power of the index, including some experiments involving phylogenetics trees from the TreeBASE.

In Chapter 3 we introduce and study the *cophenetic metrics*, defined as the L^p norm of the difference of the vectors of cophenetic values of the trees. These metrics can be meaningfully used on phylogenetic trees with nested taxa and weighted arcs. We characterize the pairs of trees at minimum cophenetic distance on several spaces of phylogenetic trees, and we establish the order of their diameter. Moreover, in the Euclidean case, we provide closed formulas for the expected value under the Yule and uniform models of the square of the corresponding cophenetic distance and we estimate the order for its variance.

Finally, in Chapter 4 we generalize the classic Colless index to multifurcating trees. Recall that the original index is defined as the sum over all internal nodes of a tree of their *balance values*: the absolute value of the difference between the number of descendant leaves of its pair of children. Consequently, its greatest strength resides in the fact that it measures directly the notion of balance of a tree. However, its generalization to multifurcating trees is not trivial, because the natural generalizations, defining the balance value of a node as some dissimilarity applied to the number of descendant leaves of its children,

yields 0 on multifurcating trees that are not fully symmetric. In this chapter, we show how to correctly generalize the Colless index to multifurcating trees, we propose six particular such generalizations, and we characterize the phylogenetic trees with minimum and maximum values for them. To perform the numeric experiments of this chapter we have created the R package “*CollessLike*”, which is available on the CRAN [74] and the latest version on GitHub [96].

The Thesis ends with a short Conclusions chapter and an Appendix where, for reproducibility, we provide the R and Python scripts used in all computations. These scripts are also available on the GitHub repository associated to this PhD Thesis [97].

Publications

The results reported in this Thesis have been published in:

- (1) Arnau Mir, Francesc Rosselló and Lucía Rotger. A new balance index for phylogenetic trees. *Mathematical Biosciences* 241 (2013), pp. 125–136.
- (2) Gabriel Cardona, Arnau Mir, Francesc Rosselló, Lucía Rotger and David Sánchez. Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinformatics* 14 (2013), article number: 3.
- (3) Gabriel Cardona, Arnau Mir, Francesc Rosselló and Lucía Rotger. The expected value of the squared cophenetic metric under the Yule and the uniform models. *Mathematical Biosciences* 295 (2018), pp. 73–85.
- (4) Arnau Mir, Lucía Rotger and Francesc Rosselló. Sound Colless-like balance indices for multifurcating trees. *PloS ONE* 13 (2018), e0203401.
- (5) Tomás M. Coronado, Arnau Mir, Francesc Rosselló and Lucía Rotger. On Sackin’s original proposal: The variance of the leaves’ depths as a phylogenetic balance index. Submitted for publication (2019).

More specifically, most of the results included in Chapter 2 are published in (1), except Section 2.6, which is contained in (5), and parts of Sections 2.5 and 2.7 that remain unpublished. All results and experiments in Chapter 3 appear in (2) or (3). Finally, the results in Chapter 4 are included in (4).

Chapter 1

Preliminaries

In this chapter, we gather several definitions and results that are used in the upcoming chapters. In the first section we give some notations and definitions related to phylogenetic trees and in the second section we describe two classical balance indices, Sackin's and Colless' indices, and we recall some of their properties. Then, in the third section we review several probabilistic models for phylogenetic trees. Finally, we devote the last section to the *lookup algorithm*, a method for the summation of series that is used in many computations in later chapters.

1.1 Phylogenetic trees

A *rooted tree* is a directed finite graph (without self-loops or repeated arcs) that contains a distinguished node, called the *root*, from which every node can be reached through exactly one directed path. Given a rooted tree T , we shall denote its sets of nodes and arcs by $V(T)$ and $E(T)$, respectively, or simply by V and E if the tree T is clear from the context.

Let T be a rooted tree. Whenever $(u, v) \in E(T)$, we say that v is a *child* of u and that u is the *parent* of v . Two nodes with the same parent are called *siblings*. This genealogical metaphor is extended to other levels of relationship in the natural way: *grandparents*, *cousins*, etc. The *out-degree* of a node $u \in V$ is the number of children of u , and we denote it by $\deg_T(u)$ or simply by $\deg(u)$. The nodes without children are the *leaves* of the tree, and the other nodes are called *internal*. An arc is *pendant* when it ends in a leaf, and *internal* when it ends in an internal node. In a rooted tree consisting of a single node and no arc, its only node is simultaneously the root and a leaf, and hence this tree has no internal node. We shall denote by $L(T)$ the set of leaves of T and by $V_{int}(T)$ its set of internal nodes. The nodes with exactly one child are called *elementary*. We assume henceforth that, unless otherwise stated, our rooted trees do not contain elementary nodes.

A *phylogenetic tree* is a representation of the shared evolutionary history

of a set of species. From the mathematical point of view, it is a leaf-labeled rooted tree, with its leaves representing the species under study, its internal nodes representing common ancestors of different groups of them, the root representing the most recent common ancestor of all of them, and the arcs representing direct descendance through mutations.

So, formally, given a non-empty set S , a *phylogenetic tree* on S is a rooted tree without elementary nodes endowed with a bijective labeling function from $L(T)$ to S ; in the context of phylogenetic trees, the elements of S are called *taxa*. To simplify the language, we shall always identify a leaf of a phylogenetic tree with its label. Although in practice S may be any set of taxa, to fix ideas we shall usually take $S = [n] :=$ with n the number of leaves or, more in general, in the phylogenetic trees with nested taxa (see later in this section), the number of labeled nodes. We shall use the term *phylogenetic tree with n leaves* to refer to a phylogenetic tree on $[n]$.

Two phylogenetic trees T_1, T_2 on the same set of taxa S are *isomorphic* when there exists an isomorphism of directed graphs between them that preserves the labelling of the leaves. We shall always make the abuse of language of saying that two isomorphic (phylogenetic or unlabeled) trees are *equal*, and hence we shall always identify any tree with its isomorphism class. To simplify the language, we shall use the term *space* to mean a set of isomorphism classes of trees of some type.

Given a set of labels S , we shall denote by \mathcal{T}_S the space (i.e. the set of isomorphism classes) of phylogenetic trees on S , and we shall denote by \mathcal{T}_n , for every $n \geq 1$, the space $\mathcal{T}_{[n]}$, i.e., the space of phylogenetic trees with n leaves. Notice that if $|S| = n$, then any bijection $S \leftrightarrow [n]$ induces a bijection $\mathcal{T}_S \leftrightarrow \mathcal{T}_n$. Moreover, we shall denote by \mathcal{T}_n^* the space of rooted trees with n leaves, and by \mathcal{T}^* the union $\bigcup_{n \geq 1} \mathcal{T}_n^*$. If $|S| = n$, there is a forgetful mapping $\pi : \mathcal{T}_S \rightarrow \mathcal{T}_n^*$ that sends every phylogenetic tree to the corresponding unlabeled tree, which we shall call its *shape*.

In the rest of this section, we shall introduce some concepts and notations on rooted trees: we understand, usually without any further notice, that they extend to phylogenetic trees through their shape.

An internal node v of a rooted tree T is *bifurcating* when it has exactly two children, i.e. when $\deg(v) = 2$. A rooted tree is *bifurcating*, or *binary*, when all its internal nodes are bifurcating. Notice that since the tree consisting of only one node does not have internal nodes, it is bifurcating. A phylogenetic tree is *bifurcating*, or *fully resolved*, when its shape is bifurcating. Whenever we want to emphasize the fact that a tree needs not be bifurcating, we shall call it *multifurcating*.

We shall denote by \mathcal{BT}_S , \mathcal{BT}_n , and \mathcal{BT}_n^* the spaces of bifurcating phylogenetic trees on a set S of taxa, of bifurcating trees with n leaves —i.e., on the set $[n]$ — and of bifurcating (unlabeled) trees with n leaves, respectively. In later chapters we shall need to know the cardinality of \mathcal{BT}_n , which is well known:

$|\mathcal{BT}_1| = 1$ and, for every $n \geq 2$,

$$|\mathcal{BT}_n| = (2n - 3)!! = (2n - 3)(2n - 5) \cdots 3 \cdot 1.$$

This formula also works in the case $n = 1$, because by definition $(-1)!! = 1$.

If there exists a path from u to v in a rooted tree T , we shall say that v is a *descendant* of u and also that u is an *ancestor* of v , and we shall denote it by $v \preceq u$. If, moreover, $u \neq v$, we shall write $v \prec u$ and we shall say that v is a *proper descendant* of u and that u is a *proper ancestor* of v . The *cluster* of a node v in T is the set $C_T(v)$ of its descendant leaves, and we shall denote by $\kappa_T(v)$ the cardinality $|C_T(v)|$ of $C_T(v)$, that is, the number of descendant leaves of v . When T is a phylogenetic tree, we identify $C_T(v)$ with the set of labels of the descendant leaves of v .

Given a node v of a rooted tree T , the *subtree of T rooted at v* is the directed subgraph of T induced on the set of descendants of v . If T is phylogenetic, then T_v is a phylogenetic tree on $C_T(v)$ with root this node v . A rooted subtree is a *cherry* when it has exactly 2 leaves, a *triplet*, when it has 3 leaves, and a *quartet*, when it has 4 leaves.

The *lowest common ancestor (LCA)* of a pair of nodes u, v in a rooted tree T , in symbols $[u, v]_T$, is the unique common ancestor of them that is a descendant of every other common ancestor of them.

The *distance* $d(u, v)$ from a node u to a descendant v of it in a rooted tree T is the number of arcs in the unique path from u to v . The *depth* $\delta_T(v)$ of a node v is the distance from the root r to the node v . The *depth* $\delta(T)$ (or simply δ when T is clear from the context) of a tree T is the largest depth of any leaf in it.

Let T_1, \dots, T_k be phylogenetic trees on pairwise disjoint sets of labels S_1, \dots, S_k , respectively. Their *root join* is the phylogenetic tree $T_1 \star \cdots \star T_k$ on $S_1 \cup \cdots \cup S_k$ obtained by adding to the disjoint union of T_1, \dots, T_k a new node r and new arcs from r to the root of each T_i . In this way, the new node r is the root of $T_1 \star \cdots \star T_k$ and the trees T_1, \dots, T_k become the subtrees of $T_1 \star \cdots \star T_k$ rooted at the children of r ; cf. Fig. 1.1. A similar construction produces a rooted tree $T_1 \star \cdots \star T_k$ from a set of (unlabeled) rooted trees T_1, \dots, T_k , in such a way that the forgetful mapping that sends every phylogenetic tree to its shape preserves this operation.

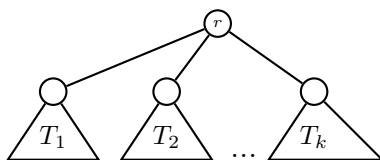


Figure 1.1: The root join $T_1 \star \cdots \star T_k$.

We need now to introduce several specific types of rooted trees. We begin

with the *comb* with n leaves, which is the unique bifurcating rooted tree with n leaves all whose internal nodes have a leaf child: see Fig. 1.2.(a). We also call a comb any phylogenetic tree whose shape is a comb in the previous sense. We shall generically denote every comb in \mathcal{T}_n , as well as their shape in \mathcal{T}_n^* , by K_n .

The *rooted star*, or simply *star*, with n leaves is the unique rooted tree with n leaves where all of them have depth 1: see Fig. 1.2.(b). Again, we also call a star any phylogenetic tree whose shape is a star in the previous sense. For consistency with later notations, we shall denote the star in \mathcal{T}_n , and its shape in \mathcal{T}_n^* , by FS_n .

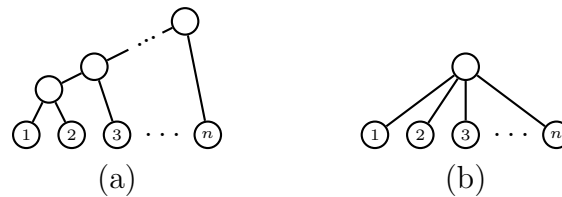


Figure 1.2: (a) A comb with n leaves, K_n . (b) The rooted star with n leaves, FS_n .

Let T be a bifurcating rooted tree. For every $v \in V_{int}(T)$, say with children v_1, v_2 , the *balance value* of v is $bal_T(v) = |\kappa_T(v_1) - \kappa_T(v_2)|$. An internal node v of T is *balanced* when $bal_T(v) \leq 1$. So, a node v with children v_1 and v_2 is balanced if, and only if, $\{\kappa_T(v_1), \kappa_T(v_2)\} = \{\lfloor \kappa_T(v)/2 \rfloor, \lceil \kappa_T(v)/2 \rceil\}$. We shall say that a bifurcating rooted tree T is *maximally balanced* when all its internal nodes are balanced. Since the only tree in \mathcal{T}_1 , consisting of a single node, does not have internal nodes, it is maximally balanced. The maximally balanced trees were named by Shao and Sokal [99] the “most balanced” bifurcating trees for any given number of leaves.

For the sake of completeness, we establish in the next lemma several easy properties on maximally balanced trees that were stated without proof in [73].

Lemma 1.1. *Let T be a bifurcating rooted tree.*

- (a) *If T has more than one leaf, then T is maximally balanced if, and only if, its root is balanced and both subtrees rooted at the children of the root are maximally balanced.*
- (b) *If T is maximally balanced, then any rooted subtree of it is maximally balanced.*
- (c) *For every number n of leaves, there exists one, and only one, maximally balanced rooted tree with n leaves up to isomorphism.*

Proof. As far as (a) goes, let T_1 and T_2 be the subtrees rooted at the children of the root r of T , so that $T = T_1 \star T_2$. Then, since $V_{int}(T) = \{r\} \cup V_{int}(T_1) \cup V_{int}(T_2)$, we have that T is maximally balanced if, and only if, r is balanced

and all internal nodes of T_1 and T_2 are balanced in T . Since the balance of a node of a rooted subtree is the same as in the larger tree, this is equivalent to r being balanced and T_1 and T_2 being maximally balanced.

As to (b), we prove it by induction on the depth of T . If $\delta(T) = 0$, then T is the only tree of depth 0: a single node, which is maximally balanced and it has no rooted subtree other than itself. Assume now that the assertion is true for trees of depth smaller than $\delta \geq 1$ and let T be a maximally balanced tree of depth $\delta(T) = \delta$. Let again T_1 and T_2 be the subtrees rooted at the children of the root of T . Notice that the depths of T_1 and T_2 are at most $\delta - 1$ and, by (a), they are both maximally balanced. Therefore, by the induction hypothesis, all their rooted subtrees are maximally balanced. Now, if T_v is a rooted subtree different from T , then its root v belongs to $V(T_1)$ or to $V(T_2)$, and if $v \in V(T_i)$, then, as we have just seen, $T_v = (T_i)_v$ is maximally balanced. Since the subtree of T rooted at the root, which is the only internal node not covered by this discussion, is T itself and hence also maximally balanced, we conclude that all rooted subtrees of T are maximally balanced.

Finally, we prove (c) by induction on n . If $n = 1$, the assertion is true because there is only one tree in \mathcal{T}_1 and it is maximally balanced. Assume now that the assertion is true for every number of leaves n' smaller than $n \geq 2$, and let T, T' two maximally balanced rooted trees with n leaves. Let T_1, T_2 be the subtrees of T rooted at the children of the root, let n_1 and n_2 be their respective numbers of leaves, and assume without any loss of generality that $n_1 \geq n_2$. In a similar way, let T'_1, T'_2 be the subtrees of T' rooted at the children of the root, n'_1 and n'_2 , their respective numbers of leaves, and assume that $n'_1 \geq n'_2$. By (a), the roots of T and T' are balanced and the trees T_1, T_2, T'_1, T'_2 are maximally balanced, and their numbers of leaves are smaller than n .

Now, since the roots of T and T' are balanced, it must happen that $n_1 = n'_1 = \lceil n/2 \rceil$ and $n_2 = n'_2 = \lfloor n/2 \rfloor$, and then, for every $i = 1, 2$, since T_i and T'_i are maximally balanced with the same number of leaves and this number of leaves is smaller than n , the induction hypothesis implies that they are equal (i.e., isomorphic). So, $T_1 = T'_1$ and $T_2 = T'_2$ and hence $T = T_1 \star T_2 = T'_1 \star T'_2 = T'$. This completes the proof of the inductive step. \square

Notice that, by (c) in the last lemma, given any set S of taxa, the shape of a maximally balanced phylogenetic tree on a set S of leaves is fixed and it only depends on the cardinality of S . Therefore, two maximally balanced phylogenetic trees with the same number of leaves differ only in their labeling. Fig. 1.3 depicts the maximally balanced trees with $n = 2, \dots, 6$ leaves.

An internal node v of a multifurcating rooted tree T is *symmetric* when, if v_1, \dots, v_k are its children, the trees T_{v_1}, \dots, T_{v_k} are isomorphic. A tree T is *fully symmetric* when all its internal nodes are symmetric, and a phylogenetic tree is *fully symmetric* when its shape is so.

Given a number n of leaves, there may exist several fully symmetric trees with n leaves. For instance, there are three fully symmetric trees with 6

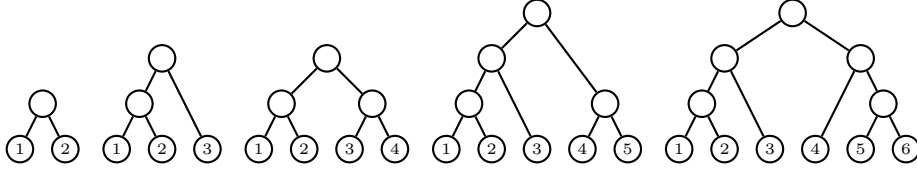


Figure 1.3: Maximally balanced trees.

leaves, depicted in Fig. 1.4. In fact, every fully symmetric tree with n leaves is characterized by an ordered factorization $n_1 \cdots n_k$ of n , with $n_1, \dots, n_k \geq 2$. More specifically, for every $k \geq 1$ and $(n_1, \dots, n_k) \in \mathbb{N}^k$ with $n_1, \dots, n_k \geq 2$, let FS_{n_1, \dots, n_k} be the tree defined, up to isomorphism, recursively as follows:

- FS_{n_1} is the rooted star with n_1 leaves.
- If $k \geq 2$, FS_{n_1, \dots, n_k} is a tree whose root has n_1 children, and the subtrees rooted at each one of these children are (isomorphic to) FS_{n_2, \dots, n_k} .

We shall also say that a tree is *of the form* FS_{n_1, \dots, n_k} when it is isomorphic to a tree FS_{n_1, \dots, n_k} obtained by means of the previous procedure. Since the tree consisting of only one node does not have any internal node, it is fully symmetric, and in this context we shall denote it, when needed, by FS_- .

The following result was stated without proof in [75].

Lemma 1.2. *Every FS_{n_1, \dots, n_k} is fully symmetric, and every fully symmetric tree is isomorphic to some FS_{n_1, \dots, n_k} .*

Proof. We prove both assertions by induction on the depth δ of the trees (which is equal to k on trees of the form FS_{n_1, \dots, n_k}). The case $\delta = 0$ is clear, because the only tree of depth 0, the single node, is fully symmetric. The case $\delta = 1$ is also clear, because every tree of depth 1 is a rooted star, which is fully symmetric of the form FS_n with n its number of leaves.

Assume now that both assertions are true for trees of depth smaller than $\delta \geq 2$. Consider a tree $FS_{n_1, \dots, n_\delta}$ of depth δ . By definition, it has the form

$$FS_{n_2, \dots, n_\delta}^{(1)} \star \cdots \star FS_{n_2, \dots, n_\delta}^{(n_1)}$$

with all $FS_{n_2, \dots, n_\delta}^{(i)}$ pairwise disjoint trees of the form $FS_{n_2, \dots, n_\delta}$. By the induction hypothesis, $FS_{n_2, \dots, n_\delta}$ is fully symmetric. Then, since the subtrees rooted at the children of the root of $FS_{n_1, \dots, n_\delta}$ are isomorphic, the root is symmetric, and since every internal node of $FS_{n_1, \dots, n_\delta}$ other than the root is an internal node of some $FS_{n_2, \dots, n_\delta}^{(i)}$ and these trees are fully symmetric, we conclude that the internal nodes different from the root are also symmetric. This proves that $FS_{n_1, \dots, n_\delta}$ is fully symmetric.

Let now T be a fully symmetric of depth δ , let r be its root and let $\deg(r) = n_1$. Since T is fully symmetric, the n_1 subtrees T_1, \dots, T_{n_1} rooted at

the children of r are isomorphic, and since their internal nodes are internal nodes of T and hence symmetric, the trees T_1, \dots, T_{n_1} are fully symmetric of depth smaller than δ . Being isomorphic to each other, the induction hypothesis implies that there exists a tree FS_{n_2, \dots, n_k} such that T_1, \dots, T_{n_1} are all of them isomorphic to it. By definition, this says that T is of type $FS_{n_1, n_2, \dots, n_k}$. This concludes the proof of the inductive step. \square

Therefore, for every n , the number of fully symmetric trees with n leaves is equal to the number $H(n)$ of ordered factorizations of n (sequence A074206 in Sloane's *On-Line Encyclopedia of Integer Sequences (OEIS)* [100]).

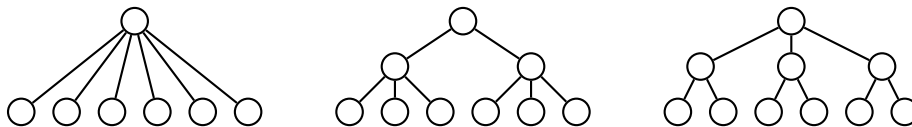


Figure 1.4: Three fully symmetric trees with 6 leaves: from left to right, FS_6 , $FS_{2,3}$ and $FS_{3,2}$.

An *ordered m -forest* on a set S is an ordered sequence of m phylogenetic trees (T_1, T_2, \dots, T_m) , each T_i on a set S_i of taxa, such that these sets S_i are pairwise disjoint and their union is S . An ordered forest is *bifurcating* when it consists of bifurcating trees. Let $\mathcal{BF}_{m,n}$ be the space of bifurcating ordered m -forests on the set $[n]$. It is known (see, for instance, [72, Lem. 1]) that for every $n \geq m \geq 1$,

$$|\mathcal{BF}_{m,n}| = \frac{(2m - m - 1)! m}{(n - m)! 2^{n-m}}. \quad (1.1)$$

In Chapter 3 we shall need to generalize the concepts of rooted trees and phylogenetic trees considered so far in two directions: on the one hand, by adding weights to the arcs, and on the other hand, by adding labeled internal nodes and, for technical reasons that are discussed therein, allowing the existence, with some restrictions, of elementary nodes. Let us consider first weighted trees. A *weighted tree* is a pair (T, ω) consisting of a rooted tree T and a *weight function* $\omega : E(T) \rightarrow \mathbb{R}_{>0}$ that associates to every arc $e \in E(T)$ a non-negative real number $\omega(e) > 0$. When working with weighted trees, we identify every *unweighted* (that is, where no weight function has been explicitly defined) tree T with the weighted tree (T, ω) with ω the weight 1 constant function. A *weighted phylogenetic tree* is simply a phylogenetic tree whose shape is a weighted tree. The isomorphisms of weighted (unlabeled or phylogenetic) trees are required to preserve the weights. We shall denote by \mathcal{WT}_n^* the space of weighted rooted trees with n leaves.

The only notions that are modified in weighted rooted trees are those related to distances. In a weighted rooted tree, the *distance* $d(u, v)$ from a node u to a descendant v is the sum of the weights of the arcs in the unique path from u to v , and the *depth* of a node is then the distance from the root to it in this sense.

As we have mentioned, in Chapter 3 we shall also allow the phylogenetic trees to have labeled internal nodes. A *phylogenetic tree with nested taxa* on S is a rooted tree T with possibly elementary nodes endowed with a partial labeling mapping from $V(T)$ to S satisfying the following two properties:

- The domain of the labeling mapping contains all leaves and all elementary nodes of T other than the root (which may be labeled or not, but it needs not be labeled even if it happens to be elementary).
- The labeling mapping is bijective from its domain to S .

In this context, the labeled internal nodes are called the *nested taxa* of the tree. Notice in particular that if a phylogenetic tree with nested taxa does not have nested taxa after all, then it cannot contain elementary nodes other than the root, and therefore it is a phylogenetic tree with, possibly, its root elementary. The isomorphisms of phylogenetic trees are required to preserve and reflect the labeling mappings, in the sense that a node is labeled if, and only if, its image is labeled and both labels are the same. We shall also consider *weighted phylogenetic trees with nested taxa*, that is, phylogenetic trees with nested taxa whose underlying shape is a weighted rooted tree.

So, given a set S of taxa, besides the spaces \mathcal{T}_S of all (unweighted) phylogenetic trees on S and $\mathcal{BT}_S \subseteq \mathcal{T}_S$ of all bifurcating phylogenetic trees on S , in Chapter 3 we shall also deal with the spaces \mathcal{WT}_S of all weighted phylogenetic trees with nested taxa on S and \mathcal{UT}_S of all unweighted phylogenetic trees with nested taxa on S . When $S = [n]$, we shall simply write \mathcal{WT}_n and \mathcal{UT}_n , respectively.

1.2 Balance indices

One of the most thoroughly studied properties of the shapes of phylogenetic trees is their *balance*, that is, the degree to which the children of internal nodes tend to have the same number of descendant leafs. The balance of a tree is usually quantified by means of a single number generically called a *balance index*. The two most popular balance indices are Sackin's [98, 99] and Colless' [29] indices, but, as we mentioned in the Introduction, there are many more such indices (cf. [43, Chap. 33]) and Shao and Sokal [99, p. 1990] explicitly advised to use more than one such index to quantify tree balance. In particular, in later chapters of this memory we shall introduce new balance indices. In this section we review the basic properties the Sackin and the Colless indices. The properties of the Sackin index will be used in Chapters 2 and 3, while in Chapter 4 we shall present a generalization of the Colless index to multifurcating trees.

A *shape index* for phylogenetic trees is a mapping $I : \bigcup_{n \geq 1} \mathcal{T}_n \rightarrow \mathbb{R}$ such that, for every $n \geq 1$ and for every $T, T' \in \mathcal{T}_n$, if $\pi(T) = \pi(T')$, then $I(T) = I(T')$.

Such a shape index is extended to phylogenetic trees on any set of taxa S by taking any bijection $\phi : S \rightarrow [n]$, where $n = |S|$, and then defining, for every $T \in \mathcal{T}_S$, $I(T) = I(\phi(T))$. So, in particular, the value of a shape index on a phylogenetic tree only depends on its shape, and not on its isomorphism class or the specific labeling of its leaves. A *shape index for bifurcating phylogenetic trees* is defined in a similar way, but restricting the domain of I to $\bigcup_{n \geq 1} \mathcal{BT}_n$.

A shape index I is *recursive* [69] when there exists a mapping

$$f_I : \bigcup_{k \geq 1} \mathbb{N}^k \rightarrow \mathbb{R}$$

such that:

- i) f_I is *symmetric*: for every $k \geq 2$, for every permutation σ of $\{1, \dots, k\}$ and for every $(n_1, \dots, n_k) \in \mathbb{N}^k$, $f_I(n_1, \dots, n_k) = f_I(n_{\sigma(1)}, \dots, n_{\sigma(k)})$.
- ii) For every phylogenetic trees T_1, \dots, T_k on disjoint sets of taxa,

$$I(T_1 \star \dots \star T_k) = \sum_{i=1}^k I(T_i) + f_I(|L(T_1)|, \dots, |L(T_k)|).$$

When I is shape index for bifurcating phylogenetic trees, these two conditions amount to impose that $f_I(n_1, n_2) = f_I(n_2, n_1)$ for every $n_1, n_2 \in \mathbb{N}$, and that if $T_1 \in \mathcal{BT}_{n_1}$ and $T_2 \in \mathcal{BT}_{n_2}$, then

$$I(T_1 \star T_2) = I(T_1) + I(T_2) + f_I(n_1, n_2).$$

Now, the *Sackin index* of a phylogenetic tree $T \in \mathcal{T}_n$ is defined as the sum of the depths of its leaves:

$$S(T) = \sum_{i=1}^n \delta_T(i).$$

Equivalently [15], it is equal to the sum of the numbers of descendant leaves of the internal nodes of the tree:

$$S(T) = \sum_{v \in V_{int}(T)} \kappa_T(v).$$

On the other hand, the *Colles index* of a bifurcating phylogenetic tree T is defined as the sum of the balance values of its internal nodes:

$$C(T) = \sum_{v \in V_{int}(T)} bal_T(v) = \sum_{v \in V_{int}(T)} |\kappa_T(v_1) - \kappa_T(v_2)|$$

where v_1 and v_2 denote the children of each $v \in V_{int}(T)$.

Both the Sackin and the Colless indices measure actually the *imbalance* of the tree, in the sense that a larger value of these indices usually corresponds to a smaller degree of balance. This can be deduced for the Colless index from its very definition as the sum of the balance values of the internal nodes. As to the Sackin index, this fact is not immediately obvious and we have to take our hat off to Sackin's intuition that made him realize that more unbalanced trees tended to have larger total sums of depths.

The Sackin index can be computed on any multifurcating tree, while the Colles index only makes sense, as it stands, for bifurcating trees. Shao and Sokal proposed in [99] to extend it to multifurcating trees by defining the balance of a multifurcating node to be 0, but this solution is clearly unsatisfactory. As we mentioned above, in Chapter 4 we shall introduce a new family of Colless-like balance indices that generalize in a specific sense the Colless index to multifurcating phylogenetic trees.

It is straightforward to notice that these two indices are shape indices in the sense defined above: they depend only on the shape of the tree, being invariant under isomorphisms and relabelings of leaves. This is desirable, because the balance of a phylogenetic tree depends only on its shape. So, the Sackin and the Colless indices of an unlabeled rooted tree are simply those of any phylogenetic tree having this rooted tree as its shape. As a matter of fact, these indices could have been introduced the other way round: first defining them for unlabeled trees and next extending them to phylogenetic trees as the corresponding indices of their shape.

Moreover, they are recursive:

- For every $T_1 \in \mathcal{T}_{n_1}, \dots, T_k \in \mathcal{T}_{n_k}$,

$$S(T_1 \star \dots \star T_k) = \sum_{i=1}^k S(T_i) + \sum_{i=1}^k n_i \quad [92].$$

Notice moreover that $\sum_{i=1}^k n_i$ is the total number of leaves of $T_1 \star \dots \star T_k$.

- For every $T_1 \in \mathcal{T}_{n_1}$ and $T_2 \in \mathcal{T}_{n_2}$,

$$C(T_1 \star T_2) = C(T_1) + C(T_2) + |n_1 - n_2| \quad [91].$$

It may happen that two phylogenetic trees with the same number of leaves but different shapes have the same value of a given balance index. We call this phenomenon a *tie*. Fig. 1.5 shows an example of a simultaneous tie for the Sackin and the Colless indices.

Although some results on the maximum and minimum values of the Sackin and the Colless indices has been common knowledge almost since their introduction, it has not been until very recently that their maxima and minima have been fully characterized.

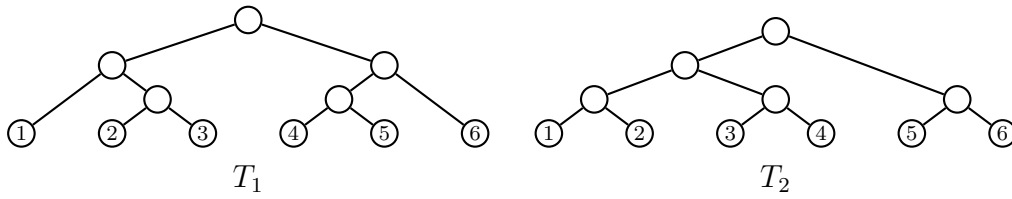


Figure 1.5: Two phylogenetic trees with 6 leaves and different shapes that have the same values of both the Colless and the Sackin indices: namely, $C(T_1) = C(T_2) = 2$ and $S(T_1) = S(T_2) = 16$.

As far as the Sackin index goes, its maximum value both in \mathcal{T}_n and in \mathcal{BT}_n , for any given number of leaves n , is achieved exactly at the combs K_n [45, Thm. 2], and its value is

$$S(K_n) = \binom{n+1}{2} - 1. \quad (1.2)$$

As to its minimum value in \mathcal{T}_n , it is clearly reached at the star FS_n , with $S(FS_1) = 0$ and, for every $n \geq 2$,

$$S(FS_n) = n,$$

because the sum of $n \geq 2$ depths is always at least n . The minimum value of the Sackin index in \mathcal{BT}_n is reached, among other trees, at the maximally balanced trees with n leaves, as it can be deduced from Thm. 5 in [45], and its value is [45, Thm. 4]

$$n(\lceil \log_2(n) \rceil + 1) - 2^{\lceil \log_2(n) \rceil}. \quad (1.3)$$

As far as the Colless index goes, it was already hinted at by Colless [29] that its maximum for any given number of leaves is achieved at the combs and since then this fact has been taken as well-known (see, for instance, [90]), but to our knowledge no proof of this property had been published before Lemma 1 in [75] (see Lemma 4.1 in this memory). This maximum value is

$$C(K_n) = \binom{n-1}{2}.$$

The minimum value of the Colless index in \mathcal{BT}_n is reached again, among other trees, at the maximally balanced trees with n leaves (see [31, Thm. 1]) and its

value can be obtained in the following way (see [31, Thm. 2]). If $n = \sum_{j=1}^{\ell} 2^{m_j}$, with $m_1 > \dots > m_{\ell}$, is the binary expansion of n , then the minimum value of the Colless index in \mathcal{BT}_n is

$$c_n = \sum_{j=2}^{\ell} 2^{m_j} (m_1 - m_j - 2(j-2)).$$

So, both the Sackin and the Colless indices take their maximum values exactly at the combs, which are the trees usually considered the most unbalanced ones [98, 99], and they both achieve their minimum value (for bifurcating trees, in the case of the Sackin index) at the bifurcating trees considered by Shao and Sokal the most balanced ones. But both Sackin's and Colless' indices achieve their minimum values in \mathcal{BT}_n also in trees that are not maximally balanced. For instance, both trees in Fig. 1.5 have the minimum values of the Sackin and the Colless indices on \mathcal{BT}_6 , and the left hand side tree is maximally balanced, but the right hand side not. The complete characterizations of the trees achieving the minimum values of these indices are given in [45], for the Sackin index, and in [31], for the Colless index. They are not needed here, so we omit them. Moreover, it is proved in [31, Prop. 8] that every tree in \mathcal{BT}_n that has minimum Colless index, then it also has minimum Sackin index, although the converse implication is not true. In Chapter 2 we shall introduce a new balance index that achieves its minimum value in \mathcal{BT}_n exactly at the maximally balanced trees.

1.3 Probabilistic models for phylogenetic trees

The balance indices, like Sackin's and Colless', only depend on the shape of the trees. Since it is believed that the shape of a phylogenetic tree reflects, at least to some extent, the evolutionary processes that have produced it [43, Chap. 33], these indices have been widely used as tools to test stochastic models of evolution: see, for instance, [13, 76, 99].

Two of the most popular stochastic models of evolutionary tree growth are the Yule and the uniform models. The *Yule*, or *Equal-Rate Markov, model* [55, 116], starts with a single node and, at every step, a leaf is chosen randomly and uniformly, and it is replaced by a cherry. Equivalently, a pendant arc is chosen randomly and uniformly and a new leaf is *added* to this arc, by which we mean that if the chosen arc is (u, x) , then it is replaced by two arcs (u, v) and (v, x) , with v a new node, and then a new leaf y is added together with a new arc (v, y) . Finally, once the desired number of leaves is reached, the taxa are assigned randomly and uniformly to the leaves. So, the Yule model corresponds to an evolutionary process where, at each step, each currently extant species can give rise with the same probability to two new species.

This process only produces bifurcating trees, and the probability of a bifurcating phylogenetic tree under this model is then defined as the probability of obtaining it through this procedure. Under this model, different trees with the same number of leaves may have different probabilities. Specifically, a tree $T \in \mathcal{BT}_n$ turns out to have probability [17, 99]

$$P_Y(T) = \frac{2^{n-1}}{n!} \prod_{v \in V_{int}(T)} \frac{1}{\kappa_T(v) - 1}. \quad (1.4)$$

In contrast, the main feature of the *uniform*, or *Proportional to Distinguishable Arrangements, model* [94] is that all phylogenetic trees with the same number of leaves have the same probability. From the point of view of tree growth [26, 70], this corresponds to a process where, starting with a node labeled 1, at the k -th step a new pendant arc, ending in the leaf labeled $k + 1$, is added either to a new root or to some edge (being all possible locations of this new pendant arc equiprobable). Notice that this is not an explicit model of evolution, only of bifurcating tree growth. Since under this uniform model all trees $T \in \mathcal{BT}_n$ are equiprobable, they all have probability

$$P_U(T) = \frac{1}{|\mathcal{BT}_n|} = \frac{1}{(2n - 3)!!}.$$

So, both the Yule and the uniform models define probability distributions on every \mathcal{BT}_n , with $n \geq 1$: we say then that they are *probabilistic models for bifurcating phylogenetic trees*. From the equations for P_Y and P_U given above it is clear that both models are *invariant under relabelings*, or *shape invariant*, in the following sense: if $T, T' \in \mathcal{BT}_n$ have the same shape, then they have the same probability.

Several properties of the distributions of Sackin's and Colless' indices under this pair of models have been studied in the literature [15, 14, 56, 63, 77, 90, 91, 92, 104]. Given a number n of leaves, let C_n and S_n be the random variables defined by choosing a tree $T \in \mathcal{BT}_n$ and computing $C(T)$ or $S(T)$, respectively. The following facts are known about the expected values of these random variables:

(1.3) Under the Yule model, their expected values are

- $E_Y(C_n) = n(H_{\lfloor n/2 \rfloor} - 1) + \lceil n/2 \rceil - \lfloor n/2 \rfloor$ [56]
- $E_Y(S_n) = 2n(H_n - 1)$ [63]

where H_n denotes the n -th *harmonic number*, $H_n = \sum_{i=1}^n \frac{1}{i}$.

(1.4) Under the uniform model, previous to our work it was only known about their expected values that

$$E_U(C_n), E_U(S_n) \sim \sqrt{\pi} n^{3/2} \quad [14].$$

In Theorem 2.27 in Chapter 2 we shall prove a closed formula for $E_U(S_n)$:

$$E_U(S_n) = n \left(\frac{(2n - 2)!!}{(2n - 3)!!} - 1 \right).$$

As far as their variance goes:

(1.5) Under the Yule model, their variances were computed in [23]:

- $\sigma_Y^2(C_n) = (5n^2 + 7n)/2 + (6n + 1)\lfloor n/2 \rfloor - 4\lfloor n/2 \rfloor^2 + 8\lfloor (n + 2)/4 \rfloor^2 + (2\lfloor n/2 \rfloor - n(n - 3))H_{\lfloor n/2 \rfloor} - 8(n + 1)\lfloor (n + 2)/4 \rfloor - 6nH_n + (n^2 + 3n - 2\lfloor n/2 \rfloor)H_{\lfloor (n+2)/4 \rfloor} - n^2H_{\lfloor n/2 \rfloor}^{(2)} - 2nH_{\lfloor n/4 \rfloor}$
- $\sigma_Y^2(S_n) = 7n^2 - 4n^2H_n^{(2)} - 2nH_n - n$

where $H_n^{(2)} = \sum_{i=1}^n \frac{1}{i^2}$.

(1.6) Under the uniform model, it is only known their limit behaviour [14]:

$$\sigma_U^2(S_n), \sigma_U^2(C_n) \sim \left(\frac{10 - 3\pi}{3}\right)n^3.$$

The Yule and the uniform model are particular cases of more general probabilistic models for bifurcating phylogenetic trees. One of these more general models that appears in later chapters of this memory is the so-called α -model, introduced by D. Ford in his PhD Thesis [46]. This is a parametric model, depending on a parameter $\alpha \in [0, 1]$ that gives it its name. This model, as the Yule model or, under its second description through a tree growth process, the uniform model, recursively builds bifurcating trees and assigns to each such tree the probability of producing it. More specifically, let us fix a desired number $n \geq 1$ of leaves. Then:

- (1) Start with the tree $T_1 \in \mathcal{BT}_1$ consisting of a single node labeled 1. Set $P'_{\alpha,1}(T_1) = 1$.
- (2) In the first step, the only tree T_2 in \mathcal{BT}_2 , the cherry, is obtained by adding a new root and a new leaf labeled 2 and arcs from the new root to the old node and the new leaf. Since $\mathcal{BT}_2 = \{T_2\}$, we set $P'_{\alpha,2}(T_2) = 1$, and this will be consistent with the fact that there is only one way to obtain this tree from T_1 .
- (3) Now, for every $m = 2, \dots, n - 1$, let $T_{m+1} \in \mathcal{BT}_{m+1}$ be obtained by adding a new leaf labeled $m + 1$ to T_m in a place chosen according to the following probability distribution:

- The probability of choosing any pendant arc is $\frac{1-\alpha}{m-\alpha}$, and in this case, then, the resulting tree T_{m+1} has probability

$$P'_{\alpha,m+1}(T_{m+1}) = \frac{1-\alpha}{m-\alpha} \cdot P'_{\alpha,m}(T_m).$$

- The probability of choosing any internal arc is $\frac{\alpha}{m-\alpha}$ and then

$$P'_{\alpha,m+1}(T_{m+1}) = \frac{\alpha}{m-\alpha} \cdot P'_{\alpha,m}(T_m).$$

- The probability of adding a new root to the tree and then pending the new leaf from this root is again $\frac{\alpha}{m-\alpha}$, and in this case we have again that

$$P'_{\alpha,m+1}(T_{m+1}) = \frac{\alpha}{m-\alpha} \cdot P'_{\alpha,m}(T_m).$$

These probabilities can be understood as assigning a weight α to all “internal places” (i.e., internal arcs or the new root) where the new leaf can be added, and a weight $1 - \alpha$ to all pendant arcs, and forcing the probability of choosing one such place to be proportional to its weight. For every $m \geq 2$, after $m - 1$ steps of the algorithm the resulting bifurcating tree T_m contains m leaves, and hence m pendant arcs, and $m - 1$ internal nodes and hence $m - 1$ internal places where the leaf can be added ($m - 2$ internal arcs and the new root). So, the total sum of weights is

$$m(1 - \alpha) + (m - 1)\alpha = m - \alpha$$

and hence to transform weights into probabilities we must divide by this value, resulting that the probability of an internal place is $\frac{\alpha}{m-\alpha}$ and that of a pendant arc is $\frac{1-\alpha}{m-\alpha}$.

- (4) When the desired number n of leaves is reached, the probability of a given tree is defined as the sum of the probabilities $P'_{\alpha,n}$ of all phylogenetic trees with that shape; that is, for every $T_n^* \in \mathcal{BT}_n^*$, its probability under the α -model is

$$P_{\alpha,n}^*(T_n^*) = \sum_{\pi(T'_n)=T_n^*} P'_{\alpha,n}(T'_n).$$

- (5) Finally, the probability $P_{\alpha,n}(T)$ of any phylogenetic tree $T \in \mathcal{BT}_n$ is obtained from the probability under $P_{\alpha,n}^*$ of its shape by splitting it equally among all phylogenetic trees in \mathcal{BT}_n with its shape:

$$P_{\alpha,n}(T) = \frac{P_{\alpha,n}^*(\pi(T))}{|\{T' \in \mathcal{BT}_n \mid \pi(T') = \pi(T)\}|}.$$

These two last steps are equivalent to applying an equiprobably chosen permutation to the leaves $\{1, \dots, n\}$ and defining the probability $P_{\alpha,n}(T)$ of a phylogenetic tree $T \in \mathcal{BT}_n$ as the probability of obtaining it through this procedure, and the probability $P_{\alpha,n}^*(T^*)$ of an unlabeled tree $T^* \in \mathcal{BT}_n^*$ as the sum of the probabilities of all phylogenetic trees with this shape. Defined in this way, the α -model is invariant under relabelings.

Notice that:

- If $\alpha = 1$, all new pendant leaves are added to internal arcs. This process generates only combs.

- If $\alpha = 1/2$, at each step all places where to add the new leaf have the same probability to be chosen. This process gives rise then to the uniform model. I.e., $P_{1/2,n} = P_U$, the probability under the uniform model.
- If $\alpha = 0$, all new leaves are added to pendant arcs. This process is equivalent to the Yule model, i.e., $P_{0,n} = P_Y$, the probability under the Yule model.

Ford provides in [46] explicit formulas for the probabilities of unlabeled trees and phylogenetic trees, but they fail in some cases and they have been amended recently in [32]. In particular, if, for every $T \in \mathcal{BT}_n$ and for every $v \in V_{int}(T)$, with children v_1, v_2 such that $\kappa_T(v_1) \leq \kappa_T(v_2)$, we let $NS_T(v) = (\kappa_T(v_1), \kappa_T(v_2))$ and we denote by $NS(T)$ the multiset

$$NS(T) = \{NS_T(v) \mid v \in V_{int}(T)\},$$

then the probability of T under the α -model is

$$P_{\alpha,n}(T) = \frac{2^{n-1}}{n! \cdot \Gamma_{\alpha}(n)} \prod_{(a,b) \in NS(T)} \left(\frac{\alpha}{2} \binom{a+b}{a} + (1-2\alpha) \binom{a+b-2}{a-1} \right),$$

where $\Gamma_{\alpha} : \mathbb{Z}^+ \rightarrow \mathbb{R}$ is the mapping defined by $\Gamma_{\alpha}(1) = 1$ and, for every $n \geq 2$,

$$\Gamma_{\alpha}(n) = \prod_{j=1}^{n-1} (j - \alpha).$$

Ford's α -model for bifurcating trees was later extended to multifurcating trees by B. Chen, D. Ford, and M. Winkel in the so-called α - γ -model [27]. Its definition is similar to the α -model, but now the probabilities of the places where the new leaves can be added depend on two parameters, α and γ with $0 \leq \gamma \leq \alpha \leq 1$. More specifically, let us fix a desired number $n \geq 1$ of leaves. Then:

- (1) Start with the tree $T_1 \in \mathcal{T}_1$ consisting of a single node labeled 1. Set $P_{\alpha,\gamma,1}(T_1) = 1$.
- (2) As in the α -model, in the first step, the only tree T_2 in \mathcal{T}_2 , the cherry, is obtained by adding a new root and a new leaf labeled 2 and arcs from the new root to both leaves. Set $P_{\alpha,\gamma,2}(T_2) = 1$.
- (3) For every $m = 1, \dots, n-1$, let $T_{m+1} \in \mathcal{T}_{m+1}$ be obtained by adding a new leaf labeled $m+1$ to T_m . Then:
 - If the new leaf is added to a pending arc,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{1-\alpha}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

- If the new leaf is added to an internal arc,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{\gamma}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

- If the new leaf is added as a child of a new root,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{\gamma}{m-\alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

- If the new leaf is added as a child of an internal node $u \in V_{int}(T_m)$,

$$P_{\alpha,\gamma,m+1}(T_{m+1}) = \frac{(\deg(u) - 1)\alpha - \gamma}{m - \alpha} \cdot P_{\alpha,\gamma,m}(T_m).$$

These probabilities can be understood as assigning a weight $1 - \alpha$ to each pendant arc, a weight γ to each internal arc and to the new root, and a weight $(k - 1)\alpha - \gamma$ to each node of out-degree $k \geq 2$ and forcing the probability of choosing one such place to be proportional to this weight. For every $m \geq 2$, after $m - 1$ steps of the algorithm the resulting tree T_m contains m leaves, and hence m pendant arcs. For every $k \geq 2$, let p_k be the number of internal nodes in T_m of out-degree k , and let $p = \sum_{k \geq 2} p_k$ be the total number of internal nodes in T_m , so that the number of places of weight γ (the internal arcs plus 1 corresponding to the new root) is precisely p . Notice moreover that

$$\sum_{k \geq 2} p_k \cdot k = |E(T_m)| = |V(T_m)| - 1 = p + m - 1.$$

So, the total sum of weights in T_m is

$$\begin{aligned} & m(1 - \alpha) + p\gamma + \sum_{k \geq 2} ((k - 1)\alpha - \gamma)p_k \\ &= m + \alpha \left(\sum_{k \geq 2} p_k \cdot k - \sum_{k \geq 2} p_k - m \right) + \gamma \left(p - \sum_{k \geq 2} p_k \right) \\ &= m + \alpha \left(\sum_{k \geq 2} p_k \cdot k - p - m \right) + \gamma(p - p) = m - \alpha. \end{aligned}$$

So, the weights have to be divided by $m - \alpha$ to obtain a proper probability distribution.

- (4) When the desired number n of leaves is reached, the probability $P_{\alpha,\gamma,n}(T_n)$ of the resulting tree T_n is the one obtained in this way. Then, the probability $P_{\alpha,\gamma,n}^*(T^*)$ of a given tree $T^* \in \mathcal{T}_n^*$ is defined as the sum of the probabilities of all phylogenetic trees with that shape:

$$P_{\alpha,\gamma,n}^*(T^*) = \sum_{\pi(T_n)=T^*} P_{\alpha,\gamma,n}(T_n).$$

Notice that if $\alpha = \gamma$, the probability of choosing an internal node as the place where to add the new pendant arc is $(k - 2)\alpha$. It can be proved then by induction on m that all internal nodes in T_m are bifurcating. Indeed, T_2 clearly satisfies this property and if all internal nodes in T_m have out-degree $k = 2$, then the probability of choosing any of them to make it multifurcating in T_{m+1} is 0, and the internal node that is generated by adding the new leaf to an arc or to a new root is bifurcating, which implies that all internal nodes in T_{m+1} are again bifurcating. And if all nodes are bifurcating and $\alpha = \gamma$, the probability of choosing any pendant arc becomes $\frac{1-\alpha}{m-\alpha}$ and the probability of choosing any internal arc or a new root is $\frac{\alpha}{m-\alpha}$, which are the probabilities used in the α -model. This entails that, for every $T_n \in \mathcal{BT}_n$, $P_{\alpha,\alpha,n}(T_n) = P'_{\alpha,n}(T_n)$ —the provisional probability of T_n defined by the recursive application of step (3) in the definition of the α -model—and, for every $T_n^* \in \mathcal{BT}_n^*$, $P_{\alpha,\alpha,n}^*(T_n^*) = P_{\alpha,n}^*(T_n^*)$. It is in this sense that we say that the α - γ -model generalizes the α -model to multifurcating trees. Notice also that the α - γ model defined in this way is not invariant under relabelings.

In Section 4.6 we shall also consider the *uniform model* on \mathcal{T}_n^* . As its name hints, this probabilistic model assigns to every tree T_n^* in \mathcal{T}_n^* the same probability,

$$P_U(T_n^*) = \frac{1}{|\mathcal{T}_n^*|}.$$

Since no closed formula for the cardinality $|\mathcal{T}_n^*|$ is known, we cannot give an explicit formula for these probabilities. Felsenstein explains in Chapter 3 in [43] how to obtain a recurrence to compute $|\mathcal{T}_n^*|$ for every $n \geq 1$, and an explicit algorithm to compute these values is provided in [114]; for more information on this sequence, including its generating function, see sequence A000669 in the OEIS [100]. It should be mentioned here that this uniform model is not equal to any α - γ -model. For instance, in Lemma 6 in [33] the probabilities of all trees in \mathcal{T}_4^* under the α - γ -model are computed explicitly, and it is easy to check that no choice of α and γ gives the same probabilities for all of them.

1.4 Hypergeometric series

The main goal of this section is to explain the *lookup algorithm* [81, p. 36], which is often applied in this memory to sum hypergeometric series, like for instance in Theorem 2.27. The main idea of the algorithm is to *standardize* the sum of a hypergeometric series, that is, to transform the sum of a hypergeometric series into a sum of a general hypergeometric series that can be expressed as a specific value of a hypergeometric function which, with some luck, can be computed using some of the many properties and specific values of hypergeometric functions established in the literature and gathered in handbooks like [1] or in WolframAlpha's *Mathematical Functions Site* [6]. This will allow us to know the value of many sums of hypergeometric series using known values

of hypergeometric functions.

We start with some definitions. A numerical series $\sum_{k \geq k_0} t_k$, with $t_k \in \mathbb{R}$ for every $k \geq k_0$, is a *hypergeometric series* when the ratio of two consecutive terms is a fixed rational function of the summation index k : i.e., when there exist polynomials $P(x), Q(x) \in \mathbb{R}[x]$ such that, for every $k \geq k_0$,

$$\frac{t_{k+1}}{t_k} = \frac{P(k)}{Q(k)}.$$

In this case, t_k is called a *hypergeometric term*.

A *general hypergeometric series* is a hypergeometric series of the form

$$\sum_{k \geq 0} \frac{(a_1)_k (a_2)_k \cdots (a_p)_k}{(b_1)_k (b_2)_k \cdots (b_q)_k} \cdot \frac{x^k}{k!} \quad (1.7)$$

where $a_1, \dots, a_p, b_1, \dots, b_q, x \in \mathbb{R}$ and $(a)_n$ denotes the *rising factorial function*, defined as follows:

$$(a)_n = \begin{cases} a(a+1)(a+2) \cdots (a+n-1) & \text{if } n \geq 1 \\ 1 & \text{if } n = 0 \end{cases} \quad (1.8)$$

Notice that the ratio of two successive terms in the general hypergeometric series (1.7) is

$$\frac{t_{k+1}}{t_k} = \frac{(k+a_1)(k+a_2) \cdots (k+a_p)}{(k+b_1)(k+b_2) \cdots (k+b_q)} \cdot \frac{x}{(k+1)},$$

and hence, for every $x \in \mathbb{R}$, it is a fixed rational function. The *hypergeometric function*

$${}_pF_q \left[\begin{matrix} a_1 & a_2 & \cdots & a_p \\ b_1 & b_2 & \cdots & b_q \end{matrix} ; x \right]$$

is simply the real function represented by the hypergeometric series (1.7) and defined on the convergence domain of the series. The real numbers a, \dots, a_p are called the *upper* parameters of the function, and b_1, \dots, b_q , the *lower* parameters of the function. The sum (1.7) is well-defined if no lower parameter belongs to $\mathbb{Z}_{\leq 0}$ and it is a convergent series if $p \leq q$, or if $p = q + 1$ and $|x| < 1$ [64]. If $p = 2$ and $q = 1$, the function becomes a traditional hypergeometric function ${}_2F_1(a, b; c; x)$.

By *standardizing* a hypergeometric series [81] we mean to write it explicitly as a general hypergeometric series in the form (1.7), up to a constant factor, when possible. To standardize a hypergeometric series, we shall use the following easy lemma:

Lemma 1.3. *Let $\sum_{k \geq 0} t_k$ and $\sum_{k \geq 0} t'_k$ be two hypergeometric series. Assume that their terms satisfy that $t_0 = t'_0 = 1$ and*

$$\frac{t_{k+1}}{t_k} = \frac{t'_{k+1}}{t'_k},$$

for every $k \geq 0$ (with the understanding that this implies that if $t_k = t'_k = 0$, then $t_{k+1} = t'_{k+1} = 0$, too). Then $t_k = t'_k$ for every $k \geq 0$, and therefore $\sum_{k \geq 0} t_k = \sum_{k \geq 0} t'_k$.

Proof. We shall prove this equality by induction on k . If $k = 0$ the equality holds by hypothesis. Assume now that $t_k = t'_k$. If both are 0, then by hypothesis $t_{k+1} = t'_{k+1} = 0$ and if $t_k = t'_k \neq 0$, then

$$t_{k+1} = t_k \cdot \frac{t_{k+1}}{t_k} = t'_k \cdot \frac{t'_{k+1}}{t'_k} = t'_{k+1}$$

as we wanted to prove. \square

The following corollary is a direct application of this lemma.

Corollary 1.4. *Let $\sum_{k \geq 0} t_k$ be a hypergeometric series such that $t_0 = 1$ and, for every $k \geq 0$,*

$$\frac{t_{k+1}}{t_k} = \frac{(k+a_1)(k+a_2)\cdots(k+a_p)}{(k+b_1)(k+b_2)\cdots(k+b_q)} \cdot \frac{x}{(k+1)}$$

for some $a_1, \dots, a_p, b_1, \dots, b_q \in \mathbb{R}$. Then

$$\sum_{k \geq 0} t_k = {}_pF_q \left[\begin{matrix} a_1 & a_2 & \cdots & a_p \\ b_1 & b_2 & \cdots & b_q \end{matrix} ; x \right].$$

The *lookup algorithm* is based on the last corollary. Its input is a sum $\sum_k t_k$. Its detailed steps are the following:

- (1) If necessary, apply a translation to the summation index k , so that the sum starts at $k = 0$ with a nonzero term t_0 .
- (2) If $t_0 \neq 1$, extract t_0 as a common factor of the sum.
- (3) If the sum is finite, say $\sum_{k=0}^{k_0-1} t_k$, write it as the difference of two series

$$\sum_{k=0}^{k_0-1} t_k = \sum_{k=0}^{\infty} t_k - \sum_{k=k_0}^{\infty} t_k = \sum_{k=0}^{\infty} t_k - \sum_{k=0}^{\infty} t_{k_0+k}$$

and proceed with both series separately.

After these three steps, we assume that we are dealing with a sum of the form $A \cdot \sum_{k=0}^{\infty} t_k$ with $A \in \mathbb{R}$ and $t_0 = 1$.

- (4) Write the ratio t_{k+1}/t_k as a rational expression $P(k)/Q(k)$, with P and Q polynomials. If this cannot be done, the series is not hypergeometric.

- (5) Completely factor the polynomials $P(x)$ and $Q(x)$ into real linear factors (if this cannot be done, i.e., if the polynomials P or Q have non-real roots, then the series is still hypergeometric, but we do not consider here this case as our hypergeometric series have their upper and lower parameters real) and write the term ratio in the form

$$\frac{P(k)}{Q(k)} = \frac{(k + a_1)(k + a_2) \cdots (k + a_p)}{(k + b_1)(k + b_2) \cdots (k + b_q)} \cdot \frac{x}{(k + 1)}$$

The factor $k + 1$ always has to be in the denominator, hence sometimes it has to be inserted and compensated in the numerator. Other extra numerical factors are included into the factor x .

- (6) By the last corollary, the series we are dealing from step (3) on is transformed into a general hypergeometric series

$$A \cdot \sum_{k=0}^{\infty} t_k = A \cdot {}_pF_q \left[\begin{matrix} a_1 & a_2 & \cdots & a_p \\ b_1 & b_2 & \cdots & b_q \end{matrix} ; x \right]$$

- (7) Compare the obtained hypergeometric function with some handbook or database of hypergeometric function properties and specific values, like for instance the texts [1, 64] or the site [6], to simplify the sum and finally compute it.

Futhermore, and in order to ease the task of the reader, we gather all properties of hypergeometric functions used in this memory. Before proceeding, and since many of these formulas involve the classical *gamma function* $\Gamma(x)$, we review its definition and the properties we use here. Since we only need it on real arguments, for simplicity we restrict ourselves to this setting.

The *gamma function* Γ is defined on the set of all real numbers outside $\mathbb{Z}_{\leq 0}$ in the following way. One first defines Γ on any positive real number $x > 0$ by means of

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

When $x > 1$, using integration by parts we obtain the following well known identity:

$$\Gamma(x) = (x - 1)\Gamma(x - 1). \quad (1.9)$$

Then, the function Γ is extended from $\mathbb{R}_{>0}$ to $\mathbb{R} \setminus \mathbb{Z}_{\leq 0}$ by means of this identity, i.e., through $\Gamma(x - 1) = \Gamma(x)/(x - 1)$.

Equation (1.9) implies by induction (since $\Gamma(1) = \int_0^{\infty} e^{-t} dt = 1$) that

$$\Gamma(n) = (n - 1)! \quad \text{for every } n \in \mathbb{N}_{\geq 1}. \quad (1.10)$$

This shows that Γ extends to $\mathbb{R} \setminus \mathbb{Z}_{\leq 0}$ the factorial of non-zero natural numbers. But it is wrong to define $\Gamma(0)$ as 1, because actually $\lim_{x \rightarrow 0^+} \Gamma(x) = \infty$.

Other properties of the Gamma function used in this memory are:

(1.11) $\Gamma(1/2) = \sqrt{\pi}$; see Formula 6.1.8 in [1].

(1.12) *Euler's reflection formula*: for every $x \notin \mathbb{Z}$,

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi \cdot x)};$$

see Formula 6.1.17 in [1]. This implies by induction on n that for every $x \notin \mathbb{Z}$

$$\Gamma(x-n) = \frac{(-1)^{n-1}\Gamma(-x)\Gamma(1+x)}{\Gamma(n+1-x)}.$$

(1.13) $\Gamma\left(n + \frac{1}{2}\right) = \frac{(2n-1)!!\sqrt{\pi}}{2^n}$; see Formula 6.1.12 in [1].

In particular, $\Gamma(3/2) = \sqrt{\pi}/2$.

(1.14) $\Gamma\left(\frac{1}{2} - n\right) = \frac{(-1)^n 2^n \sqrt{\pi}}{(2n-1)!!}$; see <http://functions.wolfram.com/06.05.03.0010.01> in [6].

In particular, $\Gamma(-1/2) = -2\sqrt{\pi}$ and $\Gamma(-3/2) = 4\sqrt{\pi}/3$.

(1.15) The Maclaurin series of Γ around n and $-n$, for $n \in \mathbb{N}$, have the form:

$$\Gamma(-n+x) = \frac{(-1)^n}{n!x}(1+O(x)), \quad \Gamma(n-x) = (n-1)! + O(x),$$

by property <http://functions.wolfram.com/06.05.06.0008.01> and Euler's reflection formula (1.12)

Now, we list the formulas on hypergeometric functions used in this work. For each formula we provide a reference.

(1.16) ${}_1F_0\left[\begin{matrix} a \\ - \end{matrix}; x\right] = (1-x)^{-a}$; see <http://functions.wolfram.com/07.19.02.0002.01>.

(1.17) If $(1+a+b)/2 > 1$,

$${}_2F_1\left[\begin{matrix} a, & b \\ (1+a+b)/2 \end{matrix}; \frac{1}{2}\right] = \frac{\Gamma((1+a+b)/2)\sqrt{\pi}}{\Gamma((1+a)/2)\Gamma((1+b)/2)};$$

see Formula 15.1.24 from [1].

(1.18) ${}_2F_1\left[\begin{matrix} a, & b \\ (a+b-1)/2 \end{matrix}; \frac{1}{2}\right]$
 $= \frac{2^{b-1}\Gamma((a+b-1)/2)}{\Gamma(b)} \left(\frac{\Gamma(b/2)}{\Gamma((a-1)/2)} + \frac{2\Gamma((b+1)/2)}{\Gamma(a/2)} + \frac{\Gamma((b+2)/2)}{\Gamma((a+1)/2)} \right)$; see <http://functions.wolfram.com/07.23.03.0023.01>.

$$(1.19) \quad {}_2F_1 \left[\begin{matrix} b, & a \\ b & \end{matrix} ; x \right] = {}_1F_0 \left[\begin{matrix} a \\ - \end{matrix} ; x \right] = (1-x)^{-a}; \text{ see Formula 15.1.8 in [1].}$$

$$(1.20) \quad c(1-x) \cdot {}_2F_1 \left[\begin{matrix} a, & b \\ c & \end{matrix} ; x \right] = c \cdot {}_2F_1 \left[\begin{matrix} a-1, & b \\ c & \end{matrix} ; x \right] - (c-b)x \cdot {}_2F_1 \left[\begin{matrix} a, & b \\ c+1 & \end{matrix} ; x \right]; \\ \text{see Formula 15.2.20 in [1].}$$

$$(1.21) \quad {}_2F_1 \left[\begin{matrix} a, & b \\ c & \end{matrix} ; x \right] = (1-x)^{-a} {}_2F_1 \left[\begin{matrix} a, & c-b \\ c & \end{matrix} ; \frac{x}{x-1} \right]; \text{ see Formula 15.3.4} \\ \text{in [1]}$$

$$(1.22) \quad {}_2F_1 \left[\begin{matrix} a, & b \\ a-b-1 & \end{matrix} ; -1 \right] \\ = \frac{2^{-2(b+1)}\Gamma(a-b-1)}{\Gamma(a-2b-1)} \left(\frac{\Gamma((a-1-2b)/2)}{\Gamma((a-1)/2)} + \frac{\Gamma((a+1-2b)/2)}{\Gamma((a+1)/2)} + \frac{2\Gamma((a-2b)/2)}{\Gamma(a/2)} \right); \\ \text{see } \text{http://functions.wolfram.com/07.23.03.0005.01}$$

$$(1.23) \quad {}_3F_2 \left[\begin{matrix} a, & b, & c \\ a-1, & e & \end{matrix} ; x \right] = {}_2F_1 \left[\begin{matrix} b, & c \\ e & \end{matrix} ; x \right] + \frac{bcx}{(a-1)e} \cdot {}_2F_1 \left[\begin{matrix} b+1, & c+1 \\ e+1 & \end{matrix} ; x \right]; \\ \text{see } \text{http://functions.wolfram.com/07.27.03.0118.01}$$

$$(1.24) \quad \text{If } d-b-c > 1,$$

$${}_3F_2 \left[\begin{matrix} a, & b, & c \\ d, & a-1 & \end{matrix} ; 1 \right] = \frac{\Gamma(d)\Gamma(d-b-c)}{\Gamma(d-b)\Gamma(d-c)} \left(1 - \frac{bc}{(a-1)(b+c-d+1)} \right);$$

see <http://functions.wolfram.com/07.27.03.0005.01>

$$(1.25) \quad \text{If } s = d + e - a - b - c,$$

$${}_3F_2 \left[\begin{matrix} a, & b, & c \\ d, & e & \end{matrix} ; 1 \right] = \frac{\Gamma(d)\Gamma(e)\Gamma(s)}{\Gamma(a)\Gamma(s+b)\Gamma(s+c)} {}_3F_2 \left[\begin{matrix} d-a, & e-a, & s \\ s+b, & s+c & \end{matrix} ; 1 \right];$$

see Expression (3.1.2) in [47, p. 59]

$$(1.26) \quad \text{If } d-a-b > -1 \text{ and } a, b, d \neq 1$$

$${}_3F_2 \left[\begin{matrix} 1, & a, & b \\ 2, & d & \end{matrix} ; 1 \right] = \frac{d-1}{(a-1)(b-1)} \left(\frac{\Gamma(d-1)\Gamma(d-a-b+1)}{\Gamma(d-a)\Gamma(d-b)} - 1 \right);$$

see <http://functions.wolfram.com/07.27.03.0021.01>

When the value of the hypergeometric function obtained through the lookup algorithm cannot be found in any database of hypergeometric functions, one must try something else to compute its value. In this Thesis it has only been the case with one hypergeometric function, and then we have solved this problem using the so-called *Gosper's algorithm* [81, 49]. The main idea of this procedure is to transform the general hypergeometric term of the series into a telescopic term.

Lemma 1.5. *Given a hypergeometric term t_k , if there exists a hypergeometric term z_k such that*

$$t_k = z_{k+1} - z_k \quad (1.11)$$

then $\sum_{k \geq 0}^n t_k = z_{n+1} - z_0$.

Proof. Let t_n and z_n be two hypergeometric terms that satisfy identity (1.11). Then, for every $n \geq 0$,

$$z_{n+1} = z_n + t_n = z_{n-1} + t_{n-1} + t_n = \cdots = z_0 + \sum_{k=0}^n t_k.$$

□

As a direct consequence of this lemma we obtain that, given a hypergeometric series $\sum_n t_n$, if z_n is a hypergeometric term such that $t_n = z_{n+1} - z_n$, then

$$\frac{z_n}{t_n} = \frac{z_n}{z_{n+1} - z_n} = \frac{1}{\frac{z_{n+1}}{z_n} - 1}$$

is a rational function of n . Let us denote it by y_n . By equation (1.11), we have that

$$\frac{t_{n+1}}{t_n} \cdot y_{n+1} - y_n = \frac{t_{n+1}}{t_n} \cdot \frac{z_{n+1}}{t_{n+1}} - \frac{z_n}{t_n} = \frac{z_{n+1} - z_n}{t_n} = 1.$$

This corresponds to the first-order linear recurrence with rational coefficients

$$y_{n+1} = \frac{t_n}{t_{n+1}} \cdot y_n + \frac{t_n}{t_{n+1}}. \quad (1.12)$$

The problem of finding hypergeometric sums is reduced to the problem of finding a rational solution of recurrences of this type, because

$$\sum_{k=0}^n t_k = z_{n+1} - z_0 = y_{n+1} t_{n+1} - z_0.$$

The procedure to solve (1.12) is based on the following key result, whose proof can be found in [81].

Theorem 1.6. *Let $r \in \mathbb{R}[n]$ be a nonzero rational function. Then, there exist three unique polynomials $a, b, c \in \mathbb{R}[n]$ satisfying*

- (a) b and c are monic;
- (b) $\gcd(a(n), b(n+h)) = 1$, for all nonnegative integers h ;
- (c) $\gcd(a(n), c(n)) = 1$;

(d) $\gcd(a(n), c(n+1)) = 1$;

and such that

$$r(n) = \frac{a(n)}{b(n)} \cdot \frac{c(n+1)}{c(n)}.$$

In particular, we have

$$\frac{t_{n+1}}{t_n} = \frac{a(n)}{b(n)} \cdot \frac{c(n+1)}{c(n)} \quad (1.13)$$

for some $a(n), b(n), c(n) \in \mathbb{R}[n]$ such that

$$\gcd(a(n), b(n+h)) = 1, \text{ for all nonnegative integers } h. \quad (1.14)$$

We shall look for a nonzero rational solution of recurrence (1.12) of the form

$$y_n = \frac{b(n-1)}{c(n)} x(n)$$

with $x(n)$ a function of n . Since

$$\frac{t_{n+1}}{t_n} \cdot y_{n+1} - y_n = 1,$$

it must happen that

$$\frac{a(n)}{b(n)} \cdot \frac{c(n+1)}{c(n)} \cdot \frac{b(n)}{c(n+1)} x(n+1) - \frac{b(n-1)}{c(n)} x(n) = 1$$

which amounts to the equation

$$a(n)x(n+1) - b(n-1)x(n) = c(n). \quad (1.15)$$

In this situation we have the following theorem, which is the base of Gosper's algorithm [49].

Theorem 1.7. *Let $a(n)$, $b(n)$ and $c(n)$ be real polynomials in n such that equation (1.14) holds. If $x(n)$ is a rational function of n satisfying (1.15), then $x(n)$ is a polynomial in n .*

Therefore, finding hypergeometric solutions of (1.11) is equivalent to finding polynomial solutions of (1.15), because once we know a nonzero polynomial solution $x(n)$ of this last equation, we will have

$$z_n = \frac{b(n-1)x(n)}{c(n)} t_n.$$

So, based on the previous results, *Gosper's algorithm* finds the value of the hypergeometric sum $\sum_k t_k$ by means of the following steps:

- (1) Write the quotient t_{k+1}/t_k as a rational expression of k .
- (2) Write this quotient in the form $\frac{a(k)}{b(k)} \cdot \frac{c(k+1)}{c(k)}$, where $a(k)$, $b(k)$ and $c(k)$ are polynomials satisfying equation (1.14), following, for instance, this algorithm:
- (2.1) Write the quotient t_{k+1}/t_k as $W \cdot f(k)/g(k)$, where f, g are relatively prime monic polynomials and W is a constant. Let $R(h)$ be the resultant of polynomials $f(k)$ and $g(k+h)$,

$$R(h) = \text{Resultant}(f(k), g(k+h)),$$

and let $\mathcal{S} = \{h_1, \dots, h_N\}$ be the set of nonnegative integer roots of $R(h)$, with $N \geq 0$ and $0 \leq h_1 < h_2 < \dots < h_N$.

- (2.2) Let $p_0(k) = f(k)$ and $q_0(k) = g(k)$ and iterate for j from 1 to N the following equations:

$$\begin{aligned} s_j(k) &= \gcd(p_{j-1}(k), q_{j-1}(k+h_j)) \\ p_j(k) &= p_{j-1}(k)/s_j(k) \\ q_j(k) &= q_{j-1}(k)/s_j(k-h_j) \end{aligned}$$

Take finally

$$a(k) = W p_N(k), \quad b(k) = q_N(k), \quad c(k) = \prod_{i=1}^N \prod_{j=1}^{h_i} s_i(k-j)$$

- (3) To find the polynomial $x(n)$, if we knew its degree d , we would only have to solve a linear system of equations from equation (1.15) to calculate its coefficients. The computation of d is easy taking into account the leading terms of the polynomials $a(k)$, $b(k)$ and $c(k)$. If such polynomial $x(n)$ does not exist, stop.

- (4) Return $\frac{b(n-1)x(n)}{c(n)}t_n$ and stop.

Once we have z_n , the sum $\sum_{k=0}^n t_k$ that we wanted to compute is equal to $z_{n+1} - z_0$.

For instance, in page 143 we shall need to compute

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

using Gosper's algorithm. In this case,

$$t_k = (4k-1) \frac{(2k-3)!!}{(2k-2)!!}$$

and hence

$$\frac{t_{k+1}}{t_k} = \frac{(4k+3)(2k-1)}{2k(4k-1)} = \frac{(k+\frac{3}{4})(k-\frac{1}{2})}{k(k-\frac{1}{4})}.$$

Thus, taking $a(k) = k - 1/2$, $b(k) = k$ and $c(k) = k - 1/4$, we have a decomposition of t_{k+1}/t_k as required in step (2) of the algorithm. Now equation (1.15) becomes

$$\left(n - \frac{1}{2}\right)x(n+1) - (n-1)x(n) = n - \frac{1}{4},$$

from which we see that we can take $x(n)$ of degree 1. Let $x(n) = \alpha n + \beta$. Then the previous equation becomes

$$\left(n - \frac{1}{2}\right)(\alpha(n+1) + \beta) - (n-1)(\alpha n + \beta) = n - \frac{1}{4},$$

i.e.

$$\frac{3\alpha}{2}n + \frac{\beta}{2} - \frac{\alpha}{2} = n - \frac{1}{4}$$

from where we obtain $\alpha = 2/3$ and $\beta = 1/6$ and hence

$$x(n) = \frac{2}{3}n + \frac{1}{6}.$$

Consequently,

$$\begin{aligned} z_n &= \frac{b(n-1)x(n)}{c(n)}t_n = \frac{(n-1)\left(\frac{2}{3}n + \frac{1}{6}\right)}{n - \frac{1}{4}} \cdot (4n-1) \frac{(2n-3)!!}{(2n-2)!!} \\ &= \frac{4}{3}(n-1)\left(2n + \frac{1}{2}\right) \frac{(2n-3)!!}{(2n-2)!!} \\ &= \frac{2}{3}(n-1)(4n+1) \frac{(2n-3)!!}{(2n-2)!!} \\ &= \frac{1}{3 \cdot 2^{2n-4}}(n-1)(4n+1) \binom{2n-3}{n-1} \end{aligned}$$

Finally,

$$\begin{aligned} \sum_{k=3}^{n-1} (4k-1) \frac{(2k-3)!!}{(2k-2)!!} &= \frac{1}{3 \cdot 2^{2n-4}}(n-1)(4n+1) \binom{2n-3}{n-1} - \frac{13}{2} \\ &= \frac{1}{3 \cdot 2^{2n+1}} \left(32(4n^2 - 3n - 1) \binom{2n-3}{n-1} - 39 \cdot 2^{2n} \right). \end{aligned}$$

Chapter 2

The total cophenetic index

In this chapter, we introduce a new balance index for phylogenetic trees, called the *total cophenetic index*. It derives from the Sackin index, by replacing in it the depths of the leaves by the depths of the lowest common ancestors of pairs of different leaves, the *cophenetic values* of the tree, hence the name of the index. As it also happens with the Sackin and Colless indices, our total cophenetic index actually measures the “imbalance” of the tree: smaller values of the index correspond to more balanced trees.

2.1 Main definitions

For every pair of leaves i, j in a phylogenetic tree T , their *cophenetic value* [102] is the depth of their lowest common ancestor (their LCA, for short) $[i, j]_T$:

$$\varphi_T(i, j) = \delta_T([i, j]_T).$$

In other words, the cophenetic value of a pair of leaves is the depth at which the leaves diverge.

Example 2.1. *If T stands for the phylogenetic tree depicted in Fig. 2.1, the cophenetic values of its pairs of leaves are:*

$$\begin{aligned} \varphi_T(1, 2) = 2, \quad \varphi_T(1, 3) = 1, \quad \varphi_T(1, 4) = \varphi_T(1, 5) = 0, \quad \varphi_T(2, 3) = 1, \\ \varphi_T(2, 4) = \varphi_T(2, 5) = 0, \quad \varphi_T(3, 4) = \varphi_T(3, 5) = 0, \quad \varphi_T(4, 5) = 1 \end{aligned}$$

Definition 2.2. *The total cophenetic index of a phylogenetic tree $T \in \mathcal{T}_n$ is the sum of the cophenetic values of its pairs of different leaves:*

$$\Phi(T) = \sum_{1 \leq i < j \leq n} \varphi_T(i, j).$$

This index can be seen as an extension of Sackin’s: instead of adding up the depths of the leaves (that is, the depths of the LCA of every leaf and itself),

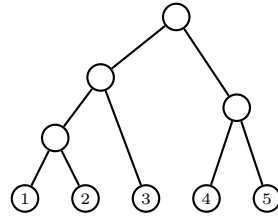


Figure 2.1: A phylogenetic tree with 5 leaves.

$\Phi(T)$ adds up the depths of the LCA of every pair of different leaves in T . Notice also that, as Sackin's and Colless' indices, $\Phi(T)$ only depends on the shape of T , and in particular it is invariant under permutations of its labels or their actual values.

Fig. 2.2 shows all possible shapes of phylogenetic trees with 5 leaves, and their total cophenetic indices. Although we shall return on it later for trees with an arbitrary number n of leaves, notice that the rooted star has the smallest total cophenetic value, 0; the bifurcating tree with the smallest total cophenetic value, 5, is the maximally balanced; and the tree with the largest total cophenetic value, 10, is the comb.

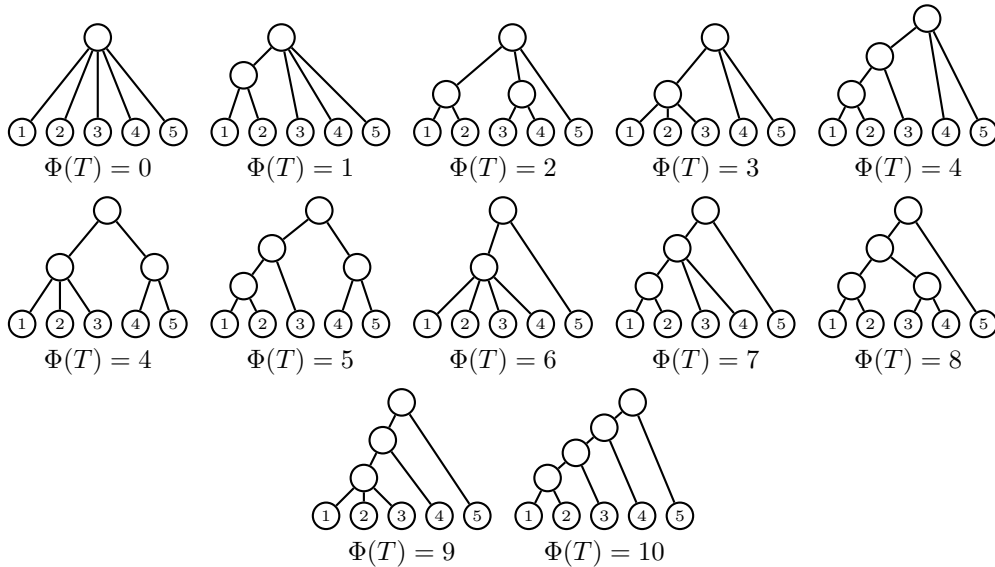


Figure 2.2: All phylogenetic trees with 5 leaves, up to relabelings, and their total cophenetic indices.

As it was the case with Sackin's index, we can express the total cophenetic index in terms of the cluster sizes $(\kappa_T(v))_{v \in V_{int}(T)}$ of the tree. As a result, we obtain the following alternative expression for $\Phi(T)$ that will be used in many proofs.

Lemma 2.3. *Let $T \in \mathcal{T}_n$ be a phylogenetic tree with root r . Then,*

$$\Phi(T) = \sum_{v \in V_{\text{int}}(T) - \{r\}} \binom{\kappa_T(v)}{2}.$$

Proof. For every $v \in V_{\text{int}}(T) - \{r\}$, consider the function $\gamma_v : L(T)^2 \rightarrow \{0, 1\}$ that tells whether a pair of leaves $i, j \in L(T)$ are contained in its cluster $C_T(v)$:

$$\gamma_v(i, j) = \begin{cases} 1 & \text{if } i, j \in C_T(v) \\ 0 & \text{otherwise} \end{cases}$$

Then, $\varphi_T(i, j) = \sum_{v \in V_{\text{int}}(T) - \{r\}} \gamma_v(i, j)$ and thus

$$\begin{aligned} \Phi(T) &= \sum_{1 \leq i < j \leq n} \sum_{v \in V_{\text{int}}(T) - \{r\}} \gamma_v(i, j) = \sum_{v \in V_{\text{int}}(T) - \{r\}} \sum_{1 \leq i < j \leq n} \gamma_v(i, j) \\ &= \sum_{v \in V_{\text{int}}(T) - \{r\}} \left| \{ \{i, j\} \mid 1 \leq i < j \leq n, i, j \in C_T(v) \} \right| \\ &= \sum_{v \in V_{\text{int}}(T) - \{r\}} \binom{|C_T(v)|}{2} = \sum_{v \in V_{\text{int}}(T) - \{r\}} \binom{\kappa_T(v)}{2}. \end{aligned}$$

□

The following lemma expresses the total cophenetic index of a tree in terms of the numbers of leaves and the total cophenetic indices of the subtrees rooted at the children of its root.

Lemma 2.4. *Let $T \in \mathcal{T}_n$ be a phylogenetic tree with root r , and let T_1, \dots, T_k , $k \geq 2$, be the subtrees rooted at the children of r , so that $T = T_1 \star \dots \star T_k$; cf. Fig 2.3. Then,*

$$\Phi(T) = \sum_{i=1}^k \Phi(T_i) + \sum_{i=1}^k \binom{|L(T_i)|}{2}.$$

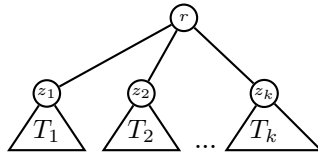


Figure 2.3: The tree T in the statement of Lemma 2.4.

Proof. Let z_i be the root of T_i , $i = 1, \dots, k$. Then, by Lemma 2.3,

$$\begin{aligned} \Phi(T) &= \sum_{v \in V_{\text{int}}(T) - \{r\}} \binom{\kappa_T(v)}{2} = \sum_{i=1}^k \sum_{v \in V_{\text{int}}(T_i)} \binom{\kappa_{T_i}(v)}{2} \\ &= \sum_{i=1}^k \left(\binom{\kappa_{T_i}(z_i)}{2} + \sum_{v \in V_{\text{int}}(T_i) - \{z_i\}} \binom{\kappa_{T_i}(v)}{2} \right) \\ &= \sum_{i=1}^k \left(\binom{|L(T_i)|}{2} + \Phi(T_i) \right). \end{aligned}$$

□

This shows that the total cophenetic index is a recursive shape index in the sense of Section 1.2 (page 17).

Next lemma shows that the total cophenetic index is local, in the sense that if two trees differ only on a rooted subtree, then the difference between their total cophenetic indices is equal to that of these subtrees. Sackin's and Colless' indices also satisfy this property.

Lemma 2.5. *Let T_0 and T'_0 be two phylogenetic trees with $L(T_0) = L(T'_0) \subseteq [n]$, let $T \in \mathcal{T}_n$ be such that its subtree rooted at some node z is T_0 , and let $T' \in \mathcal{T}_n$ be the tree obtained from T by replacing T_0 by T'_0 as its subtree rooted at z . Then*

$$\Phi(T) - \Phi(T') = \Phi(T_0) - \Phi(T'_0).$$

Proof. Without any loss of generality, assume that $L(T_0) = L(T'_0) = [m]$ with $m \leq n$. Let $k = \delta_T(z) = \delta_{T'}(z)$. Then, for every $1 \leq i < j \leq m$,

$$\varphi_T(i, j) = k + \varphi_{T_0}(i, j), \quad \varphi_{T'}(i, j) = k + \varphi_{T'_0}(i, j).$$

On the other hand, if $i > m$ or $j > m$, then $[i, j]_T = [i, j]_{T'}$ and hence $\varphi_T(i, j) = \varphi_{T'}(i, j)$. Therefore

$$\begin{aligned} \Phi(T) - \Phi(T') &= \sum_{1 \leq i < j \leq n} (\varphi_T(i, j) - \varphi_{T'}(i, j)) \\ &= \sum_{1 \leq i < j \leq m} (\varphi_T(i, j) - \varphi_{T'}(i, j)) \\ &= \sum_{1 \leq i < j \leq m} (\varphi_{T_0}(i, j) - \varphi_{T'_0}(i, j)) = \Phi(T_0) - \Phi(T'_0). \end{aligned}$$

□

The *nodal distance* $d_T(i, j)$ between a pair of leaves i, j is the length of the unique undirected path connecting them; equivalently, it is the sum of the

lengths of the paths from $[i, j]_T$ to i and j . The *total area* [72] of a tree $T \in \mathcal{T}_n$ is defined as

$$D(T) = \sum_{1 \leq i < j \leq n} d_T(i, j).$$

There is an easy relation between the total cophenetic index $\Phi(T)$, the Sackin index $S(T)$ and the total area $D(T)$, which will be used in later proofs, like for instance those of Corollary 2.22 and Theorem 2.28.

Lemma 2.6. *For every $T \in \mathcal{T}_n$,*

$$(n - 1)S(T) = 2\Phi(T) + D(T).$$

Proof. It is straightforward to check (see, for instance, [42]) that, for every $i, j \in L(T)$,

$$\delta_T(i) + \delta_T(j) = d_T(i, j) + 2\varphi_T(i, j)$$

since the paths from the root to the leaves i and j bifurcate at the LCA of the pair of leaves.

Therefore,

$$\begin{aligned} 2\Phi(T) + D(T) &= \sum_{1 \leq i < j \leq n} (2\varphi_T(i, j) + d_T(i, j)) = \sum_{1 \leq i < j \leq n} (\delta_T(i) + \delta_T(j)) \\ &= (n - 1) \sum_{i=1}^n \delta_T(i) = (n - 1)S(T). \end{aligned}$$

□

2.2 Trees with maximum and minimum Φ

In this section we determine which trees in \mathcal{T}_n and \mathcal{BT}_n have the largest and smallest total cophenetic indices. We begin by establishing several lemmas that will allow us to find the trees with maximum value of Φ on \mathcal{T}_n .

Lemma 2.7. *Let T_1, \dots, T_k , with $k \geq 3$, be an ordered forest on $[m]$. Consider the trees $T_0, T'_0 \in \mathcal{T}_n$ described in Fig. 2.4. Then, $\Phi(T'_0) - \Phi(T_0) > 0$.*

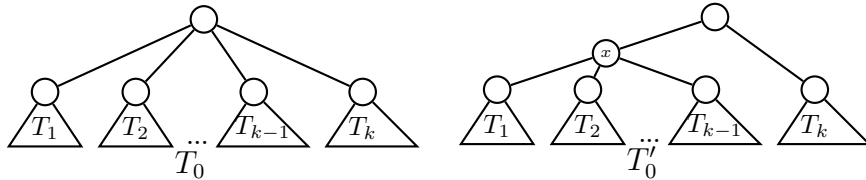


Figure 2.4: The trees T_0 and T'_0 in the statement of Lemma 2.7.

Proof. With the notations of Fig. 2.4, notice that

$$\begin{aligned} \Phi(T'_0) - \Phi(T_0) &= \sum_{v \in V_{int}(T'_0) - \{r\}} \binom{\kappa_{T'_0}(v)}{2} - \sum_{v \in V_{int}(T_0) - \{r\}} \binom{\kappa_{T_0}(v)}{2} \\ &= \binom{\kappa_{T'_0}(x)}{2} + \sum_{i=1}^k \sum_{v \in V_{int}(T_i)} \binom{\kappa_{T_i}(v)}{2} - \sum_{i=1}^k \sum_{v \in V_{int}(T_i)} \binom{\kappa_{T_i}(v)}{2} \\ &= \binom{\kappa_{T'_0}(x)}{2} > 0. \end{aligned}$$

□

Lemma 2.8. *For every $n \geq 3$,*

$$\sum_{i=2}^{n-1} \binom{i}{2} = \binom{n}{3}.$$

Proof. Notice that there exists a bijection

$$\begin{aligned} \{\{i, j, k\} \subseteq [n] \mid i < j < k\} &\rightarrow \bigsqcup_{k=3}^n \{\{i, j\} \subseteq \{1, \dots, k-1\} \mid i < j\} \times \{k\} \\ \{i, j, k\} &\mapsto (\{i, j\}, k) \end{aligned}$$

Now, the cardinality of the left-hand side set is $\binom{n}{3}$ and, for every $k = 3, \dots, n$,

$$\left| \{\{i, j\} \subseteq \{1, \dots, k-1\} \mid i < j\} \right| = \binom{k-1}{2}.$$

Therefore

$$\binom{n}{3} = \sum_{k=3}^n \binom{k-1}{2} = \sum_{i=2}^{n-1} \binom{i}{2}.$$

□

Corollary 2.9. *For every non-bifurcating phylogenetic tree $T \in \mathcal{T}_n$, there always exists a bifurcating phylogenetic tree $T' \in \mathcal{BT}_n$ such that $\Phi(T') > \Phi(T)$.*

Proof. Let $T \in \mathcal{T}_n$ be a non-bifurcating phylogenetic tree. Then it contains an internal node z whose rooted subtree looks like the tree T_0 in Lemma 2.7, for some $k \geq 3$. By Lemmas 2.5 and 2.7, and with the notations of the latter, if $T' \in \mathcal{T}_n$ is the tree obtained from T by replacing T_0 by T'_0 as its subtree rooted at z , then $\Phi(T') - \Phi(T) > 0$. If we iterate this procedure while there remain non-bifurcating internal nodes in the tree, at each step the total cophenetic index of the resulting tree increases, and when we stop we obtain a bifurcating tree whose total cophenetic index is larger than $\Phi(T)$. □

Therefore, the maximum total cophenetic index on \mathcal{T}_n is reached at a bifurcating tree. It remains thus to determine the bifurcating trees with n leaves that have the largest Φ .

Lemma 2.10. *Let $m \geq 4$, let $2 \leq k \leq m - 2$, let T_1 be any bifurcating tree on $\{k + 1, \dots, m\}$, and let T_0 and T'_0 be the phylogenetic trees in \mathcal{BT}_m depicted in Fig. 2.5. Then, $\Phi(T'_0) - \Phi(T_0) > 0$.*

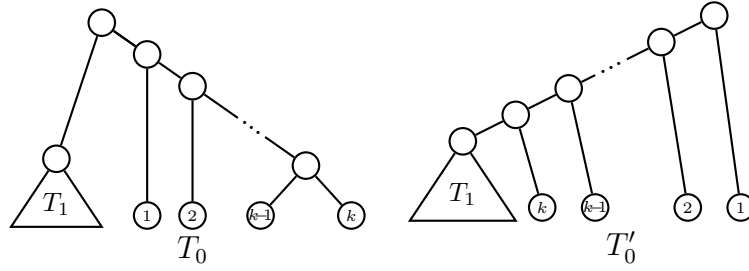


Figure 2.5: The trees T_0 and T'_0 in the statement of Lemma 2.10.

Proof. By Lemmas 2.3 and 2.4, and recalling that $|L(T_1)| = m - k$, we have that

$$\begin{aligned}\Phi(T_0) &= \binom{m-k}{2} + \Phi(T_1) + \binom{k}{2} + \binom{k-1}{2} + \dots + \binom{3}{2} + \binom{2}{2} \\ \Phi(T'_0) &= \binom{m-1}{2} + \binom{m-2}{2} + \dots + \binom{m-k+1}{2} + \binom{m-k}{2} + \Phi(T_1)\end{aligned}$$

and hence

$$\begin{aligned}\Phi(T'_0) - \Phi(T_0) &= \sum_{i=m-k+1}^{m-1} \binom{i}{2} - \sum_{i=2}^k \binom{i}{2} \\ &= \sum_{i=2}^{m-1} \binom{i}{2} - \sum_{i=2}^{m-k} \binom{i}{2} - \sum_{i=2}^k \binom{i}{2} \\ &= \binom{m}{3} - \binom{m-k+1}{3} - \binom{k+1}{3} \\ &\quad \text{(by Lemma 2.8)} \\ &= \frac{1}{2}(k-1)m(m-k-1) > 0\end{aligned}$$

because $m - k \geq 2$. □

Proposition 2.11. *The trees in \mathcal{T}_n with maximum total cophenetic index are exactly the combs K_n , and this maximum is $\Phi(K_n) = \binom{n}{3}$.*

Proof. By Corollary 2.9, any tree in \mathcal{T}_n with maximum total cophenetic index will be bifurcating. Now, we shall prove that if $T \in \mathcal{BT}_n$ is not a comb, then $\Phi(T)$ is not maximum. Since \mathcal{BT}_n is finite, this implies that the largest total cophenetic index is reached exactly at the combs.

So, let $T \in \mathcal{BT}_n$ be a bifurcating phylogenetic tree that is not a comb. Therefore, it has an internal node z of largest depth without any leaf child; in particular, all internal descendant nodes of z have some leaf child. Thus, and up to a relabeling of its leaves, the subtree of T rooted at z has the form of the tree T_0 in Fig. 2.6, for some $k \geq 2$ and some $l \geq k + 2$. But then, by Lemma 2.10 (taking as T_1 the comb subtree rooted at the parent x of the leaf k), the tree T'_0 also depicted in Fig. 2.6 has a strictly larger total cophenetic index. Then, by Lemma 2.5, if we replace in T the subtree rooted at z by this tree T'_0 , we obtain a new tree T' with $\Phi(T') > \Phi(T)$. So, if $T \in \mathcal{BT}_n$ is not a comb, then there exists another tree $T' \in \mathcal{BT}_n$ with larger total cophenetic index, as we claimed.

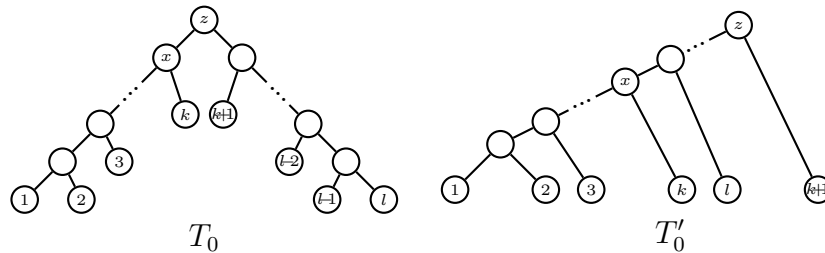


Figure 2.6: The trees T_0 and T'_0 in the proof of Proposition 2.11.

As to the value of Φ on the comb K_n with n leaves depicted in Fig. 1.2.(a), since the parent of the leaf labelled j , for $j = 2, \dots, n$, has j descendant leaves, by Lemmas 2.3 and 2.8 we have that

$$\Phi(K_n) = \sum_{j=2}^{n-1} \binom{j}{2} = \binom{n}{3}.$$

□

As far as the minimum total cophenetic index goes, we have the following result:

Proposition 2.12. *The tree in \mathcal{T}_n with minimum total cophenetic index is the rooted star tree S_n (depicted in Fig. 1.2.(b)), and this minimum is $\Phi(S_n) = 0$.*

Proof. Since the LCA of every pair of leaves of the rooted star tree S_n is the root, all cophenetic values in S_n are 0 and therefore $\Phi(S_n) = 0$. Conversely, if $T \in \mathcal{T}_n$ is not the rooted star tree, then it has some non-root internal node, whose pairs of descendant leaves have thus non-zero cophenetic value and hence $\Phi(T) > 0$. □

Therefore, the range of Φ on \mathcal{T}_n goes from 0 to $\binom{n}{3}$. This is one order of magnitude larger than the range of Sackin's and Colless' indices, whose maximum values, reached also at the combs, have both order $O(n^2)$ (see Section 1.2).

Fig. 2.7 recalls the example of Fig. 1.5 in Section 1.2, where we observed ties between the Sackin and Colless indices of two phylogenetic trees with different shape. Now, their total cophenetic indices are different, and according to them T_1 is more balanced than T_2 , as it should be desired, because T_1 is maximally balanced and T_2 is not.

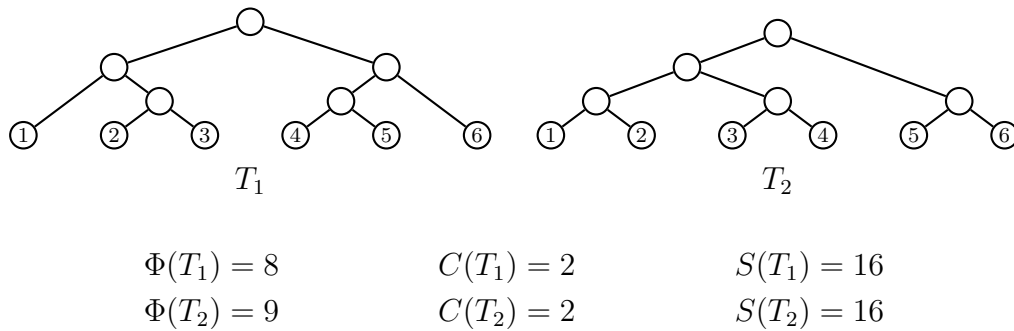


Figure 2.7: Two phylogenetic trees with different shape that have the same Colless index and the same Sackin index but different total cophenetic indices.

Let us characterize now those *bifurcating* phylogenetic trees with smallest total cophenetic index.

Lemma 2.13. *Let T_1, T_2, T_3, T_4 be an ordered bifurcating forest on $[m]$, let $n_i = |L(T_i)|$, for $i = 1, 2, 3, 4$, and assume that $n_1 \geq n_2$, $n_3 \geq n_4$ and $n_1 > n_3$. Let T_0 the phylogenetic tree depicted in Fig 2.8.(a), and let $T \in \mathcal{BT}_n$ ($n \geq m$) be a bifurcating phylogenetic tree having T_0 as the subtree rooted at some node. If $\Phi(T)$ is minimum in \mathcal{BT}_n , then $n_4 \geq n_2$.*

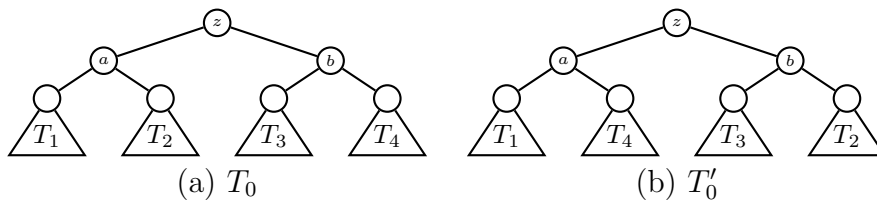


Figure 2.8: (a) The tree T_0 in the statement of Lemma 2.13. (b) The tree T'_0 in the proof of Lemma 2.13.

Proof. Assume that $n_2 > n_4$. We shall show that, in this case, a suitable interchange of subtrees rooted at cousins in T_0 produces a tree with smaller

total cophenetic index, which in particular will imply that $\Phi(T)$ cannot be the minimum in \mathcal{BT}_n .

Assume that the tree T in the statement has the subtree T_0 rooted at a node z . Consider the tree T'_0 obtained by interchanging in T_0 the subtrees T_2 and T_4 (see Fig. 2.8.(b)) and let T' be the tree obtained from T by replacing T_0 by T'_0 as its subtree rooted at z . Then, by Lemma 2.3,

$$\begin{aligned} \Phi(T') - \Phi(T) &= \Phi(T'_0) - \Phi(T_0) \\ &= \binom{\kappa_{T'_0}(a)}{2} + \binom{\kappa_{T'_0}(b)}{2} - \binom{\kappa_{T_0}(a)}{2} - \binom{\kappa_{T_0}(b)}{2} \\ &= \binom{n_1 + n_4}{2} + \binom{n_2 + n_3}{2} - \binom{n_1 + n_2}{2} - \binom{n_3 + n_4}{2} \\ &= n_1 n_4 + n_2 n_3 - n_1 n_2 - n_3 n_4 = (n_1 - n_3)(n_4 - n_2) < 0 \end{aligned}$$

which shows that $\Phi(T') < \Phi(T)$. \square

From the proof of the last lemma we deduce that if, in the tree T_0 in Fig. 2.8.(a), $|L(T_1)| \neq |L(T_3)|$ and $|L(T_2)| \neq |L(T_4)|$, and if we interchange T_2 and T_4 , then the resulting tree has always a different total cophenetic index. Recall that in these circumstances, Sackin's and Colless' indices may remain constant: cf. Fig. 2.7.

Lemma 2.14. *Let T_1, T_2 be an ordered bifurcating forest on $[m - 1]$, let $n_i = |L(T_i)|$, for $i = 1, 2$, and assume that $n_1 \geq n_2$. Let T_0 the phylogenetic tree depicted in Fig 2.9.(a), and let $T \in \mathcal{BT}_n$ be a bifurcating phylogenetic tree having T_0 as the subtree rooted at some node. If $\Phi(T)$ is minimum in \mathcal{BT}_n , then $n_1 = n_2 = 1$.*

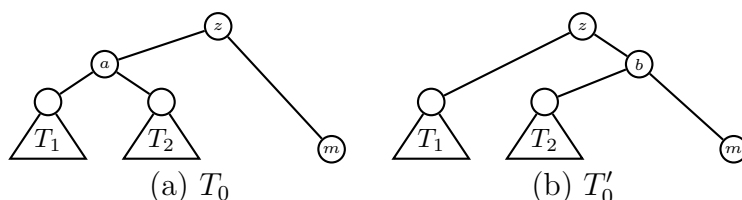


Figure 2.9: (a) The tree T_0 in the statement of Lemma 2.14. (b) The tree T'_0 in the proof of Lemma 2.14.

Proof. Assume that $n_1 > 1$ and that the tree T in the statement has the subtree T_0 rooted at a node z . Let T' be the tree obtained from T by replacing T_0 by the subtree T'_0 described in Fig. 2.9.(b). Then:

$$\begin{aligned} \Phi(T') - \Phi(T) &= \Phi(T'_0) - \Phi(T_0) \\ &= \binom{\kappa_{T'_0}(b)}{2} - \binom{\kappa_{T_0}(a)}{2} = \binom{n_2 + 1}{2} - \binom{n_1 + n_2}{2} < 0 \end{aligned}$$

which shows that $\Phi(T') < \Phi(T)$. Therefore, $\Phi(T)$ cannot be the minimum in \mathcal{BT}_n . \square

The last two lemmas imply that, unlike what happens with Sackin's and Colless' indices, any interchange of subtrees rooted at cousins that changes the balance of their grandparent always changes the total cophenetic index of a tree. This reduces the frequency of ties of Φ compared with S and C (see §2.8.1).

Theorem 2.15. *For every $T \in \mathcal{BT}_n$, $\Phi(T)$ is minimum on \mathcal{BT}_n if, and only if, T is maximally balanced.*

Proof. We shall prove that if $T \in \mathcal{BT}_n$ is not maximally balanced, then $\Phi(T)$ is not minimum. Since all maximally balanced trees in \mathcal{BT}_n have the same shape, and hence the same total cophenetic index, this will imply that the trees $T \in \mathcal{BT}_n$ with minimum $\Phi(T)$ are exactly the maximally balanced.

So, assume that $T \in \mathcal{BT}_n$ is not maximally balanced and let us prove that $\Phi(T)$ is not minimum on \mathcal{BT}_n . Let z be a non-balanced internal node in T with largest depth and assume that a and b are its children, with $\kappa_T(a) \geq \kappa_T(b) + 2$.

If b is a leaf, then $\kappa_T(a) \geq 3$ and then, by Lemma 2.14, $\Phi(T)$ is not minimum. Assume now that a and b are internal, and hence balanced by the assumption on z . Let T_0 be the subtree of T rooted at z , represented in Fig. 2.8.(a), and let $n_i = |L(T_i)|$, for $i = 1, 2, 3, 4$; without any loss of generality, we shall assume that $n_1 \geq n_2$ and $n_3 \geq n_4$ and thus, since a and b are balanced, $n_2 = n_1$ or $n_1 - 1$ and $n_4 = n_3$ or $n_3 - 1$. Then, $n_1 + n_2 = \kappa_T(a) \geq \kappa_T(b) + 2 = n_3 + n_4 + 2$ implies that $2n_1 \geq 2n_3 + 1$, and hence that $n_1 > n_3$.

Now, the facts that a and b are balanced and z is not balanced imply that $n_2 > n_4$. Indeed, if $n_4 \geq n_2$, then we would have $n_1 > n_3 \geq n_4 \geq n_2$. Since it forbids the equality $n_1 = n_2$, it would imply that $n_1 = n_2 + 1$ and therefore $n_2 = n_3 = n_4$. But then $n_1 + n_2 = 2n_2 + 1$ and $n_3 + n_4 = 2n_2$, against the assumption that z is not balanced.

So, $n_2 > n_4$. In summary, we have that $n_1 \geq n_2$, $n_3 \geq n_4$, $n_1 > n_3$ and $n_2 > n_4$, and hence, by Lemma 2.13, $\Phi(T)$ is not minimum. \square

So, the only bifurcating trees with minimum total cophenetic index are the maximally balanced trees. Let us compute now this minimum value of Φ on \mathcal{BT}_n .

Lemma 2.16. *For every n , let $f(n)$ be the minimum of Φ on \mathcal{BT}_n . Then, $f(1) = f(2) = 0$ and*

$$f(n) = f(\lceil n/2 \rceil) + f(\lfloor n/2 \rfloor) + \binom{\lceil n/2 \rceil}{2} + \binom{\lfloor n/2 \rfloor}{2}, \quad \text{for } n \geq 3.$$

Proof. The assertion when $n = 1, 2$ is obvious, so assume that $n \geq 3$. Let $T \in \mathcal{BT}_n$ be a maximally balanced tree, so that $\Phi(T) = f(n)$, and let T_1 and T_2 be the subtrees rooted at the children of its root r , with, say $|L(T_1)| \geq |L(T_2)|$. Then, on the one hand, since r is balanced, $|L(T_1)| = \lceil n/2 \rceil$ and $|L(T_2)| = \lfloor n/2 \rfloor$,

and, on the other hand, both T_1 and T_2 are again maximally balanced, and therefore $\Phi(T_1)$ and $\Phi(T_2)$ take the minimum value of Φ for their number of leaves: $\Phi(T_1) = f(\lceil n/2 \rceil)$ and $\Phi(T_2) = f(\lfloor n/2 \rfloor)$. Then, by the recurrence for $\Phi(T)$ established in Lemma 2.4,

$$\begin{aligned}\Phi(T) &= \Phi(T_1) + \Phi(T_2) + \binom{|L(T_1)|}{2} + \binom{|L(T_2)|}{2} \\ &= f(\lceil n/2 \rceil) + f(\lfloor n/2 \rfloor) + \binom{\lceil n/2 \rceil}{2} + \binom{\lfloor n/2 \rfloor}{2}\end{aligned}$$

□

Corollary 2.17. $f(n)$ is in $\Theta(n^2)$.

Proof. With the notational convention used in the statement of the Master Theorem for solving recurrences as stated in [30, Thm. 4.1], by Lemma 2.16 the sequence $f(n)$ satisfies a recurrence of the form

$$f(n) = 2f(n/2) + F(n),$$

where $F(n) = \binom{\lceil n/2 \rceil}{2} + \binom{\lfloor n/2 \rfloor}{2}$ is in $\Omega(n^2) = \Omega(n^{\log_2(2)+1})$ and it satisfies that

$$2F(n/2) \leq 2F(n).$$

Therefore, by case (3) in that theorem, $f(n)$ is in $\Theta(F(n))$, i.e., in $\Theta(n^2)$. □

Next lemma shows an alternative formula for the minimum value of Φ .

Proposition 2.18. For every $n \geq 0$, let $a(n)$ be the highest power of 2 that divides $n!$. Then, for every $n \geq 1$,

$$f(n) = \sum_{k=0}^{n-1} a(k).$$

Proof. The sequence $(a(n))_n$ is sequence A011371 in the OEIS, where we learn that it satisfies the recurrence

$$a(n) = \lfloor n/2 \rfloor + a(\lfloor n/2 \rfloor).$$

Let now $(x(n))_n$ denote the sequence of partial, cumulative sums of $(a(n))_n$, which is sequence A174605 in the OEIS:

$$x(n) = \sum_{k=0}^n a(k).$$

This sequence $(x(n))_n$ starts with $x(0) = x(1) = 0$ and it satisfies the recurrence $x(n) - x(n-1) = a(n) = \lfloor n/2 \rfloor + a(\lfloor n/2 \rfloor) = \lfloor n/2 \rfloor + x(\lfloor n/2 \rfloor) - x(\lfloor n/2 \rfloor - 1)$.

We want to prove that $f(n+1) = x(n)$, for every $n \geq 0$. Since $f(1) = f(2) = 0$, it remains to check the equality

$$f(n+1) - f(n) = \lfloor n/2 \rfloor + f(\lfloor n/2 \rfloor + 1) - f(\lfloor n/2 \rfloor), \quad \text{for } n \geq 2.$$

We prove this equality with the help of Lemma 2.16 and by distinguishing four cases, depending on the residue of $n \pmod 4$.

- If $n = 4m$, then

$$\begin{aligned} f(n+1) - f(n) &= f(2m+1) - f(2m) + \binom{2m+1}{2} - \binom{2m}{2} \\ &= f(2m+1) - f(2m) + 2m \\ &= f(\lfloor n/2 \rfloor + 1) - f(\lfloor n/2 \rfloor) + \lfloor n/2 \rfloor \end{aligned}$$

- If $n = 4m + 1$, then

$$\begin{aligned} f(n+1) - f(n) &= f(2m+1) - f(2m) + \binom{2m+1}{2} - \binom{2m}{2} \\ &= f(2m+1) - f(2m) + 2m \\ &= f(\lfloor n/2 \rfloor + 1) - f(\lfloor n/2 \rfloor) + \lfloor n/2 \rfloor \end{aligned}$$

- If $n = 4m + 2$, then

$$\begin{aligned} f(n+1) - f(n) &= f(2m+2) - f(2m+1) + \binom{2m+2}{2} - \binom{2m+1}{2} \\ &= f(2m+2) - f(2m+1) + 2m+1 \\ &= f(\lfloor n/2 \rfloor + 1) - f(\lfloor n/2 \rfloor) + \lfloor n/2 \rfloor \end{aligned}$$

- If $n = 4m + 3$, then

$$\begin{aligned}
& f(n+1) - f(n) \\
&= f(2m+2) + f(2m+2) + \binom{2m+2}{2} + \binom{2m+2}{2} \\
&\quad - \left(f(2m+2) + f(2m+1) + \binom{2m+2}{2} + \binom{2m+1}{2} \right) \\
&= f(2m+2) - f(2m+1) + \binom{2m+2}{2} - \binom{2m+1}{2} \\
&= f(2m+2) - f(2m+1) + 2m+1 \\
&= f(\lfloor n/2 \rfloor + 1) - f(\lfloor n/2 \rfloor) + \lfloor n/2 \rfloor
\end{aligned}$$

This completes the proof. \square

In particular, this provides a new meaning and a new recurrence for sequence A174605 in the OEIS.

2.3 Expected value of Φ under the Yule model

Let Φ_n be the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes its total cophenetic index $\Phi(T)$. In this section we determine the expected value of Φ_n under the Yule model. To do this, we shall make use of the following lemma, which can be useful to study the expected value under the Yule model of other recursive shape indices for bifurcating phylogenetic trees.

Lemma 2.19. *Let I be recursive shape index for bifurcating phylogenetic trees, and in particular let $f_I : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$ be the symmetric mapping such that, for every pair of phylogenetic trees T, T' on disjoint sets of taxa S, S' , respectively,*

$$I(T \star T') = I(T) + I(T') + f_I(|S|, |S'|).$$

Let I_n be the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes $I(T)$, and let $E_Y(I_n)$ be its expected value under the Yule model. Then, for every $n \geq 2$,

$$E_Y(I_n) = \frac{1}{n-1} \left(2 \sum_{k=1}^{n-1} E_Y(I_k) + \sum_{k=1}^{n-1} f_I(k, n-k) \right).$$

Proof. First of all, notice that if $T_k \in \mathcal{BT}(S_k)$, for some $S_k \subsetneq [n]$ with $|S_k| = k$, and $T'_{n-k} \in \mathcal{BT}_{[n] \setminus S_k}$, then

$$P_Y(T_k \star T'_{n-k}) = \frac{2}{(n-1) \binom{n}{k}} P_Y(T_k) P_Y(T'_{n-k}) \quad (2.1)$$

where P_Y denotes the probability of a phylogenetic tree under the Yule model. This assertion is a direct consequence of the explicit probabilities of T_k, T'_{n-k}

and $T_k \star T'_{n-k}$ under the Yule model given by formula (1.4) in Section 1.3, and the fact that $V_{int}(T_k \star T'_{n-k}) = V_{int}(T_k) \cup V_{int}(T'_{n-k}) \cup \{r\}$ (where r denotes the root of $T_k \star T'_{n-k}$), these unions being disjoint. Indeed, by the aforementioned formula

$$\begin{aligned}
P_Y(T_k \star T'_{n-k}) &= \frac{2^{n-1}}{n!} \prod_{v \in V_{int}(T_k \star T'_{n-k})} \frac{1}{\kappa_T(v) - 1} \\
&= \frac{2^{n-1}}{n!} \cdot \frac{1}{\kappa_T(r) - 1} \prod_{v \in V_{int}(T_k)} \frac{1}{\kappa_T(v) - 1} \prod_{v \in V_{int}(T'_{n-k})} \frac{1}{\kappa_T(v) - 1} \\
&= \frac{2^{n-1}}{n!} \cdot \frac{1}{n-1} \cdot \frac{P_Y(T_k)k!}{2^{k-1}} \cdot \frac{P_Y(T'_{n-k})(n-k)!}{2^{n-k-1}} \\
&= \frac{2}{(n-1) \binom{n}{k}} P_Y(T_k) P_Y(T'_{n-k}).
\end{aligned}$$

Let us compute now $E_Y(I_n)$, for $n \geq 2$, using its very definition. The starting point in this computation is the fact that every $T \in \mathcal{BT}_n$ can be obtained by choosing a number k of leaves between 1 and $n-1$, a subset S_k of k labels of $[n]$, a bifurcating tree T_k on S_k and a bifurcating tree T'_{n-k} on $S_k^c = [n] \setminus S_k$, and then taking their root join $T_k \star T'_{n-k}$. Actually, every $T \in \mathcal{BT}_n$ is obtained twice in this way, depending on whether the result of our first choice of set of labels turns out to be S_k or S_k^c .

$$\begin{aligned}
E_Y(I_n) &= \sum_{T \in \mathcal{BT}_n} I(T) \cdot P_Y(T) \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subseteq [n] \\ |S_k|=k}} \sum_{T_k \in \mathcal{BT}(S_k)} \sum_{T'_{n-k} \in \mathcal{BT}(S_k^c)} I(T_k \star T'_{n-k}) \cdot P_Y(T_k \star T'_{n-k}) \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathcal{BT}_{\{1, \dots, k\}}} \sum_{T'_{n-k} \in \mathcal{BT}_{\{k+1, \dots, n\}}} I(T_k \star T'_{n-k}) \cdot P_Y(T_k \star T'_{n-k}) \\
&\text{(by the invariance under leaf relabelings)} \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k}) \\
&\quad + f_I(k, n-k)) \cdot \frac{2}{(n-1) \binom{n}{k}} P_Y(T_k) P_Y(T'_{n-k})
\end{aligned}$$

(by condition (b) in the statement, formula (2.1) and, again, the invariance under leaf relabelings)

$$\begin{aligned}
&= \frac{1}{n-1} \sum_{k=1}^{n-1} \sum_{T_k} \sum_{T'_{n-k}} (I(T_k) + I(T'_{n-k}) + f_I(k, n-k)) P_Y(T_k) P_Y(T'_{n-k}) \\
&= \frac{1}{n-1} \sum_{k=1}^{n-1} \left(\sum_{T_k} \sum_{T'_{n-k}} I(T_k) P_Y(T_k) P_Y(T'_{n-k}) \right)
\end{aligned}$$

$$\begin{aligned}
& + \sum_{T_k} \sum_{T'_{n-k}} I(T'_{n-k}) P_Y(T_k) P_Y(T'_{n-k}) \\
& + \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k) P_Y(T_k) P_Y(T'_{n-k}) \\
= & \frac{1}{n-1} \sum_{k=1}^{n-1} \left(\left(\sum_{T_k} I(T_k) P_Y(T_k) \right) \left(\sum_{T'_{n-k}} P_Y(T'_{n-k}) \right) \right. \\
& + \left(\sum_{T_k} P_Y(T_k) \right) \left(\sum_{T'_{n-k}} I(T'_{n-k}) P_Y(T'_{n-k}) \right) \\
& \left. + f_I(k, n-k) \left(\sum_{T_k} P_Y(T_k) \right) \left(\sum_{T'_{n-k}} P_Y(T'_{n-k}) \right) \right) \\
= & \frac{1}{n-1} \sum_{k=1}^{n-1} \left(\sum_{T_k} I(T_k) P_Y(T_k) + \sum_{T'_{n-k}} I(T'_{n-k}) P_Y(T'_{n-k}) + f_I(k, n-k) \right) \\
= & \frac{1}{n-1} \sum_{k=1}^{n-1} (E_Y(I_k) + E_Y(I_{n-k}) + f_I(k, n-k)) \\
= & \frac{1}{n-1} \left(2 \sum_{k=1}^{n-1} E_Y(I_k) + \sum_{k=1}^{n-1} f_I(k, n-k) \right)
\end{aligned}$$

as we claimed. \square

Theorem 2.20. *Under the Yule model, the expected value of Φ_n is*

$$E_Y(\Phi_n) = n(n+1) - 2nH_n,$$

where H_n denotes the n -th harmonic number, $H_n = \sum_{i=1}^n 1/i$.

Proof. If $n = 1$, the equality in the statement holds because its two sides are equal to 0 (notice that the only tree in \mathcal{BT}_1 has total cophenetic index 0). Let us consider now the case $n \geq 2$. Lemma 2.4 implies that Φ satisfies the hypothesis of Lemma 2.19 with $f_\Phi(k, n-k) = \binom{k}{2} + \binom{n-k}{2}$. Then

$$\sum_{k=1}^{n-1} f_\Phi(k, n-k) = \sum_{k=1}^{n-1} \left(\binom{k}{2} + \binom{n-k}{2} \right) = 2 \sum_{k=1}^{n-1} \binom{k}{2} = 2 \binom{n}{3},$$

and hence, for every $n \geq 2$,

$$\begin{aligned}
E_Y(\Phi_n) &= \frac{2}{n-1} \sum_{k=1}^{n-1} E_Y(\Phi_k) + \frac{2}{n-1} \binom{n}{3} \\
&= \frac{2}{n-1} E_Y(\Phi_{n-1}) + \frac{n-2}{n-1} \cdot \frac{2}{n-2} \sum_{k=1}^{n-2} E_Y(\Phi_k) + \frac{2}{n-1} \binom{n}{3}
\end{aligned}$$

$$\begin{aligned}
&= \frac{2}{n-1} E_Y(\Phi_{n-1}) + \frac{n-2}{n-1} \left(E_Y(\Phi_{n-1}) - \frac{2}{n-2} \binom{n-1}{3} \right) + \frac{2}{n-1} \binom{n}{3} \\
&= \frac{n}{n-1} E_Y(\Phi_{n-1}) + \frac{2}{n-1} \left(\binom{n}{3} - \binom{n-1}{3} \right) \\
&= \frac{n}{n-1} E_Y(\Phi_{n-1}) + n - 2.
\end{aligned}$$

To solve this equation, divide both sides of it by n :

$$\frac{1}{n} E_Y(\Phi_n) = \frac{1}{n-1} E_Y(\Phi_{n-1}) + \frac{n-2}{n}.$$

Setting $x_n = E_Y(\Phi_n)/n$, the sequence $(x_n)_n$ satisfies

$$x_n = x_{n-1} + \frac{n-2}{n}, \text{ starting with } x_1 = 0.$$

Therefore

$$x_n = \sum_{i=2}^n \frac{i-2}{i} = (n-1) - 2 \sum_{i=2}^n \frac{1}{i} = (n-1) - 2 \left(\sum_{i=1}^n \frac{1}{i} - 1 \right) = n+1 - 2H_n$$

and thus, finally,

$$E_Y(\Phi_n) = nx_n = n(n+1 - 2H_n),$$

as we claimed. \square

Using that $H_n = \ln(n) + \gamma + 1/(2n) + O(1/n^2)$ (see, for instance, [50, p. 264]), where $\gamma = 0.577215\dots$ is the Euler-Mascheroni constant, we obtain the following result:

Corollary 2.21. $E_Y(\Phi_n) = n^2 - 2n \ln(n) + (1 - 2\gamma)n - 1 + O(1/n)$.

So, the order $O(n^2)$ of the expected value under the Yule model of the total cophenetic index on \mathcal{BT}_n is larger than the order $O(n \log(n))$ of the expected values of Sackin's and Colless' indices; see formulas (1.3) in Section 1.3.

Let S_n stand for the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes its Sackin index $S(T)$. Notice that, since $E_Y(S_n) = 2n(H_n - 1)$ by (1.3) in Section 1.3, we have that

$$E_Y(\Phi_n) + E_Y(S_n) = n(n-1).$$

From the expected values of the Sackin and the total cophenetic indices, we can deduce the expected value of the total area D on \mathcal{BT}_n under the Yule model.

Corollary 2.22. *Let D_n be the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes its total area $D(T)$. Under the Yule model, its expected value is*

$$E_Y(D_n) = 2n(n+1)H_n - 4n^2.$$

Proof. From Lemma 2.6 we deduce that

$$2\Phi_n + D_n = (n - 1)S_n,$$

and therefore

$$\begin{aligned} E_Y(D_n) &= (n - 1)E_Y(S_n) - 2E_Y(\Phi_n) \\ &= (n - 1)2n(H_n - 1) - 2(n(n + 1) - 2nH_n) \\ &= 2n(n + 1)H_n - 4n^2. \end{aligned}$$

□

Remark 2.23. In equation (35) in [77, p. 143] it is claimed that

$$E_Y(D_n) = 2n(n + 1)(H_n - 1) - \frac{5}{2}n(n - 1),$$

which cannot be correct: since all three trees $T \in \mathcal{BT}_3$ have $D(T) = 8$, it must happen that $E_Y(D_3) = 8$, while the expression given in loc. cit. yields $E_Y(D_3) = 5$. And incidentally, our formula does yield the correct value in this case.

To double-check the formula given in Theorem 2.20, we have computed the values of $E_Y(\Phi_n)$, for $n = 3, \dots, 8$, from the cophenetic indices and the probabilities under the Yule model of all trees in the corresponding \mathcal{BT}_n , and they agree with the figures given by our formula. The R script used to compute these “real” values and to compare them with the values obtained through our formula is available in Appendix A.4.1 and on the GitHub repository associated to this PhD Thesis [97].

2.4 Expected value of Φ under the uniform model

In this section we determine the expected value of Φ_n under the uniform model. This expected value of Φ_n will be easily deduced, through Lemma 2.6, from the expected value of the total area, which was obtained in [72], and the expected value of the Sackin index, which was previously unknown and we obtain in Theorem 2.27 below.

Under the uniform model, all trees in \mathcal{BT}_n have the same probability: namely, $1/(2n - 3)!!$. Therefore, the expected value of S_n under the uniform model is

$$E_U(S_n) = \frac{\sum_{T \in \mathcal{BT}_n} S(T)}{(2n - 3)!!}.$$

So, we need to compute the numerator in this fraction.

Lemma 2.24. *For every $n \geq 2$,*

$$\sum_{T \in \mathcal{BT}_n} S(T) = n \sum_{k=1}^{n-1} \frac{(2n-k-3)!k^2}{(n-k-1)!2^{n-k-1}}.$$

Proof. For every $k = 1, \dots, n-1$, set

$$c_{k,n} = \left| \{T \in \mathcal{BT}_n \mid \delta_T(1) = k\} \right|.$$

Notice that, for every $1 \leq i \leq n$, we also have

$$c_{k,n} = \left| \{T \in \mathcal{BT}_n \mid \delta_T(i) = k\} \right|.$$

Then

$$\begin{aligned} \sum_{T \in \mathcal{BT}_n} S(T) &= \sum_{T \in \mathcal{BT}_n} \sum_{i=1}^n \delta_T(i) = \sum_{i=1}^n \sum_{T \in \mathcal{BT}_n} \delta_T(i) \\ &= \sum_{i=1}^n \sum_{k=1}^{n-1} k \cdot \left| \{T \in \mathcal{BT}_n \mid \delta_T(i) = k\} \right| \\ &= \sum_{i=1}^n \sum_{k=1}^{n-1} k \cdot c_{k,n} = n \sum_{k=1}^{n-1} k \cdot c_{k,n}. \end{aligned}$$

It remains to compute $c_{k,n}$ for $k \geq 1$. To do so, notice that every tree $T \in \mathcal{BT}_n$ such that $\delta(1) = k$ will have the form described in Fig. 2.10. Therefore, it is determined by the ordered k -forest T_1, T_2, \dots, T_k on $\{2, \dots, n\}$, and thus, using the formula (1.1) in Section 1.1,

$$c_{k,n} = |\mathcal{BF}_{k,n-1}| = \frac{(2n-k-3)!k}{(n-k-1)!2^{n-k-1}},$$

from which the expression in the statement follows. □

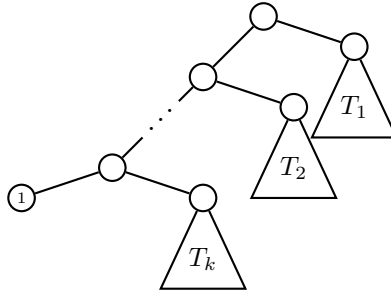


Figure 2.10: The structure of a tree T with $\delta_T(1) = k$.

Now, to compute $E_U(S_n)$ we shall make use of the following two technical lemmas, which shall also be used in the next chapter.

Lemma 2.25. $\sum_{i=0}^{n-2} \frac{(n+i-2)!}{i!2^i} = (n-2)!2^{n-2}.$

Proof. Following the notations that we shall introduce in the next lemma, set

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

Let us compute its value:

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

These two sums can be computed using the lookup algorithm. Let us start with

$$\sum_{i=0}^{\infty} \frac{(n+i-2)!}{i!} \left(\frac{1}{2}\right)^i.$$

Set

$$t_i = \frac{(n+i-2)!}{i!}.$$

Then $t_0 = (n-2)!$ and

$$\frac{t_{i+1}}{t_i} = \frac{i+n-1}{i+1}$$

and therefore, by the lookup algorithm,

$$\sum_{i=0}^{\infty} \frac{(n+i-2)!}{i!} \left(\frac{1}{2}\right)^i = (n-2)! {}_1F_0 \left[\begin{matrix} n-1 \\ - \end{matrix} ; \frac{1}{2} \right].$$

As to

$$\sum_{i=0}^{\infty} \frac{(2n+i-3)!}{(i+n-1)!} \left(\frac{1}{2}\right)^{n-1+i} = \sum_{i=0}^{\infty} \frac{(2n+i-3)!}{(i+n-1)!2^{n-1}} \left(\frac{1}{2}\right)^i,$$

set now

$$t_i = \frac{(2n+i-3)!}{(i+n-1)!2^{n-1}}.$$

Then

$$t_0 = \frac{(2n-3)!}{(n-1)!2^{n-1}}$$

and

$$\frac{t_{i+1}}{t_i} = \frac{i+2n-2}{i+n} = \frac{(i+2n-2)(i+1)}{(i+n)(i+1)}$$

and therefore, by the lookup algorithm,

$$\sum_{i=0}^{\infty} \frac{(2n+i-3)!}{(i+n-1)!} \left(\frac{1}{2}\right)^{n-1+i} = \frac{(2n-3)!}{(n-1)!2^{n-1}} {}_2F_1 \left[\begin{matrix} 2n-2, & 1 \\ n & \end{matrix} ; \frac{1}{2} \right].$$

So, in summary,

$$U_{n,0} = (n-2)! {}_1F_0 \left[\begin{matrix} n-1 \\ - \end{matrix} ; \frac{1}{2} \right] - \frac{(2n-3)!}{(n-1)!2^{n-1}} {}_2F_1 \left[\begin{matrix} 2n-2, & 1 \\ n & \end{matrix} ; \frac{1}{2} \right].$$

Now, by Formula (1.16),

$${}_1F_0 \left[\begin{matrix} n-1 \\ - \end{matrix} ; \frac{1}{2} \right] = \left(1 - \frac{1}{2}\right)^{-(n-1)} = 2^{n-1}$$

and, by Formulas (1.17) and (1.13) and using that $\Gamma(1) = 1$,

$${}_2F_1 \left[\begin{matrix} 2n-2, & 1 \\ n & \end{matrix} ; \frac{1}{2} \right] = \frac{\sqrt{\pi}\Gamma(n)}{\Gamma(n-1/2)\Gamma(1)} = \frac{(n-1)!2^{n-1}}{(2n-3)!!}$$

and therefore

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

as we claimed. □

Lemma 2.26. For every $m \geq 0$, let

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

Then, for every $m \geq 1$

$$\begin{aligned} U_{n,m} &= (n-1)! \cdot 2^{n-2} + \sum_{j=1}^{m-1} \left[(n-1) \binom{m-1}{j} + \binom{m-1}{j-1} \right] U_{n,j} \\ &\quad - (n-1)^{m-1} (2n-3)!!. \end{aligned}$$

Proof. We start by developing $U_{n,m}$ using that $m \geq 1$:

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

$$\begin{aligned}
&= \sum_{i=0}^{n-2} \frac{(i+1)^{m-1}(n+i-1)!}{i! \cdot 2^{i+1}} - \frac{(n-1)^{m-1}(2n-3)!}{(n-2)!2^{n-1}} \\
&= \sum_{i=0}^{n-2} \frac{n+i-1}{2} \cdot \frac{(i+1)^{m-1}(n+i-2)!2^{-i}}{i!} - \frac{(n-1)^{m-1}(2n-3)!2^{-n+1}}{(n-2)!} \\
&= \sum_{i=0}^{n-2} \left[\frac{n+i-1}{2} \sum_{j=0}^{m-1} \binom{m-1}{j} \frac{i^j(n+i-2)!}{i! \cdot 2^i} \right] - \frac{(n-1)^{m-1}(2n-3)!!}{2} \\
&= \frac{n-1}{2} \sum_{i=0}^{n-2} \sum_{j=0}^{m-1} \binom{m-1}{j} \frac{i^j(n+i-2)!}{i! \cdot 2^i} \\
&\quad + \frac{1}{2} \sum_{i=0}^{n-2} \sum_{j=0}^{m-1} \binom{m-1}{j} \frac{i^{j+1}(n+i-2)!}{i! \cdot 2^i} - \frac{(n-1)^{m-1}(2n-3)!!}{2} \\
&= \frac{n-1}{2} \sum_{j=0}^{m-1} \binom{m-1}{j} \sum_{i=0}^{n-2} \frac{i^j(n+i-2)!}{i! \cdot 2^i} \\
&\quad + \frac{1}{2} \sum_{j=1}^m \binom{m-1}{j-1} \sum_{i=0}^{n-2} \frac{i^j(n+i-2)!}{i! \cdot 2^i} - \frac{(n-1)^{m-1}(2n-3)!!}{2} \\
&= \frac{n-1}{2} \sum_{j=0}^{m-1} \binom{m-1}{j} U_{n,j} + \frac{1}{2} \sum_{j=1}^m \binom{m-1}{j-1} U_{n,j} - \frac{(n-1)^{m-1}(2n-3)!!}{2} \\
&= \frac{n-1}{2} \cdot U_{n,0} + \frac{1}{2} \sum_{j=1}^{m-1} \left[(n-1) \binom{m-1}{j} + \binom{m-1}{j-1} \right] U_{n,j} + \frac{1}{2} U_{n,m} \\
&\quad - \frac{(n-1)^{m-1}(2n-3)!!}{2}
\end{aligned}$$

Isolating $U_{n,m}$, we obtain

$$U_{n,m} = (n-1)U_{n,0} + \sum_{j=1}^{m-1} \left[(n-1) \binom{m-1}{j} + \binom{m-1}{j-1} \right] U_{n,j} - (n-1)^{m-1}(2n-3)!!$$

and using $U_{n,0} = (n-2)! \cdot 2^{n-2}$, as we proved in the last lemma, we finally obtain the expression in the statement. \square

Theorem 2.27. *The expected value of the random variable S_n under the uniform model is*

$$E_U(S_n) = n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right).$$

Proof. When $n = 1$, $E_U(S_1) = 0$, because the only tree in \mathcal{BT}_1 has Sackin index 0, and $1 \cdot (0!!/(-1)!! - 1) = 0$, because, by definition, $0!! = (-1)!! = 1$. Therefore, the equality in the statement is satisfied in this case. Let us consider

now the case $n \geq 2$. In this case, by the last lemma, we have that

$$\begin{aligned}
E_U(S_n) &= \frac{\sum_{T \in \mathcal{BT}_n} S(T)}{(2n-3)!!} = \frac{n}{(2n-3)!!} \sum_{k=1}^{n-1} \frac{(2n-k-3)!k^2}{(n-k-1)!2^{n-k-1}} \\
&= \frac{n}{(2n-3)!!} \sum_{j=1}^{n-1} \frac{(n+j-3)!(n-j)^2}{(j-1)!2^{j-1}} \\
&= \frac{n}{(2n-3)!!} \sum_{i=0}^{n-2} \frac{(n+i-2)!(n-i-1)^2}{i!2^i} \\
&= \frac{n}{(2n-3)!!} \sum_{i=0}^{n-2} \frac{((n-1)^2 - 2(n-1)i + i^2)(n+i-2)!2^{-i}}{i!} \\
&= \frac{n}{(2n-3)!!} ((n-1)^2 U_{n,0} - 2(n-1)U_{n,1} + U_{n,2})
\end{aligned}$$

where, with the notations of the last lemma,

$$U_{n,m} = \sum_{i=0}^{n-2} \frac{i^m (n+i-2)!2^{-i}}{i!}, \quad m = 0, 1, 2.$$

Let us compute these values. We already know $U_{n,0}$ from Lemma 2.25:

$$U_{n,0} = (n-2)!2^{n-2}.$$

We compute now the values of $U_{n,1}$ and $U_{n,2}$ using Lemma 2.26:

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

Returning back to the computation of $E_U(S_n)$, we have that

$$\begin{aligned}
E_U(S_n) &= \frac{n}{(2n-3)!!} ((n-1)^2 U_{n,0} - 2(n-1)U_{n,1} + U_{n,2}) \\
&= \frac{n}{(2n-3)!!} \left((n-1)^2 2^{n-2} (n-2)! - 2(n-1)((n-1)!2^{n-2} - (2n-3)!!) \right. \\
&\quad \left. + 2^{n-2}(n+1)(n-1)! - (2n-1)!! \right) \\
&= \frac{n}{(2n-3)!!} (2^{n-1}(n-1)! - (2n-3)!!) = n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right),
\end{aligned}$$

as we claimed. \square

In particular, we can deduce from this exact expression for $E_U(S_n)$ the limit formula recalled in (1.4) in Section 1.3

$$E_U(S_n) \sim \sqrt{\pi n}^{3/2}.$$

Indeed, using Stirling's approximation for large factorials we have that

$$\begin{aligned} E_U(S_n) &= n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right) = n \left(\frac{2^{n-1}(n-1)!2^{n-1}(n-1)!}{(2n-2)!} - 1 \right) \\ &\sim n \left(\frac{(2^{n-1}\sqrt{2\pi n}(n-1)^{n-1})^2}{\sqrt{2\pi 2n}(2n-2)^{2n-2}} - 1 \right) = n \left(\frac{2\pi n(2n-2)^{2n-2}}{2\sqrt{\pi n}(2n-2)^{2n-2}} - 1 \right) \\ &= n(\sqrt{\pi n} - 1) \sim \sqrt{\pi n}^{3/2} \end{aligned}$$

We have now the following result.

Theorem 2.28. *Under the uniform model, the expected value of Φ_n is*

$$E_U(\Phi_n) = \frac{1}{2} \binom{n}{2} \left(\frac{(2n-2)!!}{(2n-3)!!} - 2 \right).$$

Proof. The expected values under the uniform model of S_n and D_n are:

$$\begin{aligned} E_U(S_n) &= n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right) && \text{by Theorem 2.27} \\ E_U(D_n) &= \binom{n}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} && [72] \end{aligned}$$

Then, by Lemma 2.6,

$$\begin{aligned} E_U(\Phi_n) &= \frac{n-1}{2} E_U(S_n) - \frac{1}{2} E_U(D_n) \\ &= \frac{n-1}{2} \cdot n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right) - \frac{1}{2} \binom{n}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} \\ &= \frac{1}{2} \binom{n}{2} \left(\frac{(2n-2)!!}{(2n-3)!!} - 2 \right). \end{aligned}$$

□

Corollary 2.29. $E_U(\Phi_n) \sim \frac{\sqrt{\pi}}{4} n^{5/2}$.

Proof. Notice that

$$\begin{aligned} E_U(\Phi_n) &= \frac{n(n-1)}{4} \left(\frac{(2n-2)!!}{(2n-3)!!} - 2 \right) = \frac{n(n-1)}{4} \left(\frac{1}{n} E_U(S_n) - 1 \right) \\ &\sim \frac{n(n-1)}{4} (\sqrt{\pi n}^{1/2} - 1) \sim \frac{\sqrt{\pi}}{4} n^{5/2}. \end{aligned}$$

□

Again, to double-check the formula given in Theorem 2.28, we have computed the values of $E_U(\Phi_n)$, for $n = 3, \dots, 8$, from the cophenetic indices of all trees in the corresponding \mathcal{BT}_n , and they agree with the figures given by our formula. The R script used to compute the “real” values and to compare them with the values obtained through our formula is available in Appendix A.4.2 and on the GitHub repository [97].

2.5 On the variance of Φ under the uniform model

In the last sections we have obtained formulas for the expected value of the random variable Φ_n under the uniform and the Yule models. As far as its variance goes, an explicit formula for it under the Yule model was given in [23, Cor. 3]:

$$\sigma_Y^2(\Phi_n) = \frac{1}{12}(n^4 - 10n^3 + 131n^2 - 2n) - 4n^2 H_n^{(2)} - 6nH_n, \quad (2.2)$$

where $H_n^{(2)} = \sum_{i=1}^n 1/i^2$. This implies that $\sigma_Y^2(\Phi_n)$ grows in $O(n^4)$ [23, Cor. 4].

In this section we are interested in its variance $\sigma_U^2(\Phi_n)$ under the uniform model. Using that

$$\sigma_U^2(\Phi_n) = E_U(\Phi_n^2) - E_U(\Phi_n)^2,$$

where Φ_n^2 is the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes $\Phi(T)^2$, to obtain $\sigma_U^2(\Phi_n)$ it remains to compute $E_U(\Phi_n^2)$. Unfortunately, we have not been able to find a closed formula for this expected value, but in this section we shall obtain a recurrence that allows its computation for every n . The basis of this recurrence will be the following lemma, which provides expressions for the expected value under the uniform model of a recursive shape index for bifurcating phylogenetic trees I , as well as of its square I^2 , similar in spirit to the expression for the expected value of I under the Yule model given in Lemma 2.19.

Lemma 2.30. *Let I a recursive shape index for bifurcating phylogenetic trees as in Lemma 2.19. Let I_n and I_n^2 be the random variables that choose a tree $T \in \mathcal{T}_n$ and compute $I(T)$ and $I(T)^2$, respectively. To simplify the notations, for every $1 \leq k \leq n-1$, set*

$$C_{k,n-k} = \frac{1}{2} \binom{n}{k} \frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}.$$

Then, for every $n \geq 2$,

$$\begin{aligned} E_U(I_n) &= \sum_{k=1}^{n-1} C_{k,n-k} (2E_U(I_k) + f_I(k, n-k)) \\ E_U(I_n^2) &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(I_k^2) + f_I(k, n-k)^2 \right. \\ &\quad \left. + 4f_I(k, n-k)E_U(I_k) + 2E_U(I_k)E_U(I_{n-k}) \right) \end{aligned}$$

Proof. Since, for every $m \geq 1$, the probability under the uniform model of a bifurcating tree with m leaves is $1/(2m-3)!!$, we have that, for every

$T_k \in \mathcal{BT}(S_k)$, with $S_k \subsetneq [n]$ of cardinality k , and for every $T'_{n-k} \in \mathcal{BT}_{[n] \setminus S_k}$,

$$P_U(T_k \star T'_{n-k}) = \frac{1}{(2n-3)!!}, \quad P_U(T_k)P_U(T'_{n-k}) = \frac{1}{(2k-3)!!(2(n-k)-3)!!}$$

and hence, with the notation introduced in the statement,

$$\begin{aligned} P_U(T_k \star T'_{n-k}) &= \frac{(2k-3)!! \cdot (2(n-k)-3)!!}{(2n-3)!!} P_U(T_k)P_U(T'_{n-k}) \\ &= \frac{2C_{k,n-k}}{\binom{n}{k}} P_U(T_k)P_U(T'_{n-k}). \end{aligned}$$

Notice by the way that, since $\binom{n}{k} = \binom{n}{n-k}$, the coefficient $C_{k,n-k}$ is symmetric, that is, $C_{k,n-k} = C_{n-k,n}$.

Then, if we develop $E_U(I_n)$, for $n \geq 2$, as we did with $E_Y(I_n)$ in the proof of Lemma 2.19, replacing the probabilities under the Yule model by the probabilities under the uniform model, we obtain

$$\begin{aligned} E_U(I_n) &= \sum_{T \in \mathcal{BT}_n} I(T) \cdot P_U(T) \\ &= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subsetneq [n] \\ |S_k|=k}} \sum_{T_k \in \mathcal{BT}(S_k)} \sum_{T'_{n-k} \in \mathcal{BT}(S_k^c)} I(T_k \star T'_{n-k}) \cdot P_U(T_k \star T'_{n-k}) \\ &= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k}) \\ &\quad + f_I(k, n-k)) \cdot \frac{2C_{k,n-k}}{\binom{n}{k}} P_U(T_k)P_U(T'_{n-k}) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k}) \\ &\quad + f_I(k, n-k)) P_U(T_k)P_U(T'_{n-k}) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} (E_U(I_k) + E_U(I_{n-k}) + f_I(k, n-k)) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} (2E_U(I_k) + f_I(k, n-k)) \end{aligned}$$

where in the last step we have used the symmetry of $C_{k,n-k}$.

As far as $E_U(I_n^2)$ goes, we can develop it in a similar way (cf. the proof of [23, Lem. 2]) for $n \geq 2$, as follows:

$$\begin{aligned} E_U(I_n^2) &= \sum_{T \in \mathcal{BT}_n} I(T)^2 \cdot P_U(T) \\ &= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subsetneq [n] \\ |S_k|=k}} \sum_{T_k \in \mathcal{BT}(S_k)} \sum_{T'_{n-k} \in \mathcal{BT}(S_k^c)} I(T_k \star T'_{n-k})^2 \cdot P_U(T_k \star T'_{n-k}) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k})) \\
&\quad + f_I(k, n-k)^2 \cdot \frac{2C_{k,n-k}}{\binom{n}{k}} P_U(T_k) P_U(T'_{n-k}) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k})) \\
&\quad + f_I(k, n-k)^2 P_U(T_k) P_U(T'_{n-k}) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \sum_{T_k} \sum_{T'_{n-k}} (I(T_k)^2 + I(T'_{n-k})^2 + f_I(k, n-k)^2 \\
&\quad + 2I(T_k)I(T'_{n-k}) + 2f_I(k, n-k)I(T_k) \\
&\quad + 2f_I(k, n-k)I(T'_{n-k})) P_U(T_k) P_U(T'_{n-k}) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \left(\sum_{T_k} \sum_{T'_{n-k}} I(T_k)^2 P_U(T_k) P_U(T'_{n-k}) \right. \\
&\quad + \sum_{T_k} \sum_{T'_{n-k}} I(T'_{n-k})^2 P_U(T_k) P_U(T'_{n-k}) \\
&\quad + \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k)^2 P_U(T_k) P_U(T'_{n-k}) \\
&\quad + 2 \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k) I(T_k) P_U(T_k) P_U(T'_{n-k}) \\
&\quad + 2 \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k) I(T'_{n-k}) P_U(T_k) P_U(T'_{n-k}) \\
&\quad \left. + 2 \sum_{T_k} \sum_{T'_{n-k}} I(T_k) I(T'_{n-k}) P_U(T_k) P_U(T'_{n-k}) \right) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \left(\sum_{T_k} I(T_k)^2 P_U(T_k) + \sum_{T'_{n-k}} I(T'_{n-k})^2 P_U(T'_{n-k}) \right. \\
&\quad + f_I(k, n-k)^2 + 2f_I(k, n-k) \sum_{T_k} I(T_k) P_U(T_k) \\
&\quad + 2f_I(k, n-k) \sum_{T'_{n-k}} I(T'_{n-k}) P_U(T'_{n-k}) \\
&\quad \left. + 2 \left(\sum_{T_k} I(T_k) P_U(T_k) \right) \left(\sum_{T'_{n-k}} I(T'_{n-k}) P_U(T'_{n-k}) \right) \right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{n-1} C_{k,n-k} \left(E_U(I_k^2) + E_U(I_{n-k}^2) + f_I(k, n-k)^2 + 2f_I(k, n-k)E_U(I_k) \right. \\
&\quad \left. + 2f_I(k, n-k)E_U(I_{n-k}) + 2E_U(I_k)E_U(I_{n-k}) \right) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(I_k^2) + f_I(k, n-k)^2 + 4f_I(k, n-k)E_U(I_k) \right. \\
&\quad \left. + 2E_U(I_k)E_U(I_{n-k}) \right)
\end{aligned}$$

where in the last step we have used the symmetry of $C_{k,n-k}$ and $f_I(k, n-k)$. \square

The announced recurrence for $E_U(\Phi_n^2)$ is given by the following proposition:

Proposition 2.31. *For every $n \geq 2$,*

$$\begin{aligned}
E_U(\Phi_n^2) &= 2 \sum_{k=1}^{n-1} C_{k,n-k} E_U(\Phi_k^2) \\
&\quad + \frac{1}{6} \binom{n}{2} \left(\frac{49n^3 - 57n^2 - 22n + 24}{8} \cdot \frac{(2n-4)!!}{(2n-3)!!} - \frac{63n^2 - 95n + 28}{5} \right).
\end{aligned}$$

Proof. If we apply Lemma 2.30 taking as I the total cophenetic value Φ , we have

$$\begin{aligned}
E_U(\Phi_n^2) &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(\Phi_k^2) + f_\Phi(k, n-k)^2 + 4f_\Phi(k, n-k)E_U(\Phi_k) \right. \\
&\quad \left. + 2E_U(\Phi_k)E_U(\Phi_{n-k}) \right)
\end{aligned}$$

where

$$f_\Phi(k, n-k) = \binom{k}{2} + \binom{n-k}{2}, \quad E_U(\Phi_k) = \frac{1}{2} \binom{k}{2} \left(\frac{(2k-2)!!}{(2k-3)!!} - 2 \right).$$

Let us simplify this recurrence. To begin with, we have that

$$\begin{aligned}
&f_\Phi(k, n-k)^2 + 4f_\Phi(k, n-k)E_U(\Phi_k) + 2E_U(\Phi_k)E_U(\Phi_{n-k}) \\
&= \left(\binom{k}{2} + \binom{n-k}{2} \right)^2 + 2 \left(\binom{k}{2} + \binom{n-k}{2} \right) \binom{k}{2} \left(\frac{(2k-2)!!}{(2k-3)!!} - 2 \right) \\
&\quad + \frac{1}{2} \binom{k}{2} \binom{n-k}{2} \left(\frac{(2k-2)!!}{(2k-3)!!} - 2 \right) \left(\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} - 2 \right) \\
&= \binom{k}{2}^2 + \binom{n-k}{2}^2 + 2 \binom{k}{2} \binom{n-k}{2} + 2 \binom{k}{2}^2 \frac{(2k-2)!!}{(2k-3)!!} \\
&\quad + 2 \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} - 4 \binom{k}{2}^2 - 4 \binom{k}{2} \binom{n-k}{2}
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{2} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} + 2 \binom{k}{2} \binom{n-k}{2} \\
& - \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} - \binom{k}{2} \binom{n-k}{2} \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \\
& = \binom{n-k}{2}^2 - 3 \binom{k}{2}^2 + 2 \binom{k}{2}^2 \frac{(2k-2)!!}{(2k-3)!!} \\
& + \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} - \binom{k}{2} \binom{n-k}{2} \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \\
& + \frac{1}{2} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!}
\end{aligned}$$

and then, using again that $C_{k,n-k} = C_{n-k,k}$,

$$\begin{aligned}
& \sum_{k=1}^{n-1} C_{k,n-k} \left(f_{\Phi}(k, n-k)^2 + 4f_{\Phi}(k, n-k)E_U(\Phi_k) + 2E_U(\Phi_k)E_U(\Phi_{n-k}) \right) \\
& = \sum_{k=1}^{n-1} C_{k,n-k} \left(\binom{n-k}{2}^2 - 3 \binom{k}{2}^2 + 2 \binom{k}{2}^2 \frac{(2k-2)!!}{(2k-3)!!} \right. \\
& \quad + \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} - \binom{k}{2} \binom{n-k}{2} \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \\
& \quad \left. + \frac{1}{2} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \right) \\
& = \sum_{k=1}^{n-1} C_{k,n-k} \left(-2 \binom{k}{2}^2 + 2 \binom{k}{2}^2 \frac{(2k-2)!!}{(2k-3)!!} \right. \\
& \quad \left. + \frac{1}{2} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \right)
\end{aligned}$$

so that, finally

$$\begin{aligned}
E_U(\Phi_n^2) & = 2 \sum_{k=1}^{n-1} C_{k,n-k} E_U(\Phi_k^2) + \sum_{k=1}^{n-1} C_{k,n-k} \left(2 \binom{k}{2}^2 \frac{(2k-2)!!}{(2k-3)!!} \right. \\
& \quad \left. + \frac{1}{2} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} - 2 \binom{k}{2}^2 \right). \tag{2.3}
\end{aligned}$$

To compute the independent term in this recurrence, we compute the three sums that form it.

Claim 2.32.

$$\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} = \frac{n!2^{n-5}}{(2n-3)!!} \binom{n-1}{3}.$$

Proof of the Claim:

$$\begin{aligned}
& \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \\
&= \sum_{k=2}^{n-2} C_{k,n-k} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \\
&= \sum_{k=2}^{n-2} \frac{n!(2k-3)!!(2(n-k)-3)!!k!(n-k)!2^{k-1}(k-1)!2^{n-k-1}(n-k-1)!}{2(2n-3)!!k!(n-k)!2^2(k-2)!(n-k-2)!(2k-3)!!(2(n-k)-3)!!} \\
&= \frac{n!2^{n-5}}{(2n-3)!!} \sum_{k=2}^{n-2} (k-1)(n-k-1) \\
&= \frac{n!2^{n-5}}{(2n-3)!!} \left((n-1) \sum_{k=2}^{n-2} (k-1) - \sum_{k=2}^{n-2} (k-1)k \right) \\
&= \frac{n!2^{n-5}}{(2n-3)!!} \left((n-1) \sum_{k=1}^{n-3} k - 2 \sum_{k=2}^{n-2} \binom{k}{2} \right) \\
&= \frac{n!2^{n-5}}{(2n-3)!!} \left((n-1) \binom{n-2}{2} - 2 \binom{n-1}{3} \right) = \frac{n!2^{n-5}}{(2n-3)!!} \binom{n-1}{3}
\end{aligned}$$

Claim 2.33.

$$\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 \cdot \frac{(2k-2)!!}{(2k-3)!!} = \binom{n}{2}^2 \frac{2^{n-2} \cdot (n-1)!}{(2n-3)!!} - \binom{n}{2} \frac{(2n-1)(6n-7)}{15}.$$

Proof of the Claim: We start by developing this sum:

$$\begin{aligned}
& \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 \cdot \frac{(2k-2)!!}{(2k-3)!!} \\
&= \sum_{k=1}^{n-1} \frac{n!(2k-3)!!(2(n-k)-3)!!k^2(k-1)^2(2k-2)!!}{2(2n-3)!!k!(n-k)!4(2k-3)!!} \\
&= \frac{n!}{(2n-3)!!8} \sum_{k=1}^{n-1} \frac{(2(n-k)-2)!k^2(k-1)^22^{k-1}(k-1)!}{(n-k-1)!2^{n-k-1}k!(n-k)!} \\
&= \frac{n!}{(2n-3)!!2^{n+3}} \sum_{k=1}^{n-1} \frac{(2(n-k)-2)!k(k-1)^22^{2k}}{(n-k-1)!(n-k)!} \\
&= \frac{n!}{(2n-3)!!2^{n+3}} \sum_{j=1}^{n-1} \frac{(2j-2)!(n-j)(n-j-1)^22^{2n-2j}}{(j-1)!j!} \\
&= \frac{n!2^{n-3}}{(2n-3)!!} \sum_{j=1}^{n-1} \frac{(2j-2)!(n-j)(n-j-1)^2}{2^{2j}(j-1)!j!} = (*)
\end{aligned}$$

We shall compute now the sum

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

using the lookup algorithm. Setting

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

we have that

$$t_0 = \frac{(n-2)^2(n-1)}{4}, \quad \frac{t_{j+1}}{t_j} = \frac{(j+1/2)(j-n+3)^2(j+1)}{(j+2)(j-n+1)(j-n+2)(j+1)},$$

and $t_{n-2} = 0$. So, by the lookup algorithm

$$\begin{aligned} \sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)(n-j-2)^2}{2^{2j+2}j!(j+1)!} &= \sum_{j=0}^{n-3} \frac{(2j)!(n-j-1)(n-j-2)^2}{2^{2j+2}j!(j+1)!} \\ &= \frac{(n-2)^2(n-1)}{4} {}_4F_3 \left[\begin{matrix} \frac{1}{2} & 1 & 3-n & 3-n \\ 2 & 1-n & 2-n & \end{matrix} ; 1 \right]. \end{aligned} \quad (2.4)$$

Let us compute now this general hypergeometric series of type ${}_4F_3$ from its very definition:

$${}_4F_3 \left[\begin{matrix} \frac{1}{2} & 1 & 3-n & 3-n \\ 2 & 1-n & 2-n & \end{matrix} ; 1 \right] = \sum_{k=0}^{\infty} \frac{(\frac{1}{2})_k (1)_k (3-n)_k (3-n)_k}{(2)_k (1-n)_k (2-n)_k k!}$$

where:

$$\begin{aligned} \left(\frac{1}{2}\right)_k &= \frac{1}{2} \left(\frac{1}{2} + 1\right) \cdots \left(\frac{1}{2} + k - 1\right) = \frac{(2k-1)!!}{2^k} \\ (1)_k &= 1 \cdot 2 \cdots (1+k-1) = k! \\ (3-n)_k &= (3-n) \cdot (4-n) \cdots (3-n+k-1) = \frac{(-1)^k (n-3)!}{(n-k-3)!} \\ (2)_k &= 2 \cdot 3 \cdots (2+k-1) = (k+1)! \\ (1-n)_k &= (1-n) \cdot (2-n) \cdots (1-n+k-1) = \frac{(-1)^k (n-1)!}{(n-k-1)!} \\ (2-n)_k &= (2-n) \cdot (3-n) \cdots (2-n+k-1) = \frac{(-1)^k (n-2)!}{(n-k-2)!} \end{aligned}$$

and therefore

$$\begin{aligned} &{}_4F_3 \left[\begin{matrix} \frac{1}{2} & 1 & 3-n & 3-n \\ 2 & 1-n & 2-n & \end{matrix} ; 1 \right] \\ &= \sum_{k=0}^{\infty} \frac{(2k-1)!! k! (-1)^{2k} (n-3)!^2 (n-k-1)! (n-k-2)!}{2^k (n-k-3)!^2 (k+1)! (-1)^k (n-1)! (-1)^k (n-2)! k!} \\ &= \frac{(n-3)!^2}{(n-1)! (n-2)!} \sum_{k=0}^{n-3} \frac{(2k-1)!! (n-k-1)! (n-k-2)!}{2^k (k+1)! (n-k-3)!^2} \\ &= \frac{(n-3)!^2}{(n-1)! (n-2)!} \sum_{j=1}^{n-2} \frac{(2j-3)!! (n-j)! (n-j-1)!}{2^{j-1} j! (n-j-2)!^2} \\ &= \frac{(n-3)!^2}{(n-1)! (n-2)!} \left(\sum_{j=0}^{n-2} \frac{(2j-3)!! (n-j)! (n-j-1)!}{2^{j-1} j! (n-j-2)!^2} + \frac{2(n-1)! n!}{(n-2)!^2} \right) \end{aligned}$$

(in the last equality we have used that $(-3)!! = -1$ because of the recurrence $n!! = (n+2)!!/(n+2)$ that is imposed on the double factorial to extend it to negative integers; see [20]). Let us apply now the lookup algorithm to the sum

$$\sum_{j=0}^{n-2} \frac{(2j-3)!!(n-j)!(n-j-1)!}{2^{j-1}j!(n-j-2)!^2}.$$

Setting

$$t_j = \frac{(2j-3)!!(n-j)!(n-j-1)!}{2^{j-1}j!(n-j-2)!^2} = \frac{(2j-3)!!(n-j)(n-j-1)^2}{2^{j-1}j!}$$

we have that

$$t_0 = -\frac{2(n-1)n!}{(n-2)!^2}, \quad \frac{t_{j+1}}{t_j} = \frac{(j-1/2)(j+2-n)^2}{(j-n)(j-n+1)(j+1)}, \quad t_{n-1} = 0,$$

and hence

$$\begin{aligned} & \sum_{j=0}^{n-2} \frac{(2j-3)!!(n-j)!(n-j-1)!}{2^{j-1}j!(n-j-2)!^2} \\ &= -\frac{2(n-1)n!}{(n-2)!^2} {}_3F_2 \left[\begin{matrix} -\frac{1}{2} & 2-n & 2-n \\ -n & 1-n \end{matrix} ; 1 \right]. \end{aligned} \quad (2.5)$$

Now, we can compute this general hypergeometric series applying Identity (1.24), and we obtain (after reordering its parameters)

$${}_3F_2 \left[\begin{matrix} 2-n & -\frac{1}{2} & 2-n \\ -n & 1-n \end{matrix} ; 1 \right] = \frac{\Gamma(-n)\Gamma(-\frac{3}{2})}{\Gamma(\frac{1}{2}-n)\Gamma(-2)} \left(1 + \frac{n-2}{5(n-1)} \right). \quad (2.6)$$

Notice that the condition “ $d-b-c > 1$ ” assume in (1.24) is not satisfied here: we overcome this drawback by taking limits. In this expression, by identity (1.14),

$$\Gamma\left(\frac{1}{2}-n\right) = \frac{(-2)^n \sqrt{\pi}}{(2n-1)!!}, \quad \Gamma\left(-\frac{3}{2}\right) = \frac{4\sqrt{\pi}}{3}$$

(notice that the second equality is a particular case of the first, taking $n=2$). The fraction $\Gamma(-n)/\Gamma(-2)$ is indeterminate of the form $0/0$, but we can solve this indetermination by using the first formula in (1.15) and taking limits:

$$\begin{aligned} \frac{\Gamma(-n)}{\Gamma(-2)} &= \lim_{x \rightarrow 0} \frac{\Gamma(-n+x)}{\Gamma(-2+x)} = \lim_{x \rightarrow 0} \frac{(-1)^n(1+O(x))/(n!x)}{(1+O(x))/(2x)} \\ &= \lim_{x \rightarrow 0} \frac{2(-1)^n(1+O(x))}{n!(1+O(x))} = \frac{2(-1)^n}{n!}. \end{aligned} \quad (2.7)$$

Using these values in Identity (2.6) we obtain

$$\begin{aligned} {}_3F_2 \left[\begin{matrix} 2-n & -\frac{1}{2} & 2-n \\ -n & 1-n \end{matrix} ; 1 \right] &= \frac{2(-1)^n 4\sqrt{\pi}(2n-1)!!}{n!3(-2)^n \sqrt{\pi}} \left(1 + \frac{n-2}{5(n-1)} \right) \\ &= \frac{2^{3-n}(6n-7)(2n-1)!!}{15n!(n-1)}. \end{aligned}$$

We reverse now the cascade of computations. First, by Identity (2.5),

$$\begin{aligned} \sum_{j=0}^{n-2} \frac{(2j-3)!!(n-j)!(n-j-1)!}{2^{j-1}j!(n-j-2)!^2} &= -\frac{2(n-1)!n!}{(n-2)!^2} \cdot \frac{2^{3-n}(6n-7)(2n-1)!!}{15n!(n-1)} \\ &= -\frac{2^{4-n}(6n-7)(2n-1)!!}{15(n-2)!}. \end{aligned}$$

The value of ${}_4F_3$ is, thus,

$$\begin{aligned} {}_4F_3 \left[\begin{matrix} \frac{1}{2} & 1 & 3-n & 3-n \\ 2 & 1-n & 2-n & \end{matrix} ; 1 \right] \\ &= \frac{(n-3)!^2}{(n-1)!(n-2)!} \left(\sum_{j=0}^{n-2} \frac{(2j-3)!!(n-j)!(n-j-1)!}{2^{j-1}j!(n-j-2)!^2} + \frac{2(n-1)!n!}{(n-2)!^2} \right) \\ &= \frac{(n-3)!^2}{(n-1)!(n-2)!} \left(-\frac{2^{4-n}(6n-7)(2n-1)!!}{15(n-2)!} + \frac{2(n-1)!n!}{(n-2)!^2} \right) \\ &= \frac{30(n-1)n! - 2^{4-n}(6n-7)(2n-1)!!}{15(n-2)^2(n-1)!}. \end{aligned}$$

Using this value in Identity (2.4), we obtain

$$\begin{aligned} \sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)(n-j-2)^2}{2^{2j+2}j!(j+1)!} \\ &= \frac{(n-2)^2(n-1)}{4} \cdot \frac{30(n-1)n! - 2^{4-n}(6n-7)(2n-1)!!}{15(n-2)^2(n-1)!} \\ &= \frac{30(n-1)n! - 2^{4-n}(6n-7)(2n-1)!!}{60(n-2)!}. \end{aligned}$$

Finally, at last:

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 \cdot \frac{(2k-2)!!}{(2k-3)!!} &= (*) \\ &= \frac{n!2^{n-3}}{(2n-3)!!} \cdot \frac{(30(n-1)n! - 2^{4-n}(6n-7)(2n-1)!!)}{60(n-2)!} \\ &= \frac{n(n-1)[15 \cdot 2^{n-2}(n-1)n! - 2(6n-7)(2n-1)!!]}{60(2n-3)!!} \\ &= \binom{n}{2} \frac{2^{n-2} \cdot (n-1)!}{(2n-3)!!} - \binom{n}{2} \frac{(2n-1)(6n-7)}{15}. \end{aligned}$$

Remark 2.34. We want to point out here that if we compute

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

with *Mathematica*, it yields

$$\binom{n}{2} \left((n-1) - \frac{2^{4-2n}(2n-1)(6n-7)(2n-2)!}{15 \cdot n!(n-1)!} \right)$$

and it is easy to see that it agrees with our result for that sum. We do not know if there is a way simpler than ours to obtain this expression.

Claim 2.35.

$$\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 = \frac{1}{2} \binom{n}{2}^2 - \frac{2^{n-8} n! (15n^2 - 27n + 10)}{(2n-3)!!}$$

Proof of the Claim: We start by developing this sum:

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 &= \sum_{k=2}^{n-1} \frac{n! (2k-3)!! (2n-2k-3)!! k^2 (k-1)^2}{2 \cdot (2n-3)!! k! (n-k)! 4} \\ &= \frac{n!}{8 \cdot (2n-3)!!} \sum_{k=2}^{n-1} \frac{(2k-2)! (2n-2k-2)! k^2 (k-1)^2}{2^{k-1} (k-1)! 2^{n-k-1} (n-k-1)! k! (n-k)!} \\ &= \frac{n!}{(2n-3)!! 2^{n+1}} \sum_{k=2}^{n-1} \frac{(2k-2)! (2n-2k-2)! k}{(k-2)!^2 (n-k-1)! (n-k)!} \\ &= \frac{n!}{(2n-3)!! 2^{n+1}} \sum_{j=0}^{n-3} \frac{(2j+2)! (2n-2j-6)! (j+2)}{j!^2 (n-j-3)! (n-j-2)!} = (*) \end{aligned}$$

We shall apply the lookup algorithm to compute the last sum. Let

$$t_j = \frac{(2j+2)! (2n-2j-6)! (j+2)}{j!^2 (n-j-3)! (n-j-2)!}$$

so that

$$t_0 = \frac{4 \cdot (2n-6)!}{(n-2)! (n-3)!}, \quad \frac{t_{j+1}}{t_j} = \frac{(j+3)(j+3/2)(j-n+2)}{(j-n+7/2)(j+1)^2},$$

and $t_{n-2} = 0$. But in this case it is wrong to write

$$\sum_{j=0}^{n-3} \frac{(2j+2)! (2n-2j-6)! (j+2)}{j!^2 (n-j-3)! (n-j-2)!} = \frac{4 \cdot (2n-6)!}{(n-2)! (n-3)!} \cdot {}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2} - n \end{matrix} ; 1 \right]$$

because, since

$${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2} - n \end{matrix} ; 1 \right] = \sum_{k=0}^{\infty} \frac{\left(\frac{3}{2}\right)_k (3)_k (2-n)_k}{(1)_k \left(\frac{7}{2} - n\right)_k k!}$$

and $(2-n)_k = 0$ if, and only if, $k \geq n-1$, we have actually that

$${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2} - n \end{matrix} ; 1 \right] = \sum_{k=0}^{n-2} \frac{\left(\frac{3}{2}\right)_k (3)_k (2-n)_k}{(1)_k \left(\frac{7}{2} - n\right)_k k!},$$

while the upper limit of our original sum is $n - 3$. Therefore,

$$\begin{aligned} & \sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{j!^2(n-j-3)!(n-j-2)!} \\ &= \frac{4 \cdot (2n-6)!}{(n-2)!(n-3)!} \left({}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2}-n \end{matrix} ; 1 \right] \right. \\ & \quad \left. - \frac{\left(\frac{3}{2}\right)_{n-2} (3)_{n-2} (2-n)_{n-2}}{(1)_{n-2} \left(\frac{7}{2}-n\right)_{n-2} (n-2)!} \right) \end{aligned} \quad (2.8)$$

Let us compute the subtrahend inside the parentheses:

$$\begin{aligned} \left(\frac{3}{2}\right)_{n-2} &= \frac{3}{2} \left(\frac{3}{2} + 1\right) \cdots \left(\frac{3}{2} + n - 3\right) = \frac{(2n-3)!!}{2^{n-2}} \\ (3)_{n-2} &= 3 \cdot 4 \cdots (3+n-3) = \frac{n!}{2} \\ (2-n)_{n-2} &= (2-n) \cdot (3-n) \cdots (2-n+n-3) = (-1)^{n-2} (n-2)! \\ (1)_{n-2} &= 1 \cdot 2 \cdots (1+n-3) = (n-2)! \\ \left(\frac{7}{2}-n\right)_{n-2} &= \left(\frac{7}{2}-n\right) \left(\frac{7}{2}-n+1\right) \cdots \left(\frac{7}{2}-n+n-3\right) = (-1)^{n-3} \frac{(2n-7)!!}{2^{n-2}} \end{aligned}$$

and therefore

$$\begin{aligned} & \frac{\left(\frac{3}{2}\right)_{n-2} (3)_{n-2} (2-n)_{n-2}}{(1)_{n-2} \left(\frac{7}{2}-n\right)_{n-2} (n-2)!} \\ &= \frac{(2n-3)!! n! (-1)^{n-2} (n-2)! 2^{n-2}}{2^{n-1} (n-2)! (-1)^{n-3} (2n-7)!! (n-2)!} \\ &= -\frac{(2n-3)(2n-5)n(n-1)}{2} \end{aligned} \quad (2.9)$$

Now, to compute ${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2}-n \end{matrix} ; 1 \right]$, we use Identity (1.25) and we obtain

$${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2}-n \end{matrix} ; 1 \right] = \frac{\Gamma(1)\Gamma\left(\frac{7}{2}-n\right)\Gamma(-2)}{\Gamma\left(\frac{3}{2}\right)\Gamma(1)\Gamma(-n)} {}_3F_2 \left[\begin{matrix} -\frac{1}{2} & 2-n & -2 \\ 1 & -n \end{matrix} ; 1 \right]. \quad (2.10)$$

In the right-hand side expression, we know —using (1.13) and (1.14)— that

$$\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}, \quad \Gamma\left(\frac{7}{2}-n\right) = \frac{(-1)^{n-3} 2^{n-3}}{(2n-7)!!} \sqrt{\pi}$$

and, from (2.7), we also know that

$$\frac{\Gamma(-2)}{\Gamma(-n)} = \frac{(-1)^n n!}{2}.$$

Finally, since $(-2)_k = 0$ for every $k \geq 3$,

$$\begin{aligned} & {}_3F_2 \left[\begin{matrix} -\frac{1}{2} & 2-n & -2 \\ 1 & -n \end{matrix} ; 1 \right] \\ &= 1 + \frac{(-1/2)_1(2-n)_1(-2)_1}{(1)_1(-n)_1 1!} + \frac{(-1/2)_2(2-n)_2(-2)_2}{(1)_2(-n)_2 2!} \\ &= 1 + \frac{(-1/2)(2-n)(-2)}{-n} + \frac{(-1/2)(1/2)(2-n)(3-n)(-2)(-1)}{2(-n)(-n+1)2} \\ &= \frac{15n^2 - 27n + 10}{8n(n-1)}. \end{aligned}$$

Therefore, the expression (2.10) becomes

$$\begin{aligned} & {}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 1 & \frac{7}{2}-n \end{matrix} ; 1 \right] = \frac{(-1)^{n-3} 2^{n-3} \sqrt{\pi} 2 (-1)^n n!}{(2n-7)!! \sqrt{\pi} 2} \cdot \frac{(15n^2 - 27n + 10)}{8n(n-1)} \\ &= -\frac{2^{n-6} (n-2)! (15n^2 - 27n + 10)}{(2n-7)!!} \end{aligned}$$

Replacing this value and the one given by (2.9) in equality (2.8), we obtain

$$\begin{aligned} & \sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{j!^2(n-j-3)!(n-j-2)!} \\ &= \frac{4 \cdot (2n-6)!}{(n-2)!(n-3)!} \left(-\frac{2^{n-6} (n-2)! (15n^2 - 27n + 10)}{(2n-7)!!} \right. \\ & \quad \left. + \frac{(2n-3)(2n-5)n(n-1)}{2} \right) \\ &= \frac{(2n-3)!n(n-1)}{(n-2)!^2} - 2^{2n-7} (15n^2 - 27n + 10) \end{aligned}$$

and, finally at last,

$$\begin{aligned} & \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 = (*) \\ &= \frac{n!}{(2n-3)!! 2^{n+1}} \left(\frac{(2n-3)!n(n-1)}{(n-2)!^2} - 2^{2n-7} (15n^2 - 27n + 10) \right) \\ &= \frac{1}{2} \binom{n}{2}^2 - \frac{2^{n-8} n! (15n^2 - 27n + 10)}{(2n-3)!!} \end{aligned}$$

It is time to return to the expression for $E_U(\Phi_n^2)$ given by equation (2.3). Its independent term turns out to be

$$\begin{aligned} & 2 \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 \frac{(2k-2)!!}{(2k-3)!!} \\ & \quad + \frac{1}{2} \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} \binom{n-k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \end{aligned}$$

$$\begin{aligned}
& - 2 \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2}^2 \\
& = 2 \left(\binom{n}{2}^2 \frac{2^{n-2} \cdot (n-1)!}{(2n-3)!!} - \binom{n}{2} \frac{(2n-1)(6n-7)}{15} \right) \\
& \quad + \frac{1}{2} \cdot \frac{n!2^{n-5}}{(2n-3)!!} \binom{n-1}{3} \\
& \quad - 2 \left(\frac{1}{2} \binom{n}{2}^2 - \frac{2^{n-8}n!(15n^2 - 27n + 10)}{(2n-3)!!} \right) \\
& = \frac{1}{6} \binom{n}{2} \left(\frac{49n^3 - 57n^2 - 22n + 24}{8} \cdot \frac{(2n-4)!!}{(2n-3)!!} - \frac{63n^2 - 95n + 28}{5} \right)
\end{aligned}$$

which finally proves the identity in the statement. \square

Unfortunately, we have not been able to derive from the recurrence in the last proposition an explicit formula for $E_U(\Phi_n^2)$, but it allows to compute recurrently $E_U(\Phi_n^2)$, starting with the obvious initial condition $E_U(\Phi_1^2) = 0$, and then to compute the variance of Φ_n under the uniform model by means of

$$\sigma_U^2(\Phi_n) = E_U(\Phi_n^2) - E_U(\Phi_n)^2.$$

We have computed these variances up to $n = 1000$, using the recurrence above for $E_U(\Phi_n^2)$ and Theorem 2.28 for $E_U(\Phi_n)$. Table 2.1 gives the values of $\sigma_U^2(\Phi_n)$ for $n = 3, \dots, 20$ obtained in this way. The code and the rest of the values obtained are available on the GitHub repository associated to this PhD Thesis [97]. The R script used to compute them is also available in Appendix A.4.3. To double-check the recurrence, we have also computed the values of $\sigma_U^2(\Phi_n)$, for $n = 3, \dots, 8$, from the cophenetic indices of all trees in the corresponding \mathcal{BT}_n , and they agree with the figures give in Table 2.1. The R script used to compute these exact values and to compare them with the values obtained through our recurrence is also available in Appendix A.4.3 and on the GitHub repository [97].

n	3	4	5	6	7	8
$\sigma_U^2(\Phi_n)$	0	0.64	4.77551	19.58277	58.97521	146.2314
n	9	10	11	12	13	14
$\sigma_U^2(\Phi_n)$	316.7786	621.0986	1127.730	1926.353	3130.941	4882.977
n	15	16	17	18	19	20
$\sigma_U^2(\Phi_n)$	7354.712	10752.48	15320.03	21341.97	29147.09	39111.90

Table 2.1: $\sigma_U^2(\Phi_n)$ for $n = 3, \dots, 20$.

We have estimated the main order in the expansion of $\sigma_U^2(\Phi_n)$ as a function of n , by performing the minimum squares linear regression of $\ln(\sigma_U^2(\Phi_n))$ as a

function of $\ln(n)$ for $n = 900, \dots, 1000$, and the result has been

$$\ln(\sigma_U^2(\Phi_n)) \approx -3.8743868 + 5.0657352 \cdot \ln(n),$$

with a determination coefficient $R^2 \approx 1$. We conclude then that, according to our approximations, $\sigma_U^2(\Phi_n)$ is in $O(n^{5.0657})$. We conjecture that, actually, $\sigma_U^2(\Phi_n)$ is in $O(n^5)$, the order of $E_U(\Phi_n)^2$. Fig. 2.11 displays $\ln(\sigma_U^2(\Phi_n))$ as a function of $\ln(n)$, together with the corresponding regression line.

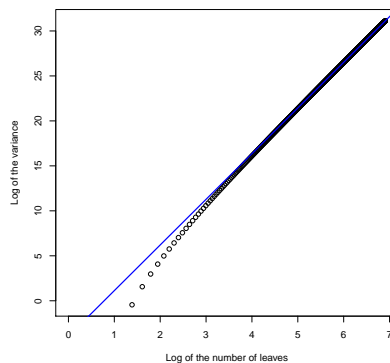


Figure 2.11: Log-log plot of $\sigma_U^2(\Phi_n)$.

2.6 On the variance of S under the uniform model

The techniques applied in the last section to study $E_U(\Phi_n^2)$ can also be used to obtain a recurrence for $E_U(S_n^2)$ that allows to compute recurrently these values, and hence those of $\sigma_U^2(S_n)$. The variance of the Sackin index under the Yule model was computed in As far as the uniform model goes, Blum, François and Janson established in [14, Rem. 3] its limit behaviour, which we have recalled in (1.6):

$$\sigma_U^2(S_n) \sim \left(\frac{10 - 3\pi}{3}\right)n^3,$$

and Rogers provided in [92] a inefficient procedure to compute this variance from a recurrence to compute the distribution of S_n . To our knowledge, no efficient recurrent formula allowing the computation of this variance for every n is known so far, and we have considered it of interest to provide this recurrence in this thesis.

Recall from the previous section that, for every $1 \leq k \leq n - 1$,

$$C_{k,n-k} = \frac{1}{2} \binom{n}{k} \frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}.$$

Proposition 2.36. *For every $n \geq 2$,*

$$E_U(S_n^2) = 2 \sum_{k=1}^{n-1} C_{k,n-k} E_U(S_k^2) + \frac{5n2^{n-2}n!}{(2n-3)!!} - n(5n-2)$$

Proof. Applying Lemma 2.30 taking as I the Sackin index S , for which

$$f_S(k, n-k) = n, \quad E_U(S_k) = k \left(\frac{(2k-2)!!}{(2k-3)!!} - 1 \right),$$

we obtain

$$\begin{aligned} E_U(S_n^2) &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(S_k^2) + f_S(k, n-k)^2 + 4f_S(k, n-k)E_U(S_k) \right. \\ &\quad \left. + 2E_U(S_k)E_U(S_{n-k}) \right) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(S_k^2) + n^2 + 4nk \left(\frac{(2k-2)!!}{(2k-3)!!} - 1 \right) \right. \\ &\quad \left. + 2k(n-k) \left(\frac{(2k-2)!!}{(2k-3)!!} - 1 \right) \left(\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} - 1 \right) \right) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(S_k^2) + n^2 + 4nk \frac{(2k-2)!!}{(2k-3)!!} - 4nk \right. \\ &\quad \left. + 2k(n-k) \frac{(2k-2)!!}{(2k-3)!!} \cdot \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} + 2k(n-k) \right. \\ &\quad \left. - 2k(n-k) \frac{(2k-2)!!}{(2k-3)!!} - 2k(n-k) \cdot \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \right) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(S_k^2) + 4nk \frac{(2k-2)!!}{(2k-3)!!} - 2k^2 \right. \\ &\quad \left. + 2k(n-k) \frac{(2k-2)!!}{(2k-3)!!} \cdot \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \right. \\ &\quad \left. - 4k(n-k) \frac{(2k-2)!!}{(2k-3)!!} \right) = (*) \end{aligned}$$

because, by the symmetry of $C_{k,n-k}$,

$$\sum_{k=1}^{n-1} C_{k,n-k} k(n-k) \frac{(2k-2)!!}{(2k-3)!!} = \sum_{k=1}^{n-1} C_{k,n-k} k(n-k) \cdot \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!}$$

and

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} (n^2 - 4nk + 2k(n-k)) &= \sum_{k=1}^{n-1} C_{k,n-k} ((n-k)^2 - 3k^2) \\ &= -2 \sum_{k=1}^{n-1} C_{k,n-k} k^2. \end{aligned}$$

Simplifying one step further the sum (*), we finally obtain

$$E_U(S_n^2) = 2 \sum_{k=1}^{n-1} C_{k,n-k} E_U(S_k^2) + \sum_{k=1}^{n-1} C_{k,n-k} \left(4k^2 \frac{(2k-2)!!}{(2k-3)!!} - 2k^2 + 2k(n-k) \frac{(2k-2)!!}{(2k-3)!!} \cdot \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \right). \quad (2.11)$$

To compute the independent term in this recurrence, we calculate the three sums that form it.

Claim 2.37.

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

Proof of the Claim:

$$\begin{aligned} & \sum_{k=1}^{n-1} C_{k,n-k} k(n-k) \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \\ &= \sum_{k=1}^{n-1} \frac{n!(2k-3)!!(2(n-k)-3)!!k(n-k)(2k-2)!!(2(n-k)-2)!!}{2 \cdot (2n-3)!!k!(n-k)!(2k-3)!!(2(n-k)-3)!!} \\ &= \frac{n!}{2 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{k(n-k)2^{k-1}(k-1)!2^{n-k-1}(n-k-1)!}{k!(n-k)!} \\ &= \frac{n!}{2 \cdot (2n-3)!!} \cdot (n-1)2^{n-2} = \frac{n!(n-1)2^{n-3}}{(2n-3)!!} \end{aligned}$$

Claim 2.38.

$$\sum_{k=1}^{n-1} C_{k,n-k} k^2 \cdot \frac{(2k-2)!!}{(2k-3)!!} = \frac{n2^{n-2}n!}{(2n-3)!!} - \frac{n(2n-1)}{2}$$

Proof of the Claim: We start by developing this sum:

$$\begin{aligned} & \sum_{k=1}^{n-1} C_{k,n-k} k^2 \cdot \frac{(2k-2)!!}{(2k-3)!!} = \sum_{k=1}^{n-1} \frac{n!(2k-3)!!(2(n-k)-3)!!k^2(2k-2)!!}{2(2n-3)!!k!(n-k)!(2k-3)!!} \\ &= \frac{n!}{2 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{(2(n-k)-2)!k^22^{k-1}(k-1)!}{(n-k-1)!2^{n-k-1}k!(n-k)!} \\ &= \frac{n!2^n}{2 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{(2(n-k)-2)!k}{2^{2n-2k}(n-k-1)!(n-k)!} \\ &= \frac{n!2^{n-1}}{(2n-3)!!} \sum_{j=1}^{n-1} \frac{(2j-2)!(n-j)}{2^{2j}(j-1)!j!} \end{aligned}$$

We shall compute now the sum

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

using the lookup algorithm. Setting

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

we have that

$$t_0 = \frac{n-1}{4}, \quad \frac{t_{j+1}}{t_j} = \frac{(j+1/2)(j-n+2)(j+1)}{(j+2)(j-n+1)(j+1)},$$

and $t_{n-1} = 0$, and hence, by the lookup algorithm,

$$\sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)}{2^{2j+2}j!(j+1)!} = \frac{n-1}{4} \cdot {}_3F_2 \left[\begin{matrix} \frac{1}{2} & 1 & 2-n \\ 2 & 1-n \end{matrix} ; 1 \right]. \quad (2.12)$$

We can compute this ${}_3F_2$ hypergeometric series using the Identity (1.26):

$${}_3F_2 \left[\begin{matrix} \frac{1}{2} & 1 & 2-n \\ 2 & 1-n \end{matrix} ; 1 \right] = \frac{-n}{(-\frac{1}{2})(1-n)} \left(\frac{\Gamma(-n)\Gamma(-1/2)}{\Gamma(1/2-n)\Gamma(-1)} - 1 \right)$$

where, by (1.14),

$$\Gamma\left(-\frac{1}{2}\right) = -2\sqrt{\pi}, \quad \Gamma\left(\frac{1}{2}-n\right) = \frac{(-2)^n\sqrt{\pi}}{(2n-1)!!}$$

and the value of the indeterminate fraction $\Gamma(-n)/\Gamma(-1)$ can be computed using the first MacLaurin series in (1.15) and taking limits:

$$\begin{aligned} \frac{\Gamma(-n)}{\Gamma(-1)} &= \lim_{x \rightarrow 0} \frac{\Gamma(-n+x)}{\Gamma(-1+x)} = \lim_{x \rightarrow 0} \frac{(-1)^n(1+O(x))/(n!x)}{-(1+O(x))/x} \\ &= \lim_{x \rightarrow 0} \frac{(-1)^{n-1}(1+O(x))}{n!(1+O(x))} = \frac{(-1)^{n-1}}{n!}. \end{aligned}$$

Therefore,

$$\begin{aligned} {}_3F_2 \left[\begin{matrix} \frac{1}{2} & 1 & 2-n \\ 2 & 1-n \end{matrix} ; 1 \right] &= -\frac{2n}{n-1} \left(\frac{-2\sqrt{\pi}(-1)^{n-1}(2n-1)!!}{n!(-2)^n\sqrt{\pi}} - 1 \right) \\ &= \frac{2^{n-1}n! - (2n-1)!!}{2^{n-2}(n-1)!(n-1)} \end{aligned}$$

Then, Identity (2.12) yields

$$\sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)}{2^{2j+2}j!(j+1)!} = \frac{(n-1)(2^{n-1}n! - (2n-1)!!)}{4 \cdot 2^{n-2}(n-1)!(n-1)} = \frac{2^{n-1}n! - (2n-1)!!}{2^n(n-1)!}$$

and, finally,

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} k^2 \cdot \frac{(2k-2)!!}{(2k-3)!!} &= \frac{n!2^{n-1}}{(2n-3)!!} \sum_{j=1}^{n-1} \frac{(2j-2)!(n-j)}{2^{2j}(j-1)!j!} \\ &= \frac{n!2^{n-1}}{(2n-3)!!} \cdot \frac{2^{n-1}n! - (2n-1)!!}{2^n(n-1)!} = \frac{n2^{n-2}n!}{(2n-3)!!} - \frac{n(2n-1)}{2}. \end{aligned}$$

Claim 2.39.

$$\sum_{k=1}^{n-1} C_{k,n-k} k^2 = \frac{n^2}{2} - \frac{2^{n-3}n!}{(2n-3)!!}$$

Proof of the Claim: We start by developing this sum:

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} k^2 &= \sum_{k=1}^{n-1} \frac{n!(2k-3)!!(2n-2k-3)!!k^2}{2 \cdot (2n-3)!!k!(n-k)!} \\ &= \frac{n!}{2 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{(2k-2)!(2n-2k-2)!k^2}{2^{k-1}(k-1)!2^{n-k-1}(n-k-1)!k!(n-k)!} \\ &= \frac{n!}{(2n-3)!!2^{n-1}} \sum_{k=1}^{n-1} \frac{(2k-2)!(2n-2k-2)!k}{(k-1)!^2(n-k-1)!(n-k)!} \\ &= \frac{n!}{(2n-3)!!2^{n-1}} \sum_{j=0}^{n-2} \frac{(2j)!(2n-2j-4)!(j+1)}{j!^2(n-j-2)!(n-j-1)!} \end{aligned}$$

We shall apply the lookup algorithm to compute the sum on the right hand side. Let

$$t_j = \frac{(2j)!(2n-2j-4)!(j+1)}{j!^2(n-j-2)!(n-j-1)!}$$

so that

$$t_0 = \frac{(2n-4)!}{(n-2)!(n-1)!}, \quad \frac{t_{j+1}}{t_j} = \frac{(j+2)(j+1/2)(j-n+1)}{(j-n+5/2)(j+1)^2},$$

and $t_{n-1} = 0$. But, as we already encountered in the proof of Claim 2.35, it is wrong to write

$$\sum_{j=0}^{n-2} \frac{(2j)!(2n-2j-4)!(j+1)}{j!^2(n-j-2)!(n-j-1)!} = \frac{(2n-4)!}{(n-2)!(n-1)!} \cdot {}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} & 1-n \\ 1 & \frac{5}{2}-n \end{matrix} ; 1 \right]$$

because, since, by definition,

$${}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} & 1-n \\ 1 & \frac{5}{2}-n \end{matrix} ; 1 \right] = \sum_{k=0}^{\infty} \frac{(2)_k \left(\frac{1}{2}\right)_k (1-n)_k}{(1)_k \left(\frac{5}{2}-n\right)_k k!}$$

and $(1-n)_k = 0$ for every $k \geq n$, we have actually that

$${}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} & 1-n \\ 1 & \frac{5}{2}-n \end{matrix} ; 1 \right] = \sum_{k=0}^{n-1} \frac{(2)_k \left(\frac{1}{2}\right)_k (1-n)_k}{(1)_k \left(\frac{5}{2}-n\right)_k k!}$$

while the upper limit of our original sum is $n - 2$. Therefore,

$$\begin{aligned} & \sum_{j=0}^{n-2} \frac{(2j)!(2n-2j-4)!(j+1)}{j!^2(n-j-2)!(n-j-1)!} \\ &= \frac{(2n-4)!}{(n-2)!(n-1)!} \left({}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} & 1-n \\ 1 & \frac{5}{2}-n & \end{matrix} ; 1 \right] \right. \\ & \quad \left. - \frac{(2)_{n-1} \left(\frac{1}{2}\right)_{n-1} (1-n)_{n-1}}{(1)_{n-1} \left(\frac{5}{2}-n\right)_{n-1} (n-1)!} \right) \end{aligned} \quad (2.13)$$

Let us compute the subtrahend inside the parentheses:

$$\begin{aligned} (2)_{n-1} &= 2 \cdot 3 \cdots (2+n-2) = n! \\ \left(\frac{1}{2}\right)_{n-1} &= \frac{1}{2} \left(\frac{1}{2}+1\right) \cdots \left(\frac{1}{2}+n-2\right) = \frac{(2n-3)!!}{2^{n-1}} \\ (1-n)_{n-1} &= (1-n) \cdot (2-n) \cdots (1-n+n-2) = (-1)^{n-1} (n-1)! \\ (1)_{n-1} &= 1 \cdot 2 \cdots (1+n-2) = (n-1)! \\ \left(\frac{5}{2}-n\right)_{n-1} &= \left(\frac{5}{2}-n\right) \left(\frac{5}{2}-n+1\right) \cdots \left(\frac{5}{2}-n+n-2\right) = \frac{(-1)^{n-2} (2n-5)!!}{2^{n-1}} \end{aligned}$$

and therefore

$$\begin{aligned} & \frac{(2)_{n-1} \left(\frac{1}{2}\right)_{n-1} (1-n)_{n-1}}{(1)_{n-1} \left(\frac{5}{2}-n\right)_{n-1} (n-1)!} \\ &= \frac{n!(2n-3)!!(-1)^{n-1}(n-1)!2^{n-1}}{2^{n-1}(n-1)!(-1)^{n-2}(2n-5)!!(n-1)!} = -n(2n-3) \end{aligned}$$

On the other hand, we shall compute ${}_3F_2 \left[\begin{matrix} \frac{1}{2} & 2 & 1-n \\ 1 & \frac{5}{2}-n & \end{matrix} ; 1 \right]$ permutating its parameters and using again Identity (1.24). But we cannot apply this identity directly to this hypergeometric sum, because if we do so, we obtain

$${}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} & 1-n \\ \frac{5}{2}-n & 1 & \end{matrix} ; 1 \right] = \frac{\Gamma(5/2-n)\Gamma(1)}{\Gamma(2-n)\Gamma(3/2)} \left(1 - \frac{\frac{1}{2}(1-n)}{1 \cdot 0} \right)$$

where $\Gamma(2-n) = \infty$. To overcome this problem, we apply this identity to

compute ${}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} + x & 1 - n \\ \frac{5}{2} - n & 1 & \end{matrix} ; 1 \right]$ and then we take limit for $x \rightarrow 0$:

$$\begin{aligned} {}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} & 1 - n \\ \frac{5}{2} - n & 1 & \end{matrix} ; 1 \right] &= \lim_{x \rightarrow 0} {}_3F_2 \left[\begin{matrix} 2 & \frac{1}{2} + x & 1 - n \\ \frac{5}{2} - n & 1 & \end{matrix} ; 1 \right] \\ &= \lim_{x \rightarrow 0} \frac{\Gamma(\frac{5}{2} - n) \Gamma(1 - x)}{\Gamma(2 - n + x) \Gamma(\frac{3}{2})} \left(1 - \frac{(1 - n)(\frac{1}{2} + x)}{x} \right) \\ &= \lim_{x \rightarrow 0} \frac{\Gamma(\frac{5}{2} - n) (1 + O(x))}{\Gamma(\frac{3}{2}) \left(\frac{(-1)^{n-1}}{(n-2)!} + O(1) \right)} \left(\frac{2nx + n - 1}{2x} \right) \\ &= \lim_{x \rightarrow 0} \frac{\Gamma(\frac{5}{2} - n) (1 + O(x)) (2nx + n - 1)}{2\Gamma(\frac{3}{2}) \left(\frac{(-1)^{n-1}}{(n-2)!} + O(x) \right)} \\ &= \frac{\Gamma(\frac{5}{2} - n) (n - 1)}{2\Gamma(\frac{3}{2}) \frac{(-1)^{n-1}}{(n-2)!}} = -\frac{2^{n-2}(n-1)!}{(2n-5)!!} \end{aligned}$$

where, in the last equality, we have used that, by (1.13) and (1.14),

$$\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}, \quad \Gamma\left(\frac{5}{2} - n\right) = \frac{(-1)^{n-2} 2^{n-2} \sqrt{\pi}}{(2n-5)!!}.$$

Therefore, Identity (2.13) yields

$$\sum_{j=0}^{n-2} \frac{(2j)!(2n-2j-4)!(j+1)}{j!^2(n-j-2)!(n-j-1)!} = \frac{(2n-4)!}{(n-2)!(n-1)!} \left(n(2n-3) - \frac{2^{n-2}(n-1)!}{(2n-5)!!} \right)$$

and, finally,

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

It is time to return to the independent term of the expression for $E_U(S_n^2)$ given by equation (2.11):

$$\begin{aligned} 4 \sum_{k=1}^{n-1} C_{k,n-k} k^2 \frac{(2k-2)!!}{(2k-3)!!} - 2 \sum_{k=1}^{n-1} C_{k,n-k} k^2 + \\ + 2 \sum_{k=1}^{n-1} C_{k,n-k} k(n-k) \frac{(2k-2)!!}{(2k-3)!!} \cdot \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} \\ = 4 \left(\frac{n 2^{n-2} n!}{(2n-3)!!} - \frac{n(2n-1)}{2} \right) + \frac{n!(n-1) 2^{n-2}}{(2n-3)!!} - 2 \left(\frac{n^2}{2} - \frac{2^{n-3} n!}{(2n-3)!!} \right) \\ = \frac{5n 2^{n-2} n!}{(2n-3)!!} - n(5n-2). \end{aligned}$$

This completes the proof of the identity in the statement. \square

We have been neither able to derive from the recurrence in the last proposition an explicit formula for $E_U(S_n^2)$. But, as it was also the case with the total cophenetic index, this recurrence allows one to compute recurrently $E_U(S_n^2)$, starting with the obvious initial condition $E_U(S_1^2) = 0$, and then to compute the variance of S_n under the uniform model, for each desired n , by means of

$$\sigma_U^2(S_n) = E_U(S_n^2) - E_U(S_n)^2.$$

We have computed these variances up to $n = 1000$. Table 2.2 gives the values of $\sigma_U^2(S_n)$ for $n = 3, \dots, 20$. The R script used to compute them and the rest of the values obtained are available on the GitHub repository associated to this PhD Thesis [97]; the R script is also available in Appendix A.4.4.

n	3	4	5	6	7	8
$\sigma_U^2(S_n)$	0	0.16	0.7755	2.2358	4.9991	9.5765
n	9	10	11	12	13	14
$\sigma_U^2(S_n)$	16.5219	26.4242	39.9017	57.5982	80.1794	108.3305
n	15	16	17	18	19	20
$\sigma_U^2(S_n)$	142.7538	184.1671	233.3023	290.9038	357.7276	434.5405

Table 2.2: $\sigma_U^2(S_n)$ for $n = 3, \dots, 20$.

To double-check the recurrence, we have computed the values of $\sigma_U^2(S_n)$, for $n = 3, \dots, 8$, from the Sackin indices of all trees in the corresponding \mathcal{BT}_n , and they agree with the figures given by our recurrence. The R script used in these computations is also available in Appendix A.4.4 and on the GitHub repository [97].

We have estimated the main order in the expansion of $\sigma_U^2(S_n)$ as a function of n , by performing the minimum squares linear regression of $\ln(\sigma_U^2(S_n))$ as a function of $\ln(n)$ for $n = 900, \dots, 1000$, and the result has been

$$\ln(\sigma_U^2(S_n)) \approx -2.347121 + 3.078995 \cdot \ln(n),$$

with a determination coefficient $R^2 \approx 1$. So, our regression yields that $\sigma_U^2(S_n)$ is in $O(n^{3.078995})$, which we consider consistent with the $O(n^3)$ figure given in (1.6). Fig. 2.12 plots $\ln(\sigma_U^2(S_n))$ as a function of $\ln(n)$, together with the corresponding regression line (thick) and the line defined by (1.6) (dashed).

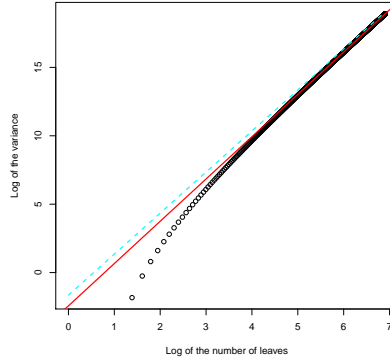


Figure 2.12: Log-log plot of $\sigma_U^2(S_n)$. The thick red line represents the linear regression of the log-log points, and the dashed light blue line is the log-log graph of the cubic defined by (1.6).

2.7 On the covariance of Φ and S under the uniform model

The covariance between Φ and S under the Yule model was also computed in [23] (see Corollary 8 therein):

$$\text{Cov}_Y(S_n, \Phi_n) = 4n(nH_n^{(2)} + H_n) + \frac{1}{6}n(n^2 - 51n + 2).$$

In this section we are interested in this covariance under the uniform model. Our contribution is, again, a recurrence that allows to compute recurrently $E_U(S_n \cdot \Phi_n)$. These values, together with the knowledge of $E_U(\Phi_n)$ and $E_U(S_n)$, can be used then to compute $\text{Cov}_U(S_n, \Phi_n)$ and then Pearson's correlation $\rho_U(S_n, \Phi_n)$. The key lemma, that plays in this section the role of Lemma 2.19 in the previous two sections, is the following:

Lemma 2.40. *Let I and J be two recursive shape indices for bifurcating phylogenetic trees as in Lemma 2.19. Let I_n and J_n be, respectively, the random variables that choose a tree $T \in \mathcal{T}_n$ and compute $I(T)$ and $J(T)$. Then, for every $n \geq 2$,*

$$\begin{aligned} E_U(I_n J_n) = \sum_{k=1}^{n-1} C_{k,n-k} & \left(2E_U(I_k J_k) + 2E_U(I_k)E_U(J_{n-k}) \right. \\ & + 2f_J(k, n-k)E_U(I_k) + 2f_I(k, n-k)E_U(J_k) \\ & \left. + f_I(k, n-k)f_J(k, n-k) \right) \end{aligned}$$

$$\text{where } C_{k,n-k} = \frac{1}{2} \binom{n}{k} \frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}.$$

Proof. Remember from the proof of Lemma 2.30 that, for every $T_k \in \mathcal{BT}(S_k)$, with $S_k \subsetneq [n]$ of cardinality k , and for every $T'_{n-k} \in \mathcal{BT}_{[n] \setminus S_k}$,

$$P_U(T_k \star T'_{n-k}) = \frac{2C_{k,n-k}}{\binom{n}{k}} P_U(T_k) P_U(T'_{n-k}).$$

We develop now $E_U(I_n \cdot J_n)$, for $n \geq 2$, as we did with $E_U(I_n^2)$ in the proof of Lemma 2.30:

$$\begin{aligned} E_U(I_n \cdot J_n) &= \sum_{T \in \mathcal{BT}_n} I(T) J(T) \cdot P_U(T) \\ &= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subsetneq [n] \\ |S_k|=k}} \sum_{T_k \in \mathcal{BT}(S_k)} \sum_{T'_{n-k} \in \mathcal{BT}(S_k^c)} I(T_k \star T'_{n-k}) J(T_k \star T'_{n-k}) P_U(T_k \star T'_{n-k}) \\ &= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k}) + f_I(k, n-k)) \\ &\quad \cdot (J(T_k) + J(T'_{n-k}) + f_J(k, n-k)) \frac{2C_{k,n-k}}{\binom{n}{k}} P_U(T_k) P_U(T'_{n-k}) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \sum_{T_k \in \mathcal{BT}_k} \sum_{T'_{n-k} \in \mathcal{BT}_{n-k}} (I(T_k) + I(T'_{n-k}) + f_I(k, n-k)) \\ &\quad \cdot (J(T_k) + J(T'_{n-k}) + f_J(k, n-k)) P_U(T_k) P_U(T'_{n-k}) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \sum_{T_k} \sum_{T'_{n-k}} \left(I(T_k) J(T_k) + I(T'_{n-k}) J(T'_{n-k}) + I(T_k) J(T'_{n-k}) \right. \\ &\quad \left. + I(T'_{n-k}) J(T_k) + f_J(k, n-k) I(T_k) + f_J(k, n-k) I(T'_{n-k}) \right. \\ &\quad \left. + f_I(k, n-k) J(T_k) + f_I(k, n-k) J(T'_{n-k}) \right. \\ &\quad \left. + f_I(k, n-k) f_J(k, n-k) \right) P_U(T_k) P_U(T'_{n-k}) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \left(\sum_{T_k} \sum_{T'_{n-k}} I(T_k) J(T_k) P_U(T_k) P_U(T'_{n-k}) \right. \\ &\quad \left. + \sum_{T_k} \sum_{T'_{n-k}} I(T'_{n-k}) J(T'_{n-k}) P_U(T_k) P_U(T'_{n-k}) \right. \\ &\quad \left. + \sum_{T_k} \sum_{T'_{n-k}} I(T_k) J(T'_{n-k}) P_U(T_k) P_U(T'_{n-k}) \right. \\ &\quad \left. + \sum_{T_k} \sum_{T'_{n-k}} I(T'_{n-k}) J(T_k) P_U(T_k) P_U(T'_{n-k}) \right. \\ &\quad \left. + \sum_{T_k} \sum_{T'_{n-k}} f_J(k, n-k) I(T_k) P_U(T_k) P_U(T'_{n-k}) \right) \end{aligned}$$

$$\begin{aligned}
& + \sum_{T_k} \sum_{T'_{n-k}} f_J(k, n-k) I(T'_{n-k}) P_U(T_k) P_U(T'_{n-k}) \\
& + \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k) J(T_k) P_U(T_k) P_U(T'_{n-k}) \\
& + \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k) J(T'_{n-k}) P_U(T_k) P_U(T'_{n-k}) \\
& + \sum_{T_k} \sum_{T'_{n-k}} f_I(k, n-k) f_J(k, n-k) P_U(T_k) P_U(T'_{n-k}) \Big) \\
= & \sum_{k=1}^{n-1} C_{k,n-k} \left[\sum_{T_k} I(T_k) J(T_k) P_U(T_k) + \sum_{T'_{n-k}} I(T'_{n-k}) J(T'_{n-k}) P_U(T'_{n-k}) \right. \\
& + \left(\sum_{T_k} I(T_k) P_U(T_k) \right) \left(\sum_{T'_{n-k}} J(T'_{n-k}) P_U(T'_{n-k}) \right) \\
& + \left(\sum_{T'_{n-k}} I(T'_{n-k}) P_U(T'_{n-k}) \right) \left(\sum_{T_k} J(T_k) P_U(T_k) \right) \\
& + f_J(k, n-k) \sum_{T_k} I(T_k) P_U(T_k) + f_J(k, n-k) \sum_{T'_{n-k}} I(T'_{n-k}) P_U(T'_{n-k}) \\
& + f_I(k, n-k) \sum_{T_k} J(T_k) P_U(T_k) + f_I(k, n-k) \sum_{T'_{n-k}} J(T'_{n-k}) P_U(T'_{n-k}) \\
& \left. + f_I(k, n-k) f_J(k, n-k) \right] \\
= & \sum_{k=1}^{n-1} C_{k,n-k} \left(E_U(I_k J_k) + E_U(I_{n-k} J_{n-k}) + E_U(I_k) E_U(J_{n-k}) \right. \\
& + E_U(I_{n-k}) E_U(J_k) + f_J(k, n-k) E_U(I_k) + f_J(k, n-k) E_U(I_{n-k}) \\
& + f_I(k, n-k) E_U(J_k) + f_I(k, n-k) E_U(J_{n-k}) \\
& \left. + f_I(k, n-k) f_J(k, n-k) \right) \\
= & \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(I_k J_k) + 2E_U(I_k) E_U(J_{n-k}) + 2f_J(k, n-k) E_U(I_k) \right. \\
& \left. + 2f_I(k, n-k) E_U(J_k) + f_I(k, n-k) f_J(k, n-k) \right)
\end{aligned}$$

using the symmetry of $C_{k,n-k}$, $f_I(k, n-k)$, and $f_J(k, n-k)$.

□

Proposition 2.41. *For every $n \geq 2$,*

$$E_U(\Phi_n \cdot S_n) = 2 \sum_{k=1}^{n-1} C_{k,n-k} E_U(\Phi_k \cdot S_k) + \frac{2^{n-5} n! (13n^2 - 9n - 2)}{(2n-3)!!} - \binom{n}{2} (5n-2)$$

Proof. We apply the last lemma taking as I and J the total cophenetic index Φ and the Sackin index S . Recall that

$$f_\Phi(k, n-k) = \binom{k}{2} + \binom{n-k}{2}, \quad E_U(\Phi_k) = \frac{1}{2} \binom{k}{2} \left(\frac{(2k-2)!!}{(2k-3)!!} - 2 \right)$$

$$f_S(k, n-k) = n, \quad E_U(S_k) = k \left(\frac{(2k-2)!!}{(2k-3)!!} - 1 \right)$$

We obtain

$$\begin{aligned} E_U(\Phi_n S_n) &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(\Phi_k S_k) + n \binom{k}{2} \left(\frac{(2k-2)!!}{(2k-3)!!} - 2 \right) \right. \\ &\quad \left. + \binom{k}{2} \left(\frac{(2k-2)!!}{(2k-3)!!} - 2 \right) (n-k) \left(\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} - 1 \right) \right. \\ &\quad \left. + 2 \left(\binom{k}{2} + \binom{n-k}{2} \right) k \left(\frac{(2k-2)!!}{(2k-3)!!} - 1 \right) + n \left(\binom{k}{2} + \binom{n-k}{2} \right) \right) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \left(2E_U(\Phi_k S_k) + n \binom{k}{2} + n \binom{n-k}{2} - 4k \binom{k}{2} \right. \\ &\quad \left. - 2k \binom{n-k}{2} + (n-k) \binom{k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \right. \\ &\quad \left. - 2(n-k) \binom{k}{2} \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} + 3k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right. \\ &\quad \left. + 2k \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right) \end{aligned}$$

Now, using the symmetry of $C_{k,n-k}$ we have that

$$\begin{aligned} &\sum_{k=1}^{n-1} C_{k,n-k} \left(n \binom{k}{2} + n \binom{n-k}{2} - 4k \binom{k}{2} - 2k \binom{n-k}{2} \right) \\ &= \sum_{k=1}^{n-1} C_{k,n-k} \left(n \binom{k}{2} + n \binom{k}{2} - 4k \binom{k}{2} - 2(n-k) \binom{k}{2} \right) \\ &= -2 \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} k \end{aligned}$$

and that

$$\begin{aligned}
& \sum_{k=1}^{n-1} C_{k,n-k} \left(-2(n-k) \binom{k}{2} \frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} + 3k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right. \\
& \quad \left. + 2k \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \left(-2k \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} + 3k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right. \\
& \quad \left. + 2k \binom{n-k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right) = 3 \sum_{k=1}^{n-1} C_{k,n-k} k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!}
\end{aligned}$$

and therefore

$$\begin{aligned}
E_U(\Phi_n S_n) &= 2 \sum_{k=1}^{n-1} C_{k,n-k} E_U(\Phi_k S_k) + \sum_{k=1}^{n-1} C_{k,n-k} \left(3k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} \right. \\
& \quad \left. - 2 \binom{k}{2} k + (n-k) \binom{k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \right) \quad (2.14)
\end{aligned}$$

Let us compute the three sums that form the independent term in this recurrence.

Claim 2.42.

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

Proof of the Claim:

$$\begin{aligned}
& \sum_{k=1}^{n-1} C_{k,n-k} (n-k) \binom{k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \\
&= \sum_{k=1}^{n-1} \frac{n!(2k-3)!!(2n-2k-3)!!(n-k)k(k-1)(2k-2)!!(2n-2k-2)!!}{4 \cdot (2n-3)!!k!(n-k)!(2k-3)!!(2(n-k)-3)!!} \\
&= \frac{n!}{4 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{2^{k-1}(k-1)!2^{n-k-1}(n-k-1)!}{(k-2)!(n-k-1)!} \\
&= \frac{n!2^{n-4}}{(2n-3)!!} \sum_{k=1}^{n-1} (k-1) = \frac{2^{n-4} \cdot n!}{(2n-3)!!} \binom{n-1}{2}
\end{aligned}$$

Claim 2.43.

$$\sum_{k=1}^{n-1} C_{k,n-k} k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} = \binom{n}{2} \frac{2^{n-2} \cdot n!}{(2n-3)!!} - \frac{2}{3} \binom{n}{2} (2n-1).$$

Proof of the Claim:

$$\begin{aligned}
& \sum_{k=1}^{n-1} C_{k,n-k} k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} \\
&= \sum_{k=1}^{n-1} \frac{n!(2k-3)!!(2n-2k-3)!!k^2(k-1)(2k-2)!!}{4 \cdot (2n-3)!!k!(n-k)!(2k-3)!!} \\
&= \frac{n!}{4 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{(2n-2k-2)!k2^{k-1}(k-1)!}{(n-k-1)!2^{n-k-1}(k-2)!(n-k)!} \\
&= \frac{2^n \cdot n!}{4 \cdot (2n-3)!!} \sum_{k=1}^{n-1} \frac{(2n-2k-2)!k(k-1)}{2^{2n-2k}(n-k-1)!(n-k)!} \\
&= \frac{2^{n-2} \cdot n!}{(2n-3)!!} \sum_{j=1}^{n-1} \frac{(2j-2)!(n-j)(n-j-1)}{2^{2j}(j-1)!j!} = (*)
\end{aligned}$$

We shall compute now the sum

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

using the lookup algorithm. Setting

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

we have that

$$t_0 = \frac{(n-2)(n-1)}{4}, \quad t_{j+1} = \frac{(j+1/2)(j-n+3)(j+1)}{(j+2)(j-n+1)(j+1)},$$

and $t_{n-2} = 0$. So, by the lookup algorithm

$$\sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)(n-j-2)}{2^{2j+2}j!(j+1)!} = \frac{(n-2)(n-1)}{4} {}_3F_2 \left[\begin{matrix} \frac{1}{2} & 1 & 3-n \\ 2 & 1-n \end{matrix} ; 1 \right].$$

We can compute this ${}_3F_2$ hypergeometric series using the Identity (1.26):

$$\begin{aligned}
{}_3F_2 \left[\begin{matrix} 1 & \frac{1}{2} & 3-n \\ 2 & 1-n \end{matrix} ; 1 \right] &= \frac{2n}{2-n} \left(\frac{\Gamma(-n)\Gamma(-3/2)}{\Gamma(-2)\Gamma(1/2-n)} - 1 \right) \\
&= \frac{2n}{2-n} \left(\frac{\frac{2(-1)^n 4\sqrt{\pi}}{n!} \frac{3}{3}}{\frac{(-2)^n \sqrt{\pi}}{(2n-1)!!}} - 1 \right) = \frac{3 \cdot n!2^{n-3} - (2n-1)!!}{3(n-2)2^{n-4} \cdot (n-1)!}
\end{aligned}$$

(recall that the values of $\Gamma(-3/2)$, $\Gamma(1/2-n)$, and $\Gamma(-n)/\Gamma(-2)$ have already been given in pages 68–68). Then

$$\begin{aligned}
\sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)(n-j-2)}{2^{2j+2}j!(j+1)!} &= \frac{(n-2)(n-1)}{4} \cdot \frac{3 \cdot n!2^{n-3} - (2n-1)!!}{3(n-2)2^{n-4} \cdot (n-1)!} \\
&= \frac{3 \cdot n!2^{n-3} - (2n-1)!!}{3 \cdot 2^{n-2} \cdot (n-2)!}
\end{aligned}$$

and, finally,

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} &= (*) \\ &= \frac{2^{n-2} \cdot n!}{(2n-3)!!} \cdot \frac{3 \cdot n! 2^{n-3} - (2n-1)!!}{3 \cdot 2^{n-2} \cdot (n-2)!} = \binom{n}{2} \frac{2^{n-2} \cdot n!}{(2n-3)!!} - \frac{2}{3} \binom{n}{2} (2n-1). \end{aligned}$$

Remark 2.44. We want to point out here that Mathematica also knows to compute the sum

$$\sum_{j=0}^{n-2} \frac{(2j)!(n-j-1)(n-j-2)}{2^{2j+2}j!(j+1)!},$$

giving the value

$$\frac{n(n-1)(3 \cdot 2^{2n}n!^2 - 16(2n-1)(2n-2)!}{3 \cdot 2^{2n+1}n!^2},$$

which is easy to see to agree with our result for that sum.

Claim 2.45.

$$\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} k = \frac{1}{2} \binom{n}{2} n - \frac{2^{n-5}n!(3n-2)}{(2n-3)!!}$$

Proof of the Claim: Let us develop this sum:

$$\begin{aligned} \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} k &= \sum_{k=2}^{n-1} C_{k,n-k} \binom{k}{2} k = \sum_{k=2}^{n-1} \frac{n!(2k-3)!!(2n-2k-3)!!k^2(k-1)}{4 \cdot (2n-3)!!k!(n-k)!} \\ &= \frac{n!}{4 \cdot (2n-3)!!} \sum_{k=2}^{n-1} \frac{(2k-2)!(2n-2k-2)!k}{2^{k-1}(k-1)!2^{n-k-1}(n-k-1)!(k-2)!(n-k)!} \\ &= \frac{n!}{2^n \cdot (2n-3)!!} \sum_{k=2}^{n-1} \frac{(2k-2)!(2n-2k-2)!k}{(k-1)!(n-k-1)!(k-2)!(n-k)!} \\ &= \frac{n!}{2^n \cdot (2n-3)!!} \sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{(j+1)!(n-j-3)!j!(n-j-2)!} = (*) \end{aligned}$$

We shall compute now the sum

$$\sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{(j+1)!(n-j-3)!j!(n-j-2)!}$$

using an argument similar to the one developed in the proof of Claim 2.35.

Let

$$t_j = \frac{(2j+2)!(2n-2j-6)!(j+2)}{(j+1)!(n-j-3)!j!(n-j-2)!}$$

Then

$$t_0 = \frac{4 \cdot (2n-6)!}{(n-3)!(n-2)!}, \quad \frac{t_{j+1}}{t_j} = \frac{(j+3/2)(j+3)(j-n+2)}{(j-n+7/2)(j+2)(j+1)}$$

and $t_{n-2} = 0$. Then, in principle, the lookup algorithm seems to imply that

$$\begin{aligned} & \sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{(j+1)!(n-j-3)!j!(n-j-2)!} \\ &= \frac{4 \cdot (2n-6)!}{(n-3)!(n-2)!} \cdot {}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right] \end{aligned}$$

but it is wrong for the same reason as a similar identity was false in the proof of Claim 2.35: since

$${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right] = \sum_{k=0}^{\infty} \frac{(\frac{3}{2})_k (3)_k (2-n)_k}{(2)_k (\frac{7}{2}-n)_k k!}$$

and $(2-n)_k = 0$ for every $k \geq n-1$, it turns out that

$${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right] = \sum_{k=0}^{n-2} \frac{(\frac{3}{2})_k (3)_k (2-n)_k}{(2)_k (\frac{7}{2}-n)_k k!}$$

while the upper limit of our original sum was $n-3$. Therefore, the correct conclusion of the lookup algorithm is

$$\begin{aligned} & \sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{(j+1)!(n-j-3)!j!(n-j-2)!} \\ &= \frac{4 \cdot (2n-6)!}{(n-3)!(n-2)!} \left({}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right] - \frac{(\frac{3}{2})_{n-2} (3)_{2-n} (2-n)_{n-2}}{(2)_{2-n} (\frac{7}{2}-n)_{n-2} (n-2)!} \right) \end{aligned}$$

where, as we saw in the proof of Claim 2.35,

$$\begin{aligned} \left(\frac{3}{2}\right)_{n-2} &= \frac{(2n-3)!!}{2^{n-2}}, \quad (3)_{n-2} = \frac{n!}{2}, \quad (2-n)_{n-2} = (-1)^{n-2} (n-2)!, \\ \left(\frac{7}{2}-n\right)_{n-2} &= (-1)^{n-3} \frac{(2n-7)!!}{2^{n-2}} \end{aligned}$$

and

$$(2)_{n-2} = 2 \cdot 3 \cdots (2+n-3) = (n-1)!$$

Thus,

$$\begin{aligned} & \frac{(\frac{3}{2})_{n-2} (3)_{n-2} (2-n)_{n-2}}{(2)_{n-2} (\frac{7}{2}-n)_{n-2} (n-2)!} \\ &= \frac{(2n-3)!! n! (-1)^n (n-2)! 2^{n-2}}{2^{n-2} 2 \cdot (n-1)! (-1)^{n-1} (2n-7)!! (n-2)!} = -\frac{1}{2} (2n-3)(2n-5)n \end{aligned}$$

As far as the hypergeometric series ${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right]$ goes, we compute it using Expression (1.25):

$${}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right] = \frac{\Gamma(2)\Gamma(\frac{7}{2}-n)\Gamma(-1)}{\Gamma(\frac{3}{2})\Gamma(2)\Gamma(1-n)} {}_3F_2 \left[\begin{matrix} \frac{1}{2} & 2-n & -1 \\ 2 & 1-n \end{matrix} ; 1 \right]$$

In this expression we already know by (1.10), (1.13), and (1.14) that

$$\Gamma(2) = 1, \quad \Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}, \quad \Gamma\left(\frac{7}{2}-n\right) = \frac{(-1)^{n-3}2^{n-3}\sqrt{\pi}}{(2n-7)!!},$$

and we can compute $\Gamma(-1)/\Gamma(1-n)$ using (1.15) and taking limits:

$$\begin{aligned} \frac{\Gamma(-1)}{\Gamma(1-n)} &= \lim_{x \rightarrow 0} \frac{\Gamma(-1+x)}{\Gamma(1-n+x)} = \lim_{x \rightarrow 0} \frac{-(1+O(x))/x}{(-1)^{n-1}(1+O(x))/((n-1)!x)} \\ &= \lim_{x \rightarrow 0} \frac{-(n-1)!(1+O(x))}{(-1)^{n-1}(1+O(x))} = (-1)^n(n-1)! \end{aligned}$$

Finally, since $(-1)_k = 0$ for every $k \geq 2$,

$${}_3F_2 \left[\begin{matrix} \frac{1}{2} & 2-n & -1 \\ 2 & 1-n \end{matrix} ; 1 \right] = 1 + \frac{(1/2)_1(2-n)_1(-1)_1}{(2)_1(1-n)_1!} = \frac{3n-2}{4(n-1)}$$

Combining all these values we obtain

$$\begin{aligned} &{}_3F_2 \left[\begin{matrix} \frac{3}{2} & 3 & 2-n \\ 2 & \frac{7}{2}-n \end{matrix} ; 1 \right] \\ &= \frac{(-1)^{n-3}2^{n-3}\sqrt{\pi}(-1)^n(n-1)!2}{(2n-7)!!\sqrt{\pi}} \cdot \frac{3n-2}{4(n-1)} = -\frac{2^{n-4}(n-2)!(3n-2)}{(2n-7)!!} \end{aligned}$$

and hence

$$\begin{aligned} &\sum_{j=0}^{n-3} \frac{(2j+2)!(2n-2j-6)!(j+2)}{(j+1)!(n-j-3)!j!(n-j-2)!} \\ &= \frac{4 \cdot (2n-6)!}{(n-3)!(n-2)!} \left(\frac{1}{2}(2n-3)(2n-5)n - \frac{2^{n-4}(n-2)!(3n-2)}{(2n-7)!!} \right) \\ &= \frac{(2n-3)!n}{(n-2)!^2} - 2^{2n-5}(3n-2) \end{aligned}$$

from where we finally obtain

$$\begin{aligned} &\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} k = (*) = \\ &= \frac{n!}{2^n \cdot (2n-3)!!} \left(\frac{(2n-3)!n}{(n-2)!^2} - 2^{2n-5}(3n-2) \right) \\ &= \frac{1}{2} \binom{n}{2} n - \frac{2^{n-5}n!(3n-2)}{(2n-3)!!} \end{aligned}$$

It is time to return to the expression for $E_U(\Phi_n \cdot S_n)$ given by equation (2.14). Its independent term turns out to be

$$\begin{aligned}
& 3 \sum_{k=1}^{n-1} C_{k,n-k} k \binom{k}{2} \frac{(2k-2)!!}{(2k-3)!!} - 2 \sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{2} k \\
& \quad + \sum_{k=1}^{n-1} C_{k,n-k} (n-k) \binom{k}{2} \frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!} \\
& = 3 \left(\binom{n}{2} \frac{2^{n-2} \cdot n!}{(2n-3)!!} - \frac{2}{3} \binom{n}{2} (2n-1) \right) \\
& \quad - 2 \left(\frac{1}{2} \binom{n}{2} n - \frac{2^{n-5} n! (3n-2)}{(2n-3)!!} \right) + \frac{2^{n-4} \cdot n!}{(2n-3)!!} \binom{n-1}{2} \\
& = \frac{2^{n-5} n! (13n^2 - 9n - 2)}{(2n-3)!!} - \binom{n}{2} (5n-2)
\end{aligned}$$

which finally proves the identity in the statement. \square

Again, we have not been able to derive from the recurrence in the last proposition an explicit formula for $E_U(S_n \cdot \Phi_n)$, but it allows to compute recurrently this expected value, starting with the obvious initial condition $E_U(S_1 \cdot \Phi_1) = 0$, and then to use these values to compute the covariance of S_n and Φ_n under the uniform model by means of

$$Cov_U(S_n, \Phi_n) = E_U(S_n \Phi_n) - E_U(S_n) E_U(\Phi_n).$$

We have computed these covariances up to $n = 1000$. Table 2.3 gives the values of $Cov_U(S_n, \Phi_n)$ for $n = 4, \dots, 20$. As in the previous sections, the R script used to compute them and the rest of the values obtained is available on the GitHub repository [97] and in Appendix A.4.5. And, again, to double-check our recurrence, we have computed the values of $Cov_U(S_n, \Phi_n)$, for $n = 3, \dots, 8$, from the Sackin and total cophenetic indices of all trees in the corresponding \mathcal{BT}_n , and they agree with the figures given by our recurrence. The R scripts used in these computations is also available in Appendix A.4.5 and on the GitHub repository [97].

n	3	4	5	6	7	8
$Cov_U(S_n, \Phi_n)$	0	0.3200	1.9184	6.5805	17.0441	37.0899
n	9	10	11	12	13	14
$Cov_U(S_n, \Phi_n)$	71.6117	126.6718	209.5473	328.7683	494.1515	716.8288
n	15	16	17	18	19	20
$Cov_U(S_n, \Phi_n)$	1009.273	1385.318	1860.185	2450.493	3174.282	4051.022

Table 2.3: $Cov_U(S_n, \Phi_n)$ for $n = 3, \dots, 20$.

We have estimated the main order in the expansion of $Cov_U(S_n, \Phi_n)$ as a function of n , by performing the minimum squares linear regression of

$\ln(\text{Cov}_U(S_n, \Phi_n))$ as a function of $\ln(n)$ for $n = 900, \dots, 1000$, and the result has been

$$\ln(\text{Cov}_U(S_n, \Phi_n)) \approx -3.133400 + 4.070915 \cdot \ln(n),$$

with a determination coefficient $R^2 \approx 1$. We conclude then that, according to our approximations, $\text{Cov}_U(S_n, \Phi_n)$ is in $O(n^{4.070915})$, and we conjecture that, actually, $\text{Cov}_U(S_n, \Phi_n)$ is in $O(n^4)$, the order of $E_U(S_n)E_U(\Phi_n)$. Fig. 2.13 displays $\ln(\text{Cov}_U(S_n, \Phi_n))$ as a function of $\ln(n)$, together with the corresponding regression line.

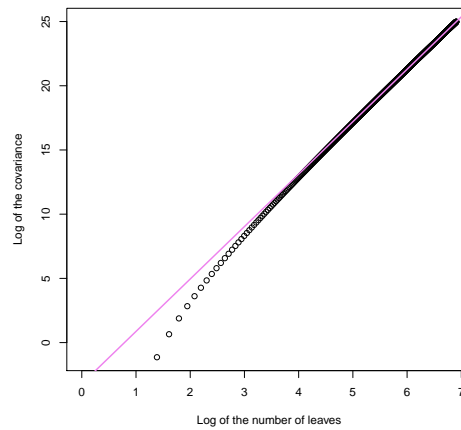


Figure 2.13: Log-log plot of $\text{Cov}_U(S_n, \Phi_n)$.

Now, since we know how to compute recurrently $\text{Cov}_U(S_n, \Phi_n)$, $\sigma_U^2(S_n)$ and $\sigma_U^2(\Phi_n)$, we can compute Pearson's correlation ρ of S and Φ under the uniform model for any desired $n \geq 4$, by means of

$$\rho_U(S_n, \Phi_n) = \frac{\text{Cov}_U(S_n, \Phi_n)}{\sigma_U(S_n) \cdot \sigma_U(\Phi_n)}.$$

Table 2.4 gives the values of $\rho_U(S_n, \Phi_n)$ for $n = 4, \dots, 20$.

n	4	5	6	7	8	9
$\rho_U(S_n, \Phi_n)$	1	0.99685	0.99450	0.99264	0.99113	0.98986
n	10	11	12	13	14	15
$\rho_U(S_n, \Phi_n)$	0.98878	0.98783	0.98700	0.98626	0.98559	0.98499
n	16	17	18	19	20	
$\rho_U(S_n, \Phi_n)$	0.98444	0.98394	0.98347	0.98304	0.98264	

Table 2.4: $\rho_U(S_n, \Phi_n)$ for $n = 4, \dots, 20$.

2.8 Numerical experiments

We devote this section to a pair of numerical experiments related to the cophenetic index: in §2.8.1 we show numerically that Φ has less ties than Sackin's or Colless' indices, and in §2.8.2 we use Φ to test the Yule and uniform models on the bifurcating trees contained in the TreeBASE [83, 109].

2.8.1 The discriminative power of Φ

In this subsection we analyze numerically the discriminative power of Φ compared with that of Sackin's and Colless' indices, by estimating, for each one of these indices, the probability of giving the same result on trees with different shape. For simplicity, for each $I = C, S, \Phi$ we have actually estimated the probability that a pair of trees $T_1, T_2 \in \mathcal{BT}_n$ have $I(T_1) = I(T_2)$. Notice that if T_1 and T_2 do have the same shape, then all these indices must be equal on them: therefore, any difference in these probabilities must be due to pairs of trees with different shape but having the same index.

For every $n = 3, \dots, 50$ we have chosen uniformly a sample of N random pairs of trees in \mathcal{BT}_n (for $n = 3, \dots, 7$, we took $N = |\mathcal{BT}_n|$ and, for $n \geq 8$, we took $N = 3000$), and computed, for $I = C, S, \Phi$,

$$\hat{p}_n(I) = \frac{\text{number of pairs } (T_1, T_2) \text{ in the sample of } \mathcal{BT}_n \text{ s. t. } I(T_1) = I(T_2)}{N}.$$

Fig. 2.14 summarizes the results. It plots $\log(\hat{p}_n(I))$ for the three balance indices as a function of $\log(n)$. We can see that Φ has the lowest relative frequency of ties.

So, our simulation shows that the discriminative power of Φ outperforms that of Sackin's and Colless' indices. We already saw some hints of this property in the previous sections: for instance, in Theorem 2.15, where we proved that the only trees $T \in \mathcal{BT}_n$ that have minimum $\Phi(T)$ are the maximally balanced, something that is not true in general for Sackin's and Colless' indices (recall Fig. 2.7); or in the lemmas previous to their proof of that theorem, where we saw that any interchange of subtrees rooted at cousins that modifies the balance of their grandparent also modifies the value of Φ . This greater resolution power of Φ makes it a better candidate to be used to test evolutionary hypotheses.

The R script used to compute these simulations and the estimated probability is also available in Appendix A.4.6 and on the GitHub repository [97].

2.8.2 A test on TreeBASE

In this subsection we report on a simple test to check which of the Yule or uniform models is the one that better fits the TreeBASE [83, 109] using the total cophenetic index.

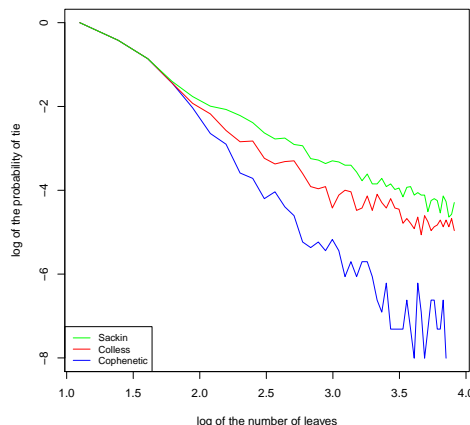


Figure 2.14: Log-log plot of the estimated probability of a tie for three balance indices.

We downloaded (December 13-14, 2015) all phylogenetic trees in the TreeBASE database [83, 109] using the function `search_treebase()` of the R package `treebase` [16]. We obtained 13,008 trees, from which 80 had format problems that prevented R from reading them, so we restricted ourselves to the remaining 12,928 trees. To simplify the language, we shall still refer here and in the next chapters to this slightly smaller subset of phylogenetic trees as “all trees in TreeBASE”. Only 4469 among these 12,928 trees in TreeBASE are rooted and bifurcating.

Now, in this experiment we have taken the numbers n of leaves for which the TreeBASE contains at least 20 bifurcating phylogenetic trees with n leaves, and for each such n we have computed the mean of the total cophenetic indices of the corresponding bifurcating trees. Fig. 2.15 plots the log of these means as a function of n . We have added the curves of the log of the expected values of Φ_n under the Yule distribution (lower curve) and under the uniform distribution (upper curve), again as a function of n . Finally, we have taken the intervals $E_Y(\Phi_n) \pm \sigma_Y(\Phi_n)$ and $E_U(\Phi_n) \pm \sigma_U(\Phi_n)$ as reference intervals for Φ_n under the Yule and the uniform model, using the formulas from Theorems 2.20 and 2.28, Formula 2.2 (from [23, Cor. 3]) and the computations for the variance under the uniform model performed in Section 2.5. This figure shows that the total cophenetic indices of the bifurcating phylogenetic trees in TreeBASE seem to be better explained by the uniform model than by the Yule model.

The R script used to compute the mean of Φ of the bifurcating trees in TreeBASE is also available in Appendix A.4.7 and on the GitHub repository [97].

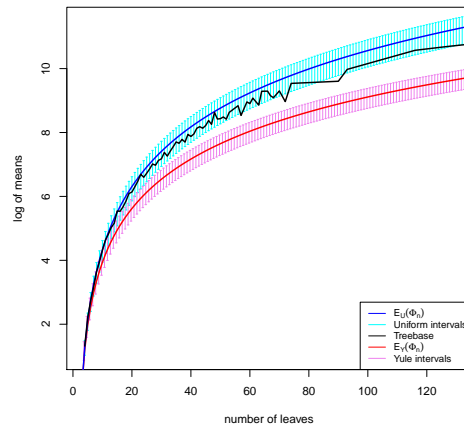


Figure 2.15: Log plots of the mean of the total cophenetic index of the bifurcating trees in TreeBASE with a fixed number n of leaves, of $E_Y(\Phi_n)$ (lower curve), $E_U(\Phi_n)$ (upper curve) and its reference intervals.

Chapter 3

The cophenetic metrics

Phylogenetic tree comparison metrics are an important tool in the study of evolution, and hence the definition of such metrics is an interesting problem in phylogenetics. In this context, Sokal and Rohlf proposed in a paper in the journal *Taxon* more than fifty years ago [102] to measure quantitatively the difference between a pair of phylogenetic trees by first encoding them by means of their half-matrices of cophenetic values and then comparing these matrices. In this chapter, we develop this idea for weighted phylogenetic trees with nested taxa. As we mentioned in §1.1, we allow these phylogenetic trees to have an unlabeled elementary root. Besides the fact that in some contexts it is not uncommon to add unlabeled elementary roots to phylogenetic trees, the main reason to allow for the root to be elementary is not for the sake of generality, but because, even if we forbid it in the statements, we would need to consider trees with elementary root in some proofs, like for instance in Theorem 3.2.

3.1 The cophenetic vectors

Recall from §1.1 (see page 16) that, given a non-empty set of taxa S , we denote by \mathcal{WT}_S the space of weighted phylogenetic trees with nested taxa on S and by \mathcal{WT}_n the space $\mathcal{WT}_{[n]}$. But notice that the subindex n in \mathcal{WT}_n does not stand for the number of leaves of the trees, but for their number of labeled nodes. As usual, to simplify the language we identify a labeled node with its taxon. Unweighted trees are understood as weighted by setting all their arcs' weights equal to 1. Then, the spaces \mathcal{UT}_n of (unweighted) phylogenetic trees with nested taxa on $[n]$ and \mathcal{T}_n of phylogenetic trees with n leaves are identified, respectively, with the subspaces of \mathcal{WT}_n consisting of all its trees without elementary root and with all their arcs' weights equal to 1, and moreover, in the case of \mathcal{T}_n , without nested taxa.

Given a tree $T \in \mathcal{WT}_S$ and a pair of taxa $i, j \in S$, the *cophenetic value* $\varphi_T(i, j)$ of i and j in T is the depth of the lowest common ancestor $[i, j]_T$ of the nodes in T labeled with i and j . If $i = j$, then $\varphi_T(i, j)$ is simply the depth

of the node labeled with i . Recall that in a weighted tree the depth of a node is the sum of the weights of the arcs in the path from the root to the node. Since $\varphi_T(i, j) = \varphi_T(j, i)$, in practice we shall only consider the phylogenetic values of pairs of taxa i, j with $i \leq j$ for some predetermined order on S (for instance, the usual order of integer numbers if $S = [n]$, the alphabetic order if the taxa in S are names, ...). We understand that this order on S defines in the usual way an alphabetic order on S^2 , where $(i, j) < (i', j')$ if $i < i'$ or if $i = i'$ and $j < j'$.

The *cophenetic vector* of $T \in \mathcal{WT}_n$ is then the vector of cophenetic values of pairs (i, j) of taxa,

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

and the *strict cophenetic vector* of T is the vector of cophenetic values of pairs (i, j) of *different* taxa:

$$\tilde{\varphi}(T) = (\varphi_T(i, j))_{1 \leq i < j \leq n} \in \mathbb{R}^{n(n-1)/2}.$$

In both cases we consider the elements of these vectors alphabetically ordered in (i, j) .

Example 3.1. If T is the unweighted phylogenetic tree with nested taxa depicted in Fig. 3.1, then $\varphi(T)$ is the vector obtained by alphabetically ordering in (i, j) the elements above the main diagonal in Table 3.1 and $\tilde{\varphi}(T)$ is obtained by alphabetically ordering in (i, j) the elements strictly above the main diagonal in that table.

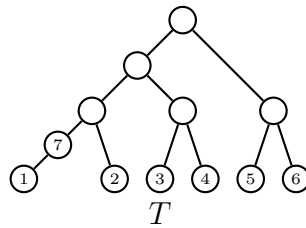


Figure 3.1: An unweighted phylogenetic tree with nested taxa.

The cophenetic vector $\varphi(T)$ of $T \in \mathcal{WT}_n$ can be computed in optimal $O(n^2)$ time (assuming a constant cost for the addition of real numbers) by computing, for each internal node v , its depth $\delta_T(v)$ through a preorder traversal of T , and the pairs of taxa of which v is the LCA through a postorder traversal of the tree.

It turns out that, for every S , the cophenetic vectors single out the members of \mathcal{WT}_S , as the following result shows.

$i \setminus j$	1	2	3	4	5	6	7
1	4	2	1	1	0	0	3
2		3	1	1	0	0	2
3			3	2	0	0	1
4				3	0	0	1
5					2	1	0
6						2	0
7							3

Table 3.1: Cophenetic values of the pairs of taxa in the tree T in Fig. 3.1.

Theorem 3.2. *For every non-empty set of taxa S and for every $T, T' \in \mathcal{WT}_S$, if $\varphi(T) = \varphi(T')$, then $T = T'$.*

Proof. Let us denote by $V(T)$ the set of nodes of a tree $T \in \mathcal{WT}_S$. We shall prove by complete induction on $|S| + |V(T)|$, i.e. on the sum of the number of labeled nodes and the total number of nodes of the tree, that a tree $T \in \mathcal{WT}_S$ can be reconstructed from its cophenetic vector $\varphi(T)$.

When $|S| + |V(T)| = 2$, the assertion is obvious, since there is only one phylogenetic tree with a single node up to its label. Assume now that the assertion is true for all trees $T' \in \mathcal{WT}_{S'}$ on any S' with $|S'| + |V(T')| < n$, and let $T \in \mathcal{WT}_S$ with $|S| + |V(T)| = n$.

If there is some $i \in S$ such that $\varphi_T(i, i) = 0$, then the root of T is labeled with i . This implies, in particular, that i is the only taxon such that $\varphi_T(i, i) = 0$. Consider then the tree $T' \in \mathcal{WT}_{S'}$ on $S' = S \setminus \{i\}$ obtained from T by removing the label i from the root. The vector $\varphi(T')$ is obtained from $\varphi(T)$ by simply removing the entries in it corresponding to pairs involving the taxon i (which are all 0), and hence it is determined by $\varphi(T)$. Then, since $|S'| + |V(T')| = |S| + |V(T)| - 1 = n - 1$, the induction hypothesis implies that T' can be reconstructed from $\varphi(T')$, and then T is obtained by labeling the root of T' with i , which is possible because this root is unlabeled since $\varphi_{T'}(j, j) > 0$ for every $j \in S'$.

If there is no $i \in S$ such that $\varphi_T(i, i) = 0$, the root r of T is unlabeled. Let $w_0 = \min\{\varphi_T(i, j) \mid i, j \in S\}$. If $w_0 = 0$ (and hence, this minimum is reached only at pairs of different taxa), then there are pairs of different taxa whose LCA is the root, and therefore the root is not elementary. In this case, consider the subtrees T_1, \dots, T_m , $m \geq 2$, rooted at the children of r . It is clear that two taxa i, j appear in two different such subtrees if, and only if, their LCA is r , that is, if, and only if, $\varphi_T(i, j) = 0$. Therefore, if we consider the equivalence relation \sim on S defined by $i \sim j$ if, and only if, $\varphi_T(i, j) > 0$, its equivalence classes will be the sets of taxa S_1, \dots, S_m of the subtrees T_1, \dots, T_m , and they are completely determined by $\varphi(T)$.

Now, for every $k = 1, \dots, m$, let $w_k = \min\{\varphi_T(i, j) \mid i, j \in S_k\}$. This value will be the weight of the arc from r to the root r_k of the subtree T_k

with set of taxa S_k . So, $\varphi(T_k)$ is obtained from $\varphi(T)$ by taking only the entries in it corresponding to pairs (i, j) with $i, j \in S_k$, and subtracting w_k from them. Then, $|S_k| + |V(T_k)| < n$ and thus, by induction, each T_k can be reconstructed from $\varphi(T_k)$. Finally, T is obtained by connecting the roots of the trees T_1, \dots, T_m obtained in this way to a new root r through arcs weighted w_1, \dots, w_m , respectively.

Assume finally that $w_0 > 0$. This means that the root of T is elementary and unlabeled, and the weight of the arc incident to it is w_0 . Let T' be the phylogenetic tree on S obtained by removing this root together with the arc incident to it. The vector $\varphi(T')$ is obtained from $\varphi(T)$ by simply subtracting w_0 to all its entries, and hence it is determined by $\varphi(T)$. Then, $|S| + |V(T')| = |S| + |V(T)| - 1 = n - 1$ and thus, by induction, T' can be reconstructed from $\varphi(T')$, and finally T is obtained from T' by adding a new elementary and unlabeled root and connecting it to the root of T' through an arc of weight w_0 . \square

A suitable adaptation of the proof of the last theorem proves that the strict cophenetic vectors single out the unweighted phylogenetic trees without nested taxa. We establish this fact in Corollary 3.3 below; we include its proof to ease the task of the reader. But, before proceeding with this result, notice that in order to single out phylogenetic trees with non constant weights in the arcs or with nested taxa, it is necessary to take into account also the depths of the leaves. Actually, for example, there is no way to reconstruct from $\tilde{\varphi}(T)$ the weights of the pendant arcs: the depths of the leaves are needed. Or, without being able to compare depths with cophenetic values, there is no way to say whether a taxon is nested or not. More specifically, for instance, the three trees in Fig. 3.2 have the same cophenetic value of $(1, 2)$, and hence the same strict cophenetic vector, but they are not isomorphic as weighted phylogenetic trees.

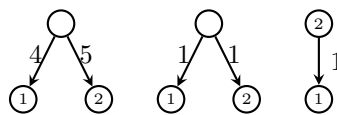


Figure 3.2: Three non-isomorphic trees with the same strict cophenetic vector.

Corollary 3.3. *For every $n \geq 1$ and for every $T, T' \in \mathcal{T}_n$, if $\tilde{\varphi}(T) = \tilde{\varphi}(T')$, then $T = T'$.*

Proof. We shall prove by complete induction on n that a tree $T \in \mathcal{T}_n$ can be reconstructed from its strict cophenetic vector $\tilde{\varphi}(T)$.

When $n = 1$, it is obvious, because \mathcal{T}_1 has only one member. Assume now that the assertion is true for all phylogenetic trees T' of any number of leaves smaller than n , and consider a tree $T \in \mathcal{T}_n$. Let T_1, \dots, T_m , $m \geq 2$, be its subtrees rooted at the children of its root r . As in the previous proof,

it is clear that two leaves i, j belong to two different such subtrees if, and only if, their LCA is r , that is, if, and only if, $\varphi_T(i, j) = 0$. Therefore, if we consider the equivalence relation \sim on S defined by $i \sim j$ if, and only if, $\varphi_T(i, j) > 0$, its equivalence classes will be the sets of taxa S_1, \dots, S_m of the subtrees T_1, \dots, T_m , and they are completely determined by $\tilde{\varphi}(T)$. Then, each $\tilde{\varphi}(T_k)$ is obtained from $\tilde{\varphi}(T)$ by taking only the entries in it corresponding to pairs (i, j) with $i, j \in S_k$, and subtracting 1 from them. By induction, each T_k can be reconstructed from $\tilde{\varphi}(T_k)$. Finally, T is obtained by connecting the roots of the trees T_1, \dots, T_m obtained in this way to a new root r . \square

3.2 The definition of the cophenetic metrics

We have proved in Theorem 3.2 that the mapping

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

that sends each $T \in \mathcal{WT}_n$ to its cophenetic vector $\varphi(T)$, is injective up to isomorphism. As it is well known, this allows to induce metrics on \mathcal{WT}_n from metrics defined on powers of \mathbb{R} . In particular, every L^p norm $\|\cdot\|_p$ on $\mathbb{R}^{n(n+1)/2}$, $p \geq 1$, induces a *cophenetic metric* $d_{\varphi,p}$ on \mathcal{WT}_n by means of

$$d_{\varphi,p}(T_1, T_2) = \|\varphi(T_1) - \varphi(T_2)\|_p, \quad T_1, T_2 \in \mathcal{WT}_n.$$

Recall that

$$\|(x_1, \dots, x_m)\|_p = \sqrt[p]{|x_1|^p + \dots + |x_m|^p},$$

and so, for instance,

$$d_{\varphi,1}(T_1, T_2) = \sum_{1 \leq i \leq j \leq n} |\varphi_{T_1}(i, j) - \varphi_{T_2}(i, j)|,$$

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

are the cophenetic metrics on \mathcal{WT}_n induced by the Manhattan L^1 and the Euclidean L^2 norms. One can also use Donoho's L^0 "norm" (which, actually, is not a proper norm)

$$\|(x_1, \dots, x_m)\|_0 = |\{i \mid i = 1, \dots, m, x_i \neq 0\}|$$

to induce a metric $d_{\varphi,0}(T_1, T_2)$ on \mathcal{WT}_n , which turns out to be simply the Hamming distance between $\varphi(T_1)$ and $\varphi(T_2)$.

Example 3.4. Consider the phylogenetic trees $T, T' \in \mathcal{T}_4$ depicted in Fig. 3.3. Their total cophenetic vectors are

$$\varphi(T) = (2, 1, 0, 0, 2, 0, 0, 2, 1, 2), \quad \varphi(T') = (1, 0, 0, 0, 2, 1, 1, 3, 2, 3)$$

and hence $d_{\varphi,1}(T, T') = 7$ and $d_{\varphi,2}(T, T') = \sqrt{7}$.

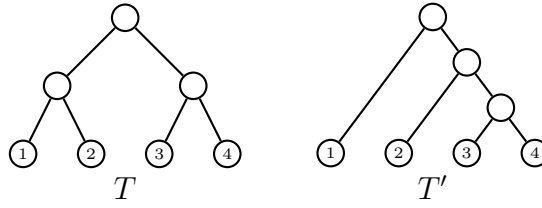


Figure 3.3: Two phylogenetic trees with 4 leaves.

As we have seen in the previous section, the cophenetic vector of a phylogenetic tree in \mathcal{WT}_n can be computed in $O(n^2)$ time. For every $T_1, T_2 \in \mathcal{WT}_n$, and assuming a constant cost for the addition and product of real numbers, the cost of computing $d_{\varphi,0}(T_1, T_2)$ (as the number of non-zero entries of $\varphi(T_1) - \varphi(T_2)$) is $O(n^2)$, and the cost of computing $d_{\varphi,p}(T_1, T_2)^p$, for $p \geq 1$ (as the sum of the p -th powers of the entries of the difference $\varphi(T_1) - \varphi(T_2)$) is $O(n^2 + \log_2(p)n^2)$, which is again $O(n^2)$ if we understand $\log(p)$ as part of the constant factor. Finally, the cost of computing $d_{\varphi,p}(T_1, T_2)$, $p \geq 1$, as the p -th root of $d_{\varphi,p}(T_1, T_2)^p$ will depend on p and on the accuracy with which this root is computed. Assuming a constant cost for the computation of p -th roots with a given accuracy (notice that, in practice, for low p and accuracy, this step will be dominated by the computation of $d_{\varphi,p}(T_1, T_2)^p$), the total cost of computing $d_{\varphi,p}(T_1, T_2)$ is $O(n^2)$.

Next examples show some features of these cophenetic metrics.

Example 3.5. Let $T \in \mathcal{WT}_n$, let (u, v) be an arc of T with u or v unlabeled and let w_0 be its weigh. Let T' be the tree in \mathcal{WT}_n obtained by contracting (u, v) : that is, by removing the node v and the arc (u, v) , labeling u with the label of v if it was labeled, and replacing every arc (v, x) in T by an arc (u, x) with its same weight. Notice that, in the passage from T to T' , for every $i, j \in S$:

- If both i, j are descendants of v in T , then $\varphi_{T'}(i, j) = \varphi_T(i, j) - w_0$.
- In any other case, $\varphi_{T'}(i, j) = \varphi_T(i, j)$.

As a consequence,

$$\varphi_T(i, j) - \varphi_{T'}(i, j) = \begin{cases} w_0 & \text{if } i, j \preceq v \\ 0 & \text{otherwise} \end{cases}$$

and therefore, if n_v is the number of descendant taxa of v ,

$$d_{\varphi,0}(T, T') = \binom{n_v + 1}{2}, \quad d_{\varphi,p}(T, T') = w_0 \sqrt[p]{\binom{n_v + 1}{2}} \text{ if } p \geq 1.$$

So the contraction of an arc in a tree T (which is Robinson-Foulds α -operation [89]) yields a new tree T' at a cophenetic distance from T that depends increasingly on the number of descendant taxa of the head of the contracted arc and, if $p \geq 1$, on the weight of the removed arc. \square

Example 3.6. Let $T_0, T'_0 \in \mathcal{WT}_m$, for some $m < n$, let $T \in \mathcal{WT}_n$ be such that its subtree rooted at some node z is T_0 , and let $T' \in \mathcal{WT}_n$ be the tree obtained by replacing in T this subtree T_0 by T'_0 . In particular, the taxa in T_0 and T'_0 are $1, \dots, m$. Then, for every $i, j \in [n]$ with $i \leq j$ and $i \leq m$,

$$\varphi_T(i, j) = \begin{cases} \delta_T(z) + \varphi_{T_0}(i, j) & \text{if } i, j \leq m \\ \varphi_T(z, j) & \text{if } i \leq m < j \end{cases}$$

and the same holds in T' , replacing T and T_0 by T' and T'_0 , respectively. Since, moreover, $\delta_T(z) = \delta_{T'}(z)$, $\varphi_T(z, j) = \varphi_{T'}(z, j)$ for every $j > m$, and $\varphi_T(i, j) = \varphi_{T'}(i, j)$ for every $i, j > m$, we conclude that

$$\varphi(T) - \varphi(T') = \varphi(T_0) - \varphi(T'_0)$$

and hence

$$d_{\varphi, p}(T, T') = d_{\varphi, p}(T_0, T'_0).$$

So, the cophenetic metrics are local, as other popular metrics like the Robinson-Foulds metric [88, 89] or the rooted triples metric [34], but unlike other popular metrics, like for instance the classical nodal metrics for bifurcating trees [41, 112] or the splitted nodal metrics for weighted multifurcating trees [21].

3.3 Minimum values

Our next goal is to find the smallest non-zero value of $d_{\varphi, p}$ on several spaces of phylogenetic trees, and the pairs of trees at which it is reached. These pairs of trees at minimum distance can be understood as “adjacent” in the corresponding metric space, and their characterization yields a first step towards understanding how cophenetic metrics measure the difference between two trees.

Notice that this problem makes no sense for weighted phylogenetic trees. For instance, if we add or subtract an $\varepsilon > 0$ to the weight of a pendant arc in a tree T , without changing its topology, the distance between T and the resulting tree will be ε , which can be as small as desired (or 1 when $p = 0$, which is its smallest possible value). So, we only consider this problem on \mathcal{UT}_n , \mathcal{T}_n , and \mathcal{BT}_n .

In order to simplify the statements, set

$$D_p(T_1, T_2) = \begin{cases} d_{\varphi, 0}(T_1, T_2) & \text{if } p = 0 \\ d_{\varphi, p}(T_1, T_2)^p & \text{if } p \geq 1 \end{cases}$$

It is clear that, for each p , the maximum and minimum values of D_p and $d_{\varphi, p}$ are reached at exactly the same pairs of trees. Therefore, in the rest of this section we shall focus on these D_p .

The following simple result will be used in the proof of the next propositions.

Lemma 3.7. *Let $p \geq 1$. If $D_p(T_1, T_2) = D_0(T_1, T_2)$ for every pair of different trees T_1, T_2 in \mathcal{UT}_n , \mathcal{T}_n or \mathcal{BT}_n such that $D_0(T_1, T_2)$ is minimum on its corresponding space of trees, then the minimum non-zero value of D_p on this space of trees is equal to the minimum non-zero value of D_0 on it, and it is reached at exactly the same pairs of trees.*

Proof. Let $T_1, T_2 \in \mathcal{UT}_n$ (respectively, in \mathcal{T}_n or \mathcal{BT}_n) such that $D_0(T_1, T_2)$ is minimum and let $p \geq 1$. Then, for every $T'_1, T'_2 \in \mathcal{UT}_n$ (respectively, in \mathcal{T}_n or \mathcal{BT}_n) we have that

$$D_p(T'_1, T'_2) \geq D_0(T'_1, T'_2) \geq D_0(T_1, T_2) = D_p(T_1, T_2)$$

where the second inequality holds by the assumption on T_1, T_2 and last equality holds by the statement's hypothesis. As to the first inequality, notice that if $\varphi_{T'_1}(i, j) \neq \varphi_{T'_2}(i, j)$, then $|\varphi_{T'_1}(i, j) - \varphi_{T'_2}(i, j)| \geq 1$ and therefore

$$\begin{aligned} D_p(T'_1, T'_2) &= \sum_{\substack{(i,j) \text{ s.t.} \\ \varphi_{T'_1}(i,j) \neq \varphi_{T'_2}(i,j)}} |\varphi_{T'_1}(i, j) - \varphi_{T'_2}(i, j)|^p \\ &\geq \sum_{\substack{(i,j) \text{ s.t.} \\ \varphi_{T'_1}(i,j) \neq \varphi_{T'_2}(i,j)}} 1^p \\ &= |\{(i, j) \mid \varphi_{T'_1}(i, j) \neq \varphi_{T'_2}(i, j)\}| = D_0(T'_1, T'_2) \end{aligned}$$

This shows that $D_p(T'_1, T'_2) \geq D_p(T_1, T_2)$ for every $T'_1, T'_2 \in \mathcal{UT}_n$ (respectively, in \mathcal{T}_n or \mathcal{BT}_n), as we claimed. \square

We can proceed now with the detection of the least non-zero values of D_p , for $p \in \{0\} \cup [1, \infty[$, on \mathcal{UT}_n , \mathcal{T}_n , and \mathcal{BT}_n , together with an explicit description of the pairs of trees where these minimum values are reached. We begin with \mathcal{UT}_n .

Proposition 3.8. *The minimum non-zero value of D_p on \mathcal{UT}_n , for $p \in \{0\} \cup [1, \infty[$ and $n \geq 2$, is 1. And for every $T, T' \in \mathcal{UT}_n$, $D_p(T, T') = 1$ if, and only if, one of them is obtained from the other by contracting a pendant arc.*

Proof. By Lemma 3.7, it is enough to prove that the minimum non-zero value of D_0 is 1, that it is reached exactly at the pairs of trees with one of them obtained from the other by contracting a pendant arc, and that for every such pair of trees $T, T' \in \mathcal{UT}_n$ we have that $D_p(T, T') = 1$ for every $p \geq 1$.

By Example 3.5, if we contract in a tree $T \in \mathcal{UT}_n$ a pendant arc with unlabeled source, we obtain a new tree T' such that $D_p(T, T') = 1$, for every $p \in \{0\} \cup [1, \infty[$, and this is of course the smallest possible non-zero value of D_p on \mathcal{UT}_n . It remains to prove that this is the only way we can obtain a pair of trees such that $D_0(T, T') = 1$.

So, let $T, T' \in \mathcal{UT}_n$ be such that $D_0(T, T') = 1$, that is, such that $\varphi(T)$ and $\varphi(T')$ differ in only one entry. Without any loss of generality, assume

that $\varphi(T) = \varphi(T') + m \cdot e_{i,j}$ for some $m \geq 1$ and $1 \leq i, j \leq n$ (where $e_{i,j}$ stands for the vector of length $n(n+1)/2$ with all entries 0 except an 1 in the entry corresponding to the pair (i, j)); that is, T and T' are such that $\varphi_T(i, j) = \varphi_{T'}(i, j) + m$, for some $m \geq 1$, and $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j)$. Let us prove first of all that $m = 1$. So, assume that $m \geq 2$ and let us reach a contradiction.

Since $\varphi_T(i, j) > 0$, there exists some taxon $k \neq i, j$ that is a descendant in T of the parent of $[i, j]_T$. In other words, such that $[i, k]_T = [j, k]_T$ is the parent of $[i, j]_T$. But then

$$\begin{aligned}\varphi_{T'}(i, k) &= \varphi_T(i, k) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j) + (m - 1) > \varphi_{T'}(i, j) \\ \varphi_{T'}(j, k) &= \varphi_T(j, k) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j) + (m - 1) > \varphi_{T'}(i, j)\end{aligned}$$

which cannot hold simultaneously: if $\varphi_{T'}(i, k) > \varphi_{T'}(i, j)$, then $\varphi_{T'}(j, k) = \varphi_{T'}(i, j)$. This shows that $m = 1$, and thus $\varphi(T) = \varphi(T') + e_{i,j}$.

Let us prove now that it cannot happen that $i \neq j$. Indeed, assume that $i \neq j$. If $\varphi_{T'}(i, j) = \delta_{T'}(i)$, then

$$\varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \delta_{T'}(i) + 1 = \delta_T(i) + 1,$$

which is impossible. This implies that $\varphi_{T'}(i, j) < \delta_{T'}(i), \delta_{T'}(j)$. If, now, $\varphi_{T'}(i, j) < \delta_{T'}(i) - 1$, then there will exist some leaf k such that $[i, k]_{T'}$ is the child of $[i, j]_{T'}$ in the path from $[i, j]_{T'}$ to i . Then $\varphi_{T'}(i, k) = \varphi_{T'}(i, j) + 1$ and $\varphi_{T'}(j, k) = \varphi_{T'}(i, j)$, which entail that

$$\varphi_T(i, k) = \varphi_{T'}(i, k) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j) > \varphi_{T'}(i, j) = \varphi_{T'}(j, k) = \varphi_T(j, k),$$

which is also impossible. So, if $i \neq j$, the only possibility is that $\varphi_{T'}(i, j) = \delta_{T'}(i) - 1 = \delta_{T'}(j) - 1$, but then it would imply that $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \delta_T(i) = \delta_T(j)$ and hence that $[i, j]_T = i = j$, which is again impossible.

So, if $\varphi(T) = \varphi(T') + e_{i,j}$ then it must happen that $i = j$. In this case, moreover, i must be a leaf in T with unlabeled parent. Indeed, if i is not a leaf, then there is some leaf k such that $i = [i, k]_T$ and hence $\delta_T(i) = \varphi_T(i, k)$. Then, $\delta_{T'}(i) = \delta_T(i) - 1 = \varphi_T(i, k) - 1 = \varphi_{T'}(i, k) - 1$, which is impossible. So, i is a leaf in T . And if its parent is labeled, say with l , then $\delta_T(i) = \delta_T(l) + 1$ and $\delta_T(l) = \varphi_T(i, l)$. Thus, in T' , $\delta_{T'}(i) = \delta_T(i) - 1 = \delta_T(l) = \delta_{T'}(l)$ and $\delta_{T'}(i) = \delta_T(l) = \varphi_T(i, l) = \varphi_{T'}(i, l)$, which is also impossible, since it would imply that $[i, l]_{T'} = i = l$.

So, finally, it must happen that i is a leaf in T and its parent is not labeled. Let T_0 be the phylogenetic tree obtained from T by contracting the pendant arc ending in i . Then $\varphi(T_0) = \varphi(T) - e_{i,i} = \varphi(T')$, and this implies, by Theorem 3.2, that $T_0 = T'$.

This finishes the proof that the only pairs of trees $T, T' \in \mathcal{WT}_n$ such that $D_0(T, T') = 1$ are those where one of them is obtained from the other by the contraction of a pendant arc with unlabeled source node. Since these pairs of trees also satisfy that $D_p(T, T') = 1$ for every $p \geq 1$, this completes the proof of the proposition. \square

So, not every tree in \mathcal{UT}_n has neighbors at cophenetic distance 1: only those trees with some leaf whose parent is unlabeled. Now, it is not difficult to check that a tree $T \in \mathcal{UT}_n$ such that all its leaves have labeled parents has some tree T' such that $D_p(T, T') = 2$, which is the minimum value of D_p on \mathcal{UT}_n greater than 1. One such T' is obtained by choosing a pendant arc in T and interchanging the labels of its source and its target nodes. Thus, by exchanging taxa i and j , where i was the parent of j in T , we have that $\varphi_T(i, i) = \varphi_{T'}(i, i) + 1$ and $\varphi_T(i, j) = \varphi_{T'}(j, j) - 1$ and then $D_p(T, T') = |1| + |-1| = 2$.

Let us consider now the space \mathcal{T}_n of usual multifurcating phylogenetic trees with n leaves, and in particular without nested taxa.

Proposition 3.9. *The minimum non-zero value of D_p on \mathcal{T}_n , for $p \in \{0\} \cup [1, \infty[$ and $n \geq 3$, is 3. And for every $T, T' \in \mathcal{T}_n$, $D_p(T, T') = 3$ if, and only if, one of them is obtained from the other by means of one of the following two operations:*

- (a) *Contracting an arc ending in the parent of a cherry (see Fig. 3.4)*
- (b) *Pruning a leaf that is a sibling of the root of a cherry and pending it from the root of the cherry (see Fig. 3.5)* □

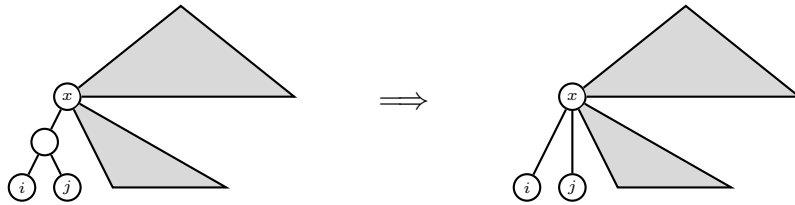


Figure 3.4: Contraction of an arc ending in the parent of a cherry.

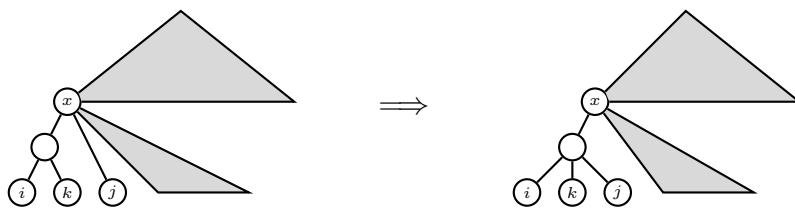


Figure 3.5: Pruning and regrafting a leaf that is a sibling of a cherry to make it a sibling of the leaves in the cherry.

We have split the proof of this proposition into Lemmas 3.10 to 3.16. But before starting with them, notice that if $n \geq 3$, there are always pairs of trees $T, T' \in \mathcal{T}_n$ such that $D_p(T, T') = 3$ for every $p \in \{0\} \cup [1, \infty[$: for instance, by Example 3.5, when T' is obtained from T by contracting an arc ending in the root of a cherry. So, the minimum non-zero value of $D_p(T, T')$ on \mathcal{T}_n is at most 3.

Lemma 3.10. *If $T, T' \in \mathcal{T}_n$ are such that $D_0(T, T') > 0$, then there exists a pair of different taxa $i \neq j$ such that $\varphi_T(i, j) \neq \varphi_{T'}(i, j)$.*

Proof. If $\varphi_T(i, j) = \varphi_{T'}(i, j)$ for every $i \neq j$, then, by Corollary 3.3, $T = T'$ and therefore $D_0(T, T') = 0$. \square

So, every pair of phylogenetic trees in \mathcal{T}_n at non-zero D_0 distance must have a pair of different leaves with different cophenetic values.

Lemma 3.11. *Let $T, T' \in \mathcal{T}_n$ be such that $\varphi_T(i, j) = \varphi_{T'}(i, j) + m$, for some $1 \leq i < j \leq n$ and some $m \geq 1$. Let $k \neq i, j$ be a leaf such that there exists a path from $[i, j]_{T'}$ to $[i, k]_{T'}$ of length l , for some $l \geq 1$. Then:*

(a) *If $\varphi_T(i, k) = \varphi_{T'}(i, k)$, then $\varphi_T(j, k) \geq \varphi_{T'}(j, k) + \min\{m, l\}$*

(b) *If $\varphi_T(j, k) = \varphi_{T'}(j, k)$, then $\varphi_T(i, k) = \varphi_{T'}(i, k) - l$*

Proof. From the assumptions we have $\varphi_{T'}(i, k) = \varphi_{T'}(i, j) + l = \varphi_{T'}(j, k) + l$. Now:

(a) Assume that $\varphi_T(i, k) = \varphi_{T'}(i, k)$. Then,

$$\varphi_T(i, k) = \varphi_{T'}(i, k) = \varphi_{T'}(i, j) + l = \varphi_T(i, j) - (m - l),$$

and then

• If $m > l$, then $\varphi_T(i, k) < \varphi_T(i, j)$, that is, $[i, j]_T \prec [i, k]_T$, and thus

$$\varphi_T(j, k) = \varphi_T(i, k) = \varphi_{T'}(i, k) = \varphi_{T'}(j, k) + l.$$

• If $m = l$, then $\varphi_T(i, k) = \varphi_T(i, j)$, that is, $[i, k]_T = [i, j]_T$, and thus

$$\varphi_T(j, k) \geq \varphi_T(i, j) = \varphi_{T'}(i, j) + m = \varphi_{T'}(j, k) + m.$$

• If $m < l$, then $\varphi_T(i, k) > \varphi_T(i, j)$, that is, $[i, k]_T \prec [i, j]_T$, and thus

$$\varphi_T(j, k) = \varphi_T(i, j) = \varphi_{T'}(i, j) + m = \varphi_{T'}(j, k) + m.$$

(b) Assume that $\varphi_T(j, k) = \varphi_{T'}(j, k)$. Then

$$\varphi_T(j, k) = \varphi_{T'}(j, k) = \varphi_{T'}(i, j) = \varphi_T(i, j) - m,$$

so that $[i, j]_T \prec [j, k]_T$, and thus

$$\varphi_T(i, k) = \varphi_T(j, k) = \varphi_{T'}(j, k) = \varphi_{T'}(i, j) = \varphi_{T'}(i, k) - l.$$

\square

Lemma 3.12. *Let $T, T' \in \mathcal{T}_n$ be such that $\varphi_T(i, j) = \varphi_{T'}(i, j) + m$, for some $1 \leq i < j \leq n$ and some $m \geq 1$. Let N be the number of leaves k such that $k \neq i, j$ and either $[i, k]_{T'} \prec [i, j]_{T'}$ or $[j, k]_{T'} \prec [i, j]_{T'}$. Then,*

$$D_0(T, T') \geq N + 1.$$

Proof. If $k \neq i, j$ is such that $[i, k]_{T'} \prec [i, j]_{T'}$, so that $\varphi_{T'}(i, k) = \varphi_{T'}(i, j) + l$ for some $l \geq 1$, then, by the previous lemma, $\varphi_T(i, k) \neq \varphi_{T'}(i, k)$ or $\varphi_T(j, k) \neq \varphi_{T'}(j, k)$, and thus this leaf k contributes at least 1 to $D_0(T, T')$. By symmetry, any leaf k such that $[j, k]_{T'} \prec [i, j]_{T'}$ also contributes at least 1 to $D_0(T, T')$. Therefore, if there are N such leaves, they contribute at least N to $D_0(T, T')$. Finally, the pair (i, j) contributes 1 to $D_0(T, T')$ by assumption. We conclude that $D_0(T, T') \geq N + 1$. \square

Lemma 3.13. *Let $T, T' \in \mathcal{T}_n$ be such that $D_0(T, T') \leq 3$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + m$, for some $1 \leq i < j \leq n$ and some $m \geq 1$, then $m = 1$.*

Proof. If $\delta_{T'}(i) = \delta_T(i)$, then $\delta_{T'}(i) = \delta_T(i) > \varphi_T(i, j) = \varphi_{T'}(i, j) + m$ which implies that there are at least m leaves k such that $[i, k]_{T'} \prec [i, j]_{T'}$. Then, by the last lemma, $D_0(T, T') \geq m + 1$. Now, if $\delta_{T'}(j) = \delta_T(j)$, then for the same reason there are at least m leaves k such that $[j, k]_{T'} \prec [i, j]_{T'}$ and they increase $D_0(T, T')$ to at least $2m + 1$, while if $\delta_{T'}(j) \neq \delta_T(j)$, then $D_0(T, T') \geq m + 2$. We conclude then that if $\delta_{T'}(i) = \delta_T(i)$, then $m = 1$. By symmetry, if $\delta_{T'}(j) = \delta_T(j)$, then $m = 1$, either.

Finally, if $\delta_{T'}(i) \neq \delta_T(i)$ and $\delta_{T'}(j) \neq \delta_T(j)$, and since $\varphi_T(i, j) \neq \varphi_{T'}(i, j)$, we have that $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, i), (j, j), (i, j)$. Let now $k \neq i, j$ be a taxon such that $[i, k]_T = [j, k]_T$ is the parent of $[i, j]_T$ in T . Then

$$\varphi_{T'}(i, k) = \varphi_T(i, k) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j) + (m - 1)$$

and therefore, if $m \geq 2$, $\varphi_{T'}(i, k) > \varphi_{T'}(i, j)$ and then, by Lemma 3.11, either $\varphi_T(i, k) \neq \varphi_{T'}(i, k)$ or $\varphi_T(j, k) \neq \varphi_{T'}(j, k)$, which, as we have seen, is impossible. Thus, $m = 1$ in all cases. \square

Lemma 3.14. *Let $T, T' \in \mathcal{T}_n$ be such that $D_0(T, T') \leq 3$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, then $(\delta_{T'}(i) - \varphi_{T'}(i, j)) + (\delta_{T'}(j) - \varphi_{T'}(i, j)) \leq 3$.*

Proof. Let us assume that $(\delta_{T'}(i) - \varphi_{T'}(i, j)) + (\delta_{T'}(j) - \varphi_{T'}(i, j)) \geq 4$ and let us reach a contradiction.

Assume first that $\delta_{T'}(i) \geq \varphi_{T'}(i, j) + 3$. Then, there are at least two leaves k_1, k_2 such that $[i, k_1]_{T'}, [i, k_2]_{T'} \prec [i, j]_{T'}$. Since, by Lemma 3.11, each such leaf contributes at least 1 to $D_0(T, T') \leq 3$, we conclude that there must be exactly two such leaves and, moreover, $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k_1), (j, k_1), (i, k_2), (j, k_2)$. But then, on the one hand, $\delta_T(j) = \delta_{T'}(j)$ and, on the other hand, $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$ (otherwise, there would be some

other leaf k such that $[j, k]_{T'} \prec [i, j]_{T'}$, which, again by Lemma 3.11 would satisfy that $\varphi_T(i, k) \neq \varphi_{T'}(i, k)$ or $\varphi_T(j, k) \neq \varphi_{T'}(j, k)$. Combining these two equalities we obtain $\delta_T(j) = \varphi_T(i, j)$, which is impossible in a tree without nested taxa. This proves that $\delta_{T'}(i) \leq \varphi_{T'}(i, j) + 2$ and, by symmetry, that $\delta_{T'}(j) \leq \varphi_{T'}(i, j) + 2$, as we claimed.

Thus, it remains to prove that the case $\delta_{T'}(i) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 2$ is impossible. So, assume this case holds, and let's reach a contradiction. By Lemma 3.12, if $D_0(T, T') \leq 3$ and $\delta_{T'}(i) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 2$, then there can exist only one extra leaf k pending from the parent of i and one extra leaf l pending from the parent of j : see Fig. 3.6, where the grey triangle stands for the (possibly empty) subtree consisting of all other descendants of $[i, j]_{T'}$. Moreover, since $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$ and since both k and l contribute at least 1 to $D_0(T, T') \leq 3$, we conclude that $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k), (j, k), (i, l), (j, l)$. In particular

$$\begin{aligned} \varphi_T(k, l) &= \varphi_{T'}(k, l) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 \\ \delta_T(i) &= \delta_{T'}(i) = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1 \\ \delta_T(j) &= \delta_T(k) = \delta_T(l) = \varphi_T(i, j) + 1 \text{ for the same reason} \end{aligned}$$

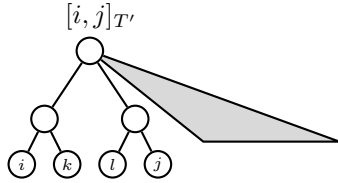


Figure 3.6: The subtree of T' rooted at $[i, j]_{T'}$ in the proof of Lemma 3.14.

Now we shall prove that, in this situation, each one of k, l contributes actually at least 2 to $D_0(T, T')$, and therefore $D_0(T, T') \geq 5$, which contradicts the assumption that $D_0(T, T') \leq 3$. We distinguish several cases.

(1) Assume that $\varphi_T(i, k) = \varphi_{T'}(i, k)$. Then, by Lemmas 3.11.(a) and 3.13, $\varphi_T(j, k) = \varphi_{T'}(j, k) + 1$, and hence

$$\begin{aligned} \varphi_T(i, k) &= \varphi_{T'}(i, k) = \varphi_{T'}(j, k) + 1 = \varphi_T(j, k) \\ \varphi_T(i, k) &= \varphi_{T'}(i, k) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j) \\ \delta_T(i) &= \delta_T(j) = \delta_T(k) = \delta_T(l) = \varphi_T(i, j) + 1 \\ \varphi_T(k, l) &= \varphi_T(i, j) - 1 \end{aligned}$$

Thus, the subtree of T rooted at $[k, l]_T$ contains a subtree of the form described in Fig. 3.7, for at least one leaf h . But then

$$\varphi_{T'}(l, h) = \varphi_T(l, h) = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \varphi_{T'}(l, j)$$

which is impossible, since it would imply that h is another descendant of $[l, j]_{T'}$. Therefore, $\varphi_T(i, k) \neq \varphi_{T'}(i, k)$ and, by symmetry, $\varphi_T(j, l) \neq \varphi_{T'}(j, l)$.

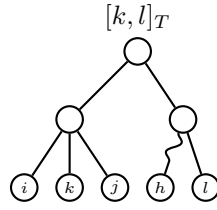


Figure 3.7: A subtree of the subtree of T rooted at $[k, l]_T$ in case (1) in the proof of Lemma 3.14.

(2) Assume now that $\varphi_T(i, l) = \varphi_{T'}(i, l)$. Then, by Lemma 3.11.(b), $\varphi_T(j, l) = \varphi_{T'}(j, l) - 1$, and then

$$\begin{aligned} \varphi_T(i, l) &= \varphi_{T'}(i, l) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 \\ \varphi_T(j, l) &= \varphi_{T'}(j, l) - 1 = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 \\ \varphi_T(k, l) &= \varphi_T(i, j) - 1 \\ \delta_T(i) &= \delta_T(j) = \delta_T(k) = \delta_T(l) = \varphi_T(i, j) + 1 \end{aligned}$$

Therefore, the subtree of T rooted at $[k, l]_T$ contains a subtree of the form described in Fig. 3.8, for at least one leaf h . Moreover, $h \neq k$ because $\varphi_T(h, l) > \varphi_T(j, l) = \varphi_T(k, l)$. But then, again,

$$\varphi_{T'}(l, h) = \varphi_T(l, h) = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \varphi_{T'}(l, j)$$

which is impossible by the same reason as in (1). Therefore, $\varphi_T(i, l) \neq \varphi_{T'}(i, l)$ and, by symmetry, $\varphi_T(j, k) \neq \varphi_{T'}(j, k)$.

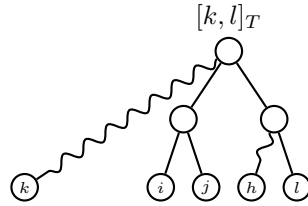


Figure 3.8: A subtree of the subtree of T rooted at $[k, l]_T$ in case (2) in the proof of Lemma 3.14.

So, $\varphi_T(i, k) \neq \varphi_{T'}(i, k)$, $\varphi_T(i, l) \neq \varphi_{T'}(i, l)$, $\varphi_T(j, k) \neq \varphi_{T'}(j, k)$, $\varphi_T(j, l) \neq \varphi_{T'}(j, l)$, and $\varphi_T(i, j) \neq \varphi_{T'}(i, j)$ by assumption, and thus $D_0(T, T') \geq 5$. \square

Summarizing the last lemmas, we have proved so far that if $D_0(T, T') \leq 3$ and $\varphi_T(i, j) \neq \varphi_{T'}(i, j)$, then, up to interchanging T and T' , $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$ and either i and j are sibling in T' or one of these leaves is a sibling of the parent of the other one in T' . Next two lemmas cover these two remaining cases.

Lemma 3.15. *Let $T, T' \in \mathcal{T}_n$ be such that $D_0(T, T') \leq 3$, and assume that $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$. If i and j are sibling in T' , then they are also sibling in T , they have no other sibling in T , and T' is obtained from T by contracting the arc ending in $[i, j]_T$. And then, $D_0(T, T') = 3$.*

Proof. If $\delta_{T'}(i) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 1$, then it must happen that $\delta_T(i) = \delta_{T'}(i) + 1$ and $\delta_T(j) = \delta_{T'}(j) + 1$. Indeed, if $\delta_T(i) \leq \delta_{T'}(i)$, then $\delta_T(i) \leq \varphi_{T'}(i, j) + 1 = \varphi_T(i, j)$, which is impossible. Therefore, $\delta_T(i) > \delta_{T'}(i)$ and by symmetry $\delta_T(j) > \delta_{T'}(j)$. Since $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $D_0(T, T') \leq 3$ implies that $\varphi_T(x, y) = \varphi_{T'}(x, y)$, for every $(x, y) \neq (i, j), (i, i), (j, j)$. Now, if, say $\delta_T(i) \geq \delta_{T'}(i) + 2$, then

$$\delta_T(i) \geq \delta_{T'}(i) + 2 = \varphi_{T'}(i, j) + 3 = \varphi_T(i, j) + 2$$

and there would exist some leaf k such that $[i, k]_T$ is a child of $[i, j]_T$. But then

$$\varphi_{T'}(i, k) = \varphi_T(i, k) = \varphi_T(i, j) + 1 = \varphi_{T'}(i, j) + 2 = \delta_{T'}(i) + 1,$$

which is impossible. This proves that $\delta_T(i) = \delta_{T'}(i) + 1$ and, by symmetry, $\delta_T(j) = \delta_{T'}(j) + 1$.

So, in summary, $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $\delta_T(i) = \delta_{T'}(i) + 1$, $\delta_T(j) = \delta_{T'}(j) + 1$ and $\varphi_T(x, y) = \varphi_{T'}(x, y)$, for every $(x, y) \neq (i, j), (i, i), (j, j)$, and in particular $D_0(T, T') = 3$.

Now, $\delta_T(i) = \delta_{T'}(i) + 1 = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1$, and by symmetry, $\delta_T(j) = \varphi_T(i, j) + 1$, either. Therefore, i and j are sibling in T . Let us see that they have no other sibling in this tree. Indeed, if k is a sibling of i and j in T , then

$$\varphi_{T'}(i, k) = \varphi_T(i, k) = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \delta_{T'}(i)$$

which is impossible.

Let x be the parent of $[i, j]_T$, and assume that the subtree T_0 of T rooted at x is as described in Fig. 3.9.(a), for some subtree \widehat{T} . Moreover, let T'_0 be the (possibly empty) subtree of T' rooted at $[i, j]_{T'}$, which is as described in Fig. 3.9.(b) for some subtree \widehat{T}' . We shall prove that $\widehat{T} = \widehat{T}'$.

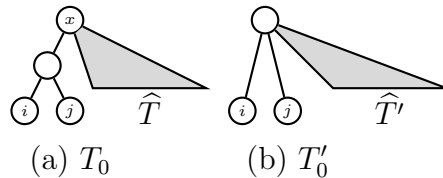


Figure 3.9: (a) The subtree T_0 of T rooted at the parent of $[i, j]_T$ in the proof of Lemma 3.15. (b) The subtree T'_0 of T' rooted at $[i, j]_{T'}$ in the proof of the same Lemma.

For every $k \in L(\widehat{T})$,

$$\varphi_{T'}(i, k) = \varphi_T(i, k) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j),$$

which entails that $k \in L(\widehat{T}')$. Conversely, if $k \in L(\widehat{T}')$, then

$$\varphi_T(i, k) = \varphi_{T'}(i, k) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1,$$

which entails that $k \in L(\widehat{T})$. Thus, $L(\widehat{T}) = L(\widehat{T}')$. And finally, for every (not necessarily different) $k, l \in L(\widehat{T})$,

$$\begin{aligned} \varphi_{\widehat{T}}(k, l) &= \varphi_T(k, l) - \delta_T(x) = \varphi_T(k, l) - \varphi_T(i, j) + 1 \\ &= \varphi_{T'}(k, l) - \varphi_{T'}(i, j) = \varphi_{\widehat{T}'}(k, l), \end{aligned}$$

which implies by Theorem 3.2 that $\widehat{T} = \widehat{T}'$ (notice that \widehat{T} and \widehat{T}' can have elementary roots).

Finally, let us prove now that T and T' are exactly the same except for T_0 and T'_0 . More specifically, let T_1 and T'_1 be obtained by replacing in T and T' the subtrees T_0 and T'_0 by a single leaf x . Since for every $p, q \notin L(T_0) = L(T'_0)$,

$$\begin{aligned} \varphi_{T'_1}(p, q) &= \varphi_{T'}(p, q) = \varphi_T(p, q) = \varphi_{T_1}(p, q), \\ \varphi_{T'_1}(x, p) &= \varphi_{T'}(i, p) = \varphi_T(i, p) = \varphi_{T_1}(p, x), \end{aligned}$$

we deduce, again by Theorem 3.2, that $T_1 = T'_1$.

This completes the proof that T' is obtained from T by replacing in it the subtree T_0 rooted at the parent x of $[i, j]_T$ by the subtree T'_0 obtained from T_0 by contracting the arc $(x, [i, j]_T)$. \square

Lemma 3.16. *Let $T, T' \in \mathcal{T}_n$ be such that $D_0(T, T') \leq 3$. Assume that $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, and that j is a sibling of the parent of i in T' . Then, the subtree of T' rooted at $[i, j]_{T'}$ is the tree T'_0 depicted in Fig. 3.10.(a), for some taxon $k \neq i, j$ and some (possibly empty) subtree \widehat{T}' , and T is obtained from T' by replacing T'_0 by the tree T_0 depicted in Fig. 3.10.(b). And then, $D_0(T, T') = 3$.*

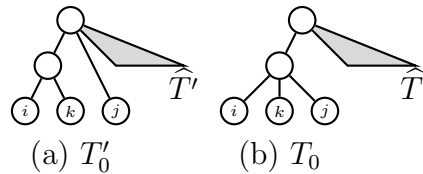


Figure 3.10: (a) The subtree T'_0 of T' rooted at $[i, j]_{T'}$ in the statement of Lemma 3.16. (b) The subtree T_0 which replaces T'_0 in T in the same statement.

Proof. We assume that $\delta_{T'}(i) = \varphi_{T'}(i, j) + 2$ and $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$. This implies that there exists at least one leaf k such that $[i, k]_{T'} \prec [i, j]_{T'}$. Now, we have that $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $\delta_T(j) > \delta_{T'}(j)$ (because, otherwise, $\delta_T(j) \leq \delta_{T'}(j) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j)$, which is impossible), and

$$|\varphi_T(i, k) - \varphi_{T'}(i, k)| + |\varphi_T(j, k) - \varphi_{T'}(j, k)| \geq 1$$

(because if $\varphi_T(i, k) = \varphi_{T'}(i, k)$ and $\varphi_T(j, k) = \varphi_{T'}(j, k)$, then $\varphi_T(i, k) = \varphi_{T'}(i, k) > \varphi_{T'}(j, k) = \varphi_T(j, k)$ would imply $\varphi_T(i, j) = \varphi_T(k, j) = \varphi_{T'}(j, k) = \varphi_{T'}(i, j)$, against the assumption that $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$). Then, the assumption $D_0(T, T') \leq 3$ entails that $\varphi_T(i, k) = \varphi_{T'}(i, k)$ or $\varphi_T(j, k) = \varphi_{T'}(j, k)$, and that $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k), (j, k), (j, j)$ (and, in particular, k is the only leaf different from i such that $[i, k]_{T'} \prec [i, j]_{T'}$). Moreover, we have that $D_0(T, T') = 3$.

Let us see now that $\delta_T(j) = \delta_{T'}(j) + 1$. Indeed, if $\delta_T(j) \geq \delta_{T'}(j) + 2$, then

$$\delta_T(j) \geq \delta_{T'}(j) + 2 = \varphi_{T'}(i, j) + 3 = \varphi_T(i, j) + 2$$

and there would exist some leaf l such that $[j, l]_T$ is a child of $[i, j]_T$. But then

$$\varphi_{T'}(j, l) = \varphi_T(j, l) = \varphi_T(i, j) + 1 = \varphi_{T'}(i, j) + 2 = \delta_{T'}(j) + 1$$

and we reach a contradiction.

So, in summary, the subtree T'_0 of T' rooted at $[i, j]_{T'}$ is as described in Fig. 3.10.(a), and $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $\delta_T(j) = \delta_{T'}(j) + 1$, $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k), (j, k), (j, j)$, and either $\varphi_T(i, k) = \varphi_{T'}(i, k)$ or $\varphi_T(j, k) = \varphi_{T'}(j, k)$. Now, we discuss these two possibilities.

(a) If $\varphi_T(j, k) = \varphi_{T'}(j, k)$, then $\varphi_T(i, k) = \varphi_{T'}(i, k) - 1$ by Lemma 3.11.(b). In this case

$$\begin{aligned} \varphi_T(i, k) &= \varphi_{T'}(i, k) - 1 = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 \\ \varphi_T(j, k) &= \varphi_{T'}(j, k) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 \\ \delta_T(i) &= \delta_{T'}(i) = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1 \\ \delta_T(j) &= \delta_{T'}(j) + 1 = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1 \\ \delta_T(k) &= \delta_{T'}(k) = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1 \end{aligned}$$

This means that the subtree of T rooted at $[i, k]_T = [j, k]_T$ contains a subtree of the form described in Fig. 3.11, for at least some new leaf h . But then

$$\varphi_{T'}(k, h) = \varphi_T(k, h) = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \varphi_{T'}(i, k)$$

which is impossible in T' , because i and k are the only descendants of $[i, k]_{T'}$ in T' . So, this case is impossible.

(b) If $\varphi_T(i, k) = \varphi_{T'}(i, k)$, then $\varphi_T(j, k) = \varphi_{T'}(j, k) + 1$ by Lemmas 3.11.(a) and 3.13. In this case

$$\begin{aligned} \varphi_T(i, k) &= \varphi_{T'}(i, k) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j) \\ \varphi_T(j, k) &= \varphi_{T'}(j, k) + 1 = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j) \\ \delta_T(i) &= \delta_T(j) = \delta_T(k) = \varphi_T(i, j) + 1 \text{ as in (a)} \end{aligned}$$

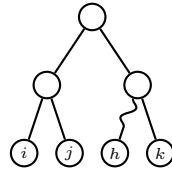


Figure 3.11: A subtree contained in the subtree of T rooted at $[i, k]_T$ in case (a) in the proof of Lemma 3.16.

This implies that i, j, k are sibling in T . If l is any other sibling of them in T , then

$$\varphi_{T'}(i, l) = \varphi_T(i, l) = \varphi_T(i, k) = \varphi_{T'}(i, k)$$

which entails that l is another descendant of $[i, k]_{T'}$ in T' , which is impossible. Therefore, the subtree T_0 of T rooted at the parent of $[i, j]_T$ has the form depicted in Fig. 3.12, for some subtree \widehat{T} .

Finally, the same argument as in the last part of the proof of the last lemma shows that $\widehat{T} = \widehat{T}'$, and that if T_1 and T'_1 are obtained by replacing in T and T' the subtrees T_0 and T'_0 by a single leaf x , then $T_1 = T'_1$. We leave the details to the reader.

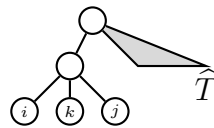


Figure 3.12: The subtree T_0 rooted at the parent of $[i, j]_T$ in case (b) in the proof of Lemma 3.16.

This completes the proof that T and T' are as described in the statement. \square

We have proved so far that the minimum value of D_0 on \mathcal{T}_n is 3, and we have characterized those pairs of trees $T, T' \in \mathcal{T}_n$ such that $D_0(T, T') = 3$. To extend this result to every D_p , $p \geq 1$, it is enough to check that every pair of trees in \mathcal{T}_n such that $D_0(T, T') = 3$ also satisfies that $D_p(T, T') = 3$ for every $p \geq 1$, which is straightforward: every pair of these trees has only three differences between their cophenetic vectors and all of them are 1 or -1 . Then, adding up their absolute values raised to the p -th power, we obtain that $D_p(T, T') = 3$. This completes the proof of Proposition 3.9.

So, every tree $T \in \mathcal{T}_n$ has neighbors T' such that $D_p(T, T') = 3$. Indeed, take the internal node v in T of largest depth, so that all its children are leaves. If v has exactly two children, one such neighbor of T is obtained by contracting the arc ending in v : see Fig. 3.13.

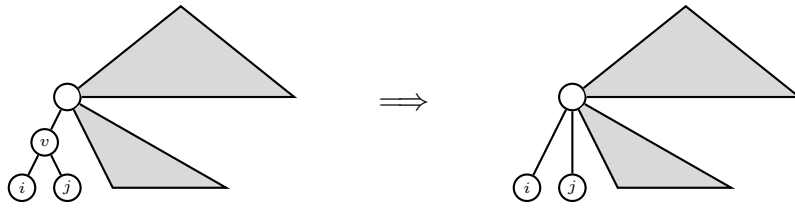


Figure 3.13: Contraction of an arc ending in v , when v has two children.

If v has more than two children, one such neighbor of T is obtained by replacing any two children of v by a cherry pending from v (that is, taking two children i, j of v , removing the arcs (v, i) and (v, j) , and then adding a new node w and arcs (v, w) , (w, i) , and (w, j)): see Fig. 3.14.

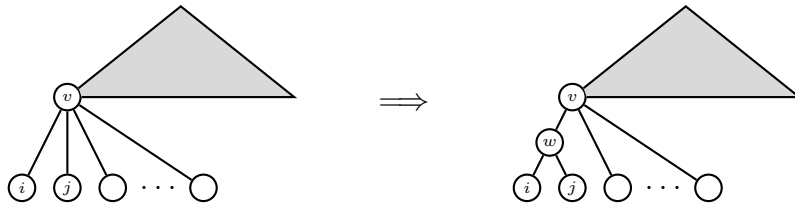


Figure 3.14: Replacing any two children of v by a cherry.

Let us consider finally the bifurcating phylogenetic trees case.

Proposition 3.17. *The minimum non-zero value of D_p on \mathcal{BT}_n , for $p \in \{0\} \cup [1, \infty[$ and $n \geq 3$, is 4. And for every $T, T' \in \mathcal{BT}_n$, $D_p(T, T') = 4$ if, and only if, one of them is obtained from the other by means of one of the following operations:*

- (a) *Reorganizing a triplet (see Fig. 3.15)*
- (b) *Reorganizing a maximally balanced quartet (see Fig. 3.16)*

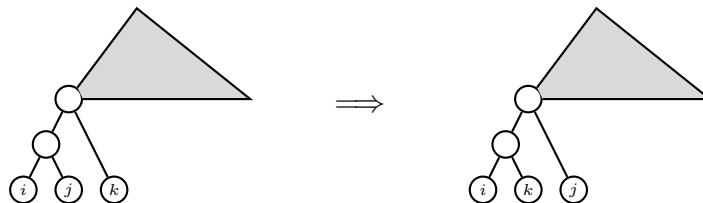


Figure 3.15: Reorganizing a triplet.

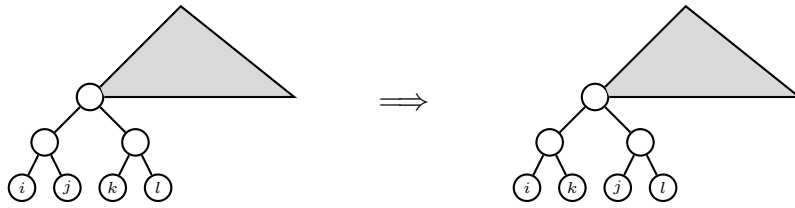


Figure 3.16: Reorganizing a maximally balanced quartet.

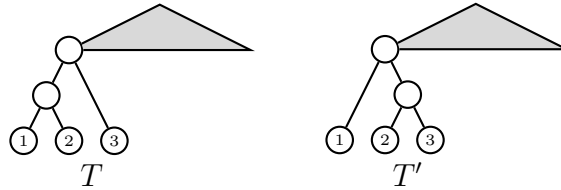


Figure 3.17: A pair of bifurcating trees such that $D_p(T, T') = 4$. The grey triangles represent the same tree.

As in Proposition 3.9, we also split this proof into several lemmas. First of all, notice that there are pairs of trees $T, T' \in \mathcal{BT}_n$ such that $D_p(T, T') = 4$ for every $p \in \{0\} \cup [1, \infty[$: see, for instance, Fig. 3.17. Therefore, the minimum value of D_p on \mathcal{BT}_n is at most 4.

Notice also that Lemma 3.10 also applies in \mathcal{BT}_n , and therefore, if $T, T' \in \mathcal{BT}_n$ are such that $D_0(T, T') > 0$, then there exist two taxa $i \neq j$ such that $\varphi_T(i, j) \neq \varphi_{T'}(i, j)$. And, of course, Lemma 3.11 also applies in \mathcal{BT}_n .

Lemma 3.18. *Let $T, T' \in \mathcal{BT}_n$ be such that $D_0(T, T') \leq 4$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + m$, for some $1 \leq i < j \leq n$ and some $m \geq 1$, then $m = 1$.*

Proof. Assume that $\varphi_T(i, j) = \varphi_{T'}(i, j) + m$ with $m \geq 2$, and let us reach a contradiction.

If $\delta_{T'}(i) = \delta_T(i)$, then $\delta_{T'}(i) > \varphi_T(i, j) = \varphi_{T'}(i, j) + m$, and therefore there exist leaves x_1, \dots, x_m such that $\varphi_T(i, x_l) = \varphi_{T'}(i, j) + l$, for $l = 1, \dots, m$. By Lemma 3.11, each such leaf x_l adds at least 1 to $D_0(T, T')$. Therefore $D_0(T, T') \geq 1 + m$. Now, if moreover $\delta_{T'}(j) = \delta_T(j)$, then there also exist leaves y_1, \dots, y_m such that $\varphi_T(j, y_l) = \varphi_{T'}(i, j) + l$, for $l = 1, \dots, m$, and each such leaf y_l also adds at least 1 to $D_0(T, T')$, which entails $D_0(T, T') \geq 1 + 2m \geq 5$. So, if $D_0(T, T') \leq 4$, it must happen that $\delta_{T'}(i) \neq \delta_T(i)$ or $\delta_{T'}(j) \neq \delta_T(j)$ (or both). Let assume that $\delta_{T'}(j) \neq \delta_T(j)$.

Now, $\varphi_T(i, j) = \varphi_{T'}(i, j) + m \geq m$, and then there exist leaves z_1, \dots, z_m such that $\varphi_T(i, z_l) = \varphi_T(j, z_l) = \varphi_T(i, j) - l$, for $l = 1, \dots, m$. If $\varphi_T(i, z_l) = \varphi_{T'}(i, z_l)$, then

$$\varphi_{T'}(i, z_l) = \varphi_T(i, z_l) = \varphi_T(i, j) - l = \varphi_{T'}(i, j) + (m - l) \geq \varphi_{T'}(i, j)$$

and therefore, by Lemma 3.11, $\varphi_{T'}(j, z_l) \neq \varphi_T(j, z_l)$, and thus, each such leaf z_l adds at least 1 to $D_0(T, T')$, which entails $D_0(T, T') \geq 2 + m$. Therefore, if $D_0(T, T') \leq 4$ and $m \geq 2$, it must happen $m = 2$ and, moreover, $\varphi_T(a, b) = \varphi_{T'}(a, b)$ for every $(a, b) \neq (i, j), (j, j), (i, z_1), (i, z_2), (j, z_1), (j, z_2)$.

In particular, $\delta_T(i) = \delta_{T'}(i)$, which as we have seen implies that there are at least two leaves x_1, x_2 such that $i \prec [i, x_2]_{T'} \prec [i, x_1]_{T'} \prec [i, j]_{T'}$. Since

$$\varphi_{T'}(z_1, z_2) = \varphi_T(z_1, z_2) = \varphi_T(i, j) - 2 = \varphi_{T'}(i, j)$$

implies that (up to interchanging z_1 and z_2) $i \prec [i, z_1]_{T'} \prec [i, j]_{T'}$ and $j \prec [j, z_2]_{T'} \prec [i, j]_{T'}$, we conclude that $\{x_1, x_2, z_1, z_2\}$ are at least 3 different leaves and hence they contribute at least 3 to $D_0(T, T')$, making $D_0(T, T') \geq 5$. \square

Lemma 3.19. *Let $T, T' \in \mathcal{BT}_n$ be such that $D_0(T, T') \leq 4$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, then $\delta_{T'}(i), \delta_{T'}(j) \leq \varphi_{T'}(i, j) + 2$.*

Proof. Let us assume that $\delta_{T'}(i) \geq \varphi_{T'}(i, j) + 3$, and let us reach a contradiction. The case when $\delta_{T'}(j) \geq \varphi_{T'}(i, j) + 3$ is symmetrical.

Since $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1 > 0$, there exists some taxon k_0 such that $[i, k_0]_T$ is the parent of $[i, j]_T$. Let us distinguish several cases.

(a) Assume first that $\varphi_T(i, k_0) = \varphi_{T'}(i, k_0)$. Then, $\varphi_{T'}(i, k_0) = \varphi_T(i, k_0) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j)$ implies that $[j, k_0]_{T'} \prec [i, j]_{T'}$ and thus $\varphi_{T'}(j, k_0) > \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 = \varphi_T(j, k_0)$ and in particular, by the previous lemma $\varphi_{T'}(j, k_0) = \varphi_T(j, k_0) + 1 = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1$. Now, since $D_0(T, T') \leq 4$, by Lemma 3.13 the number of leaves $a \neq i, j, k_0$ such that $a \prec [i, j]_{T'}$ is at most 2.

If $\delta_{T'}(i) \geq \varphi_{T'}(i, j) + 3$, then there exist leaves k_1, k_2 such that $\varphi_{T'}(i, k_1) = \varphi_{T'}(i, j) + 1$ and $\varphi_{T'}(i, k_2) = \varphi_{T'}(i, j) + 2$ and then $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k_0), (j, k_0), (k_1, i), (k_1, j), (k_2, i), (k_2, j)$. In particular, no leaf other than i, j, k_0, k_1, k_2 descends from $[i, j]_{T'}$. But then

$$\begin{aligned} \varphi_T(k_1, k_0) &= \varphi_{T'}(k_1, k_0) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1, \\ \varphi_T(k_2, k_0) &= \varphi_T(i, j) - 1, \\ \varphi_T(k_1, k_2) &= \varphi_{T'}(k_1, k_2) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j) \end{aligned}$$

imply that, up to interchanging k_1 and k_2 , $i \prec [i, k_1]_T \prec [i, j]_T$ and $j \prec [j, k_2]_T \prec [i, j]_T$, and then

$$\delta_{T'}(j) = \delta_T(j) > \varphi_T(i, j) + 1 = \varphi_{T'}(i, j) + 2$$

implies the existence of at least another leaf h such that $j \prec [j, h]_{T'} \prec [j, k_0]_{T'} \prec [i, j]_{T'}$, which, as we have mentioned, is impossible. So, this case cannot happen.

(b) Assume now that $\varphi_T(j, k_0) = \varphi_{T'}(j, k_0)$. By symmetry with the previous case, this implies that $\varphi_{T'}(i, k_0) = \varphi_{T'}(i, j) + 1$, $\varphi_{T'}(i, k_0) = \varphi_T(i, k_0) + 1$ and that the number of leaves $a \neq i, j, k_0$ such that $a \prec [i, j]_{T'}$ is at most 2. Now we have three new subcases to discuss.

- (b.1) If $\delta_{T'}(i) = \varphi_{T'}(i, j) + 4$, so that there exist leaves $k_1, k_2 \neq i$ such that $\varphi_{T'}(i, k_0), \varphi_{T'}(i, k_1), \varphi_{T'}(i, k_2) > \varphi_{T'}(i, j)$, and no leaf other than i, j, k_0, k_1, k_2 descends from $[i, j]_{T'}$, then $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k_0), (j, k_0), (k_1, i), (k_1, j), (k_2, i), (k_2, j)$. But in this case it must happen that $\delta_T(j) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j)$, which is impossible. So, this case cannot happen.
- (b.2) If $\delta_{T'}(i) = \varphi_{T'}(i, j) + 3$ and $\delta_{T'}(j) = \varphi_{T'}(i, j) + 2$, so that there exist leaves k_1, k_2 such that $\varphi_{T'}(j, k_1) = \varphi_{T'}(i, j) + 1$, $\varphi_{T'}(i, k_2) = \varphi_{T'}(i, j) + 2$ and, recall, $\varphi_{T'}(i, k_0) = \varphi_{T'}(i, j) + 1$, then $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k_0), (j, k_0), (k_1, i), (k_1, j), (k_2, i), (k_2, j)$. But then

$$\varphi_T(k_1, k_0) = \varphi_{T'}(k_1, k_0) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1$$

implies that $k_1 \prec [i, j]_T$, and then

$$\begin{aligned} \delta_T(j) &= \delta_{T'}(j) = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1, \\ \delta_T(k_1) &= \delta_{T'}(k_1) = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1 \end{aligned}$$

imply that j and k_1 are the only children of $[i, j]_T$, which is, of course, impossible. So, this case cannot happen, either.

- (b.3) If $\delta_{T'}(i) = \varphi_{T'}(i, j) + 3$ and $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$, then on the one hand there exists a leaf k_1 such that $\varphi_{T'}(i, k_1) = \varphi_{T'}(j, k_0) + 1 = \varphi_{T'}(i, j) + 2$ and, on the other hand, as we have seen in (b.1), $\delta_T(j) > \delta_{T'}(j)$. Then, $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (j, j), (i, k_0), (j, k_0), (k_1, i), (k_1, j)$, and in particular no leaf other than i, j, k_0, k_1 descends from $[i, j]_{T'}$.

Now,

$$\varphi_T(k_1, k_0) = \varphi_{T'}(k_1, k_0) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j)$$

implies that $k_1 \not\prec [i, j]_T$, and

$$\delta_T(i) = \delta_{T'}(i) = \varphi_{T'}(i, j) + 3 = \varphi_T(i, j) + 2$$

implies that there exists a leaf $h \neq k_0, k_1$ such that $i \prec [i, h]_T \prec [i, j]_T$ and hence

$$\varphi_{T'}(i, h) = \varphi_T(i, h) > \varphi_T(i, j) + 1 = \varphi_{T'}(i, j)$$

would entail that $h \prec [i, j]_{T'}$, which is impossible. Thus, this case cannot happen, either.

- (c) Assume finally that $\varphi_T(i, k_0) \neq \varphi_{T'}(i, k_0)$ and $\varphi_T(j, k_0) \neq \varphi_{T'}(j, k_0)$. The contribution to D_0 of the pairs $(i, j), (i, k_0), (j, k_0)$ is at least 3, and therefore there can only exist at most one other pair of leaves with different cophenetic value in T and in T' . Since every $x \neq i, j$ such that $x \prec [i, j]_{T'}$ defines at

least one such pair, we conclude that if $\delta_{T'}(i) \geq \varphi_{T'}(i, j) + 3$, then, it must happen that $[i, k_0]_{T'} \prec [i, j]_{T'}$ and that there can only exist one leaf $k_1 \neq k_0, i$ such that $[i, k_1]_{T'} \prec [i, j]_{T'}$, and then, moreover $[i, k_0]_{T'} \neq [i, k_1]_{T'}$. In this case, $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every $(x, y) \neq (i, j), (i, k_0), (j, k_0), (k_1, i), (k_1, j)$. But then, in particular, $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$ and $\delta_T(j) = \delta_{T'}(j)$, which implies $\delta_T(i) = \varphi_T(i, j)$, which is impossible

This finishes the proof that, if $D_0(T, T') \leq 4$, then $\delta_{T'}(i) \leq \varphi_{T'}(i, j) + 2$ and $\delta_{T'}(j) \leq \varphi_{T'}(i, j) + 2$. \square

Lemma 3.20. *Let $T, T' \in \mathcal{BT}_n$ be such that $D_0(T, T') \leq 4$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, then i, j are sibling in T .*

Proof. Let k_0 be any leaf such that $[i, k_0]_T = [j, k_0]_T$ is the parent of $[i, j]_T$ in T . If $\varphi_T(i, k_0) = \varphi_{T'}(i, k_0)$, then $\varphi_{T'}(i, k_0) = \varphi_T(i, k_0) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j)$ implies that $[j, k_0]_{T'} \prec [i, j]_{T'}$ and thus $\varphi_{T'}(j, k_0) > \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 = \varphi_T(j, k_0)$. Therefore, $|\varphi_T(i, k_0) - \varphi_{T'}(i, k_0)| + |\varphi_T(j, k_0) - \varphi_{T'}(j, k_0)| \geq 1$.

Assume now that i, j are not sibling in T , and let h be a leaf such that $[i, h]_T$ is a child of $[i, j]_T$. If $\varphi_T(i, h) \leq \varphi_{T'}(i, h)$, then

$$\delta_{T'}(i) \geq \varphi_{T'}(i, h) + 1 \geq \varphi_T(i, h) + 1 = \varphi_T(i, j) + 2 = \varphi_{T'}(i, j) + 3$$

which is impossible by the previous lemma. Therefore, $\varphi_T(i, h) > \varphi_{T'}(i, h)$, and by Lemma 3.18, $\varphi_T(i, h) = \varphi_{T'}(i, h) + 1$.

In a similar way, if $\delta_T(i) = \delta_{T'}(i)$, then

$$\delta_{T'}(i) = \delta_T(i) \geq \varphi_T(i, h) + 1 = \varphi_T(i, j) + 2 = \varphi_{T'}(i, j) + 3$$

which is again impossible by the previous lemma. Therefore, $\delta_T(i) \neq \delta_{T'}(i)$, too. So, $(i, j), (i, k_0), (j, k_0), (i, i)$, and (i, h) contribute at least 4 to $D_0(T, T')$. Since the latter is at most 4, this implies that $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every other pair of leaves (x, y) . But then,

$$\begin{aligned} \varphi_{T'}(j, h) &= \varphi_T(j, h) = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 \\ \varphi_{T'}(i, h) &= \varphi_T(i, h) - 1 = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 \end{aligned}$$

which is impossible. Therefore, i and j are sibling in T . \square

Lemma 3.21. *Let $T, T' \in \mathcal{BT}_n$ be such that $D_0(T, T') \leq 4$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, then i, j are not sibling in T' .*

Proof. Assume that i, j are sibling in T' , and recall that we already know that they are sibling in T . Let k_0 be any leaf such that $[i, k_0]_T = [j, k_0]_T$ is the parent of $[i, j]_T$ in T . If $\varphi_T(i, k_0) = \varphi_{T'}(i, k_0)$, then

$$\varphi_{T'}(i, k_0) = \varphi_T(i, k_0) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j)$$

which is impossible if i, j are sibling in T' . Thus, $\varphi_T(i, k_0) \neq \varphi_{T'}(i, k_0)$ and, by symmetry, $\varphi_T(j, k_0) \neq \varphi_{T'}(j, k_0)$. On the other hand, if $\delta_T(i) = \delta_{T'}(i)$, then

$$\delta_T(i) = \delta_{T'}(i) = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j)$$

which is also impossible. Therefore, $\delta_T(i) \neq \delta_{T'}(i)$ and, by symmetry, $\delta_T(j) \neq \delta_{T'}(j)$. But, then, $D_0(T, T') \geq 5$. \square

Summarizing what we know so far, we have proved that if $D_0(T, T') \leq 4$ and $\varphi_T(i, j) \neq \varphi_{T'}(i, j)$, with $i \neq j$, then, up to interchanging T and T' as well as i and j :

- (a) $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$,
- (b) i, j are sibling in T ,
- (c) $\delta_{T'}(i) = \varphi_{T'}(i, j) + 2$ and $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$, or $\delta_{T'}(i) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 2$.

Next two lemmas cover the two possibilities mentioned in (c).

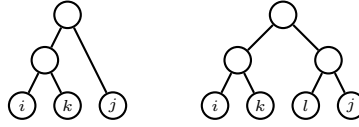


Figure 3.18: The only possibilities for the subtree of T' rooted at $[i, j]_{T'}$ if $D_0(T, T') \leq 4$ and $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$.

Lemma 3.22. *Let $T, T' \in \mathcal{BT}_n$ be such that $D_0(T, T') \leq 4$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, and moreover $\delta_{T'}(i) = \varphi_{T'}(i, j) + 2$ and $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$, then the subtree of T' rooted at $[i, j]_{T'}$ is a triplet as the one depicted in the left hand side of Fig. 3.18 and T is obtained from T' by interchanging j and the sibling of i : cf. Fig. 3.19. And, then, $D_0(T, T') = 4$.*

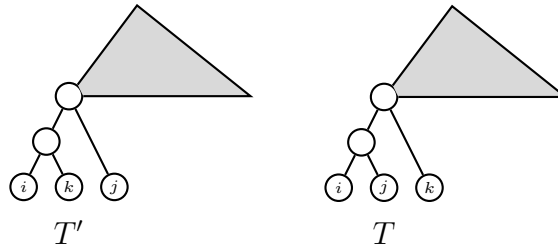


Figure 3.19: The only pairs of trees T, T' such that $D_0(T, T') \leq 4$ and $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, when the subtree of T' rooted at $[i, j]_{T'}$ is a triplet.

Proof. Assume that $D_0(T, T') \leq 4$, $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $\delta_{T'}(i) = \varphi_{T'}(i, j) + 2$, and $\delta_{T'}(j) = \varphi_{T'}(i, j) + 1$. Since i, j are sibling in T by Lemma 3.20,

$$\delta_T(j) = \varphi_T(i, j) + 1 = \varphi_{T'}(i, j) + 2 = \delta_{T'}(j) + 1.$$

Now, since $\delta_{T'}(i) = \varphi_{T'}(i, j) + 2$, there exists some leaf k such that $[i, k]_{T'}$ is a child of $[i, j]_{T'}$, being j its other child. Then, since i, j are sibling in T ,

$$\varphi_T(i, k) < \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \varphi_{T'}(i, k),$$

and therefore, by Lemma 3.18, $\varphi_T(i, k) = \varphi_{T'}(i, k) - 1 = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1$, which implies that $[i, k]_T$ is the parent of $[i, j]_T$ in T .

Since $D_0(T, T') \leq 4$, $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $\delta_T(j) = \delta_{T'}(j) + 1$ and $\varphi_T(i, k) = \varphi_{T'}(i, k) - 1$ for every leaf k such that $[i, k]_{T'}$ is a child of $[i, j]_{T'}$, there exist at most two leaves satisfying this last property. Assume for a moment that there exist two such leaves, say k_1 and k_2 . Then they must be sibling in T' and $[k_1, k_2]_{T'}$ must be a child of $[i, k_1]_{T'} = [i, k_2]_{T'}$, and $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every (x, y) other than (i, j) , (j, j) , (i, k_1) , (i, k_2) . In particular

$$\varphi_T(k_1, k_2) = \varphi_{T'}(k_1, k_2) = \varphi_{T'}(i, j) + 2 = \varphi_T(i, j) + 1 = \varphi_T(i, k_1) + 2.$$

But then there would have to exist a third leaf k_3 such that $[k_1, k_2]_T \prec [k_1, k_3]_T \prec [k_1, i]_T$ and then

$$\varphi_{T'}(i, k_3) = \varphi_T(i, k_3) = \varphi_T(i, k_1) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j)$$

which is impossible, because the only leaves that descend from $[i, j]_{T'}$ are i, j, k_1, k_2 . This leads to a contradiction, which implies that, actually, there is exactly one leaf k such that $[i, k]_{T'}$ is a child of $[i, j]_{T'}$, and hence that the subtree of T' rooted at $[i, j]_{T'}$ is a triplet as depicted in the left hand side of Fig. 3.18.

Now, if $\delta_T(k) \geq \varphi_T(i, j) + 1$, then there would exist at least some other leaf $l \prec [i, k]_T$ and arguing as above we would have that $\varphi_T(i, l) \neq \varphi_{T'}(i, l)$ ($\varphi_{T'}(i, l) = \varphi_T(i, l) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j)$) is impossible because the only leaves descending from $[i, j]_{T'}$ are i, j, k) and, by symmetry, $\varphi_T(j, l) \neq \varphi_{T'}(j, l)$, and we would reach $D_0(T, T') \geq 5$. Therefore,

$$\delta_T(k) = \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \delta_{T'}(k) - 1.$$

So, in summary, $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, $\delta_T(j) = \delta_{T'}(j) + 1$, $\varphi_T(i, k) = \varphi_{T'}(i, k) - 1$, and $\delta_T(k) = \delta_{T'}(k) - 1$, and $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every (x, y) other than (i, j) , (j, j) , (i, k) , (k, k) . Moreover, in T , k is the other child of the parent of $[i, j]_T$.

So, the subtree T_0 of T rooted at the parent of $[i, j]_T$ is obtained by interchanging j and k in the subtree T'_0 of T' rooted at $[i, j]_{T'}$. Finally, let us prove now that T and T' are exactly the same except for T_0 and T'_0 . More

specifically, let T_1 and T'_1 be obtained by replacing in T and T' the subtrees T_0 and T'_0 by a single leaf x . Since for every $p, q \notin \{i, j, k\}$,

$$\begin{aligned}\varphi_{T'_1}(p, q) &= \varphi_{T'}(p, q) = \varphi_T(p, q) = \varphi_{T_1}(p, q), \\ \varphi_{T'_1}(x, p) &= \varphi_{T'}(i, p) = \varphi_T(i, p) = \varphi_{T_1}(x, p),\end{aligned}$$

we deduce, by Theorem 3.2, that $T_1 = T'_1$.

This completes the proof that the subtree of T' rooted at $[i, j]$ is a triplet and that T is obtained from T' by interchanging the leaf j and its nephew. \square

Lemma 3.23. *Let $T, T' \in \mathcal{BT}_n$ be such that $D_0(T, T') \leq 4$. If $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, for some $1 \leq i < j \leq n$, and moreover $\delta_{T'}(i) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 2$, then the subtree of T' rooted at $[i, j]_{T'}$ is a maximally balanced quartet as the one depicted in the right hand side of Fig. 3.18 and T is obtained from T' by interchanging j and the sibling of i : cf. Fig. 3.20. And, then, $D_0(T, T') = 4$.*

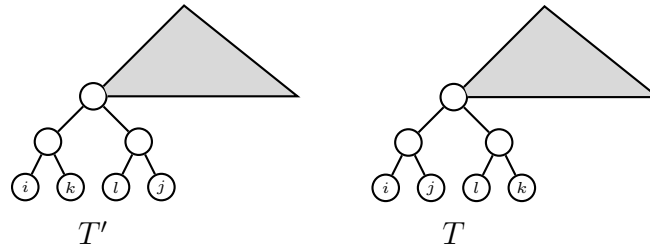


Figure 3.20: The only pairs of trees T, T' such that $D_0(T, T') \leq 4$ and $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$, when the subtree of T' rooted at $[i, j]_{T'}$ is a quartet.

Proof. Assume that $D_0(T, T') \leq 4$, $\varphi_T(i, j) = \varphi_{T'}(i, j) + 1$ and $\delta_{T'}(i) = \delta_{T'}(j) = \varphi_{T'}(i, j) + 2$. Then, there exist leaves k and l such that $[i, k]_{T'}$ and $[j, l]_{T'}$ are the children of $[i, j]_{T'}$ in T' .

Now, since i, j are sibling in T ,

$$\varphi_T(i, k) < \varphi_T(i, j) = \varphi_{T'}(i, j) + 1 = \varphi_{T'}(i, k)$$

and therefore, by Lemma 3.18, $\varphi_T(i, k) = \varphi_{T'}(i, k) - 1$, and in particular $\varphi_T(i, k) = \varphi_T(i, j) - 1$. By symmetry, $\varphi_T(j, l) = \varphi_{T'}(j, l) - 1$ and hence $\varphi_T(j, l) = \varphi_T(i, j) - 1$, too. Therefore, both k and l are descendants of the parent of $[i, j]_T$. But then,

$$\varphi_{T'}(k, l) = \varphi_{T'}(i, j) = \varphi_T(i, j) - 1 < \varphi_T(k, l)$$

and therefore, by Lemma 3.18, $\varphi_T(k, l) = \varphi_{T'}(k, l) + 1 = \varphi_{T'}(i, j) + 1 = \varphi_T(i, j)$.

At this point, $D_0(T, T') \leq 4$ entails that $\varphi_T(x, y) = \varphi_{T'}(x, y)$ for every (x, y) other than $(i, j), (i, k), (j, l), (k, l)$. Moreover, on the one hand, it also

implies that k is the only leaf such that $[i, k]_{T'}$ is a child of $[i, j]_{T'}$ and l is the only leaf such that $[j, l]_{T'}$ is a child of $[i, j]_{T'}$, because any other such leaf would contribute at least 2 more units to D_0 and we have already reached $D_0(T, T') = 4$. Therefore, the subtree of T' rooted at $[i, j]_{T'}$ is the totally balanced quartet depicted in the right hand side of Fig. 3.18. On the other hand, i, k, j, l are the only descendant leaves of the parent of $[i, j]_T$ in T . Indeed, if h is another descendant leaf of the parent of $[i, j]_T$, then

$$\varphi_{T'}(i, h) = \varphi_T(i, h) = \varphi_T(i, j) - 1 = \varphi_{T'}(i, j)$$

and therefore h would be another descendant of $[i, j]_{T'}$.

Then, since, moreover, $\varphi_T(i, k) = \varphi_T(i, j) - 1$, $\varphi_T(j, l) = \varphi_T(i, j) - 1$, and $\varphi_T(k, l) = \varphi_T(i, j)$, we conclude that the subtree T_0 of T rooted at $[i, j]_T$ is the totally balanced quartet obtained from the subtree T'_0 of T' rooted at $[i, j]_{T'}$ by interchanging j and k . Finally, arguing as in the last part of the proof of the previous lemma, we deduce that T and T' are exactly the same except for T_0 and T'_0 . \square

These two lemmas complete the proof of the fact that the minimum value of D_0 on \mathcal{BT}_n is 4 and the characterization of the pairs of trees $T, T' \in \mathcal{BT}_n$ such that $D_0(T, T') = 4$. To extend this result to every D_p with $p \geq 1$, it is enough to check that every pair of bifurcating trees such that $D_0(T, T') = 4$ also satisfies that $D_p(T, T') = 4$ for every $p \geq 1$, which is straightforward: from the characterization of the pairs of trees $T, T' \in \mathcal{BT}_n$ such that $D_0(T, T') = 4$ we deduce that their cophenetic vectors $\varphi(T)$ and $\varphi(T')$ only differ in four entries, and the differences in all four cases are ± 1 . Then, $D_p(T, T') = 4$ for every $p \geq 1$. This completes the proof of Proposition 3.17.

So again, every tree $T \in \mathcal{BT}_n$ has neighbors T' such that $D_p(T, T') = 4$. Indeed, take an internal node v in T of largest depth, so that its two children are leaves. Let w be the parent of v . Then, either the other child of w is a leaf, in which case w is the root of a triple and reorganizing its taxa we obtain a neighbor of T , or the other child of w is the parent of a cherry (it will have the same, maximum, depth as v), in which case w is the root of a maximally balanced quartet and reorganizing its taxa we obtain a neighbor of T .

3.4 Diameters

We focus now on the diameters of the cophenetic metrics, that is, the largest value of $d_{\varphi, p}$ on the different spaces of unweighted phylogenetic trees; as in the case of the minimum non-zero value, and for the same reasons, the problem of finding the diameter makes no sense for weighted trees. Unfortunately, we have not been able to find exact formulas for it, but we have obtained its order, which we give in the next result.

Theorem 3.24. *The diameter of $d_{\varphi,p}$ on \mathcal{UT}_n , \mathcal{T}_n , and \mathcal{BT}_n is in $\Theta(n^2)$ if $p = 0$ and in $\Theta(n^{(p+2)/p})$ if $p \geq 1$.*

In particular, the diameter of $d_{\varphi,1}$ on these spaces is in $\Theta(n^3)$, and the diameter of $d_{\varphi,2}$ is in $\Theta(n^2)$.

As we did with the main results in the last section, we have split the proof of this theorem into several lemmas. We consider first the case when $p = 1$, which will be used later to prove the case when $p > 1$. For every $T \in \mathcal{UT}_n$, let

$$S(T) = \sum_{i=1}^n \delta_T(i), \quad \Phi(T) = \sum_{1 \leq i < j \leq n} \varphi_T(i, j).$$

S and Φ are the extensions to \mathcal{UT}_n of the *Sackin index* (see Section 1.2) and the *total cophenetic index* (see Chapter 2) for phylogenetic trees without nested taxa, respectively. Notice that $\|\varphi(T)\|_1 = S(T) + \Phi(T)$. We have the following results on these indices:

- We know that the minimum values of S and Φ on \mathcal{T}_n are both reached at the rooted star tree with n leaves, and these minimum values are, respectively,
 - $\min S(\mathcal{T}_n) = n$
 - $\min \Phi(\mathcal{T}_n) = 0$
- It is straightforward to check that the minimum values of S and Φ on \mathcal{UT}_n are both reached at the rooted star tree with $n - 1$ leaves and labeled root, and these minimum values are, respectively,
 - $\min S(\mathcal{UT}_n) = n - 1$
 - $\min \Phi(\mathcal{UT}_n) = 0$
- The minimum values of S and Φ on \mathcal{BT}_n are both reached at the maximally balanced trees with n leaves (see [45] for the Sackin index and Theorem 2.15 for the total cophenetic index) and these minimum values are, respectively,
 - $\min S(\mathcal{BT}_n) = n(\lceil \log_2(n) \rceil + 1) - 2^{\lceil \log_2(n) \rceil}$ (see Equation (1.3))
 - $\min \Phi(\mathcal{BT}_n) = \sum_{k=0}^{n-1} a(k)$, where $a(k)$ is the highest power of 2 that divides $n!$ (see Proposition 2.18)

From the first formula it is clear that $\min S(\mathcal{BT}_n)$ is in $\Theta(n \log(n))$, and we have proved in Corollary 2.17 that $\min \Phi(\mathcal{BT}_n)$ is in $\Theta(n^2)$.

- The maximum values of S and Φ on both \mathcal{T}_n and \mathcal{BT}_n are reached at the combs with n leaves (see [45] for the Sackin index and Proposition 2.11 for the total cophenetic index), and these maximum values are, respectively,

- $\max S(\mathcal{T}_n) = \max S(\mathcal{BT}_n) = \binom{n+1}{2} - 1$ (see Equation (1.2))
- $\max \Phi(\mathcal{T}_n) = \max \Phi(\mathcal{BT}_n) = \binom{n}{3}$ (see Proposition 2.11)

So, they are in $\Theta(n^2)$ and $\Theta(n^3)$, respectively.

- Given any tree in \mathcal{UT}_n with a nested taxon, if we replace this nested taxon by a new leaf labeled with it pending from the node previously labeled with it (cf. Fig. 3.21), we obtain a new tree in \mathcal{UT}_n with strictly larger value of S and the same value of Φ . This shows that the maximum values of S and Φ on \mathcal{UT}_n are reached at trees in \mathcal{T}_n , and hence at the combs with n leaves. Therefore, they are also in $\Theta(n^2)$ and $\Theta(n^3)$, respectively.

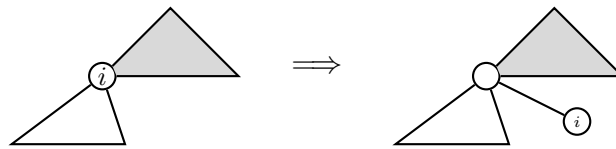


Figure 3.21: This operation increases the value of S and does not modify the value of Φ .

From these properties and the fact that $\|\varphi(T)\|_1 = S(T) + \Phi(T)$ for every $T \in \mathcal{UT}_n$, we deduce the following result.

Lemma 3.25. *The minimum value of $\|\varphi(T)\|_1$ on \mathcal{UT}_n and \mathcal{T}_n is in $\Theta(n)$. The minimum value of $\|\varphi(T)\|_1$ on \mathcal{BT}_n is in $\Theta(n^2)$. The maximum value of $\|\varphi(T)\|_1$ on \mathcal{UT}_n , \mathcal{T}_n , and \mathcal{BT}_n is in $\Theta(n^3)$. \square*

Now, we can apply this lemma to find the order of the diameter of $d_{\varphi,1}$ on the different spaces of unweighted phylogenetic trees with n leaves. To simplify the notations, from time to time in the rest of this section we shall use X_n to denote any space \mathcal{UT}_n , \mathcal{T}_n or \mathcal{BT}_n , and we shall denote, for every $p \in \{0\} \cup [0, \infty)$, the diameter of $d_{\varphi,p}$ on X_n by $\Delta_p(X_n)$, and the set of values $\|\varphi(T)\|_p$ with $T \in X_n$ by $\|\varphi(X_n)\|_p$.

Lemma 3.26. *The diameter of $d_{\varphi,1}$ on \mathcal{UT}_n , \mathcal{T}_n , and \mathcal{BT}_n is in $\Theta(n^3)$.*

Proof. Let $T_1, T_2 \in X_n$. Then, on the one hand,

$$d_{\varphi,1}(T_1, T_2) = \|\varphi(T_1) - \varphi(T_2)\|_1 \leq \|\varphi(T_1)\|_1 + \|\varphi(T_2)\|_1 \leq 2 \cdot \max \|\varphi(X_n)\|_1,$$

which is in $\Theta(n^3)$. This implies that $\Delta_1(X_n)$ is in $O(n^3)$. On the other hand, if $\|\varphi(T_1)\|_1 \geq \|\varphi(T_2)\|_1$, then

$$d_{\varphi,1}(T_1, T_2) = \|\varphi(T_1) - \varphi(T_2)\|_1 \geq \|\varphi(T_1)\|_1 - \|\varphi(T_2)\|_1$$

and therefore $\Delta_1(X_n) \geq \max \|\varphi(X_n)\|_1 - \min \|\varphi(X_n)\|_1$, which is also in $\Theta(n^3)$. This implies that $\Delta_1(X_n)$ is in $\Omega(n^3)$ and hence in $\Theta(n^3)$, as we claimed. \square

Let us consider now the case $p > 1$. By Hölder's inequality, we have that, for every $x = (x_1, \dots, x_m) \in \mathbb{R}^m$, $\|x\|_1 \leq m^{(p-1)/p} \|x\|_p$. Indeed, recall that Hölder's inequality [12, §17] states that, for every real numbers $a_1, \dots, a_m, b_1, \dots, b_m$ and for every positive real number $r > 1$,

$$\sum_{i=1}^m |a_i| \cdot |b_i| \leq \left(\sum_{i=1}^m |a_i|^r \right)^{1/r} \left(\sum_{i=1}^m |b_i|^{r/(r-1)} \right)^{(r-1)/r}.$$

Applying this inequality with $a_i = |x_i|$, $b_i = 1$ and $r = p$, we obtain

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^m |x_i| = \sum_{i=1}^m |x_i| \cdot 1 \leq \left(\sum_{i=1}^m |x_i|^p \right)^{1/p} \left(\sum_{i=1}^m 1^{p/(p-1)} \right)^{(p-1)/p} \\ &= \|x\|_p \cdot m^{(p-1)/p}. \end{aligned}$$

Returning to the cophenetic metrics, applying this inequality, we have that, for every pair of trees $T_1, T_2 \in X_n$,

$$d_{\varphi,1}(T_1, T_2) \leq \binom{n+1}{2}^{(p-1)/p} d_{\varphi,p}(T_1, T_2).$$

and therefore

$$\Delta_1(X_n) \leq \binom{n+1}{2}^{(p-1)/p} \Delta_p(X_n)$$

from where we deduce that

$$\Delta_p(X_n) \geq \Delta_1(X_n) \cdot \binom{n+1}{2}^{-1+\frac{1}{p}},$$

and this lower bound is in $\Theta(n^{3-2+2/p}) = \Theta(n^{(p+2)/p})$. This implies that $\Delta_p(X_n)$ is in $\Omega(n^{(p+2)/p})$.

To prove the converse inequality, let

$$\varphi^{(p)}(T) = \sum_{1 \leq i < j \leq n} \varphi_T(i, j)^p.$$

We have that, for every $T_1, T_2 \in X_n$,

$$\begin{aligned} d_{\varphi,p}(T_1, T_2) &= \|\varphi(T_1) - \varphi(T_2)\|_p \leq \|\varphi(T_1)\|_p + \|\varphi(T_2)\|_p \\ &= \sqrt[p]{\varphi^{(p)}(T_1)} + \sqrt[p]{\varphi^{(p)}(T_2)} \leq 2 \sqrt[p]{\max \varphi^{(p)}(X_n)}, \end{aligned}$$

which implies that $\Delta_p(X_n) \leq 2 \sqrt[p]{\max \varphi^{(p)}(X_n)}$. Therefore, to prove that the diameter of $d_{\varphi,p}$ on each X_n is bounded from above by $O(n^{(p+2)/p})$, it is enough to prove that $\max \varphi^{(p)}(X_n)$ is in $O(n^{p+2})$. We will do it in the next lemma.

Lemma 3.27. *For every $p \geq 1$, the maximum value of $\varphi^{(p)}$ on \mathcal{UT}_n , \mathcal{T}_n or \mathcal{BT}_n is reached at the combs, and its value is in $\Theta(n^{p+2})$.*

Proof. Arguing as in the case $p = 1$, we have that the maximum value of $\varphi^{(p)}(T)$ on \mathcal{UT}_n is reached on trees in \mathcal{T}_n , because if we replace each nested taxon in a tree by a new leaf labeled with the same taxon hanging from the former nested taxon, as in Fig. 3.21, the value of $\varphi^{(p)}$ increases. On the other hand, if a tree $T \in \mathcal{T}_n$ contains a node with $k \geq 3$ children, as in the left hand side of Fig. 3.22, and we replace its subtree rooted at this node as described in the right hand side of Fig. 3.22, we obtain a new tree $T' \in \mathcal{T}_n$ with larger $\varphi^{(p)}$ value: the values of $\varphi(i, j)^p$ for $i, j \in L(T_1) \cup \dots \cup L(T_{k-1})$ increase, and the other values of $\varphi(i, j)^p$ do not change. This implies that for every non-bifurcating phylogenetic tree $T \in \mathcal{T}_n$, there always exists a bifurcating phylogenetic tree $T' \in \mathcal{BT}_n$ such that $\varphi^{(p)}(T') > \varphi^{(p)}(T)$ and in particular that the maximum value of $\varphi^{(p)}(T)$ on \mathcal{UT}_n or \mathcal{T}_n is actually reached on \mathcal{BT}_n .

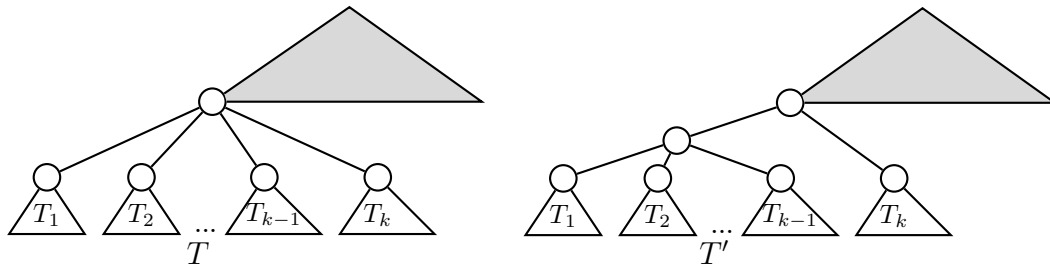


Figure 3.22: $\varphi^{(p)}(T') > \varphi^{(p)}(T)$.

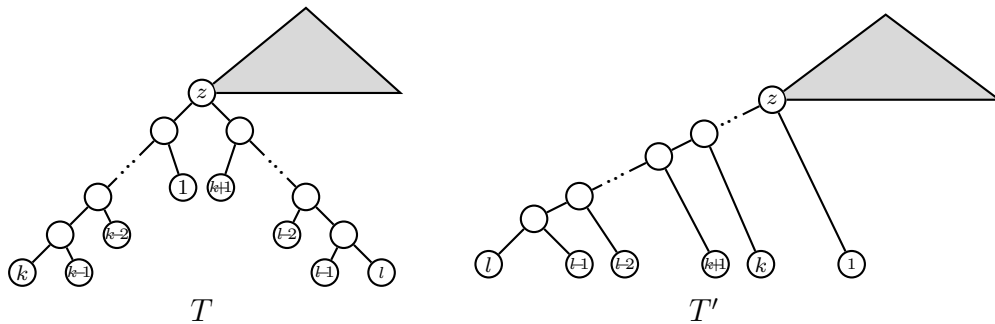


Figure 3.23: $\varphi^{(p)}(T') > \varphi^{(p)}(T)$.

Let now $T \in \mathcal{BT}_n$ and assume that it is not a comb. Therefore, it has an internal node z of largest depth without any leaf child; in particular, all internal descendant nodes of z have some leaf child. Thus, and up to a relabeling of its leaves, T has the form represented in the left hand side of Fig. 3.23, for some $k \geq 2$ and some $l \geq k + 2$. Consider then the tree T' depicted in the right hand side of Fig. 3.23, where the grey triangle represents the same tree in both sides. It turns out that $\varphi^{(p)}(T') - \varphi^{(p)}(T) > 0$. Indeed, if q denotes the depth of the

node z in both trees, then

$$\varphi_{T'}(i, j)^p - \varphi_T(i, j)^p = \begin{cases} (q+i)^p - (q+i+1)^p & \text{if } 1 \leq i = j \leq k-1 \\ 0 & \text{if } i = j = k \\ (q+i)^p - (q+i-k+1)^p & \text{if } k+1 \leq i = j \leq l-1 \\ (q+l-1)^p - (q+l-k)^p & \text{if } i = j = l \\ (q+i-1)^p - (q+i)^p & \text{if } 1 \leq i < j \leq k \\ (q+i-1)^p - (q+i-k)^p & \text{if } k+1 \leq i < j \leq l \\ (q+i-1)^p - q^p & \text{if } 1 \leq i \leq k < j \leq l \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} \varphi^{(p)}(T') - \varphi^{(p)}(T) &= \sum_{i=1}^{k-1} ((q+i)^p - (q+i+1)^p) \\ &+ \sum_{\substack{i=k+1 \\ k-1}}^{l-1} ((q+i)^p - (q+i-k+1)^p) + (q+l-1)^p - (q+l-k)^p \\ &+ \sum_{\substack{i=1 \\ l-1}}^{k-1} (k-i)((q+i-1)^p - (q+i)^p) + \sum_{i=1}^k (l-k)((q+i-1)^p - q^p) \\ &+ \sum_{i=k+1}^{l-1} (l-i)((q+i-1)^p - (q+i-k)^p) \\ &= (q+1)^p - (q+k)^p + \sum_{i=1}^{l-k-1} ((q+k+i)^p - (q+1+i)^p) \\ &+ (q+l-1)^p - (q+l-k)^p + \sum_{i=1}^{k-1} (k-i)((q+i-1)^p - (q+i)^p) \\ &+ \sum_{\substack{i=1 \\ k}}^{l-k-1} (l-k-i)((q+k+i-1)^p - (q+i)^p) \\ &+ \sum_{i=1}^k (l-k)((q+i-1)^p - q^p) \end{aligned}$$

To prove that this sum is non-zero, let us write it as

$$\varphi^{(p)}(T') - \varphi^{(p)}(T) = S_1 + S_2 + S_3,$$

where

$$\begin{aligned} S_1 &= \sum_{\substack{i=1 \\ l-k-1}}^{k-1} (k-i)((q+i-1)^p - (q+i)^p) + \sum_{i=1}^k (l-k)((q+i-1)^p - q^p) \\ S_2 &= \sum_{\substack{i=1 \\ l-k-1}}^{l-k-1} ((q+k+i)^p - (q+1+i)^p) \\ &+ \sum_{i=1}^{l-k-1} (l-k-i)((q+k+i-1)^p - (q+i)^p) \\ S_3 &= (q+1)^p - (q+k)^p + (q+l-1)^p - (q+l-k)^p \end{aligned}$$

Then

$$\begin{aligned}
S_1 &= \sum_{i=1}^{k-1} (k-i)((q+i-1)^p - (q+i)^p) + \sum_{i=1}^k (l-k)((q+i-1)^p - q^p) \\
&= \sum_{i=1}^{k-1} (k-i)(q+i-1)^p - \sum_{i=1}^{k-1} (k-i)(q+i)^p \\
&\quad + \sum_{i=1}^k (l-k)((q+i-1)^p - q^p) \\
&= \sum_{i=1}^{k-1} (k-i)(q+i-1)^p - \sum_{i=2}^k (k-i+1)(q+i-1)^p \\
&\quad + (l-k) \sum_{i=1}^k (q+i-1)^p - k(l-k)q^p \\
&= \sum_{i=1}^{k-1} (l-k-1)(q+i-1)^p + kq^p - (q+k-1)^p \\
&\quad + (l-k)(q+k-1)^p - k(l-k)q^p \\
&= (l-k-1) \sum_{i=1}^k ((q+i-1)^p - q^p) > 0 \\
S_2 &= \sum_{i=1}^{l-k-1} ((q+k+i)^p - (q+1+i)^p) \\
&\quad + \sum_{i=1}^{l-k-1} (l-k-i)((q+k+i-1)^p - (q+i)^p) \\
&= \sum_{i=1}^{l-k-1} ((q+k+i)^p - (q+1+i)^p) \\
&\quad + \sum_{i=0}^{l-k-1} (l-k-i-1)((q+k+i)^p - (q+i+1)^p) \\
&= \sum_{i=1}^{l-k-1} (l-k-i)((q+k+i)^p - (q+1+i)^p) \\
&\quad + (l-k-1)((q+k)^p - (q+1)^p) \\
&> (l-k-1)((q+k)^p - (q+1)^p).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\varphi^{(p)}(T') - \varphi^{(p)}(T) &= S_1 + S_2 + S_3 \\
&> (l-k-1)((q+k)^p - (q+1)^p) + (q+1)^p - (q+k)^p + (q+l-1)^p \\
&\quad - (q+l-k)^p \\
&= (l-k-2)((q+k)^p - (q+1)^p) + (q+l-1)^p - (q+l-k)^p > 0.
\end{aligned}$$

This implies that no tree other than a comb can have the largest $\varphi^{(p)}$ value in \mathcal{BT}_n , and hence also in \mathcal{T}_n and \mathcal{UT}_n .

Finally, if K_n denotes the comb with n leaves in Fig. 1.2.(a) (see also the comb K in the closer Fig. 3.24 below),

$$\varphi_{K_n}(i, j)^p = \begin{cases} (n-1)^p & \text{if } i = j = 1 \\ (n-i+1)^p & \text{if } 2 \leq i = j \leq n \\ (n-j)^p & \text{if } 1 \leq i < j \leq n \end{cases}$$

and thus

$$\begin{aligned} \varphi^{(p)}(K_n) &= (n-2) \cdot 1^p + (n-3) \cdot 2^p + \cdots + 2 \cdot (n-3)^p + 1 \cdot (n-2)^p \\ &\quad + 1^p + 2^p + \cdots + (n-2)^p + (n-1)^p + (n-1)^p \\ &= (n-1) \cdot 1^p + (n-2) \cdot 2^p + \cdots + 3 \cdot (n-3)^p + 2 \cdot (n-2)^p \\ &\quad + (n-1)^p + (n-1)^p \\ &= \sum_{k=1}^{n-1} (n-k) \cdot k^p + (n-1)^p \end{aligned}$$

Now, it turns out that

$$\sum_{k=1}^{n-1} k^m = \frac{1}{m+1} n^{m+1} + O(n^m). \quad (3.1)$$

This property is well known for natural numbers $m \in \mathbb{N}$ [111]. For arbitrary real numbers $m > 0$, it derives from the fact that

$$\int_1^{n-1} (x-1)^m dx \leq \sum_{k=1}^{n-1} k^m \leq \int_1^{n-1} x^m dx,$$

and then

$$\begin{aligned} \int_1^{n-1} (x-1)^m dx &= \frac{1}{m+1} (n-2)^{m+1} = \frac{1}{m+1} n^{m+1} + O(n^m) \\ \int_1^{n-1} x^m dx &= \frac{1}{m+1} ((n-1)^{m+1} - 1) = \frac{1}{m+1} n^{m+1} + O(n^m) \end{aligned}$$

So, by identity (3.1), we have that

$$\begin{aligned} \sum_{k=1}^{n-1} (n-k) \cdot k^p + (n-1)^p &= n \sum_{k=1}^{n-1} k^p - \sum_{k=1}^{n-1} k^{p+1} + O(n^p) \\ &= \left(\frac{1}{p+1} - \frac{1}{p+2} \right) n^{p+2} + O(n^{p+1}) \end{aligned}$$

and hence $\varphi^{(p)}(K_n)$ is in $\Theta(n^{p+2})$. \square

Therefore, $\Delta_p(X_n)$ is bounded from above by $O(n^{(p+2)/p})$ and from below by $\Omega(n^{(p+2)/p})$ and therefore it is indeed in $\Theta(n^{(p+2)/p})$.

We finally prove the case $p = 0$, which needs a completely different argument.

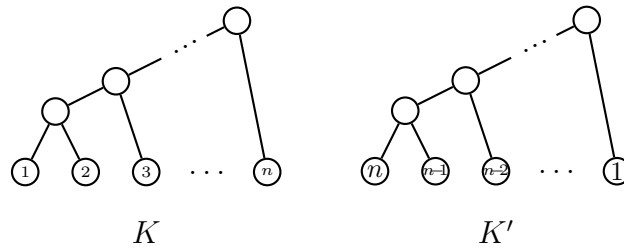


Figure 3.24: The combs used in the proof of Lemma 3.28.

Lemma 3.28. *The diameter of $d_{\varphi,0}$ on \mathcal{UT}_n , \mathcal{T}_n and \mathcal{BT}_n is in $\Theta(n^2)$.*

Proof. Since the cophenetic vector of a tree $T \in \mathcal{UT}_n$ lies in $\mathbb{R}^{n(n+1)/2}$, it is clear that $d_{\varphi,0}(T_1, T_2) \leq n(n+1)/2$, for every $T_1, T_2 \in \mathcal{UT}_n$. Now, consider the pair of combs with n leaves depicted in Fig. 3.24. We have that

$$\begin{aligned} \varphi_K(i, j) &= n - j & \varphi_{K'}(i, j) &= i - 1 & \text{for every } 1 \leq i < j \leq n \\ \varphi_K(i, i) &= n - i + 1 & \varphi_{K'}(i, i) &= i & \text{for every } 2 \leq i \leq n - 1 \\ \varphi_K(1, 1) &= n - 1 & \varphi_{K'}(1, 1) &= 1 \\ \varphi_K(n, n) &= 1 & \varphi_{K'}(n, n) &= n - 1 \end{aligned}$$

Then, the pairs (i, j) , $1 \leq i \leq j \leq n$, such that $\varphi_K(i, j) = \varphi_{K'}(i, j)$ are, on the one hand, the pair (i, i) such that $2i = n + 1$ (which exists only if n is odd) and, on the other hand, the pairs (i, j) such that $1 \leq i \leq n/2$ and $i + j = n + 1$, and there are $n/2$ such pairs if n is even and $(n - 1)/2$ if n is odd. So, the number of pairs such that $\varphi_K(i, j) = \varphi_{K'}(i, j)$ is at most $(n + 1)/2$, and therefore $d_{\varphi,0}(K, K') \geq (n^2 - 1)/2$. So, the diameter of $d_{\varphi,0}$ on \mathcal{UT}_n is bounded from above by $n(n + 1)/2$, and its diameter on \mathcal{BT}_n is bounded from below by $(n^2 - 1)/2$, which implies that the diameter of $d_{\varphi,0}$ on \mathcal{UT}_n , \mathcal{T}_n and \mathcal{BT}_n is in $\Theta(n^2)$. \square

3.5 Expected values in the Euclidean case

Once the dissimilarity 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 value: is it much larger, much smaller, similar? [105] 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, [18, 22, 34, 57, 72, 105].

The expected value of a distance depends on the probability distribution on the space of phylogenetic trees under consideration. Again, we consider the two most popular probabilistic models for bifurcating phylogenetic trees, the uniform and the Yule model (see Section 1.3). In this section we provide

explicit formulas for the expected values of the square of the cophenetic metric $d_{\varphi,2}$ under these two models.

So, let D_n^2 be the random variable that chooses a pair of trees $T, T' \in \mathcal{BT}_n$ and computes $d_{\varphi,2}(T, T')^2$. We shall reduce 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 $T \in \mathcal{BT}_n$ and computes its Sackin index,
- Φ_n , the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes its total cophenetic index,
- $\bar{\Phi}_n^{(2)}$, the random variable that chooses a tree $T \in \mathcal{BT}_n$ and computes

$$\bar{\Phi}_n^{(2)}(T) = \sum_{1 \leq i < j \leq n} \varphi_T(i, j)^2.$$

For the models under consideration, the expected values of these variables are related to that of D_n^2 by the next proposition.

Proposition 3.29. *Let P be a probabilistic model for bifurcating phylogenetic trees invariant under relabelings and let E denote the expected value under this model. Then,*

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

Proof. To simplify the notations, let

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

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

$$\begin{aligned} E(D_n^2) &= \sum_{(T, T') \in \mathcal{BT}_n^2} d_{\varphi,2}(T, T')^2 P(T)P(T') \\ &= \sum_{(T, T') \in \mathcal{BT}_n^2} \left(\sum_{1 \leq i < j \leq n} (\varphi_T(i, j) - \varphi_{T'}(i, j))^2 \right) P(T)P(T') \\ &= \sum_{1 \leq i < j \leq n} \sum_{(T, T') \in \mathcal{BT}_n^2} (\varphi_T(i, j)^2 + \varphi_{T'}(i, j)^2 - 2\varphi_T(i, j)\varphi_{T'}(i, j)) P(T)P(T') \\ &= \sum_{1 \leq i < j \leq n} \left(\sum_{(T, T') \in \mathcal{BT}_n^2} \varphi_T(i, j)^2 P(T)P(T') + \sum_{(T, T') \in \mathcal{BT}_n^2} \varphi_{T'}(i, j)^2 P(T)P(T') \right. \\ &\quad \left. - 2 \sum_{(T, T') \in \mathcal{BT}_n^2} \varphi_T(i, j)\varphi_{T'}(i, j) P(T)P(T') \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{1 \leq i < j \leq n} \left(\sum_{T \in \mathcal{BT}_n} \varphi_T(i, j)^2 P(T) + \sum_{T' \in \mathcal{BT}_n} \varphi_{T'}(i, j)^2 P(T') \right. \\
&\quad \left. - 2 \left(\sum_{T \in \mathcal{BT}_n} \varphi_T(i, j) P(T) \right) \left(\sum_{T' \in \mathcal{BT}_n} \varphi_{T'}(i, j) P(T') \right) \right) \\
&= \sum_{1 \leq i < j \leq n} \left(2 \sum_{T \in \mathcal{BT}_n} \varphi_T(i, j)^2 P(T) - 2 \left(\sum_{T \in \mathcal{BT}_n} \varphi_T(i, j) P(T) \right)^2 \right) \\
&= 2 \sum_{T \in \mathcal{BT}_n} \left(\sum_{1 \leq i < j \leq n} \varphi_T(i, j)^2 \right) P(T) - 2 \sum_{1 \leq i < j \leq n} \left(\sum_{T \in \mathcal{BT}_n} \varphi_T(i, j) P(T) \right)^2 \\
&\quad - 2 \sum_{1 \leq i \leq n} \left(\sum_{T \in \mathcal{BT}_n} \varphi_T(i, i) P(T) \right)^2 \\
&= 2 \sum_{T \in \mathcal{BT}_n} \bar{\Phi}^{(2)}(T) P(T) - 2 \binom{n}{2} \left(\sum_{T \in \mathcal{BT}_n} \varphi_T(1, 2) P(T) \right)^2 \\
&\quad - 2n \left(\sum_{T \in \mathcal{BT}_n} \delta_T(1) P(T) \right)^2 \\
&\text{(by the invariance under relabelings of } P) \\
&= 2E(\bar{\Phi}_n^{(2)}) - n(n-1)E(\varphi_n)^2 - 2nE(\delta_n)^2.
\end{aligned}$$

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 again the invariance under relabelings of P :

$$E(\delta_n) = E(S_n)/n, \quad E(\varphi_n) = E(\Phi_n)/\binom{n}{2}.$$

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

3.5.1 Expected value of D_n^2 under the Yule model

Since the Yule model is invariant under relabelings, Proposition 3.29 implies that

$$E_Y(D_n^2) = 2E_Y(\bar{\Phi}_n^{(2)}) - 2 \cdot \frac{E_Y(S_n)^2}{n} - 4 \cdot \frac{E_Y(\Phi_n)^2}{n(n-1)}.$$

In this expression we already know (from (1.4) and Theorem 2.20, respectively) the expected values of S_n and Φ_n :

$$E_Y(S_n) = 2n(H_n - 1), \quad E_Y(\Phi_n) = n(n-1) - 2n(H_n - 1).$$

Using these values in the expression for $E_Y(D_n^2)$ given above, we obtain

$$E_Y(D_n^2) = 2E_Y(\bar{\Phi}_n^{(2)}) + 16n(H_n - 1) - 4n(n-1) - \frac{8n(n+1)(H_n - 1)^2}{n-1}. \quad (3.2)$$

So, to obtain $E_Y(D_n^2)$, it remains to compute $E_Y(\bar{\Phi}_n^{(2)})$.

Proposition 3.30. *For every $n \geq 2$, $E_Y(\bar{\Phi}_n^{(2)}) = 5n(n-1) - 8n(H_n - 1)$.*

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

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

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

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

To compute $E_Y(\bar{\Phi}_n^{(2)})$, we shall use an argument similar to the one used in the proof of Lemma 2.19. To begin with, from the very definition of $E_Y(\bar{\Phi}_n^{(2)})$ we have that

$$\begin{aligned} E_Y(\bar{\Phi}_n^{(2)}) &= \sum_{T \in \mathcal{BT}_n} \bar{\Phi}^{(2)}(T) \cdot P_Y(T) \\ &= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\substack{S_k \subsetneq [n] \\ |S_k|=k}} \sum_{T_k \in \mathcal{BT}(S_k)} \sum_{T'_{n-k} \in \mathcal{BT}(S_k^c)} \bar{\Phi}^{(2)}(T_k \star T'_{n-k}) \cdot P_Y(T_k \star T'_{n-k}). \end{aligned} \quad (3.3)$$

In this expression, we know from equation (2.1) that, for every $\emptyset \neq S_k \subsetneq [n]$ with $|S_k| = k$, if $T_k \in \mathcal{BT}(S_k)$ and $T'_{n-k} \in \mathcal{BT}(S_k^c)$, then,

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

On the other hand, we have the following recursive expression for $\bar{\Phi}^{(2)}(T_k \star T'_{n-k})$

$$\begin{aligned} \bar{\Phi}^{(2)}(T_k \star T'_{n-k}) &= \bar{\Phi}^{(2)}(T_k) + \bar{\Phi}^{(2)}(T'_{n-k}) + 2\bar{\Phi}(T_k) + 2\bar{\Phi}(T'_{n-k}) \\ &\quad + \binom{k+1}{2} + \binom{n-k+1}{2}. \end{aligned} \quad (3.4)$$

Indeed, let us assume without any loss of generality, that $S_k = \{1, \dots, k\}$ and $T_k \in \mathcal{BT}(S_k)$ and $T'_{n-k} \in \mathcal{BT}(S_k^c)$. Then:

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

and therefore

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

Then, using the identities (2.1) and (3.4) in the expression (3.3), we obtain:

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

In particular

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

and therefore

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

$$\begin{aligned}
&= \frac{n-2}{n-1} E_Y(\bar{\Phi}_{n-1}^{(2)}) + \frac{2}{n-1} E_Y(\bar{\Phi}_{n-1}^{(2)}) + \frac{4}{n-1} E_Y(\bar{\Phi}_{n-1}^{(2)}) + n \\
&= \frac{n}{n-1} E_Y(\bar{\Phi}_{n-1}^{(2)}) + \frac{4}{n-1} (n-1)(n-2) + n \\
&= \frac{n}{n-1} E_Y(\bar{\Phi}_{n-1}^{(2)}) + 5n - 8.
\end{aligned}$$

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

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

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

$$x_n = \sum_{k=2}^n \left(5 - \frac{8}{k}\right) = 5(n-1) - 8(H_n - 1)$$

from where we deduce that $E_Y(\bar{\Phi}_n^{(2)}) = 5n(n-1) - 8n(H_n - 1)$, as we claimed. \square

Replacing the value of $E_Y(\bar{\Phi}_n^{(2)})$ obtained in the last proposition in identity (3.2), we obtain the main result in this subsection:

Theorem 3.31. *For every $n \geq 2$, the expected value of D_n^2 under the Yule model is*

$$E_Y(D_n^2) = \frac{2n}{n-1} (3(n-1)^2 - 4(n+1)(H_n - 1)^2).$$

\square

Since $H_n \sim \ln(n)$, this formula implies that

$$E_Y(D_n^2) \sim 6n^2.$$

To double-check the formula given in Theorem 3.31, we have computed the exact values of $E_Y(\Phi_n)$, for $n = 3, \dots, 7$, and they agree with the figures given by our formula. Table 3.2 below gives these values, together with those of the corresponding variance $\sigma_Y^2(D_n^2)$ (both rounded to 5 decimal digits). The Python and R scripts used to compute them are available in Appendix A.5.1 and on the GitHub repository associated to this PhD Thesis [97].

n	3	4	5	6	7
$E_Y(D_n^2)$	2.66667	9.40741	21.18333	38.71200	62.55619
$\sigma_Y^2(D_n^2)$	3.55556	29.13032	117.63306	339.28881	797.15834

Table 3.2: The exact values of the mean and variance of D_n^2 under the Yule model for $n = 3, \dots, 7$.

3.5.2 Expected value of D_n^2 under the uniform model

Since the uniform model is also invariant under relabelings, Proposition 3.29 implies that

$$E_U(D_n^2) = 2E_U(\overline{\Phi}_n^{(2)}) - 2 \cdot \frac{E_U(S_n)^2}{n} - 4 \cdot \frac{E_U(\Phi_n)^2}{n(n-1)}. \quad (3.5)$$

In this expression we already know (Theorems 2.27 and 2.28, respectively) the expected values of S_n and Φ_n :

$$E_U(S_n) = n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right), \quad E_U(\Phi_n) = \frac{1}{2} \binom{n}{2} \left(\frac{(2n-2)!!}{(2n-3)!!} - 2 \right) \quad (3.6)$$

So, to obtain $E_U(D_n^2)$, it remains to compute $E_U(\overline{\Phi}_n^{(2)})$, which we do in the next proposition.

Proposition 3.32. *For every $n \geq 2$,*

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

We split the proof of this result into several lemmas, using an argument similar to the one used Section 2.4 to compute $E_U(S_n)$. For every $k = 1, \dots, n-1$, let

$$\begin{aligned} c_{k,n} &= |\{T \in \mathcal{BT}_n \mid \delta_T(1) = k\}| \\ &= |\{T \in \mathcal{BT}_n \mid \delta_T(i) = k\}| \text{ for every } 1 \leq i \leq n \end{aligned}$$

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

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

The upper limits for k in $c_{k,n}$ and $f_{k,n}$ are $n-1$ and $n-2$, respectively, because the largest depth of a leaf and of an internal node in a bifurcating tree with n leaves are $n-1$ and $n-2$, respectively.

Lemma 3.33. *For every $n \geq 2$,*

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

Proof. Since, under the uniform model, $P_U(T) = 1/(2n-3)!!$ for every $T \in \mathcal{BT}_n$,

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

where

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

□

A formula for $c_{k,n}$ was obtained in the proof of Lemma 2.24:

$$c_{k,n} = \frac{(2n - k - 3)! \cdot k}{(n - k - 1)! 2^{n-k-1}}. \quad (3.7)$$

As far as $f_{k,n}$ goes, we have the following result.

Lemma 3.34. *For every $n \geq 2$, $f_{0,n} = (2n - 4)!!$ and*

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

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 $[1, 2]_T$ 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 $T' \in \mathcal{BT}_{n-1}$ 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 $k \geq 1$, we shall study the structure of a tree $T \in \mathcal{BT}_n$ such that $\varphi_T(1, 2) = k$; to simplify the notations, let us denote by x the node $[1, 2]_T$, 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 $S_0 \subseteq [n]$ 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. 3.25.

So, the tree T is determined by:

- A number $0 \leq m \leq n - k - 2$, so that $m + 2$ will be the number of leaves of the phylogenetic tree T_0 rooted at $[1, 2]_T$
- 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 bifurcating k -forest (T_1, T_2, \dots, T_k) on $[n] \setminus L(T_0)$. The number of such ordered k -forests is, by Equation (1.1),

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

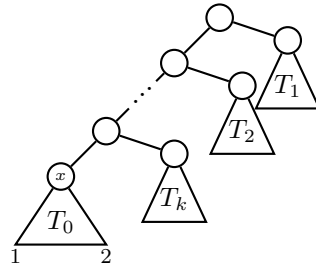


Figure 3.25: The structure of a tree T with $\varphi_T(1, 2) = k$.

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

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

We use the lookup algorithm to compute this sum. Let

$$t_m = \frac{4^m(2n - 2m - k - 5)!}{(n - m - 2)!(n - m - k - 2)!}.$$

Then

$$t_0 = \frac{(2n - k - 5)!}{(n - 2)!(n - k - 2)!}$$

and

$$\begin{aligned} \frac{t_{m+1}}{t_m} &= \frac{4^{m+1}(2n - 2m - k - 7)!(n - m - 2)!(n - m - k - 2)!}{(n - m - 3)!(n - m - k - 3)!4^m(2n - 2m - k - 5)!} \\ &= \frac{4(n - m - 2)(n - m - k - 2)}{(2n - 2m - k - 5)(2n - 2m - k - 6)} \\ &= \frac{(m + 2 - n)(m + k + 2 - n)(m + 1)}{(m + \frac{k+5}{2} - n)(m + \frac{k}{2} + 3 - n)(m + 1)} \end{aligned}$$

This implies, by the lookup algorithm, that

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

as we claimed. \square

We must compute now the sums $\sum_{k=1}^{n-1} k^2 \cdot c_{k,n}$ and $\sum_{k=1}^{n-2} k^2 \cdot f_{k,n}$. In these computations we shall use twice the following sum:

Lemma 3.35.

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

Proof. With the notation

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

introduced in Lemma 2.26, we have that

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

The value of $U_{n,0}$ was given in Lemma 2.25,

$$U_{n,0} = (n-2)!2^{n-2} = (2n-4)!!,$$

and the values of $U_{n,1}$ and $U_{n,2}$ were computed in the proof of Theorem 2.27, and they are

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

As far as $U_{n,3}$ goes, using Lemma 2.26 we have that

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

Therefore

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

as we claimed. □

Lemma 3.36. For every $n \geq 2$,

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

Proof. By equation (3.7) and the last lemma

$$\begin{aligned} \sum_{k=1}^{n-1} k^2 c_{k,n} &= \sum_{k=1}^{n-1} \frac{k^3(2n-k-3)!}{(n-k-1)!2^{n-k-1}} \\ &= (4n-1)(2n-3)!! - 3(2n-2)(2n-4)!! \end{aligned}$$

□

Lemma 3.37. For every $n \geq 2$,

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

Proof. To simplify the notations, set $S_n = \sum_{k=1}^{n-2} k^2 f_{k,n}$. As we have seen in the proof of Lemma 3.34,

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

and therefore

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

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

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

where

$$\begin{aligned} S'_{p+1} - S'_p &= \sum_{k=1}^{p-2} \frac{k^3}{2^k} \sum_{m=1}^{p-k-1} \frac{1}{4^m m} \binom{k+2m-1}{k+m} \\ &\quad - \sum_{k=1}^{p-3} \frac{k^3}{2^k} \sum_{m=1}^{p-k-2} \frac{1}{4^m m} \binom{k+2m-1}{k+m} \\ &= \frac{(p-2)^3}{2^p} + \sum_{k=1}^{p-3} \frac{k^3}{2^k} \sum_{m=1}^{p-k-1} \frac{1}{4^m m} \binom{k+2m-1}{k+m} \\ &\quad - \sum_{k=1}^{p-3} \frac{k^3}{2^k} \sum_{m=1}^{p-k-2} \frac{1}{4^m m} \binom{k+2m-1}{k+m} \end{aligned}$$

$$\begin{aligned}
&= \frac{(p-2)^3}{2^p} + \sum_{k=1}^{p-3} \frac{k^3}{2^k(p-k-1)4^{p-k-1}} \binom{2p-k-3}{p-1} \\
&= \frac{(p-2)^3}{2^p} + \frac{1}{2^{2p-2}} \sum_{k=1}^{p-3} \frac{k^3(2p-k-3)!}{2^{-k}(p-k-1)(p-1)!(p-k-2)!} \\
&= \frac{(p-2)^3}{2^p} + \frac{1}{2^{2p-2}(p-1)!} \sum_{k=1}^{p-3} \frac{k^3(2p-k-3)!}{2^{-k}(p-k-1)!} \\
&= \frac{(p-2)^3}{2^p} + \frac{1}{2^{2p-2}(p-1)!} \sum_{k=1}^{p-3} \frac{(p-k-2)^3(p+k-1)!}{2^{k-p+2}(k+1)!} \\
&= \frac{(p-2)^3}{2^p} + \frac{1}{2^{p-1}(p-1)!} \sum_{k=2}^{p-2} \frac{(p-k-1)^3(p+k-2)!}{2^k k!} \\
&= \frac{(p-2)^3}{2^p} + \frac{1}{2^{p-1}(p-1)!} \left[\sum_{k=0}^{p-2} \frac{(p-k-1)^3(p+k-2)!}{2^k k!} \right. \\
&\quad \left. - (p-1)^3(p-2)! - \frac{1}{2}(p-2)^3(p-1)! \right] \\
&= -\frac{(p-1)^2}{2^{p-1}} + \frac{1}{2^{p-1}(p-1)!} \sum_{k=0}^{p-2} \frac{(p-k-1)^3(p+k-2)!}{2^k k!} \\
&= -\frac{(p-1)^2}{2^{p-1}} + \frac{1}{(2p-2)!!} ((4p-1)(2p-3)!! - 3(2p-2)!!) \\
&\quad \text{(by Lemma 3.35)} \\
&= -\frac{(p-1)^2}{2^{p-1}} + (4p-1) \frac{(2p-3)!!}{(2p-2)!!} - 3.
\end{aligned}$$

Therefore

$$\begin{aligned}
S'_n &= \sum_{p=3}^{n-1} \left((4p-1) \frac{(2p-3)!!}{(2p-2)!!} - \frac{(p-1)^2}{2^{p-1}} - 3 \right) \\
&= \sum_{p=3}^{n-1} (4p-1) \frac{(2p-3)!!}{(2p-2)!!} - \sum_{k=2}^{n-2} \frac{k^2}{2^k} - 3(n-3) \\
&= \sum_{p=3}^{n-1} (4p-1) \frac{(2p-3)!!}{(2p-2)!!} - \frac{11}{2} - \frac{2+n^2}{2^{n-2}} - 3(n-3)
\end{aligned}$$

Now, applying *Gosper's algorithm* (see Section 1.4) we have that

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

and then

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

Thus, finally,

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

□

We are finally in condition to prove Proposition 3.32. Indeed, by Lemmas 3.33, 3.36, and 3.37, we have that

$$\begin{aligned} E_U(\overline{\Phi}_n^{(2)}) &= \frac{1}{(2n-3)!!} \left(n \sum_{k=1}^{n-1} k^2 \cdot c_{k,n} + \binom{n}{2} \sum_{k=1}^{n-2} k^2 \cdot f_{k,n} \right) \\ &= \frac{1}{(2n-3)!!} \left(n((4n-1)(2n-3)!! - 3(2n-2)!!) \right. \\ &\quad \left. + \binom{n}{2} \left(\frac{1}{3}(4n+1)(2n-3)!! - \frac{3}{2}(2n-2)!! \right) \right) \\ &= \frac{1}{6}n(4n^2 + 21n - 7) - \frac{3n(n+3)}{4} \cdot \frac{(2n-2)!!}{(2n-3)!!} \end{aligned}$$

Theorem 3.38. *For every $n \geq 2$, the expected value of D_n^2 under the uniform model is*

$$\begin{aligned} E_U(D_n^2) &= \frac{1}{3}(4n^3 + 18n^2 - 10n) - \frac{n(n+3)}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} \\ &\quad - \frac{n(n+7)}{4} \left(\frac{(2n-2)!!}{(2n-3)!!} \right)^2. \end{aligned}$$

Proof. If we replace $E_U(S_n)$, $E_U(\Phi_n)$, and $E_U(\overline{\Phi}_n^{(2)})$ in (3.5) by their values given in (3.6) and Proposition 3.32, we obtain:

$$\begin{aligned} E_U(D_n^2) &= 2E_U(\overline{\Phi}_n^{(2)}) - 2 \cdot \frac{E_U(S_n)^2}{n} - 4 \cdot \frac{E_U(\Phi_n)^2}{n(n-1)} \\ &= 2 \left(\frac{1}{6}n(4n^2 + 21n - 7) - \frac{3}{4}n(n+3) \frac{(2n-2)!!}{(2n-3)!!} \right) \end{aligned}$$

$$\begin{aligned}
& -\frac{2}{n}n^2 \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right)^2 - \frac{4}{n(n-1)} \left(\frac{1}{2} \binom{n}{2} \left(\frac{(2n-2)!!}{(2n-3)!!} - 2 \right) \right)^2 \\
= & \frac{1}{3}(4n^3 + 21n^2 - 7n) - 3 \frac{n(n+3)}{2} \frac{(2n-2)!!}{(2n-3)!!} - 2n \left(\frac{(2n-2)!!}{(2n-3)!!} - 1 \right)^2 \\
& - \frac{4}{n(n-1)} \frac{n^2(n-1)^2}{16} \left(\frac{(2n-2)!!}{(2n-3)!!} - 2 \right)^2 \\
= & \frac{1}{3}(4n^3 + 21n^2 - 7n) - 3 \frac{n(n+3)}{2} \frac{(2n-2)!!}{(2n-3)!!} \\
& - 2n \left(\left(\frac{(2n-2)!!}{(2n-3)!!} \right)^2 + 1 - 2 \frac{(2n-2)!!}{(2n-3)!!} \right) \\
& - \frac{n(n-1)}{4} \left(\left(\frac{(2n-2)!!}{(2n-3)!!} \right)^2 + 4 - 4 \frac{(2n-2)!!}{(2n-3)!!} \right) \\
= & \frac{1}{3}(4n^3 + 21n^2 - 7n - 6n - 3n(n-1)) \\
& + \left(-\frac{3n(n+3)}{2} + 4n + n(n-1) \right) \frac{(2n-2)!!}{(2n-3)!!} \\
& - \left(2n + \frac{n(n-1)}{4} \right) \left(\frac{(2n-2)!!}{(2n-3)!!} \right)^2 \\
= & \frac{1}{3}(4n^3 + 18n^2 - 10n) - \frac{n(n+3)}{2} \frac{(2n-2)!!}{(2n-3)!!} - \frac{n(n+7)}{4} \left(\frac{(2n-2)!!}{(2n-3)!!} \right)^2
\end{aligned}$$

□

Since $(2n-2)!!/(2n-3)!! \sim \sqrt{\pi n}$, as we saw in page 60, Theorem 3.38 implies that

$$E_U(D_n^2) \sim \left(\frac{4}{3} - \frac{\pi}{4} \right) n^3.$$

To double-check the formula given in Theorem 3.38, we have computed the exact values of $E_U(\Phi_n)$, for $n = 3, \dots, 7$, and they agree with the figures given by our formula. Table 3.3 below gives these values, together with those of the corresponding variance $\sigma_U^2(D_n^2)$ (both rounded to 5 decimal digits). The Python and R scripts used to compute them are available in Appendix A.5.1 and on the GitHub repository associated to this PhD Thesis [97].

n	3	4	5	6	7
$E_U(D_n^2)$	2.66667	10.56000	26.23673	52.30234	91.40863
$\sigma_U^2(D_n^2)$	3.55556	34.08640	159.50314	539.50829	1502.72330

Table 3.3: The exact values of the mean and variance of D_n^2 under the uniform model for $n = 3, \dots, 7$.

3.6 On the variance of D_n^2

The spread of D_n^2 around its expected value can be quantified by means of its variance. Unfortunately, we have not been able to derive a closed formula for this variance under any model. So, we provide instead an estimation of its order, both under the uniform and the Yule models, based on simulations.

Since $\sigma^2(D_n^2) = E(D_n^4) - E(D_n^2)^2$, the computation of this variance involves the computation of the expected value of D_n^4 . Developing this expected value as in Proposition 3.29, 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 we have not been able to convert it, either for the Yule or the uniform model, into a closed formula depending only on n , not even to a recurrence similar to those given in Sections 2.5 to 2.7. Therefore, in order to be able to, at least, estimate the asymptotic order of $E(D_n^4)$ and $\sigma^2(D_n^2)$, we have taken the Monte Carlo path.

More specifically, both for the Yule and the uniform models, and for every $n = 3, \dots, 100$, we have randomly generated $N = 10000$ pairs of bifurcating trees $(T, T') \in \mathcal{BT}_n \times \mathcal{BT}_n$, we have computed the value of $d_{\varphi,2}(T, T')^2$ and $d_{\varphi,2}(T, T')^4$ for each such pair (T, T') , we have computed the arithmetic means $\overline{D_n^2}$ and $\overline{D_n^4}$ of these N values, and, finally, the variance of the values $d_{\varphi,2}(T, T')^2$ using the expression

$$\overline{\sigma^2(D_n^2)} = \overline{D_n^4} - \overline{D_n^2}^2.$$

This value is an estimation of $\sigma^2(D_n^2)$ under the corresponding model.

Finally, we have computed the slope α of the regression line of $\ln(\overline{\sigma^2(D_n^2)})$ as a function of $\log(n)$ using the values for $n = 50, \dots, 100$. We have only considered the largest values of n because if smaller values were also included in the regression, the determination 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:

- $\ln(\overline{\sigma_Y^2(D_n^2)}) \sim 4.1522n$ with $R^2 = 0.99714$
- $\ln(\overline{\sigma_U^2(D_n^2)}) \sim 6.3883n$ with $R^2 = 0.99931$

The intermediate results of all these computations, as well as the Python and R scripts used to compute them, are available in the GitHub repository [97] and also the scripts in Appendix A.5.2.

So, we estimate that $\sigma_Y^2(D_n^2) \approx O(n^{4.15})$ and $\sigma_U^2(D_n^2) \approx O(n^{6.39})$, and the large R^2 value tell us that these orders explain quite well the estimated expected values up to $n = 100$.

Fig. 3.26 displays $\ln(\sigma_Y^2(D_n^2))$ and $\ln(\sigma_U^2(D_n^2))$ as a function of $\ln(n)$, together with the corresponding regression lines.

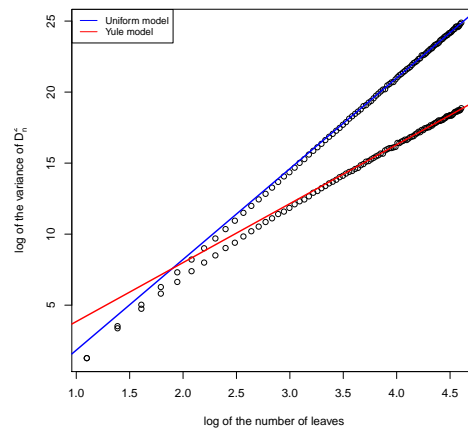


Figure 3.26: Log-log plot of $\ln(\sigma_Y^2(D_n^c))$ and $\ln(\sigma_U^2(D_n^c))$.

Chapter 4

The Colless-like indices

In this chapter, we generalize the classical Colless index for bifurcating trees to a family of *Colless-like* balance indices defined on multifurcating phylogenetic trees. Each such Colless-like index $\mathfrak{C}_{D,f}$ is determined by the choice of a dissimilarity D and a weight function $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$. We shall prove that taking $f(n) = \ln(n + e)$ or $f(n) = e^n$ as weight functions, the resulting index $\mathfrak{C}_{D,f}$ is equal to 0 exactly on the fully symmetric trees. Next, for each one of these two functions f and for three popular dissimilarities D (the variance, the standard deviation, and the mean deviation from the median), we find the phylogenetic trees that achieve the maximum value of $\mathfrak{C}_{D,f}$ for their number n of leaves. The results show that the growth pace of the function f influences the notion of “balance” measured by the indices it defines.

4.1 The Colless index

As we have seen in Section 1.2, the *Colless index* $C(T)$ of a bifurcating phylogenetic tree T with n leaves is defined as follows: if, for every $v \in V_{int}(T)$, we denote by v_1 and v_2 its two children and by $\kappa(v_1)$ and $\kappa(v_2)$ their respective numbers of descendant leaves, then

$$C(T) = \sum_{v \in V_{int}(T)} bal_T(v) = \sum_{v \in V_{int}(T)} |\kappa(v_1) - \kappa(v_2)|.$$

The Colless index of an unlabeled tree is simply defined as the Colless index of any phylogenetic tree with this shape.

It is well-known that the maximum Colless index on the set of bifurcating trees with n leaves is reached at the comb K_n (see Fig. 1.2 (a)), and it is

$$C(K_n) = \binom{n-1}{2}$$

(see, for instance, [90]). In fact, this maximum is only reached at the comb. Since we have not been able to find an explicit reference for this last result in the literature and we shall make use of it later, we provide a proof here.

Lemma 4.1. *For every bifurcating tree T with n leaves, if $T \neq K_n$, then $C(T) < C(K_n)$.*

Proof. Let T a bifurcating tree with n leaves different from the comb K_n . Let x be an internal node of smallest depth in it without any leaf child, and let $T_1 \star T_2$ and $T_3 \star T_4$ be the subtrees rooted at its children (see Fig. 4.1); for every $i = 1, 2, 3, 4$, let t_i be the number of leaves of T_i . Assume, without any loss of generality, that $t_1 \leq t_2$ and $t_1 + t_2 \leq t_3 + t_4$. Let then T' be the tree obtained by pruning T_2 from T and inserting it in the other arc starting in x (see again Fig. 4.1).

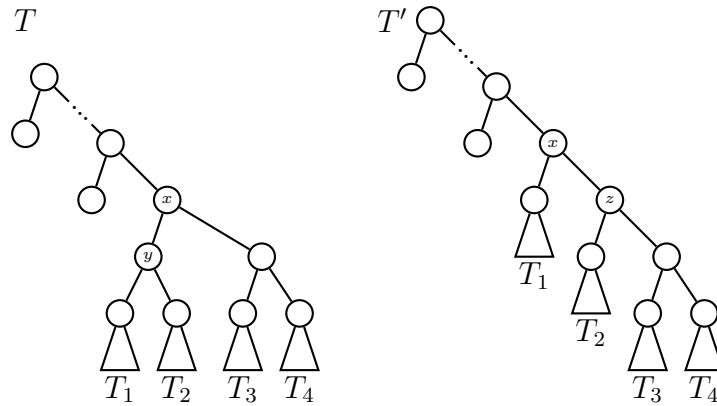


Figure 4.1: The trees T and T' in the proof of Lemma 4.1.

Then $C(T') > C(T)$. Indeed, the only nodes whose children change their numbers of descendant leaves from T to T' are (cf. Fig. 4.1): the node x ; the parent y of the roots of T_1 and T_2 in T , which is removed in T' ; and the parent z of the root of T_2 in T' , which does not exist in T . Therefore,

$$\begin{aligned}
 C(T') - C(T) &= \text{bal}_{T'}(z) + \text{bal}_{T'}(x) - \text{bal}_T(y) + \text{bal}_T(x) \\
 &= |t_3 + t_4 - t_2| + [t_3 + t_4 + t_2 - t_1] - |t_2 - t_1| - |t_3 + t_4 - t_2 - t_1| \\
 &= t_3 + t_4 - t_2 + t_3 + t_4 + t_2 - t_1 - t_2 + t_1 - t_3 - t_4 + t_2 + t_1 \\
 &= t_1 + t_3 + t_4 > 0.
 \end{aligned}$$

So, this procedure takes a bifurcating tree with n leaves $T \neq K_n$ and produces a new bifurcating tree T' with the same number n of leaves and strictly larger Colless index. Since the number of bifurcating trees with n leaves is finite, the Colless index cannot increase indefinitely, which means that if we iterate this procedure, we must eventually stop at a comb K_n . And since the Colless index strictly increases at each iteration, we conclude that if $T \neq K_n$, then $C(T) < C(K_n)$.

□

4.2 The Colless-like indices

Let $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ be a function that sends each natural number to a positive real number. The f -size of a tree $T \in \mathcal{T}^*$ is defined as

$$\sigma_f(T) = \sum_{v \in V(T)} f(\deg(v)).$$

If $T \in \mathcal{T}_S$, for some set of taxa S , then $\sigma_f(T)$ is defined as the f -size of its shape: $\sigma_f(T) = \sigma_f(\pi(T))$.

Therefore, $\sigma_f(T)$ is the sum of the out-degrees of all nodes in T , with these degrees weighted by means of the function f . Examples of f -sizes include:

- The *number of leaves*, κ , which is obtained by taking $f(0) = 1$ and $f(n) = 0$ if $n > 0$.
- The *order* (the number of nodes), τ , which corresponds to $f(n) = 1$ for every $n \in \mathbb{N}$.
- The usual *size* (the number of arcs), θ , which corresponds to $f(n) = n$ for every $n \in \mathbb{N}$.

The following lemma shows that σ_f is a recursive shape index in the sense of page 17.

Lemma 4.2. *Let $T \in \mathcal{T}_n$ be a phylogenetic tree with root r , and let T_1, \dots, T_k , $k \geq 2$, be the subtrees rooted at the children of r . Then,*

$$\sigma_f(T) = \sum_{i=1}^k \sigma_f(T_i) + f(k).$$

Proof. If we consider separately the root r of T and the nodes of each subtree T_i , we obtain:

$$\sigma_f(T) = \sum_{v \in V(T)} f(\deg(v)) = f(\deg(r)) + \sum_{i=1}^k \sum_{v \in V(T_i)} f(\deg(v)) = f(k) + \sum_{i=1}^k \sigma_f(T_i)$$

as we claimed. □

Table 4.2 in Section 4.7 gives the abstract values of $\sigma_f(T)$ for every $T \in \mathcal{T}_n^*$ with $n = 2, 3, 4, 5$.

Example 4.3. *If T is a bifurcating tree with n leaves, and hence with $n - 1$ internal nodes, all of them of out-degree 2, then*

$$\sigma_f(T) = n \cdot f(0) + (n - 1)f(2).$$

Example 4.4. For every fully symmetric tree FS_{n_1, \dots, n_k} (see Chapter 1 and the example depicted in Fig. 1.4),

$$\sigma_f(FS_{n_1, \dots, n_k}) = n_1 \cdots n_k \cdot f(0) + n_1 \cdots n_{k-1} \cdot f(n_k) + \cdots + n_1 \cdot f(n_2) + f(n_1).$$

Indeed, notice that, by construction, the root r of FS_{n_1, \dots, n_k} has out-degree n_1 , its n_1 children have out-degree n_2 , the $n_2 n_1$ nodes of depth 2 have out-degree n_3 , and so on until we reach the $n_1 \cdots n_k$ leaves, all of them of depth k and out-degree 0. Then, separating the nodes of FS_{n_1, \dots, n_k} by their depth, we obtain

$$\begin{aligned} \sigma_f(FS_{n_1, \dots, n_k}) &= \sum_{l=0}^k \sum_{\substack{v \in V(FS_{n_1, \dots, n_k}) \\ \delta(v)=l}} f(\deg(v)) \\ &= f(n_1) + n_1 \cdot f(n_2) + n_1 n_2 \cdot f(n_3) + \cdots + n_1 \cdots n_{k-1} \cdot f(n_k) \\ &\quad + n_1 \cdots n_k \cdot f(0) \end{aligned}$$

Definition 4.5. Let $\mathbb{R}^+ = \bigcup_{k \geq 1} \mathbb{R}^k = \{(x_1, \dots, x_k) \mid k \geq 1, x_1, \dots, x_k \in \mathbb{R}\}$ be the set of all non-empty finite-length sequences of real numbers. A dissimilarity on \mathbb{R}^+ is any mapping $D : \mathbb{R}^+ \rightarrow \mathbb{R}_{\geq 0}$ satisfying the following conditions: for every $(x_1, \dots, x_k) \in \mathbb{R}^+$,

- $D(x_1, \dots, x_k) = D(x_{\sigma(1)}, \dots, x_{\sigma(k)})$, for every permutation σ of $\{1, \dots, k\}$;
- $D(x_1, \dots, x_k) = 0$ if, and only if, $x_1 = \cdots = x_k$.

The theory developed in this chapter works for any dissimilarity. Nevertheless, the dissimilarities that we shall explicitly use in practice are the *mean deviation from the median*,

$$\text{MDM}(x_1, \dots, x_k) = \frac{1}{k} \sum_{i=1}^k |x_i - \text{Median}(x_1, \dots, x_k)|,$$

the (sample) variance,

$$\text{var}(x_1, \dots, x_k) = \frac{1}{k-1} \sum_{i=1}^k (x_i - \text{Mean}(x_1, \dots, x_k))^2,$$

and the (sample) standard deviation,

$$\text{sd}(x_1, \dots, x_k) = +\sqrt{\text{var}(x_1, \dots, x_k)}.$$

Let D be a dissimilarity on \mathbb{R}^+ , $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ a function, and σ_f the corresponding f -size, and let $T \in \mathcal{T}^*$. For every internal node v in T , with children v_1, \dots, v_k , the (D, f) -balance value of v is

$$\text{bal}_{D,f}(v) = D(\sigma_f(T_{v_1}), \dots, \sigma_f(T_{v_k})).$$

So, $\text{bal}_{D,f}(v)$ measures, through D , the spread of the f -sizes of the subtrees rooted at the children of v . In particular, $\text{bal}_{D,f}(v) = 0$ if, and only if, $\sigma_f(T_{v_1}) = \cdots = \sigma_f(T_{v_k})$.

Definition 4.6. Let D be a dissimilarity on \mathbb{R}^+ and $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ a function. For every $T \in \mathcal{T}^*$, its Colless-like index relative to D and f , $\mathfrak{C}_{D,f}(T)$, is the sum of the (D, f) -balance values of the internal nodes of T :

$$\mathfrak{C}_{D,f}(T) = \sum_{v \in V_{int}(T)} bal_{D,f}(v).$$

If $T \in \mathcal{T}_S$, for some set of labels S , then $\mathfrak{C}_{D,f}(T)$ is defined as $\mathfrak{C}_{D,f}(T) = \mathfrak{C}_{D,f}(\pi(T))$.

Example 4.7. If we take $D = \text{MDM}$ and f the constant mapping 1, so that $\sigma_f = \tau$, the usual order of a tree, then

$$\begin{aligned} \mathfrak{C}_{\text{MDM},\tau}(T) &= \sum_{v \in V_{int}(T)} \text{MDM}(\tau_{v_1}, \dots, \tau_{v_{\deg(v)}}) \\ &= \sum_{v \in V_{int}(T)} \frac{1}{\deg(v)} \sum_{i=1}^{\deg(v)} |\tau_{v_1} - \text{Median}(\tau_{v_1}, \dots, \tau_{v_{\deg(v)}})|, \end{aligned}$$

where, for every $v \in V_{int}(T)$, $v_1, \dots, v_{\deg(v)}$ denote its children and $\tau_{v_1}, \dots, \tau_{v_{\deg(v)}}$ their numbers of descendant nodes.

Notice that $\mathfrak{C}_{D,f}$ becomes larger as the f -sizes of the subtrees rooted at siblings get more different, and therefore it behaves as a balance index for trees, in the same way as, for instance, the Colless index for bifurcating trees: the smaller the value of $\mathfrak{C}_{D,f}(T)$, the more balanced is T relative to the f -size σ_f .

Proposition 4.8. Let $T \in \mathcal{T}_n$ be a phylogenetic tree with root r , and let T_1, \dots, T_k , $k \geq 2$, be the subtrees rooted at the children of r . Then,

$$\mathfrak{C}_{D,f}(T) = \sum_{i=1}^k \mathfrak{C}_{D,f}(T_i) + D(\sigma_f(T_1), \dots, \sigma_f(T_k)).$$

Proof. Let w_1, \dots, w_k be the children of the root r of T . Then (denoting as usual by $v_1, \dots, v_{\deg(v)}$ the children of a generic internal node v of T)

$$\begin{aligned} \mathfrak{C}_{D,f}(T) &= \sum_{v \in V_{int}(T)} bal_{D,f}(v) = \sum_{v \in V_{int}(T)} D(\sigma_f(T_{v_1}), \dots, \sigma_f(T_{v_{\deg(v)}})) \\ &= bal_{D,f}(r) + \sum_{v \in V_{int}(T) \setminus \{r\}} D(\sigma_f(T_{v_1}), \dots, \sigma_f(T_{v_{\deg(v)}})) \\ &= D(\sigma_f(T_{w_1}), \dots, \sigma_f(T_{w_k})) + \sum_{i=1}^k \sum_{v \in V_{int}(T_i)} D(\sigma_f(T_{v_1}), \dots, \sigma_f(T_{v_{\deg(v)}})) \\ &= D(\sigma_f(T_{w_1}), \dots, \sigma_f(T_{w_k})) + \sum_{i=1}^k \sum_{v \in V_{int}(T_i)} D(\sigma_f((T_i)_{v_1}), \dots, \sigma_f((T_i)_{v_{\deg(v)}})) \\ &= D(\sigma_f(T_{w_1}), \dots, \sigma_f(T_{w_k})) + \sum_{i=1}^k \mathfrak{C}_{D,f}(T_i) \end{aligned}$$

as we claimed. \square

Table 4.2 in Section 4.7 also gives the abstract values of $\mathfrak{C}_{D,f}(T)$, for $D = \text{MDM}$, var , and sd , and for every $T \in \mathcal{T}_n^*$ with $n = 2, 3, 4, 5$.

The next result shows that, if we take $D = \text{MDM}$ or $D = \text{sd}$, then any index $\mathfrak{C}_{D,f}$ restricted to only bifurcating trees defines, up to a constant factor, the usual Colless index.

Proposition 4.9. *Let T be a bifurcating tree with n leaves and $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ any function. Then,*

$$\mathfrak{C}_{\text{MDM},f}(T) = \frac{f(0) + f(2)}{2} \cdot C(T), \quad \mathfrak{C}_{\text{sd},f}(T) = \frac{f(0) + f(2)}{\sqrt{2}} \cdot C(T).$$

Proof. Notice that, for every $x, y \in \mathbb{R}$, $\text{MDM}(x, y) = \frac{1}{2}|x - y|$ and $\text{sd}(x, y) = \frac{1}{\sqrt{2}}|x - y|$. Indeed,

$$\text{MDM}(x, y) = \frac{1}{2} \left(\left| x - \frac{x+y}{2} \right| + \left| y - \frac{x+y}{2} \right| \right) = \frac{1}{2} \left(\frac{|x-y|}{2} + \frac{|y-x|}{2} \right) = \frac{|x-y|}{2}$$

and (remember that our variance and standard deviation are the sample versions of these statistics)

$$\begin{aligned} \text{var}(x, y) &= \left(x - \frac{x+y}{2} \right)^2 + \left(y - \frac{x+y}{2} \right)^2 = \left(\frac{x-y}{2} \right)^2 + \left(\frac{y-x}{2} \right)^2 \\ &= 2 \cdot \frac{(x-y)^2}{4} = \frac{(x-y)^2}{2} \end{aligned}$$

and hence

$$\text{sd}(x, y) = \sqrt{\frac{(x-y)^2}{2}} = \frac{|x-y|}{\sqrt{2}}.$$

We shall prove the statement for MDM ; the proof for sd is identical, replacing the 2 in the denominator by $\sqrt{2}$. For every internal node v in a bifurcating tree T , if v_1 and v_2 denote its children,

$$\begin{aligned} \text{bal}_{\text{MDM},f}(v) &= \frac{1}{2} |\sigma_f(T_{v_1}) - \sigma_f(T_{v_2})| \\ &= \frac{1}{2} |((f(0) + f(2))\kappa(v_1) - f(2)) - ((f(0) + f(2))\kappa(v_2) - f(2))| \\ &\quad \text{(by Example 4.3)} \\ &= \frac{f(0) + f(2)}{2} \cdot |\kappa(v_1) - \kappa(v_2)| \end{aligned}$$

and therefore

$$\begin{aligned} \mathfrak{C}_{\text{MDM},f}(T) &= \sum_{v \in V_{\text{int}}(T)} \text{bal}_{\text{MDM},f}(v) = \frac{f(0) + f(2)}{2} \cdot \sum_{v \in V_{\text{int}}(T)} |\kappa(v_1) - \kappa(v_2)| \\ &= \frac{f(0) + f(2)}{2} \cdot C(T), \end{aligned}$$

as we claimed. \square

If we define the *quadratic Colless index* of a bifurcating tree T as

$$C^{(2)}(T) = \sum_{v \in V_{int}(T)} (\kappa(v_1) - \kappa(v_2))^2$$

(where, for every $v \in V_{int}(T)$, v_1, v_2 denote its children), then a similar argument proves the following result.

Proposition 4.10. *Let T be a bifurcating tree with n leaves and $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ any function. Then,*

$$\mathfrak{C}_{var,f}(T) = \frac{(f(0) + f(2))^2}{2} \cdot C^{(2)}(T).$$

Proof. As we have seen in the proof of the previous proposition, for every $x, y \in \mathbb{R}$, $\text{var}(x, y) = \frac{1}{2}(x - y)^2$. Now, for every internal node v in a bifurcating tree T , if v_1 and v_2 denote its children,

$$\begin{aligned} \text{bal}_{var,f}(v) &= \frac{1}{2} \left(\sigma_f(T_{v_1}) - \sigma_f(T_{v_2}) \right)^2 \\ &= \frac{1}{2} \left(((f(0) + f(2))\kappa(v_1) - f(2)) - ((f(0) + f(2))\kappa(v_2) - f(2)) \right)^2 \\ &\quad \text{(by Example 4.3)} \\ &= \frac{(f(0) + f(2))^2}{2} \cdot (\kappa(v_1) - \kappa(v_2))^2 \end{aligned}$$

and therefore

$$\begin{aligned} \mathfrak{C}_{var,f}(T) &= \sum_{v \in V_{int}(T)} \text{bal}_{var,f}(v) = \frac{(f(0) + f(2))^2}{2} \cdot \sum_{v \in V_{int}(T)} (\kappa(v_1) - \kappa(v_2))^2 \\ &= \frac{(f(0) + f(2))^2}{2} \cdot C^{(2)}(T), \end{aligned}$$

as we claimed. □

As far as the cost of computing Colless-like indices goes, we have the following result.

Proposition 4.11. *If the cost of computing $D(x_1, \dots, x_k)$ is in $O(k)$ and the cost of computing each $f(k)$ is at most in $O(k)$, then, for every $T \in \mathcal{T}_n^*$, the cost of computing $\mathfrak{C}_{D,f}(T)$ is in $O(n)$.*

Proof. Assume that every $f(k)$ is computed in time at most $O(k)$. For every $k \geq 2$, let m_k be the number of internal nodes in T of out-degree k . Since, by Lemma 4.2, the sizes $\sigma_f(T_v)$ satisfy that if v has children v_1, \dots, v_k , then $\sigma_f(T_v) = \sum_{i=1}^k \sigma_f(T_{v_i}) + f(k)$, we can compute the whole vector $(\sigma_f(T_v))_{v \in V(T)}$ in time $O(n + \sum_{k \geq 2} m_k \cdot k) = O(n)$ by traversing the tree in post-order.

Assume now that $D(x_1, \dots, x_k)$ can be computed in time $O(k)$. Then, for every internal node v of out-degree k , $bal_{D,f}(v) = D(\sigma_f(T_{v_1}), \dots, \sigma_f(T_{v_k}))$ can be computed in time $O(k)$, by simply reading the k f -sizes of its children (which are already computed) and applying D to them. This shows that the whole vector $(bal_{D,f}(v))_{v \in V(T)}$ can be computed again in time $O(\sum_{k \geq 2} m_k \cdot k) = O(n)$. Finally, we compute $\mathfrak{C}_{D,f}(T)$ by adding up the entries of $(bal_{D,f}(v))_{v \in V(T)}$, which still can be done in time $O(n)$. \square

The dissimilarities mentioned previously in this section can be computed in a number of sums and multiplications that is linear in the length of the input vector, and the specific functions f that we shall consider in the next section, basically exponentials and logarithms, can be approximated to any desired precision in constant time by using addition and look-up tables [115].

4.3 Sound Colless-like indices

It is clear that for every dissimilarity D , for every function $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ and for every fully symmetric tree FS_{n_1, \dots, n_k} , $\mathfrak{C}_{D,f}(FS_{n_1, \dots, n_k}) = 0$ because $bal_{D,f}(v) = 0$ for every $v \in V_{int}(FS_{n_1, \dots, n_k})$. Indeed, notice that, for every node w in FS_{n_1, \dots, n_k} of depth l , the rooted subtree of FS_{n_1, \dots, n_k} rooted at w is isomorphic to FS_{n_{l+1}, \dots, n_k} (a single node FS_- if $\sigma(w) = k$, because the leaves of FS_{n_1, \dots, n_k} are exactly its nodes of maximum depth, k). This implies in particular that, for every internal node v of FS_{n_1, \dots, n_k} , the subtrees rooted at its children are pairwise isomorphic and therefore they have the same f -size.

Now, we shall say that a Colless-like index $\mathfrak{C}_{D,f}$ is *sound* when the converse implication is true:

Definition 4.12. *A Colless-like index $\mathfrak{C}_{D,f}$ is sound when, for every $T \in \mathcal{T}^*$, $\mathfrak{C}_{D,f}(T) = 0$ if, and only if, T is fully symmetric.*

In other words, $\mathfrak{C}_{D,f}$ is sound when, according to it, the most balanced trees are exactly the fully symmetric trees.

The Colless index C and its quadratic version $C^{(2)}$ are sound for *bifurcating* trees. Unfortunately, this is not always so for Colless-like indices for multi-furcating trees. It is not so even for direct generalizations of C or $C^{(2)}$. For instance, $\mathfrak{C}_{MDM,\kappa}$, $\mathfrak{C}_{sd,\kappa}$ and $\mathfrak{C}_{var,\kappa}$, where κ denotes the number of leaves, are not sound; neither are $\mathfrak{C}_{MDM,\tau}$, $\mathfrak{C}_{sd,\tau}$ and $\mathfrak{C}_{var,\tau}$, where τ denotes the number of nodes; and they are not sound even when replacing τ by θ , the usual size, which is simply $\tau - 1$. For example, the tree T in Fig. 4.2 is not fully symmetric, but $\mathfrak{C}_{MDM,\kappa}(T) = \mathfrak{C}_{var,\kappa}(T) = \mathfrak{C}_{sd,\kappa}(T) = \mathfrak{C}_{MDM,\tau}(T) = \mathfrak{C}_{var,\tau}(T) = \mathfrak{C}_{sd,\tau}(T) = 0$, because for every internal node in T , the subtrees rooted at its children have all the same number of leaves and the same order and size.

The following lemma shows that the soundness of $\mathfrak{C}_{D,f}(T) = 0$ does not depend on D , but only on f .

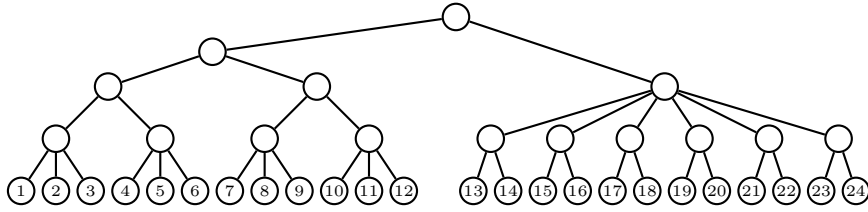


Figure 4.2: A non fully symmetric tree T with $\mathfrak{C}_{\text{MDM},\kappa}(T) = \mathfrak{C}_{\text{var},\kappa}(T) = \mathfrak{C}_{sd,\kappa}(T) = \mathfrak{C}_{\text{MDM},\tau}(T) = \mathfrak{C}_{\text{var},\tau}(T) = \mathfrak{C}_{sd,\tau}(T) = 0$.

Lemma 4.13. $\mathfrak{C}_{D,f}$ is sound if, and only if, $\sigma_f(T_1) \neq \sigma_f(T_2)$ for every pair of different fully symmetric trees T_1, T_2 .

Proof. The “only if” implication follows from the fact that if there exist two different (i.e., non isomorphic) fully symmetric trees T_1, T_2 such that $\sigma_f(T_1) = \sigma_f(T_2)$, then the tree $T = T_1 \star T_2$ is not fully symmetric, but

$$\mathfrak{C}_{D,f}(T) = \mathfrak{C}_{D,f}(T_1) + \mathfrak{C}_{D,f}(T_2) + D(\sigma_f(T_1), \sigma_f(T_2)) = 0.$$

Conversely, assume that, for every pair of fully symmetric trees T_1, T_2 , if $\sigma_f(T_1) = \sigma_f(T_2)$ then $T_1 = T_2$. We shall prove by complete induction on n that if T is a tree with n leaves such that $\mathfrak{C}_{D,f}(T) = 0$, then T is fully symmetric. The base case is obvious, because if T has only one leaf, it is the tree consisting of a single node and therefore it is fully symmetric. Now, assume that $n > 1$ and hence that T has depth at least 1. Let T_1, \dots, T_{n_1} , $n_1 \geq 2$, be its subtrees rooted at the children of its root, so that $T = T_1 \star \dots \star T_{n_1}$. Then,

$$0 = \mathfrak{C}_{D,f}(T) = \sum_{i=1}^{n_1} \mathfrak{C}_{D,f}(T_i) + D(\sigma_f(T_1), \dots, \sigma_f(T_{n_1}))$$

implies, on the one hand, that $\mathfrak{C}_{D,f}(T_1) = \dots = \mathfrak{C}_{D,f}(T_{n_1}) = 0$, and hence, by induction, that T_1, \dots, T_{n_1} are fully symmetric, and, on the other hand, that $D(\sigma_f(T_1), \dots, \sigma_f(T_{n_1})) = 0$, and hence that $\sigma_f(T_1) = \dots = \sigma_f(T_{n_1})$, which, by assumption, implies that $T_1 = \dots = T_{n_1}$. So, T_1, \dots, T_{n_1} are isomorphic copies of the same tree FS_{n_2, \dots, n_k} and therefore T is a fully symmetric tree of the form $FS_{n_1, n_2, \dots, n_k}$. \square

The following problem now arises:

Problem. To find functions $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $\mathfrak{C}_{D,f}$ is sound.

Unfortunately, many natural functions f do not define sound Colless-like indices, as the following examples show.

Example 4.14. If $f(n) = an^2 + bn + c$, for any a, b, c , then $\mathfrak{C}_{D,f}$ is not sound, because, for example, $\sigma_f(FS_{2,2,2,7}) = \sigma_f(FS_{14,4}) = 420a + 70b + 71c$.

Example 4.15. If $f(n) = n^d$, for any $d \geq 0$, then $\mathfrak{C}_{D,f}$ is not sound. Indeed, for every $d \geq 3$ (the case when $d \leq 2$ is a particular case of the last example), take

- $k = 2^d + 1$ and $l = 2$;
- $n_i = 2^{(d-1)^i d^{k-i-1}}$ for $i = 1, \dots, k-1$;
- $n_k = 2$;
- $m_1 = 2^{(d-1)d^{k-2}+1}$;
- $m_2 = 2^{((d-1)^2(d^{k-2}-(d-1)^{k-2})+d-1)/d}$; notice that this exponent is an integer number, because k is odd and therefore d divides $(d-1)^k + 1$.

Then

$$n_1 \cdots n_{i-1} \cdot n_i^d = n_1^d$$

and hence, on the one hand,

$$\begin{aligned} n_1^d + \cdots + n_1 \cdots n_{k-2} \cdot n_{k-1}^d &= (k-1)n_1^d = 2^d \cdot 2^{(d-1)d^{k-1}} \\ &= \left(2^{1+(d-1)d^{k-2}}\right)^d = m_1^d, \end{aligned}$$

and, on the other hand,

$$\begin{aligned} n_1 \cdots n_{k-1} \cdot n_k^d &= n_1^{\frac{(1-(\frac{d-1}{d})^{k-1})}{(1-\frac{d-1}{d})}} \cdot n_k^d = n_1^{\frac{d^{k-1}-(d-1)^{k-1}}{d^{k-2}}} n_k^d \\ &= 2^{(d-1)(d^{k-1}-(d-1)^{k-1})+d} = m_1 m_2^d. \end{aligned}$$

Therefore, $\sigma_{n^d}(FS_{n_1, \dots, n_k}) = \sigma_{n^d}(FS_{m_1, m_2})$.

Of course, for any given d there may exist “smaller” counterexamples: for instance, $\sigma_{n^3}(FS_{2,10,4}) = \sigma_{n^3}(FS_{6,8}) = 3288$ and $\sigma_{n^4}(FS_{2,6,2,3}) = \sigma_{n^4}(FS_{8,3}) = 4744$.

Example 4.16. If $f(n) = \log_a(n)$ (for some $a > 1$) when $n > 0$, and $f(0) = 0$, then $\mathfrak{C}_{D,f}$ is not sound: for instance, $\sigma_f(FS_{2,2}) = \sigma_f(FS_8) = \log_a(8)$. In a similar way, if $f(n) = \log_a(n+1)$ (for some $a > 1$), then $\mathfrak{C}_{D,f}$ is not sound, either: for instance, $\sigma_f(FS_{2,3,3}) = \sigma_f(FS_{5,7}) = \log_a(196608)$.

On the positive side, we shall show now two functions that define sound indices. The following lemmas will be useful to prove it.

Lemma 4.17. For every $k, l \geq 2$ and $n_1, n_2, \dots, n_k, m_1, m_2, \dots, m_l \geq 2$, if $\sigma_f(FS_{n_1, \dots, n_k}) = \sigma_f(FS_{m_1, \dots, m_l})$, $n_1 \cdots n_k = m_1 \cdots m_l$, and $n_k = m_l$, then $\sigma_f(FS_{n_1, \dots, n_{k-1}}) = \sigma_f(FS_{m_1, \dots, m_{l-1}})$.

Proof. If $n_1 \cdots n_k = m_1 \cdots m_l$ and $n_k = m_l$, then $n_1 \cdots n_{k-1} = m_1 \cdots m_{l-1}$. If, moreover, $\sigma_f(FS_{n_1, n_2, \dots, n_k}) = \sigma_f(FS_{m_1, m_2, \dots, m_l})$, that is,

$$\begin{aligned} n_1 \cdots n_k f(0) + n_1 \cdots n_{k-1} f(n_k) + n_1 \cdots n_{k-2} f(n_{k-1}) + \cdots + f(n_1) \\ = m_1 \cdots m_l f(0) + m_1 \cdots m_{l-1} f(m_l) + m_1 \cdots m_{l-2} f(m_{l-1}) + \cdots + f(m_1), \end{aligned}$$

then

$$\begin{aligned} n_1 \cdots n_{k-2} f(n_{k-1}) + \cdots + n_1 f(n_2) + f(n_1) \\ = m_1 \cdots m_{l-2} f(m_{l-1}) + \cdots + m_1 f(m_2) + f(m_1) \end{aligned}$$

and hence

$$\begin{aligned} \sigma_f(FS_{n_1, n_2, \dots, n_{k-1}}) \\ = n_1 \cdots n_{k-1} f(0) + n_1 \cdots n_{k-2} f(n_{k-1}) + \cdots + n_1 f(n_2) + f(n_1) \\ = m_1 \cdots m_{l-1} f(0) + m_1 \cdots m_{l-2} f(m_{l-1}) + \cdots + m_1 f(m_2) + f(m_1) \\ = \sigma_f(FS_{m_1, \dots, m_{l-1}}) \end{aligned}$$

as we claimed. □

Lemma 4.18. *If $n_1, \dots, n_k \geq 2$, then*

$$1 + n_1 + n_1 n_2 + \cdots + n_1 \cdots n_{k-1} < n_1 \cdots n_k.$$

Proof. By induction on k . If $k = 1$, the statement says that $1 < n_1$, which is true by assumption. Assume now that the statement is true for any $n_1, \dots, n_k \geq 2$, and let $n_{k+1} \geq 2$. Then,

$$\begin{aligned} 1 + n_1 + n_1 n_2 + \cdots + n_1 \cdots n_{k-1} + n_1 \cdots n_k < n_1 \cdots n_k + n_1 \cdots n_k \\ = 2n_1 \cdots n_k \leq n_1 \cdots n_k \cdot n_{k+1}. \end{aligned}$$

□

Proposition 4.19. *If $f(n) = e^n$, then $\mathfrak{C}_{D,f}$ is sound.*

Proof. Assume that there exist two non-isomorphic fully symmetric trees FS_{n_1, \dots, n_k} and FS_{m_1, \dots, m_l} such that

$$\sigma_{e^n}(FS_{n_1, \dots, n_k}) = \sigma_{e^n}(FS_{m_1, \dots, m_l}),$$

that is, such that

$$\begin{aligned} n_1 \cdots n_k + n_1 \cdots n_{k-1} e^{n_k} + \cdots + n_1 e^{n_2} + e^{n_1} \\ = m_1 \cdots m_l + m_1 \cdots m_{l-1} e^{m_l} + \cdots + m_1 e^{m_2} + e^{m_1}. \end{aligned} \tag{4.1}$$

Assume that l is the smallest depth of a fully symmetric tree with e^n -size equal to the e^n -size of another fully symmetric tree non-isomorphic to it.

Since e is transcendental, equality (4.1) implies the equality of polynomials in $\mathbb{Z}[x]$

$$\begin{aligned} n_1 \cdots n_k + n_1 \cdots n_{k-1} x^{n_k} + \cdots + n_1 x^{n_2} + x^{n_1} \\ = m_1 \cdots m_l + m_1 \cdots m_{l-1} x^{m_l} + \cdots + m_1 x^{m_2} + x^{m_1}. \end{aligned}$$

If $l = 1$, the right-hand side polynomial is simply $m_1 + x^{m_1}$ and then the equality of polynomials implies that $k = 1$ and $n_1 = m_1$, which contradicts the assumption that $FS_{n_1, \dots, n_k} \neq FS_{m_1, \dots, m_l}$. Now assume that $l \geq 2$. This equality of polynomials implies the equality of their independent terms: $n_1 \cdots n_k = m_1 \cdots m_l$. On the other hand, the non-zeroth power of x with the largest coefficient in the left-hand side polynomial is x^{n_k} (because all coefficients are non-negative, and, by Lemma 4.18, $n_1 \cdots n_{k-1}$ alone is larger than the sum $n_1 \cdots n_{k-2} + \cdots + n_1 + 1$ of all other coefficients of non-zeroth powers of x) and, by the same reason, the non-zeroth power of x with the largest coefficient in the right-hand side polynomial is x^{m_l} . The equality of polynomials implies then that $n_k = m_l$ and hence, by Lemma 4.17, that $\sigma_{e^n}(FS_{n_1, \dots, n_{k-1}}) = \sigma_{e^n}(FS_{m_1, \dots, m_{l-1}})$, against the assumption on l . We reach a contradiction that implies that there do not exist any two non-isomorphic fully symmetric trees with the same e^n -size. By Lemma 4.13, this implies that \mathfrak{C}_{D, e^n} is sound. \square

The same argument shows that $\mathfrak{C}_{D, f}$ is sound for every exponential function $f(n) = r^n$ with base r a transcendental real number. However, if r is not transcendental, then \mathfrak{C}_{D, r^n} need not be sound. For instance, $\sigma_{2^n}(FS_{2,3}) = \sigma_{2^n}(FS_{3,2}) = 26$ and $\sigma_{\sqrt{2}^n}(FS_{8,10}) = \sigma_{\sqrt{2}^n}(FS_{12,8}) = 352$.

Proposition 4.20. *If $f(n) = \ln(n + e)$, then $\mathfrak{C}_{D, f}$ is sound.*

Proof. The argument is similar to that of the previous proof. Let $f(n) = \ln(n + e)$ and assume that there exist two non-isomorphic fully symmetric trees FS_{n_1, \dots, n_k} and FS_{m_1, \dots, m_l} such that $\sigma_f(FS_{n_1, \dots, n_k}) = \sigma_f(FS_{m_1, \dots, m_l})$, that is, such that

$$\begin{aligned} n_1 \cdots n_k + n_1 \cdots n_{k-1} \ln(n_k + e) + \cdots + \ln(n_1 + e) \\ = m_1 \cdots m_l + m_1 \cdots m_{l-1} \ln(m_l + e) + \cdots + \ln(m_1 + e). \end{aligned} \quad (4.2)$$

Assume that l is the smallest depth of a fully symmetric tree with f -size equal to the f -size of a fully symmetric tree non-isomorphic to it.

Applying the exponential function to both sides of equality (4.2), we obtain

$$\begin{aligned} e^{n_1 \cdots n_k} (n_k + e)^{n_1 \cdots n_{k-1}} \cdots (n_2 + e)^{n_1} (n_1 + e) \\ = e^{m_1 \cdots m_l} (m_l + e)^{m_1 \cdots m_{l-1}} \cdots (m_2 + e)^{m_1} (m_1 + e). \end{aligned}$$

Since e is transcendental, this implies the equality of polynomials in $\mathbb{Z}[x]$

$$\begin{aligned} x^{n_1 \cdots n_k} (n_k + x)^{n_1 \cdots n_{k-1}} \cdots (n_2 + x)^{n_1} (n_1 + x) \\ = x^{m_1 \cdots m_l} (m_l + x)^{m_1 \cdots m_{l-1}} \cdots (m_2 + x)^{m_1} (m_1 + x), \end{aligned}$$

which, since $n_1, \dots, n_k, m_1, \dots, m_l \geq 2$, on its turn implies the equalities

$$\begin{aligned} x^{n_1 \cdots n_k} &= x^{m_1 \cdots m_l}, \text{ i.e., } n_1 \cdots n_k = m_1 \cdots m_l, \\ (x + n_k)^{n_1 \cdots n_{k-1}} \cdots (x + n_2)^{n_1} (x + n_1) \\ &= (x + m_l)^{m_1 \cdots m_{l-1}} \cdots (x + m_2)^{m_1} (x + m_1). \end{aligned}$$

If $l = 1$, the right-hand side polynomial in the second equality is simply $x + m_1$ and then this equality of polynomials implies that $k = 1$ and $n_1 = m_1$, which contradicts the assumption that $FS_{n_1, \dots, n_k} \neq FS_{m_1, \dots, m_l}$. Now assume that $l \geq 2$. From the first equality we know that $n_1 \cdots n_k = m_1 \cdots m_l$. But, the root of the left-hand side polynomial in the second equality with largest multiplicity is $-n_k$ (because, by Lemma 4.18, the contribution of $n_1 \cdots n_{k-1}$ to the multiplicity of $-n_k$ through the factor $(x + n_k)^{n_1 \cdots n_{k-1}}$ is, by itself, greater than the degree of $(x + n_{k-1})^{n_1 \cdots n_{k-2}} \cdots (x + n_2)^{n_1} (x + n_1)$) and, similarly, the root of the right-hand side polynomial in the second equality with largest multiplicity is $-m_l$. Therefore, the equality of both polynomials implies that $n_k = m_l$ and hence, by Lemma 4.17, $\sigma_f(FS_{n_1, \dots, n_{k-1}}) = \sigma_f(FS_{m_1, \dots, m_{l-1}})$, against the assumption on l . As in the previous proof, this contradiction implies that $\mathfrak{C}_{D,f}$, with $f(n) = \ln(n + e)$, is sound. \square

The same argument proves that, for every transcendental number $r > 1$, the function $f(n) = \log_r(n + r)$ defines sound indices $\mathfrak{C}_{D,f}$. However, if r is not transcendental, then such a $\mathfrak{C}_{D,f}$ need not be sound. For instance, $\sigma_{\log_2(n+2)}(FS_{9,6}) = \sigma_{\log_2(n+2)}(FS_{20,2}) = 81 + \log_2(11)$.

Summarizing, both functions $f(n) = \ln(n + e)$ and $f(n) = e^n$ define, for every dissimilarity D , a Colless-like index $\mathfrak{C}_{D,f}$ that reaches its minimum value on each \mathcal{T}_n , 0, at exactly the fully symmetric trees.

4.4 Maximally unbalanced trees

In this section we give the maximum values of $\mathfrak{C}_{D,f}$ on \mathcal{T}_n when $D = \text{MDM}$, var or sd and $f(n) = \ln(n + e)$ or $f(n) = e^n$; we devote a subsection to each one of these two functions f . Recall these maxima define the range of each $\mathfrak{C}_{D,f}$ on \mathcal{T}_n , and, dividing by them, we can define normalized Colless-like indices that can be used to compare the balance of trees with different numbers of leaves.

4.4.1 The case of $f(n) = \ln(n + e)$

The function $f(n) = \ln(n + e)$ is covered by the following theorem.

Theorem 4.21. *Let f be a function $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $0 < f(k) < f(k - 1) + f(2)$, for every $k \geq 3$. Then, for every $n \geq 2$, the indices $\mathfrak{C}_{\text{MDM},f}$, $\mathfrak{C}_{sd,f}$ and $\mathfrak{C}_{\text{var},f}$ achieve their maximum values on \mathcal{T}_n exactly at the combs K_n . These*

maximum values are, respectively,

$$\begin{aligned}\mathfrak{C}_{\text{MDM},\sigma_f}(K_n) &= \frac{f(0) + f(2)}{4}(n-1)(n-2), \\ \mathfrak{C}_{sd,\sigma_f}(K_n) &= \frac{f(0) + f(2)}{2\sqrt{2}}(n-1)(n-2), \\ \mathfrak{C}_{\text{var},\sigma_f}(K_n) &= \frac{(f(0) + f(2))^2}{12}(n-1)(n-2)(2n-3).\end{aligned}$$

It is straightforward to check that the function $f(n) = \ln(n+e)$ satisfies the hypothesis of Theorem 4.21. Indeed, for the inequality $f(k) < f(k-1) + f(2)$, notice that $\ln(k+e) < \ln(k+e-1) + \ln(2)$ if, and only if, $k+e < 2(k+e-1)$, and this last inequality is true for every $k \in \mathbb{N}$ because $e > 2$. Therefore, $\mathfrak{C}_{\text{MDM},\ln(n+e)}$, $\mathfrak{C}_{\text{var},\ln(n+e)}$, and $\mathfrak{C}_{sd,\ln(n+e)}$ take their maximum values on \mathcal{T}_n^* at the comb K_n . In other words, the combs are the most unbalanced trees according to these indices. Table 4.3 in Section 4.7 gives the values of $\mathfrak{C}_{\text{MDM},\ln(n+e)}$, $\mathfrak{C}_{\text{var},\ln(n+e)}$, and $\mathfrak{C}_{sd,\ln(n+e)}$ on \mathcal{T}_n^* , for $n = 2, 3, 4, 5$, and the positions of the different trees in each \mathcal{T}_n^* according to the increasing order of the corresponding index.

We have split the proof of Theorem 4.21 into 3 subsections, one for each dissimilarity. Throughout this subsection, $f : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ stands for any mapping such that $0 < f(k) < f(k-1) + f(2)$ for every $k \geq 3$. Notice that if f satisfies this condition then $f(k) > 0$ not only for every $k \geq 3$, but also for $k = 2$, because $0 < f(3) < 2f(2)$. Moreover, such a mapping f also satisfies that

$$f(k) < f(k-j) + j \cdot f(2) \quad \text{for every } 1 \leq j \leq k-2$$

because while $k - (i-1) \geq 3$, the inequalities $f(k - (i-1)) < f(k-i) + f(2)$ give rise to the sequence of inequalities

$$\begin{aligned}f(k) &< f(k-1) + f(2) < f(k-2) + f(2) + f(2) = f(k-2) + 2f(2) \\ &< f(k-3) + f(2) + 2f(2) = f(k-3) + f(2) + 3f(2) < \dots\end{aligned}$$

To simplify the notations, we shall denote σ_f by σ .

Proof of the thesis of Theorem 4.21 for $\mathfrak{C}_{\text{MDM},f}$

To further simplify the notations, we shall denote in this subsection $\mathfrak{C}_{\text{MDM},f}$ by \mathfrak{C} and the function $\text{bal}_{\text{MDM},f}$ on a tree T by bal_T or simply by bal when it is not necessary to specify the tree. We shall often use, without any further mention, that $\text{MDM}(x, y) = |x - y|/2$, which was established in the proof of Proposition 4.9.

Lemma 4.22. *For every $(x_1, \dots, x_{2n+1}) \in \mathbb{R}^{2n+1}$ with $n \geq 1$, if x_i is the median of $\{x_1, \dots, x_n\}$, then*

$$\text{MDM}(x_1, \dots, x_{2n+1}) \leq \text{MDM}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{2n+1}).$$

Moreover, this inequality is strict unless $x_1 = \dots = x_{2n+1}$.

Proof. After rearranging x_1, \dots, x_{2n+1} if necessary, we assume that

$$x_1 \leq \dots \leq x_n \leq x_{n+1} \leq x_{n+2} \leq \dots \leq x_{2n+1}, \quad (4.3)$$

in which case their median is x_{n+1} , and the median of $x_1, \dots, x_n, x_{n+2}, \dots, x_{2n+1}$ is $M = (x_n + x_{n+2})/2$. We want to prove that

$$\text{MDM}(x_1, \dots, x_{2n+1}) \leq \text{MDM}(x_1, \dots, x_n, x_{n+2}, \dots, x_{2n+1}).$$

This inequality is true, because

$$\begin{aligned} \text{MDM}(x_1, \dots, x_{2n+1}) &= \frac{1}{2n+1} \sum_{i=1}^{2n+1} |x_i - x_{n+1}| \\ &= \frac{1}{2n+1} \left(\sum_{i=1}^n (x_{n+1} - x_i) + \sum_{i=n+2}^{2n+1} (x_i - x_{n+1}) \right) = \frac{1}{2n+1} \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right) \\ \text{MDM}(x_1, \dots, x_n, x_{n+2}, \dots, x_{2n+1}) &= \frac{1}{2n} \left(\sum_{i=1}^n |x_i - M| + \sum_{i=n+2}^{2n+1} |x_i - M| \right) \\ &= \frac{1}{2n} \left(\sum_{i=1}^n (M - x_i) + \sum_{i=n+2}^{2n+1} (x_i - M) \right) = \frac{1}{2n} \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right), \end{aligned}$$

and, clearly,

$$\frac{1}{2n+1} \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right) \leq \frac{1}{2n} \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right).$$

Now, assume that this inequality is an equality:

$$\begin{aligned} \frac{1}{2n+1} \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right) &= \frac{1}{2n} \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right) \\ \iff 2n \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right) &= (2n+1) \left(\sum_{i=n+2}^{2n+1} x_i - \sum_{i=1}^n x_i \right) \\ \iff \sum_{i=1}^n x_i &= \sum_{i=n+2}^{2n+1} x_i \end{aligned}$$

But the inequalities (4.3) imply that

$$\sum_{i=1}^n x_i \leq n \cdot x_{n+1} \leq \sum_{i=n+2}^{2n+1} x_i$$

and then the equality between the left-hand side and the right-hand side sums, together with the inequalities (4.3), imply that $x_1 = \dots = x_n = x_{n+1} = x_{n+1} = \dots = x_{2n+1}$. \square

Unfortunately, the thesis of this lemma is false for vectors of numbers of even length. For instance, consider the vector $(1, 1, 2, 2)$. If we remove any single element, its MDM decreases:

$$\text{MDM}(1, 1, 2, 2) = \frac{1}{2}, \quad \text{MDM}(1, 1, 2) = \text{MDM}(1, 2, 2) = \frac{1}{3}.$$

But, providentially, we can always increase the MDM of an even quantity of real numbers by removing *two* of them.

Lemma 4.23. *For every $(x_1, \dots, x_{2n}) \in \mathbb{R}^{2n}$ with $n \geq 2$, if x_i, x_j , with $i < j$, are the middle values of $\{x_1, \dots, x_{2n}\}$, then*

$$\text{MDM}(x_1, \dots, x_{2n}) \leq \text{MDM}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_{2n}).$$

Moreover, this inequality is strict unless $\{x_1, \dots, x_{2n}\}$ consists either of $2n$ copies of a single element or of n copies of two different elements.

Proof. After rearranging x_1, \dots, x_{2n} if necessary, we assume that $x_1 \leq \dots \leq x_{2n}$, so that their middle values are x_n, x_{n+1} , and hence their median is $M = (x_n + x_{n+1})/2$ and the median of $x_1, \dots, x_{n-1}, x_{n+2}, \dots, x_{2n}$ is $M' = (x_{n-1} + x_{n+2})/2$. We want to prove that

$$\text{MDM}(x_1, \dots, x_{2n}) \leq \text{MDM}(x_1, \dots, x_{n-1}, x_{n+2}, \dots, x_{2n}).$$

And, indeed,

$$\begin{aligned} \text{MDM}(x_1, \dots, x_{2n}) &\leq \text{MDM}(x_1, \dots, x_{n-1}, x_{n+2}, \dots, x_{2n}) \\ \iff \frac{1}{2n} \sum_{i=1}^{2n} |x_i - M| &\leq \frac{1}{2n-2} \left(\sum_{i=1}^{n-1} |x_i - M'| + \sum_{i=n+2}^{2n} |x_i - M'| \right) \\ \iff (2n-2) \left(\sum_{i=1}^n (M - x_i) + \sum_{i=n+1}^{2n} (x_i - M) \right) \\ &\leq 2n \left(\sum_{i=1}^{n-1} (M' - x_i) + \sum_{i=n+2}^{2n} (x_i - M') \right) \\ \iff (n-1) \left(\sum_{i=n+1}^{2n} x_i - \sum_{i=1}^n x_i \right) &\leq n \left(\sum_{i=n+2}^{2n} x_i - \sum_{i=1}^{n-1} x_i \right) \\ &= n \left(\sum_{i=n+1}^{2n} x_i - \sum_{i=1}^n x_i \right) - n(x_{n+1} - x_n) \\ \iff n(x_{n+1} - x_n) &\leq \sum_{i=n+1}^{2n} x_i - \sum_{i=1}^n x_i = \sum_{i=1}^n (x_{n+i} - x_{n+1-i}) \end{aligned}$$

and this last inequality is true because, since $x_1 \leq \dots \leq x_{2n}$, $x_{n+1} - x_n \leq x_{n+i} - x_{n+1-i}$ for every $i = 1, \dots, n$. Moreover, these inequalities entail that

$$n(x_{n+1} - x_n) = \sum_{i=1}^n (x_{n+i} - x_{n+1-i})$$

if, and only if, $x_{n+1} - x_n = x_{n+i} - x_{n+1-i}$ for every $i = 1, \dots, n$. Since $x_{n+i} \geq x_{n+1}$ and $x_{n+1-i} \leq x_n$ for each $i = 1, \dots, n$, this last condition is equivalent to $x_1 = \dots = x_n$ and $x_{n+1} = \dots = x_{2n}$. \square

Lemma 4.24. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $f(k) > 0$, for every $k \geq 2$, and let T be a tree of the form $T_1 \star \dots \star T_k$, with $k \geq 3$ (see Fig. 1.1).*

- (a) *If k is an odd number and if $\sigma(T_1)$ is the median of $\{\sigma(T_1), \dots, \sigma(T_k)\}$, then the tree $T' = T_1 \star (T_2 \star \dots \star T_k)$ (cf. Fig. 4.3) satisfies that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*
- (b) *If k is an even number and if $\sigma(T_1), \sigma(T_2)$ are the middle values of the set $\{\sigma(T_1), \dots, \sigma(T_k)\}$, with $\sigma(T_1) \leq \sigma(T_2)$, then the tree $T'' = T_1 \star (T_2 \star (T_3 \star \dots \star T_k))$ (cf. Fig. 4.3) satisfies that $\mathfrak{C}(T'') > \mathfrak{C}(T)$.*

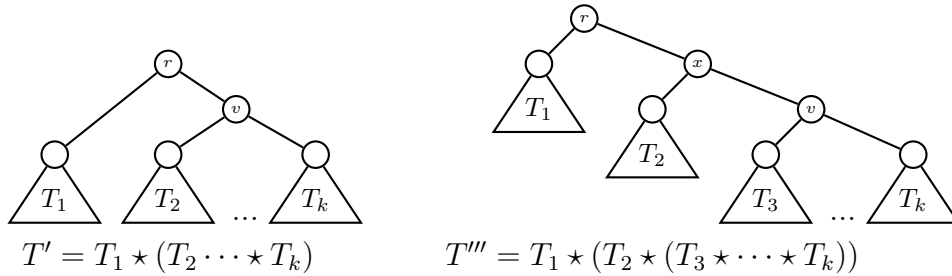


Figure 4.3: The trees T' and T'' in Lemma 4.24.

Proof. Let $t_i = \sigma(T_i)$, for every $i = 1, \dots, k$.

As to (a), we assume that t_1 is the median of $\{t_1, \dots, t_k\}$. The only nodes in T or T' with different *bal* value in both trees are the roots r and the new node v in T' . Therefore,

$$\begin{aligned}
 \mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(v) + \text{bal}_{T'}(r) - \text{bal}_T(r) \\
 &= \text{MDM}(t_2, \dots, t_k) + \frac{1}{2} \left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| - \text{MDM}(t_1, \dots, t_k) \\
 &\geq \frac{1}{2} \left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| > 0,
 \end{aligned}$$

where the first inequality is a consequence of Lemma 4.22 and the second inequality is strict because, since t_1 is the median of $\{t_1, \dots, t_k\}$ and $k \geq 3$, there is some $i \geq 2$ such that $t_i \geq t_1$ and hence

$$\left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| = \sum_{i=2}^k t_i + f(k-1) - t_1 \geq f(k-1) > 0.$$

As far as (b) goes, the only nodes in T or T'' with different bal value in both trees are the roots r and the new nodes x, v in T'' . Therefore,

$$\begin{aligned} \mathfrak{C}(T'') - \mathfrak{C}(T) &= bal_{T''}(v) + bal_{T''}(x) + bal_{T''}(r) - bal_T(r) \\ &= \text{MDM}(t_3, \dots, t_k) + \frac{1}{2} \left| \sum_{i=3}^k t_i + f(k-2) - t_2 \right| \\ &\quad + \frac{1}{2} \left| \sum_{i=2}^k t_i + f(k-2) + f(2) - t_1 \right| - \text{MDM}(t_1, \dots, t_k) \\ &\geq \frac{1}{2} \left(\left| \sum_{i=3}^k t_i + f(k-2) - t_2 \right| + \left| \sum_{i=2}^k t_i + f(k-2) + f(2) - t_1 \right| \right) > 0, \end{aligned}$$

where the first inequality is a consequence of Lemma 4.23 and the second inequality is strict because, by assumption, $t_1 \leq t_2$ and $k \geq 3$, and therefore

$$\left| \sum_{i=2}^k t_i + f(k-2) + f(2) - t_1 \right| = \sum_{i=2}^k t_i + f(k-2) + f(2) - t_1 \geq f(k-2) + f(2) > 0.$$

□

Lemma 4.25. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $f(2) > 0$. Consider the trees T and T' depicted in Fig. 4.4, where, in both trees, all nodes in the path from r to x are bifurcating, and T' is obtained from T by simply interchanging the subtrees T_l and T_{l-1} . If $\sigma(T_l) < \sigma(T_{l-1})$ and $\sigma(T_l) \leq \sigma(T_0)$, then $\mathfrak{C}(T') > \mathfrak{C}(T)$.*

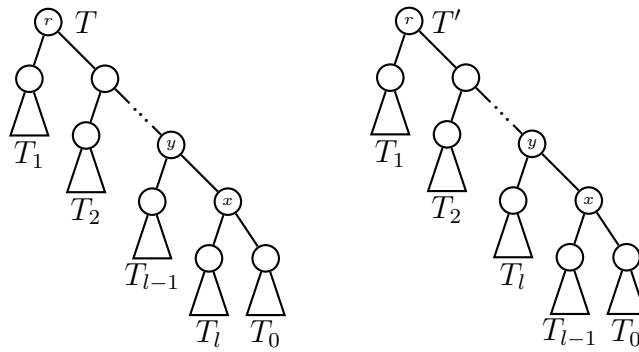


Figure 4.4: The trees T and T' in Lemma 4.25.

Proof. Let $t_i = \sigma(T_i)$, for every $i = 0, \dots, l$, so that $t_l < t_{l-1}$ and $t_l \leq t_0$. Since $\sigma(T_y) = \sigma(T'_y)$, the only nodes in T or T' with different bal value in both trees

are x and y , and therefore,

$$\begin{aligned} \mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(x) + \text{bal}_{T'}(y) - \text{bal}_T(x) - \text{bal}_T(y) \\ &= \frac{1}{2}|t_{l-1} - t_0| + \frac{1}{2}|t_{l-1} + t_0 + f(2) - t_l| - \frac{1}{2}|t_l - t_0| \\ &\quad - \frac{1}{2}|t_l + t_0 + f(2) - t_{l-1}| = (*) \end{aligned}$$

Now we must distinguish two cases:

- If $t_l < t_{l-1} \leq t_0$, then

$$\begin{aligned} (*) &= \frac{1}{2}((t_0 - t_{l-1}) + (t_{l-1} + t_0 + f(2) - t_l) - (t_0 - t_l) \\ &\quad - (t_l + t_0 + f(2) - t_{l-1})) = \frac{1}{2}(t_{l-1} - t_l) > 0 \end{aligned}$$

- If $t_l \leq t_0 \leq t_{l-1}$ and $t_l < t_{l-1}$, then

$$\begin{aligned} (*) &= \frac{1}{2}((t_{l-1} - t_0) + (t_{l-1} + t_0 + f(2) - t_l) - (t_0 - t_l) \\ &\quad - |t_l + t_0 + f(2) - t_{l-1}|) \\ &= \begin{cases} \frac{1}{2}(2t_{l-1} - t_0 + f(2) - (t_l + t_0 + f(2) - t_{l-1})) = \frac{1}{2}(3t_{l-1} - 2t_0 - t_l) \\ \geq \frac{1}{2}(t_{l-1} - t_l) > 0 & \text{(if } t_l + t_0 + f(2) - t_{l-1} \geq 0) \\ \frac{1}{2}(2t_{l-1} - t_0 + f(2) - (t_{l-1} - t_l - t_0 - f(2))) \\ = \frac{1}{2}(t_{l-1} + t_l + 2f(2)) > 0 & \text{(if } t_l + t_0 + f(2) - t_{l-1} \leq 0) \end{cases} \end{aligned}$$

and therefore, in all cases, $\mathfrak{C}(T') - \mathfrak{C}(T) > 0$, as we claimed. \square

Lemma 4.26. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $0 < f(k) < f(k-1) + f(2)$, for every $k \geq 3$, and let T be the tree depicted in Fig. 4.5, where $l \geq 1$, x_1 is the root, all nodes in the path from x_1 to x_l are bifurcating, and $k \geq 3$. Assume moreover that $\sigma(S_1) \leq \sigma(S_2) \leq \dots \leq \sigma(S_l)$.*

(a) *Assume that k is odd and that $\sigma(T_1)$ is the median of $\{\sigma(T_1), \dots, \sigma(T_k)\}$.*

(a.1) *If $\sigma(S_l) \leq \sigma(T_1)$, then the tree T' depicted in Fig. 4.6, obtained by pruning the subtree T_1 and inserting it in the arc ending in x , satisfies that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*

(a.2) *If $\sigma(S_l) > \sigma(T_1)$, then the tree T'' depicted in Fig. 4.6, obtained by pruning the subtree T_1 and inserting it in the arc ending in x_l , satisfies that $\mathfrak{C}(T'') > \mathfrak{C}(T)$.*

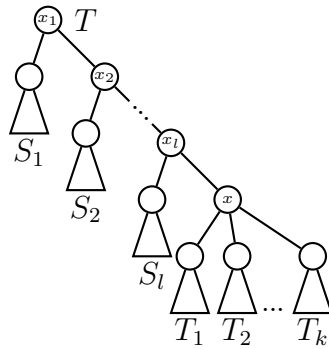


Figure 4.5: The tree T in Lemma 4.26.

(b) Assume that k is even and that $\sigma(T_1), \sigma(T_2)$ are the middle values of the set $\{\sigma(T_1), \dots, \sigma(T_k)\}$.

(b.1) If $\sigma(S_l) \leq \sigma(T_1 \star T_2)$, then the tree T' depicted in Fig. 4.7, obtained by pruning the subtrees T_1 and T_2 and then inserting $T_1 \star T_2$ in the arc ending in x , satisfies that $\mathfrak{C}(T') > \mathfrak{C}(T)$.

(b.2) If $\sigma(S_l) > \sigma(T_1 \star T_2)$, then the tree T'' depicted in Fig. 4.7, obtained by pruning the subtrees T_1 and T_2 and then inserting $T_1 \star T_2$ in the arc ending in x_1 , satisfies that $\mathfrak{C}(T'') > \mathfrak{C}(T)$.

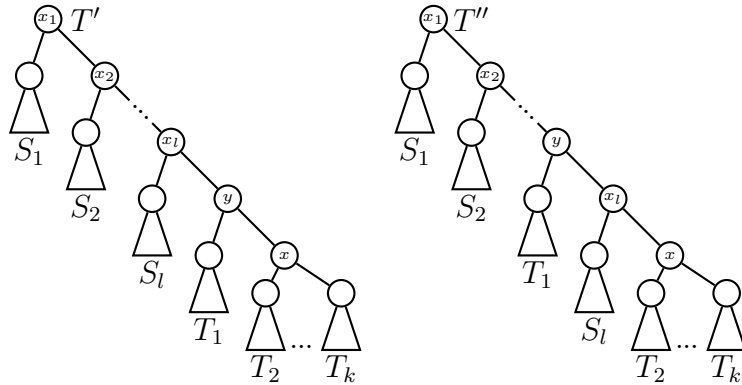


Figure 4.6: The trees T', T'' in Lemma 4.26.(a).

Proof. For every $i = 1, \dots, l$, let x_i denote the parent of the root of S_i in all trees in the statement. Moreover, for every $i = 1, \dots, l$, let $s_i = \sigma(S_i)$ and, for every $i = 1, \dots, k$, let $\sigma(T_i) = t_i$, and let $t = \sum_{i=1}^k t_i$. Recall that we are assuming throughout this proof that $s_1 \leq \dots \leq s_l$.

(a) Assume that $k \geq 3$ is odd and that t_1 is the median of $\{t_1, t_2, \dots, t_k\}$. By Lemma 4.22, this implies that $\text{MDM}(t_1, \dots, t_k) \leq \text{MDM}(t_2, \dots, t_k)$, which is the property that we are actually going to use in the proof.

As far as assertion (a.1) goes, let us assume that $s_l \leq t_1$ and therefore that $s_i \leq t$ for every $i = 1, \dots, l$. In this case, the only nodes in T or T' with different bal value in both trees are x_1, \dots, x_l, x and the new node y in T' . Then:

$$\begin{aligned}
\mathfrak{C}(T') - \mathfrak{C}(T) &= bal_{T'}(x) - bal_T(x) + bal_{T'}(y) + \sum_{i=1}^l (bal_{T'}(x_i) - bal_T(x_i)) \\
&= \text{MDM}(t_2, \dots, t_k) - \text{MDM}(t_1, \dots, t_k) + \frac{1}{2} \left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| \\
&\quad + \frac{1}{2} \sum_{i=1}^l \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\
&\quad \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\
&\geq \frac{1}{2} \sum_{i=1}^l \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\
&\quad \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\
&= \frac{1}{2} \sum_{i=1}^l \left(\left(t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right) \right. \\
&\quad \quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right) \right) \\
&\quad \text{(because } s_i \leq s_l \leq t_1 \leq t \text{ for every } i = 1, \dots, l) \\
&= \frac{l}{2} (f(k-1) + f(2) - f(k)) > 0
\end{aligned}$$

As far as assertion (a.2) goes, let us assume that $s_l > t_1$. Again, the only nodes in T or T'' with different bal value in both trees are x_1, \dots, x_l, x and the new node y . Therefore:

$$\begin{aligned}
\mathfrak{C}(T'') - \mathfrak{C}(T) &= bal_{T''}(x) - bal_T(x) + bal_{T''}(y) + bal_{T''}(x_l) - bal_T(x_l) \\
&\quad + \sum_{i=1}^{l-1} (bal_{T''}(x_i) - bal_T(x_i)) \\
&= \text{MDM}(t_2, \dots, t_k) - \text{MDM}(t_1, \dots, t_k)
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{2} \left| \sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 \right| \\
& + \frac{1}{2} \left| \sum_{i=2}^k t_i + f(k-1) - s_l \right| - \frac{1}{2} |t + f(k) - s_l| \\
& + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\
& \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\
\geq & \frac{1}{2} \left(\left| \sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 \right| - |t + f(k) - s_l| \right) \\
& + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\
& \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\
= & \frac{1}{2} \left[\sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 - |t + f(k) - s_l| \right) \\
& + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left(t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right) \right. \\
& \quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right) \right) \Big]
\end{aligned}$$

(because $s_i \leq s_l$, for every $i = 1, \dots, l$, and $s_l > t_1$)

$$\begin{aligned}
& = \frac{1}{2} \left(\sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 - |t + f(k) - s_l| \right) \\
& \quad + \frac{l-1}{2} (f(k-1) + f(2) - f(k)) \\
& \geq \frac{1}{2} \left(\sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 - |t + f(k) - s_l| \right)
\end{aligned}$$

$$= \begin{cases} \frac{1}{2}(2(s_l - t_1) + f(k - 1) + f(2) - f(k)) > 0 & (\text{if } s_l \leq t + f(k)) \\ \frac{1}{2}\left(2 \sum_{i=2}^k t_i + f(k - 1) + f(2) + f(k)\right) > 0 & (\text{if } s_l \geq t + f(k)) \end{cases}$$

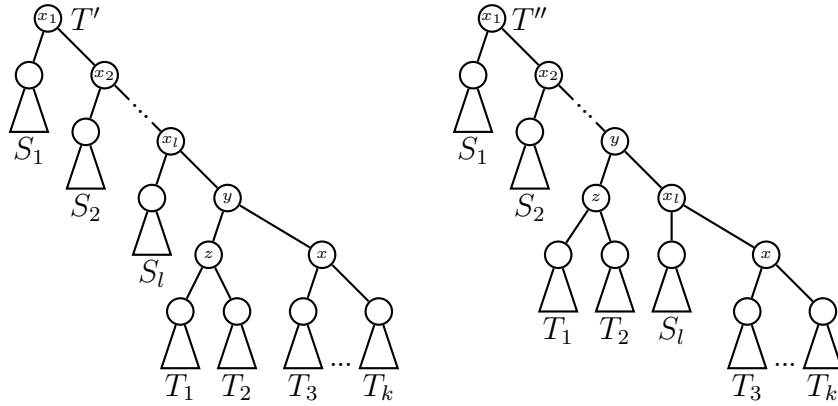


Figure 4.7: The trees in Lemma 4.26.(b).

(b) Assume now that $k \geq 3$ is even, and hence $k \geq 4$, and that t_1, t_2 are the middle values of $\{t_1, t_2, \dots, t_k\}$. By Lemma 4.23, this implies that $\text{MDM}(t_1, \dots, t_k) \leq \text{MDM}(t_3, \dots, t_k)$, which is the property that we are actually going to use in the proof.

As far as assertion (b.1) goes, let us assume that $s_l \leq t_1 + t_2 + f(2) \leq t + f(2)$. The only nodes in T or T' with different *bal* value in both trees are x_1, \dots, x_l, x and the new nodes y and z in T' . Therefore:

$$\begin{aligned} \mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(x) - \text{bal}_T(x) + \text{bal}_{T'}(z) + \text{bal}_{T'}(y) + \text{bal}_{T'}(x_l) - \text{bal}_T(x_l) \\ &\quad + \sum_{i=1}^{l-1} (\text{bal}_{T'}(x_i) - \text{bal}_T(x_i)) \\ &= \text{MDM}(t_3, \dots, t_k) - \text{MDM}(t_1, \dots, t_k) + \frac{1}{2}|t_2 - t_1| \\ &\quad + \frac{1}{2} \left| \sum_{i=3}^k t_i + f(k - 2) - (t_2 + t_1 + f(2)) \right| \\ &\quad + \frac{1}{2} |t + f(k - 2) + 2f(2) - s_l| - \frac{1}{2} |t + f(k) - s_l| \\ &\quad + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left| t + \sum_{j=i+1}^l s_j + f(k - 2) + (l - i + 2)f(2) - s_i \right| \right) \end{aligned}$$

$$\begin{aligned}
& - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \\
& \geq \frac{1}{2} (t + f(k-2) + 2f(2) - s_l - |t + f(k) - s_l|) \\
& \quad + \frac{1}{2} \sum_{i=1}^{l-1} \left(t + \sum_{j=i+1}^l s_j + f(k-2) + (l-i+2)f(2) - s_i \right. \\
& \quad \quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right) \right)
\end{aligned}$$

(because $s_l \leq t + f(2)$ and $s_i \leq s_l$, for every $i = 1, \dots, l-1$)

$$\begin{aligned}
& = \frac{1}{2} (t + f(k-2) + 2f(2) - s_l - |t + f(k) - s_l|) \\
& \quad + \frac{l-1}{2} (f(k-2) + 2f(2) - f(k)) \\
& \geq \frac{1}{2} (t + f(k-2) + 2f(2) - s_l - |t + f(k) - s_l|)
\end{aligned}$$

(because $k \geq 4$ implies that $f(k) < f(k-2) + 2f(2)$)

$$= \begin{cases} \frac{1}{2} (f(k-2) + 2f(2) - f(k)) > 0 & \text{if } s_l \leq t + f(k) \\ \frac{1}{2} (2(t + f(2) - s_l) + f(k-2) + f(k)) > 0 & \text{if } s_l \geq t + f(k) \end{cases}$$

As far as assertion (b.2) goes, let us assume now that $s_l > t_1 + t_2 + f(2)$. Again, the only nodes in T or T'' with different *bal* value in both trees are x_1, \dots, x_l, x and the new nodes y, z . Therefore:

$$\begin{aligned}
\mathfrak{C}(T'') - \mathfrak{C}(T) & = \text{bal}_{T''}(x) - \text{bal}_T(x) + \text{bal}_{T''}(z) + \text{bal}_{T''}(x_l) - \text{bal}_T(x_l) \\
& \quad + \text{bal}_{T''}(y) + \sum_{i=1}^{l-1} (\text{bal}_{T''}(x_i) - \text{bal}_T(x_i)) \\
& = \text{MDM}(t_3, \dots, t_k) - \text{MDM}(t_1, \dots, t_k) + \frac{1}{2} |t_2 - t_1| \\
& \quad + \frac{1}{2} \left| \sum_{i=3}^k t_i + f(k-2) - s_l \right| - \frac{1}{2} |t + f(k) - s_l| \\
& \quad + \frac{1}{2} \left| \sum_{i=3}^k t_i + s_l + f(k-2) + f(2) - (t_1 + t_2 + f(2)) \right|
\end{aligned}$$

$$\begin{aligned}
 & + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left| t + \sum_{j=i+1}^l s_j + f(k-2) + (l-i+2)f(2) - s_i \right| \right. \\
 & \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\
 \geq & \frac{1}{2} \left(\sum_{i=3}^k t_i + s_l + f(k-2) + f(2) - (t_1 + t_2 + f(2)) - |t + f(k) - s_l| \right) \\
 & + \frac{1}{2} \sum_{i=1}^{l-1} \left(t + \sum_{j=i+1}^l s_j + f(k-2) + (l-i+2)f(2) - s_i \right. \\
 & \quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right) \right) \\
 = & \frac{1}{2} \left(\sum_{i=3}^k t_i + s_l + f(k-2) - (t_1 + t_2) - |t + f(k) - s_l| \right) \\
 & + \frac{l-1}{2} (f(k-2) + 2f(2) - f(k)) \\
 & \text{(because } k \geq 4 \text{ implies that } f(k) < f(k-2) + 2f(2)) \\
 \geq & \frac{1}{2} \left(\sum_{i=3}^k t_i + s_l + f(k-2) - (t_1 + t_2) - |t + f(k) - s_l| \right) \\
 = & \begin{cases} \frac{1}{2} (2s_l - 2(t_1 + t_2) + f(k-2) - f(k)) \\ > \frac{1}{2} (2f(2) + f(k-2) - f(k)) > 0 & \text{if } s_l \leq t + f(k) \\ \frac{1}{2} \left(2 \sum_{i=3}^k t_i + f(k-2) + f(k) \right) > 0 & \text{if } s_l \geq t + f(k) \end{cases}
 \end{aligned}$$

□

Corollary 4.27. *For every non-bifurcating tree $T \in \mathcal{T}_n$, there always exists a bifurcating tree $T' \in \mathcal{BT}_n$ such that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*

Proof. We shall prove by complete induction on the sum S of the out-degrees of the non-bifurcating internal nodes in a tree $T \in \mathcal{T}_n$ that there always exists a bifurcating tree $T' \in \mathcal{BT}_n$ such that $\mathfrak{C}(T') \geq \mathfrak{C}(T)$. Moreover, it will be clear from the proof that if T isn't bifurcating, then T' can be chosen so that this inequality is strict.

The assertion to be proved by induction is obviously true if $S = 0$ (which means that T is bifurcating), so assume that $S > 0$. Let x be an internal

non-bifurcating node of T such that all nodes in the path from the root r to x , except x itself, are bifurcating.

If x is the root, then we apply Lemma 4.24 and we obtain a tree T_0 with smaller S and larger \mathfrak{C} . Then, by induction, there exists a bifurcating tree T' such that $\mathfrak{C}(T) < \mathfrak{C}(T_0) \leq \mathfrak{C}(T')$.

If x is not the root r , let r, x_2, \dots, x_l, x be the path from the root to x : all these nodes except x are bifurcating. By Lemma 4.25, if we rearrange the subtrees rooted at the children of the nodes r, x_2, \dots, x_l in increasing order of their σ -sizes, the \mathfrak{C} value of the resulting tree increases without modifying the value of S ; let \widehat{T} be the tree obtained in this way. Next, by Lemma 4.26, in \widehat{T} we can either prune a subtree T_1 rooted at one child of x and insert it in an arc in the path from r to x (adding a new bifurcating node to the tree), or we can prune two subtrees T_1, T_2 rooted at two children of x and insert $T_1 \star T_2$ in an arc in the path from r to x (adding two new bifurcating nodes to the tree), in both cases in such a way that the resulting tree T_0 has a larger \mathfrak{C} and a smaller S . Then, by induction, there exists a bifurcating tree T' such that $\mathfrak{C}(T) \leq \mathfrak{C}(\widehat{T}) < \mathfrak{C}(T_0) \leq \mathfrak{C}(T')$.

This finishes the proof by induction. \square

Therefore, the maximum \mathfrak{C} value on \mathcal{T}_n is reached at a bifurcating tree, where, by Proposition 4.9, it is equal to $(f(0) + f(2))/2$ times the Colless index, with $f(0) + f(2) \geq f(2) > 0$. Then, since, by Lemma 4.1, the maximum Colless index of a bifurcating tree with n leaves is reached exactly at the comb K_n , the same is true for \mathfrak{C} . So, the maximum value of \mathfrak{C} on \mathcal{T}_n is reached exactly at K_n , and it is

$$\mathfrak{C}(K_n) = \frac{f(0) + f(2)}{2} C(K_n) = \frac{f(0) + f(2)}{4} (n-1)(n-2).$$

Proof of the thesis of Theorem 4.21 for $\mathfrak{C}_{sd,f}$

The proof of Theorem 4.21 for $D = sd$, the sample standard deviation, is very similar to the one provided for $D = \text{MDM}$ in the previous subsection, but simpler, because Lemmas 4.22 and 4.23 are replaced by Lemma 4.28 below, which guarantees that it is always enough to remove a suitable element in a non-constant numeric vector of length at least 3, in order to increase its variance.

In this subsection, \mathfrak{C} stands for $\mathfrak{C}_{sd,f}$ and we shall denote $bal_{sd,f}$ on a tree T by bal_T or simply by bal when it is not necessary to specify the tree. We shall often use, without any further mention, that $sd(x, y) = |x - y|/\sqrt{2}$, which as established in the proof of Proposition 4.9.

Lemma 4.28. *For every $(x_1, \dots, x_n) \in \mathbb{R}^n$ with $n \geq 3$, if x_i is the value in the set $\{x_1, \dots, x_n\}$ closest to its mean, then*

$$\text{var}(x_1, \dots, x_n) \leq \text{var}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n),$$

and thus, taking positive square roots,

$$sd(x_1, \dots, x_n) \leq sd(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n).$$

Moreover, these inequalities are strict unless either $x_1 = \dots = x_n$ or n is even and $\{x_1, \dots, x_n\}$ consists of $n/2$ copies of two different elements.

Proof. Let $\bar{x} = (x_1 + \dots + x_n)/n$ and, after rearranging x_1, \dots, x_n if necessary, assume that $(x_n - \bar{x})^2 \leq (x_i - \bar{x})^2$, for every $i = 1, \dots, n-1$. We shall prove that

$$\text{var}(x_1, \dots, x_n) \leq \text{var}(x_1, \dots, x_{n-1}).$$

Indeed, let $\bar{x}' = (x_1 + \dots + x_{n-1})/(n-1)$. Notice that

$$\bar{x}' = \frac{x_1 + \dots + x_{n-1} + x_n - x_n}{n-1} = \frac{n \cdot \bar{x} - x_n}{n-1} = \bar{x} + \frac{\bar{x} - x_n}{n-1}.$$

Then

$$\text{var}(x_1, \dots, x_{n-1}) \geq \text{var}(x_1, \dots, x_n) \iff (n-1) \sum_{i=1}^{n-1} (x_i - \bar{x}')^2 \geq (n-2) \sum_{i=1}^n (x_i - \bar{x})^2.$$

Now:

$$\begin{aligned} (n-1) \sum_{i=1}^{n-1} (x_i - \bar{x}')^2 &= (n-1) \sum_{i=1}^{n-1} \left(x_i - \bar{x} + \frac{1}{n-1} (x_n - \bar{x}) \right)^2 \\ &= (n-1) \sum_{i=1}^{n-1} \left((x_i - \bar{x})^2 + \frac{2}{n-1} (x_i - \bar{x})(x_n - \bar{x}) + \frac{1}{(n-1)^2} (x_n - \bar{x})^2 \right) \\ &= (n-1) \sum_{i=1}^{n-1} (x_i - \bar{x})^2 + 2(x_n - \bar{x}) \sum_{i=1}^{n-1} (x_i - \bar{x}) + (x_n - \bar{x})^2 \\ &= (n-1) \sum_{i=1}^{n-1} (x_i - \bar{x})^2 + 2(x_n - \bar{x})(\bar{x} - x_n) + (x_n - \bar{x})^2 \\ &= (n-1) \sum_{i=1}^{n-1} (x_i - \bar{x})^2 - (x_n - \bar{x})^2 \\ &= (n-2) \sum_{i=1}^{n-1} (x_i - \bar{x})^2 + \sum_{i=1}^{n-1} (x_i - \bar{x})^2 - (x_n - \bar{x})^2 \\ &\geq (n-2) \sum_{i=1}^{n-1} (x_i - \bar{x})^2 + (n-1)(x_n - \bar{x})^2 - (x_n - \bar{x})^2 \\ &= (n-2) \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

that is,

$$\frac{\sum_{i=1}^{n-1} (x_i - \bar{x}')^2}{n-2} \geq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1},$$

as we wanted to prove. This inequality is an equality if, and only if, $(x_i - \bar{x})^2 = (x_n - \bar{x})^2$ for every $i = 1, \dots, n-1$, and it is easy to check that this condition holds exactly when either $x_1 = \dots = x_n$ or n is even and $\{x_1, \dots, x_n\}$ consists of $n/2$ copies of two different elements. Indeed, notice that

$$(x_i - \bar{x})^2 = (x_n - \bar{x})^2 \iff \begin{cases} x_i - \bar{x} = x_n - \bar{x} \iff x_i = x_n \\ \text{or} \\ x_i - \bar{x} = \bar{x} - x_n \iff x_i = 2\bar{x} - x_n \end{cases}$$

Assume now that there are $n_1 > 0$ numbers x_i that are equal to x_n and $n - n_1 > 0$ numbers x_i that are equal to $2\bar{x} - x_n$. Then

$$\begin{aligned} n \cdot \bar{x} &= n_1 \cdot x_n + (n - n_1)(2\bar{x} - x_n) = (2n_1 - n)x_n + 2(n - n_1)\bar{x} \\ &\iff (2n_1 - n)\bar{x} = (2n_1 - n)x_n \\ &\iff \begin{cases} \bar{x} = x_n, \text{ i.e. } x_n = 2\bar{x} - x_n \\ \text{or} \\ n = 2n_1 \end{cases} \end{aligned}$$

which proves that either all numbers x_i are equal to x_n or half of them are equal to x_n and the other half to $2\bar{x} - x_n$. \square

We prove now a series of lemmas that play in this proof the same role as Lemmas 4.24 to 4.26 in the last subsection.

Lemma 4.29. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $f(k) > 0$, for every $k \geq 2$, and let T be a tree of the form $T_1 \star \dots \star T_k$, with $k \geq 3$. If $\sigma(T_1)$ is the value in the set $\{\sigma(T_1), \dots, \sigma(T_k)\}$ closest to its mean, then taking the tree $T' = T_1 \star (T_2 \star \dots \star T_k)$ depicted in the left-hand side of Fig. 4.3, we obtain that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*

Proof. Let $t_i = \sigma(T_i)$, for every $i = 1, \dots, k$. The only nodes in T or T' with different *bal* value in both trees are the roots and the new node v in T' . Therefore,

$$\begin{aligned} \mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(r) + \text{bal}_{T'}(v) - \text{bal}_T(r) \\ &= \frac{1}{\sqrt{2}} \left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| + \text{sd}(t_2, \dots, t_k) - \text{sd}(t_1, \dots, t_k) \\ &\geq \frac{1}{\sqrt{2}} \left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| \geq 0, \end{aligned}$$

where the first inequality is a consequence of Lemma 4.28. Moreover, by the aforementioned lemma, this first inequality is strict unless either $t_1 = t_2 = \dots = t_k$ or (up to reordering the trees T_2, \dots, T_k) $k = 2m \geq 4$, $t_1 = \dots = t_m$ and $t_{m+1} = \dots = t_k$, and in both cases

$$\left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| = \sum_{i=3}^k t_i + f(k-1) \geq f(k-1) > 0$$

Therefore, we always have that $\mathfrak{C}(T') - \mathfrak{C}(T) > 0$. □

The proof of the following lemma is the same as that of Lemma 4.25 (up to replacing the fractions $1/2$ by $1/\sqrt{2}$).

Lemma 4.30. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $f(2) > 0$. Consider the trees T and T' depicted in Fig. 4.4, where T' is obtained from T by simply interchanging the subtrees T_l and T_{l-1} . If $\sigma(T_l) < \sigma(T_{l-1})$ and $\sigma(T_l) \leq \sigma(T_0)$, then $\mathfrak{C}(T') > \mathfrak{C}(T)$. □*

Lemma 4.31. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $0 < f(k) < f(k-1) + f(2)$, for every $k \geq 3$, and let T be the tree depicted in Fig. 4.5, where $l \geq 1$, x_1 is the root, all nodes in the path from x_1 to x_l are bifurcating, and $k \geq 3$. Assume moreover that $\sigma(S_1) \leq \sigma(S_2) \leq \dots \leq \sigma(S_l)$ and that $\sigma(T_1)$ is the value in the set $\{\sigma(T_1), \dots, \sigma(T_k)\}$ closest to its mean.*

- (a) *If $\sigma(S_l) \leq \sigma(T_1)$, then the tree T' depicted in Fig. 4.6, obtained by pruning the subtree T_1 and inserting it in the arc ending in x , is such that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*
- (b) *If $\sigma(S_l) > \sigma(T_1)$, then the tree T'' depicted in Fig. 4.6, obtained by pruning the subtree T_1 and inserting it in the arc ending in x_l , is such that $\mathfrak{C}(T'') > \mathfrak{C}(T)$.*

Proof. For every $i = 1, \dots, l$, let $s_i = \sigma(S_i)$ and, for every $i = 1, \dots, k$, $t_i = \sigma(T_i)$. Let, moreover, $t = t_1 + \dots + t_k$. So, we are assuming that $s_1 \leq \dots \leq s_l$ and that $sd(t_1, \dots, t_k) \leq sd(t_2, \dots, t_k)$. Moreover, for every $i = 1, \dots, l$, we shall call x_i the parent of the root of S_i in all three trees T, T', T'' .

As far as assertion (a) goes, the only nodes in T or T' with different *bal* value in both trees are x_1, \dots, x_l, x and the new node y . Therefore:

$$\begin{aligned} \mathfrak{C}(T') - \mathfrak{C}(T) &= bal_{T'}(x) - bal_T(x) + bal_{T'}(y) + \sum_{i=1}^l (bal_{T'}(x_i) - bal_T(x_i)) \\ &= sd(t_2, \dots, t_k) - sd(t_1, \dots, t_k) + \frac{1}{\sqrt{2}} \left| \sum_{i=2}^k t_i + f(k-1) - t_1 \right| \\ &\quad + \sum_{i=1}^l \frac{1}{\sqrt{2}} \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\ &\quad \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\ &\geq \frac{1}{\sqrt{2}} \sum_{i=1}^l \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \end{aligned}$$

$$\begin{aligned}
& - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \\
& = \frac{l}{\sqrt{2}}(f(k-1) + f(2) - f(k)) > 0
\end{aligned}$$

where, as in the proof of Lemma 4.26.(a.1), the last equality is a consequence of the fact that, for every $i = 1, \dots, l$, $s_i \leq s_l \leq t_1 \leq t$.

Let us prove now assertion (b). Again, the only nodes in T or T'' with different *bal* value in both trees are x_1, \dots, x_l, x and the new node y . Therefore,

$$\begin{aligned}
\mathfrak{C}(T') - \mathfrak{C}(T) & = \text{bal}_{T''}(x) - \text{bal}_T(x) + \text{bal}_{T''}(x_l) + \text{bal}_{T''}(y) - \text{bal}_T(x_l) \\
& + \sum_{i=1}^{l-1} (\text{bal}_{T''}(x_i) - \text{bal}_T(x_i)) \\
& = \text{sd}(t_2, \dots, t_k) - \text{sd}(t_1, \dots, t_k) + \frac{1}{\sqrt{2}} \left| \sum_{i=2}^k t_i + f(k-1) - s_l \right| \\
& + \frac{1}{\sqrt{2}} \left| \sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 \right| - \frac{1}{\sqrt{2}} |t + f(k) - s_l| \\
& + \frac{1}{\sqrt{2}} \sum_{i=1}^{l-1} \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\
& \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) \\
& \geq \frac{1}{\sqrt{2}} \left(\left| \sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 \right| - |t + f(k) - s_l| \right) \\
& + \frac{1}{\sqrt{2}} \sum_{i=1}^{l-1} \left(\left| t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right| \right. \\
& \quad \left. - \left| t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right| \right) > 0
\end{aligned}$$

where the last strict inequality is derived as in the proof of Lemma 4.26.(a.2). \square

Then, using Lemmas 4.29 to 4.31 and arguing as in the proof of Corollary 4.27, we deduce that, for every non-bifurcating tree $T \in \mathcal{T}_n$, there always exists a bifurcating tree $T' \in \mathcal{T}_n$ such that $\mathfrak{C}(T') > \mathfrak{C}(T)$. Starting from this fact, the same argument that completes the proof of Theorem 4.21 for $\mathfrak{C}_{\text{MDM},f}$ also completes it for $\mathfrak{C}_{\text{sd},f}$.

Proof of the thesis of Theorem 4.21 for $\mathfrak{C}_{\text{var},f}$

The proof is similar to those described in the previous two subsections, using Lemma 4.28 and proving a series of lemmas that show how to increase the Colless-like index of a non-bifurcating tree by making it “more bifurcating.” To simplify the notations, in this subsection we shall denote $\mathfrak{C}_{\text{var},f}$ by \mathfrak{C} and we shall denote $\text{bal}_{\text{var},f}$ on a tree T by bal_T or simply by bal when it is not necessary to specify the tree. We shall often use, without any further mention, that $\text{var}(x, y) = (x - y)^2/2$, which as established in the proof of Proposition 4.9.

Lemma 4.32. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $f(k) > 0$, for every $k \geq 2$, and let T be a tree of the form $T_1 \star \dots \star T_k$, with $k \geq 3$. If $\sigma(T_1)$ is the value in the set $\{\sigma(T_1), \dots, \sigma(T_k)\}$ closest to its mean, then taking the tree $T' = T_1 \star (T_2 \star \dots \star T_k)$ depicted in the left-hand side of Fig. 4.3, we obtain that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*

Proof. Let $t_i = \sigma(T_i)$, for every $i = 1, \dots, k$. The only nodes in T or T' with different bal value in both trees are the roots and the new node v in T' . Therefore,

$$\begin{aligned} \mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(v) + \text{bal}_{T'}(r) - \text{bal}_T(r) \\ &= \text{var}(t_2, \dots, t_k) + \frac{1}{2} \left(\sum_{i=2}^k t_i + f(k-1) - t_1 \right)^2 - \text{var}(t_1, \dots, t_k) \\ &\geq \frac{1}{2} \left(\sum_{i=2}^k t_i + f(k-1) - t_1 \right)^2 \geq 0. \end{aligned}$$

Now, the first inequality is strict unless either $t_1 = t_2 = \dots = t_k$ or (up to reordering the trees T_2, \dots, T_k) $k = 2m \geq 4$, $t_1 = \dots = t_m$ and $t_{m+1} = \dots = t_k$ (see Lemma 4.28), and in both cases the last inequality is strict (cf. the proof of Lemma 4.29). \square

Lemma 4.33. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $f(2) > 0$. Consider the trees T and T' depicted in Fig. 4.4, where, in both trees, the nodes in the path connecting the root r with x are bifurcating, and T' is obtained from T by simply interchanging the subtrees T_l and T_{l-1} . If $\sigma(T_l) < \sigma(T_{l-1})$, then $\mathfrak{C}(T') > \mathfrak{C}(T)$.*

Proof. Let $t_i = \sigma(T_i)$, for every $i = 0, \dots, l$, so that $t_l < t_{l-1}$. The only nodes in T or T' with different bal value in both trees are x and y , and therefore,

$$\begin{aligned} \mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(x) + \text{bal}_{T'}(y) - \text{bal}_T(x) - \text{bal}_T(y) \\ &= \frac{1}{2}(t_{l-1} - t_0)^2 + \frac{1}{2}(t_{l-1} + t_0 + f(2) - t_l)^2 - \frac{1}{2}(t_l - t_0)^2 \\ &\quad - \frac{1}{2}(t_l + t_0 + f(2) - t_{l-1})^2 \\ &= \frac{1}{2}(t_{l-1} - t_l)(t_{l-1} + t_l + 2t_0 + 4f(2)) > 0 \end{aligned}$$

because $t_{l-1} > t_l$ and $f(2) > 0$. \square

Lemma 4.34. *Let f be a mapping $\mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ such that $0 < f(k) < f(k-1) + f(2)$, for every $k \geq 3$, and let T be the tree depicted in Fig. 4.5, where $l \geq 1$, x_1 is the root, all nodes in the path from x_1 to x_l are bifurcating, and $k \geq 3$. Assume moreover that $\sigma(S_1) \leq \sigma(S_2) \leq \dots \leq \sigma(S_l)$ and that $\sigma(T_1)$ is the value in the set $\{\sigma(T_1), \dots, \sigma(T_k)\}$ closest to its mean. Then:*

- (a) *If $\sigma(S_l) \leq \sigma(T_1)$, then the tree T' depicted in Fig. 4.6, obtained by pruning the subtree T_1 and inserting it in the arc ending in x , is such that $\mathfrak{C}(T') > \mathfrak{C}(T)$.*
- (b) *If $\sigma(S_l) > \sigma(T_1)$, then the tree T'' depicted in Fig. 4.6, obtained by pruning the subtree T_1 and inserting it in the arc ending in x_l , is such that $\mathfrak{C}(T'') > \mathfrak{C}(T)$.*

Proof. For every $i = 1, \dots, l$, let $s_i = \sigma(S_i)$ and, for every $i = 1, \dots, k$, $\sigma(T_i) = t_i$, and let $t = t_1 + \dots + t_k$. We are assuming that $s_1 \leq \dots \leq s_l$ and that $\text{var}(t_1, \dots, t_k) \leq \text{var}(t_2, \dots, t_k)$. Moreover, for every $i = 1, \dots, l$, we shall call x_i the parent of the root of S_i in all three trees T, T', T'' .

As far as assertion (a) goes, the only nodes in T or T' with different *bal* value in both trees are x_1, \dots, x_l, x and the new node y . Therefore,

$$\begin{aligned}
\mathfrak{C}(T') - \mathfrak{C}(T) &= \text{bal}_{T'}(x) - \text{bal}_T(x) + \text{bal}_{T'}(y) + \sum_{i=1}^l (\text{bal}_{T'}(x_i) - \text{bal}_T(x_i)) \\
&= \text{var}(t_2, \dots, t_k) - \text{var}(t_1, \dots, t_k) + \frac{1}{2} \left(\sum_{i=2}^k t_i + f(k-1) - t_1 \right)^2 \\
&\quad + \sum_{i=1}^l \left(\frac{1}{2} \left(t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right)^2 \right. \\
&\quad \left. - \frac{1}{2} \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right)^2 \right) \\
&\geq \frac{1}{2} \sum_{i=1}^l \left(\left(t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right)^2 \right. \\
&\quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right)^2 \right) \\
&= \frac{1}{2} (f(k-1) + f(2) - f(k)) \sum_{i=1}^l \left(2 \left(t + \sum_{j=i+1}^l s_j - s_i \right) + f(k-1) + f(k) \right. \\
&\quad \left. + (2(l-i) + 1)f(2) \right) > 0,
\end{aligned}$$

where this last expression is > 0 because $f(k-1) + f(2) - f(k) > 0$ and, for every $i = 1, \dots, l$, $s_i \leq s_l \leq t_1 \leq t$.

Let us prove now assertion (b). Again, the only nodes in T or T'' with different bal value in both trees are x_1, \dots, x_l, x and the new node y . Therefore,

$$\begin{aligned}
\mathfrak{C}(T') - \mathfrak{C}(T) &= bal_{T''}(x) - bal_T(x) + bal_{T''}(y) + bal_{T''}(x_l) - bal_T(x_l) \\
&\quad + \sum_{i=1}^{l-1} (bal_{T''}(x_i) - bal_T(x_i)) \\
&= \text{var}(t_2, \dots, t_k) - \text{var}(t_1, \dots, t_k) + \frac{1}{2} \left(\sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 \right)^2 \\
&\quad + \frac{1}{2} \left(\sum_{i=2}^k t_i + f(k-1) - s_l \right)^2 - \frac{1}{2} (t + f(k) - s_l)^2 \\
&\quad + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left(t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right)^2 \right. \\
&\quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right)^2 \right) \\
&\geq \frac{1}{2} \left(\left(\sum_{i=2}^k t_i + s_l + f(k-1) + f(2) - t_1 \right)^2 - (t + f(k) - s_l)^2 \right) \\
&\quad + \frac{1}{2} \sum_{i=1}^{l-1} \left(\left(t + \sum_{j=i+1}^l s_j + f(k-1) + (l-i+1)f(2) - s_i \right)^2 \right. \\
&\quad \left. - \left(t + \sum_{j=i+1}^l s_j + f(k) + (l-i)f(2) - s_i \right)^2 \right) \\
&= \frac{1}{2} \left(2 \sum_{i=2}^k t_i + f(k-1) + f(2) + f(k) \right) (f(k-1) + f(2) - f(k) + 2(s_l - t_1)) \\
&\quad + \frac{1}{2} (f(k-1) + f(2) - f(k)) \sum_{i=1}^{l-1} \left(2 \left(t + \sum_{j=i+1}^l s_j - s_i \right) + f(k-1) + f(k) \right. \\
&\quad \left. + (2(l-i) + 1)f(2) \right) > 0
\end{aligned}$$

where this last expression is > 0 because $f(k-1) + f(2) - f(k) > 0$, $s_l > t_1$ and, for every $i = 1, \dots, l-1$, $s_i \leq s_l$. \square

Then, using Lemmas 4.32 to 4.34 and arguing as in the proof of Corollary 4.27, it can be proved that, for every non-bifurcating tree $T \in \mathcal{T}_n$, there always

exists a bifurcating tree $T' \in \mathcal{BT}_n$ such that $\mathfrak{C}(T') > \mathfrak{C}(T)$. Therefore, the maximum \mathfrak{C} value is reached at some bifurcating tree. Since, for bifurcating trees T , $\mathfrak{C}(T) = \frac{(f(0)+f(2))^2}{2} \cdot C^{(2)}(T)$ (see Proposition 4.10), and $f(0) + f(2) > 0$, it remains to prove that the bifurcating tree in \mathcal{BT}_n with maximum $C^{(2)}$ is exactly the comb. The proof of this fact follows closely that of Lemma 4.1.

Corollary 4.35. *For every bifurcating tree $T \in \mathcal{T}_n$, if $T \neq K_n$, then $C^{(2)}(K_n) > C^{(2)}(T)$.*

Proof. Using the argument of the proof of Lemma 4.1, it is enough to prove that if T and T' are the trees depicted in Fig. 4.1, then, under the assumptions therein, $C^{(2)}(T') > C^{(2)}(T)$. And, indeed (using the notations therein),

$$\begin{aligned} C^{(2)}(T') - C^{(2)}(T) &= (t_3 + t_4 - t_2)^2 + (t_3 + t_4 + t_2 - t_1)^2 - (t_2 - t_1)^2 \\ &\quad - (t_3 + t_4 - t_2 - t_1)^2 \\ &= (t_3 + t_4 - t_1)(t_3 + t_4 + t_1 + 2t_2) > 0 \end{aligned}$$

where the last inequality holds because, by assumption, $t_1, t_2, t_3, t_4 > 0$ and $t_1 + t_2 \leq t_3 + t_4$. \square

Now, it is straightforward to check that

$$C^{(2)}(K_n) = \sum_{k=1}^{n-2} k^2 = \frac{1}{6}(n-1)(n-2)(2n-3),$$

from where we obtain

$$\mathfrak{C}(K_n) = \frac{(f(0) + f(2))^2}{2} \cdot C^{(2)}(K_n) = \frac{(f(0) + f(2))^2}{12}(n-1)(n-2)(2n-3),$$

as we claimed in the statement.

4.4.2 The case of $f(n) = e^n$

As far as the case when $f(n) = e^n$ is concerned, we have the following result.

Theorem 4.36. *For every $n \geq 2$:*

- (a) *If $n \neq 4$, then both $\mathfrak{C}_{\text{MDM}, e^n}$ and $\mathfrak{C}_{\text{sd}, e^n}$ reach their maximum on \mathcal{T}_n exactly at the trees of shape $FS_1 \star FS_{n-1}$ (see Fig. 4.8), and these maximum values are*

$$\begin{aligned} \mathfrak{C}_{\text{MDM}, e^n}(FS_1 \star FS_{n-1}) &= \frac{1}{2}(e^{n-1} + n - 2), \\ \mathfrak{C}_{\text{sd}, e^n}(FS_1 \star FS_{n-1}) &= \frac{1}{\sqrt{2}}(e^{n-1} + n - 2). \end{aligned}$$

(b) Both $\mathfrak{C}_{\text{MDM},e^n}$ and \mathfrak{C}_{sd,e^n} reach their maximum on \mathcal{T}_4 exactly at the combs K_4 , and these maximum values are

$$\mathfrak{C}_{\text{MDM},e^n}(K_4) = \frac{3}{2}(e^2 + 1),$$

$$\mathfrak{C}_{sd,e^n}(K_4) = \frac{3}{\sqrt{2}}(e^2 + 1).$$

(c) $\mathfrak{C}_{\text{var},e^n}$ always reaches its maximum on \mathcal{T}_n exactly at the trees of shape $FS_1 \star FS_{n-1}$, and the maximum value is

$$\mathfrak{C}_{\text{var},e^n}(FS_1 \star FS_{n-1}) = \frac{1}{2}(e^{n-1} + n - 2)^2.$$

□

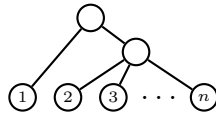


Figure 4.8: The tree $FS_1 \star FS_{n-1}$.

So, according to $\mathfrak{C}_{\text{MDM},e^n}$, $\mathfrak{C}_{\text{var},e^n}$, and \mathfrak{C}_{sd,e^n} , the trees of the form $FS_1 \star FS_{n-1}$ are the most unbalanced (except for $n = 4$ and $D = \text{MDM}$ or sd , in which case the most unbalanced tree is the comb). Table 4.4 in Section 4.7 gives the values of these indices on \mathcal{T}_n^* , for $n = 2, 3, 4, 5$, and the positions of the different trees in each \mathcal{T}_n^* according to the increasing order of the corresponding index.

We have also split the proof of Theorem 4.36 into 3 subsections, one for each dissimilarity. To simplify the notations, throughout this subsection we shall denote σ_{e^n} by σ . We shall often use, usually without any further notice, that $\sigma(FS_1) = e^0 = 1$.

Proof of the thesis of Theorem 4.36 for $\mathfrak{C}_{\text{MDM},e^n}$

In this subsection, \mathfrak{C} will stand for $\mathfrak{C}_{\text{MDM},e^n}$.

Lemma 4.37. *Let $n_1, \dots, n_k \in \mathbb{N}_{>0}$ and $n = n_1 + \dots + n_k$, and assume that $2 \leq k \leq n - 1$.*

- (a) $e^{n_1} + \dots + e^{n_k} \leq e^{n-k+1} + (k - 1)e$ and the equality is strict if, and only if, there is more than one exponent $n_i \geq 2$.
- (b) $e^{n-k+1} + e^k + (k - 1)e < e^n$

Proof. To begin with, notice that if $1 \leq x \leq \min(a, b)$, then

$$e^a + e^b \leq e^{a+b-x} + e^x, \quad (4.4)$$

because

$$e^b - e^x = (e^{b-x} - 1)e^x \leq (e^{b-x} - 1)e^a = e^{a+b-x} - e^a.$$

Moreover, if $x < \min(a, b)$ then the inequality is clearly strict.

Now, applying $k - 1$ times inequality (4.4) with $x = 1$, we obtain

$$e^{n_1} + \dots + e^{n_k} \leq e^{n_1+(n_2-1)+(n_3-1)+\dots+(n_k-1)} + (k-1)e = e^{n-k+1} + (k-1)e,$$

which implies (a). Moreover, this inequality is an equality if, and only if, $n_i = 1$ for every i except at most one. Indeed, remember that (4.4) is strict if, and only if, $x < a$ and $x < b$. Thus, taking $x = 1$, the inequality which we iterate $k - 1$ times

$$e^{n_1+\dots+n_j+(j-1)} + e^{n_{j+1}} \leq e^{n_1+\dots+n_j+n_{j+1}+j} + e$$

is an equality if, and only if, $n_1 + \dots + n_j + (j - 1) = 1$ or $n_{j+1} = 1$, i.e. (since $n_i \geq 1$ for every i), if, and only if $j = 1$ and $n_1 = 1$ or $n_{j+1} = 1$. So, in the first application of the inequality, i.e. when $j = 1$, we have an equality if, and only if, $n_1 = 1$ or $n_2 = 1$, and in the next applications, i.e., when $j \geq 2$, we have an equality if and only if $n_{j+1} = 1$. So, after the $k - 1$ applications, we have an equality if, and only if, $n_3 = \dots = n_k = 1$ and $n_1 = 1$ or $n_2 = 1$.

As far as inequality (b) goes, since $2 \leq k \leq n - 1$, and hence, in particular, $n \geq 3$ (and using, in the second inequality, that $x + 1 < e^x$ for every $x \in \mathbb{R}_{>0}$), we have

$$\begin{aligned} e^{n-k+1} + (k-1)e + e^k &\leq 2e^{n-1} + (n-2)e < 2e^{n-1} + e^{n-3} \cdot e = e^{n-2}(2e+1) \\ &< e^{n-2} \cdot e^2 = e^n \end{aligned}$$

as we claimed. \square

Lemma 4.38. *Let $n_1, \dots, n_k, l \in \mathbb{N}$ be such that $k \geq 1$, $k + l \geq 2$, each $n_i \geq 2$, and let $n = n_1 + \dots + n_k$. Then*

$$e^{n_1} + \dots + e^{n_k} + e^{k+l} < e^{n+l}$$

Proof. The case $l = 0$ is a particular instance of the combination of both inequalities in the last lemma. So, we assume henceforth that $l \geq 1$. If $k = 1$, so that $n = n_1$, then applying inequality (4.4) with $x = 2$ we have

$$e^n + e^{l+1} \leq e^{n+l-1} + e^2 \leq e^{n+l-1} + e^{n+l-1} = 2e^{n+l-1} < e^{n+l}$$

where the second inequality holds because $n + l \geq 3$.

Assume finally that $k \geq 2$. Applying $k - 1$ times inequality (4.4) with $x = 2$, we obtain

$$e^{n_1} + \dots + e^{n_k} + e^{k+l} \leq e^{n-2(k-1)} + (k-1)e^2 + e^{k+l}$$

Now, since $2 \leq k \leq n/2$, and hence $n \geq 4$,

$$\begin{aligned} e^{n-2(k-1)} + (k-1)e^2 + e^{k+l} &\leq e^{n-2} + (n/2 - 1)e^2 + e^{l+n/2} \\ &< e^{n-2} + e^{n/2-2} \cdot e^2 + e^{l+n/2} \\ \text{(because } x-1 < e^{x-2} \text{ if } x > 0) \\ &= e^{n-2} + e^{n/2} + e^{l+n/2} \leq e^{n-2} + e^{n+l-2} + e^2 \\ \text{(applying inequality (4.4) to } e^{n/2} + e^{l+n/2} \text{ and } k = 2) \\ &< 3e^{n+l-2} < e^{n+l} \end{aligned}$$

where the second last inequality is a consequence of $e^{n-2} < e^{n+l-2}$ (because $l \geq 1$) and $e^2 \leq e^{n+l-2}$ (because $l \geq 1$ and $n \geq 4$ and therefore $2 < 3 \leq n+l-2$). \square

Lemma 4.39. *The largest e^n -size of a tree in \mathcal{T}_n is $e^n + n$, and it is reached exactly at the rooted star FS_n .*

Proof. The cases $n = 1, 2$ are obvious, because then $\mathcal{T}_n = \{FS_n\}$. Consider now the case $n \geq 3$. We shall prove that for every $T \in \mathcal{T}_n$, if $T \neq FS_n$, there is a tree $T' \in \mathcal{T}_n$ with $\sigma(T') > \sigma(T)$. This shows that no tree other than FS_n can have the maximum e^n -size.

So, let $T = T_1 \star \dots \star T_m \in \mathcal{T}_n \setminus \{FS_n\}$, with $m \geq 2$. Let $l \geq 0$ be such that, for every $i = 1, \dots, l$, the subtree T_i consists of a single node, and, for every $i = l+1, \dots, m$, $T_i = T_{i,1} \star \dots \star T_{i,n_i}$ with $n_i \geq 2$; cf. Fig. 4.9. Since $T \neq FS_n$, $l < m$. Let now T' be the tree

$$T' = T_1 \star \dots \star T_l \star T_{l+1,1} \star \dots \star T_{l+1,n_{l+1}} \star \dots \star T_{m,1} \star \dots \star T_{m,n_m}.$$

Then,

$$\begin{aligned} \sigma(T) &= e^m + \sum_{i=1}^l \sigma(T_i) + \sum_{i=l+1}^m \left(e^{n_i} + \sum_{j=1}^{n_i} \sigma(T_{i,j}) \right) \\ &= \sum_{i=l+1}^m e^{n_i} + e^{(m-l)+l} + \sum_{i=1}^l \sigma(T_i) + \sum_{i=l+1}^m \sum_{j=1}^{n_i} \sigma(T_{i,j}) \\ &< e^{l+n_{l+1}+\dots+n_m} + \sum_{i=1}^l \sigma(T_i) + \sum_{i=l+1}^m \sum_{j=1}^{n_i} \sigma(T_{i,j}) = \sigma(T') \end{aligned}$$

where the inequality is due to Lemma 4.38. \square

Lemma 4.40. *For every $T \in \mathcal{T}_n$ with $n \neq 1, 3$,*

$$2\mathfrak{C}(T) + \sigma(T) \leq 2\mathfrak{C}(FS_n) + \sigma(FS_n) = e^n + n,$$

and the inequality is strict if $T \neq FS_n$.

When $n = 1$, $2\mathfrak{C}(FS_1) + \sigma(FS_1) = 1$ (instead of $e^1 + 1$), and when $n = 3$, the maximum value of $2\mathfrak{C} + \sigma$ is $2\mathfrak{C}(K_3) + \sigma(K_3) = 3e^2 + 4$ and it is reached exactly at K_3 .

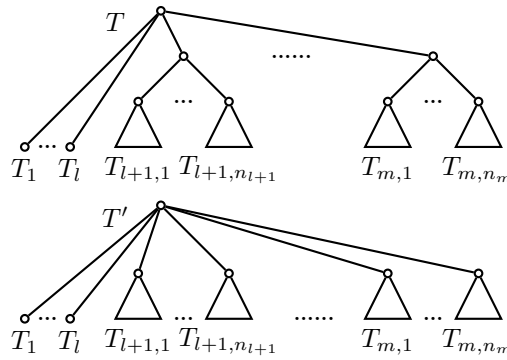


Figure 4.9: The trees T and T' in the proof of Lemma 4.39.

Proof. The cases $n = 1, 2$ are obvious, because then $\mathcal{T}_n = \{FS_n\}$, and the cases $n = 3, 4, 5$ can be checked in Table 4.4 in Section 4.7. Notice that $1 < e^1 + 1$ and $3e^2 + 4 < (e^3 + 3) + 4$; we shall use these two inequalities below. We prove now the general case $n \geq 6$ using the cases $n = 1, \dots, 5$ and complete induction on n .

Let $T = T_1 \star \dots \star T_k$, with $k \geq 2$ and $T_i \in \mathcal{T}_{n_i}$ for every $i = 1, \dots, k$, so that $n = n_1 + \dots + n_k \geq 6$. If $k = n$, then $n_i = 1$ for every i and $T = FS_n$, in which case $2\mathfrak{C}(T) + \sigma(T) = e^n + n$. So, we shall assume that $k \leq n - 1$, and we shall prove that, in this case $2\mathfrak{C}(T) + \sigma(T) < e^n + n$.

After renumbering the subtrees T_i if necessary, assume that there exists $l \geq 0$ such that $n_i = 3$ for every $i \leq l$ and $n_i \neq 3$ for every $i = l + 1, \dots, k$. We shall prove first of all that

$$\text{MDM}(\sigma(T_1), \dots, \sigma(T_k)) \leq \frac{1}{k} \sum_{i=1}^k (e^{n_i} + n_i - 2) \tag{4.5}$$

Indeed, let $M = \text{Median}(\sigma(T_1), \dots, \sigma(T_k))$. Then,

$$\begin{aligned} \text{MDM}(\sigma(T_1), \dots, \sigma(T_k)) &= \frac{1}{k} \sum_{i=1}^k |\sigma(T_i) - M| \leq \frac{1}{k} \sum_{i=1}^k |\sigma(T_i) - 2| \\ &\leq \frac{1}{k} \sum_{i=1}^k (e^{n_i} + n_i - 2) \end{aligned}$$

where the first inequality is due to the fact that M is the real number that minimizes the function $x \mapsto \sum_{i=1}^k |\sigma(T_i) - x|$, and the second inequality holds because $|\sigma(T_i) - 2| \leq e^{n_i} + n_i - 2$ for every $i = 1, \dots, k$; on its turn, this inequality is a consequence, when $n_i \geq 2$, of the fact that $\sigma(T_i) \geq n_i$ (because each leaf of the tree contributes 1 to its size) and Lemma 4.39, and, when $n_i = 1$, of the fact that if $T_1 \in \mathcal{T}_1$, then $\sigma(T_i) = 1$ and hence $|\sigma(T_i) - 2| = 1 < e^1 + 1 - 2$.

Now,

$$\begin{aligned}
2\mathfrak{C}(T) + \sigma(T) &= 2\left(\sum_{i=1}^k \mathfrak{C}(T_i) + \text{MDM}(\sigma(T_1), \dots, \sigma(T_k))\right) + \sum_{i=1}^k \sigma(T_i) + e^k \\
&= \sum_{i=1}^k (2\mathfrak{C}(T_i) + \sigma(T_i)) + 2\text{MDM}(\sigma(T_1), \dots, \sigma(T_k)) + e^k \\
&= \sum_{i=1}^l (2\mathfrak{C}(T_i) + \sigma(T_i)) + \sum_{i=l+1}^k (2\mathfrak{C}(T_i) + \sigma(T_i)) \\
&\quad + 2\text{MDM}(\sigma(T_1), \dots, \sigma(T_k)) + e^k \\
&\leq \sum_{i=1}^l (3e^2 + 4) + \sum_{i=l+1}^k (e^{n_i} + n_i) + \frac{2}{k} \sum_{i=1}^k (e^{n_i} + n_i - 2) + e^k \\
&\text{(by the case } n = 3, \text{ the induction hypothesis, and inequality (4.5))} \\
&\leq \sum_{i=1}^l (e^3 + 3 + 4) + \sum_{i=l+1}^k (e^{n_i} + n_i) + \sum_{i=1}^k (e^{n_i} + n_i - 2) + e^k \\
&\text{(because } k \geq 2 \text{ and } 3e^2 + 4 < e^3 + 3 + 4) \\
&= \sum_{i=1}^k (e^{n_i} + n_i) + 4l + \sum_{i=1}^k (e^{n_i} + n_i - 2) + e^k \\
&= 2 \sum_{i=1}^k e^{n_i} + 2n + 4l - 2k + e^k \leq 2 \sum_{i=1}^k e^{n_i} + e^k + 2n + 2k \\
&\leq 2e^{n-(k-1)} + e^k + 2(e+1)k + 2n - 2e \\
&\text{(by the first inequality in Lemma 4.37)}
\end{aligned}$$

Thus, it remains to prove that, for every $n \geq 6$ and for every $2 \leq k \leq n - 1$,

$$2e^{n-(k-1)} + e^k + 2(e+1)k + 2n - 2e < e^n + n. \quad (4.6)$$

Since, for every $n \geq 6$, the function

$$\begin{aligned}
f_n(x) &= e^n + n - (2e^{n-(x-1)} + e^x + 2(e+1)x + 2n - 2e) \\
&= e^n - 2e^{n-(x-1)} - e^x - 2(e+1)x - n + 2e
\end{aligned}$$

is concave, because

$$f_n''(x) = -2e^{n+1-x} - e^x < 0,$$

its minimum value on the closed interval $[2, n - 1]$ is reached at one of its ends. So, in order to prove inequality (4.6) for every $n \geq 6$ and for every

$k = 2, \dots, n-1$, it is enough to prove that $f_n(2) > 0$ and $f_n(n-1) > 0$ for every $n \geq 6$. And, indeed

- $f_n(2) = e^n - 2e^{n-1} - n - e^2 - 2e - 4 > 0$ because the function $g(x) = e^x - 2e^{x-1} - x - e^2 - 2e - 4$ is increasing on $\mathbb{R}_{\geq 2}$ and $g(5) > 0$.
- $f_n(n-1) = e^n - e^{n-1} - (2e+3)n - 2e^2 + 4e + 2 > 0$ by a similar reason.

This finishes the proof of the statement. \square

Now we can proceed with the proof of Theorem 4.36.(a) for $D = \text{MDM}$. The cases $n = 2, 3, 4, 5$ can be checked in Table 4.4 in Section 4.7. Notice in particular that, when $n = 4$, the maximum is

$$\mathfrak{C}(K_4) = \frac{3}{2}(e^2 + 1) < \frac{1}{2}(e^3 + 2) + 2;$$

we shall use this inequality below. We prove now, using the cases $n = 1, \dots, 5$ and complete induction on n , that, for every $n \geq 6$,

$$\textit{The tree in } \mathcal{T}_n^* \textit{ with maximum } \mathfrak{C} \textit{ is } FS_1 \star FS_{n-1}, \textit{ with } \mathfrak{C}(FS_1 \star FS_{n-1}) = \frac{1}{2}(e^{n-1} + n - 2)$$

Recall that, as in the previous subsection, Lemma 4.24 implies that the maximum \mathfrak{C} value on \mathcal{T}_n^* is reached at a tree with bifurcating root. So, let $T = T_1 \star T_2 \in \mathcal{T}_n$, with $T_1 \in \mathcal{T}_{n_1}$ and $T_2 \in \mathcal{T}_{n_2}$. We must distinguish two cases:

a) Assume that $n_1 = 1$, and therefore $n_2 = n - 1 \geq 5$. In this case,

$$\begin{aligned} \mathfrak{C}(T) &= \mathfrak{C}(T_2) + \frac{1}{2}(\sigma(T_2) - 1) = \frac{1}{2}(2\mathfrak{C}(T_2) + \sigma(T_2) - 1) \\ &\leq \frac{1}{2}(e^{n_2} + n_2 - 1) = \frac{1}{2}(e^{n-1} + n - 2) \end{aligned}$$

by Lemma 4.40. Moreover, the equality holds only when $T_2 = FS_{n-1}$.

b) Assume that $n_1, n_2 \geq 2$ and, without any loss of generality, that $\sigma(T_2) \leq \sigma(T_1)$. Then,

$$\begin{aligned} \mathfrak{C}(T) &= \mathfrak{C}(T_1) + \mathfrak{C}(T_2) + \frac{1}{2}(\sigma(T_1) - \sigma(T_2)) \\ &< \frac{1}{2}(e^{n_1-1} + n_1 - 2) + 2 + \frac{1}{2}(e^{n_2-1} + n_2 - 2) + 2 + \frac{1}{2}(e^{n_1} + n_1 - n_2) = (*) \end{aligned}$$

This inequality is due to the following facts. On the one hand, $n_2 \leq \sigma(T_2)$ and $\sigma(T_1) \leq e^{n_1} + n_1$, by Lemma 4.39, and hence $\sigma(T_1) - \sigma(T_2) \leq e^{n_1} + n_1 - n_2$. On the other hand, by the induction hypothesis, $\mathfrak{C}(T_i) \leq \frac{1}{2}(e^{n_i-1} + n_i - 2) < \frac{1}{2}(e^{n_i-1} + n_i - 2) + 2$, unless $n_i = 4$, in which case we still have $\mathfrak{C}(T_i) \leq \mathfrak{C}(K_4) < \frac{1}{2}(e^{n_i-1} + n_i - 2) + 2$.

Let us continue:

$$(*) = \frac{1}{2}((1 + e)e^{n_1-1} + e^{n_2-1} + 2n_1 + 4) \leq \frac{1}{2}((2 + e)e^{n-3} + 2n)$$

because $n_1, n_2 \leq n - 2$. So, it remains to prove that, for every $n \geq 6$,

$$(2 + e)e^{n-3} + 2n < e^{n-1} + n - 2$$

This is equivalent to

$$(e^2 - e - 2)e^{n-3} - n - 2 > 0,$$

which is easy to prove, for instance noticing that $f(x) = (e^2 - e - 2)e^{x-3} - x - 2$ is increasing on $\mathbb{R}_{\geq 3}$ and that $f(5) > 0$. This finishes the proof of Theorem 4.36 for $D = \text{MDM}$.

Proof of the thesis of Theorem 4.36 for \mathfrak{C}_{sd,e^n}

The proof of this case follows closely that of the case when $D = \text{MDM}$ given in the last subsection. To begin with, it turns out that a key lemma similar to Lemma 4.40 also holds when $D = sd$. To simplify the notations, \mathfrak{C} stands in this subsection for \mathfrak{C}_{sd,e^n} .

Lemma 4.41. *For every $T \in \mathcal{T}_n$ with $n \neq 1, 3$,*

$$\sqrt{2} \cdot \mathfrak{C}(T) + \sigma(T) \leq \sqrt{2} \cdot \mathfrak{C}(FS_n) + \sigma(FS_n) = e^n + n.$$

and the inequality is strict if $T \neq FS_n$.

When $n = 1$, $\sqrt{2} \cdot \mathfrak{C}(FS_1) + \sigma(FS_1) = 1$, and when $n = 3$, the maximum value of $\sqrt{2} \cdot \mathfrak{C} + \sigma$ is $\sqrt{2} \cdot \mathfrak{C}(K_3) + \sigma(K_3) = 3e^2 + 4$.

Proof. The cases $n = 1, 2$ are, as always, obvious because then $\mathcal{T}_n = \{FS_n\}$, and the cases $n = 3, 4, 5$ can be checked in Table 4.4 in Section 4.7. We shall use that $1 < e^1 + 1$ and the following inequalities:

$$\sqrt{2} \cdot \mathfrak{C}(K_3) + \sigma(K_3) = 3e^2 + 4 < (e^3 + 3) + 4 \tag{4.7a}$$

$$\sigma(K_3) = 2e^2 + 3 < (e^3 + 3) - 5 \tag{4.7b}$$

We prove the general case $n \geq 6$ by induction on n using an argument very similar to the one given in the proof of Lemma 4.40. Let $T = T_1 \star \dots \star T_k$, with $k \geq 2$ and $T_i \in \mathcal{T}_{n_i}$ for every $i = 1, \dots, k$, so that $n = n_1 + \dots + n_k$. If $k = n$, then $n_i = 1$ for every i and $T = FS_n$, in which case $\sqrt{2} \cdot \mathfrak{C}(T) + \sigma(T) = e^n + n$. So, we shall assume that $k \leq n - 1$. Without any loss of generality, we assume that there exists $l \geq 0$ such that $T_i = K_3$ if, and only if, $i \leq l$.

Now, it turns out that

$$sd(\sigma(T_1), \dots, \sigma(T_k)) \leq \frac{1}{\sqrt{k-1}} \left(\sum_{i=1}^k (e^{n_i} + n_i - 2) - 5l \right) \tag{4.8}$$

Indeed, let $m = (\sigma(T_1) + \dots + \sigma(T_k))/k$. Then,

$$\text{var}(\sigma(T_1), \dots, \sigma(T_k)) = \frac{1}{k-1} \sum_{i=1}^k (\sigma(T_i) - m)^2 \leq \frac{1}{k-1} \sum_{i=1}^k (\sigma(T_i) - 2)^2$$

because m is the real number that minimizes the function $x \mapsto \sum_{i=1}^k (\sigma(T_i) - x)^2$.

Taking square roots,

$$\begin{aligned} sd(\sigma(T_1), \dots, \sigma(T_k)) &\leq \sqrt{\frac{1}{k-1} \sum_{i=1}^k (\sigma(T_i) - 2)^2} \leq \frac{1}{\sqrt{k-1}} \sum_{i=1}^k |\sigma(T_i) - 2| \\ &\leq \frac{1}{\sqrt{k-1}} \left(\sum_{i=1}^k (e^{n_i} + n_i - 2) - 5l \right) \end{aligned}$$

where the second last inequality is a consequence of Lemma 4.39, inequality (4.7b), and the fact, already used in the previous subsection, that $|\sigma(FS_1) - 2| = 1 < e + 1 - 2$. Then

$$\begin{aligned} \sqrt{2}\mathfrak{C}(T) + \sigma(T) &= \sqrt{2} \left(\sum_{i=1}^k \mathfrak{C}(T_i) + sd(\sigma(T_1), \dots, \sigma(T_k)) \right) + \sum_{i=1}^k \sigma(T_i) + e^k \\ &= \sum_{i=1}^k (\sqrt{2}\mathfrak{C}(T_i) + \sigma(T_i)) + \sqrt{2}sd(\sigma(T_1), \dots, \sigma(T_k)) + e^k \\ &\leq \sum_{i=1}^k (e^{n_i} + n_i) + 4l + \frac{\sqrt{2}}{\sqrt{k-1}} \left(\sum_{i=1}^k (e^{n_i} + n_i - 2) - 5l \right) + e^k \\ &\quad \text{(by the induction hypothesis and inequalities (4.7a) and (4.8))} \\ &\leq \sum_{i=1}^k e^{n_i} + n + 4l + \sqrt{2} \left(\sum_{i=1}^k e^{n_i} + n - 2k - 5l \right) + e^k \\ &\leq (1 + \sqrt{2}) \sum_{i=1}^k e^{n_i} + (1 + \sqrt{2})n - 2\sqrt{2}k + e^k \\ &\leq (1 + \sqrt{2})e^{n-(k-1)} + (1 + \sqrt{2})(k-1)e + e^k + (1 + \sqrt{2})n - 2\sqrt{2}k \\ &\quad \text{(because of the first inequality in Lemma 4.37)} \\ &= (1 + \sqrt{2})e^{n-(k-1)} + (1 + \sqrt{2})n + e^k + (e + \sqrt{2}e - 2\sqrt{2})k - (1 + \sqrt{2})e \end{aligned}$$

Thus, it remains to prove that, for every $n \geq 6$ and for every $2 \leq k \leq n-1$,

$$(1 + \sqrt{2})e^{n-(k-1)} + (1 + \sqrt{2})n + e^k + (e + \sqrt{2}e - 2\sqrt{2})k - (1 + \sqrt{2})e < e^n + n \quad (4.9)$$

Now, for every $n \geq 1$, the function

$$\begin{aligned} f_n(x) &= e^n + n - \left((1 + \sqrt{2})e^{n-(x-1)} + e^x + (1 + \sqrt{2})n + (e + \sqrt{2}e - 2\sqrt{2})x \right. \\ &\quad \left. - (1 + \sqrt{2})e \right) \\ &= e^n - (1 + \sqrt{2})e^{n-(x-1)} - e^x - \sqrt{2}n - (e + \sqrt{2}e - 2\sqrt{2})x + (1 + \sqrt{2})e \end{aligned}$$

is concave, and therefore the minimum value of $f_n(x)$ on the closed interval $[2, n - 1]$ will be reached at one of its ends. So, in order to prove inequality (4.9) for every $n \geq 6$ and every $k = 2, \dots, n - 1$, it is enough to prove that $f_n(2) > 0$ and $f_n(n - 1) > 0$ for every $n \geq 6$. And, indeed

- $f_n(2) = e^n - (1 + \sqrt{2})e^{n-1} - \sqrt{2}n - (e^2 + \sqrt{2}e + e - 4\sqrt{2}) > 0$ because the function $g(x) = e^x - (1 + \sqrt{2})e^{x-1} - \sqrt{2}x - (e^2 + \sqrt{2}e + e - 4\sqrt{2})$ is increasing on $\mathbb{R}_{\geq 3}$ and $g(5) > 0$.
- $f_n(n-1) = e^n - (1 + \sqrt{2})e^2 - e^{n-1} - \sqrt{2}n - (e + \sqrt{2}e - 2\sqrt{2})(n-1) + (1 + \sqrt{2})e$ by a similar reason.

This finishes the proof of the lemma. □

From here on, the proof of Theorem 4.36 for $D = sd$ proceeds as the one for $D = \text{MDM}$ given in the previous subsection, using Lemma 4.41 instead of Lemma 4.40; to ease the task of the reader we provide this proof. The cases $n = 2, 3, 4, 5$ can be checked in Table 4.4 in Section 4.7. Notice in particular that, when $n = 4$, the maximum \mathfrak{C} value is

$$\mathfrak{C}(K_4) = \frac{3}{\sqrt{2}}(e^2 + 1) < \frac{1}{\sqrt{2}}(e^3 + 2) + \frac{5}{2} \tag{4.10}$$

we shall use it below.

We prove now, using the cases $n = 1, \dots, 5$ and complete induction on n , that, for every $n \geq 6$,

The tree in \mathcal{T}_n^ with maximum \mathfrak{C} is $FS_1 \star FS_{n-1}$, with $\mathfrak{C}(FS_1 \star FS_{n-1}) = \frac{1}{\sqrt{2}}(e^{n-1} + n - 2)$*

To begin with, notice that Lemma 4.29 implies that the maximum \mathfrak{C} value on \mathcal{T}_n^* is reached at a tree with bifurcating root. So, let $T = T_1 \star T_2 \in \mathcal{T}_n$, with $T_1 \in \mathcal{T}_{n_1}$ and $T_2 \in \mathcal{T}_{n_2}$. We must distinguish two cases:

a) Assume that $n_1 = 1$, and therefore $n_2 = n - 1 \geq 5$. In this case,

$$\begin{aligned} \mathfrak{C}(T) &= \mathfrak{C}(T_2) + \frac{1}{\sqrt{2}}(\sigma(T_2) - 1) = \frac{1}{\sqrt{2}}(\sqrt{2}\mathfrak{C}(T_2) + \sigma(T_2) - 1) \\ &\leq \frac{1}{\sqrt{2}}(e^{n-1} + n - 1 - 1) = \frac{1}{\sqrt{2}}(e^{n-1} + n - 2) \end{aligned}$$

by Lemma 4.40. Moreover, the equality holds only when $T_2 = FS_{n-1}$.

b) Assume that $n_1, n_2 \geq 2$ and, without any loss of generality, that $\sigma(T_2) \leq \sigma(T_1)$. Then,

$$\begin{aligned} \mathfrak{C}(T) &= \mathfrak{C}(T_1) + \mathfrak{C}(T_2) + \frac{1}{\sqrt{2}}(\sigma(T_1) - \sigma(T_2)) \\ &< \frac{1}{\sqrt{2}}(e^{n_1-1} + n_1 - 2) + \frac{5}{2} + \frac{1}{\sqrt{2}}(e^{n_2-1} + n_2 - 2) + \frac{5}{2} \\ &\quad + \frac{1}{\sqrt{2}}(e^{n_1} + n_1 - n_2) = (*) \end{aligned}$$

This inequality is due to the following facts. On the one hand, $n_2 \leq \sigma(T_2)$ and $\sigma(T_1) \leq e^{n_1} + n_1$, by Lemma 4.39, and hence $\sigma(T_1) - \sigma(T_2) \leq e^{n_1} + n_1 - n_2$. On the other hand, by the induction hypothesis,

$$\mathfrak{C}(T_i) \leq \frac{1}{\sqrt{2}}(e^{n_i-1} + n_i - 2) < \frac{1}{\sqrt{2}}(e^{n_i-1} + n_i - 2) + \frac{5}{2}$$

except when $n_i = 4$, in which case we still have, by inequality (4.10),

$$\mathfrak{C}(T_i) \leq \mathfrak{C}(K_4) < \frac{1}{\sqrt{2}}(e^{n_i-1} + n_i - 2) + \frac{5}{2}.$$

Let us continue

$$\begin{aligned} (*) &= \frac{1}{\sqrt{2}}((1+e)e^{n_1-1} + e^{n_2-1} + 2n_1 - 4) + 5 \\ &\leq \frac{1}{\sqrt{2}}((2+e)e^{n-3} + 2n + 5\sqrt{2} - 8) \\ &\quad (\text{because } n_1, n_2 \leq n - 2) \\ &< \frac{1}{\sqrt{2}}((2+e)e^{n-3} + 2n) < \frac{1}{\sqrt{2}}(e^{n-1} + n - 2) \end{aligned}$$

because $(2+e)e^{n-3} + 2n < e^{n-1} + n - 2$, as it was proven in the last step of the proof of Theorem 4.36 for $D = \text{MDM}$ in the last subsection. This finishes the proof of Theorem 4.36 for $D = sd$.

Proof of the thesis of Theorem 4.36 for $\mathfrak{C}_{\text{var},e^n}$

The stated maximum value of $\mathfrak{C}_{\text{var},e^n}$ on \mathcal{T}_n , for $n = 2, \dots, 5$, can be checked in Table 4.4 in Section 4.7. As far as the case when $n \geq 6$, it is a direct consequence of the corresponding result for $D = sd$, established in the previous subsection.

Indeed, to begin with, notice that, since, for every node v in a tree T , $\text{bal}_{\text{var},f}(v) = \text{bal}_{sd,f}(v)^2$, we have that, for every tree T ,

$$\mathfrak{C}_{\text{var},f}(T) = \sum_{v \in V_{\text{int}}(T)} \text{bal}_{sd,f}(v)^2 \leq \left(\sum_{v \in V_{\text{int}}(T)} \text{bal}_{sd,f}(v) \right)^2 = \mathfrak{C}_{sd,f}(T)^2.$$

So, for every $T \in \mathcal{T}_n^*$ with $n \geq 6$,

$$\begin{aligned} \mathfrak{C}_{\text{var},e^n}(T) &\leq \mathfrak{C}_{sd,e^n}(T)^2 \leq \mathfrak{C}_{sd,e^n}(FS_1 \star FS_{n-1})^2 = \frac{1}{2}(e^{n-1} + n - 2)^2 \\ &= \mathfrak{C}_{\text{var},e^n}(FS_1 \star FS_{n-1}) \end{aligned}$$

where the second inequality is strict if $T \neq FS_1 \star FS_{n-1}$.

4.5 The R package *CollessLike*

We have written an R package called *CollessLike*, available on the CRAN [74] and on the GitHub [96], that computes the Colless-like indices and their normalized version, as well as several other balance indices, and simulates the distribution of these indices on \mathcal{T}_n^* under the α - γ -model (see Section 1.3). This package contains functions that:

- Compute the following balance indices for multifurcating trees: the Sackin index S , the total cophenetic index Φ , and the Colless-like index $\mathfrak{C}_{D,f}$ for several predefined dissimilarities D and functions f as well as for any user-defined ones.

Our functions also compute the normalized versions (obtained by subtracting their minimum value and dividing by the width of their range, so that they take values in $[0, 1]$) of S , Φ and the Colless-like indices $\mathfrak{C}_{D,f}$ for which we have computed the range in Theorems 4.21 and 4.36. Recall that, for every $n \geq 2$:

- The range of S on \mathcal{T}_n^* goes from $S(FS_n) = n$ to $S(K_n) = (n+2)(n-1)/2$ (see Section 1.2)
- The range of Φ on \mathcal{T}_n^* goes from $\Phi(FS_n) = 0$ to $\Phi(K_n) = \binom{n}{3}$ (see Section 2.2)

Therefore, for every $T \in \mathcal{T}_n^*$, the normalized Sackin and total cophenetic index are, respectively,

$$S_{norm}(T) = \frac{S(T) - n}{\frac{1}{2}(n+2)(n-1) - n}, \quad \Phi_{norm}(T) = \frac{\Phi(T)}{\binom{n}{3}},$$

while, for instance, the normalized version of $\mathfrak{C}_{MDM, \ln(n+e)}$ is

$$\mathfrak{C}_{MDM, \ln(n+e), norm}(T) = \frac{\mathfrak{C}_{MDM, \ln(n+e)}(T)}{\frac{1+\ln(e+2)}{4}(n-1)(n-2)}.$$

- Given two natural numbers n and N , produce a random sample of N values of a balance index S , Φ , or $\mathfrak{C}_{D,f}$ on trees in \mathcal{T}_n^* generated following an α - γ -model: the parameters N , n , α , γ (with $0 \leq \gamma \leq \alpha \leq 1$) can be set by the user.

Due to the computational cost of this function, we have stored the values of S , Φ , and $\mathfrak{C}_{MDM, \ln(n+e)}$ (denoted henceforth simply by \mathfrak{C}) on the random samples of $N = 5000$ trees in each \mathcal{T}_n^* (for every $n = 3, \dots, 50$ and for every $\alpha, \gamma \in \{0, 0.1, 0.2, \dots, 0.9, 1\}$ with $\gamma \leq \alpha$) generated in the study performed in the next section. In this way, if the user is interested in this range of numbers of leaves and this range of parameters, they can estimate

the distribution of the corresponding balance index efficiently and quickly. This database is available on the GitHub repository associated to the CollessLike package [96].

- Given a tree $T \in \mathcal{T}_n^*$, estimate the percentile $q_{T,n,\alpha,\gamma}$ of its balance index S , Φ , or $\mathfrak{C}_{D,f}$ with respect to the distribution of this index on \mathcal{T}_n^* under some α - γ -model. If n, α, γ are among those mentioned in the previous item, for the sake of efficiency this function uses the database of computed indices to simulate the distribution of the balance index on \mathcal{T}_n^* under this α - γ -model.

For instance, the unlabeled tree $T \in \mathcal{T}_8^*$ in Fig. 4.10 is the shape of a phylogenetic tree randomly generated under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$ (using `set.seed(1000)` for reproducibility). The values of its balance indices are given in the figure's caption.

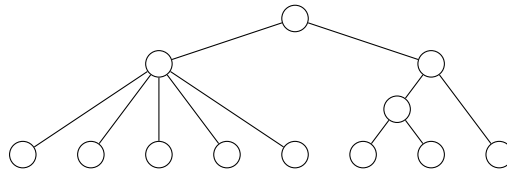


Figure 4.10: A tree with 8 leaves randomly generated under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$. Its indices are $\mathfrak{C}(T) = 1.746$, $S(T) = 18$, and $\Phi(T) = 14$, and its normalized indices are $\mathfrak{C}_{norm}(T) = 0.06518$, $S_{norm}(T) = 0.3704$, and $\Phi_{norm}(T) = 0.25$.

Fig. 4.11 displays the estimation of the density function of the balance indices \mathfrak{C} , S , and Φ under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$ on \mathcal{T}_8 , obtained using the 5000 random trees gathered in our database. The percentiles of the tree of Fig. 4.10 are given by the area to the left of the vertical lines.

Fig. 4.12 shows a percentile plot of \mathfrak{C} , S , and Φ under the α - γ -model for $\alpha = 0.7$ and $\gamma = 0.4$ on \mathcal{T}_8^* . Moreover, the estimated percentiles of the balance indices of the tree of Fig. 4.10 are also shown in the figure.

A special case of the α - γ -model, corresponding to the case $\alpha = \gamma$, is Ford's α -model for bifurcating phylogenetic trees (see Section 1.3). This model includes as special cases the Yule model (when $\alpha = \gamma = 0$) and the uniform model (when $\alpha = \gamma = 1/2$), also described in Section 1.3. So, this package allows also to study these models.

For example, the unlabeled tree in Fig. 4.13 has been generated (with `set.seed(1000)`) using $n = 8$ and $\alpha = \gamma = 0.5$, which corresponds to the uniform model. The Fig. 4.14 depicts the estimation of the density functions and of the percentile plots of \mathfrak{C} , S , and Φ on \mathcal{T}_8 under this model, as well as the percentile values of the tree.

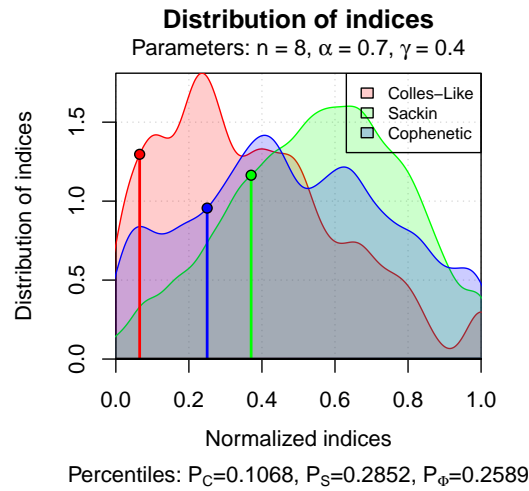


Figure 4.11: The estimated density function of the distribution of \mathfrak{C} , S and Φ on \mathcal{T}_8^* under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$. The percentiles of the tree in Fig. 4.10 are also represented.

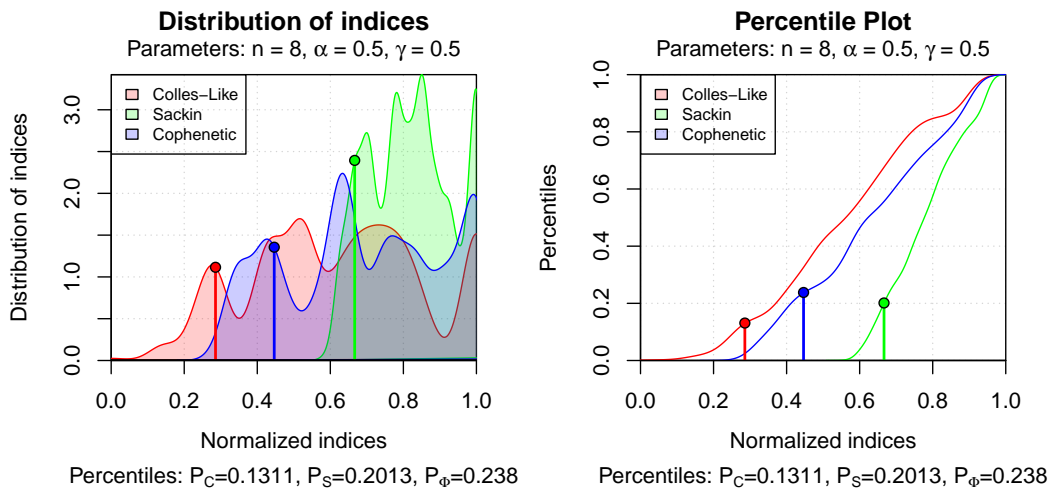


Figure 4.14: The estimated density function of the distribution of the three balance indices on \mathcal{T}_8 of the tree represented in Fig.4.13 under the uniform model, and their percentile plot.

4.5.1 A real example

To illustrate the behaviour of the Colless-like \mathfrak{C} , Sackin S and cophenetic Φ balance indices on a real tree, we have chosen a tree of the Primate phylogeny (see [108] for details). This tree is a part of a superstructure which represents the phylogeny for the Primates order and combines different smaller trees with partial overlap. In this tree, the topology of the genera *Trachypithecus*, *Presbytis*,

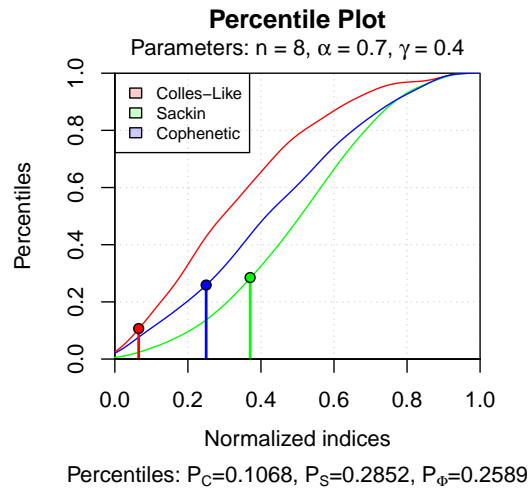


Figure 4.12: Percentile plot of the distribution of \mathfrak{C} , S and Φ on \mathcal{T}_8 under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$. The percentiles of the tree of Fig. 4.10 are also highlighted.

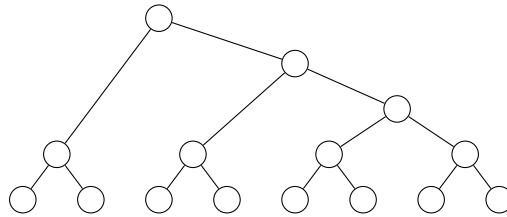


Figure 4.13: A bifurcating tree randomly generated under the uniform model.

Semnopithecus, *Pygathrix*, *Nasalis*, *Colobus* and *Procolobus* is displayed. The tree is depicted in Fig. 4.15.

The three balance indices of this tree and their normalized values are the following:

$$\begin{aligned} \mathfrak{C}(T) &= 81.48, & S(T) &= 161, & \Phi(T) &= 655, \\ \mathfrak{C}_{norm}(T) &= 0.16898, & S_{norm}(T) &= 0.32593, & \Phi_{norm}(T) &= 0.17926. \end{aligned}$$

To establish a relationship of the previous tree with the α - γ -model we have computed the percentile of the tree for every $(\alpha, \gamma) \in \{0, 0.1, 0.2, \dots, 0.9, 1\}^2$ with $\gamma \leq \alpha$, in order to check the values of the parameters (α, γ) for which the tree has higher values. The heatmap of the percentiles of the Colless balance index depicted in Fig. 4.16 shows the results. The parameters yielding the highest percentiles are given in Table 4.1 Under the models defined by these parameters, the Primate phylogeny must be understood as highly unbalanced, because there are many trees with a lower balance index.

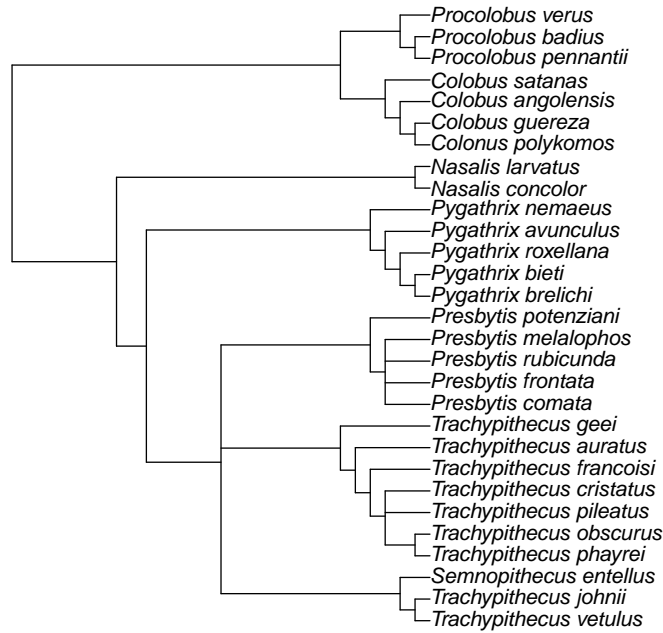


Figure 4.15: Subtree of the Primate phylogeny. In this case, it is represented the topology of the genera *Trachypithecus*, *Presbytis*, *Semnopithecus*, *Pygathrix*, *Nasalis*, *Colobus* and *Procolobus*.

The R script used to compute this study is also available in Appendix A.6.2 and on the GitHub repository [97].

α	γ	Percentile
0.9	0	0.9031
1	0.2	0.8725
1	0.1	0.8620
1	0.3	0.8607

Table 4.1: The four higher percentiles of the heatmap represented in Fig. 4.16.

4.6 Experimental results on TreeBASE

To assess the performance of $\mathfrak{C}_{\text{MDM}, \ln(n+e)}$, which we abbreviate again by \mathfrak{C} , we have considered the downloaded TreeBASE database already used in §2.8.2, containing 12928 phylogenetic trees.

Then, for every phylogenetic tree T in this set, we have computed its Colless-like index $\mathfrak{C}(T)$, its Sackin index $S(T)$, and its total cophenetic index $\Phi(T)$. We have studied the behaviour of the mean and variance of the balance indices as a function of the number of leaves of the trees, the number of ties of the balance

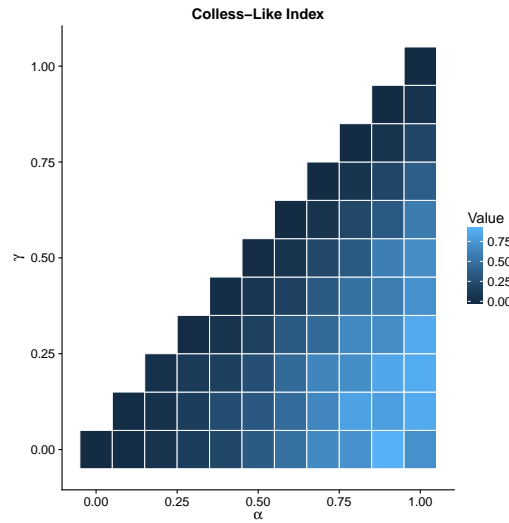


Figure 4.16: Heatmap of the percentiles of the Colless-like balance index of tree of Fig. 4.15 for every $(\alpha, \gamma) \in \{0, 0.1, 0.2, \dots, 0.9, 1\}^2$ with $\gamma \leq \alpha$ of the α - γ -model.

indices as a measure of quality, and how similar are them by computing the Spearman rank correlation. Moreover, we have tested whether the distribution of the Colless-like index of the phylogenetic trees in TreeBASE fits with the theoretical distributions defined by uniform model or the α - γ -model for some parameters α, γ . All analysis have been performed with R [107] and the scripts are available in Appendix A.6 and on the GitHub repository [97].

4.6.1 Mean and variance as a function of the number of leaves of the trees

For every number of leaves n , we have computed the mean and the variance of \mathfrak{C} , S and Φ on all trees with n leaves in TreeBASE. Then, we have computed the regression of these values as a function of n .

For the means, the best fits have been:

- *Colless-like index*: $\bar{\mathfrak{C}} \approx 0.5354 \cdot n^{1.5846}$, with a coefficient of determination of $R^2 = 0.9869$ and a p-value for the exponent $p \approx 0$.
- *Sackin index*: $\bar{S} \approx 1.4519 \cdot n^{1.4358}$, with $R^2 = 0.9953$ and $p \approx 0$.
- *Total cophenetic index*: $\bar{\Phi} \approx 0.1895 \cdot n^{2.5477}$, with $R^2 = 0.9945$ and $p \approx 0$.

Fig. 4.17 depicts these mean values of \mathfrak{C} , S , and Φ as functions of n .

Thus, S and \mathfrak{C} have similar mean growth rates, while Φ has a mean growth rate one order higher in magnitude. This difference vanishes if we normalize

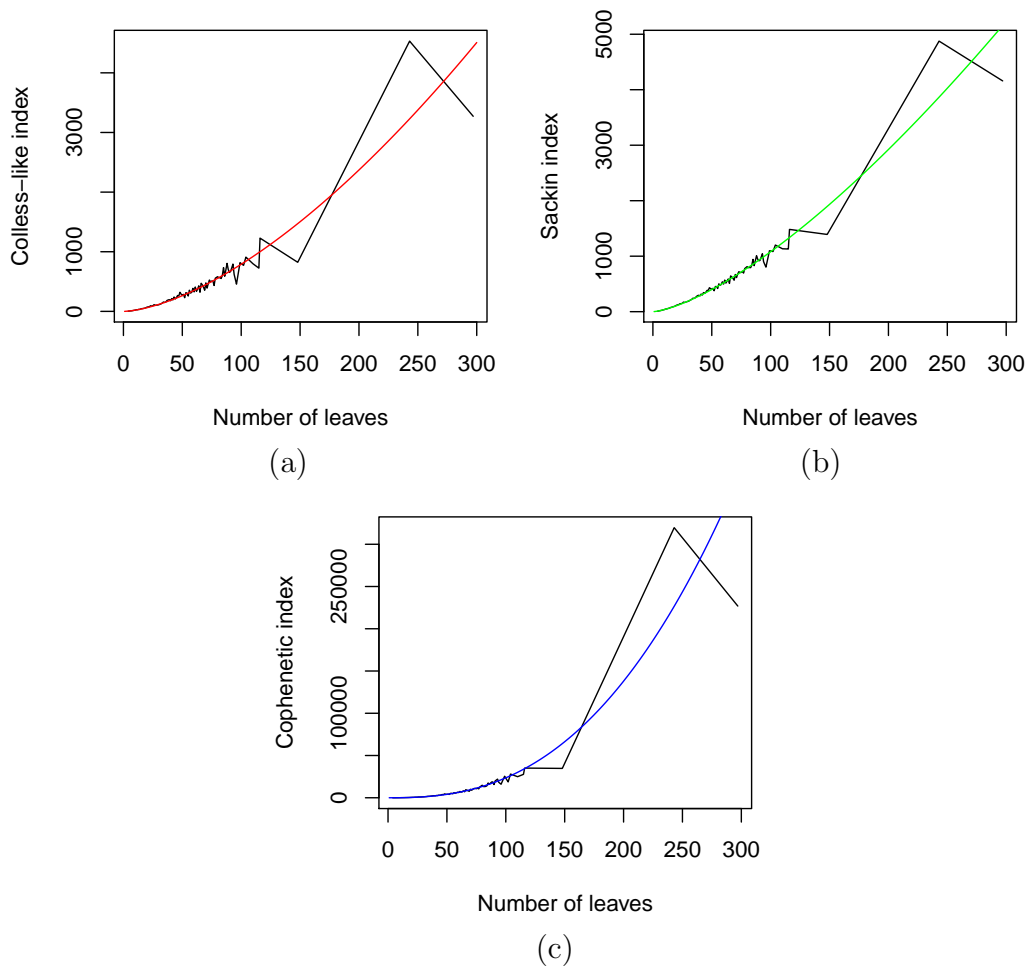


Figure 4.17: Growth of the mean value of \mathfrak{C} (a), S (b), and Φ (c) on TreeBASE, as functions of the trees' numbers of leaves n .

the indices by their range width, which is $O(n^2)$ for \mathfrak{C} and S , and $O(n^3)$ for Φ :

$$\overline{\mathfrak{C}_{norm}} \approx 1.6722 \cdot n^{-0.5686}$$

$$\overline{S_{norm}} \approx 2.314 \cdot n^{-0.5346}$$

$$\overline{\Phi_{norm}} \approx 2.2649 \cdot n^{-0.6055}$$

As for the behaviour of the variances, the best fits are the following:

- *Colless index*: $\text{Var}(\mathfrak{C}) \approx 0.07609 \cdot n^{3.1280}$, with $R^2 = 0.9620$ and $p \approx 0$.
- *Sackin index*: $\text{Var}(S) \approx 0.03208 \cdot n^{3.2225}$, with $R^2 = 0.9575$ and $p \approx 0$.
- *Total cophenetic index*: $\text{Var}(\Phi) \approx 0.00407 \cdot n^{5.2071}$, with $R^2 = 0.9812$ and $p \approx 0$.

The results are in the same line as before, with the variances of \mathfrak{C} and S having similar growth rates, and the variance of Φ having a growth rate two orders of magnitude higher. This difference vanishes again when we normalize the indices:

$$\begin{aligned}\text{Var}(\mathfrak{C}_{norm}) &\approx 0.74233 \cdot n^{-1.1785} \\ \text{Var}(S_{norm}) &\approx 0.2288 \cdot n^{-0.9082} \\ \text{Var}(\Phi_{norm}) &\approx 0.5814 \cdot n^{-1.0993}\end{aligned}$$

So, in summary, \mathfrak{C} has, on TreeBASE and relative to the range of values, a slightly larger mean growth rate and a slightly smaller variance growth rate than the other two indices.

4.6.2 Numbers of ties

The number of ties (that is, of pairs of different trees with the same index value) of a balance index is an interesting measure of quality, because the smaller its frequency of ties, the bigger its ability to rank the balance of any pair of different trees. Although, in our opinion, this ability need not always be an advantage: for instance, neither Φ nor S take the same, minimum, value on all different fully symmetric trees with the same numbers of leaves (for example, $S(FS_6) = 6$ but $S(FS_{2,3}) = S(FS_{3,2}) = 12$; and $\Phi(FS_6) = 0$, but $\Phi(FS_{3,2}) = 3$ and $\Phi(FS_{2,3}) = 6$; cf. Fig. 1.4), while \mathfrak{C} applied to any fully symmetric tree is always 0. In this case, we believe that these ties are fair.

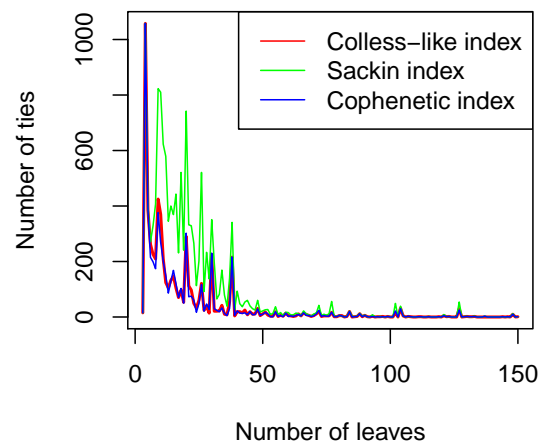


Figure 4.18: Numbers of ties of \mathfrak{C} , S , and Φ in TreeBASE, as functions of the trees' numbers of leaves n .

Anyway, for every number of leaves n and for every one of all three indices under scrutiny, we have computed the numbers of pairs of trees with n leaves

in TreeBASE having the same value of the corresponding index (in the case of \mathfrak{C} , up to 16 decimal digits). Fig. 4.18 plots the frequencies of ties of \mathfrak{C} , S and Φ as functions of n . As it can be seen in this figure, \mathfrak{C} and Φ have a similar number of ties, and consistently less ties than S .

4.6.3 Spearman's rank correlation

In order to measure whether all three indices sort the trees according to their balance in the same way or not, we have computed the Spearman's rank correlation coefficient [103] of the indices on all trees in TreeBASE, as well as grouping them by their number of leaves n .

The global Spearman's rank correlation coefficient of \mathfrak{C} and S is 0.9765, and that of \mathfrak{C} and Φ is 0.9619. The graphics in Fig. 4.19 plot these coefficients as functions of n . As it can be seen, Spearman's rank correlation coefficient for \mathfrak{C} and S grows with n , approaching to 1, while the coefficient for \mathfrak{C} and Φ shows a decreasing tendency with n .

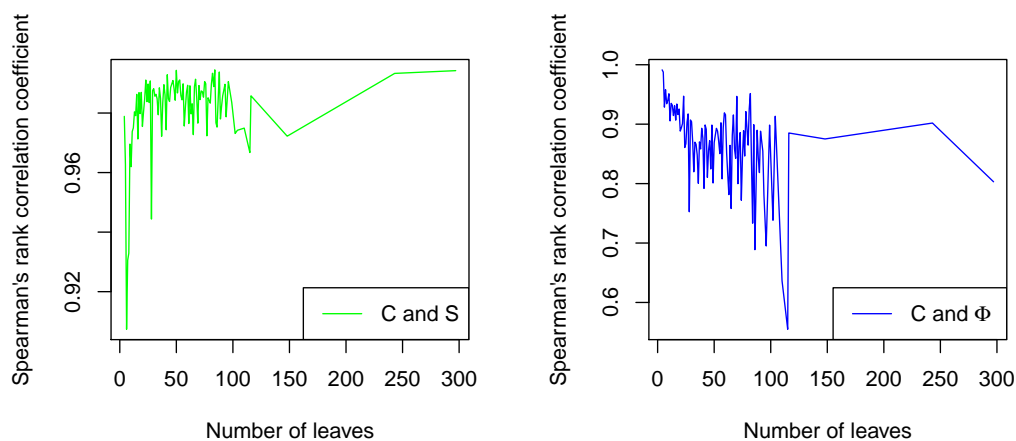


Figure 4.19: Spearman's rank correlation coefficient of \mathfrak{C} and S (left) and of \mathfrak{C} and Φ (right) in TreeBASE, as functions of the trees' numbers of leaves n .

4.6.4 Does TreeBASE fit the uniform model or the α - γ -model?

In this section, we test whether the distribution of the Colless-like index of the phylogenetic trees in TreeBASE agrees with its theoretical distribution under either the uniform model or some α - γ -model for multifurcating phylogenetic trees (see Section 1.3). To do it, we use the normalized version \mathfrak{C}_{norm} of \mathfrak{C} , which can be used simultaneously on trees with different numbers of leaves.

To estimate the theoretical distribution of this index under the two aforementioned theoretical models, for every $n = 3, \dots, 50$ we have generated, on the one hand, 10,000 random phylogenetic trees in \mathcal{T}_n^* under the uniform model using the algorithm described in [79], and, on the other hand, 5000 random phylogenetic trees in \mathcal{T}_n^* under the α - γ -model for every pair of parameters $(\alpha, \gamma) \in \{0, 0.1, 0.2, \dots, 0.9, 1\}^2$ with $\gamma \leq \alpha$. We have computed the value of \mathfrak{C}_{norm} on all these trees, and we have used the distribution of these values as an estimation of the corresponding theoretical distribution. To test whether the distribution of the normalized Colless-like index on TreeBASE (or on some subset of it: see below) fits one of these theoretical distributions, we have performed two non-parametric statistical tests on the observed set of indices of TreeBASE and the corresponding simulated set of indices: Pearson's chi-squared test and the Kolmogorov-Smirnov test, using bootstrapping techniques in the latter to avoid problems with ties.

As a first approach, we have performed these tests on the whole set of trees in TreeBASE. The p-values obtained in all tests, be it for the uniform model or for any considered pair (α, γ) , have turned out to be negligible. Then, we conclude confidently that the distribution of the normalized Colless-like index on TreeBASE does not fit either the uniform model or any α - γ -model when we round α, γ to one decimal place. For instance, Fig. 4.20 displays the distribution of \mathfrak{C}_{norm} on TreeBASE and its estimated theoretical distribution under the uniform model. As it can be seen, these distributions are quite different, which supports the conclusion of the statistical test.

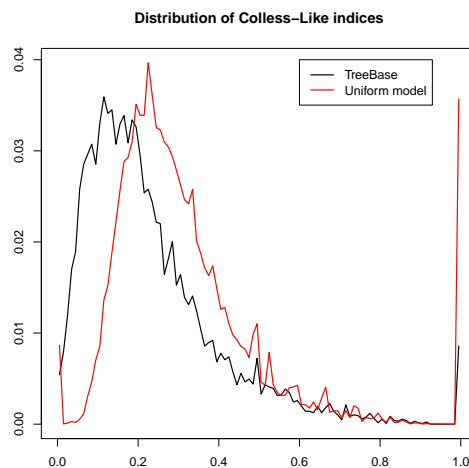


Figure 4.20: The distribution of \mathfrak{C}_{norm} on all trees in TreeBASE (black line) and its estimated theoretical distribution under the uniform model (red line).

Fig. 4.21 displays the distribution of \mathfrak{C}_{norm} for all trees in TreeBASE and its estimated theoretical distribution under the α - γ -model for the pair of

parameters α, γ that gave the largest p-values in the goodness of fit tests, which are $\alpha = 0.7$ and $\gamma = 0.4$. Although graphically both distributions are quite similar, the p-values of the Pearson chi-squared test and of the Kolmogorov-Smirnov test are virtually zero. One might think that the high “peaks” of the theoretical distribution near 0 and 1 could have influenced the outcome of these statistical tests. For this reason, we have repeated them without taking into account these “extreme” values, and the results have been the same.

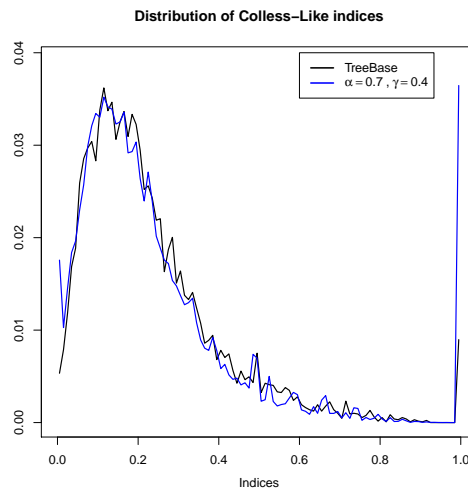


Figure 4.21: The distribution of \mathfrak{C}_{norm} on all trees in TreeBASE (black line) and its estimated theoretical distribution under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$ (blue line).

Since TreeBASE gathers phylogenetic trees of different types and from different sources, we have also considered subsets of it defined by means of attributes. More specifically, besides the whole TreeBASE as explained above, we have also considered the following subsets of it:

- All trees in TreeBASE up to repetitions: we have removed 513 repeated trees (which represent about a 4% of the total).
- All trees with their **kind** attribute equal to “Species”. This **kind** attribute can take three values: “Barcode tree”, “Gene Tree” and “Species Tree”.
- All trees with their **kind** attribute equal to “Species” and their **type** attribute equal to “Consensus”. This **type** attribute can take two values: “Consensus” and “Single”.
- All trees with their **kind** attribute equal to “Species” and their **type** attribute equal to “Single”.

We have repeated the study explained above for these four subsets of TreeBASE, comparing the distribution of the normalized Colless-like indices of their trees with the estimated theoretical distributions by means of goodness-of-fit tests, and the results have been the same, that is, all p-values have also turned out to be negligible. Our conclusion is, then, that neither the whole TreeBASE nor any of these four subsets of it seem to fit either the uniform model or some α - γ -model.

4.7 Tables of values for some examples of $\mathfrak{C}_{D,f}$

Tree	δ_f	$\mathfrak{C}_{\text{MDM},f}$	$\mathfrak{C}_{\text{var},f}$	$\mathfrak{C}_{sd,f}$
	$x_2 + 2x_0$	0	0	0
	$x_3 + 3x_0$	0	0	0
	$2x_2 + 3x_0$	$\frac{1}{2}(x_2 + x_0)$	$\frac{1}{2}(x_2 + x_0)^2$	$\frac{1}{\sqrt{2}}(x_2 + x_0)$
	$x_4 + 4x_0$	0	0	0
	$x_3 + x_2 + 4x_0$	$\frac{1}{2}(x_3 + 2x_0)$	$\frac{1}{2}(x_3 + 2x_0)^2$	$\frac{1}{\sqrt{2}}(x_3 + 2x_0)$
	$x_3 + x_2 + 4x_0$	$\frac{1}{3}(x_2 + x_0)$	$\frac{1}{3}(x_2 + x_0)^2$	$\frac{1}{\sqrt{3}}(x_2 + x_0)$
	$3x_2 + 4x_0$	0	0	0
	$3x_2 + 4x_0$	$\frac{3}{2}(x_2 + x_0)$	$\frac{5}{2}(x_2 + x_0)^2$	$\frac{3}{\sqrt{2}}(x_2 + x_0)$
	$x_5 + 5x_0$	0	0	0
	$x_4 + x_2 + 5x_0$	$\frac{1}{2}(x_4 + 3x_0)$	$\frac{1}{2}(x_4 + 3x_0)^2$	$\frac{1}{\sqrt{2}}(x_4 + 3x_0)$
	$x_4 + x_2 + 5x_0$	$\frac{1}{4}(x_2 + x_0)$	$\frac{1}{4}(x_2 + x_0)^2$	$\frac{1}{2}(x_2 + x_0)$
	$2x_3 + 5x_0$	$\frac{1}{3}(x_3 + 2x_0)$	$\frac{1}{3}(x_3 + 2x_0)^2$	$\frac{1}{\sqrt{3}}(x_3 + 2x_0)$
	$x_3 + 2x_2 + 5x_0$	$\frac{1}{2} x_3 - x_2 + x_0 $	$\frac{1}{2}(x_3 - x_2 + x_0)^2$	$\frac{1}{\sqrt{2}} x_3 - x_2 + x_0 $
	$x_3 + 2x_2 + 5x_0$	$\frac{1}{2}(2x_3 + x_2 + 5x_0)$	$\frac{1}{2}(x_3 + 2x_0)^2 + \frac{1}{2}(x_3 + x_2 + 3x_0)^2$	$\frac{1}{2}(2x_3 + x_2 + 5x_0)$
	$x_3 + 2x_2 + 5x_0$	$\frac{1}{6}(3x_3 + 5x_2 + 11x_0)$	$\frac{1}{3}(x_2 + x_0)^2 + \frac{1}{2}(x_3 + x_2 + 3x_0)^2$	$\frac{1}{\sqrt{3}}(x_2 + x_0) + \frac{1}{\sqrt{2}}(x_3 + x_2 + 3x_0)$
	$x_3 + 2x_2 + 5x_0$	$\frac{1}{3}(x_2 + x_0)$	$\frac{1}{3}(x_2 + x_0)^2$	$\frac{1}{\sqrt{3}}(x_2 + x_0)$
	$x_3 + 2x_2 + 5x_0$	$\frac{7}{6}(x_2 + x_0)$	$\frac{11}{6}(x_2 + x_0)^2$	$\frac{\sqrt{3}+2\sqrt{2}}{\sqrt{6}}(x_2 + x_0)$
	$4x_2 + 5x_0$	$x_2 + x_0$	$(x_2 + x_0)^2$	$\sqrt{2}(x_2 + x_0)$
	$4x_2 + 5x_0$	$\frac{3}{2}(x_2 + x_0)$	$\frac{9}{2}(x_2 + x_0)^2$	$\frac{3}{\sqrt{2}}(x_2 + x_0)$
	$4x_2 + 5x_0$	$3(x_2 + x_0)$	$7(x_2 + x_0)^2$	$3\sqrt{2}(x_2 + x_0)$

Table 4.2: Abstract values of δ_f , $\mathfrak{C}_{\text{MDM},f}$, $\mathfrak{C}_{\text{var},f}$, and $\mathfrak{C}_{sd,f}$ on \mathcal{T}_n^* for $n = 2, 3, 4, 5$. For space reasons, we denote $f(i)$ by x_i .





















Tree	δ_{\ln}	$\mathfrak{C}_{\text{MDM},\ln}$	Pos.	$\mathfrak{C}_{\text{var},\ln}$	Pos.	$\mathfrak{C}_{sd,\ln}$	Pos.
	3.5514	0	(1)	0	(1)	0	(1)
	4.7437	0	(1)	0	(1)	0	(1)
	6.1029	1.2757	(2)	3.2549	(2)	1.8041	(2)
	5.9048	0	(1)	0	(1)	0	(1)
	8.6543	0	(1)	0	(1)	0	(1)
	7.2951	0.8505	(2)	2.1700	(2)	1.4731	(2)
	7.2951	1.8718	(3)	7.0075	(3)	2.6472	(3)
	8.6543	3.8272	(4)	16.2747	(4)	5.4124	(4)
	7.0436	0	(1)	0	(1)	0	(1)
	9.8466	0.5961	(2)	0.7107	(2)	0.8430	(2)
	8.4563	0.6379	(3)	1.6275	(3)	1.2757	(3)
	9.8466	0.8505	(4)	2.1700	(4)	1.4731	(4)
	8.4873	1.2479	(5)	4.6717	(5)	2.1614	(5)
	8.4563	2.4524	(6)	12.0287	(7)	3.4682	(6)
	11.2058	2.5514	(7)	6.5099	(6)	3.6083	(7)
	9.8466	2.9767	(8)	11.9348	(8)	4.7503	(8)
	11.2058	3.8272	(9)	29.2944	(11)	5.4124	(9)
	9.8466	3.9980	(10)	21.9842	(9)	5.9244	(10)
	9.8466	5.0194	(11)	26.8218	(10)	7.0985	(11)
	11.2058	7.6543	(12)	45.5691	(12)	10.8249	(12)

Table 4.3: Numerical values (rounded to 4 decimal places) of δ_{\ln} , where \ln stands for the function $n \mapsto \ln(n + e)$, and of $\mathfrak{C}_{D,\ln}$, for $D = \text{MDM}$, var or sd , on \mathcal{T}_n^* , for $n = 2, 3, 4, 5$. The columns labeled “Pos.” give the position of the tree in its \mathcal{T}_n^* in increasing order of the Colless-like balance index corresponding to the column on its left. The rows are sorted, for each n , in increasing order of $\mathfrak{C}_{\text{MDM},\ln}$.





















Tree	δ_{e^n}	$\mathfrak{C}_{\text{MDM},e^n}$	Pos.	$\mathfrak{C}_{\text{var},e^n}$	Pos.	\mathfrak{C}_{sd,e^n}	Pos.
	9.3891	0	(1)	0	(1)	0	(1)
	23.0855	0	(1)	0	(1)	0	(1)
	17.7781	4.1945	(2)	35.1881	(2)	5.9320	(2)
	58.5982	0	(1)	0	(1)	0	(1)
	26.1672	0	(1)	0	(1)	0	(1)
	31.4746	2.7964	(2)	23.4588	(2)	4.8434	(2)
	31.4746	11.0428	(3)	243.8855	(4)	15.6168	(3)
	26.1672	12.5836	(4)	175.9407	(3)	17.7959	(4)
	153.4132	0	(1)	0	(1)	0	(1)
	39.8636	6.8482	(4)	93.7968	(5)	9.6849	(4)
	66.9872	2.0973	(2)	17.5941	(2)	4.1945	(2)
	39.8636	2.7964	(3)	23.4588	(3)	4.8434	(3)
	45.1711	7.3618	(5)	162.5903	(7)	12.7511	(6)
	66.9872	28.7991	(12)	1658.7734	(12)	40.7280	(12)
	34.5562	8.3891	(6)	70.3763	(4)	11.8639	(5)
	39.8636	9.7872	(7)	129.0231	(6)	15.6188	(7)
	34.5562	12.5836	(8)	316.6932	(8)	17.7959	(8)
	39.8636	18.0336	(9)	487.8092	(9)	26.3922	(9)
	39.8636	26.2801	(11)	708.2359	(11)	37.1656	(11)
	34.5562	25.1672	(10)	492.6338	(10)	35.5918	(10)

Table 4.4: Numerical values (rounded to 4 decimal places) of δ_{e^n} and of \mathfrak{C}_{D,e^n} , for $D = \text{MDM}$, var or sd , on \mathcal{T}_n^* , for $n = 2, 3, 4, 5$. The columns labeled “Pos.” give the position of the tree in its \mathcal{T}_n^* in increasing order of the Colless-like balance index corresponding to the column on its left. The rows are sorted, for each n , in increasing order of $\mathfrak{C}_{\text{MDM},\ln}$ (i.e., in the same order as in Table 4.3).

Conclusions and future work

The study of indices quantifying the graph-theoretical properties of phylogenetic trees and of metrics allowing for the comparison of phylogenetic trees is motivated by the hypothesis that the shape of a phylogenetic tree is a reflection of the properties of the evolutionary processes that have given rise to it. The main contribution of this PhD Thesis is the addition to the set of available techniques for the analysis and comparison of phylogenetic trees of two balance indices and a metric. More in detail, in this Thesis:

- We have defined the *total cophenetic index* Φ and we have proved that it is a good alternative to other popular balance indices like Sackin's and Colless' indices. Among the nice properties of the total cophenetic index, let us emphasize the following ones: it is defined for multifurcating trees; it achieves its maximum value exactly at the combs, which are considered the most unbalanced trees of any given number of leaves; it achieves its minimum value among the multifurcating trees exactly at the star trees and among the bifurcating trees at the maximally balanced trees, which are considered the most balanced trees of any given number of leaves, being the first balance index defined so far satisfying this last property; we have been able to compute closed formulas for its expected value under the Yule and the uniform models of phylogenetic tree growth and a simple recurrence for its variance under the uniform model (its variance under the Yule model has been computed elsewhere [23]), which makes it useful in evolutionary hypothesis tests with these models as null models, and we have provided a proof-of-concept experiment in this sense; and we have shown experimentally that it has a lower rate of ties than the Sackin and Colless indices. Let us mention moreover that, as a by-product of our study of Φ , we have obtained a closed formula for the expected value of the Sackin index under the uniform model, for which only the order of growth in the number n of leaves was known so far [14].
- We have defined a family of *Colless-like indices* that provide the first sound extension to multifurcating trees of the Colless index for bifurcating trees, in such a way that, on the one hand, when restricted to bifurcating trees they give the classical Colless index up to a constant factor and, on the other hand, the only trees with any given number of leaves n

that achieve the minimum value of these indices, 0, are exactly the fully symmetric trees with n leaves. These Colless-like indices depend on the choice of a dissimilarity function and of a notion of *size* of a rooted tree, and we show that this choice may affect how they measure the balance of a tree. In particular, we provide a subfamily of these Colless-like indices for which the maximum values are achieved at the combs. In connection with these indices, we introduce in this Thesis our R package “CollessLike”, available on the CRAN, that allows to perform goodness of fit tests of a phylogenetic tree with null model any α - γ -model for multifurcating trees.

- Inspired by an old paper by Sokal and Rohlf [102] and by our work on the total cophenetic index, we have defined a new family of metrics for phylogenetic trees, the *cophenetic metrics* $d_{\varphi,p}$, with $p \in \{0\} \cup [1, \infty[$ parameterizing these metrics through their dependence on an L^p norm. These metrics can be used to compare phylogenetic trees with nested taxa and weights on the arcs, provided that they have the same sets of labels. In the non-weighted case (which is the only case when these questions make sense) we have computed their minimum non-zero value, the neighborhood of any given phylogenetic tree, and the order of their diameter on different spaces of phylogenetic trees. Moreover, we have computed closed formulas for the expected value under the Yule and the uniform models of the square of the metric $d_{\varphi,2}$ on the space of bifurcating phylogenetic trees and we have estimated its variance. This allows the use also of this metric in evolutionary hypothesis tests with these models as null models. The use of metrics in this connection seems to be new in the literature.

Some of the developments of the topics treated in this Thesis that we consider worth to study in the future are:

- To extend the study of the behavior of the total cophenetic index to some parameterized models of evolutionary tree growth, like Ford’s α -model or Aldous’ β -model [3, 4] in the bifurcating case, and the α - γ -model and the uniform model in the multifurcating case. Let us mention in this connection that the limit distribution of Φ under the Yule model has been recently established by K. Bartoszek [9].
- To obtain analytically some information on the expected value and the variance of some Colless-like index under the α - γ -model or the multifurcating uniform model.
- To expand the CollessLike package by incorporating also the aforementioned β -model.
- To study the balance of *multilabeled trees*, rooted trees T without elementary nodes endowed with a surjective (but not necessarily injective)

labeling function from $L(T)$ to a set of labels S [60]. This involves, on the one hand, to define balance indices that capture a suitable notion of “balance” for multilabeled trees and to characterize the trees achieving the maximum and minimum value on suitable spaces of such trees (for instance, fixing the number of leaves, or the number of labels, or the multiplicities of the different labels) and, on the other hand, to define probabilistic models of multilabeled tree growth, for instance involving mutations and duplications, and to determine the behavior of these balance indices under these models.

- To study the balance of *taxonomic trees* [28]. These rooted multifurcating trees, whose paradigm are the usual taxonomies with a fixed number of taxonomic ranks, have all their leaves of the same depth (the number of ranks in the taxonomy) but they may have *elementary* nodes, that is, internal nodes of out-degree 1 (corresponding to intermediate taxonomic ranks containing organisms of only one immediately lower rank). This study should follow the same steps as in the case of multilabeled trees: to define suitable balance indices and probabilistic models for them and to determine the behavior of the former under the latter.
- To study the balance of *rooted phylogenetic networks* [62]. The first step would be to clarify the notion of balance in the context of phylogenetic networks, since in them paths from nodes to leaves need no longer be unique and moreover there are different notions of descendance of a leaf x from a node v (for instance, *strict*, when all paths from the root to x contain the node v ; of *tree type*, when there is a path from v to x that avoids reticulation nodes, with more than one parent; etc.). Probably the best plan would be to focus first on spaces of phylogenetic trees with some strong topological restriction that makes them “close to trees”, like the *galled trees* [52, 95], the *tree-child networks* [25] or the *LGT-networks* [25]. In these three cases, moreover, one has available algorithms to build all such networks with a given number of leaves [24, 51, 84] that could lay the basis for the definition of probabilistic models for them.

Bibliography

- [1] Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1964. Available at <http://people.math.sfu.ca/~cbm/aands/> (Last visited, 26/12/2018).
- [2] Cyrus Afrasiabi, Bushra Samad, David Dineen, Christopher Meacham, and Kimmen Sjölander. The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Research*, 41(W1):W242–W248, 2013.
- [3] David J. Aldous. *Random discrete structures*, volume 76 of *The IMA Volumes in Mathematics and its Applications*, chapter Probability distributions on cladograms, pages 1–18. Springer, 1996.
- [4] David J. Aldous. Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science*, 16:23–34, 2001.
- [5] Benjamin L. Allen and Mike Steel. Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics*, 5:1–15, 2001.
- [6] Wolfram Alpha. The Mathematical Functions Site, 1998–2019. Available online at <http://functions.wolfram.com/>.
- [7] J. David Archibald. *Aristotle’s Ladder, Darwin’s Tree: The Evolution of Visual Metaphors for Biological Order*. Columbia University Press, 2104.
- [8] Mukul S. Bansal, J. Gordon Burleigh, Oliver Eulenstein, and David Fernández-Baca. Robinson-Foulds Supertrees. *Algorithms for Molecular Biology*, 5(1):18, 2010.
- [9] Krzysztof Bartoszek. Exact and approximate limit behaviour of the Yule tree’s cophenetic index. *Mathematical Biosciences*, 303:26–45, 2018.
- [10] David A. Baum and Susan Offner. Phylogenics & tree-thinking. *The American Biology Teacher*, 70(4):222–230, 2008.
- [11] David A. Baum, Stacey DeWitt Smith, and Samuel S. S. Donovan. The tree-thinking challenge. *Science*, 310(5750):979–980, 2005.

-
- [12] Edwin F. Beckenbach and Richard Bellman. *Inequalities*, volume 30. Springer Science & Business Media, 2012.
- [13] Michael G. B. Blum and Olivier François. Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance. *Systematic Biology*, 55:685–691, 2006.
- [14] Michael G. B. Blum, Olivier François, and Svante Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Annals of Applied Probability*, 16:2195–2214, 2006.
- [15] Michael G.B. Blum and Olivier François. On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Mathematical Biosciences*, 195:141–153, 2005.
- [16] Carl Boettiger and Duncan Temple Lang. Treebase: an R package for discovery, access and manipulation of online phylogenies. *Methods in Ecology and Evolution*, 3(6):1060–1066, 2012.
- [17] James K. M. Brown. Probabilities of Evolutionary Trees. *Systematic Biology*, 43:78–91, 1994.
- [18] David Bryant and Mike Steel. Computing the Distribution of a Tree Metric. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 16:420–426, 2009.
- [19] Georges-Louis Leclerc Comte de Buffon. *Histoire naturelle generale et particuliere*, volume 5. chez F. Dufart, 1755.
- [20] David Callan. A combinatorial survey of identities for the double factorial. *arXiv preprint arXiv:0906.1317*, 2009.
- [21] Gabriel Cardona, Mercè Llabrés, Francesc Rosselló, and Gabriel Valiente. Nodal distances for rooted phylogenetic trees. *Journal of Mathematical Biology*, 61(2):253–276, 2010.
- [22] Gabriel Cardona, Arnau Mir, and Francesc Rosselló. The expected value under the Yule model of the squared path-difference distance. *Applied Mathematics Letters*, 25:2031–2036, 2012.
- [23] Gabriel Cardona, Arnau Mir, and Francesc Rosselló. Exact formulas for the variance of several balance indices under the Yule model. *Journal of Mathematical Biology*, 67(6-7):1833–1846, 2013.
- [24] Gabriel Cardona, Joan Carles Pons, and Celine Scornavacca. Generation of Binary Tree-Child phylogenetic networks. *PLoS Computational Biology*, 15(9):1–29, 2019.

- [25] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. Comparison of Tree-Child Phylogenetic Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):552–569, 2009.
- [26] Luigi L. Cavalli-Sforza and Anthony W. F. Edwards. Phylogenetic analysis: Models and estimation procedures. *Evolution*, 21(3):550–570, 1967.
- [27] Bo Chen, Daniel Ford, and Matthias Winkel. A new family of Markov branching trees: the alpha-gamma model. *Electronic Journal of Probability*, 14:400–430, 2009.
- [28] K. Robert Clarke and Richard M. Warwick. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, 35(4):523–531, 1998.
- [29] Donald H. Colless. Review of “Phylogenetics: The Theory and Practice of Phylogenetic Systematics”. *Systematic Zoology*, 31:100–104, 1982.
- [30] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms (3rd edition)*. The MIT Press, 2009.
- [31] Tomás M. Coronado, Mareike Fischer, Lina Herbst, Francesc Rosselló, and Kristina Wicke. On the minimum value of the Colless index and the bifurcating trees that achieve it. *arXiv preprint arXiv:1907.05064*, 2019.
- [32] Tomás M. Coronado, Arnau Mir, and Francesc Rosselló. The Probabilities of Trees and Cladograms under Ford’s α -Model. *The Scientific World Journal*, 2018:1916094, 2018.
- [33] Tomás M. Coronado, Arnau Mir, Francesc Rosselló, and Gabriel Valiente. A balance index for phylogenetic trees based on rooted quartets. *Journal of Mathematical Biology*, pages 1105–1148, 2019.
- [34] Douglas E. Critchlow, Dennis K. Pearl, and Chunlin Qian. The Triples Distance for Rooted Bifurcating Phylogenetic Trees. *Systematic Biology*, 45(3):323–334, 1996.
- [35] Charles Darwin. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859.
- [36] Ruchira S. Datta, Christopher Meacham, Bushra Samad, Christoph Neyer, and Kimmen Sjölander. Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Research*, 37(suppl.2):W84–W89, 2009.
- [37] W. Ford Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999.

-
- [38] W. Ford Doolittle and Tyler D. P. Brunet. What is the Tree of Life? *PLoS Genetics*, 12(4):e1005912, 2016.
- [39] Antoine-Nicolas Duchesne. *Histoire naturelle des fraisiers contenant les vues d'économie réunies à la botanique, et suivie de remarques particulières sur plusieurs points qui ont rapport à l'histoire naturelle générale.* chez Didot le jeune, 1766.
- [40] James S. Farris. A Successive Approximations Approach to Character Weighting. *Systematic Zoology*, 18:374–385, 1969.
- [41] James S. Farris. On Comparing the Shapes of Taxonomic Trees. *Systematic Zoology*, 22:50–54, 1973.
- [42] James S. Farris, Arnold G. Kluge, and Michael J. Eckardt. A Numerical Approach to Phylogenetic Systematics. *Systematic Zoology*, 19:172–189, 1970.
- [43] Joseph Felsenstein. *Inferring Phylogenies.* Sinauer Associates Inc, 2004.
- [44] C. R. Finden and A. D. Gordon. Obtaining common pruned trees. *Journal of Classification*, 2:255–276, 1985.
- [45] Mareike Fischer. Extremal values of the Sackin balance index for rooted binary trees. *arXiv preprint arXiv:1801.10418*, 2018.
- [46] Daniel J. Ford. Probabilities on cladograms: introduction to the alpha model. *arXiv preprint arXiv:math/0511246*, 2005.
- [47] George Gasper and Mizan Rahman. *Basic Hypergeometric Series*, volume 96. Cambridge University Press, 2004.
- [48] Wayne Goddard, Ewa Kubicka, Grzegorz Kubicki, and F. R. McMorris. The agreement metric for labeled binary trees. *Mathematical Biosciences*, 123:215–226, 1994.
- [49] R. William Gosper. Decision procedure for indefinite hypergeometric summation. *Proceedings of the National Academy of Sciences*, 75(1):40–42, 1978.
- [50] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A foundation for computer science (2n edition).* Addison-Wesley, 1994.
- [51] Andreas DM Gunawan, Jeyaram Rathin, and Louxin Zhang. Counting and Enumerating Galled Networks. *arXiv preprint arXiv:1812.08569*, 2018.

- [52] Dan Gusfield, Satish Eddhu, and Charles Langley. Efficient reconstruction of phylogenetic networks with constrained recombination. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference*, pages 363–374, 2003.
- [53] Julia Handl, Joshua Knowles, and Douglas B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21:3201–3212, 2005.
- [54] Katherine G. Harcourt-Brown, Paul N. Pearson, and Mark Wilkinson. The imbalance of paleontological trees. *Paleobiology*, 27(2):188–204, 2001.
- [55] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3:44–77, 1971.
- [56] Stephen B. Heard. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46(6):1818–1826, 1992.
- [57] M. D. Hendy, C. H. C. Little, and David Penny. Comparing Trees with Pendant Vertices Labelled. *SIAM Journal of Applied Mathematics*, 44:1054–1065, 1984.
- [58] Jody Hey. Using phylogenetic trees to study speciation and extinction. *Evolution*, 46(3):627–640, 1992.
- [59] Cody E. Hinchliff, Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):12764–12769, 2015.
- [60] Katharina T. Huber and Vincent Moulton. Phylogenetic networks from multi-labelled trees. *Journal of Mathematical Biology*, 52(5):613–632, 2006.
- [61] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, et al. A new view of the tree of life. *Nature Microbiology*, 1(5):16048, 2016.
- [62] Daniel Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks. Concepts, Algorithms and Applications*. Cambridge University Press, Cambridge, UK, 2010.
- [63] Mark Kirxpatrick and Montgomery Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47:1171–1181, 1993.

- [64] Wolfram Koepf. *Hypergeometric Summation*. Springer, 1998.
- [65] Nandini Krishnamurthy, Duncan P. Brown, Dan Kirshner, and Kimmen Sjölander. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome biology*, 7(9):R83, 2006.
- [66] Jean-Baptiste de Monet Lamarck. *Philosophie zoologique, ou Exposition des considérations relatives à l’histoire naturelle des animaux*, volume 1. Dentu, 1809.
- [67] Brendan B. Larsen, Elizabeth C. Miller, Matthew K. Rhodes, and John J. Wiens. Inordinate fondness multiplied and redistributed: The number of species on Earth and the new pie of life. *The Quarterly Review of Biology*, 92(3):229–265, 2017.
- [68] Heng Li, Avril Coghlan, Jue Ruan, Lachlan James Coin, Jean-Karim Hériché, Lara Osmotherly, Ruiqiang Li, Tao Liu, Zhang Zhang, Lars Bolund, Gane Ka-Shu Wong, Weimou Zheng, Paramvir Dehal, Jun Wang, Jun Wang, and Richard Durbin. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research*, 34(Database issue):D572—80, 2006.
- [69] Frederick Matsen. Optimization Over a Class of Tree Shape Statistics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(3):506–512, 2007.
- [70] Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164:81–92, 2000.
- [71] David P. Mindell. The tree of life: metaphor, model, and heuristic device. *Systematic biology*, 62(3):479–489, 2013.
- [72] Arnau Mir and Francesc Rosselló. The mean value of the squared path-difference distance for rooted phylogenetic trees. *Journal of Mathematical Analysis and Applications*, 371:168–176, 2010.
- [73] Arnau Mir, Francesc Rosselló, and Lucía Rotger. A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1):125–136, 2013.
- [74] Arnau Mir, Francesc Rosselló, and Lucía Rotger. *CollessLike: Distribution and Percentile of Sackin, Cophenetic and Colless-Like Balance Indices of Phylogenetic Trees*. <https://CRAN.R-project.org/package=CollessLike>, 2018. R package version 1.0.
- [75] Arnau Mir, Lucía Rotger, and Francesc Rosselló. Sound Colless-like balance indices for multifurcating trees. *PloS one*, 13(9):e0203401, 2018.

- [76] Arne O. Mooers and Stephen B. Heard. Inferring Evolutionary Process from Phylogenetic Tree Shape. *The Quarterly Review of Biology*, 72(1):31–54, 1997.
- [77] Willem H. Mulder. Probability distributions of ancestries and genealogical distances on stochastically generated rooted binary trees. *Journal of Theoretical Biology*, 280:139–145, 2011.
- [78] Martha I. Nelson and Edward C. Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8(3):196–205, 2007.
- [79] Neal L. Oden and Kwang-Tsao Shao. An algorithm to equiprobably generate all directed trees with k labeled terminal nodes and unlabeled interior nodes. *Bulletin of Mathematical Biology*, 46(3):379–387, 1984.
- [80] Maureen A. O’Malley and Eugene V. Koonin. How stands the Tree of Life a century and a half after *The Origin?* *Biology Direct*, 6(1):32, 2011.
- [81] Marko Petkovšek, Herbert S. Wilf, and Doron Zeilberger. $A = B$. AK Peters Ltd., 1996. Available online at <https://www.math.upenn.edu/~wilf/AeqB.html>.
- [82] J. B. Phipps. Dendrogram Topology. *Systematic Zoology*, 20:306–308, 1971.
- [83] William H. Piel, Lucie Chan, Mark J. Dominus, Jin Ruan, Rutger Aldo Vos, and Val Tannen. Treebase v. 2: A database of phylogenetic knowledge. In *e-BioSphere*, 2009.
- [84] Joan Carles Pons, Celine Scornavacca, and Gabriel Cardona. Generation of level- k lgt networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.
- [85] Mark A. Ragan. Trees and networks before and after Darwin. *Biology Direct*, 4(1):43, 2009.
- [86] Mark A. Ragan, James O. McInerney, and James A. Lake. The network of life: genome beginnings and evolution, 2009.
- [87] Guillermo Restrepo, Héber Mesa, and Eugenio J Llanos. Three Dissimilarity Measures to Contrast Dendrograms. *Journal of Chemical Information and Modeling*, 47:761–770, 2007.
- [88] David F. Robinson and Leslie R. Foulds. Comparison of weighted labelled trees. In *Combinatorial Mathematics VI. Lecture Notes in Mathematics, vol 748*, pages 119–126. Springer Berlin Heidelberg, 1979.
- [89] David F. Robinson and Leslie R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981.

-
- [90] James S. Rogers. Response of Colless's Tree Imbalance to Number of Terminal Taxa. *Systematic Biology*, 42(1):102–105, 1993.
- [91] James S. Rogers. Central Moments and Probability Distributions of Colless's Coefficient of Tree Imbalance. *Evolution*, 48:2026–2036, 1994.
- [92] James S. Rogers. Central Moments and Probability Distributions of Three Measures of Phylogenetic Tree Imbalance. *Systematic Biology*, 45(1):99–110, 1996.
- [93] F. James Rohlf and Robert R. Sokal. Comparing Numerical Taxonomic Studies. *Systematic Biology*, 30(4):459–490, 1981.
- [94] Donn E. Rosen. Vicariant Patterns and Historical Explanation in Biogeography. *Systematic Biology*, 27(2):159–188, 1978.
- [95] Francesc Rosselló and Gabriel Valiente. All that glitters is not galled. *Mathematical Biosciences*, 221(1):54–59, 2009.
- [96] Lucía Rotger. CollessLike. <https://github.com/LuciaRotger/CollessLike>, 2018.
- [97] Lucía Rotger. PhD-Code. <https://github.com/LuciaRotger/PhD-Code>, 2019.
- [98] MJ Sackin. “Good” and “Bad” phenograms. *Systematic Zoology*, 21(2):225–226, 1972.
- [99] Kwang-Tsao Shao and Robert R. Sokal. Tree Balance. *Systematic Zoology*, 39:266–276, 1990.
- [100] Neil Sloane. The On-line Encyclopedia of Integer Sequences, 2010. Available online at <http://oeis.org/>.
- [101] Robert R. Sokal. A phylogenetic analysis of the Caminalcules I: The data base. *Systematic Biology*, 32:159–184, 1983.
- [102] Robert R. Sokal and F. James Rohlf. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11:33–40, 1962.
- [103] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.
- [104] Mike Steel and Andy McKenzie. Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 170:91–112, 2001.
- [105] Mike Steel and David Penny. Distributions of Tree Comparison Metrics—Some New Results. *Systematic Biology*, 42:126–141, 1993.

- [106] M. Stich and S. C. Manrubia. Topological properties of phylogenetic trees in evolutionary models. *The European Physical Journal B*, 70(4):583–592, jul 2009.
- [107] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2008. <http://www.R-project.org>.
- [108] Rutger Aldo Vos. *Inferring large phylogenies: The big tree problem*. Doctoral dissertation, Biological Sciences Department-Simon Fraser University, 2006.
- [109] Rutger Aldo Vos, James P. Balhoff, Jason A. Caravas, Mark T. Holder, Hilmar Lapp, Wayne P. Maddison, Peter E. Midford, Anurag Priyam, Jeet Sukumaran, Xuhua Xia, and Arlin Stoltzfus. Nexml: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, 61:675–689, 2012.
- [110] Michael S. Waterman and Temple F. Smith. On the similarity of dendrograms. *Journal of Theoretical Biology*, 73:789–800, 1978.
- [111] Eric W. Weisstein. Power sum. <http://mathworld.wolfram.com/PowerSum.html>.
- [112] William T. Williams and Harold T. Clifford. On the Comparison of Two Classifications of the Same Set of Elements. *Taxon*, 20:519–522, 1971.
- [113] Taoyang Wu and Kwok Pui Choi. On joint subtree distributions under two evolutionary models. *Theoretical Population Biology*, 108:13–23, 2016.
- [114] Yang Xiang, Zoe Jingyu Zhu, and Yu Li. Enumerating Unlabeled and Root Labeled Trees for Causal Model Acquisition. In *Advances in Artificial Intelligence*, pages 158–170. Springer, Berlin, 2009.
- [115] Xiaobo Yan, Tao Tang, Yu Deng, Jing Du, and Xuejun Yang. Evaluation of Transcendental Functions on Imagine Architecture. In *International Conference on Parallel Processing*, page 53. IEEE, 2007.
- [116] George Udny Yule. A Mathematical Theory of Evolution, based on the Conclusions of Dr J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213:21–87, 1924.
- [117] Yang Zhong, Christopher A. Meacham, and Sakti Pramanik. A general method for tree-comparison based on subtree similarity and its use in a taxonomic database. *Biosystems*, 42:1–8, 1997.

- [118] Sha Zhu, James H. Degnan, and Mike Steel. Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theoretical Population Biology*, 79(4):220–227, 2011.
- [119] Sha Zhu, Coung Than, and Taoyang Wu. Clades and clans: a comparison study of two evolutionary models. *Journal of Mathematical Biology*, 71(1):99–124, 2015.

Appendix A

Scripts

For reproducibility, in this Appendix we provide the code used in the experiments and simulations reported in the previous chapters, including the code that generates the plots resulting from these computations. The code and the files produced with it are also available on the GitHub repository associated to this PhD Thesis [97]. In particular, it should be understood that all data tables generated with the code explained in this Appendix are available on this GitHub repository.

A.1 Packages required

The functions used in this appendix need the following R packages to be installed and loaded:

```
library(Zseq)
library(gmp)
library(ape)
library(igraph)
library(CollessLike)
```

A.2 List of all binary trees

We have obtained all phylogenetic trees in \mathcal{BT}_n for $n = 3, \dots, 8$ using the Python package *phylonetwork*:

```
import phylonetwork.generators as gen
from phylonetwork.distances import cophenetic_distance as cophdist
from math import factorial

for n in range(3,9):
```

```

taxa = [str(i+1) for i in range(n)]
tg=gen.all_trees(taxa=taxa, binary= True, nested_taxa= False)
trees = list(tg)
newicks = []
file = open("bintrees-n"+str(n)+". ", "w+")
for i in range(len(trees)):
    newicks.append(trees[i].eNewick())
    print >>file, newicks[i]
file.close()

```

The resulting lists of trees are available in the *List of Trees* folder of the PhD Thesis GitHub repository.

A.3 General functions

The following functions are needed in some computations performed in the next two sections.

```

big.factorial = function(n){
  if(n<2) return(1)
  return(Factorial(n+1) [n+1])
}

big.double.factorial = function(n){
  if(n<2) return(1)
  m = (n+2+n%%2)/2
  return(Factorial.Double(m, odd=(n%%2==1)) [m])
}

big.binomial = function(n,k){
  return(big.factorial(n)/(big.factorial(k)*big.factorial(n-k)))
}

Cknk = function(k,n){
  return(big.binomial(n,k)*((big.double.factorial(2*k-3)*
    big.double.factorial(2*(n-k)-3))/
    (2*big.double.factorial(2*n-3))))
}

```


A.4 Scripts from Chapter 2

A.4.1 Computation of $E_Y(\Phi_n)$

The formula in Theorem 2.20 can be computed with the following function:

```
harmonic=function(n){return(sum(1/(1:n)))}
EYPhi=function(n){
  return(n*(n+1-2*harmonic(n)))
}
```

For $n = 3, \dots, 20$, the results are:

```
sapply(3:20,EYPhi)
```

```
## [1] 1.000000 3.333333 7.166667 12.600000 19.700000
## [6] 28.514286 39.078571 51.420635 65.562698 81.522944
## [11] 99.316522 118.956255 140.453130 163.816672 189.055214
## [16] 216.176109 245.185893 276.090414
```

To double-check that formula, we have computed the values of $E_Y(\Phi_n)$, for $n = 3, \dots, 8$, from the cophenetic indices of all trees in the corresponding \mathcal{BT}_n . To do that, we have used the full content of \mathcal{BT}_n , for $n = 3, \dots, 8$, obtained in Section A.2. Then, on the one hand, we have computed the probability of each tree under the Yule model with the following function:

```
yule.prob = function(tree){
  if (class(tree)=="phylo")
    tree=graph.edgelist(tree$edge, directed=TRUE)
  sp = shortest.paths(tree,mode = "out")
  deg = degree(tree,mode="out")
  leaves = which(deg==0)
  n = length(leaves)
  k.node = function(node){
    subtree = which(sp[node,]<Inf)
    return(length(intersect(leaves,subtree)))
  }
  kappas = sapply(which(deg>0), k.node)
  value = (2^(n-1)/as.numeric(big.factorial(n)))*
    prod(1/(kappas-1))
  return(value)
}
```

And, on the other hand, we have computed the total cophenetic index of each tree with the function `cophen.index` contained in our R package *CollessLike* (see §A.6.1). Finally, we have computed the desired expected value for each n as the sum over all phylogenetic trees in \mathcal{BT}_n of the product of their total cophenetic index and their probability:

```
exp.yule = c()
for(n in 3:8){
  trees=read.tree(file=paste("./bintrees-n",n,".txt",sep=""))
  indices = sapply(trees, cophen.index)
  probs = sapply(trees, yule.prob)
  exp.yule[n]=sum(indices*probs)
}
exp.yule

## [1] 1.000000 3.333333 7.166667 12.600000 19.700000
## [6] 28.514286
```

So, the results agree with the figures given by our formula.

A.4.2 Computation of $E_U(\Phi_n)$

The formula in Theorem 2.28 can be computed with the following function (it uses the function `big.double.factorial` explained in Section A.3):

```
EUPhi = function(n){
  return(as.numeric((n*(n-1)/4)*
    (big.double.factorial(2*n-2)/
    big.double.factorial(2*n-3)-2)))
}
```

For $n = 3, \dots, 20$ the results are:

```
sapply(3:20,EUPhi)

## [1] 1.000000 3.600000 8.285714 15.476190 25.545455 38.834499
## [7] 55.658741 76.313040 101.075256 130.208893 163.965117 202.584342
## [13] 246.297504 295.327098 349.888046 410.188417 476.430046 548.809061
```

To double-check that formula, we have computed the values of each $E_U(\Phi_n)$, for $n = 3, \dots, 8$, as the arithmetic mean of the total cophenetic indices of all phylogenetic trees in \mathcal{BT}_n computed with our function `cophen.index`.

```

exp.uni = c()
for(n in 3:8){
  trees=read.tree(file=paste("./bintrees-n",n, ".txt",sep=""))
  indices = sapply(trees, cophen.index)
  exp.uni[n]=mean(indices)
}
exp.uni

```

```

## [1] 1.000000 3.600000 8.285714 15.476190
## [5] 38.834499

```

Again, the results agree with the figures given by our formula.

A.4.3 Computation of $\sigma_U^2(\Phi_n)$

Computing the variance of Φ_n using our formula

We can compute $\sigma_U^2(\Phi_n)$ using the recurrence for $E_U(\Phi_n^2)$, the exact formula for $E_U(\Phi_n)$, and the identity

$$\sigma_U^2(\Phi_n) = E_U(\Phi_n^2) - E_U(\Phi_n)^2.$$

The following functions are needed to compute this variance, in addition to those in Section A.3.

```

EUPhi = function(n){
  return(as.numeric((n*(n-1)/4)*
    (big.double.factorial(2*n-2)/
    big.double.factorial(2*n-3)-2)))
}

term.Phi = function(n){
  return(mul.bigq(as.bigq(n*(n-1)/2), (mul.bigq(as.bigq(
    (49*n^3-57*n^2-22*n+24)/48),
    big.double.factorial(2*n-4)/big.double.factorial(
    2*n-3))-as.bigq((63*n^2-95*n+28)/30))))
}

compute.EUPhi2 = function(n.max=500){
  terms = lapply(2:n.max, term.Phi)
  terms = c(0, terms)
  exp.values = list(0)
  for(n in 2:n.max){
    sums = 0

```

```

    if(n>2){
      for(k in 2:(n-1)){
        sums = sums + Cknk(k,n)*exp.values[[k]]
      }
      sums = 2*sums
    }
    sums = sums + terms[[n]]
    exp.values[[n]] = sums
    print(n)
  }
  exp.values = sapply(exp.values, as.numeric)
  write.table(exp.values,file = paste("C2-EU(Phi2)",n.max,".txt",
                                     sep = ""),row.names = F,col.names = F)

  return(exp.values)
}

compute.varUPhi = function(exp.values,n.max=500){
  var.form = function(i)return(exp.values[i]-EUPhi(i)^2)
  var.values = sapply(1:n.max, var.form)
  write.table(var.values,file = paste("C2-varU(Phi)",n.max,".txt",
                                     sep = ""),row.names = F,col.names = F)

  return(var.values)
}

```

We have computed these variances up to $n = 1000$ with the following commands:

```

exp.values.Phi = compute.EUPhi2(1000)
var.values.Phi = compute.varUPhi(exp.values.Phi,1000)

```

For $n = 3, \dots, 20$ the results have been:

```
exp.values.Phi [3:20]
```

```

## [1]      1.00000      13.60000      73.42857      259.09524      711.54545
## [6]    1654.34965    3414.67413    6444.77869   11343.93737   18880.70867
## [11]  30015.50122  45923.39304  68017.17207  97970.57161 137741.67942
## [16] 189596.50276 256132.67433 340303.28664

```

```
var.values.Phi [3:20]
```

```

## [1]      0.00000      0.64000      4.77551      19.58277      58.97521
## [6]    146.23135    316.77865    621.09863   1127.72999   1926.35278
## [11]  3130.94150  4882.97724  7354.71170 10752.47676 15320.03476
## [16] 21341.96545 29147.08600 39111.90118

```

The rest of the values are available in the files “C2-EU(Phi2)1000.txt” and “C2-varU(Phi)1000.txt”.

We have estimated the main order in the expansion of $\sigma_U^2(\Phi_n)$ as a function of n , by performing the minimum squares linear regression of $\ln(\sigma_U^2(\Phi_n))$ as a function of $\ln(n)$ for $n = 900, \dots, 1000$:

```
summary(lm(log(var.values.Phi[900:1000])~log(900:1000)))

##
## Call:
## lm(formula = log(var.values.Phi[900:1000]) ~ log(900:1000))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.330e-05 -1.131e-05  4.161e-06  1.340e-05  1.671e-05
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.8743868  0.0003353  -11555  <2e-16 ***
## log(900:1000)  5.0657352  0.0000489  103586  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.509e-05 on 99 degrees of freedom
## Multiple R-squared:      1, Adjusted R-squared:      1
## F-statistic: 1.073e+10 on 1 and 99 DF, p-value: < 2.2e-16
```

The commands below produce Fig. 2.11 that displays $\ln(\sigma_U^2(\Phi_n))$ as a function of $\ln(n)$, together with the corresponding regression line.

```
plot(log(1:1000),log(var.values.Phi),
      xlab="log of the number of leaves",
      ylab="log of the variance")
reg.phi=lm(log(var.values.Phi[500:1000])~log(500:1000))
abline(reg.phi,col="blue",lwd=2)
```

Computing the variance of Φ_n from the cophenetic indices

To double-check our recurrence, we have computed the values of $\sigma_U^2(\Phi_n)$, for $n = 3, \dots, 8$, from the cophenetic indices of all trees in the corresponding \mathcal{BT}_n . To do this, we have carried out a similar process as in §A.4.2, replacing the arithmetic mean by the (true) variance:

```

var.n = function (vec)
  return(var(vec)*(length(vec)-1)/length(vec))
trees = list()
all.cophen.index = list()
real.var.Phi = c()
for(n in 3:8){
  trees[[n]] = read.tree(file = paste("bintrees-n",n,".txt",
                                     sep = ""))
  all.cophen.index[[n]] = sapply(trees[[n]], cophen.index)
  real.var.Phi[n] = var.n(all.cophen.index[[n]])
  print(paste("var(Phi",n,") = ",real.var.Phi[n],sep = ""))
}
real.var.Phi

## [1] 0.00000 0.64000 4.77551 19.58277 58.97521
## [5] 58.97521 146.23135

```

The results agree with the figures given by our recurrence.

A.4.4 Computation of $\sigma_U^2(S_n)$

Computing the variance of S_n using our formula

We can compute $\sigma_U^2(S_n)$ using the recurrence for $E_U(S_n^2)$, the exact formula for $E_U(S_n)$, and the identity

$$\sigma_U^2(S_n) = E_U(S_n^2) - E_U(S_n)^2.$$

The following functions are needed to compute this variance, in addition to those in Section A.3.

```

EUS = function(n){
  return(as.numeric(n*(big.double.factorial(2*n-2)/
                                     big.double.factorial(2*n-3)-1)))
}

term.S = function(n){
  return((5*n*2^(n-2)*big.factorial(n))/(
         big.double.factorial(2*n-3))-n*(5*n-2))
}

compute.EUS2 = function(n.max=500){
  terms = lapply(2:n.max,term.S)
  terms = c(0,terms)
}

```

```

exp.values = list(0)
for(n in 2:n.max){
  sums = 0
  if(n>2){
    for(k in 2:(n-1)){
      sums = sums + Cknk(k,n)*exp.values[[k]]
    }
    sums = 2*sums
  }
  sums = sums + terms[[n]]
  exp.values[[n]] = sums
  print(n)
}
exp.values = sapply(exp.values, as.numeric)
write.table(exp.values,file = paste("C2-EU(S2)",n.max,".txt",
                                   sep = ""),row.names = F,col.names = F)
return(exp.values)
}

compute.varUS = function(exp.values, n.max){
  var.form = function(i)return(exp.values[i]-EUS(i)^2)
  var.values = sapply(1:n.max,var.form)
  write.table(var.values,file = paste("C2-varU(S)",n.max,".txt",
                                     sep = ""),row.names = F,col.names = F)
  return(var.values)
}

```

We have computed these variances up to $n = 1000$ with the following commands:

```

exp.values.S = compute.EUS2(1000)
var.values.S = compute.varUS(exp.values.S,1000)

```

For $n = 3, \dots, 20$ the results have been:

```
exp.values.S[3:20]
```

```

## [1] 25.0000 77.6000 177.2857 340.0952 582.4545
## [6] 921.0816 1372.9245 1955.1189 2684.9571 3579.8650
## [11] 4657.3837 5935.1556 7430.9128 9162.4673 11147.7035
## [16] 3404.5711 15951.0795 18805.2931

```

```
var.values.S[3:20]
```

```
## [1] 0.0000000 0.1600000 0.7755102 2.2358277 4.9990817
## [6] 9.5765183 16.5219346 26.4241938 39.9016992 57.5981796
## [11] 80.1793886 108.3304640 142.7537743 184.1671371 233.3023247
## [16] 290.9037954 357.7276063 434.5404734
```

The rest of the values are available in the files “C2-EU(S2)1000.txt” and “C2-varU(S)1000.txt”.

We have estimated the main order in the expansion of $\sigma_U^2(S_n)$ as a function of n , by performing the minimum squares linear regression of $\ln(\sigma_U^2(S_n))$ as a function of $\ln(n)$ for $n = 900, \dots, 1000$,

```
summary(lm(log(var.values.S[900:1000])~log(900:1000)))
```

```
##
## Call:
## lm(formula = log(var.values.S[900:1000]) ~ log(900:1000))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.907e-05 -1.327e-05  4.879e-06  1.573e-05  1.961e-05
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.347e+00  3.935e-04  -5965  <2e-16 ***
## log(900:1000)  3.079e+00  5.739e-05  53646  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.771e-05 on 99 degrees of freedom
## Multiple R-squared:  1, Adjusted R-squared:  1
## F-statistic: 2.878e+09 on 1 and 99 DF, p-value: < 2.2e-16
```

The following code produces Fig. 2.12.

```
plot(log(1:1000),log(var.values.S),
      xlab="log of the number of leaves",
      ylab="log of the variance")
reg.S=lm(log(var.values.S[500:1000])~log(500:1000))
abline(reg.S,col="red",lwd=2)
sackin.approx = ((10-3*pi)/3)*(1:1000)^3
lines(log(1:1000),log(sackin.approx),col="cyan",lty=2,lwd=2)
```


Computing the variance of S from the Sackin indices

To double-check the recurrence, we have computed the values of $\sigma_U^2(S_n)$, for $n = 3, \dots, 8$, from the Sackin indices of all trees in the corresponding \mathcal{BT}_n . We have proceeded as in the previous subsection, using the function `sackin.index` in the package *CollessLike* (see §A.6.1) to compute the Sackin indices:

```
var.n = function (vec)
  return(var(vec)*(length(vec)-1)/length(vec))
trees = list()
all.sackin.index = list()
real.var.sackin = c()
for(n in 3:8){
  trees[[n]] = read.tree(file = paste("bintrees-n",n,".txt",
                                     sep = ""))
  all.sackin.index[[n]] = sapply(trees[[n]],sackin.index)
  real.var.sackin[n] = var.n(all.sackin.index[[n]])
  print(paste("var(S_",n,") = ",real.var.sackin[n],sep = ""))
}
real.var.sackin
```

```
## [1] 0.0000000 0.1600000 0.7755102 2.2358277
## [5] 4.9990817 9.5765183
```

So, the results agree again with the figures given by our recurrence.

A.4.5 Computation of $Cov_U(S_n, \Phi_n)$

Computing the covariance of S_n and Φ_n using our formula

We can compute $Cov_U(S_n, \Phi_n)$ using the recurrence for $E_U(S_n \Phi_n)$, the exact formulas for $E_U(S_n)$ and $E_U(\Phi_n)$, and the identity

$$Cov_U(S_n, \Phi_n) = E_U(S_n \Phi_n) - E_U(S_n)E_U(\Phi_n).$$

The following functions are needed to compute this covariance, in addition to EUPhi from Section A.4.2, EUS from Section A.4.4, and the functions in Section A.3.

```
term.cov = function(n){
  return((((13*n^2-9*n-2)*2^(n-5)*big.factorial(n))/(
          big.double.factorial(2*n-3))-(n*(n-1)/2)*(5*n-2))
}

compute.EUcov = function(n.max=500){
```

```

terms = lapply(2:n.max,term.cov)
terms = c(0,terms)
exp.values = list(0)
for(n in 2:n.max){
  sums = 0
  if(n>2){
    for(k in 2:(n-1)){
      sums = sums + Cknk(k,n)*exp.values[[k]]
    }
    sums = 2*sums
  }
  sums = sums + terms[[n]]
  exp.values[[n]] = sums
  print(n)
}
exp.values = sapply(exp.values, as.numeric)
write.table(exp.values,file=paste("C2-EU(SxPhi)",n.max,".txt",
                                sep = ""),row.names = F,col.names = F)
return(exp.values)
}

compute.cov = function(exp.values,n.max = 500){
  cov.form = function(i)return(exp.values[i]-EUS(i)*EUPhi(i))
  cov.values = sapply(1:n.max, cov.form)
  write.table(cov.values,file=paste("C2-covU(SPhi)",n.max,".txt",
                                sep = ""),row.names = F,col.names = F)
  return(cov.values)
}

```

We have computed these covariances and correlations up to $n = 1000$ with the following commands:

```

exp.values.cov = compute.EUcov(1000)
cov.values = compute.cov(exp.values.cov,1000)

```

For $n = 3, \dots, 20$ the results have been:

```

exp.values.cov[3:20]

```

```

## [1]      5.0000     32.0000    112.0000    291.0476    630.9091
## [6]  1209.5478   2121.4881   3478.1045   5407.8581   8056.4954
## [11] 11587.2180  16180.8299  22035.8667  29368.7106  38413.6931
## [16] 49423.1877  62667.6942  78435.9159

```

```
cov.values[3:20]
```

```
## [1] 0.000000 0.320000 1.918367 6.580499 17.044077
## [6] 37.089909 71.611701 126.671836 209.547294 328.768314
## [11] 494.151534 716.828809 1009.272571 1385.318372 1860.185104
## [16] 2450.493255 3174.281512 4051.021944
```

The rest of the values are available in the files “C2-EU(SxPhi)1000.txt” and “C2-covU(SPhi)1000.txt”.

We have estimated the main order in the expansion of $Cov_U(S_n, \Phi_n)$ as a function of n , by performing the minimum squares linear regression of $\ln(Cov_U(S_n, \Phi_n))$ as a function of $\ln(n)$ for $n = 900, \dots, 1000$,

```
summary(lm(log(cov.values[900:1000])~log(900:1000)))
```

```
##
## Call:
## lm(formula = log(cov.values[900:1000]) ~ log(900:1000))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.553e-05 -1.207e-05  4.439e-06  1.430e-05  1.783e-05
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.1333995  0.0003579   -8756  <2e-16 ***
## log(900:1000)  4.0709153  0.0000522  77993  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.611e-05 on 99 degrees of freedom
## Multiple R-squared: 1, Adjusted R-squared: 1
## F-statistic: 6.083e+09 on 1 and 99 DF, p-value: < 2.2e-16
```

The code below produces Fig. 2.13, which displays $\ln(Cov_U(S_n, \Phi_n))$ as a function of $\ln(n)$ together with the corresponding regression line.

```
plot(log(1:1000),log(cov.values),
      xlab="log of the number of leaves",
      ylab="log of the variance")
reg.cov=lm(log(cov.values[500:1000])~log(500:1000))
abline(reg.cov,col="violet",lwd=2)
```

Computing the covariance of S and Φ from the values of the indices

To double-check the recurrence, we have computed the values of $Cov_U(S_n, \Phi_n)$, for $n = 3, \dots, 8$, from the Sackin and the total cophenetic indices of all trees in the corresponding \mathcal{BT}_n :

```
covariancesU = function(n){
  len = length(all.sackin.index[[n]])
  value = cov(all.sackin.index[[n]], all.cophen.index[[n]] *
              (len-1)/len)
  return(value)
}
real.cov.values = sapply(3:7, covariancesU)
real.cov.values
```

```
## [1] 0.000000 0.320000 1.918367 6.580499
## [5] 17.044077 37.089909
```

The results agree again with the figures given by our recurrence.

Computing the correlation of S and Φ from the values of the indices

Since we know how to compute recurrently $Cov_U(S_n, \Phi_n)$, $\sigma_U^2(S_n)$ and $\sigma_U^2(\Phi_n)$, we can compute Pearson's correlation ρ of S_n and Φ_n under the uniform model for any desired $n \geq 4$, by means of the identity

$$\rho_U(S_n, \Phi_n) = \frac{Cov_U(S_n, \Phi_n)}{\sigma_U(S_n) \cdot \sigma_U(\Phi_n)}.$$

```
pearson.cor = function(n.max){
  return(cov.values[4:n.max]/sqrt(var.values.S[4:n.max] *
                                   var.values.Phi[4:n.max]))
}
```

For $n = 4, \dots, 20$ the results are:

```
pearson.cor(20)
```

```
## [1] 1.000000 0.9968461 0.9944951 0.9926443 0.9911325
## [6] 0.9898641 0.9887783 0.9878336 0.9870012 0.9862597
## [11] 0.9855934 0.9849901 0.9844400 0.9839358 0.9834711
## [16] 0.9830410 0.9826413
```

A.4.6 Computation of the estimated probability of a tie

We have estimated the probability that a pair of trees $T_1, T_2 \in \mathcal{BT}_n$ have $I(T_1) = I(T_2)$, for $I = C, S, \Phi$. To do that, for every $n = 3, \dots, 50$ we have chosen uniformly a set of N random pairs of trees in \mathcal{BT}_n (for $n = 3, \dots, 7$, we took $N = |\mathcal{BT}_n|$ and, for $n \geq 8$, we took $N = 3000$), and computed, for each $I = C, S, \Phi$,

$$\hat{p}_n(I) = \frac{\text{number of pairs } (T_1, T_2) \text{ with } n \text{ leaves such that } I(T_1) = I(T_2)}{N}.$$

The balance indices have been computed with the function `balance.indices` from the package *CollessLike* (see §A.6.1), with the parameter `binary.Colless` set to `TRUE`. The following functions compute these probabilities \hat{p}_n :

```
are.tie = function(xx,yy) return(xx==yy)

exact.ties = function(){
  trees = list()
  all.indices = list()
  num.ties = c(0,0,0)
  prob.ties = list()
  for(n in 3:7){
    trees[[n]] = read.tree(file = paste("bintrees-n",n,".txt",
                                         sep=""))

    total.trees = length(trees[[n]])
    total.pairs = total.trees*(total.trees-1)/2
    all.indices[[n]] = matrix(sapply(trees[[n]],
                                     balance.indices2),ncol=3,byrow=T)
    num.ties[1]=sum(outer(all.indices[[n]][,1],
                          all.indices[[n]][,1],are.tie))
    num.ties[2]=sum(outer(all.indices[[n]][,2],
                          all.indices[[n]][,2],are.tie))
    num.ties[3]=sum(outer(all.indices[[n]][,3],
                          all.indices[[n]][,3],are.tie))

    num.ties = (num.ties-total.trees)/2
    prob.ties[[n]] = num.ties/total.pairs
    print(paste("Ties for n =",n," : ",
                paste(c("p_C=", "p_S=", "p_Phi"),
                      round(prob.ties[[n]],4),collapse=" ", sep="")))
  }
  return(prob.ties)
}

sim.ties.n = function(n,num.pairs.sim=3000){
  num.ties = c(0,0,0)
```

```

for(i in 1:num.pairs.sim){
  t1 = rtree(n,rooted=TRUE)
  continue = TRUE
  while(continue){
    t2 = rtree(n,rooted=TRUE)
    continue = all.equal(t1,t2,use.length=FALSE,
                        use.tip.label=FALSE)
  }
  t1.indices = balance.indices2(t1)
  t2.indices = balance.indices2(t2)
  num.ties = num.ties + (t1.indices==t2.indices)
}
print(paste("n =",n))
print(paste("Ties :",num.ties))
prob.ties = num.ties/num.pairs.sim
print(paste("Prob :",round(prob.ties,4)))
return(prob.ties)
}

```

Now, with the following commands we compute the probabilities for each $n = 3, \dots, 50$:

```

ties.1 = exact.ties()
ties.2 = lapply(8:50, sim.ties.n,num.pairs.sim=3000)
ties = matrix(c(unlist(ties.1),unlist(ties.2)),ncol=3,byrow=TRUE)
colnames(ties)=c("Colless","Sackin","Cophenetic")
rownames(ties)=3:50

```

For $n = 4, \dots, 20$ the results are:

```
ties[1:18,]
```

	Colles	Sackin	Cophenetic
3	1.0000000	1.0000000	1.0000000
4	0.6571429	0.6571429	0.6571429
5	0.4230769	0.4230769	0.4230769
6	0.2372881	0.2463680	0.2372881
7	0.1459233	0.1716375	0.1312296
8	0.1133333	0.1363333	0.0710000
9	0.0763333	0.1260000	0.0550000
10	0.0583333	0.1090000	0.0276667
11	0.0593333	0.0920000	0.0243333
12	0.0393333	0.0720000	0.0150000

	Colles	Sackin	Cophenetic
13	0.0343333	0.0623333	0.0176667
14	0.0363333	0.0636667	0.0123333
15	0.0370000	0.0546667	0.0100000
16	0.0273333	0.0530000	0.0053333
17	0.0200000	0.0390000	0.0046667
18	0.0190000	0.0376667	0.0053333
19	0.0200000	0.0346667	0.0043333
20	0.0120000	0.0370000	0.0056667

The full table is available at “C2-table-ties.txt”. Fig. 2.14, which summarizes the results, has been produced with the following commands:

```
plot(log(3:50),log(ties[,3]),type="l",
     xlab="log of the number of leaves",
     ylab="log of the probability of tie", col="blue")
lines(log(3:50),log(ties[,1]),col="red")
lines(log(3:50),log(ties[,2]),col="green")
legend("bottomleft", legend=c("Sackin","Colless","Cophenetic"),
      col=c("green","red", "blue"),lty=1, cex=0.8)
```

A.4.7 Testing Φ_n on TreeBASE

In this subsection we explain how we have performed the test reported in §2.8.2. We have loaded the data table containing the Newick representations of all trees in TreeBASE, which we had previously downloaded using the function `search_treebase()` of the R package *treebase* and saved in the *List of Trees* folder of the PhD Thesis GitHub repository as a text file and as an R object. So, we have two ways to import these data:

```
# Option 1
tb.ape = read.tree(file = "./tb-newicks.txt")
# Option 2
load("./treeBASE-database.RData")
```

We have considered only those numbers n of leaves for which the TreeBASE contains at least 20 binary phylogenetic trees with n leaves, and for each such n we have computed the mean of the total cophenetic indices of the corresponding binary trees:

```
bin.tb.ape=tb.ape[sapply(tb.ape,is.rooted)]
bin.tb.ape=bin.tb.ape[sapply(bin.tb.ape,is.binary)]
```

```

bin.tb.n = sapply(bin.tb.ape,Ntip)
leaves=as.numeric(names(which(table(bin.tb.n)>20)))
bin.tb.mean = c()
indices.tb = list()
for(k in leaves){
  trees = bin.tb.ape[bin.tb.n==k]
  indices.tb[[k]] = sapply(trees, copen.index)
  value = mean( indices.tb[[k]] )
  bin.tb.mean = rbind(bin.tb.mean,c(k,value))
}

```

The results of these computations are available in the file “C2-table-tb-means.txt”.

The following code computes $E_Y(\Phi_n)$ and $E_U(\Phi_n)$ for $n = 3, \dots, 140$, using the functions EYPhi and EUPhi from Section A.4.1 and Section A.4.2, respectively:

```

range.plot = 3:140
eyphi.values = sapply(range.plot, EYPhi)
euphi.values = sapply(range.plot, EUPhi)

```

Using the computations of the variance of Φ_n under the uniform model (from Section A.4.3) and the exact formula for $\sigma_Y^2(\Phi_n)$ (Formula 2.2) we can obtain the reference intervals for Φ_n :

```

harmonic2 = function(n){return(sum(1/((1:n)^2)))}
varYPhi = function(n){
  return((n^4-10*n^3+131*n^2-2*n)/12-4*n^2*harmonic2(n)-
    6*n*harmonic(n))
}
varYPhi.values = sapply(range.plot,varYPhi)
intY = cbind(range.plot,log(eyphi.values-sqrt(varYPhi.values)),
  log(eyphi.values+1*sqrt(varYPhi.values)))
intU = cbind(range.plot,log(euphi.values-
  sqrt(var.values.Phi[range.plot])),
  log(euphi.values+1*sqrt(var.values.Phi[range.plot])))

draw.intervals =
  function(range.plot,int.yule,int.uniform,delta=0){
    epsilon = 0.3
    for(i in range.plot){
      lines(c(i ,i ),int.uniform[i-2,2:3],col="cyan")
      lines(c(i-epsilon,i+epsilon),rep(int.uniform[i-2,2],2),

```



```

        col="cyan")
lines(c(i-epsilon,i+epsilon),rep(int.uniform[i-2,3],2),
      col="cyan")
lines(c(i ,i )-delta,int.yule[i-2,2:3],col="violet")
lines(c(i-epsilon,i+epsilon)-delta,rep(int.yule[i-2,2],2),
      col="violet")
lines(c(i-epsilon,i+epsilon)-delta,rep(int.yule[i-2,3],2),
      col="violet")
}
}

```

Finally, the following code produces Fig. 2.15:

```

plot(NULL,NULL,col="blue",xlab="number of leaves",
     ylab="log of means",xlim=c(3,130),ylim=c(1,11.5),
     type="l",lwd=2)
draw.intervals(range.plot,intY,intU,delta=0.3)
lines(range.plot,log(eyphi.values),col="red",lwd=2)
lines(range.plot,log(euphi.values),col="blue",lwd=2)
lines(bin.tb.mean[,1],log(bin.tb.mean[,2]),type="l",lwd=2)
legend("bottomright", legend=c(expression(E[U]*(Phi[n])),
  "Uniform intervals","Treebase",expression(E[Y]*(Phi[n])),
  "Yule intervals"),col=c("blue","cyan","black","red",
  "violet"),lty=1,cex=0.8)

```

A.5 Scripts from Chapter 3

A.5.1 Computation of $E(D_n^2)$

The formulas in Theorem 3.31 and Theorem 3.38, corresponding to the expected value of D_n^2 under the Yule and uniform models, respectively, can be computed with the following functions:

```

harmonic=function(n){return(sum(1/(1:n)))}
EYD2n = function(n){
  return((2*n/(n-1))*(3*n^2-10*n-1+8*(n+1)*harmonic(n)-
    4*(n+1)*harmonic(n)^2))
}

EUD2n = function(n){
  return((4*n^3+18*n^2-10*n)/3+as.numeric(-as.bigq((n*(n+3))/2)*
    (big.double.factorial(2*n-2)/

```

```

        big.double.factorial(2*n-3))
    -as.bigq((n*(n+7))/4)*((big.double.factorial(2*n-2)/
        big.double.factorial(2*n-3))^2)))
}

```

For $n = 3, \dots, 20$ the results are:

```

# Yule model
sapply(3:20, EYD2n)

## [1] 2.666667 9.407407 21.183333 38.712000 62.556190
## [6] 93.172128 130.938761 176.176855 229.162086 290.134368
## [11] 359.304706 436.860362 522.968823 617.780914 721.433274
## [16] 834.050354 955.746046 1086.625029

```

```

# uniform model
sapply(3:20, EUD2n)

## [1] 2.666667 10.560000 26.236735 52.302343 91.408632
## [6] 146.247151 219.543237 314.051159 432.550230 577.841679
## [11] 752.746096 960.101325 1202.760711 1483.591615 1805.474154
## [16] 2171.300112 2583.971999 3046.402233

```

To double-check the formulas, we have computed the values of $d_{\varphi,2}(T, T')^2$, for $n = 3, \dots, 7$, from the cophenetic distance between all pairs of trees in the corresponding \mathcal{BT}_n . The cophenetic vectors of the phylogenetic trees have been computed with the function `cophen.vect` in the R package *CollessLike*. In the Yule case, we have used the function `yule.prob` explained in §A.4.1 to compute the probabilities. Finally, the expected values and the variances of the square of the cophenetic distance for each n have been computed in the usual way:

```

real.exp.var = function(n.max=7){
  means = matrix(0,ncol=2,nrow=8)
  colnames(means) = c("uniform", "Yule")
  vars = matrix(0,ncol=2,nrow=8)
  colnames(vars) = c("uniform", "Yule")
  for(n in 3:n.max){
    trees = read.tree(file = paste("bintrees-n",n, ".txt", sep=""))
    total.trees = length(trees)
    probs = sapply(trees, yule.prob)
    pairs.probs = c()
    all.vectors = lapply(trees, cophen.vect)
  }
}

```

```

values = c()
for(i in 1:(total.trees)){
  for(j in (1):total.trees){
    values = c(values, sum((all.vectors[[i]] -
                          all.vectors[[j]])^2))
    pairs.probs = c(pairs.probs, probs[i]*probs[j])
  }
}
means[n,1]=mean(values)
means[n,2]=sum(pairs.probs*values)
vars[n,1]=mean(values^2)-means[n,1]^2
vars[n,2]=sum(pairs.probs*values^2)-means[n,2]^2
print(paste("n =",n))
print(means[n,])
print(vars[n,])
}
results = cbind(3:7, means[3:7,2], vars[3:7,2], means[3:7,1],
               vars[3:7,1])
colnames(results) = c("n", "EY(D2n)", "varY(D2n)", "EU(D2n)",
                    "varU(D2n)")

return(results)
}
results=real.exp.var()

```

We have obtained the following results. They agree with the figures given by our formulas.

n	3	4	5	6	7
$E_Y(D_n^2)$	2.66667	9.40741	21.18333	38.71200	62.55619
$E_U(D_n^2)$	2.66667	10.56000	26.23673	52.30234	91.40863

In the previous chunk of code, we have also computed the exact values for the variance:

n	3	4	5	6	7
$\sigma_Y^2(D_n^2)$	3.55556	29.13032	117.63306	339.28881	797.15834
$\sigma_U^2(D_n^2)$	3.55556	34.08640	159.50314	539.50829	1502.72330

A.5.2 Computation of $\sigma^2(D_n^2)$

In order to estimate the asymptotic order of $E(D_n^4)$ and $\sigma^2(D_n^2)$, both for the Yule and the uniform models, and for every $n = 3, \dots, 100$, we have randomly

generated $N = 10000$ pairs of binary trees $(T, T') \in \mathcal{T}_n \times \mathcal{T}_n$ using the R package *apTreeshape*, and converted them into *phylo* objects for the R package *ape*:

```
require(apTreeshape)
generate.trees = function(n,model,repetitions=10000){
  if(model=="yule")trees=rtreeshape(repetitions*2,n,model="yule")
  if(model=="uniform")trees = rtreeshape(repe*2,n,model="pda")
  trees = lapply(trees,as.phylo)
  return(trees)
}
```

We have computed the value of $d_{\varphi,2}(T, T')^2$ and $d_{\varphi,2}(T, T')^4$ for each such pair (T, T') with the following function,

```
compute.values.pairs = function(n,model,repetitions=10000){
  euc.dist2 = function(pair){
    m = length(pair)/2
    value = sum((pair[1:m] - pair[(m+1):(2*m)])^2)
    return(value)
  }
  trees=generate.trees(n,model,repetitions)
  vectors = lapply(trees, cophen.vector)
  vectors = matrix(unlist(vectors),byrow=T,nrow=repetitions)
  result = apply(vectors,1,euc.dist2)
  result = c(mean(result),mean(result^2))
  return(c(result,result[2]-result[1]^2))
}
```

We have computed the arithmetic means $\overline{D_n^2}$ and $\overline{D_n^4}$ of these N values, and, finally, the variance of the values $d_{\varphi,2}(T, T')^2$ using the identity

$$\widehat{\sigma}^2(D_n^2) = \overline{D_n^4} - \overline{D_n^2}^2.$$

This value is an estimation of $\sigma^2(D_n^2)$ under the corresponding model. Next commands show how we have estimated these variances:

```
varD2n = c()
for(k in 3:100){
  values.yule = compute.values.pairs(k,"yule",10000)
  values.uniform = compute.values.pairs(k,"uniform",10000)
  varD2n = rbind(varD2n,c(k,values.yule[2:3],values.uniform[2:3]))
}
colnames(varD2n) = c("n","Yule_EDn4","Yule_varDn2",
                    "uniform_EDn4","uniform_varDn2")
```

Since these computations take a long time to finish, we have parallelized them with the R package *parallel*. For $n = 3, \dots, 20$ the results have been:

```
varD2n[1:13,]
```

n	$E_Y(D_n^4)$	$\overline{\sigma_Y^2}(D_n^2)$	$E_U(D_n^4)$	$\overline{\sigma_U^2}(D_n^2)$
3	10.6032	3.5765	10.6816	3.5506
4	116.7946	29.1438	143.3660	33.8602
5	559.7510	115.1738	852.0843	152.8521
6	1837.4050	334.2533	3299.1028	531.5538
7	4648.0487	761.6796	9872.4776	1498.6903
8	10330.2279	1621.1155	25175.2292	3679.6130
9	19785.9321	2975.1427	57131.8612	8153.9221
10	35948.3028	4931.6335	115022.2093	16236.0131
11	60700.8108	8278.0529	216900.0930	31384.2511
12	95272.0040	12040.3310	391444.9727	56683.5990
13	148901.4119	18688.7925	669652.7456	98755.8404
14	218716.4557	27015.5924	1088410.0273	161050.1757
15	312615.8741	37466.0192	1698106.1014	254385.1408

The rest of the values are available in the file “C3-table-expDn4-varDn2.txt”.

We have computed the slope α of the regression line of $\log(\overline{\sigma^2}(D_n^2))$ as a function of $\log(n)$ using the values for $n = 50, \dots, 100$ with the following commands.

```
#var_Y(D2n)
```

```
reg.yule = lm(log(varD2n[48:98,3])~log(50:100))
summary(reg.yule)
```

```
##
## Call:
## lm(formula = log(varD2n[48:98, 3]) ~ log(50:100))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.113573 -0.034866  0.004031  0.033017  0.092599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.30850    0.13540  -2.278   0.0271 *
## log(50:100)  4.15220    0.03147 131.931 <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04537 on 49 degrees of freedom
## Multiple R-squared:  0.9972, Adjusted R-squared:  0.9971
## F-statistic: 1.741e+04 on 1 and 49 DF,  p-value: < 2.2e-16
```

```
#var_U(D2n)
reg.uniform = lm(log(varD2n[48:98,5])~log(50:100))
summary(reg.uniform)
```

```
##
## Call:
## lm(formula = log(varD2n[48:98, 5]) ~ log(50:100))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09680 -0.02160  0.00325  0.02406  0.09204
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.55620    0.10217  -44.59  <2e-16 ***
## log(50:100)  6.38830    0.02375  269.01  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03423 on 49 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9993
## F-statistic: 7.237e+04 on 1 and 49 DF,  p-value: < 2.2e-16
```

Finally, the following commands produce Fig. 3.26:

```
plot(log(3:100),log(varD2n[,5]),
      xlab="log of the number of leaves",
      ylab=expression("log of the variance of "*D[n]^2))
abline(reg.uniform,col="blue",lwd=2)
points(log(3:100),log(varD2n[,3]))
abline(reg.yule,col="red",lwd=2)
legend("topleft",legend=c("Uniform model","Yule model"),
      col=c("blue","red"),lty=1,cex=0.8)
```

A.6 Scripts from Chapter 4

A.6.1 The R package *CollessLike*

We have written the R package *CollessLike*, available on the CRAN [74] and on the GitHub [96], that computes the Colless-like indices and their normalized version, as well as several other balance indices, and simulates the distribution of these indices under the α - γ -model. We describe its contents in this subsection.

The following function computes the Sackin index of a tree. The value can be normalized with `norm=TRUE`.

```
sackin.index <- function(tree,norm=FALSE){
  if(class(tree)=="character")
    tree=read.tree(text = tree)
  if (class(tree)=="phylo")
    tree=graph.edgelist(tree$edge, directed=TRUE)
  if(class(tree)!="igraph")
    stop("Not an igraph object. Please introduce a newick
         string, an ape tree or an igraph tree.")
  root.node = which(degree(tree,mode="in")==0)
  deg.out = degree(tree,mode="out")
  if(deg.out[root.node]==1){ #exists a root-edge
    tree = delete.vertices(tree,root.node)
    deg.out = degree(tree,mode="out")
    root.node = which(degree(tree,mode="in")==0)
  }
  leaves = which(deg.out==0)
  root.list = get.shortest.paths(tree,root.node)$vpath
  depths = unlist(lapply(root.list,function(xx){length(xx)-1}))
  SACKIN=sum(depths[leaves])
  if(norm){
    N = length(leaves)
    max.s = N*(N-1)/2 + N-1
    SACKIN = (SACKIN-N)/(max.s-N)
  }
  return(SACKIN)
}
```

The following function computes the total cophenetic index of a tree. The value can be normalized with `norm=TRUE`.

```
cophen.index <- function(tree,norm=FALSE){
  if(class(tree)=="character")
    tree=read.tree(text = tree)
```

```

if (class(tree)=="phylo")
  tree=graph.edgelist(tree$edge, directed=TRUE)
if(class(tree)!="igraph")
  stop("Not an igraph object. Please introduce a newick
       string, an ape tree or an igraph tree.")
root.node = which(degree(tree,mode="in")==0)
deg.out = degree(tree,mode="out")
if(deg.out[root.node]==1){ #exists a root-edge
  tree = delete.vertices(tree,root.node)
  deg.out = degree(tree,mode="out")
  root.node = which(degree(tree,mode="in")==0)
}
leaves = which(deg.out==0)
root.list = get.shortest.paths(tree,root.node)$vpath
N = length(leaves)
COPHEN = 0
for(i in 1:(N-1))
  for(j in (i+1):N){
    aux = length(intersect(root.list[[leaves[i]]],
                          root.list[[leaves[j]]]))-1
    COPHEN = COPHEN + aux
  }
if(norm){
  max.c = N*(N-1)*(N-2)/6
  COPHEN = COPHEN/max.c
}
return(COPHEN)
}

```

The following function computes the Colless-like index of a tree. The value can be normalized with `norm=TRUE`. By default, the f -size is $f(n) = \ln(n + e)$, if `f.size="exp"` then $f(n) = e^n$. It can also be a user-defined function but in this case, the index cannot be normalized. On the other hand, the default value of the dissimilarity is MDM (mean deviation from the median). Other values can be set as `diss="sd"` for the sample standard deviation or `diss="var"` for the sample variance. It can also be a user-defined function but, again, in this case the value cannot be normalized.

```

colless.like.index <-function(tree,f.size="ln",diss="MDM",
norm=FALSE){
  if(class(tree)=="character")
    tree=read.tree(text = tree)
  if (class(tree)=="phylo")
    tree=graph.edgelist(tree$edge, directed=TRUE)

```



```

if(class(tree)!="igraph")
  stop("Not an igraph object. Please introduce a newick
        string, an ape tree or an igraph tree.")
root.node = which(degree(tree,mode="in")==0)
deg.out = degree(tree,mode="out")
case.norm = 0
if(class(f.size)=="character"){
  if(f.size=="ln"){
    f.size = function(nn) return(log(nn+exp(1)))
    case.norm = 1
  }
  else if((f.size=="exp")||(f.size=="e")){
    f.size = function(nn) return( exp(nn) )
    case.norm = 4
  }
  else stop("The f-size introduced is not correct.")
}
if(class(diss)=="character"){
  if((diss=="MDM")||(diss=="mdm")){
    diss = function(xx)
      return(sum(abs(xx-median(xx)))/length(xx))
    case.norm = case.norm*1
  }
  else if(diss=="var"){
    diss = function(xx) return(sum((xx-mean(xx))^2)/
      (length(xx)-1))

    case.norm = case.norm*2
  }
  else if(diss=="sd"){
    diss = function(xx) return(sqrt(sum((xx-mean(xx))^2)/
      (length(xx)-1)))

    case.norm = case.norm*3
  }
  else stop("The dissimilarity introduced is not correct.")
}
int.nodes = (1:length(V(tree)))[deg.out>0]
decendants = neighborhood(tree,1,int.nodes,mode = "out")
fun.nodes.deltas = function(nodes){
  aux = neighborhood(tree,length(deg.out)-1,nodes,
    mode = "out")[[1]]
  return(sum(f.size(deg.out[aux])))
}
fun.children = function(children){
  children = children[-1]
}

```

```

    result = unlist(lapply(children,fun.nodes.deltas))
    return(result)
}
deltas = lapply(decendents,fun.children)
Vdiss = lapply(deltas, diss)
COLLESS = sum(unlist(Vdiss))
if(norm){
  if(case.norm==0) warning("Indices can not be normalized")
  else{
    N = length(which(deg.out==0))
    # ln MDM
    if(case.norm==1)
      max.cl = (f.size(0) + f.size(2))*(N-1)*(N-2)/4
    # ln var
    if(case.norm==2)
      max.cl = (f.size(0) + f.size(2) )^2*(N-1)*(N-2)*
        (2*N-3)/12
    # ln sd
    if(case.norm==3)
      max.cl = (f.size(0) + f.size(2))*(N-1)*(N-2)/
        (2*sqrt(2))
    # e^n var
    if(case.norm==8) max.cl = (f.size(N-1)+N-2)^2/2
    if(N==4){
      # e^n MDM
      if(case.norm==4) max.cl = (f.size(2)+1)*3/2
      # e^n sd
      if(case.norm==12) max.cl =(f.size(2)+1)*3/sqrt(2)
    }
    else{
      # e^n MDM
      if(case.norm==4) max.cl = (f.size(N-1)+N-2)/2
      # e^n sd
      if(case.norm==12) max.cl = (f.size(N-1)+N-2)/sqrt(2)
    }
    COLLESS = COLLESS /max.cl
  }
}
return(COLLESS)
}

```

The next function assembles the three previous functions. So, it computes the three indices and returns a single array with the three values. The results can be normalized. If `binary.Colless=TRUE`, it computes the classical Colless

index for binary trees (previously checking that the input tree is binary).

```

balance.indices<-function(tree,norm=FALSE,binary.Colless=FALSE){
  if(class(tree)=="character")
    tree=read.tree(text = tree)
  if (class(tree)=="phylo")
    tree=graph.edgelist(tree$edge, directed=TRUE)
  if(class(tree)!="igraph")
    stop("Not an igraph object. Please introduce a newick
         string, an ape tree or an igraph tree.")
  root.node = which(degree(tree,mode="in")==0)
  deg.out = degree(tree,mode="out")
  # COLLESS.MDM.LN
  D.MDM = function(xx)
    return(sum(abs(xx-median(xx)))/length(xx))
  f.ln = function(n) return(log(n+exp(1)))
  int.nodes = (1:length(V(tree)))[deg.out>0]
  decendants = neighborhood(tree,1,int.nodes,mode = "out")
  fun.nodes.deltas = function(nodes){
    aux = neighborhood(tree,length(deg.out)-1,nodes,
                       mode = "out")[[1]]
    return(sum(f.ln(deg.out[aux])))
  }
  fun.children = function(children){
    children = children[-1]
    result = unlist(lapply(children,fun.nodes.deltas))
    return(result)
  }
  deltas = lapply(decendants,fun.children)
  Vdis = lapply(deltas, D.MDM)
  COLLESS = sum(unlist(Vdis))
  if(deg.out[root.node]==1){ #exists root-edge
    tree = delete.vertices(tree,root.node)
    deg.out = degree(tree,mode="out")
    root.node = which(degree(tree,mode="in")==0)
  }
  leaves = which(deg.out==0)
  root.list = get.shortest.paths(tree,root.node)$vpath
  # SACKIN #
  depths = unlist(lapply(root.list,
                        function(xx){length(xx)-1}))
  SACKIN=sum(depths[leaves])
  # COPHENETIC #
  N = length(leaves)

```

```

COPHEN = 0
for(i in 1:(N-1))
  for(j in (i+1):N){
    aux = length(intersect(root.list[[leaves[i]]],
                          root.list[[leaves[j]]]))-1
    COPHEN = COPHEN + aux
  }
result = c("Colles-Like"=COLLESS,"Sackin"=SACKIN,
          "Cophenetic"=COPHEN)
if(binary.Colless){
  if(sum(!(deg.out %in% c("0","2")))==0)
    result[1] = result[1]/((log(0+exp(1))+log(2+exp(1)))/2)
  else warning("The tree introduced is not binary,
               Colless-like index for multifurcated trees is
               computed.")
}
else{
  if(norm){
    max.cl = ( log(0+exp(1)) + log(2+exp(1)) )*(N-1)*(N-2)/4
    max.s = N*(N-1)/2 + N-1
    max.c = N*(N-1)*(N-2)/6
    result[1] = result[1]/max.cl
    result[2] = (result[2]-N)/(max.s-N)
    result[3] = result[3]/max.c
  }
}
return(result)
}

```

The following function computes the total cophenetic vector of a tree.

```

cophen.vector <- function(tree,set.of.labels=NULL){
  if(class(tree)=="character")
    tree=read.tree(text = tree)
  if (class(tree)=="phylo"){
    if(is.null(set.of.labels))
      set.of.labels = tree$tip.label
    tree=graph.edgelist(tree$edge, directed=TRUE)
  }
  if(class(tree)!="igraph")
    stop("Not an igraph object. Please introduce a newick
         string, an ape tree or an igraph tree.")
  if(is.null(set.of.labels))
    stop("Please insert the set of labels or a phylo object")
}

```

```

root.node = which(degree(tree,mode="in")==0)
deg.out = degree(tree,mode="out")
if(deg.out[root.node]==1){ #exists a root-edge
  tree = delete.vertices(tree,root.node)
  deg.out = degree(tree,mode="out")
  root.node = which(degree(tree,mode="in")==0)
}
leaves = which(deg.out==0)
if(length(set.of.labels)!=length(leaves))
  stop("Please insert the correct set of labels or a phylo
      object")
root.list = get.shortest.paths(tree,root.node)$vpath
# COPHENETIC #
N = length(leaves)
COPHEN = c()
ordered.leaves=order(set.of.labels)
for(i in 1:(N-1)){
  leaf.i = ordered.leaves[i]
  COPHEN = c(COPHEN,length(root.list[[leaf.i]])-1)
  for(j in (i+1):N){
    leaf.j = ordered.leaves[j]
    aux = length(intersect(root.list[[leaves[leaf.i]]],
                          root.list[[leaves[leaf.j]]]))-1
    COPHEN = c(COPHEN,aux)
  }
}
COPHEN = c(COPHEN,length(root.list[[ordered.leaves[N]]])-1)
return(COPHEN)
}

```

Given α , γ and the number of leaves n , the following function generates a random phylogenetic tree with $n \geq 3$ leaves with the probability distribution defined by the α - γ -model.

```

a.g.model <-function(n,alpha,gamma){
  if(n<3)
    stop("n<3")
  else{
    if((alpha>1)|| (alpha<0)|| (gamma>1)|| (gamma<0))
      stop("alpha and gamma must been between 0 and 1")
    else{
      if(alpha<gamma)
        stop("alpha < gamma")
      else{

```

```

edge.matrix = matrix(c(1,2,2,3,2,4),byrow=T,ncol=2)
n.nodes = 4
n.edges = 3
for(n.leaves in 3:n){#Add new leaf
  #Assign probabilities
  probabilities = rep(0,n.nodes+n.edges)
  degrees = rep(0,n.nodes)
  degree.table = table(edge.matrix[,1])
  degrees[as.numeric(names(degree.table))]=degree.table

  leaves = which(degrees==0)
  leaf.edge = which(edge.matrix[,2]%in%leaves)

  probabilities[1:n.edges + n.nodes] = gamma
  probabilities[leaf.edge+n.nodes] = 1-alpha
  probabilities[which(degrees>1)] =
    (degrees[degrees>1]-1)*alpha-gamma
  probabilities = probabilities/(n.leaves-alpha)

  random = sample(c(1:n.nodes,1:n.edges+n.nodes),
    1,prob=probabilities)

  if(random<=n.nodes){#a node is selected
    edge.matrix = rbind(edge.matrix,c(random,
      n.nodes+1))

    n.nodes = n.nodes+1
    n.edges = n.edges+1
  }
  else{#an edge is selected
    random = random - n.nodes
    edge.matrix = rbind(edge.matrix,
      c(edge.matrix[random,1],n.nodes+1))
    edge.matrix = rbind(edge.matrix,
      c(n.nodes+1,edge.matrix[random,2]))
    edge.matrix = rbind(edge.matrix,
      c(n.nodes+1,n.nodes+2))
    edge.matrix = edge.matrix[-random,]
    n.nodes = n.nodes+2
    n.edges = n.edges+2
  }
}
tree = graph.edgelist(edge.matrix)
deg.out = degree(tree,mode="out")
root.node = which(degree(tree,mode="in")==0)

```

```

        if(deg.out[root.node]==1){ #Erase the root-edge
            tree = delete.vertices(tree,root.node)
        }
        return(tree)
    }
}
}
}
}

```

The following function generates a list of trees with the probability distribution defined by an α - γ -model and then it computes their Colless-like, Sackin and total cophenetic indices.

```

indices.simulation<-function(n,alpha=NA,gamma=NA,repetitions=1000,
norm=FALSE){
  only.one=FALSE
  if(is.na(alpha)){
    parameters = expand.grid(seq(0,1,0.1),seq(0,1,0.1),n)
    parameters = parameters[which(parameters[,1]>=
                                parameters[,2]),]
  }
  else{
    if(is.na(gamma)){
      parameters = expand.grid(alpha,seq(0,alpha,0.1),n)
      parameters = parameters[which(parameters[,1]>=
                                parameters[,2]),]
    }
    else{
      if((alpha>1)|| (alpha<0)|| (gamma>1)|| (gamma<0))
        stop("alpha and gamma must been between 0 and 1")
      else
        if(alpha<gamma)
          stop("alpha < gamma")
        else{
          parameters = c(alpha,gamma,n)
          only.one = TRUE
        }
    }
  }
}
generator = function(idx,n,alpha,gamma){
  return(a.g.model(n,alpha,gamma))
}
iterate.ford = function(tab){ #tab = [alpha,gamma,n]
  if(tab[1]>=tab[2]){

```

```

        alpha = tab[1]
        gamma = tab[2]
        n = tab[3]
        print(paste("n :",n," alpha :",alpha,
                    " gamma :",gamma))
        tree.list = lapply(1:repetitions,generator,n,
                           alpha,gamma)
        result = matrix(unlist(lapply(tree.list,
                                       balance.indices,norm=norm)),ncol=3,byrow=T)
        colnames(result) = c("COLLES.MDM.LN","SACKIN",
                              "COPHENETIC")
    }
    return(result)
}
if(only.one){
    result = iterate.ford(parameters)
}
else{
    parameters2=lapply(1:(dim(parameters)[1]),
                      function(i) as.numeric(parameters[i,]))
    result = lapply(parameters2,iterate.ford)
    paste.param = function(tab){
        return(paste("a",tab[1],"g",tab[2],sep=""))
    }
    names(result) = apply(parameters,1,paste.param)
}
return(result)
}

```

The following function, given α , γ and a phylogenetic tree, plots the distribution of the normalized versions of the Colless-like, Sackin and total cophenetic indices under the α - γ -model on \mathcal{T}_n . It also computes the percentiles of the indices of the tree under this α - γ -model. Two plots are available: one represents the percentile plots of the normalized balance indices (`percentile.plot=TRUE`), and the other represents the density plots of the normalized balance indices (`percentile.plot=FALSE`). In order to compute the distribution and percentiles, this function needs a database of trees generated under the α - γ -model. Our database is available on the GitHub at <https://github.com/LuciaRotger/CollessLike/tree/master/CollessLikeDataBase>. The trees stored in our database have between 3 and 50 leaves and the values of the parameters `alpha` and `gamma` are in $\{0, 0.1, \dots, 1\}$ such that `gamma` \leq `alpha`. If the introduced parameters are not in the list, a new computation is done with them and a new dataset of trees is generated, and their indices are also computed. The number of trees generated can be modified by the parameter

repetitions (see `indices.simulation` for more information). This computation may take some time, therefore you can compute the trees separately with `indices.simulation`, save their values and then call this function by setting them as parameter `set.indices`.

```
distribution <- function(tree,alpha=NA,gamma=NA,set.indices=NULL,
  new.simulation=FALSE,repetitions=1000,
  legend.location="topright",cex=0.75,
  percentile.plot=FALSE,db.path=getwd() ){
  ## Class of object "tree"
  if(class(tree)=="character")
    tree=read.tree(text = tree)
  if (class(tree)=="phylo")
    tree=graph.edgelist(tree$edge, directed=TRUE)
  if(class(tree)!="igraph")
    stop("Not an igraph object. Please introduce a newick
      string, an ape tree or an igraph tree.")
  n = sum(degree(tree,mode="out")==0)
  ## parameters alpha & gamma
  if(new.simulation){
    print("This process might take a long time. If you want
      to save the indices simulation, please run
      'indices.simulation' directly and then call
      'distribution' by setting the resulting table as
      the parameter 'set.indices'")
    print("Remember, our indices data base is available to
      download at: https://github.com/LuciaRotger/
      CollessLike/tree/master/CollessLikeDataBase")
    warning("New simulation required")
    indices.list=indices.simulation(n,alpha,gamma,repetitions)
    txt = bquote(paste("Parameters: ",alpha," = ",.(alpha),
      ", ",gamma," = ",.(gamma)))
  }
  else{
    if(is.null(set.indices)){
      if(alpha<gamma){
        print("Remember, our indices data base is available
          to download at: https://github.com/LuciaRotger/
          CollessLike/tree/master/CollessLikeDataBase")
        stop("alpha < gamma")
      }
    }
    else{
      if((alpha>1)||(alpha<0)||(gamma>1)||(gamma<0)){
        print("Remember, our indices data base is available
```

```

        to download at:https://github.com/LuciaRotger/CollessLike/tree/master/CollessLikeDataBase")
    stop("alpha and gamma must be between 0 and 1")
}
else{
    txt = bquote(paste("Parameters: n = ",.(n),",",
                      ",alpha," = ",.(alpha),",",
                      ",gamma," = ",.(gamma)))
    if(paste("n",n,sep="")%in%dir(db.path)){
        file = paste("CollessLikeDataBase_n",n,"_a",
                    alpha*100,"_g",gamma*100, "_r5000.txt",sep="")
        folder = paste(db.path,"n",n,"/",sep="")
        if(file %in% dir(folder)){
            indices.list=read.table(file=paste(folder,file,
                                                sep=""), header=TRUE)
        }
        else {
            print("Remember, our indices data base is
                  available to download at:https://github.com/LuciaRotger/CollessLike/tree/master/CollessLikeDataBase")
            stop(paste("The file '",file,
                      "' is not located at '",folder,"'",sep=""))
        }
    }
    else {
        print("Remember, our indices data base is
              available to download at: https://github.com/LuciaRotger/CollessLike/tree/master/CollessLikeDataBase")
        stop(paste("The folder 'n",n,
                  "' is not located at '",db.path,"'",sep=""))
    }
}
}
}
else{
    print("Indices Database introduced by user")
    txt=""
    indices.list = set.indices
}
}
if(max(indices.list)>1){
    ## maximum

```

```

max.cl = ( log(0+exp(1)) + log(2+exp(1)) )*(n-1)*(n-2)/4
max.s = n*(n-1)/2 + n-1
max.c = n*(n-1)*(n-2)/6
indices.list[,1] = round(indices.list[,1]/max.cl,4)
indices.list[,2] = round((indices.list[,2]-n)/(max.s-n),4)
indices.list[,3] = round(indices.list[,3]/max.c,4)
}
# densities
d.cl= density(indices.list[,1])
d.s = density(indices.list[,2])
d.c = density(indices.list[,3])
xlim = range(c(0,1))
ylim = range(c(0,d.cl$y,d.s$y,d.c$y))

#tree
tree.indices = balance.indices(tree)
tree.indices = round(c(tree.indices[1]/max.cl,
                      (tree.indices[2]-n)/(max.s-n),
                      tree.indices[3]/max.c),4)

f.cl=approxfun(d.cl$x,d.cl$y)
f.s=approxfun(d.s$x,d.s$y)
f.c=approxfun(d.c$x,d.c$y)
tree.densities = round(c(f.cl(tree.indices[1]),
                        f.s(tree.indices[2]),
                        f.c(tree.indices[3])),4)
tree.densities[is.na(tree.densities)]=0

a.cl = cumsum(d.cl$y)
a.cl=a.cl/max(a.cl)
a.s= cumsum(d.s$y)
a.s= a.s/max(a.s)
a.c= cumsum(d.c$y)
a.c= a.c/max(a.c)
#tree index plots percs
percs = c(a.cl[which(d.cl$x/max(d.cl$x)>tree.indices[1])[1]],
          + a.s[which(d.s$x/max(d.s$x)>tree.indices[2])[1]],
          + a.c[which(d.c$x/max(d.c$x)>tree.indices[3])[1]])
percs[is.na(percs)]=1
percs = round(percs,4)

print(paste("Tree with n=",n," leaves",sep=""))
print(paste("Colles-like: ",tree.indices[1],
           " (density:", tree.densities[1] ,"),

```

```

        Percentile:", percs[1] ,sep="")
print(paste("Sackin: ",tree.indices[2],
        " (density:", tree.densities[2] ,"),
        Percentile:", percs[2] ,sep=""))
print(paste("Cophenetic: ",tree.indices[3],
        " (density:", tree.densities[3] ,"),
        Percentile:", percs[3] ,sep=""))

#plots
par(xpd=FALSE)

if(!percentile.plot){
  plot(-1,-1 , xlab = "", ylab="Distribution of indices",
        xlim = xlim, ylim = ylim,xaxs = 'i', yaxs='i',
        main = 'Distribution of indices',
        panel.first = grid() )

  polygon(d.cl, density = -1, col=rgb(1,0,0,0.2),
        border = "red",lwd = 1)
  polygon(d.s, density = -1, col=rgb(0,1,0,0.2),
        border = "green",lwd = 1)
  polygon(d.c, density = -1, col=rgb(0,0,1,0.2),
        border = "blue",lwd = 1)
  legend(legend.location,c("Colles-Like","Sackin",
        "Cophenetic"), fill = c(rgb(1,0,0,0.2),
        rgb(0,1,0,0.2),rgb(0,0,1,0.2)),bty = 'n',
        border = NA),cex=cex)

  lines(rep(tree.indices[1],2),c(0,tree.densities[1]),
        col=rgb(1,0,0),lwd =2)
  lines(rep(tree.indices[2],2),c(0,tree.densities[2]),
        col=rgb(0,1,0),lwd =2)
  lines(rep(tree.indices[3],2),c(0,tree.densities[3]),
        col=rgb(0,0,1),lwd =2)

  lines(xlim,c(0,0),lwd=2)

  points(tree.indices[1],tree.densities[1],pch=21,
        bg=rgb(1,0,0))
  points(tree.indices[2],tree.densities[2],pch=21,
        bg=rgb(0,1,0))
  points(tree.indices[3],tree.densities[3],pch=21,
        bg=rgb(0,0,1))
}

```

```

else{
  plot(-1,-1 , xlab = "", ylab="Percentiles",
       xlim = xlim, ylim = c(0,1),xaxs = 'i', yaxs='i',
       main = 'Percentile Plot', panel.first = grid() )

  lines(d.cl$x/max(d.cl$x), a.cl, col=rgb(1,0,0))
  lines( d.s$x/max(d.s$x),  a.s, col=rgb(0,1,0))
  lines( d.c$x/max(d.c$x),  a.c, col=rgb(0,0,1))

  legend("topleft",c("Colles-Like","Sackin","Cophenetic"),
        fill = c(rgb(1,0,0,0.2),rgb(0,1,0,0.2),
                 rgb(0,0,1,0.2)),bty = 'n',border = NA),cex=cex)

  lines(rep(tree.indices[1],2),c(0,percs[1]),
        col=rgb(1,0,0),lwd =2)
  lines(rep(tree.indices[2],2),c(0,percs[2]),
        col=rgb(0,1,0),lwd =2)
  lines(rep(tree.indices[3],2),c(0,percs[3]),
        col=rgb(0,0,1),lwd =2)

  lines(xlim,c(0,0),lwd=1)

  points(tree.indices[1],percs[1],pch=21,bg=rgb(1,0,0))
  points(tree.indices[2],percs[2],pch=21,bg=rgb(0,1,0))
  points(tree.indices[3],percs[3],pch=21,bg=rgb(0,0,1))

}
mtext( txt , line = 0.3)
mtext("Normalized indices",line = 2.5,side = 1)
mtext(bquote(paste("Percentiles: ",P[C],"=",.(percs[1]),
                  ", ", P[S],"=",.(percs[2]),",",P[Phi],
                  "=",.(percs[3]) )),line = 4,side = 1)
return(percs)
}

```

For instance, we have generated the random tree depicted in Fig. 4.10 under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$ with the following commands (using `set.seed(1000)` for reproducibility)

```

set.seed(1000)
tree=a.g.model(8,0.7,0.4)
plot(tree,layout=layout.reingold.tilford(tree,root=1))

```

We can compute the three balance indices (Colles-like, Sackin and total

cophenetic) on this tree and their normalized values:

```
balance.indices(tree)

## Colles-Like      Sackin  Cophenetic
##    1.746074    18.000000    14.000000
```

```
balance.indices(tree,norm = TRUE)

## Colles-Like      Sackin  Cophenetic
##    0.0651759    0.3703704    0.2500000
```

Then, Fig. 4.11, displaying the estimation of the density function of the three balance indices under the α - γ -model with $\alpha = 0.7$ and $\gamma = 0.4$ on \mathcal{T}_8 , has been generated as follows:

```
database.location = "...../CollessLikeDataBase/"
distribution(tree,0.7,0.4,db.path = database.location)
```

Fig. 4.12, which shows a percentile plot of the three balance indices under the α - γ -model for $\alpha = 0.7$ and $\gamma = 0.4$ on \mathcal{T}_8^* and the estimated percentiles of the balance indices of the tree, has been produced with the following command:

```
distribution(tree,0.7,0.4,db.path = database.location,
             percentile.plot = TRUE)
```

The unlabeled tree of Fig. 4.13 has been generated (with `set.seed(1000)`) using $n = 8$ and $\alpha = \gamma = 0.5$, which corresponds to the uniform model. The information on it and Fig. 4.14 have been obtained with the following code:

```
set.seed(1000)
tree.uni=a.g.model(8,0.5,0.5)
plot(tree.uni,layout=layout.reingold.tilford(tree.uni,root=1))
```

```
balance.indices(tree.uni)

## Colles-Like      Sackin  Cophenetic
##    7.654334    26.000000    25.000000
```

```
balance.indices(tree.uni,norm = TRUE)

## Colles-Like      Sackin  Cophenetic
##    0.2857143    0.6666667    0.4464286
```

```
distribution(tree.uni,0.5,0.5,db.path = database.location)
```

```
## [1] "Tree with n=8 leaves"
## [1] "Colles-like: 0.2857 (density:1.1156), Percentile:0.1311"
## [1] "Sackin: 0.6667 (density:2.3944), Percentile:0.2013"
## [1] "Cophenetic: 0.4464 (density:1.3557), Percentile:0.238"
```

```
distribution(tree.uni,0.5,0.5,db.path = database.location,
             percentile.plot = TRUE)
```

```
## [1] "Tree with n=8 leaves"
## [1] "Colles-like: 0.2857 (density:1.1156), Percentile:0.1311"
## [1] "Sackin: 0.6667 (density:2.3944), Percentile:0.2013"
## [1] "Cophenetic: 0.4464 (density:1.3557), Percentile:0.238"
```

A.6.2 A real example

The tree considered in Section 4.5.1 is the following:

```
t1t2="((((Colonus_polykomos,Colobus_guereza)
,Colobus_angolensis),Colobus_satanas),(
(Procolobus_pennantii,Procolobus_badius),Procolobus_verus))"
t3 = "(Nasalis_concolor,Nasalis_larvatus)"
t4 = "(((Pygathrix_brelichi,Pygathrix_bieti)
,Pygathrix_roxellana),Pygathrix_avunculus),Pygathrix_nemaeus)"
t5 = "(Presbytis_potenziani,
(Presbytis_comata,Presbytis_frontata,Presbytis_rubicunda,Presbytis_melalophos))"
t6 = "((((Trachypithecus_phayrei,Trachypithecus_obscurus)
,Trachypithecus_pileatus,Trachypithecus_cristatus)
,Trachypithecus_francoisi),Trachypithecus_auratus)
,Trachypithecus_geei)"
t7 = "(Trachypithecus_vetulus,Trachypithecus_johnii)
,Semnopithecus_entellus)"
t8 = paste("(",t7,"",t6,"",t5,")",sep="")
t9 = paste("(",t8,"",t4,")",sep="")
t10 = paste("(",t9,"",t3,")",sep="")
all.txt= paste("(",t10,"",t1t2,")",sep="")
tree=read.tree(text = all.txt)
```

The following command depicts this tree in Fig. 4.15:

```
plot(tree)
```

The three balance indices of this tree and their normalized values are obtained as follows:

```
balance.indices(tree)
```

```
## Colles-Like      Sackin  Cophenetic
##      81.4844      161.0000      655.0000
```

```
balance.indices(tree, norm = T)
```

```
## Colles-Like      Sackin  Cophenetic
##    0.1689766    0.3259259    0.1792556
```

To establish a relationship of the previous tree with the α - γ -model we have computed the percentile of the tree for every $(\alpha, \gamma) \in \{0, 0.1, 0.2, \dots, 0.9, 1\}^2$ with $\gamma \leq \alpha$ and checked the values of the parameters (α, γ) for which the tree has higher values.

```
percentile.matrix = matrix(NA, nrow = 11, ncol = 11,
                           dimnames = list(paste("a1", seq(0, 1, 0.1), sep = "_"),
                                             paste("ga", seq(0, 1, 0.1), sep = "_")))
for(a in seq(0, 1, 0.1)){
  for(g in seq(0, a, 0.1)){
    pers = distribution(tree, a, g, db.path = database.location)
    percentile.matrix[a*10+1, g*10+1] = pers[1]
  }
}
write.table(percentile.matrix, row.names=TRUE, col.names=TRUE,
            file="C4-real-example-percentiles.txt")
```

The results are available at “C4-real-example-percentiles.txt” and the percentile plots at “C4-real-example-percentile-plots.pdf”.

The heatmap of the Fig. 4.16 is obtained with the following code:

```
require(ggplot2)
require(reshape2)
m1 = melt(percentile.matrix[, ], na.rm=T)
names(m1)=c("Alpha", "Gamma", "Value" )
a.g.range=seq(0, 1, by = 0.1)
gp1=ggplot(data=m1, aes(x=Alpha, y=Gamma, fill=Value))+
  geom_tile(color="white")
gp1 + labs(title = "Colless-Like Index", x=bquote(alpha),
           y=bquote(gamma))
```

The parameters yielding the highest percentiles are:

α	γ	Percentile
0.9	0.0	0.9031
1.0	0.2	0.8725
1.0	0.1	0.8620
1.0	0.3	0.8607

A.6.3 Computation of the mean and variance

First of all, we upload the TreeBASE database

```
load("./treeBASE-database.RData", verbose = T)
```

Then, we compute each one of the three indices of all trees and also their normalized version. The third vector is the number of leaves of each tree.

```
tb.idx = t(sapply(tb.ape, balance.indices))
tb.idx.norm = t(sapply(tb.ape, balance.indices, norm = TRUE))
tb.n = t(sapply(tb.ape, Ntip))
write.table(tb.idx, row.names = FALSE, file="C4-tb-indices.txt")
write.table(tb.idx.norm, row.names = FALSE,
            file="C4-tb-indices-norm.txt")
write.table(tb.n, "tb-n.txt")
```

The results are available in the files “C4-tb-indices.txt”, “C4-tb-indices-norm.txt” and “C4-tb-n.txt”.

Now, we compute the mean and variance of every index:

```
tb.means = list(c(), c())
tb.vars = list(c(), c())
for(n in 3:max(tb.n)){
  number.trees = which(tb.n==n)
  if(length(number.trees)>0){
    aux = list(tb.idx[number.trees,],
              tb.idx.norm[number.trees,])
    if(!is.null(dim(aux[[1]]))){
      means = list(colMeans(aux[[1]]), colMeans(aux[[2]]))
      vars = list(apply(aux[[1]], 2, var), apply(aux[[2]], 2, var))
    }
  }
  else{
    means= aux
    vars = list(c(0,0,0), c(0,0,0))
  }
}
```

```

num = length(number.trees)
tb.means[[1]] = rbind(tb.means[[1]],c(n,means[[1]],num))
tb.vars[[1]] = rbind(tb.vars[[1]],c(n,vars[[1]],num))
tb.means[[2]] = rbind(tb.means[[2]],c(n,means[[2]],num))
tb.vars[[2]] = rbind(tb.vars[[2]],c(n,vars[[2]],num))
}
}
colnames(tb.means[[1]]) [1]="Num.Leaves"
colnames(tb.means[[1]]) [5]="Num.Trees"
colnames(tb.means[[2]]) [1]="Num.Leaves"
colnames(tb.means[[2]]) [5]="Num.Trees"
colnames(tb.vars[[1]]) [1]="Num.Leaves"
colnames(tb.vars[[1]]) [5]="Num.Trees"
colnames(tb.vars[[2]]) [1]="Num.Leaves"
colnames(tb.vars[[2]]) [5]="Num.Trees"
write.table(tb.means[[1]],file="./C4-tb-means-ALL.txt",
            col.names=T,row.names=F)
write.table(tb.means[[2]],file="./C4-tb-means-norm-ALL.txt",
            col.names=T,row.names=F)
write.table(tb.vars[[1]],file="./C4-tb-vars-ALL.txt",
            col.names=T,row.names=F)
write.table(tb.vars[[2]],file="./C4-tb-vars-norm-ALL.txt",
            col.names=T,row.names=F)

```

The results of these computations are available in the files “C4-tb-means-ALL.txt”, “C4-tb-means-norm-ALL.txt”, “C4-tb-vars-ALL.txt”, and “C4-tb-vars-norm-ALL.txt”.

We have chosen the trees with $n < 300$ number of leaves and we have considered only those with more than 30 trees:

```

final.pos = which(tb.means[[1]][,1]==300)
morethan30=which(tb.means[[1]][1:final.pos,5]>30)
tb.means.reg = tb.means[[1]][morethan30,]
tb.means.reg.norm = tb.means[[2]][morethan30,]
tb.vars.reg = tb.vars[[1]][morethan30,]
tb.vars.reg.norm = tb.vars[[2]][morethan30,]
write.table(tb.means.reg,file="./C4-tb-means-regression.txt",
            col.names=T,row.names=F)
write.table(tb.means.reg.norm,col.names=T,row.names=F,
            file="./C4-tb-means-regression-norm.txt")
write.table(tb.vars.reg,file="./C4-tb-vars-regression.txt",
            col.names=T,row.names=F)
write.table(tb.vars.reg.norm,col.names=T,row.names=F,
            file="./C4-tb-vars-regression-norm.txt")

```

The results are available at “C4-tb-means-regression.txt”, “C4-tb-means-regression-norm.txt”, “C4-tb-vars-regression.txt” and “C4-tb-vars-regression-norm.txt”.

We have computed the regressions for the Colless-like index. These are the regressions of its mean values:

```
reg.cl=summary(lm(log(tb.means.reg[,2])~log(tb.means.reg[,1])))
reg.cl
```

```
##
## Call:
## lm(formula = log(tb.means.reg[, 2]) ~ log(tb.means.reg[, 1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58088 -0.05385  0.01801  0.09498  0.33918
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.62477    0.07181   -8.7 9.17e-14 ***
## log(tb.means.reg[, 1])  1.58463    0.01860   85.2 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1495 on 96 degrees of freedom
## Multiple R-squared:  0.9869, Adjusted R-squared:  0.9868
## F-statistic: 7260 on 1 and 96 DF,  p-value: < 2.2e-16
```

```
reg.cl.norm=summary(lm(log(tb.means.reg.norm[,2])~
                      log(tb.means.reg.norm[,1])))
reg.cl.norm
```

```
##
## Call:
## lm(formula = log(tb.means.reg.norm[, 2]) ~ log(tb.means.reg.norm[,
##      1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48405 -0.06863  0.00563  0.06702  0.50399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.51414    0.06350   8.097 1.77e-12 ***
## log(tb.means.reg.norm[, 1]) -0.56860    0.01645 -34.573 < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1322 on 96 degrees of freedom
## Multiple R-squared:  0.9257, Adjusted R-squared:  0.9249
## F-statistic: 1195 on 1 and 96 DF,  p-value: < 2.2e-16
```

These are the regressions of the variances of the values of the Colless-like index:

```
reg.cl.var=summary(lm(log(tb.vars.reg[,2])~log(tb.vars.reg[,1])))
reg.cl.var
```

```
##
## Call:
## lm(formula = log(tb.vars.reg[, 2]) ~ log(tb.vars.reg[, 1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.1727 -0.1281  0.0881  0.2379  1.9181
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.57579    0.24489  -10.52  <2e-16 ***
## log(tb.vars.reg[, 1])  3.12798    0.06342   49.32  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5099 on 96 degrees of freedom
## Multiple R-squared:  0.962, Adjusted R-squared:  0.9616
## F-statistic: 2432 on 1 and 96 DF,  p-value: < 2.2e-16
```

```
reg.cl.var.norm=summary(lm(log(tb.vars.reg.norm[,2])~
                        log(tb.vars.reg.norm[,1])))
reg.cl.var.norm
```

```
##
## Call:
## lm(formula = log(tb.vars.reg.norm[, 2]) ~ log(tb.vars.reg.norm[,
##      1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7861 -0.1327  0.0120  0.1928  2.2477
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.29796    0.22059  -1.351    0.18
## log(tb.vars.reg.norm[, 1]) -1.17848    0.05713 -20.628 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4593 on 96 degrees of freedom
## Multiple R-squared:  0.8159, Adjusted R-squared:  0.814
## F-statistic: 425.5 on 1 and 96 DF,  p-value: < 2.2e-16
```

The following code produces Fig. 4.17(a):

```
plot(tb.means.reg[,1],tb.means.reg[,2],type = "l",
      xlab="Number of leaves",ylab="Colless-like index")
lines(1:300,exp(reg.cl$coefficients[1,1])*(1:300)^
      reg.cl$coefficients[2,1],col="red")
```

As to the Sackin index, these are the regressions of its mean values:

```
reg.sa=summary(lm(log(tb.means.reg[,3])~log(tb.means.reg[,1])))
reg.sa
```

```
##
## Call:
## lm(formula = log(tb.means.reg[, 3]) ~ log(tb.means.reg[, 1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.309558 -0.024226  0.008616  0.048661  0.232089
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.37285    0.03870   9.635 9.03e-16 ***
## log(tb.means.reg[, 1]) 1.43583    0.01002 143.267 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08057 on 96 degrees of freedom
## Multiple R-squared:  0.9953, Adjusted R-squared:  0.9953
## F-statistic: 2.053e+04 on 1 and 96 DF,  p-value: < 2.2e-16
```

```
reg.sa.norm=summary(lm(log(tb.means.reg.norm[,3])~
                      log(tb.means.reg.norm[,1])))
reg.sa.norm
```

```
##
## Call:
## lm(formula = log(tb.means.reg.norm[, 3]) ~ log(tb.means.reg.norm[,
##     1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33568 -0.03219  0.00722  0.05739  0.24989
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.83896   0.04166   20.14 <2e-16 ***
## log(tb.means.reg.norm[, 1]) -0.53463   0.01079  -49.55 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08675 on 96 degrees of freedom
## Multiple R-squared:  0.9624, Adjusted R-squared:  0.962
## F-statistic:  2455 on 1 and 96 DF,  p-value: < 2.2e-16
```

These are the regressions of the variances of the values of the Sackin index:

```
reg.sa.var=summary(lm(log(tb.vars.reg[,3])~log(tb.vars.reg[,1])))
reg.sa.var
```

```
##
## Call:
## lm(formula = log(tb.vars.reg[, 3]) ~ log(tb.vars.reg[, 1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5707 -0.1466  0.1062  0.2778  1.8728
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.43962   0.26752  -12.86 <2e-16 ***
## log(tb.vars.reg[, 1])  3.22249   0.06929   46.51 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.557 on 96 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9571
## F-statistic:  2163 on 1 and 96 DF,  p-value: < 2.2e-16
```

```
reg.sa.var.norm=summary(lm(log(tb.vars.reg.norm[,3])~
                          log(tb.vars.reg.norm[,1])))
reg.sa.var.norm

##
## Call:
## lm(formula = log(tb.vars.reg.norm[, 3]) ~ log(tb.vars.reg.norm[,
##      1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3981 -0.1552  0.0853  0.2259  2.0207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.4750     0.2490  -5.923 4.92e-08 ***
## log(tb.vars.reg.norm[, 1]) -0.9082     0.0645 -14.081 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5185 on 96 degrees of freedom
## Multiple R-squared:  0.6738, Adjusted R-squared:  0.6704
## F-statistic: 198.3 on 1 and 96 DF,  p-value: < 2.2e-16
```

The following code produces Fig. 4.17(b):

```
plot(tb.means.reg[,1],tb.means.reg[,3],type = "l",
     xlab="Number of leaves",ylab="Sackin index")
lines(1:300,exp(reg.sa$coefficients[1,1])*(1:300)^
      reg.sa$coefficients[2,1],col="green")
```

Finally, these are the regressions of the mean values for the total cophenetic index:

```
reg.co=summary(lm(log(tb.means.reg[,4])~log(tb.means.reg[,1])))
reg.co

##
## Call:
## lm(formula = log(tb.means.reg[, 4]) ~ log(tb.means.reg[, 1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61240 -0.04981  0.03907  0.09208  0.34377
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.66353    0.07471  -22.27  <2e-16 ***
## log(tb.means.reg[, 1])  2.54769    0.01935  131.67  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1556 on 96 degrees of freedom
## Multiple R-squared:  0.9945, Adjusted R-squared:  0.9944
## F-statistic: 1.734e+04 on 1 and 96 DF,  p-value: < 2.2e-16

reg.co.norm=summary(lm(log(tb.means.reg.norm[,4])~
                      log(tb.means.reg.norm[,1])))
reg.co.norm

##
## Call:
## lm(formula = log(tb.means.reg.norm[, 4]) ~ log(tb.means.reg.norm[,
##      1]))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51558 -0.05255 -0.00597  0.06874  0.50857
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.81751    0.05723   14.29  <2e-16 ***
## log(tb.means.reg.norm[, 1]) -0.60554    0.01482  -40.85  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1192 on 96 degrees of freedom
## Multiple R-squared:  0.9456, Adjusted R-squared:  0.945
## F-statistic: 1669 on 1 and 96 DF,  p-value: < 2.2e-16
```

These are the regressions of the variances of the values of the total cophenetic index:

```
reg.co.var=summary(lm(log(tb.vars.reg[,4])~log(tb.vars.reg[,1])))
reg.co.var

##
## Call:
## lm(formula = log(tb.vars.reg[, 4]) ~ log(tb.vars.reg[, 1]))
##
## Residuals:
```



```
##      Min      1Q  Median      3Q      Max
## -3.6544 -0.1626  0.1001  0.2883  1.9319
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.50441    0.28383  -19.39  <2e-16 ***
## log(tb.vars.reg[, 1])  5.20711    0.07351   70.83  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.591 on 96 degrees of freedom
## Multiple R-squared:  0.9812, Adjusted R-squared:  0.981
## F-statistic:  5018 on 1 and 96 DF,  p-value: < 2.2e-16

reg.co.var.norm=summary(lm(log(tb.vars.reg.norm[,4])~
                          log(tb.vars.reg.norm[,1])))
reg.co.var.norm

##
## Call:
## lm(formula = log(tb.vars.reg.norm[, 4]) ~ log(tb.vars.reg.norm[,
##      1]))
##
## Residuals:
##      Min      1Q  Median      3Q      Max
## -3.2678 -0.1254  0.0406  0.2242  2.2615
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.54233    0.24535   -2.21  0.0294 *
## log(tb.vars.reg.norm[, 1]) -1.09934    0.06354  -17.30  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5108 on 96 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7546
## F-statistic:  299.3 on 1 and 96 DF,  p-value: < 2.2e-16
```

And the code producing Fig. 4.17(c):

```
plot(tb.means.reg[,1],tb.means.reg[,4],type = "l",
      xlab="Number of leaves",ylab="Cophenetic index")
lines(1:300,exp(reg.co$coefficients[1,1])*(1:300)^
      reg.co$coefficients[2,1],col="blue")
```

A.6.4 Computation of the number of ties

In this section we compute the number of ties of the three balance indices under consideration. For every number of leaves n and for every index, we have computed the numbers of pairs of trees with n leaves in TreeBASE having the same value of the corresponding index:

```
ties.cl=c()
ties.sa=c()
ties.co=c()
for(i in 3:max(tb.n)){
  aux=tb.idx[tb.n==i,]
  ties.cl=rbind(ties.cl,c(i,sum(choose(table(aux[,1]),2))))
  ties.sa=rbind(ties.sa,c(i,sum(choose(table(aux[,2]),2))))
  ties.co=rbind(ties.co,c(i,sum(choose(table(aux[,3]),2))))
}
```

The following commands produce Fig. 4.18:

```
plot(3:150,ties.cl[1:148,2],type="l",xlab="Number of leaves",
     ylab="Number of ties",col="red",lwd=2)
lines(3:150,ties.sa[1:148,2],col="green",lwd=1)
lines(3:150,ties.co[1:148,2],col="blue",lwd=1,lty=1)
legend("topright",legend=c("Colless-like index",
                           "Sackin index","Cophenetic index"),
       lty=c("solid","solid","solid"),
       col=c("red","green","blue"))
```

A.6.5 Computation of Spearman's rank correlation

The global Spearman's rank correlation coefficient of \mathfrak{C} and S is

```
cor(tb.idx[,1],tb.idx[,2],method="spearman")
```

```
## [1] 0.97645
```

And the global Spearman's rank correlation coefficient of \mathfrak{C} and Φ is

```
cor(tb.idx[,1],tb.idx[,3],method="spearman")
```

```
## [1] 0.9618565
```

We compute now the Spearman rank correlation coefficient of the indices on all trees in TreeBASE grouping them by their number of leaves n . As usual, we have considered only those numbers of leaves with more than 30 trees:

```
spearman.sackin=c()
for(i in 1:max(tb.n)){
  aux=tb.idx[tb.n==i,]
  if(dim(aux)[1]>30){
    aux2=rank(aux[,1])
    aux3=rank(aux[,2])
    spearman.sackin=rbind(spearman.sackin,
      c(i,cor(aux2,aux3,method="spearman")))
  }
}
colnames(spearman.sackin)=c("Num.Leaves","SpearmanRank")
write.table(spearman.sackin,file="./C4-tb-spearman-CL-S.txt",
  col.names=T,row.names=F)

spearman.coph=c()
for(i in 1:max(tb.n)){
  aux=tb.idx[tb.n==i,]
  if(dim(aux)[1]>30){
    aux2=rank(aux[,1])
    aux3=rank(aux[,3])
    spearman.coph=rbind(spearman.coph,
      c(i,cor(aux2,aux3,method="spearman")))
  }
}
colnames(spearman.sackin)=c("Num.Leaves","SpearmanRank")
write.table(spearman.coph,file="./C4-tb-spearman-CL-C.txt",
  col.names=T,row.names=F)
```

All values are available in the files “C4-tb-spearman-CL-S.txt” and “C4-tb-spearman-CL-C.txt”.

Fig. 4.19 is produced with the following commands:

```
plot(spearman.sackin[,1],spearman.sackin[,2],type="l",
  col="green",xlab="number of leaves",
  ylab="Spearman's rank correlation coefficient")
legend("bottomright",legend="C and S",lty="solid",col="green")
summary(lm(spearman.sackin[,2]~spearman.sackin[,1]))
plot(spearman.coph[,1],spearman.coph[,2],type="l",
  col="blue",xlab="number of leaves",
  ylab="Spearman's rank correlation coefficient")
```

```
legend("bottomright",legend=bquote(paste("C and ",Phi," ")),
      lty= "solid",col="blue" )
```

We also can compute the regression of the results:

```
summary(lm(spearman.sackin[,2] ~ spearman.sackin[,1]))
```

```
##
## Call:
## lm(formula = spearman.sackin[, 2] ~ spearman.sackin[, 1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.069523 -0.004901  0.003692  0.007571  0.013644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.764e-01  2.149e-03  454.235  <2e-16 ***
## spearman.sackin[, 1] 8.648e-05  2.988e-05   2.894  0.0047 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0128 on 96 degrees of freedom
## Multiple R-squared:  0.08025,    Adjusted R-squared:  0.07067
## F-statistic: 8.377 on 1 and 96 DF,  p-value: 0.004704
```

```
summary(lm(spearman.coph[,2] ~ spearman.coph[,1]))
```

```
##
## Call:
## lm(formula = spearman.coph[, 2] ~ spearman.coph[, 1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27550 -0.01853  0.01222  0.03897  0.15033
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.9012445  0.0110186  81.793  < 2e-16 ***
## spearman.coph[, 1] -0.0006160  0.0001532  -4.022  0.000115 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.06563 on 96 degrees of freedom
## Multiple R-squared: 0.1442, Adjusted R-squared: 0.1353
## F-statistic: 16.17 on 1 and 96 DF, p-value: 0.0001151
```

A.6.6 A test on the distribution of TreeBASE

In this case, we focus our study on the distribution of the normalized Colless-like index on the TreeBASE, so we consider only the following array of its normalized values:

```
tb.colless=tb.idx.norm[,1]
```

The following functions extract the data from our database and perform for every pair of `alpha` and `gamma` a `chisq.test`. Then, we study in detail each interesting case.

```
read.idx = function(n,alpha,gamma,norm=T){
  file = paste("CollessLikeDataBase_n",n,"_a",alpha*100,"_g",
              gamma*100,"_r5000.txt",sep="")
  folder = paste(database.location,"n",n,"/",sep="")
  if(file %in% dir(folder))
    indices.list=read.table(file=paste(folder,file,sep=""),
                          header=TRUE)
  if(norm) indices.list = indices.list[,1]/(( log(0+exp(1)) +
      log(2+exp(1)) )*(n-1)*(n-2)/4)
  else indices.list = indices.list[,1]
  return(indices.list)
}

pis.ag = function(alpha,gamma,intervals,dens=F){
  all.trees.study = unlist(lapply(3:50,read.idx,alpha=alpha,
                                gamma=gamma))
  brs = seq(0,1,length.out = intervals+1)
  fe = hist(all.trees.study,breaks = brs,plot=F)
  pis = fe$counts/sum(fe$counts)
  if(dens){
    dens = density(all.trees.study)
    new.pi = c()
    for(i in 1:intervals){
      aux=integrate(splinefun(dens$x,dens$y), brs[i],
                  brs[i+1])$value
      new.pi= c(new.pi,aux)
    }
    new.pi=new.pi/sum(new.pi)
  }
```

```

    return(list(pis,new.pi,dens))
  }
  else return(pis)
}

parameters.study = function(fo, intervals,dens=F,info=FALSE){
  ntb = sum(fo$counts)
  parameters = expand.grid(seq(0,1,0.1),seq(0,1,0.1))
  parameters = parameters[which(parameters[,1]>=parameters[,2]),]
  pvalues=c()
  all.info=list()
  for(i in 1:65){
    if(i!=11){
      pis = pis.ag(parameters[i,1],parameters[i,2],intervals,
                  dens = dens)

      table.info=grouping(fo$counts,pis*ntb,fo$breaks)
      out.test=chisq.test(table.info[,1],p=table.info[,2]/ntb)
      pvalues=c(pvalues,out.test$p.value)
      print(paste("a:",parameters[i,1],", g:",parameters[i,2],
                  "-->",out.test$p.value))
      if(info)print(out.test)
      all.info[[i]]=list(table.info,out.test)
    }
    else{
      print("a: 1 , g: 0 --> ERROR")
      pvalues=c(pvalues,-1)
    }
  }
  print("a: 1 , g: 1 --> ERROR")
  parameters=cbind(parameters,c(pvalues,-1))
  return(list(parameters,all.info))
}

grouping = function(xx,yy,breaks){
  ois = xx[]
  eis = yy[]
  to.modify = which(eis<5)
  while(length(to.modify)>0){
    to.modify = to.modify[1]
    N = length(eis)
    if(to.modify>1){
      if(to.modify<N){
        aux1 = eis[to.modify-1]+eis[to.modify]
        aux2 = eis[to.modify]+eis[to.modify+1]

```

```

    if(aux1<aux2){
      eis[to.modify-1] = aux1
      ois[to.modify-1] = ois[to.modify-1]+ois[to.modify]
      breaks = breaks[-to.modify]
    }
    else{
      eis[to.modify+1] = aux2
      ois[to.modify+1] = ois[to.modify]+ois[to.modify+1]
      breaks = breaks[-(to.modify+1)]
    }
    eis = eis[-to.modify]
    ois = ois[-to.modify]
  }
  else{ ## to.modify=N
    eis[N-1] = eis[N]+eis[N-1]
    eis = eis[-N]
    ois[N-1] = ois[N]+ois[N-1]
    ois = ois[-N]
    breaks = breaks[-(N)]
  }
}
else{ ## to.modify=1
  eis[2] = eis[1]+eis[2]
  eis = eis[-1]
  ois[2] = ois[1]+ois[2]
  ois = ois[-1]
  breaks = breaks[-2]
}
to.modify=which(eis<5)
}
N=length(breaks)
return(cbind(ois,eis,linf=breaks[1:(N-1)],lsup=breaks[2:N]))
}

```

The following code executes the study

```

intervals=100
fo = hist(tb.colless,breaks = intervals,plot=F)
ntb = sum(fo$counts)
results.study=parameters.study(fo,intervals)

```

```

## [1] "a: 0 , g: 0 --> 0"
## [1] "a: 0.1 , g: 0 --> 0"
## [1] "a: 0.2 , g: 0 --> 0"

```

```
## [1] "a: 0.3 , g: 0 --> 0"
## [1] "a: 0.4 , g: 0 --> 0"
## [1] "a: 0.5 , g: 0 --> 0"
## [1] "a: 0.6 , g: 0 --> 0"
## [1] "a: 0.7 , g: 0 --> 0"
## [1] "a: 0.8 , g: 0 --> 0"
## [1] "a: 0.9 , g: 0 --> 0"
## [1] "a: 1 , g: 0 --> ERROR"
## [1] "a: 0.1 , g: 0.1 --> 0"
## [1] "a: 0.2 , g: 0.1 --> 0"
## [1] "a: 0.3 , g: 0.1 --> 0"
## [1] "a: 0.4 , g: 0.1 --> 0"
## [1] "a: 0.5 , g: 0.1 --> 0"
## [1] "a: 0.6 , g: 0.1 --> 0"
## [1] "a: 0.7 , g: 0.1 --> 0"
## [1] "a: 0.8 , g: 0.1 --> 0"
## [1] "a: 0.9 , g: 0.1 --> 0"
## [1] "a: 1 , g: 0.1 --> 0"
## [1] "a: 0.2 , g: 0.2 --> 0"
## [1] "a: 0.3 , g: 0.2 --> 0"
## [1] "a: 0.4 , g: 0.2 --> 0"
## [1] "a: 0.5 , g: 0.2 --> 0"
## [1] "a: 0.6 , g: 0.2 --> 0"
## [1] "a: 0.7 , g: 0.2 --> 0"
## [1] "a: 0.8 , g: 0.2 --> 0"
## [1] "a: 0.9 , g: 0.2 --> 0"
## [1] "a: 1 , g: 0.2 --> 0"
## [1] "a: 0.3 , g: 0.3 --> 0"
## [1] "a: 0.4 , g: 0.3 --> 0"
## [1] "a: 0.5 , g: 0.3 --> 0"
## [1] "a: 0.6 , g: 0.3 --> 0"
## [1] "a: 0.7 , g: 0.3 --> 0"
## [1] "a: 0.8 , g: 0.3 --> 0"
## [1] "a: 0.9 , g: 0.3 --> 0"
## [1] "a: 1 , g: 0.3 --> 0"
## [1] "a: 0.4 , g: 0.4 --> 0"
## [1] "a: 0.5 , g: 0.4 --> 0"
## [1] "a: 0.6 , g: 0.4 --> 0"
## [1] "a: 0.7 , g: 0.4 --> 1.01304353682268e-113"
## [1] "a: 0.8 , g: 0.4 --> 0"
## [1] "a: 0.9 , g: 0.4 --> 0"
## [1] "a: 1 , g: 0.4 --> 0"
## [1] "a: 0.5 , g: 0.5 --> 0"
## [1] "a: 0.6 , g: 0.5 --> 0"
## [1] "a: 0.7 , g: 0.5 --> 0"
## [1] "a: 0.8 , g: 0.5 --> 2.42464512709513e-168"
## [1] "a: 0.9 , g: 0.5 --> 0"
## [1] "a: 1 , g: 0.5 --> 0"
## [1] "a: 0.6 , g: 0.6 --> 0"
## [1] "a: 0.7 , g: 0.6 --> 0"
## [1] "a: 0.8 , g: 0.6 --> 0"
## [1] "a: 0.9 , g: 0.6 --> 0"
## [1] "a: 1 , g: 0.6 --> 0"
```



```
## [1] "a: 0.7 , g: 0.7 --> 0"
## [1] "a: 0.8 , g: 0.7 --> 0"
## [1] "a: 0.9 , g: 0.7 --> 0"
## [1] "a: 1 , g: 0.7 --> 0"
## [1] "a: 0.8 , g: 0.8 --> 0"
## [1] "a: 0.9 , g: 0.8 --> 0"
## [1] "a: 1 , g: 0.8 --> 0"
## [1] "a: 0.9 , g: 0.9 --> 0"
## [1] "a: 1 , g: 0.9 --> 0"
## [1] "a: 1 , g: 1 --> ERROR"
```

The following cases are those with p-value different from 0:

```
results.study[[1]][which(results.study[[1]][,3]>0),]
```

```
##      Var1 Var2 c(pvalues, -1)
## 52   0.7  0.4  1.013044e-113
## 64   0.8  0.5  2.424645e-168
```

```
p42 = pis.ag(0.7,0.4,intervals)
```

Although the p-value is very small, we plot the results in Fig. 4.21 as follows:

```
plot(-1,-1,xlim =c(0,1),ylim=c(0,0.04),xlab="Indices",ylab="",
      pch=20,col="white",
      main="Distribution of Colless-Like indices")
lines(fo$mids,fo$counts/ntb,lwd=2)
lines(fo$mids,p42,pch=20,col="blue",lwd=2)
legend(c(0.4,0.75),c(0.04,0.033),legend = c("TreeBase",
      expression(paste(alpha, " = 0.7, ", gamma, "= 0.4"))),
      col=c("black", "blue"),lwd=2 ,cex=0.75)
```

Besides the whole TreeBASE as explained above, we have also considered different subsets of it:

```
tb.kind=function(tr)return(tr$kind)
kind=unlist(lapply(tb.ape ,tb.kind))
tb.type=function(tr)return(tr$type)
type=unlist(lapply(tb.ape ,tb.type))
idx.spe=which(kind=="Species Tree")
idx.gen=which(kind=="Gene Tree")
idx.spe.con=intersect(idx.spe,which(type=="Consensus"))
idx.spe.sin=intersect(idx.spe,which(type=="Single"))
tb.spe=tb.colless[idx.spe]
```

```

tb.spe.n = tb.n[idx.spe,]
tb.spe.con=tb.colless[idx.spe.con]
tb.spe.con.n = tb.n[idx.spe.con,]
tb.spe.sin=tb.colless[idx.spe.sin]
tb.spe.sin.n = tb.n[idx.spe.sin,]

erase.attributes = function(tree){
  tree$id = NULL
  tree$Tr.id = NULL
  tree$type = NULL
  tree$kind = NULL
  tree$quality = NULL
  return(tree)
}
tb.ape.aux = lapply(tb.ape,erase.attributes)

repetitions = duplicated(tb.ape.aux)
pos.repetitions = (1:length(repetitions))[repetitions]
tb.ape.no.reps = tb.ape[!repetitions]

tb.qua=function(tr)return(tr$quality)
qua=unlist(lapply(tb.ape,tb.qua))
idx.qua=which(qua=="Species Tree")
tb.spe=tb.colless[setdiff(idx.spe,pos.repetitions)]
tb.spe.n = tb.n[setdiff(idx.spe,pos.repetitions), ]
tb.spe.con=tb.colless[setdiff(idx.spe.con,pos.repetitions)]
tb.spe.con.n = tb.n[setdiff(idx.spe.con,pos.repetitions), ]
tb.spe.sin=tb.colless[setdiff(idx.spe.sin,pos.repetitions)]
tb.spe.sin.n = tb.n[setdiff(idx.spe.sin,pos.repetitions), ]

```

We have repeated the study explained above for these subsets of TreeBASE, comparing the distribution of the normalized Colless-like indices of their trees with the estimated theoretical distributions by means of goodness-of-fit tests, and the results have been the same, that is, all p-values have also turned out to be negligible.