Universitat
de les Illes Balears

DOCTORAL THESIS

2020

# New results on old and new balance indices

Tomás Martínez Coronado

Universitat
de les Illes Balears

# DOCTORAL THESIS

2020

Doctoral Programme in Information and Communications Technology

# New results on old and new balance indices

Tomás Martínez Coronado
Director: Francesc Andreu Rosselló Llompart
Advisor: Francesc Andreu Rosselló Llompart

Doctor by the Universitat de les Illes Balears

*Para mi abuelo,*
*que me enseñó a multiplicar y a dividir.*

# Agradecimientos

> 'Men work together,' I told him
> from the heart,
> 'Whether they work together
> or apart.'
>
> Robert Frost, *A tuft of flowers*

Todo trabajo científico es, de manera implícita, un agradecimiento y, con suerte, a veces incluso una modesta retribución. En primer lugar, y de manera inmediata, a la sociedad en cuyo seno se ha realizado y sin la que hubiera sido en todo punto imposible. En segundo lugar, y de manera ideal, a toda la humanidad. Esta investigación, como tantas otras, ha sido llevada a cabo en una universidad pública. El que firma estas líneas ha sido educado toda su vida en instituciones públicas, desde la escuela primaria hasta la universidad. Por tanto, es de manera indirecta el producto de todos y cada uno de los maestros y profesores que he tenido.

La prudencia aconsejaría acabar aquí los agradecimientos, ya que todo el que se atreve a concretar un agradecimiento debe, necesariamente, aceptar que corre el riesgo de dejarse a alguien. Pero merece la pena intentarlo.

En primer lugar, debo agradecérselo a mi director, Cesc, que ha soportado de manera estoica todas mis extravagancias estilísticas, y que ha intentado de manera a menudo infructuosa convencerme de no incluirlas. Los casos en los que no lo ha conseguido son sólo culpa de mi obstinada cabezonería. Lo mejor que hay en este trabajo se beneficia de manera incalculable de su guía y revisión. Su perfeccionismo no puede ponerse en duda. Si, pese a todo, algunos errores han persistido hasta la versión final, queda enteramente bajo mi responsabilidad.

A Pedro y Lucía, mis dos compañeros de despacho, y a Biel, Onofre, Raquel, Jordi, Mar y aquellos que, como yo, se han echado a la espalda el enorme peso de componer una tesis doctoral. A Mercè, Irene y Ana Belén, con quienes he compartido asignaturas y que me han ayudado en todo momento a descubrir la que probablemente sea la más importante de las facetas del académico: la de enseñar a otros lo que otros le han enseñado a uno. Gracias por vuestra paciencia infinita. También a Arnau Mir, a quien agradezco en especial su arte sumatorio, pero también a Biel Cardona, Jairo, Pere, Ricardo y, en una palabra, a todo el BIOCOM. Y a Iván, Pablo, Marina, Geoffroy, Lucie... y todos aquellos con los que he comido alguna vez bajo los árboles del campus a lo largo de estos cuatro años.

A mis amigos y amigas, demasiados para ser incluidos en una relación exhaustiva. Algunos de ellos han aparecido ya. Toda mención lleva aparejada un diminuto agravio comparativo, así que me contentaré con decir que se lo agradezco a todos y cada uno de ellos. Me hacen sentir querido. No obstante, no puedo dejar de mencionar a La Mano, en este décimo aniversario de la primera manofanía. Mano.

Como es natural, un lugar preeminente deben ocuparlo los miembros de mi familia. Mi madre, cuya ternura y apertura de mente me han permitido explorar todos los caminos que he querido. Y han sido muchos. Mi abuelo, quien con su austero amor manchego me ha apoyado a lo largo y ancho de mi carrera, desde el principio. Gracias por su apoyo y críticas, a menudo constructivas. Gracias a mi padre y Estelle, mi otra madre, y Guillem, mi hermano, por dejarme discutir ideas y darme lasaña los domingos por la noche. A mi abuela, quien tristemente no ha podido ver mi tesis acabada, y a Lino. Y también a mi familia en Inca, en especial a Francisco y Vanesa, con quienes he aprendido cosas que van más allá de lo académico. Y, en general, a todos los miembros de mi familia.

Finalmente, quiero agradecérselo a mi *karass*. Para Gerard sólo tengo una palabra: *Bro*. Para Martina tengo muchas más, pero ya se las he dicho casi todas.

Esta investigación ha sido financiada por el Ministerio de Ciencia, Innovación y Universidades y el Fondo Europeo de Desarrollo Regional (FEDER) a través de los proyectos *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).

# Contents

# Abstract

The main motivation behind the quantitative study of phylogenetic tree shapes is the belief that they reflect properties of the evolutionary processes that have derived them. The contribution of our research is the addition, to the existing set of quantitative techniques in this field, of two new balance indices, as well as the proof of some results concerning two old ones.

The minimum value of the Colless index, as well as the trees attaining it, have been unknown ever since the introduction of this index in 1982. We solve this problem by providing a full characterization of these trees and closed formulæ for the minimum value of the Colless index. We also introduce a new balance index for bifurcating trees, the Quadratic Colless index, defined as the sum of the squares, not the absolute values, of the difference in the number of leaves of the subtrees rooted at each internal node of a given tree. This new measure happens to be easier to manipulate, and we have proved that the maximally balanced tree and the caterpillar are exactly the trees attaining its minimum and maximum values, respectively. We also show that it has better statistical properties than those of the original Colless index, and we have been able to compute its expected value and variance under both the Yule and Uniform models.

In his 1972 paper on tree balance, Sackin proposed the use of the variation of the leaves' depths as a measure of the balance of a tree. Although somewhat popular in the decades of 1970 and 1980, this measure was never thoroughly studied and is now almost completely forgotten. We study some of its properties, characterizing the trees attaining its maximum value as being the caterpillars, and providing a quasi-linear algorithm to compute the bifurcating trees attaining its minimum value. Nevertheless, we also show that these are almost never maximally balanced. We also provide closed formulæ for its expected value under the Uniform and Yule models, as well as for the variance of the Sackin and Cophenetic indices and the Total Area under the Uniform model.

In the last of the central chapters of this memoir we introduce a new balance index for multifurcating trees: the Quartet index. We find the multifurcating and bifurcating trees attaining its extreme values: exactly the stars and caterpillars in the multifurcating case, and the maximally balanced trees and the caterpillars in the bifurcating case. We also give a recurrence to compute its maximum value for bifurcating trees. Thus, we prove that its range of values is the largest among the balance indices existing in the literature. Furthermore, we give its expected value and variance under both the $\beta$ and $\alpha$-$\gamma$ probabilistic models for phylogenetic trees. To our knowledge, this is the first shape index for phylogenetic trees whose first moments under the $\alpha$-$\gamma$-model are known. We end this chapter by pointing out that this index can be easily generalized to other families of directed graphs and still preserve its good statistical properties.

# Resumen

LA PRINCIPAL motivación tras el estudio cuantitativo de las formas subyacentes a los árboles filogenéticos es la creencia de que reflejan propiedades de los procesos evolutivos que los han derivado. La contribución de nuestra investigación es la adición, al conjunto de técnicas cuantitativas existentes, de dos nuevos índices de equilibrio, además de probar algunos resultados sobre dos antiguos índices de equilibrio.

El valor mínimo del índice de Colless, junto con los árboles que lo alcanzan, han sido desconocidos desde la introducción de éste en 1982. Nosotros resolvemos este problema, presentando una caracterización completa de dichos árboles, así como fórmulas cerradas para calcular su valor mínimo. Además, presentamos un nuevo índice de equilibrio para árboles binarios, el índice de Colless Cuadrático, definido como la suma de los cuadrados, y no de los valores absolutos, de la diferencia entre los números de hojas de los subárboles enraizados en cada nodo interno de un árbol dado. Esta nueva medida resulta ser más fácil de manipular, y hemos demostrado que el árbol máxime equilibrado y el árbol oruga son exactamente los árboles que alcanzan sus valores mínimo y máximo, respectivamente. También probamos que tiene mejores propiedades estadísticas, y calculamos su esperanza y varianza bajo los modelos de Yule y Uniforme.

En su artículo fundacional de 1972, Sackin propuso el uso de la variación de la profundidad de las hojas como medida del equilibrio de un árbol. Aunque esta medida fue más o menos popular en las décadas de 1970 y 1980, nunca se estudió en profundidad y ha sido casi completamente olvidada. Estudiamos algunas de sus propiedades, caracterizando los árboles alcanzando su valor máximo como los árboles oruga, y presentando un algoritmo casi-lineal para construir aquellos que alcanzan su valor mínimo. Sin embargo, también demostramos que estos últimos casi nunca son máxime equilibrados. Acabamos el capítulo proporcionando fórmulas cerradas para su esperanza bajo los modelos de Yule y Uniforme, además de la varianza de los índices de Sackin y Cofenético y el Área Total bajo el modelo Uniforme.

En el último de los capítulos centrales de esta memoria introducimos un nuevo índice de equilibrio para árboles multifurcados: el índice de Cuartetos. Encontramos los árboles multifurcados y binarios que alcanzan sus valores extremos: los árboles estrella y oruga en el primer caso, y los máxime equilibrados y oruga en el segundo. También damos una recurrencia para calcular su valor máximo para árboles binarios. Así, probamos que su rango de valores es el mayor de entre los índices de equilibrio existentes en la literatura. Además, calculamos su esperanza y varianza bajo los modelos probabilísticos de árboles filogenéticos $\beta$ y $\alpha$-$\gamma$. Por lo que sabemos, es el primer índice topológico de árboles filogenéticos del que se conocen sus primeros momentos bajo el modelo $\alpha$-$\gamma$. Finalmente, señalamos que este índice puede ser fácilmente generalizado a otras familias de grafos dirigidos preservando sus buenas propiedades estadísticas.

# Resum

La principal motivació rere l'estudi quantiatiu de les formes subjacents als arbres filogenètics és la creença que aquestes reflecteixen propietats dels procesos evolutius de què es deriven. La contribució de la nostra recerca és l'adició, al conjunt de tècniques quantitatives existents, de dos nous índexos d'equilibri, a més de provar alguns resultats sobre dos d'antics.

El valor mínim de l'índex de Colless, així com els arbres que l'assoleixen, han sigut desconeguts des de la introducció d'aquest en 1982. Nosaltres resolem aquest problema tot presentant una caracterització completa d'aquests arbres, així com fórmules tancades per a calcular el seu índex de Colless. A més, presentam un nou índex d'equilibri, l'índex de Colless Quadràtic, definit com la suma dels quadrats, i no dels valors absoluts, de les diferències entre els nombres de fulles dels subarbres arrelats a cada node interior d'un arbre donat. Aquesta nova mesura resulta ser més fàcil de manipular i hem sigut capaços de demostrar que l'arbre màximament equilibrat i l'arbre eruga són exactament els arbres que assoleixen els seus valors mínim i màxim, respectivament. També provam que té millors propietats estadístiques, i en calculam l'esperança i variància sota els models de Yule i Uniforme.

Al seu article fundacional de 1972, Sackin va proposar l'ús de la variació de la profunditat de les fulles com a mesura de l'equilibri d'un arbre. Encara que aquesta mesura va ser més o menys popular als decenis de 1970 i 1980, mai se va estudiar en detall i ha sigut quasi completament oblidada. N'estudiam alguna de les propietats, caracteritzant els arbres binaris que en prenen el valor màxim com les erugues, i donant-ne un algorisme quasi-lineal per a construir aquells arbres que prenen el seu valor mínim. No obstant, també demostram que aquests darrers gairebé mai no són màximament equilibrats. Acabam el capítol proporcionant fórmules tancades per a la seva esperança sota els models de Yule i Uniforme, a més de la variància dels índexos de Sackin i Cofenètic i el Àrea Total sota el model Uniforme.

En el darrer dels capítols centrals d'aquesta memòria, introduïm un nou índex d'equilibri per a arbres multifurcats: l'índex de Quartets. Trobam els arbres multifurcats i binaris que assoleixen els seus valors extrems: els arbres estrella i eruga en el primer cas, i els arbres màximament equilibrats i eruga en el segon. També donam una recurrència per a calcular el seu valor màxim per a arbres binaris. Així, provam que el seu rang de valors és el més gran d'entre els índexos de equilibri existents a la literatura. A més, calculam la seva esperança i variància sota els models probabilístics d'arbres filogenètics $\beta$ i $\alpha$-$\gamma$. Pel que sabem, aquest és el primer índex topològic d'arbres filogenètics del qual es coneixen els primers moments sota el model $\alpha$-$\gamma$. Finalment, indicam que aquest índex pot ser fàcilment generalitzat a altres famílies de grafs dirigits tot preservant les seves propietats estadístiques.

# Introduction

> Chaque siècle, chaque nation, chaque pays a ses préventions, ses maladies, ses modes, ses penchants, qui les caractérisent, et qui passent et se succedent les uns aux autres [...]
>
> Dom Augustin Calmet, *Traité sur les Apparitions des Esprits et sur les Vampires*, 1751

PHYLOGENETIC ANALYSIS is a practice used in historical sciences such as evolutive biology and historical linguistics, and the *phylogenetic*, or *evolutive*, *tree* is its main device, used to describe a temporal succession of contingent events [9]. Although Lamarck's branching diagram of animals in his *Philosophie Zoologique* (1809) and Darwin's early tree sketch (1837) are oftenly cited as beginning such a tradition, the truth is that the use of trees to represent evolution can be traced back to Justus Georg Schottel's branching table of Germanic languages (1663) (Figure 1) [74, 105].[1] Moreover, if we do not take into account evolutive but taxonomic intentions, the metaphor of the tree already appears in Conrad Gesner's *Historiæ Animalium* (1555), in which a tree is drawn and used to determine and classify species [42].

The study of historical linguistics has traditionally used the phylogenetic analysis of trees as a tool, while other areas such as comparative mythology or archæology have only recently begun doing so. For instance, phylogenetic tools have been recently used to support hypothesis as capital as the Indo-European Steppe Hypothesis [17], according to which the Indo-European peoples were primarily located in the Pontic-Caspian steppe, north of the Black Sea[2], to assess the validity of other linguistic models such as the Indo-Hittite [124] and Indo-Aryan Inner-Outer [15] hypotheses, or to try to re-

[1] It should come as no surprise the fact that the first evolutive depiction of a tree comes from linguistics and not from biology. Indeed, linguists have accepted the notion of evolution at least since Saint Isidore of Seville in the 6th century AD, and historical linguistics as a discipline already existed by the time Darwin was conceived. Another Mediæval source concerning the knowledge of the evolution of languages can be found in the recit of the voyage of William of Rubrouck [101, chap. XXI], in which he explicitly wrote that the language of the Ruthenians, the Poles, the Bohemians and the Sclavons was "the same" as that of the Vandals — he appears to be the first Western author to remark so [101]. Furthermore, he (correctly) stated that Turkish and Cuman had "its source and origin" in Uighur [101, chap. XXVI].

[2] A note is necessary here, since phylogenetic tools and methods had already been used to support the rival hypothesis that Indo-European peoples emerged from Anatolia [52]. The authors of [17] introduce a series of sensible constraints that put the results of [52] into question.

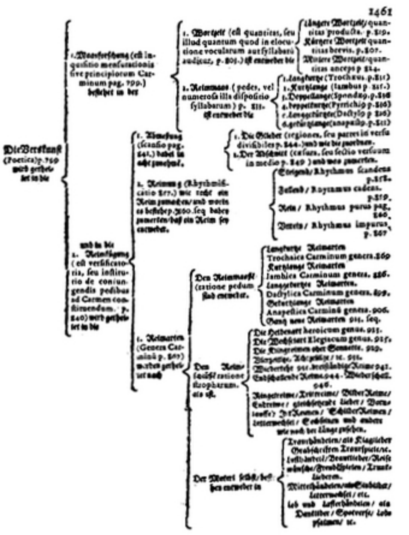


Figure 1: Justus Georg Schottel's branching table of Germanic languages at the end of *Ausführliche Arbeit Von der Teutschen HaubtSprache*, 1663 [105].

construct the Proto-Indo-European language [122]. The analysis of phylogenetic trees has also been fundamental to reconstruct the expansion of linguistic families such as the Austronesian [53, 54] or the Indo-European [17, 52]. Phylogenetic trees can, even, be used to represent the evolution of copies of Mediæval manuscripts according to linguistic characteristics and other criteria [101, Introduction].

In the realm of comparative mythology, the phylogenetic approach is often referred to as *phylomemetics* [96], a noun that draws an analogy between genes and *memes*, understood as the minimum unit of cultural information. Ever since the early observation by Carl von Sydow's —father of the knight that would famously play chess against Death in 1957's celebrated Bergman movie— that myths and stories "adapt themselves to their environment and follow the laws of natural selection" [96], biological metaphors have directed part of the comparative mythology community to the use of statistical and phylogenetic methods. For instance, methodological concept studies can be found in [118, 119], as well as the use of such methods to study the birth, spread and evolution of myths such as the recit of Polyphemus [32] or the Cosmic Hunt [118], two of the few myths appearing on both sides of the Bering strait —thus suggesting a Paleolithical origin. Phylomemetic tools have also been used to reconstruct original recits of myths, also called *ur-forms* [96], such as the ones of Pygmalion, Polyphemus and the Cosmic Hunt [96]. Notice here the analogy between, on the one hand, the reconstruction of Prehistorical languages cited in the previous paragraph and, on the other, the reconstruction of the primitive form of the narration of a myth.

However, it is in the field of evolutive biology that phylogenetic analysis is most commonly used. Nowadays, biologists extract relevant information out of phylogenetic trees in order to try to understand the underlying forces that drive speciation and extinction processes, as well as their effect on macroevolution [47]. The fundamental tenet of phylogenetics is that all species derive from the same source though diverging chains of speciations, thus forming the so-called Tree of Life. In reality, though, it would not be a tree, since hybridizations, horizontal gene transfers and other types of genome recombinations introduce reticulations that cannot be described through a branching structure, thus rather requiring a phylogenetic network [34]. But, phylogenetic trees still are the correct models of the evolution of many sets of species, and even the universal Tree of Life has been claimed to be useful as a model and as a metaphor [80]. The idealistic goal of the phylogenetic community is, thus, to reconstruct such a universal tree [30, 66]. Meanwhile, more modest studies have been pursued, such as the ones concerning the relationship between the phylogenies of hosts and parasites [58, 67, 91] or the study of historical processes responsible for today's geographic distribution of species, also known as phylogeography [29, 128]. Practical applications of the study of phylogenies vary, but we want to emphasize the ones concerning the phylodynamics of epidemics of infectious diseases [56, 73, 123], sadly *à la mode* due to the 2020 global pandemic [44, 121].

Since phylogenetic trees are the standard representation of the joint evolutionary history of groups of species, there is an understandable interest in the development of techniques allowing to measure the imprint these forces exert on them [72, 88, 114]. There are two aspects of a phylogenetic tree on which such traces can be found: its branches' lengths, determined by the timing of speciation events, and its *shape*, or *topology*, determined by the differences in the diversification rates among different subtrees, or clades [38]. However, reconstructing branch lengths associating a robust timeline to a phylogenetic tree in an accurate manner is not easy [35], whereas different phylogenetic reconstruction methods on the same sets of empirical data usually agree on the shape of the reconstructed phylogenetic tree [11, 62, 97]. Therefore, the topology of phylogenetic trees has become the focus of most studies in this regard, whether through the construction of indices quantifying topological features —cf. [46, 88, 107] and all the references on balance indices below— or the distribution of frequencies of small rooted subtrees [78, 104, 109, 126]. On a side note, the shape of phylogenetic trees and networks has also been studied in order to assess biodiversity [26, 33, 45], and we have also made a small contribution in this area [26], but it has been omitted from this memoir due to not being related to its unifying topic, the balance of phylogenetic trees, which we explain anon.

In order to perform the analysis of phylogenetic data, systematists traditionally had to rely on their own expertise. This became more and more difficult as, in the decade of 1960, large amounts of molecular data were beginning to be compiled: the promise was that these data would ultimately help evolutionists to readily reconstruct phylogenies [115]. Nevertheless, methodological concerns rapidly arose around the practice of phylogenetic reconstruction and analysis [110, 115] that led to a certain "methodological anxiety", in the words of Suárez-Díaz and Anaya-Muñoz [115], that seems to be intrinsecal to the assumptions necessary to "reconstruct the past" [110]. Such concerns revolve around the uneasy position of phylogenetic analysis between experimental and historical sciences, at home at neither [9], and its quest for objectivity or, at least, the avoidance of subjectivity [115]. Furthermore, statistical reconstruction methods such as parsimony and maximum likelihood assume underlying hypothesis on the structure of evolution that are contested to this day [110, 115].

One of the solutions proposed in order to get rid of subjectivity bias in the phylogenetic analysis is the quantification and automation of the process leading to the reconstruction or the analysis of phylogenetic trees, all in all leading to a statistical approach to the subject. However, it should be noted that this aproach is not without critique, usually centered around the fact that "the [statistical] tools at hand prevail over methodological commitments" and that "for most practitioners the software packages are literally black boxes, and the automation of procedures obscures the methodological decisions implied in those packages" [115].

Indices asigning a number to a phylogenetic tree are among the first quantitative devices appearing in the literature. For instance, the coherence (CI) and retention (RI) indices [96] of a phylogenetic tree, the first of which computes the percentage of characters in taxons that do not derive from a common ancestor, while the second measures the proportion of characters appearing during evolution that are shared by one or more taxons. These two examples of indices take into account the information associated to the tree during its process of reconstruction, but there are indices that only care about the resulting, underlying branching structure of the phylogenetic tree. These are called *shape indices*, and among them we will be concerned by *balance indices*.

Ever since Yule's early observation [127], in 1922, that taxonomic trees tend to be asymmetric, with most clades being small and only a few of them large at every taxonomic level, the idea of *balance*, understood as a propensity of the direct descendants of any given node to have the same number of descendant leaves, has become one of the most important topological notions in phylogenetics. Consider the trees with 7 leaves depicted in Figure 2: the left-hand side one is known as the *maximally balanced* (bifurcating) tree, and the other as the *caterpillar*. Intuitively, the first one is more balanced than the second one is, even if no formal, definitive definition of this concept is given.

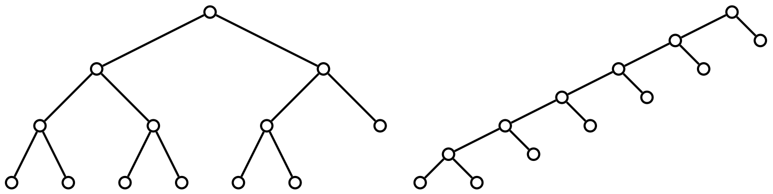


Figure 2: The maximally balanced tree and the caterpillar with 7 leaves, respectively.

Balance indices aim to give a numerical value to trees that is sensitive to this concept, in the sense that it sorts them according to it; i.e., the focus is not on the values the index itself takes upon being applied to different elements of the set of trees, but on the order it induces among them [89]. Hence, the aim is to reflect in such an ordering the propensity that diversification events may present to occur preferentially along some lineages and not others [90, 107]. Several such indices have been proposed in the literature in order to quantify the balance (or, rather often, the imbalance) of a phylogenetic tree: see, for example, [19, 41, 46, 69, 78, 85, 86, 102, 107] and the section "Measures of overall asymmetry" in [38] (pp. 562–563). These measures have been thoroughly used to test evolutionary models [3, 7, 36, 69, 88, 94, 120], compare tree shapes [5, 49, 68], assess biases in the distribution of shapes obtained through different phylogenetic tree reconstruction methods [20, 37, 63, 111, 113], as a tool to discriminate between input parameters in phylogenetic tree simulations [93, 103] or, simply, to describe phylogenies existing in the literature [16, 28, 79, 95].

The oldest, and one of the best-known, balance index is the *Sackin index* (1972) [102, 107]. It is defined as the sum of the depths of all the leaves of a given tree, i.e. given a tree $T \in \mathbf{Tree}$, its Sackin index $S(T)$ is

$$S(T) = \sum_{x \in L(T)} \delta(x),$$

where $L(T)$ is the leaf set of $T$ and $\delta(x)$ is the depth of a given leaf, $x \in L(T)$. Later on in this memoir we shall see that what Sackin actually proposed was a measure of the variation of these depths, not its sum [24]; the measure that is nowadays called the Sackin index was actually introduced by Shao and Sokal in [107]. This index has been the subject of much research. On the one hand, its extreme values have been characterized [107, 39] to be attained by the caterpillar and all trees depth-equivalent to a maximally balanced tree, and so deemed to be "sound" by taking into account what they "ought to be" according to our natural understanding of what balance means. On the other hand, its statistical properties have been thoroughly studied [85, 13, 24, 69]. In this memoir we shall add to these studies the computation of its variance under the Uniform model (on bifurcating phylogenetic trees) [106].

The second oldest, and probably the best-known, balance index existing in the literature is, to the extend of our knowledge, the *Colless index* [19]. It captures intuitively what balance means by being a rather straightforward measure of the equilibrium of each internal node: given a bifurcating tree, it is the sum, over all its internal nodes, of the absolute value of the difference between the numbers of leaves of the two subtrees rooted at them; *c'est à dire*, given a tree $T$, if $u_1, u_2$ denote the two children of an internal node $u$ and $\kappa(u_1), \kappa(u_2)$ the number of leaves of the subtrees rooted at them,

$$C(T) = \sum_{u \in \mathring{V}(T)} |\kappa(u_1) - \kappa(u_2)|, \qquad (1)$$

where $\mathring{V}(T)$ is the set of all internal nodes of $T$. This index, too, has been thoroughly researched, although it remains somewhat more mysterious than the Sackin index does. For one thing, both its expected value and its variance under the Uniform model remain unknown. Its maximum value is reached at the caterpillar [85], as was the case with the Sackin index, but the characterization of its minimum value is much more involved and it was unknown until our work, collected in Chapter 2. Notice that in both this index and Sackin's, the answer is easier found in the case of the least balanced tree than it is in the case of the most balanced one. As it is usually the case, it is the lack of a property rather than its presence that results easier to describe, thus the fact that most of the balance indices present in the literature are, rather, "imbalance" indices, since they measure what Nelson and Holmes [90] called "the propensity of evolutionary events to occur along specific lineages", underlying its absolute dependence of the shape of the tree.

In this report we will work on several such indices. The main problems we attempt to solve are: the full characterization of the trees that attain the minimum Colless index; the study of the original proposal given by Sackin in [102], that has somehow faded into oblivion; and the introduction and study of two new balance indices enjoying several useful features.

To begin with, the Preliminaries are dedicated to results necessary to the understanding of this memoir. Most of them are not new, but we want to remark that the techniques presented in Section 1.4.1 *are*, as they did not previously exist in the literature (as far as we know) and were developped in the course of this investigation. These allow us to solve a family of recurrences (Theorem 1.35) that appear in a natural way in the computation of moments of balance indices under the Uniform model for bifurcating phylogenetic trees, and that we apply later in the computation of the expected value and variance of the Quadratic Colless index (Chapter 3), the expected value of the Variance of depths (Chapter 4), and the variance of the Sackin, the Cophenetic [85] and the Total Area [81] indices (Chapter 4), all of them under the said model. In order to ease the task of the reader, we have opted to present them in the Preliminaries.

As we have already pointed out, in Chapter 2 we will characterize which trees attain the minimum Colless index for any given number of trees $n$ and provide different closed formulæ for this minimum value. This minimum is not reached by a single tree in general, and the characterization of these trees is a bit convoluted. However, in Chapter 3 we shall introduce a new index, which we shall call Quadratic Colless, defined as in Equation (1), only changing the absolute value to a square. This index has better properties than the original Colless index has on every aspect that we have been able to think of: its maxima and minima are attained exactly at the caterpillars and maximally balanced trees, respectively, and the proof of this fact is quite straightforward; and we are even able to compute both its expected value and its variance under the Yule [106] and Uniform models. Furthermore, it has much more discriminative power than both the Sackin and Colless indices have.

Then, in Chapter 4 we shall study the aforementioned original proposal of a balance index given by Sackin in [102]: the variance of the leaves' depths (see Chapter 4). As we shall see, although its maximum value is always reached by the caterpillars, it is ill-suited to be a balance index, since for numbers of leaves larger than 184 its minimum value over bifurcating trees is almost never attained at the maximally balanced trees; we will present two algorithms that compute the trees that attain this value in time $O(n \log_2 n)$.

Finally, we end this memoir by introducing a new balance index that, we believe, has in many ways better properties than those reviewed thus far and those of the Cophenetic index [85]. The Quartet index (Chapter 5) "correctly" classifies as being most and least balanced exactly the maximally balanced trees and the caterpillars, respectively, just as the Quadratic Colless and the Cophenetic indices do, but its discriminatory power has been (empirically) seen to be greater —probably due to the fact that its range of values is an order of magnitude higher than that of the aforementioned indices. But most importantly, we have been able to compute its expected value and its variance not just under the Yule and Uniform models for bifurcating phylogenetic trees, but under Aldous' $\beta$-model for bifurcating phylogenetic trees [2] and Chen-Ford-Winkel's $\alpha$-$\gamma$-model for multifurcating phylogenetic trees [18], both generalizing the Yule and the Uniform models. To our knowledge, this is the first shape index for phylogenetic trees whose first moments under the $\alpha$-$\gamma$-model are known. Furthermore, it is not only defined over multifurcating trees, but it has some natural extensions to wider sets of graphs (such as multilabelled trees or phylogenetic networks), whose properties are yet to be studied in detail.

# Publications

The results reported in this memoir have been published in:

[6] Krzysztof Bartoszek, Tomás M. Coronado, Arnau Mir and Francesc Rosselló. Squaring within the Colless index yields a better balance index. To appear in *Mathematical Biosciences* (2020). A preliminary version is available at https://arxiv.org/abs/2007.14731.

[22] 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. *Journal of Mathematical Biology* 80 (2020), pp. 1993-2054.

[23] Tomás M. Coronado, Arnau Mir and Francesc Rosselló. The Probabilities of Trees and Cladograms under Ford's $\alpha$-model. *Scientific World Journal* (2018), Article ID 1916094; doi: 10.1155/2018/1916094.

[24] 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. *BMC Bioinformatics* 21 (2020), núm. 154.

[25] Tomás M. Coronado, Arnau Mir, Francesc Rosselló and Gabriel Valiente. A balance index for phylogenetic trees based on rooted quartets. *Journal of Mathematical Biology* 79 (2019), pp. 1105-1148.

Chapter I

# Preliminaries

> For if there is any ignorance or indeed any dispute as to what are the facts from which the work opens, it is impossible that what follows should meet with acceptance or credence; but once we produce in our readers a general agreement on this point they will give ear to all the subsequent narrative.
>
> Polybius, *Histories* I, 5, 2nd century BC

Proper stories must start with the beginning. In the preface to the second edition of Newton's *Principia Mathematica* (1713), Roger Cotes wrote that "[t]hose who assume hypothesis as first principles to their speculations [...] may indeed form an ingenious romance, but a romance it will still be." In this chapter we aim to give the basic results and definitions necessary to the understanding of this work, so that "once we produce in our readers a general agreement on this point they will give ear to all the subsequent narrative."

Most results contained in this chapter have already appeared in the literature. Some of them, though, are new (or, at least, we have not been able to find explicit proofs in the literature), and we have opted to include them here for thematic coherence and in order not to burden too much the following chapters. Such results are Lemma 1.13, Theorem 1.19 (which is inspired by [39]), all the results in Section 1.3.1, Lemma 1.24 (again, we have not been able to find a suitable comprehensive reference in the literature), all the computations of probabilities of the $\alpha$ and $\beta$ models as well as the correction to some of Ford's [43] results, the discussion on binary recursive shape indices given in Section

1.3.4, and most importantly the solution of the family of recursive equations given in Theorem 1.35.

Such a chapter must be, almost by definition, fractionary: indeed, for this work will spring out of many sources, coming from the definition of (rooted) trees and its basic properties to the use of hypergeometric series, with a brief intrusion into group theory. We have thus presented these topics in four more or less independent sections. The first one deals with the concept of tree, giving as examples some families that will be later used in this work as well as presenting the concept of labellings of a tree and so of phylogenetic trees; the Newick representation of a tree is also presented. Then, the second section will revolve around the concept of balance indices, and we shall present those that were well known before this investigation began as well as a giving an abstract definition in which they all fit. The third section will be that of the probabilistic models for trees and phylogenetic trees, where we shall present some properties they might present as well as some specific models that will be used thoughout this report, and we will also be concerned with their interplay with balance indices. Finally, we shall spend the last section dealing with hypergeometric series, and we will solve a family of recursive equations that were not (to the best of our knowledge) previously solved, and will be of great use in some of the subsequent chapters.

## 1.1 On trees

This report deals in its integrity with the concept of tree, which we shall define anon. However, in order to do so we first need to recall some other concepts, of which that of *directed graph* is the most important and general. A *directed* (*finite*) *graph* is an ordered pair $G = (V, E)$, where $V$ is a non-empty finite set and $E \subseteq V^2$. We will usually denote $V$ and $E$ by $V(G)$ and $E(G)$, respectively, in order to make reference to the graph $G$. We call $V(G)$ the *set of nodes* of $G$, and $E(G)$ the *set of edges* of $G$. For any pair of nodes $u, v \in V(G)$ and a positive natural number $m \in \mathbb{N}_{\geq 1}$, a *path of length* $m$ from $u$ to $v$ is a collection of $m$ edges $(u_i, v_i) \in E(G)$, $i \in \{1, \dots, m\}$, such that $u_1 = u, v_m = v$ and for all $i \in \{1, \dots, m-1\}$, $v_i = u_{i+1}$. In particular, an edge $(u, v) \in E(G)$ is a path of length $1$ from $u$ to $v$. By convention, we will understand that, for every node $u \in V(G)$, there exists a path of length $0$ connecting $u$ and itself without leaving it. A *cycle* is a path of length $m \geq 1$ such that $u_0 = v_m$.

For any node $u \in V(G)$, we define its *in-degree* and its *out-degree* to be

$$\begin{aligned} \deg_{\text{in}}(u) &:= |\{(u_1, u_2) \in E(G) : u_2 = u\}| \\ \deg_{\text{out}}(u) &:= |\{(u_1, u_2) \in E(G) : u_1 = u\}| \end{aligned}$$

respectively; that is, the in-degree of a node is the number of edges that end in that node, whereas its out-degree is the number of edges that depart from it.

A morphism of directed graphs is an arrow $\varphi : G \to H$ consisting in a pair of maps $\varphi_V : V(G) \to V(H)$ and $\varphi_E : E(G) \to E(H)$ such that the diagram

$$\begin{tikzcd} E(G) \arrow[r, "\varphi_E"] \arrow[d, hook] & E(H) \arrow[d, hook] \\ V(G)^2 \arrow[r, "\varphi_V^2"] & V(H)^2 \end{tikzcd}$$

commutes, in the sense that if $(u, v) \in E(G)$, then $(\varphi_V(u), \varphi_V(v)) \in E(H)$ and $\varphi_E(u, v) = (\varphi_V(u), \varphi_V(v))$. An *isomorphism* of directed graphs is a morphism such that both $\varphi_V$ and $\varphi_E$ are bijective. If two graphs $G$ and $H$ are *isomorphic* (that is, if there exists an isomorphism of directed graphs $\varphi : G \to H$), we shall treat them as if they were equal. *Idem est*, we shall always make the abuse of language of calling two directed graphs *equal* when they are only isomorphic, and hence of speaking of directed graphs, when in fact we mean *isomorphism classes of* directed graphs —indeed, for otherwise we could not properly speak, for instance, of the set of all directed graphs. An *automorphism* of a graph $G$ is an isomorphism $G \to G$. We shall denote by $\mathrm{Aut}(G)$ the set of automorphisms of $G$.

We are now in a position to enonce the main definition of this work. By a *tree*, or *(tree) shape*[1], we understand a *rooted tree without elementary nodes*, i.e., a directed graph $T = (V(T), E(T))$ satisfying the following three properties: $V(T)$ contains exactly one node $\rho_T \in V(T)$ with $\deg_{\mathrm{in}}(\rho_T) = 0$, which we shall always call the *root* of $T$; for every node $u \in V(T)$, there exists a *unique path* from $\rho_T$ to $u$; and, for all nodes $u \in V(T)$, $\deg_{\mathrm{out}}(u)$ is either $0$ or greater than 1. We shall usually drop the subindex $T$ from $\rho_T$ if no ambiguity arises. It is straightforward, from the very definition of tree, to see that a tree does not contain any cycle.

Let $T$ be a tree. We call the nodes of $T$ with out-degree $0$ the *leaves* of the tree, by analogy to the leaves of a (vegetal) tree, and we denote the set formed by all of them by $L(T)$. This defines, *per negationem*, the set of *internal nodes* of $T$, $\mathring{V}(T) = V(T) \setminus L(T)$. We also classify the edges of $T$ into *pendant* and *internal*, depending on whether they end in a leaf or in an internal node, respectively. We shall always make the abuse of language of identifying a tree $(\{\rho\}, \emptyset)$ with the node $\rho$, which is simultaneously the root and the only leaf of the tree.

For any $u, v \in V(T)$, if there exists a path from $u$ to $v$, we shall say that $u$ is an *ancestor* of $v$ and that $v$ is a *descendant* of $u$. Notice that in particular we understand every node to be simultaneously an ancestor and a descendant of itself through the corresponding path of length $0$. In the case of edges, if $(u, v) \in E(T)$, we shall call $u$ the *parent* of $v$ and $v$ a *child* of $u$; we shall denote the set of all children of $u \in V(T)$ by $\mathrm{child}(u)$ and we shall say that two children of the same node are *siblings*. The length of the path from $\rho_T$ to $u \in V(T)$ is called the *depth* of $u$ and denoted by $\delta_T(u)$, and the maximum of these depths is called the *depth* of the tree, denoted by $\delta(T)$: so, $\delta(T) = \max\{\delta_T(u) : u \in V(T)\}$.

The *subtree of $T$ rooted at a node $u$*, $T_u$, is the subtree induced by the set of all the descendants of $u$; i.e., the subtree of $T$ "below" $u$. So, $V(T_u)$ is the set of all descendants of $u$, including $u$ itself, and $E(T_u) = \{(u', v') \in E(T) : u', v' \in V(T_u)\}$. We call the subtrees of $T$ rooted at the children of its root its *maximal pending subtrees*; cf. Figure 1.1.

We denote the set of all (isomorphism classes of) trees by **Tree**. The next three lemmata are well known and easy to prove, and provide useful properties concerning trees.

**Lemma 1.1.** *Let $T \in$ **Tree** be such that $|V(T)| \geq 2$. If $u \in V(T)$ is such that $u \neq \rho_T$, then* $\deg_{\mathrm{in}}(u) = 1$.

[1] We shall usually use the term "shape" when dealing with phylogenetic trees, to mark the difference between a phylogenetic tree and the tree underlying it.

**Lemma 1.2.** *Let $T \in$ **Tree** be a tree and $n$ its number of leaves. Then, $n \geq 1$, and $|V(T)| \leq 2n-1$.*

If $u \in \mathring{V}(T)$ is such that $\mathrm{child}(u) \subseteq L(T)$, we shall call the subtree $T_u$ a *$k$-fan*, where $k = |\mathrm{child}(u)|$, a notation reminiscient of Lady Windermere's alibi. We shall always refer to a 2-fan as a *cherry*. The *depth of a cherry* (or of a $k$-fan) will always be the depth of its leaves.

**Lemma 1.3.** *Let $T \in$ **Tree** be such that $|V(T)| \geq 2$. Then, it contains some $k$-fan, for some $k \geq 2$.*

For any tree $T \in$ **Tree** we define $\Delta(T)$ to be the multiset of the leaves' depths; that is, $\Delta(T) = \{\delta_T(u) : u \in L(T)\}$; notice that the condition of it being a multiset is necessary, since for any number of leaves $n \geq 2$, $T$ has, by Lemma 1.3, at least one $k$-fan for some $k \geq 2$. If two trees $T_1$ and $T_2$ are such that $\Delta(T_1) = \Delta(T_2)$, they are said to be *depth-equivalent*.

Given a subset of leaves $X \subseteq L(T)$, the *lowest common ancestor of $X$*, $\mathrm{lca}(X)$, is the node $u \in V(T)$ of maximum depth such that $X \subseteq L(T_u)$; i.e., their "most recent ancestor" in the sense that every common ancestor of all leaves in $X$ is also an ancestor of $\mathrm{lca}(X)$.

There is another way to define subtrees of a given tree $T$. For every $X \subseteq L(T)$, the *subtree of $T$ that induces $X$* is the tree obtained from $T_{\mathrm{lca}(X)}$ by keeping only the nodes, and edges, in paths from $\mathrm{lca}(X)$ to the leaves in $X$, and then recursively suppressing all *elementary* nodes, that is, all nodes with out-degree 1, as follows: if $u$ is such a node and $v, w \in V(T)$ are such that $(v,u),(u,w) \in E(T_{\mathrm{lca}(X)})$, then we replace the node $u$, along with those two edges, by a single edge $(v,w)$ (note that, by the definition of $\mathrm{lca}(X)$, the root of $T_{\mathrm{lca}(X)}$ cannot be elementary). We call such a tree $T(X)$.

Notice that in the statement of Lemma 1.2 it is said that the maximum number of nodes that a tree with $n$ leaves can have is $2n-1$. If we had imposed that $\deg_{\mathrm{out}}(u) \in \{0,2\}$ for all $u \in V(T)$, then it is an easy induction exercise to see that $|V(T)| = 2n-1$. This kind of tree is called *bifurcating*, or *binary*, and they are of the utmost importance in the whole of this work. Therefore, we emphasize the following result.

**Corollary 1.4.** *Let $T$ be a bifurcating tree with $n$ leaves. Then, $|V(T)| = 2n-1$.*

We will denote by $\mathbf{Tree}_n$ the set of (isomorphism classes of) trees with $n$ leaves, and by $\mathbf{BinTree}_n \subseteq \mathbf{Tree}_n$ the set of those of them that are bifurcating (notice that $\mathbf{BinTree}_n = \mathbf{Tree}_n$ if, and only if, $n \in \{1,2\}$).

Let $T_1, \ldots, T_k$ be $k$ trees, for some $k \geq 2$, such that $|L(T_i)| = n_i$ for every $i \in \{1, \ldots, k\}$. We define

$$T_1 * \cdots * T_k \in \mathbf{Tree}_{n_1+\cdots+n_k}$$

as the only (up to isomorphism) tree such that the trees $T_1, \ldots, T_k$ are its maximal pending subtrees: see Figure 1.1. In other words, $T_1 * \cdots * T_k$ is obtained by taking (disjoint isomorphic copies of) the trees $T_1, \ldots, T_k$, a new node $\rho$, and connecting $\rho$ to the roots $\rho_{T_i}$ of these trees through new edges $(\rho, \rho_{T_i})$. We shall call this operation the *root join* of the trees $T_1, \ldots, T_k$. Notice that $*$ is not associative, but it is commutative.

We end with two useful lemmata concerning isomorphisms of trees.

**Lemma 1.5.** *Let $\varphi : T_1 \to T_2$ be an isomorphism of trees. Then, $\varphi(\rho_{T_1}) = \rho_{T_2}$.*

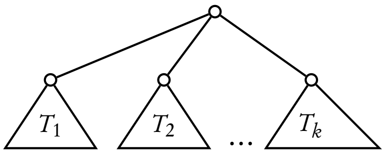


Figure 1.1: The tree $T_1 * \cdots * T_k$, with maximal pending subtrees $T_1, \ldots, T_k$.

*Proof.* Let $\rho_{T_1}$ be the root of $T_1$. If $\varphi(\rho_{T_1}) \neq \rho_{T_2}$, then $\deg_{\mathrm{in}}(\varphi(\rho_{T_1})) \neq 0$ and in particular there would exist an edge $(u, \varphi(\rho_{T_1})) \in E(T_2)$. But then, if $u' \in V(T_1)$ is the preimage of $u$ under $\varphi$, then, since $\varphi_E : E(T_1) \to E(T_2)$ is bijective, there should exist an edge $(u', \rho_{T_1})$, which would contradict the assumption that $\deg_{\mathrm{in}}(\rho_{T_1}) = 0$. □

**Lemma 1.6.** *Let $T \in \mathbf{Tree}_n$ and $\varphi_1, \varphi_2 \in \mathrm{Aut}(T)$ such that $\varphi_1|_{L(T)} = \varphi_2|_{L(T)}$. Then, $\varphi_1 = \varphi_2$.*

*Proof.* We proceed by induction on the number of leaves, $n$. The result obviously holds for $n = 1$, and so our base case is proved. We shall now suppose it to be true up to $n$ leaves, $n \geq 2$.

Let $T \in \mathbf{Tree}_{n+1}$, and $\varphi_1, \varphi_2 \in \mathrm{Aut}(T)$ be such that $\varphi_1(x) = \varphi_2(x)$ for all $x \in L(T)$. Then, since $\varphi_i$ is an isomorphism for $i \in \{1, 2\}$, we can deduce that $\varphi_1(u) = \varphi_2(u)$ for all $u \in \mathring{V}(T)$ such that $u$ has a leaf child.

Let us now consider the tree $T' \in \mathbf{Tree}_m$ for some $m \in \mathbb{N}$ defined as the tree derived from $T$ by erasing one of its $k$-fans, for some $k \in \mathbb{N}_{\geq 2}$, thus making the root of the fan, say $v$, a leaf in $T'$. Now consider the automorphisms $\varphi_1|_{T'}, \varphi_2|_{T'} \in \mathrm{Aut}(T')$, and call them $\varphi_1'$ and $\varphi_2'$, respectively. As we have already discussed, $\varphi_1'(v) = \varphi_2'(v)$ and, since with the exception of the considered $k$-fan in $T$, the leaf sets remain the same for $T$ and $T'$, we have that $\varphi_1'|_{L(T')} = \varphi_2'|_{L(T')}$. But then, by the induction hypothesis, that entails that $\varphi_1' = \varphi_2'$.

We have almost finished, since we have shown that $\varphi_1(u) = \varphi_2(u)$ for all $u \in V(T') \subsetneq V(T)$. But, as $V(T) \setminus V(T') \subseteq L(T)$, we conclude that $\varphi_1 = \varphi_2$ by means of the hypothesis in the statement of this proposition. □

### 1.1.1 Three families of trees

For every $n \in \mathbb{N}_{\geq 4}$, there are three different trees in $\mathbf{Tree}_n$ that deserve close attention. These are the *star*, the *maximally balanced (bifurcating) tree* and the *caterpillar*.

#### Stars

In $\mathbf{Tree}_n$, with $n \geq 2$, the *star* $T_n^{\mathrm{star}}$ is the tree consisting of $n + 1$ nodes: its root and $n$ leaves pending from it. Equivalently, the star with $n$ leaves is the only tree with $n$ leaves such that $|\mathring{V}(T)| = 1$. By convention, we shall always consider the unique tree in $\mathbf{Tree}_1$ to be a star.

Stars have the interesting property that they are the most symmetric family of trees, in the sense that they have the maximum number of automorphisms for any given number of leaves $n$, as the following result establishes.

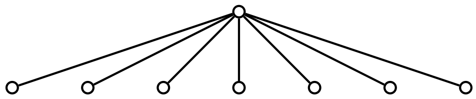


Figure 1.2: The star with seven leaves, $T_7^{\text{star}}$.

**Theorem 1.7.** *Let $T \in \textbf{Tree}_n$ a tree with n leaves. Then,*

$$\operatorname{Aut}(T) \leq n!$$

*and the equality is reached if, and only if, $T = T_n^{\text{star}}$.*

*Proof.* This result is a direct consequence of Lemma 1.13 and Theorem 1.14 below. □

### Maximally balanced (bifurcating) trees

When we restrict ourselves to bifurcating trees, the question on which are the most symmetric trees is not as easy as it was in **Tree**, since the star is not a bifurcating tree for $n \geq 3$ leaves. In any case, the number of automorphisms as a shape index presents the problem of not having a great discriminatory power: fixed $n \in \mathbb{N}$, the number of autormorphisms of a bifurcating tree $T \in \textbf{BinTree}_n$ can take at most $n - 1$ values —indeed, by Lemma 1.13 and Theorem 1.14 below, this number only depends on the number of *symmetry nodes* of $T$, those internal nodes such that both subtrees rooted at their children are isomorphic.

To circumvent this drawback, the alternative concept of "balance" of a tree is introduced and expressed in some measures, but this shall be discussed soon enough. In this section, it suffices to say that the family of trees now presented were considered by Shao and Sokal [107] to be "the most balanced trees", and they are effectively classified as "most balanced" by most balance indices. In this regard, it totally agrees with our natural intuition. In fact, as we shall discuss, it is a common practice to determine the "validity" of a balance index based on whether it classifies the maximally balanced trees as most balanced and the caterpillars —which we shall define anon— as least balanced.

To define a maximally balanced tree we need first to define the *balance* of an internal node in a bifurcating tree. For all $n \in \mathbb{N}$ and $u \in V(T)$ in any $T \in \textbf{BinTree}_n$, let $\kappa_T(u)$ be its number of descendant leaves; i.e., $\kappa_T(u) = |L(T_u)|$. Then, we define the *balance* of an internal node $u \in \mathring{V}(T)$ as

$$\operatorname{bal}_T(u) = |\kappa_T(u_1) - \kappa_T(u_2)|,$$

where $u_1, u_2 \in V(T)$ are the children of $u$. We shall drop the subindex $T$ from $\kappa_T$ and $\operatorname{bal}_T$ whenever doing so adds no ambiguity to our arguments. Then, we shall call an internal node $u \in \mathring{V}(T)$ *balanced* when $\operatorname{bal}(u) \in \{0, 1\}$. In other words, an internal node $u$ of a bifurcating tree is balanced when its two children have $\lceil \kappa(u)/2 \rceil$ and $\lfloor \kappa(u)/2 \rfloor$ descendant leaves, respectively.

A *maximally balanced* tree is then a bifurcating tree all of whose internal nodes are balanced. By convention, we shall say that the only tree in $\textbf{BinTree}_1$ is maximally balanced. There is an interesting recursive construction, given by the following result, which entails that for every given number $n$ of leaves there exists one, and only one, maximally balanced tree in $\mathbf{BinTree}_n$.

**Theorem 1.8.** *Let $n \in \mathbb{N}_{\geq 2}$. For every $T \in \mathbf{BinTree}_n$, $T$ is maximally balanced if, and only if, its maximal pending subtrees are maximally balanced with $\lceil n/2 \rceil$ and $\lfloor n/2 \rfloor$ leaves, respectively.*

*Proof.* Let $T \in \mathbf{BinTree}_n$ and let $u_1, u_2$ be the children of its root $\rho$, with $\kappa(u_1) \geq \kappa(u_2)$.

Assume first that $T$ is maximally balanced. Then, since $\rho$ is balanced, $\kappa_T(u_1) = \lceil n/2 \rceil$ and $\kappa_T(u_2) = \lfloor n/2 \rfloor$. Let now $v$ be an internal node of some rooted subtree $T_{u_i}$, $i \in \{1, 2\}$. Since $\mathrm{bal}_{T_{u_i}}(v) = \mathrm{bal}_T(v)$ and $v$ is balanced in $T$, it is also balanced in the subtree $T_{u_i}$. Therefore, $T_{u_1}$ and $T_{u_2}$ are maximally balanced. This proves the "only if" implication.

Conversely, assume that $\kappa_T(u_1) = \lceil n/2 \rceil$ and $\kappa_T(u_2) = \lfloor n/2 \rfloor$ and that both $T_{u_1}$ and $T_{u_2}$ are maximally balanced. Then, $\rho$ is balanced. Let now $v$ be an internal node of $T$ other than the root. Since $\mathring{V}(T) = \mathring{V}(T_{u_1}) \cup \mathring{V}(T_{u_2}) \cup \{\rho\}$, the node $v$ will belong to some $\mathring{V}(T_{u_i})$ and then $\mathrm{bal}_T(v) = \mathrm{bal}_{T_{u_i}}(v) \in \{0, 1\}$. This implies that all internal nodes of $T$ that are not the root are also balanced. Therefore, $T$ is maximally balanced. □

**Corollary 1.9.** *For every $n \in \mathbb{N}_{\geq 1}$, there is one, and only one, maximally balanced tree in* $\mathbf{BinTree}_n$.

*Proof.* We proceed by induction on the number of leaves, $n$. If $n \in \{1, 2, 3\}$ the result holds trivially, because the cardinality of $\mathbf{BinTree}_n$ is 1 and the only tree in it is maximally balanced. Hence, assume now that $n \geq 4$, and suppose that the result holds up to $n - 1$ leaves. Let us denote by $T^{\mathrm{bal}}_{\lfloor n/2 \rfloor}$ and $T^{\mathrm{bal}}_{\lceil n/2 \rceil}$ the maximally balanced trees with $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ leaves, which exist and are unique by the induction hypothesis. Then, by the previous result, on the one hand $T^{\mathrm{bal}}_{\lceil n/2 \rceil} * T^{\mathrm{bal}}_{\lfloor n/2 \rfloor} \in \mathbf{BinTree}_n$ is maximally balanced, and, on the other hand, if $T \in \mathbf{BinTree}_n$ is maximally balanced and if $u_1, u_2$ are the children of its root, with $\kappa_T(u_1) \geq \kappa_T(u_2)$, then $T_{u_1} = T^{\mathrm{bal}}_{\lceil n/2 \rceil}$ and $T_{u_2} = T^{\mathrm{bal}}_{\lfloor n/2 \rfloor}$ and hence $T = T^{\mathrm{bal}}_{\lceil n/2 \rceil} * T^{\mathrm{bal}}_{\lfloor n/2 \rfloor}$, which proves the uniqueness of the maximally balanced tree in $\mathbf{BinTree}_n$. □

We shall denote henceforth by $T^{\mathrm{bal}}_n$ the maximally balanced tree with $n$ leaves. When $n$ is a power of 2, Theorem 1.8 implies that the subtrees rooted at each pair of children of an internal node of $T^{\mathrm{bal}}_n$ are isomorphic, and in this case we also call $T^{\mathrm{bal}}_n$ the *fully symmetric* bifurcating tree with $n$ leaves.

In Corollary 1.16, we shall see that the number of automorphisms of a bifurcating tree is always grows exponentially with its number of symmetry nodes. Now, it would be natural, and indeed beautiful, if the maximally balanced trees represented the upper bound of the number of automorphisms for bifurcating trees with a fixed number of leaves. It is not so, as the counterexample depicted in Figure 1.4 illustrates.

Let $T^{\mathrm{bal}}_{24} \in \mathbf{BinTree}_{24}$ be the maximally balanced tree with 24 leaves, depicted in Figure 1.4. It is clear, then, that $s(T^{\mathrm{bal}}_{24}) = 15$. Now, let $T^{\mathrm{gfb}}_{24}$ be the tree depicted in the next figure. We can now count that $s(T^{\mathrm{gfb}}_{24}) = 22$, and thus that the educated guess that, for every $T \in \mathbf{BinTree}_n$, $s(T^{\mathrm{bal}}_n) > s(T)$ is false —and so, the number of automorphisms of a maximally balanced tree is not always bigger than that of other bifurcating trees.

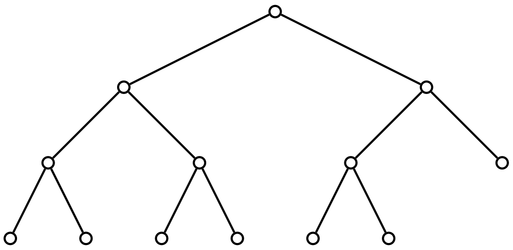


Figure 1.3: The maximally balanced tree with 7 leaves, $T_7^{\mathrm{bal}}$.

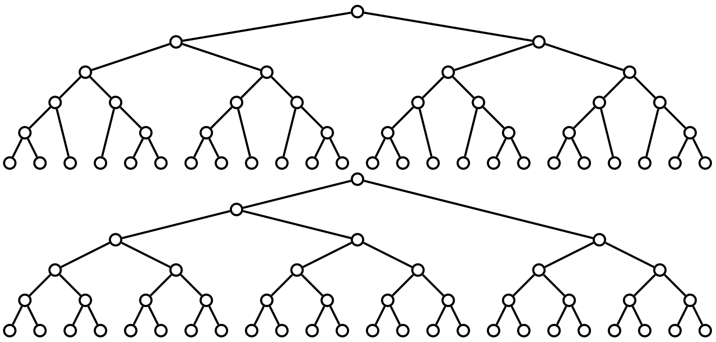


Figure 1.4: Two bifurcating trees with 24 leaves, $T_{24}^{\mathrm{bal}}$ and $T_{24}^{\mathrm{gfb}}$, respectively.

### Caterpillars

In contrast to the maximally balanced trees, caterpillars have been considered, already in the early paper by Sackin [102], to be the most imbalanced of all trees, and they are indeed the least symmetrical: they have, as we will show below, only two automorphisms no matter their number of leaves (provided it is, of course, greater than 1). But this is indeed of great importance since, as happened with the maximally balanced trees, a "good" balance index *ought to* classify these trees as being the least balanced of all trees.

A *caterpillar* of $n$ leaves is a bifurcating tree such that every internal node in it has a leaf for child; we shall consider the only tree in $\mathbf{Tree}_1$ to be a caterpillar. Equivalently, for $n \geq 2$, a *caterpillar* of $n$ leaves is a bifurcating tree with only one cherry. The equivalence between both descriptions is clear: if a bifurcating tree $T$ has two cherries, the lowest common ancestor of their roots cannot have a leaf child, and conversely, if the two children of $u \in \mathring{V}(T)$ are internal nodes, both subtrees rooted at them will contain at least one cherry, which gives at least two different cherries.

The definition of caterpillars as those bifurcating trees all whose internal nodes have a leaf child allows for their following recursive construction: a caterpillar with $n$ leaves is the root join of a leaf and a caterpillar of $n-1$ leaves. And then this construction implies, through an argument *à la* Corollary 1.9, that, for every $n \in \mathbb{N}_{\geq 1}$, there is only one caterpillar in $\mathbf{Tree}_n$, which we shall denote by $T_n^{\mathrm{cat}}$.

**Theorem 1.10.** *Let $T \in \mathbf{Tree}_n$ a tree with $n$ leaves. Then,* $\mathrm{Aut}(T) \geq 2$ *and the equality is reached if, and only if, $T = T_n^{\mathrm{cat}}$.*

*Proof.* This is a direct consequence of Lemma 1.13, Theorem 1.14, and the fact that the caterpillars are the only trees with just one cherry and no other $k$-fan, and hence the only trees with only one symmetry node that is moreover of out-degree 2. □

**Remark 1.11.** Notice that, although the caterpillar is considered to be the least symmetrical of all trees with $n$ leaves, it is actually bifurcating —thus, in particular, it is also the least symmetrical of all bifurcating trees with $n$ leaves.

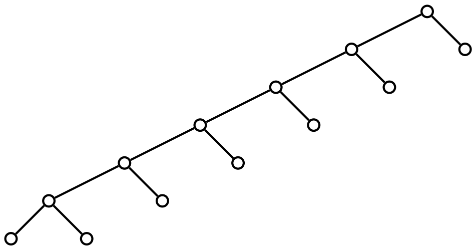


Figure 1.5: The caterpillar with seven leaves, $T_7^{\text{cat}}$.

### 1.1.2 Labels

In order to model evolutionary processes, trees usually have their leaves labelled with what we call, in general, Operational Phylogenetic Units (OPUs). These usually represent species in the biological sense, as the use of phylogenetic representations is mainly spread among the Evolutionary Biology community; but they can also represent genes [55, 64], languages or language characteristics [15, 17, 52, 53, 54, 124], cultural aspects [31], myth versions [118, 119] and even Internet memes [61].

Mathematically, let $T = (V(T), E(T))$ be a tree, and $L(T)$ be its set of leaves. Then, for some set $\Lambda$, a $\Lambda$-*labelling* of $T$ is just a function $\lambda : L(T) \to \Lambda$. A pair $(T, \lambda)$ will be called generically a *multi-labelled tree on the set* $\Lambda$, and a *phylogenetic tree* on $\Lambda$ whenever $\lambda$ is bijective, in which case $|\Lambda| = |L(T)|$. We will say that a multi-labelled tree $(T, \lambda)$ is *bifurcating* when $T$ is so. In a phylogenetic tree, we shall usually make the abuse of language of identifying a leaf and its label.

**Remark 1.12.** We will always consider the elements of the set of labels $\Lambda$ to have a "canonical" representation as strings, since they are labels. Thus, if $\Lambda \subseteq \mathbb{N}$, we shall consider the element $3 \in \Lambda$ to be also $3 \in$ **String**.

We shall distinguish two types of isomorphisms for multi-labelled trees. We begin with the "strict sense" isomorphisms. Given two multi-labelled trees $(T_1, \lambda_1)$ and $(T_2, \lambda_2)$ on a set $\Lambda$, an *isomorphism* between them is an isomorphism of trees $\varphi : T_1 \to T_2$ such that the following diagram commutes:

$$\begin{tikzcd}
L(T_1) \arrow[r, "\varphi_V|_L"] \arrow[d, "\lambda_1"'] & L(T_2) \arrow[d, "\lambda_2"] \\
\Lambda \arrow[r, "\mathrm{id}_\Lambda"] & \Lambda
\end{tikzcd}$$

In other words, an isomorphism of multi-labelled trees is an isomorphism of trees that preserves and respects the trees' labelling. When $(T_1, \lambda_1)$ and $(T_2, \lambda_2)$ are phylogenetic trees, this gives rise to the usual notion of *isomorphism of phylogenetic trees*. We shall make the abuse of language of saying that two multi-labelled, or phylogenetic, trees are *equal* when they are actually only isomorphic in this sense.

We will denote the sets of (isomorphism classes of ) phylogenetic and multi-labelled trees on a set of labels $\Lambda$ by $\mathbf{PhyloTree}(\Lambda)$ and $\mathbf{MulTree}(\Lambda)$, respectively, and we will denote by $\mathbf{BinPhyloTree}(\Lambda)$ and $\mathbf{BinMulTree}(\Lambda)$ the corresponding subsets of bifurcating such trees. More often than not, whenever we are working with a fixed number of leaves $n$, we just consider $\Lambda$ to be $[n] = \{1, \ldots, n\}$, and then we shall just write $\mathbf{PhyloTree}_n$, $\mathbf{BinTree}_n$, $\mathbf{MulTree}_n$ and $\mathbf{BinMulTree}_n$ to denote the respective sets of phylogenetic trees, bifurcating phylogenetic trees, multi-labelled trees, and bifurcating multi-labelled trees. The cardinality of $\mathbf{BinPhyloTree}_n$ is well known by a theorem of Schröder (1870),

$$|\mathbf{BinPhyloTree}_n| = (2n-3)!! = (2n-3)\cdot(2n-5)\cdots 3\cdot 1 = \frac{(2n-2)!}{(n-1)!2^{n-1}}.$$

We have a projection $\pi_1 : \mathbf{MulTree}_n \to \mathbf{Tree}_n$ defined by $\pi_1(T, \lambda) = T$. We shall say that $(T, \lambda) \in \mathbf{MulTree}_n$ is a star, a maximally balanced tree, or a caterpillar whenever $T$ is so.

Now, the second notion of isomorphism of multi-labelled trees that we shall use in this report is that of a *shape-isomorphism*, or *relabelling*. In this case, it is between multi-labelled trees on possibly different sets of labels $\Lambda_1$ and $\Lambda_2$: let $(T_1, \lambda_1), (T_2, \lambda_2)$ be two multilabelled trees, a pair $(\varphi, \varphi_\Lambda) : (T_1, \lambda_1) \to (T_2, \lambda_2)$ such that $\varphi : T_1 \to T_2$ is an isomorphism of trees and $\varphi_\Lambda : \Lambda_1 \to \Lambda_2$ is an *injective map* such that the diagram

$$\begin{tikzcd}
L(T_1) \arrow[r, "\varphi_V|_L"] \arrow[d, "\lambda_1"'] & L(T_2) \arrow[d, "\lambda_2"] \\
\Lambda_1 \arrow[r, "\varphi_\Lambda"] & \Lambda_2
\end{tikzcd}$$

commutes. In other words, a relabelling of multilabelled trees is an isomorphism of trees such that a pair of leaves in $T_1$ have the same label if, and only if, their images in $T_2$ have the same label, but it need not preserve the actual labels. Notice that when $|\Lambda_1| = |\Lambda_2|$, the map $\varphi_\Lambda$ of a shape-isomorphism will have to be a bijection. When $(T_1, \lambda_1)$ and $(T_2, \lambda_2)$ are phylogenetic trees, a relabelling between them is simply an isomorphism of their underlying shapes (Lemma 1.13).

We shall call a shape-isomorphism class of multilabelled trees a *multilabelled tree shape*, and we shall always say that two shape-isomorphic multilabelled trees *have the same multilabelled shape*, and denote by $\mathbf{MulShTree}_n$ and $\mathbf{BinMulShTree}_n$ the sets of multilabelled tree shapes and of bifurcating multilabelled tree shapes on $[n]$, respectively.

Notice that we have a projection $\pi_1 : \mathbf{MulShTree}_n \to \mathbf{Tree}_n$ defined by $\pi_1(T, \lambda) = T$,[2] that returns the underlying shape of a multilabelled tree shape, and the map $\pi^*$ :

[2] In rigor, we should have written $\pi_1([(T, \lambda)]) = T$, but we will, here and henceforth, just identify the class with its representant. We may, in the following, also commit the abuse of language of writing $T$ when we mean a specific phylogenetic tree $(T, \lambda)$ but we do not care about the specific $\lambda$ being chosen.

**MulTree**$_n$ $\longrightarrow$ **MulShTree**$_n$ that sends each multilabelled tree to its multilabelled tree shape:

$$\begin{array}{ccccc} \mathbf{MulTree}_n & \overset{\pi^*}{\longrightarrow} & \mathbf{MulShTree}_n & \overset{\pi_1}{\longrightarrow} & \mathbf{Tree}_n \\ (T,\lambda) & \mapsto & [(T,\lambda)] & \mapsto & T \end{array}$$

**Example:**

Consider the three multilabelled trees in Figure 1.6, of which $T_1$ is phylogenetic.

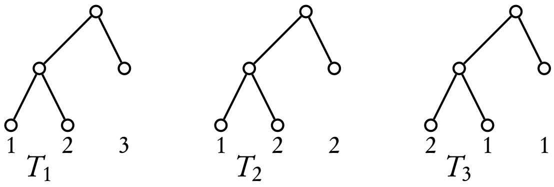


Figure 1.6: Three multilabelled trees with three leaves.

None of these trees are isomorphic in the strict sense, whereas $T_2$ and $T_3$ are shape-isomorphic. All three multilabelled trees have the same shape, but only $T_2$ and $T_3$ have the same multilabelled shape.

The concept of the root join of multilabelled trees, and indeed phylogenetic trees, generalizes easily, bearing in mind that if two phylogenetic trees share their label set, their root join may not be a phylogenetic tree. The shape of the root join of several multilabelled trees is the root join of their shapes. However, the root join of multilabelled tree shapes is not well defined, but we shall still use it as an abbreviation: that is, given a multilabelled tree shape $T \in$ **MulShTree**$_n$, we shall write it as $T = T_1 * \cdots * T_m$ to mean that $T_1, \ldots, T_m$ are the multilabelled tree shapes rooted at the children of its root.

### On relabellings of phylogenetic trees

We are now interested in the problem of counting how many different relabellings does a phylogenetic tree have. Suppose given a phylogenetic tree $(T,\lambda) \in$ **PhyloTree**$_n$; our goal is to determine how many phylogenetic trees are there in **PhyloTree**$_n$ such that they are different from $(T,\lambda)$ but share its shape. Notice that this is not as easy as simply changing $\lambda$, since there is a number of different labellings that give rise to the same phylogenetic tree. Think, for instance, about interchanging the labels 1 and 2 in tree $T_1$, as depicted by Figure 1.6. In order to count this magnitude, we shall be concerned about the problem of counting how many bijective maps $\sigma : [n] \longrightarrow [n]$ leave $(T,\lambda)$ "unchanged" —that is, send it to an isomorphic (under the strict interpretation) phylogenetic tree. Since the maps $\sigma$ are bijective, we can consider them to be members of the symmetric group $\mathfrak{S}_n$. Consider now the action of groups

$$\begin{array}{ccl} \mathfrak{S}_n & \longrightarrow & \mathrm{Aut}(\mathbf{PhyloTree}_n) \\ \sigma & \mapsto & \begin{array}{rcl} {}^{\sigma}\cdot : \mathbf{PhyloTree}_n & \longrightarrow & \mathbf{PhyloTree}_n \\ (T,\lambda) & \mapsto & (T, \sigma \circ \lambda) \end{array} \end{array}$$

where by Aut($\mathbf{PhyloTree}_n$) we mean, here, the set of bijections from $\mathbf{PhyloTree}_n$ to itself. The problem of finding how many different phylogenetic trees share the same shape can be re-stated as computing the cardinality of orb$(T, \lambda)$ under this action of groups. It is straightforward to check that, given a shape $T \in \mathbf{Tree}_n$, this cardinality does not depend on the labelling map $\lambda$, since changing of $\lambda$ (which, we recall, is bijective by definition of phylogenetic tree) induces a bijection between these orbits. We shall call henceforth this cardinality the *number of relabellings* of the shape $T \in \mathbf{Tree}_n$, and denote it by $\phi(T)$:

$$\phi(T) = |\mathrm{orb}(T, \lambda)|,$$

for some labelling $\lambda$. Now, by the Burnside Lemma, we know that

$$|\mathrm{orb}(T, \lambda)| = \frac{|\mathfrak{S}_n|}{|\mathrm{stab}(T, \lambda)|} = \frac{n!}{|\mathrm{stab}(T, \lambda)|},$$

where the elements of stab$(T, \lambda)$ are the maps $\sigma$ such that $(T, \lambda)$ is isomorphic to $(T, \sigma \circ \lambda)$.

**Lemma 1.13.** *Let* $(T, \lambda) \in \mathbf{PhyloTree}(\Lambda)$*. Then, there exists an isomorphism of groups* $\iota : \mathrm{Aut}(T) \to \mathrm{stab}(T, \lambda)$*.*

*Proof.* We shall provide such an isomorphism by the rule $\iota : \mathrm{Aut}(T) \to \mathrm{stab}(T, \lambda)$ defined by $\varphi \mapsto \lambda \circ \varphi|_{L(T)} \circ \lambda^{-1}$. Notice that, as $\varphi$ is an isomorphism, $\varphi|_{L(T)}$ is bijective and, since $\lambda$ is also bijective, $\lambda \circ \varphi|_{L(T)} \circ \lambda^{-1}$ is, too. Now, $(\lambda \circ \varphi|_{L(T)} \circ \lambda^{-1}) \circ \lambda = \lambda \circ \varphi|_{L(T)}$ and therefore $\varphi : (T, \iota(\varphi) \circ \lambda) \to (T, \lambda)$ is an isomorphism of phylogenetic trees and, hence, $\iota(\varphi) \in \mathrm{stab}(T, \lambda)$. The fact of $\iota$ being a morphism of groups is straightforward, as well as its injectivity, because $\lambda$ is bijective and, by Lemma 1.6, an automorphism is uniquely determined by its behaviour on the leaves of the tree. It remains the surjectivity to be checked. Let $\sigma \in \mathrm{stab}(T, \lambda)$, and let $\varphi : (T, \sigma \circ \lambda) \to (T, \lambda)$ be an isomorphism of phylogenetic trees —that exists, since $\sigma \in \mathrm{stab}(T, \lambda)$. Then, $\varphi : T \to T$ is an automorphism such that $\sigma \circ \lambda = \lambda \circ \varphi|_{L(T)}$, and in particular $\sigma = \lambda \circ \varphi|_{L(T)} \circ \lambda^{-1} = \iota(\varphi)$. Therefore, $\iota$ is an isomorphism. □

Next theorem, which is Proposition 2.4.2 in [106], computes the cardinality of stab$(T, \lambda)$ (and so, of Aut$(T)$).

**Theorem 1.14.** *Let* $(T, \lambda) \in \mathbf{PhyloTree}_n$ *be a phylogenetic tree with n leaves. For each internal node* $u \in \mathring{V}(T)$*, let* $D(u)$ *denote the collection of the phylogenetic subtrees rooted at the children of* $u$*. Consider the equivalence relation on* $D(u)$ *in which* $(T_1, \lambda_1), (T_2, \lambda_2) \in D(u)$ *are related if, and only if, they have the same shape. Let* $n_1(u), n_2(u), \ldots$ *denote the cardinalities of the resulting equivalence classes. Then,*

$$|\mathrm{Aut}(T)| = |\mathrm{stab}(T, \lambda)| = \prod_{u \in \mathring{V}(T)} \prod_i n_i(u)!$$

*Therefore,*

$$\phi(T) = \frac{n!}{\prod_{u \in \mathring{V}(T)} \prod_i n_i(u)!}.$$

**Remark 1.15.** Notice that the equivalence relation used in the statement of the last theorem is the kernel of the map $\pi_1|_{D(u)} : D(u) \to \mathbf{Tree}$ sending the phylogenetic trees in $D(u)$ to their shapes.

In the bifurcating case, for every $u \in \mathring{V}(T)$, the product $\prod_i n_i(u)!$ is 2 if $u$ is a *symmetry node*, that is, if the subtrees rooted at its children have the same shape, and 1 otherwise. This implies the following corollary.

**Corollary 1.16.** *Let $T \in \mathbf{BinTree}_n$ be a bifurcating tree and let $s(T)$ be its number of symmetry nodes. Then,*

$$|\mathrm{Aut}(T)| = 2^{s(T)}, \quad \phi(T) = \frac{n!}{2^{s(T)}}.$$

Notice now that, by Lemma 1.13, in order to compute the cardinality of $\mathrm{stab}(T, \lambda)$, for any given phylogenetic tree $(T, \lambda)$, it suffices to compute that of $\mathrm{Aut}(T)$, and *vice versa*; but now, due to Theorem 1.14, that amounts to counting the number of shape isomorphism classes of the maximal pending subtrees rooted at each internal node in $T$.

> **Example:**
> Consider the star with $n$ leaves, $T_n^{\mathrm{star}} \in \mathbf{Tree}_n$. It has only one internal node, $\rho$, and $n$ maximal pending subtrees, all of them isomorphic to a leaf. Thus, $n_1(\rho) = n$, and therefore
>
> $$\mathrm{Aut}(T_n^{\mathrm{star}}) = n! \text{ and } \phi(T) = 1.$$
>
> For any other tree $T \in \mathbf{Tree}_n$, its number of automorphisms must be strictly less than $n!$. Indeed, for $|\mathrm{Aut}(T)| = |\mathrm{stab}(T, \lambda)| \leq |\mathfrak{S}_n| = n!$. Now, the fact that the bound is strict derives from the fact that, for any collection of positive natural numbers $k_1, \ldots, k_m$ such that $\sum_{i=1}^m = n$, $n! \geq \prod_{i=1}^m k_i!$, with equality if, and only if, $m = 1$ and $k_1 = n$.
>
> Now consider the caterpillar with $n$ leaves. It has the characteristic property of having only one symmetry node: the root of the cherry at its bottom. Thus, its number of automorphisms is $2! = 2$. Notice that any other tree has either more symmetry nodes —hence, a bigger number of automorphisms— or a single $k$-fan with $k \geq 3$, and therefore, again, a bigger number of automorphisms.

### 1.1.3 Newick

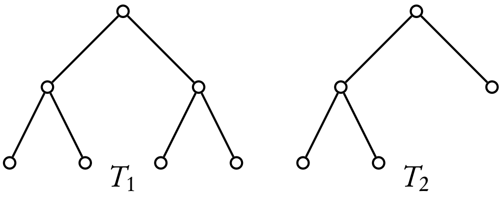
In order to represent a tree in a practical manner, or at least in one that can be easily manipulated by an easy algorithm, we will use the Newick format for a tree [14]. We define it recursively as follows:

$$\begin{aligned} \mathrm{newick} : \mathbf{Tree} &\longrightarrow \mathbf{String} \\ \ell &\longmapsto \cdot \\ T_1 * \cdots * T_k &\longmapsto (\mathrm{newick}(T_1), \ldots, \mathrm{newick}(T_k)) \end{aligned}$$

**Remark 1.17.** The usual Newick format ends the description of a tree by a semicolon, and our programs expect these strings to do so, but in this memoir we shall avoid this convention in order not to confuse this semicolon with a punctuation mark.

Example:

Let $T = T_7^{\text{bal}} \in \mathbf{BinTree}_7$ be the maximally balanced tree depicted in Figure 1.3. It can be written as $T = T_1 * T_2$, where



Proceeding in this fashion we would get to the leaves, whose Newick representation is just $\cdot$. Therefore, we would construct newick$(T_1) = ((\cdot,\cdot),(\cdot,\cdot))$ and newick$(T_2) = ((\cdot,\cdot),\cdot)$. Finally,

$$\text{newick}(T) = \text{newick}(T_1 * T_2) = (\text{newick}(T_1), \text{newick}(T_2)) = (((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),\cdot)).$$

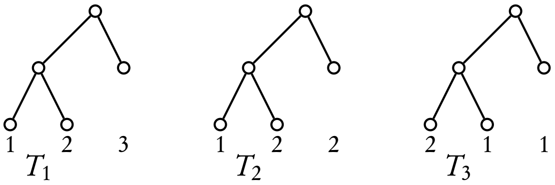
The Newick format can be extended to a function **MulTree** → **String** in an easy way, that has the perquisite of being easy to particularize to multilabelled shapes. It can be defined as follows:

$$\begin{array}{rcl} \text{newick} : \mathbf{MulTree} & \to & \mathbf{String} \\ (\{\ell\}, \lambda) & \mapsto & \lambda(\ell) \\ (T_1, \lambda_1) * \cdots * (T_k, \lambda_k) & \mapsto & (\text{newick}\ (T_1, \lambda_1), \ldots, \text{newick}\ (T_k, \lambda_k)) \end{array}$$

Finally, we shall also use Newick strings to describe the elements of **MulShTree**, by simply taking the Newick representation of a multilabelled tree shape as the Newick representation of any multilabelled tree representing it.

Example:

Recall the three trees depicted in Figure 1.6:



Then,

- newick$(\pi_1(T_1))$ = newick$(\pi_1(T_2))$ = newick$(\pi_1(T_3)) = ((\cdot,\cdot),\cdot)$.
- newick$(T_1) = ((1,2),3)$, newick$(T_2) = ((1,2),2)$, and newick$(T_3) = ((1,2),1)$.
- newick$(\pi^*(T_1)) = ((1,2),3)$ and newick$(\pi^*(T_2))$ = newick$(\pi^*(T_3)) = ((1,2),2)$.

Notice that, since the root join is an operation —and, in particular, gives rise to a unique tree—, we can easily show that the Newick representation of a tree also allows us to recover a single tree from it. Nevertheless, bear in mind that a tree, as well as a multilabelled tree and even more a multilabelled tree shape, may have more than one representation in Newick format, due to the commutativity of the root join. For instance, for the tree $T_1$ used in the last example, we also have that newick$(\pi_1(T_1)) = (\cdot, (\cdot, \cdot))$ and newick$(T_1)$ may as well be $((2,1),3)$, $(3,(1,2))$ or $(3,(2,1))$, while newick$(\pi^*(T_1))$ can be any string of the form either $((i,j),k)$ or $(i,(j,k))$ with $\{i,j,k\} = \{1,2,3\}$.

## 1.2 Measures of balance

A *balance index* is an instance of what J. Mosterín calls a *metric concept* [89], as it assigns a tree a magnitude that might not have any sense in and by itself, but allows us to compare trees in a determinate manner. They focus on a topological feature —the "balance" of a tree— for which an intuitive idea is readily available in everyone's mind. This *balance* is a measure of the propensity of a tree of having nodes such that, for every children, their rooted subtrees have the same number of leaves.

A balance index is, in general, a function $I : \mathbf{Tree} \to \mathbb{R}$ or $I : \mathbf{BinTree} \to \mathbb{R}$. If well defined, the relation $\leq_{\mathbb{R}}$ should capture some aspect of the "balance" of the trees in **Tree**, although this concept is difficult to put into words and we usually end up thinking about it in the fashion of Wittgenstein's "gut ist, was Gott befiehlt": "balance is what is measured by the $I$ index".

But what does it mean that a balance index is well, or ill, defined? This question is too big to be answered in this work, but there are certain intuitive, pre-theoretical properties that a balance index *ought to* satisfy. First of all, it should correlate with other balance indices, which is a way to ensure that they are somehow measuring *the same underlying property*, even if it cannot be unveiled; but, of course, this correlation must not be perfect, for otherwise the new index would be redundant. Secondly, it should place those trees that are conventionally considered to be "most balanced" and those considered to be "least balanced" as opposite extremes in the range of values that it can attain. Here we also find that those considered to be "least balanced" are (almost) always easier to capture than those "most balanced", and indeed all balance indices for trees considered in this work correctly classify the former even though some of them fail in classifying the latter, or at least its unicity. Thirdly, it should be easily computed, as most balance indices used in the literature are computed in linear time. Finally, a good balance index should have a reasonably big range of possible values, given the stupendously largeness the number of trees with a fixed number of leaves can get. This last reason is behind the fact that the number of automorphisms is not used to compute the balance of bifurcating trees, since Lemma 1.13 and Theorem 1.14 imply that, in the bifurcating case, it amounts to counting symmetry nodes of a tree, whose size is linear on the number of leaves.

Given a balance index $I$, it can be unreasonable to expect it to allow us to compare the balance of two trees with a different number of leaves. A possible way to circumvent this difficulty is to normalize it as follows

$$\begin{array}{rcl} \overline{I} : \mathbf{Tree} & \to & [0,1] \\ T & \mapsto & \frac{I(T)-\min\{I(T'):T'\in\mathbf{Tree}_{|L(T)|}\}}{\max\{I(T'):T'\in\mathbf{Tree}_{|L(T)|}\}-\min\{I(T'):T'\in\mathbf{Tree}_{|L(T)|}\}}. \end{array}$$

Notice that this formula has the desirable property of attaining the extreme values of the interval $[0,1]$. But, to do this, we need to know the maximum and minimum values of $I$ on each $\mathbf{Tree}_n$. Another possible way to do it, given a probabilistic model $(P_n)_n$ of phylogenetic trees (see Section 1.3) and the random variable $I_n$ that takes a tree $T \in \mathbf{PhyloTree}_n$ generated with probability $P_n(T)$ and computes the balance index $I(T)$, would be to standardize $I$ by means of the usual transformation:

$$\widehat{I}(T) = \frac{I(T) - E_P(I_n)}{\sigma_P(I_n)},$$

where $E_P(I_n)$ and $\sigma_P(I_n)$ denote the expected value and the variance of $I_n$. But, again, to do so we need to know these statistics.

In this document we will work with several balance indices, mainly the Colless index, the Quartet Index[3], the Sackin index and the Variance of depths. Up to a degree, we shall also be concerned with the Cophenetic index. Now, we introduce those of them that were already known before this research began.

### 1.2.1 The Colless index

The Colless index [19] is one of the most widely used measures of tree balance, partly due to its intuitiveness, as well as being the second oldest balance index found in the literature. It is defined only for bifurcating trees as follows [19, 22]:

$$\begin{array}{rcl} C : \mathbf{BinTree} & \to & \mathbb{N} \\ T & \mapsto & \sum_{u \in \mathring{V}(T)} \mathrm{bal}(u). \end{array}$$

Although there exists a sound extension of the Colless index for multifurcating trees [86], in this report we shall only be concerned with the original (binary) definition. Notice that, by dividing by the number of internal nodes of the tree, this gives the mean share of "imbalance" for each node, thus providing an intuitive justification for this measure.

The maximum value of the Colless index of a bifurcating tree with $n$ leaves is always reached at the caterpillars, and uniquely so [85]. This value is

$$C(T_n^{\mathrm{cat}}) = \frac{(n-1)(n-2)}{2}.$$

This result agrees with the intuition that the caterpillars are the "least balanced" of all trees [102].

We devote the integrity of Chapter 2 to the minimum value of the Colless index. Apart from the obvious result that the minimum Colless index among all bifurcating trees with $2^m$ leaves, where $m \in \mathbb{N}$, is achieved exactly at the fully symmetric trees [60, 69, 88], whose Colless index is $0$, neither this minimum nor the trees attaining it had been characterized when trees with any number of leaves were considered. We provide a full characterization of the trees that attain the minimum Colless index, as well as its value. We shall also show that, given $n \in \mathbb{N}_{\geq 1}$, the width of the range of values of the Colless index on $\mathbf{BinTree}_n$ is in $O(n^2)$.

### 1.2.2 The Sackin index

The Sackin index [107] is usually given credit to be the oldest balance index existing in the literature, even though, as we mentioned in the introduction, in its seminal paper

[3] *Née* Rooted Quartet index in [25].

[102] Sackin actually proposed another index (namely, the Variance of depths) to be used. In any case, what we currently call Sackin index is the one that got popular and widespread —and not without a fair amount of reasons—, and is defined to be the sum of the leaves' depths of a given tree [107]; that is,

$$\begin{array}{rcl} S : \mathbf{Tree} & \rightarrow & \mathbb{N} \\ T & \mapsto & \sum_{u \in L(T)} \delta(u) \end{array}$$

i.e., the sum of the multiset $\Delta(T)$ of the leaves' depths of $T$ if considered a vector of numbers. It can be proved [8, 99] that this index has the following equivalent reformulation:

$$\begin{array}{rcl} S : \mathbf{Tree} & \rightarrow & \mathbb{N} \\ T & \mapsto & \sum_{u \in \mathring{V}(T)} \kappa(u). \end{array}$$

If we take into account the number of leaves of the tree $T$, $n$, we can define

$$\overline{S}(T) = \frac{1}{n} S(T)$$

to be the *mean depth* of the leaves of $T$. This gives an intuitive justification for this index. Notice, however, that by construction the Sackin index is very prone to repeat values, since any two trees $T_1, T_2 \in \mathbf{Tree}_n$ with the same number of leaves such that $\Delta(T_1) = \Delta(T_2)$ will have the same Sackin index; it does not, thus, distinguish the shapes between them.

By Theorem 2 in [39] and the obvious fact that, for every non-binary tree there always exist a binary tree with largest Sackin index (obtained by resolving all multifurcations in it), we know that the maximum value of the Sackin index on $\mathbf{Tree}_n$ is reached exactly at the caterpillars, which, again, agrees with the observation that these trees ought to be classified as being the least balanced of all trees [102]. This maximum value is

$$S(T_n^{\mathrm{cat}}) = \frac{(n-1)(n+2)}{2},$$

for any number of leaves $n$. As to the minimum value for multifurcating trees, it is clearly reached exactly at the star $T_n^{\mathrm{star}}$ and it is

$$S(T_n^{\mathrm{star}}) = n.$$

Nevertheless, stars are never bifurcating as soon as the number of leaves exceeds 2 (which happens quite a lot), and so the question for the minimum value of the Sackin index applied to bifurcating trees is sound. This question was recently answered by M. Fischer with the next result, which is basically a restatement of Algorithm 1 and Corollary 2 in [39].

**Theorem 1.18.** *Let $T \in \mathbf{BinTree}_n$ and let $m = \lceil \log_2(n) \rceil$. Then, $S(T)$ is minimum on $\mathbf{BinTree}_n$ if, and only if, it is obtained from the fully symmetric tree with $2^m$ leaves $T_{2^m}^{\mathrm{bal}}$ by removing from it any $2^m - n$ cherries and replacing them by their roots, which become leaves of depth $m-1$. All trees obtained in this way have $2^m - n$ leaves of depth $m-1$ and $2n - 2^m$ leaves of depth $m$, and therefore the minimum value of $S$ on $\mathbf{BinTree}_n$ is*

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

Thus, the minimum and maximum values of Sackin index on $\mathbf{BinTree}_n$ grow in $O(n\log(n))$ and $O(n^2)$, respectively, and hence the range of values of the Sackin index on $\mathbf{BinTree}_n$ is in $O(n^2)$, as it was the case for the Colless index. The range of its values on $\mathbf{Tree}_n$ grows also in quadratic order, but it is wider because the minimum value is smaller.

Now, we provide the following result that we will use in Chapter 4.

**Theorem 1.19.** *Let $T \in \mathbf{BinTree}_n$. The following conditions are equivalent:*

*(i)* *$T$ has minimum Sackin index.*

*(ii)* *There exists a $d_0 \in \mathbb{N}$ such that $\delta(x) \in \{d_0, d_0+1\}$ for every $x \in L(T)$.*

*(iii)* *$|\delta(x) - \overline{S}(T)| < 1$ for every $x \in L(T)$.*

*(iv)* *$T$ is depth-equivalent to $T_n^{\mathrm{bal}}$.*

*Proof.* Let $m = \lceil \log_2(n) \rceil$. By applying Theorem 1.18, we see that the bifurcating trees that attain the minimum Sackin index have all their leaves of depth $m$ (those that remain from $T_{2^m}^{\mathrm{bal}}$) or $m-1$ (the roots of the removed cherries). Therefore, *(i)* implies *(ii)*.

We proceed now to show the equivalence between *(ii)* and *(iii)*. Since, by definition, $\overline{S}(T)$ is the mean value of the leaves' depths, *(ii)* implies *(iii)*. On the other hand, as $\delta(x) \in \mathbb{N}$ for all $x \in L(T)$, for it to be such that $|\delta(x) - \overline{S}(T)| < 1$ it must happen that $\delta(x) \in \{\lfloor \overline{S}(T) \rfloor, \lceil \overline{S}(T) \rceil\}$; therefore, *(iii)* implies *(ii)*.

In order to prove that *(ii)* implies *(i)*, we proceed by induction on the number of leaves $n$. The base case $n = 1$ is obvious, since there only exists a tree with one leaf. Therefore, we suppose this implication to be true up to $n-1$ leaves, $n \geq 2$. Let $T \in \mathbf{BinTree}_n$ be a tree satisfying condition *(ii)*, and $x_0 \in L(T)$ a leaf of maximum depth, which we shall assume to be $d_0+1$ (that is, if all the leaves were to have the same depth, it is $d_0+1$). Then, $x_0$ must be part of a cherry: consider the tree $T'$ obtained by removing that cherry and placing a leaf $y_0$ of depth $d_0$ instead. The resulting tree has $n-1$ leaves and it still satisfies condition *(ii)*, and therefore it lies under our induction hypothesis: it has, hence, minimum Sackin index, and so it has the form described in Theorem 1.18: $n-1-2^{m'-1}$ cherries at depth $m'$ and $2^{m'}-n+1$ leaves at depth $m'-1$, where $m' = \lceil \log_2(n-1) \rceil$. Note that $m' = m-1$ if $n = 2^{m-1}+1$ and $m' = m$ otherwise. Since, then, $T$ is obtained from $T'$ by pending a cherry from a leaf in it at depth $m'-1$, it is straightforward to check that $T$ is also of the form described in Theorem 1.18.

Thus concludes the proof that conditions *(i)*, *(ii)* and *(iii)* are equivalent.

Let us prove now that if a bifurcating tree $T$ satisfies *(iv)* then it satisfies *(ii)* with $d_0 = \lfloor \log_2(n) \rfloor$, by induction on the depth of the tree. This implication is trivially true when $\delta(T) = 0$, because the only tree of depth 0 has a single leaf. Assume now that the implication is true for every tree of depth at most $\delta$ and let $T \in \mathbf{BinTree}_n$ be a bifurcating tree of depth $\delta+1$ that is depth-equivalent to $T_n^{\mathrm{bal}}$. Since *(ii)* is an assertion on the depths of the leaves of $T$, and $\Delta(T) = \Delta(T_n^{\mathrm{bal}})$, in order to prove that $T$ satisfies *(ii)* we can assume without any loss of generality that $T = T_n^{\mathrm{bal}}$. Let $k = n - 2^{d_0}$, where, we recall, $d_0 = \lfloor \log_2(n) \rfloor$.

Let $T_1$ and $T_2$ be the maximal pending subtrees of $T$ and $n_1$ and $n_2$ their respective numbers of leaves, with $n_1 \geq n_2$ and $n = n_1 + n_2$. Since $T$ is maximally balanced, $T_1$ and $T_2$ are also maximally balanced and $n_1 = \lceil n/2 \rceil$ and $n_2 = \lfloor n/2 \rfloor$. Then, since

$\delta(T_1), \delta(T_2) \leq \delta = \delta(T) - 1$, by the induction hypothesis we deduce that if, for every $i \in \{1, 2\}$, we set $d_i = \lfloor \log_2(n_i) \rfloor$, then $\delta_{T_i}(x) \in \{d_i, d_i + 1\}$ for every $x \in L(T_i)$. Now:

- If $k < 2^{d_0} - 1$, then $d_1 = d_2 = d_0 - 1$.
- If $k = 2^{d_0} - 1$, then $n_1 = 2^{d_0}$, and thus $d_1 = d_0$, and $n_2 = 2^{d_0} - 1$, and thus $d_2 = d_0 - 1$; but then, $T_2$ is fully symmetric, because it is maximally balanced with $2^{d_0}$ leaves, which implies in particular that all its leaves have depth $d_0$.

Then, in both cases, $\delta_{T_i}(x) \in \{d_0 - 1, d_0\}$ for every $x \in L(T_i)$ and $i \in \{1, 2\}$. Since $\delta_T(x) = \delta_{T_i}(x) + 1$ if $x \in L(T_i)$, we conclude that $\delta_T(x) \in \{d_0, d_0 + 1\}$ for every $x \in L(T)$. This finishes the proof that *(iv)* implies *(ii)*.

Finally, we prove that *(i)* implies *(iv)*. Let $T \in \mathbf{BinTree}_n$ be of the form described in Theorem 1.18. Since we have already proved that *(iv)* implies *(ii)* and *(ii)* implies *(i)*, we know that $T_n^{\mathrm{bal}}$ is also of the form described in Theorem 1.18. But, then, $T$ and $T_n^{\mathrm{bal}}$ are depth-equivalent because all trees of the form described in Theorem 1.18 have $2^m - n$ leaves of depth $m - 1$ and $2n - 2^m$ leaves of depth $m$. □

In particular, $T_n^{\mathrm{bal}}$ achieves the minimum Sackin index on $\mathbf{BinTree}_n$, but it can be proved that there are other trees in $\mathbf{BinTree}_n$ with minimum Sackin index if, and only if, $n$ is not of the form $2^m - 1$, $2^m$, or $2^m + 1$ for some $m \in \mathbb{N}$ [39, Cor. 3].

**Remark 1.20.** As a corollary, we also draw the conclusion that any maximally balanced tree with $n = 2^{m-1} + k$ leaves has, exactly, $k$ cherries at depth $m$ and $2^{m-1} - k$ leaves at depth $m - 1$. Notice, though, that this could have also been proved by induction using Theorem 1.8.

### 1.2.3 The Cophenetic index

A balance index of which we will be but marginally concerned in this report will be the Total Cophenetix index [85], which we shall call here Cophenetic index, $\Phi$. It is defined on a given tree $T$ as the sum over each pair of different leaves of the depth of their lowest common ancestor; i.e., for any $T \in \mathbf{Tree}_n$,

$$\Phi(T) = \frac{1}{2} \sum_{\substack{x,y \in L(T) \\ x \neq y}} \delta(\mathrm{lca}(x, y)).$$

As it can be shown (Lemma 2 in [85]), the above expression is equivalent to

$$\Phi(T) = \sum_{u \in \mathring{V}(T) \setminus \{\rho\}} \binom{\kappa(u)}{2}.$$

Notice that this balance index is defined for multifurcating trees.

The maximum value of the Cophenetic index on $\mathbf{Tree}_n$ is again uniquely reached at the caterpillars [85, Prop. 10], as was the case with the Sackin and the Colless indices (the latter, in the bifurcating case). Notice that, since the caterpillars are bifurcating trees, the maximum value when we restrict the index only to bifurcating trees remains the same. This value is

$$\Phi(T_n^{\mathrm{cat}}) = \binom{n}{3}.$$

Its minimum value on $\mathbf{Tree}_n$ is obviously attained exactly at the stars, since

$$\Phi(T_n^{\text{star}}) = 0,$$

while the bifurcating trees that achieve the minimum Cophenetic index are exactly the maximally balanced trees [85, Thm. 13]. Notice that in this sense, the Cophenetic index performs "better" than the Colless and Sackin ones do, since it uniquely classifies the maximally balanced trees, i.e. "the most balanced" bifurcating trees according to Shao and Sokal [107], as being the only ones achieving the minimum Cophenetic index. Unfortunately, no closed expression is known so far that computes this value, although a recurrent one can be given [85]: $\Phi(T_1^{\text{bal}}) = 0$ and, for $n \geq 2$,

$$\Phi(T_n^{\text{bal}}) = \Phi(T_{\lceil n/2 \rceil}^{\text{bal}}) + \Phi(T_{\lfloor n/2 \rfloor}^{\text{bal}}) + \binom{\lceil \frac{n}{2} \rceil}{2} + \binom{\lfloor \frac{n}{2} \rfloor}{2}.$$

The sequence $\Phi(T_n^{\text{bal}})$ grows in $O(n^2)$, and it can be found as sequence A011371 in Sloane's Encyclopedia of Integer Sequences [108]. So, the range of values of $\Phi$ on $\mathbf{Tree}_n$ and on $\mathbf{BinTree}_n$ grows in $O(n^3)$ and so it was, until the the introduction of our Quartet index [25], the balance index with the widest range in the literature.

### 1.2.4 Binary recursive shape indices

The Colless, the Sackin and the Cophenetic indices have a useful feature that deserves an adjective: they are *recursive*.

A *recursive shape index* [76] is a map $I : \mathbf{Tree} \to \mathbb{R}$ for which there exists a symmetric function $f_I : \bigcup_{k\geq 2} \mathbb{N}^k \to \mathbb{R}$ such that, for every trees $T_1 \in \mathbf{Tree}_{n_1}, \ldots, T_k \in \mathbf{Tree}_{n_k}$,

$$I(T_1 * \cdots * T_k) = \sum_{i=1}^{k} I(T_i) + f_I(n_1, \ldots, n_k),$$

where $f_I$ being symmetric means that for every $k \geq 2$, for every $(x_1, \ldots, x_k) \in \mathbb{N}^k$ and for every $\sigma \in \mathfrak{S}_k$,

$$f_I(x_1, \ldots, x_k) = f_I(x_{\sigma(1)}, \ldots, x_{\sigma(k)}). \tag{1.1}$$

When a recursive shape index is only defined on $\mathbf{BinTree}$, we say that it is a *binary recursive shape index*. The theory of recursive shape indices was introduced in [76] and then used, among others, by Cardona et al. [13].

We have that:

- The Colless index is a binary recursive shape index with $f_C(n_1, n_2) = |n_1 - n_2|$ [22, 98].

- The Sackin index is a recursive shape index with $f_S(n_1, \ldots, n_k) = \sum_{i=1}^{k} n_i$ [99].

- The Cophenetic index is a recursive shape index with $f_\Phi(n_1, \ldots, n_k) = \sum_{i=1}^{k} \binom{n_i}{2}$ [85].

Given a shape index $I : \mathbf{Tree} \to \mathbb{R}$ or $I : \mathbf{BinTree} \to \mathbb{R}$, we extend it to $\mathbf{PhyloTree}$, or $\mathbf{BinPhyloTree}$, by associating to each phylogenetic treee the value of $I$ on its shape:

$I(T,\lambda) = I(T)$. We shall say that a map $I$ : **PhyloTree** $\to \mathbb{R}$, or $I$ : **BinPhyloTree** $\to \mathbb{R}$, is a *recursive shape index for phylogenetic trees* when it is the extension to phylogenetic trees of a recursive shape index for trees. In particular, a recursive shape index for phylogenetic trees is invariant under isomorphisms of trees and relabellings.

## 1.3 Probabilistic models

A *probabilistic model of phylogenetic trees* $(P_n)_n$ is a family of maps $P_n : \mathbf{PhyloTree}_n \to [0,1]$, $n \geq 1$, such that, for every $n \geq 1$, $\sum_{(T,\lambda)\in\mathbf{PhyloTree}_n} P_n(T,\lambda) = 1$. A *probabilistic model of trees* $(P_n^*)_n$ is a family of maps $P_n^* : \mathbf{Tree}_n \to [0,1]$, $n \geq 1$, such that, for every $n \geq 1$, $\sum_{T\in\mathbf{Tree}_n} P_n^*(T) = 1$. A *probabilistic model of bifurcating trees*, or of *phylogenetic trees*, is a probabilistic model of trees, or phylogenetic trees, such that the probability of a tree is 0 whenever it is not bifurcating.

Each probabilistic model of phylogenetic trees $(P_n)_n$ induces a probabilistic model of trees $(P_n^*)_n$ by means of the relation

$$P_n^*(T) = \sum_{(T,\lambda)\in\mathbf{PhyloTree}_n} P_n(T,\lambda).$$

We say that a probabilistic model of phylogenetic trees $(P_n)_n$ is *shape invariant* [43] (or *exchangeable* [2]) if, for every $(T_1,\lambda_1),(T_2,\lambda_2) \in \mathbf{PhyloTree}_n$, $P_n(T_1,\lambda_1) = P_n(T_2,\lambda_2)$ whenever $T_1 = T_2$. In this case, the probabilistic model of trees induced by $(P_n)_n$ satisfies that

$$P_n^*(T) = \left|\{(T_0,\lambda_0) \in \mathbf{PhyloTree}_n : T_0 = T\}\right| \cdot P_n(T,\lambda) = \phi(T)\cdot P_n(T,\lambda),$$

with the notations in Theorem 1.14, for any $(T,\lambda) \in \mathbf{PhyloTree}_n$.

Each probabilistic model of trees $(P_n^*)_n$ induces a probabilistic model of phylogenetic trees $(P_n)_n$ by splitting equally the probability of each tree among all phylogenetic trees of this shape

$$P_n(T,\lambda) = \frac{1}{\left|\{(T_0,\lambda_0) \in \mathbf{PhyloTree}_n : T_0 = T\}\right|} P_n^*(T) = \frac{1}{\phi(T)} P_n^*(T).$$

It is clear that the probabilistic model of phylogenetic trees induced in this way by a probabilistic model of trees is shape invariant.

### 1.3.1 Sampling consistency

A probabilistic model of phylogenetic trees $(P_n)_n$ is said to be *sampling consistent* [2] (or *deletion stable* [43]) when, given a phylogenetic tree $(T_0,\lambda_0) \in \mathbf{PhyloTree}_{n-1}$, the probability of obtaining it through the procedure of choosing a phylogenetic tree $(T,\lambda) \in \mathbf{PhyloTree}_n$ with probability $P_n(T,\lambda)$ and removing the leaf labelled $n$ (as well as any elementary node created in this way) is $P_{n-1}(T_0,\lambda_0)$; i.e., formally, when for every $n \geq 2$ and for every $(T_0,\lambda_0) \in \mathbf{PhyloTree}_{n-1}$,

$$P_{n-1}(T_0,\lambda_0) = \sum_{\substack{(T,\lambda)\in\mathbf{PhyloTree}_n\\ (T,\lambda)([n-1])=(T_0,\lambda_0)}} P_n(T,\lambda).$$

It is an easy induction exercise to prove that, for every $m \in \{1, \ldots, n-1\}$, a sampling consistent model $(P_n)_n$ satisfies that, for any $(T_0, \lambda_0) \in \mathbf{PhyloTree}_m$,

$$P_m(T_0, \lambda_0) = \sum_{\substack{(T,\lambda)\in\mathbf{PhyloTree}_n \\ (T,\lambda)([m])=(T_0,\lambda_0)}} P_n(T, \lambda). \tag{1.2}$$

If $(P_n)_n$ is sampling consistent *and* shape invariant, then the probabilities of the trees are not affected by the specific labels we consider, and thus for any non-empty subset $X \subseteq [n]$, and for any $(T_0, \lambda_0) \in \mathbf{PhyloTree}(X)$, we would have

$$P_X(T_0, \lambda_0) = \sum_{\substack{(T,\lambda)\in\mathbf{PhyloTree}_n \\ (T,\lambda)(X)=(T_0,\lambda_0)}} P_n(T, \lambda),$$

where $P_X : \mathbf{PhyloTree}(X) \to [0, 1]$ is induced by $P_m$ (where $|X| = m$) through any bijection $X \leftrightarrow [m]$.

We can naturally extend this concept to tree shapes. A probabilistic model of trees $(P_n^*)_n$ is *sampling consistent* if, for every $n \geq 2$, having chosen a tree $T$ with probability $P_n^*(T)$ and a leaf $x \in L(T)$ equiprobably, then the resulting tree from having $x$ removed is produced with probability given by $P_{n-1}^*$; i.e., when for any $n \geq 2$ and any $T_0 \in \mathbf{Tree}_{n-1}$,

$$P_{n-1}^*(T_0) = \sum_{T\in\mathbf{Tree}_n} \frac{|\{x \in L(T) : T(L(T) \setminus \{x\}) = T_0\}|}{n} P_n^*(T).$$

In Lemma 1.21 below we extend this relation to any subset $X \subseteq L(T)$, thus providing a generalization of Equation (1.2) for shapes. In its statement, and henceforth, $\operatorname{Part}(X)$ is the set of parts, or subsets, of any set $X$ and

$$\operatorname{Part}_m(X) = \{S \in \operatorname{Part}(X) : |S| = m\}.$$

We will then provide several lemmata on probabilistic models that will be used in this work, not having been able to find suitable references in the literature.

**Lemma 1.21.** *A probabilistic model of trees $(P_n^*)_n$ is sampling consistent if, and only if, for every $n \geq 2$, for every $1 \leq m \leq n$, and for every $T_0 \in \mathbf{Tree}_m$,*

$$P_m^*(T_0) = \sum_{T\in\mathbf{Tree}_n} \frac{|\{X \in \operatorname{Part}_m(L(T)) : T(X) = T_0\}|}{\binom{n}{m}} P_n^*(T).$$

*Proof.* The "if" implication is obvious, since it amounts to considering $m = n - 1$. Now, for the "only if" implication, we will proceed by induction on the number $n - m$ of leaves we remove. The base case $n - m = 1$ is already known, by the definition of sampling consistency. Now suppose it is true up to $n - m - 1$ removed leaves, $m \geq 1$; we want to show that

$$P_m^*(T_0) = \sum_{T\in\mathbf{Tree}_n} \frac{|\{X \in \operatorname{Part}_m(L(T)) : T(X) = T_0\}|}{\binom{n}{m}} P_n^*(T).$$

Now, we know that

$$\begin{aligned}
P_m^*(T_0) &= \sum_{T_{m+1}\in\mathbf{Tree}_{m+1}} \frac{|\{x\in L(T_{m+1}) : T(L(T_{m+1})\setminus\{x\}) = T_0\}|}{m+1} P_{m+1}^*(T_{m+1}) \\
&\quad\text{(by sampling consistency)} \\
&= \sum_{T_{m+1}\in\mathbf{Tree}_{m+1}} \left( \frac{|\{x\in L(T_{m+1}) : T(L(T_{m+1})\setminus\{x\}) = T_0\}|}{m+1} \right. \\
&\qquad \left. \cdot \sum_{T\in\mathbf{Tree}_n} \frac{|\{X\in \mathrm{Part}_{m+1}(L(T)) : T(X) = T_{m+1}\}|}{\binom{n}{m+1}} P_n^*(T) \right) \\
&\quad\text{(by the induction hypothesis)} \\
&= \sum_{T_{m+1}\in\mathbf{Tree}_{m+1}} \sum_{T\in\mathbf{Tree}_n} \frac{|\{x\in L(T_{m+1}) : T(L(T_{m+1})\setminus\{x\}) = T_0\}|}{m+1} \\
&\qquad \cdot \frac{|\{X\in \mathrm{Part}_{m+1}(L(T)) : T(X) = T_{m+1}\}|}{\binom{n}{m+1}} P_n^*(T) \\
&= \sum_{T\in\mathbf{Tree}_n} \frac{|\{(x,X)\in L(T)\times \mathrm{Part}_{m+1}(L(T)) : x\in X,\ (T(X))(X\setminus\{x\}) = T_0\}|}{(m+1)\binom{n}{m+1}} P_n^*(T) \\
&= \sum_{T\in\mathbf{Tree}_n} \frac{|\{(x,X)\in L(T)\times \mathrm{Part}_{m+1}(L(T)) : x\in X,\ T(X\setminus\{x\}) = T_0\}|}{(m+1)\binom{n}{m+1}} P_n^*(T) \\
&= \sum_{T\in\mathbf{Tree}_n} \frac{(n-m)|\{Y\in \mathrm{Part}_{m}(L(T)) : T(Y) = T_0\}|}{(m+1)\binom{n}{m+1}} P_n^*(T) \\
&= \sum_{T\in\mathbf{Tree}_n} \frac{|\{X\in \mathrm{Part}_{m}(L(T)) : T(X) = T_0\}|}{\binom{n}{m}} P_n^*(T),
\end{aligned}$$

which proves the inductive step. □

The next lemma will be useful in the proof of the last result of this section, which relates the sampling consistency of a probabilistic model of phylogenetic trees to the sampling consistency of a probabilistic model of trees.

**Lemma 1.22.** *Let $(P_n)_n$ be a shape invariant probabilistic model of phylogenetic trees. For every $(T,\lambda),(T',\lambda')\in\mathbf{PhyloTree}_{n-1}$, if $T = T'$, then*

$$\sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ (T_n,\lambda_n)([n-1])=(T,\lambda)}} P_n(T_n,\lambda_n) = \sum_{\substack{(T_n',\lambda_n')\in\mathbf{PhyloTree}_n\\ (T_n',\lambda_n')([n-1])=(T',\lambda')}} P_n(T_n',\lambda_n').$$

*Proof.* Let $\varphi : T \to T'$ be an isomorphism of trees. For every $(T'',\lambda'')\in\mathbf{PhyloTree}_{n-1}$, let

$$E_n(T'',\lambda'') = \{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n : (T_n,\lambda_n)([n-1]) = (T'',\lambda'')\}.$$

Each $(T_n,\lambda_n)\in E_n(T'',\lambda'')$ is obtained by adding a leaf labelled with $n$ to $T''$ as a new child to either an internal node, or to a new node obtained by the splitting in two of an edge, or to a new bifurcating root (whose other child would, then, be the old root). This implies the existence of a shape preserving bijection

$$f : E_n(T, \lambda) \to E_n(T', \lambda')$$

that sends each phylogenetic tree $(T_n, \lambda_n) \in E_n(T, \lambda)$ to the tree in $E_n(T', \lambda')$ obtained by adding the leaf labelled with $n$ to the place in $T'$ that corresponds to it under the isomorphism of trees $\varphi$. Then, since $(P_n)_n$ is shape invariant,

$$\sum_{(T_n,\lambda_n)\in E_n(T,\lambda)} P_n(T_n, \lambda_n) = \sum_{(T_n,\lambda_n)\in E_n(T,\lambda)} P_n(f(T_n, \lambda_n))$$

$$= \sum_{(T'_n,\lambda'_n)\in E_n(T',\lambda')} P_n(T'_n, \lambda'_n)$$

as we wanted to show. □

The next result is an intuitive, albeit important, lemma that generalizes Corollary 40 in [43].

**Lemma 1.23.** *Let $(P_n)_n$ be a shape invariant probabilistic model of phylogenetic trees and let $(P^*_n)_n$ be the corresponding probabilistic model of tree shapes. Then, $(P_n)_n$ is sampling consistent if, and only if, $(P^*_n)_n$ is.*

*Proof.* We begin by proving the "only if" implication; therefore, suppose that $(P_n)_n$ is sampling consistent. Then, for every $T \in \mathbf{Tree}_{n-1}$ and for every $(T, \lambda) \in \mathbf{PhyloTree}_{n-1}$,

$$P^*_{n-1}(T) = \phi(T)P_{n-1}(T, \lambda)$$

(by the shape invariance of $(P_n)_n$)

$$= \phi(T) \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n \\ (T_n,\lambda_n)([n-1])=(T,\lambda)}} P_n(T_n, \lambda_n)$$

(by the sampling consistency of $(P_n)_n$)

$$= \sum_{(T,\lambda)\in\mathbf{PhyloTree}_{n-1}} \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n \\ (T_n,\lambda_n)([n-1])=(T,\lambda)}} P_n(T_n, \lambda_n)$$

(by Lemma 1.22)

$$= \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n \\ \pi_1(T_n,\lambda_n)([n-1])=T}} P_n(T_n, \lambda_n) = \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n \\ \pi_1(T_n,\lambda_n)([n]\setminus\{i\})=T}} P_n(T_n, \lambda_n) \quad \text{for any } i \in [n]$$

(by the shape invariance of $(P_n)_n$)

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ \pi_1(T_n,\lambda_n)([n]\setminus\{i\})=T}} P_n(T_n,\lambda_n)\\
&= \sum_{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n} \frac{|\{i\in[n]:\pi_1(T_n,\lambda_n)([n]\setminus\{i\})=T\}|}{n} P_n(T_n,\lambda_n)\\
&= \sum_{T_n\in\mathbf{Tree}_n}\left(\frac{|\{x\in L(T_n):T_n(L(T)\setminus\{x\})=T\}|}{n}\sum_{\substack{(T,\lambda)\in\mathbf{PhyloTree}_n\\ \pi_1(T,\lambda)=T_n}} P_n(T,\lambda)\right)\\
&= \sum_{T_n\in\mathbf{Tree}_n} \frac{|\{x\in L(T_n):T_n(L(T)\setminus\{x\})=T\}|}{n} P_n^*(T_n)
\end{aligned}
$$

as we claimed.

The proof of the other implication consists in carefully reversing the argument, running backwards the sequence of equalities in the proof above. Assume that $(P_n^*)_n$ is sampling consistent and let $T\in\mathbf{Tree}_{n-1}$. Then,

$$
P_{n-1}^*(T) = \sum_{T_n\in\mathbf{Tree}_n} \frac{|\{x\in L(T_n):T_n(L(T_n)\setminus\{x\})=T\}|}{n} P_n^*(T_n)
$$

(by the sampling consistency of $(P_n)_n$)

$$
\begin{aligned}
&= \frac{1}{n}\sum_{T_n\in\mathbf{Tree}_n}\left(|\{x\in L(T_n):T_n(L(T_n)\setminus\{x\})=T\}| \sum_{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n} P_n(T_n,\lambda_n)\right)\\
&= \frac{1}{n}\sum_{T_n\in\mathbf{Tree}_n}\sum_{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n} |\{i\in[n]:T_n([n]\setminus\{i\})=T\}| P_n(T_n,\lambda_n)\\
&= \frac{1}{n}\sum_{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n} |\{i\in[n]:T_n([n]\setminus\{i\})=T\}| P_n(T_n,\lambda_n)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ \pi_1(T_n,\lambda_n)([n]\setminus\{i\})=T}} P_n(T_n,\lambda_n) = \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ \pi_1(T_n,\lambda_n)([n-1])=T}} P_n(T_n,\lambda_n)
\end{aligned}
$$

(by the shape invariance of $(P_n)_n$)

$$
\begin{aligned}
&= \sum_{(T,\lambda)\in\pi_1^{-1}(T)}\sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ (T_n,\lambda_n)([n-1])=(T,\lambda)}} P_n(T_n,\lambda_n)\\
&= \phi(T)\sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ (T_n,\lambda_n)([n-1])=(T,\lambda)}} P_n(T_n,\lambda_n)
\end{aligned}
$$

by Lemma 1.22, and then, using the shape invariance of $P_n$ and dividing both sides of the equality by $\phi(T)$, we get

$$
P_n(T_n,\lambda_n) = \frac{1}{\phi(T)}P_{n-1}^*(T) = \sum_{\substack{(T_n,\lambda_n)\in\mathbf{PhyloTree}_n\\ (T_n,\lambda_n)([n-1])=(T,\lambda)}} P_n(T_n,\lambda_n)
$$

as we wanted to prove. □

### 1.3.2 Markovianity

Let $(P_n)_n$ be a probabilistic model of phylogenetic trees. It is natural to ask ourselves whether the probability of obtaining a phylogenetic tree with $n$ leaves is related to the probability of its maximal pending subtrees in some meaningful way. We say that $(P_n)_n$ is *Markovian self-similar* [43] (or simply *Markovian*) if there exists a symmetric (in the sense of Equation (1.1)) map

$$q : \bigcup_{k\geq 2} \mathbb{N}^k \to \mathbb{R}_{\geq 0}$$

such that, for every $(T_1 * \ldots * T_k, \lambda) \in \mathbf{PhyloTree}_n$, with each $|L(T_i)| = n_i$,

$$P_n(T_1 * \ldots * T_k, \lambda) = q(n_1, \ldots, n_k) P_{n_1}(T_1, \lambda|_{T_1}) \cdots P_{n_k}(T_k, \lambda|_{T_k}).$$

It can be shown that if $(P_n)_n$ is a Markovian probabilistic model of phylogenetic trees, then the map $q$ is unique (Proposition 25 in [43]), and then it is called the *split distribution* of $(P_n)_n$.

Now suppose that a given Markovian probabilistic model is also shape invariant; then, the Markovianity of the model ought to be reflected when we forget the labelling of our trees. And this is indeed the case, as is presented by the next result for bifurcating trees (for simplicity and because this is the only case where we shall use it).

**Lemma 1.24.** *Let $(P_n)_n$ a Markovian shape invariant probabilistic model of bifurcating phylogenetic trees, and let $(P^*_n)_n$ be the induced probabilistic model of bifurcating trees. Then, if $T = T_k * T_{n-k} \in$ **BinTree**$_n$ where $T_k \in$ **BinTree**$_k$ and $T_{n-k} \in$ **BinTree**$_{n-k}$ for $1 \leq k \leq n-1$,*

*(i) If $T_k \neq T_{n-k}$,*

$$P^*_n(T_k * T_{n-k}) = \binom{n}{k} q(k, n-k) P^*_k(T_k) \cdot P^*_{n-k}(T_{n-k}).$$

*(ii) If $T_k = T_{n-k}$,*

$$P^*_n(T_k * T_{n-k}) = \frac{1}{2}\binom{n}{k} q(k, n-k) P^*_k(T_k) \cdot P^*_{n-k}(T_{n-k}).$$

*Proof.* By the shape invariance of the model, we know that $P^*_m(T) = \phi(T) P_m(T, \lambda)$ for all $m \in \mathbb{N}_{\geq 1}$ and $(T, \lambda) \in \mathbf{BinPhyloTree}_m$. Now, by the Markovianity of the model,

$$\begin{aligned} P^*_n(T_k * T_{n-k}) &= \phi(T_k * T_{n-k}) P_n(T_k * T_{n-k}, \lambda) \\ &= \phi(T_k * T_{n-k}) q(k, n-k) P_k(T_k, \lambda_k) P_{n-k}(T_{n-k}, \lambda_{n-k}) \\ &= \frac{\phi(T_k * T_{n-k})}{\phi(T_k)\phi(T_{n-k})} q(k, n-k) P^*_k(T_k) P^*_{n-k}(T_{n-k}), \end{aligned}$$

where $\lambda_k, \lambda_{n-k}$ are the restrictions of $\lambda$ to the sets of leaves of $T_k$ and $T_{n-k}$, respectively. Therefore, it is now a matter of deducing the value of the ratio

$$\frac{\phi(T_k * T_{n-k})}{\phi(T_k)\phi(T_{n-k})}.$$

Now, by Corollary 1.16, we distinguish two cases:

- If $T_k \neq T_{n-k}$ the root is not a symmetry node, and then $s(T_k * T_{n-k}) = s(T_k) + s(T_{n-k})$; thus,
$$\frac{\phi(T_k * T_{n-k})}{\phi(T_k)\phi(T_{n-k})} = \frac{\frac{n!}{2^{s(T_k*T_{n-k})}}}{\frac{k!(n-k)!}{2^{s(T_k)+s(T_{n-k})}}} = \binom{n}{k}\frac{2^{s(T_k*T_{n-k})}}{2^{s(T_k*T_{n-k})}} = \binom{n}{k}.$$
- On the other hand, if $T_k = T_{n-k}$, then the root is a symmetry node, therefore $s(T_k * T_{n-k}) = s(T_k) + s(T_{n-k}) + 1$. Hence,
$$\frac{\phi(T_k * T_{n-k})}{\phi(T_k)\phi(T_{n-k})} = \frac{\frac{n!}{2^{s(T_k*T_{n-k})}}}{\frac{k!(n-k)!}{2^{s(T_k)+s(T_{n-k})}}} = \binom{n}{k}\frac{2^{s(T_k*T_{n-k})-1}}{2^{s(T_k*T_{n-k})}} = \frac{1}{2}\binom{n}{k}.$$

□

This result will be quite useful in Chapters 3 and 4, where we shall exploit the fact that the Yule and the Uniform models for bifurcating phylogenetic trees (which we describe in the next section) are both Markovian and shape invariant.

**Remark 1.25.** Notice that this last Lemma entails in particular that if $(P_n)_n$ is Markovian and shape invariant, then $(P_n^*)_n$ is, in general, not Markovian. Indeed, let $T_n, T_n' \in$ **BinTree**$_n$ be two different tree shapes with the same number of leaves and non-zero probability. Then

$$P_{2n}^*(T_n * T_n') = \binom{2n}{n} q(n,n) P_n^*(T_n) \cdot P_n^*(T_n')$$
$$P_{2n}^*(T_n * T_n) = \frac{1}{2}\binom{2n}{n} q(n,n) P_n^*(T_n) \cdot P_n^*(T_n)$$

and therefore there does no exist a single real number $q^*(n,n)$ such that, for every $T_1, T_2 \in$ **BinTree**$_n$,
$$P_{2n}^*(T_1 * T_2) = q^*(n,n) P_n^*(T_1) \cdot P_n^*(T_2).$$

### 1.3.3 Some probabilistic models

In this section we will introduce several probabilistic models of phylogenetic trees that will be used in this work.

#### Chen-Ford-Winkel's $\alpha$-$\gamma$-model

This probabilistic model of phylogenetic trees was introduced in [18], and it will turn out that the next three probabilistic models for bifurcating phylogenetic trees that we shall describe are instances of it altough they historically preceded it. As it is inferred by its name, this model depends on two parameters $(\alpha, \gamma) \in [0,1]^2$ such that $\alpha \geq \gamma$, and it is born as a generalization of Ford's $\alpha$-model for bifurcating phylogenetic trees (which we shall describe anon) that allows the generation of random multifurcating trees. Its definition is purely algorithmic and we recall it in Algorithm 1 (for $n \geq 2$: for $n = 1$, the only tree in **PhyloTree**$_1$ has, of course, probability 1). We provide an example of application of this algorithm in Lemma 1.26.

**Algorithm 1:** $\alpha$-$\gamma$-model

```
   Input  : $n \in \mathbb{N}_{\geq 2}$
   Output: $T_n \in$ PhyloTree$_n$ and its probability $P_{\alpha,\gamma,n}(T_n)$
1  $m = 2$;
2  start with a single cherry $T_2$ labelled on [2] and $P_{\alpha,\gamma,2}(T_2) = 1$;
3  while $m < n$ do
4      from $T_m$, choose:
5          – either a pendant edge $e$, each one with probability $\frac{1-\alpha}{m-\alpha}$,
6          – or an internal edge $e$, each one with probability $\frac{\gamma}{m-\alpha}$,
7          – or an internal node $u$, each one with probability $\frac{(\deg_{\mathrm{out}}(u)-1)\alpha-\gamma}{m-\alpha}$,
8          – or to add a new root, with probability $\frac{\gamma}{m-\alpha}$ ;
9      if we have chosen an edge e then
10         split $e$ to create a new node;
11         add a new leaf adjacent to this node, with label $m+1$;
12     end
13     if we have chosen a node u then
14         add a new leaf adjacent to $u$, with label $m+1$;
15     end
16     if we have chosen to add a new root then
17         add a new root $\rho_{T_{m+1}}$ whose children are a leaf labelled $m+1$ and the
             former root of $T_m$;
18     end
19     let $T_{m+1} \in$ PhyloTree$_{m+1}$ be the resulting tree;
20     set $P_{\alpha,\gamma,m+1}(T_{m+1})$ equal to $P_{\alpha,\gamma,m}(T_m)$ multiplied by the probability of the
         choice in line 4;
21     $m = m + 1$;
22 end
23 return $T_n$ and $P_{\alpha,\gamma,n}(T_n)$;
```

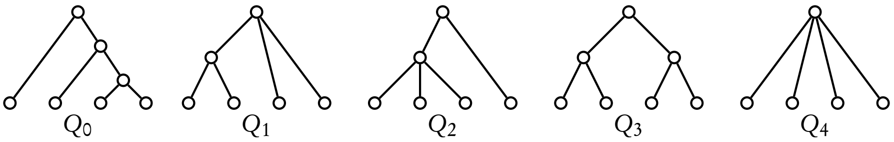


Figure 1.7: The five tree shapes in **Tree**$_4$.

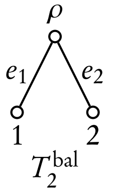
Now, it turns out that the $\alpha$-$\gamma$-model is not shape invariant in general, but the probabilistic model for trees $(P^*_{\alpha,\gamma,n})_n$ it induces is sampling consistent (Theorem 2 in [18]).

In Chapter 5 we shall need to know the probabilities under $P^*_{\alpha,\gamma,4}$ of the five different trees in **Tree**$_4$, described in Figure 1.7 together with the notations that we shall use in that chapter; notice that $Q_0 = T_4^{\mathrm{cat}}$, the caterpillar with four leaves, $Q_3 = T_4^{\mathrm{bal}}$, the maximally balanced tree with 4 leaves, and $Q_4 = T_4^{\mathrm{star}}$, the star with 4 leaves. We compute these probabilities in the following lemma, thus providing an example of explicit computation of probabilities for this model.

**Lemma 1.26.** *With the notations of Figure 1.7:*

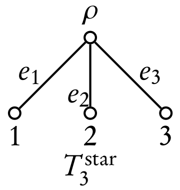
$$P^*_{\alpha,\gamma,4}(Q_0) = \frac{2(1-\alpha+\gamma)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q_1) = \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q_2) = \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q_3) = \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}$$
$$P^*_{\alpha,\gamma,4}(Q_4) = \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$

*Proof.* We begin by considering the cherry in **PhyloTree**$_2$



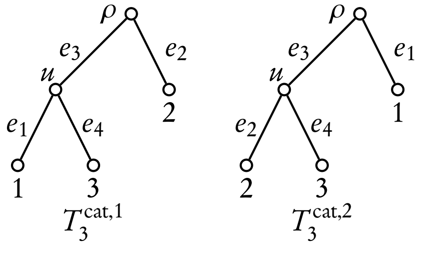
Since **PhyloTree**$_2$ contains only this tree, $P_{\alpha,\gamma,2}(T_2^{\text{bal}}) = 1$. We add now to it a leaf labelled with 3. We have several ways to do it:

- The probability of choosing the root, and then pending from it the leaf 3, is $\frac{\alpha-\gamma}{2-\alpha}$, and therefore the tree



has probability $P_{\alpha,\gamma,3}(T_3^{\text{star}}) = \frac{\alpha-\gamma}{2-\alpha}$.

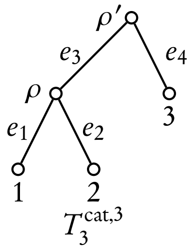
- The probabilities of choosing either $e_1$ or $e_2$ are $\frac{1-\alpha}{2-\alpha}$, since both edges are adjacent to a leaf. Choosing one or the other adding to it the leaf 3 we obtain the following trees:



Their probability is, then

$$P_{\alpha,\gamma,3}(T_3^{\text{cat},1}) = P_{\alpha,\gamma,3}(T_3^{\text{cat},2}) = \frac{1-\alpha}{2-\alpha}.$$

- The probability of choosing to add a new root $\rho'$ is $\frac{\gamma}{2-\alpha}$. The resulting tree is



Its probability is

$$P_{\alpha,\gamma,3}(T_3^{\text{cat},3}) = \frac{\gamma}{2-\alpha}.$$

From these probabilities, and considering only the tree shapes, we deduce that

$$\begin{aligned} P^*_{\alpha,\gamma,3}(T_3^{\text{star}}) &= P_{\alpha,\gamma,3}(T_3^{\text{star}}) = \frac{\alpha-\gamma}{2-\alpha} \\ P^*_{\alpha,\gamma,3}(T_3^{\text{cat}}) &= P_{\alpha,\gamma,3}(T_3^{\text{cat},1}) + P_{\alpha,\gamma,3}(T_3^{\text{cat},2}) + P_{\alpha,\gamma,3}(T_3^{\text{cat},3}) \\ &= 2\cdot\frac{1-\alpha}{2-\alpha} + \frac{\gamma}{2-\alpha} = \frac{2+\gamma-2\alpha}{2-\alpha}. \end{aligned}$$

Repeating the same process with each tree, we shall derive all phylogenetic trees in **PhyloTree**$_4$.

- The star $(1,2,3,4)$ is obtained by adding the leaf 4 to the root of the star $T_3^{\text{star}}$. Its probability is, then,

$$\frac{2\alpha-\gamma}{3-\alpha}\cdot P_{\alpha,\gamma,3}(T_3^{\text{star}}) = \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$

and since it is the only tree of shape $Q_4$, we conclude that

$$P^*_{\alpha,\gamma,4}(Q_4) = \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

- The different phylogenetic trees of shape $Q_0$, that is, the caterpillars, can be obtained as follows:

- $(((1,2),3),4)$ is obtained from $T_3^{\text{cat},3} = ((1,2),3)$ by adding to it a new root and then the leaf 4 pending from it. Its probability is, then,

$$\frac{\gamma}{3-\alpha}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},3}) = \frac{\gamma^2}{(2-\alpha)(3-\alpha)}.$$

- $(((1,3),2),4)$ and $((2,3),1),4)$ are obtained from $T_3^{\text{cat},1} = ((1,3),2)$ and $T_3^{\text{cat},2} = ((2,3),1)$, respectively, by adding to them a new root and then the leaf 4 pending from it. Their probabilities are, then,

$$\frac{\gamma}{3-\alpha}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},i}) = \frac{\gamma(1-\alpha)}{(2-\alpha)(3-\alpha)}, \quad i\in\{1,2\}.$$

- $(((1,2),4),3)$ is obtained from $T_3^{\text{cat},3}$ by adding the leaf 4 to its internal edge $e_3$. Its probability is, then,

$$\frac{\gamma}{3-\alpha}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},3})=\frac{\gamma^2}{(2-\alpha)(3-\alpha)}.$$

- $(((1,3),4),2)$ and $((2,3),4),1)$ are obtained from $T_3^{\text{cat},1}$ and $T_3^{\text{cat},2}$, respectively, by adding the leaf 4 to their internal edge $e_3$. Their probabilities are

$$\frac{\gamma}{3-\alpha}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},i})=\frac{\gamma(1-\alpha)}{(2-\alpha)(3-\alpha)},\quad i\in\{1,2\}.$$

- $(((1,4),2),3)$ and $((2,4),1),3)$ are obtained from $T_3^{\text{cat},3}$ by adding the leaf 4 to its pendant edges $e_1$ or $e_2$, respectively. Their probabilities are then

$$\frac{(1-\alpha)}{(3-\alpha)}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},3})=\frac{(1-\alpha)\gamma}{(2-\alpha)(3-\alpha)}.$$

- $(((1,4),3),2)$ and $((3,4),1),2)$ are obtained from $T_3^{\text{cat},1}$ by adding the leaf 4 to its pendant edges $e_1$ or $e_4$, respectively, and $(((2,4),3),1)$ and $((3,4),2),1)$ are obtained from $T_3^{\text{cat},2}$ by adding the leaf 4 to to its pendant edges $e_2$ or $e_4$, respectively. Therefore, their probabilities are

$$\frac{(1-\alpha)}{(3-\alpha)}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},i})=\frac{(1-\alpha)^2}{(2-\alpha)(3-\alpha)},\quad i\in\{1,2\}.$$

By adding up all these probabilities, we obtain

$$P^*_{\alpha,\gamma,4}(Q_0)=\frac{2(1-\alpha+\gamma)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}.$$

- The six phylogenetic trees of shape $Q_1$ are obtained as follows:

- The trees $((1,4),2,3)$, $((2,4),1,3)$ and $((3,4),1,2)$ are obtained by adding the leaf 4 to one of the three edges in the tree $T_3^{\text{star}}$, all of them pendant. Their probabilities are then

$$\frac{(1-\alpha)}{(3-\alpha)}\cdot P_{\alpha,\gamma,4}(T_3^{\text{star}})=\frac{(1-\alpha)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

- The tree $((1,2),3,4)$ is obtained by adding the leaf 4 to the root of the tree $T_3^{\text{cat},3}$. Its probability is then

$$\frac{(\alpha-\gamma)}{(3-\alpha)}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},3})=\frac{(\alpha-\gamma)\gamma}{(3-\alpha)(2-\alpha)}.$$

- The trees $((1,3),2,4)$ and $((2,3),1,4)$ are obtained by adding the leaf 4 to the root of the trees $T_3^{\text{cat},1}$ and $T_3^{\text{cat},2}$, respectively. Their probabilities are then

$$\frac{(\alpha-\gamma)}{(3-\alpha)}\cdot P_{\alpha,\gamma,4}(T_3^{\text{cat},i})=\frac{(\alpha-\gamma)(1-\alpha)}{(3-\alpha)(2-\alpha)},\quad i\in\{1,2\}.$$

By adding up all these probabilities, we obtain

$$P^*_{\alpha,\gamma,4}(Q_1) = \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

- The four phylogenetic trees of shape $Q_2$ are obtained as follows:

– The tree $((1,2,3),4)$ is obtained by adding a new root to $T_3^{\text{star}}$ and the leaf 4 pending from it. Its probability is, then,

$$\frac{\gamma}{3-\alpha} \cdot P_{\alpha,\gamma,3}(T_3^{\text{star}}) = \frac{\gamma(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

– The tree $((1,2,4),3)$ is obtained by adding the leaf 4 to the root of the cherry in $T_3^{\text{cat},3}$. Its probability is, then,

$$\frac{\alpha-\gamma}{3-\alpha} \cdot P_{\alpha,\gamma,3}(T_3^{\text{cat},3}) = \frac{\gamma(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

– The trees $((1,3,4),2)$ and $((2,3,4),1)$ are obtained by adding the leaf 4 to the root of the cherry in $T_3^{\text{cat},1}$ and $T_3^{\text{cat},2}$, respectively. Their probability is, then,

$$\frac{\alpha-\gamma}{3-\alpha} \cdot P_{\alpha,\gamma,3}(T_3^{\text{cat},i}) = \frac{(\alpha-\gamma)(1-\alpha)}{(3-\alpha)(2-\alpha)}, \quad i \in \{2,3\}.$$

By adding up all these probabilities, we obtain

$$P^*_{\alpha,\gamma,4}(Q_2) = \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}.$$

- Finally, the three phylogenetic trees of shape $Q_3$ are obtained as follows:

– The tree $((1,2),(3,4))$ is obtained from $T_3^{\text{cat},3}$ by adding the leaf 4 to the pendant edge $e_4$. Its probability is, then,

$$\frac{1-\alpha}{3-\alpha} \cdot P_{\alpha,\gamma,3}(T_3^{\text{cat},3}) = \frac{(1-\alpha)\gamma}{(3-\alpha)(2-\alpha)}.$$

– The trees $((1,3),(2,4))$ and $((2,3),(1,4))$ are obtained by adding the leaf 4 to the pendant edges $e_2$ and $e_1$ in $T_3^{\text{cat},1}$ and $T_3^{\text{cat},2}$, respectively. Their probability is, then,

$$\frac{1-\alpha}{3-\alpha} \cdot P_{\alpha,\gamma,3}(T_3^{\text{cat},i}) = \frac{(1-\alpha)^2}{(3-\alpha)(2-\alpha)}, \quad i \in \{2,3\}.$$

By adding up these probabilities, we obtain

$$P^*_{\alpha,\gamma,4}(Q_3) = \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}.$$

□

Notice in particular the different probabilities of the caterpillars $T_3^{\text{cat},i}$ imply that the $\alpha$-$\gamma$-model is not shape invariant unless $\gamma = 1-\alpha$.

### Ford's $\alpha$-model

Algorithm 1 presents a way to produce multifurcating trees randomly. However, if we want to define a probabilistic model that only generates bifurcating trees, it suffices to set $\alpha = \gamma$, and so the probability of adding a new leaf as a child of an internal node at step 4 becomes 0, because in this case the probability of adding a new leaf child to a node of out-degree 2 is 0, and we start with a cherry. If we, moreover, after producing a tree with $n$ leaves, relabel the tree equiprobably and we take as the probability of a phylogenetic tree the probability of producing it in this way, we obtain Ford's $\alpha$-model [43], which we shall denote by $(P_{\alpha,n})_n$. But it should be mentioned that, historically, Ford's $\alpha$-model is older than the $\alpha$-$\gamma$-model is, and the latter was born actually as a multifurcating generalization of the former. Indeed, for tree shapes, $P^*_{\alpha,\alpha,n} = P^*_{\alpha,n}$. For the sake of completeness, we provide a full algorithmic description of this model in Algorithm 2. Figure 1.8 gives the probability of the pair of bifurcating trees in **BinTree**$_4$ under $P^*_{\alpha,4}$ obtained from Lemma 1.26 (taking $\alpha = \gamma$). In the next example we compute these probabilities directly with Algorithm 2, as an example of its application, and we shall also deduce them later from Lemma 1.27 below.

**Algorithm 2:** $\alpha$-model

```
   Input  : $n \in \mathbb{N}_{\geq 2}$
   Output: $T_n \in$ BinPhyloTree$_n$ and its probability $P_{\alpha,n}(T_n)$
1  $m = 2$;
2  start with a single cherry $T_2$ labelled on $[2]$ and $P'_{\alpha,2}(T_2) = 1$;
3  while $m < n$ do
4      from $T_m$, choose:
5          – either a pendant edge $e$, each one with probability $\frac{1-\alpha}{m-\alpha}$,
6          – or an internal edge $e$, each one with probability $\frac{\alpha}{m-\alpha}$,
7          – or to add a new root, with probability $\frac{\alpha}{m-\alpha}$ ;
8      if we have chosen an edge e then
9          split $e$ to create a new node;
10         add a new leaf labelled with $m + 1$ adjacent to this node;
11     end
12     if we have chosen to add a new root then
13         add a new root $\rho_{T_{m+1}}$ whose children are a leaf labelled with $m + 1$ and
            the former root of $T_m$;
14     end
15     let $T_{m+1} \in$ PhyloTree$_{m+1}$ be the resulting tree;
16     set $P'_{\alpha,m+1}(T_{m+1})$ equal to $P'_{\alpha,\gamma,m}(T_m)$ multiplied by the probability of the
        choice in line 4;
17     $m = m + 1$;
18 end
19 set $P^*_{\alpha,n}(\pi_1(T_n)) = \sum_{T'_n \in \pi_1^{-1}(\pi_1(T_n))} P'_{\alpha,n}(T'_n)$;
20 set $P_{\alpha,n}(T_n) = \frac{2^{s(T_n)}}{n!} P^*_{\alpha,n}(\pi_1(T_n))$;
21 return $T_n$ and $P_{\alpha,n}(T_n)$;
```

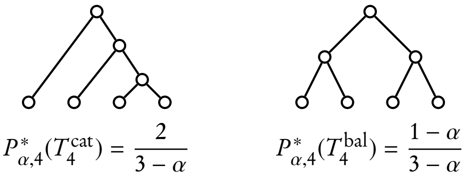

$P^*_{\alpha,4}(T_4^{\mathrm{cat}}) = \dfrac{2}{3-\alpha}$
$P^*_{\alpha,4}(T_4^{\mathrm{bal}}) = \dfrac{1-\alpha}{3-\alpha}$

Figure 1.8: The two tree shapes in **BinTree**$_4$ and their probabilities under $P^*_{\alpha,4}$.

**Example:**

Let us determine completely $P^*_{\alpha,4}$. Since **BinPhyloTree**$_4 = \{T_4^{\mathrm{cat}}, T_4^{\mathrm{bal}}\}$, it is enough to compute $P^*_{\alpha,4}(T_4^{\mathrm{bal}})$, and to do that, with the notations used in Algorithm 2, we need to compute the $P'_{\alpha,4}$ value of the three maximally balanced phylogenetic trees with 4 leaves.

We begin with the cherry $T_2 = (1,2) \in$ **BinPhyloTree**$_2$, with probability 1. From it we obtain the phylogenetic trees $T_3^{\mathrm{cat},1} = ((1,3),2)$ and $T_3^{\mathrm{cat},2} = ((2,3),1)$ by adding the leaf 3 to one of its edges, so that

$$P'_{\alpha,3}(T_3^{\mathrm{cat},1}) = P'_{\alpha,3}(T_3^{\mathrm{cat},2}) = \frac{1-\alpha}{2-\alpha},$$

and we obtain $T_3^{\mathrm{cat},3} = ((1,2),3)$ by adding to $T_2$ the leaf 3 as the child of a new root, and hence

$$P'_{\alpha,3}(T_3^{\mathrm{cat},3}) = \frac{\alpha}{2-\alpha}.$$

Let's add now the leaf 4. On the one hand, $((1,2),(3,4))$ is obtained from $T_3^{\mathrm{cat},3}$ by adding the leaf 4 to the pendant arc ending in the leaf 3, and therefore

$$P'_{\alpha,4}\big(((1,2),(3,4))\big) = \frac{1-\alpha}{3-\alpha}P'_{\alpha,3}(T_3^{\mathrm{cat},3}) = \frac{\alpha(1-\alpha)}{(2-\alpha)(3-\alpha)}.$$

On the other hand, $((1,3),(2,4))$ is obtained from $T_3^{\mathrm{cat},1}$ by adding the leaf 4 to the pendant arc ending in the leaf 2, and therefore

$$P'_{\alpha,4}\big(((1,3),(2,4))\big) = \frac{1-\alpha}{3-\alpha}P'_{\alpha,3}(T_3^{\mathrm{cat},1}) = \frac{(1-\alpha)^2}{(2-\alpha)(3-\alpha)}$$

and a similar argument shows that

$$P'_{\alpha,4}\big(((2,3),(1,4))\big) = \frac{1-\alpha}{3-\alpha}P'_{\alpha,3}(T_3^{\mathrm{cat},2}) = \frac{(1-\alpha)^2}{(2-\alpha)(3-\alpha)}.$$

Therefore,

$$\begin{aligned} P^*_{\alpha,4}(T_4^{\mathrm{bal}}) &= P'_{\alpha,4}\big(((1,2),(3,4))\big) + P'_{\alpha,4}\big(((1,3),(2,4))\big) + P'_{\alpha,4}\big(((2,3),(1,4))\big) \\ &= \frac{\alpha(1-\alpha)}{(2-\alpha)(3-\alpha)} + 2\frac{(1-\alpha)^2}{(2-\alpha)(3-\alpha)} = \frac{1-\alpha}{3-\alpha} \qquad (1.3) \end{aligned}$$

as shown in Figure 1.8. Then,

$$P^*_{\alpha,4}(T_4^{\text{cat}}) = 1 - P^*_{\alpha,4}(T_4^{\text{bal}}) = \frac{2}{3-\alpha}.$$

Finally, the actual probability of each one of the 3 fully symmetric phylogenetic trees in **BinPhyloTree**$_4$ under the $\alpha$-model is one third the probability of their shape,

$$P_{\alpha,4}\big(((1,2),(3,4))\big) = P_{\alpha,4}\big(((1,3),(2,4))\big) = P_{\alpha,4}\big(((2,3),(1,4))\big) = \frac{1-\alpha}{3(3-\alpha)},$$

and the actual probability of each one of the twelve caterpillars in **BinPhyloTree**$_4$ is

$$\frac{1}{12}P^*_{\alpha,4}(T_4^{\text{cat}}) = \frac{1}{6(3-\alpha)}.$$

Ford's $\alpha$-model for phylogenetic trees is both shape invariant (by line 20 in Algorithm 2) and sampling consistent (Proposition 42 in [43]). Moreover, it is Markovian, but Ford's proof of this fact is wrong, and we provided a correct proof in [23], which is included below.

Set $q_\alpha : \mathbb{N}^2_{\geq 1} \to \mathbb{R}$ to be [43, Lemma 27]

$$q_\alpha(a,b) = \frac{\Gamma_\alpha(a)\Gamma_\alpha(b)}{\Gamma_\alpha(a+b)} \cdot \varphi_\alpha(a,b), \tag{1.4}$$

where

$$\varphi_\alpha(a,b) = \frac{\alpha}{2}\binom{a+b}{a} + (1-2\alpha)\binom{a+b-2}{a-1}$$

and $\Gamma_\alpha : \mathbb{N}_{\geq 1} \to \mathbb{R}$ is the mapping defined by $\Gamma_\alpha(1) = 1$ and, for every $n \geq 2$, $\Gamma_\alpha(n) = (n-1-\alpha)\cdot\Gamma_\alpha(n-1)$. In other words, $\Gamma_\alpha(n) = (1-\alpha)_{n-1}$ where $(a)_m$ is the *Pochhammer symbol* defined as

$$(a)_m = \begin{cases} 1 & \text{if } k = 0 \\ a(a+1)\cdots(a+m-1) & \text{if } m \in \mathbb{N}_{\geq 1} \end{cases} \tag{1.5}$$

For every internal node $v$ in an bifurcating tree $T$, we call its *numerical split* the ordered pair $\text{NS}_T(v) = (\kappa_T(v_1), \kappa_T(v_2))$, where $\text{child}(v) = \{v_1, v_2\}$ with $\kappa_T(v_1) \geq \kappa_T(v_2)$. The *multiset of numerical splits* of $T$ is $\text{NS}(T) = \{\text{NS}_T(v) : v \in \mathring{V}(T)\}$. For instance (cf. Figure 1.3)

$$\text{NS}(T_7^{\text{bal}}) = \{(1,1),(1,1),(1,1),(2,2),(2,1),(4,3)\}.$$

The following lemma provides an explicit formula for $P_{\alpha,n}(T,\lambda)$, for every $n \geq 1$ and $(T,\lambda) \in$ **BinPhyloTree**$_n$.

**Lemma 1.27.** *For every $(T,\lambda) \in$ **BinPhyloTree**$_n$, its probability under the $\alpha$-model is*

$$P_{\alpha,n}(T,\lambda) = \frac{2^{n-1}}{n! \cdot \Gamma_\alpha(n)} \prod_{(a,b)\in \text{NS}(T)} \varphi_\alpha(a,b).$$

*Proof.* To prove this result, we shall need to make a detour on the set of bifurcating ordered trees, which shall only be used in this proof. An *ordered tree* is a pair $(T, <_T)$ with $T \in \mathbf{Tree}$ and $<_T = (<_v)_{v\in \mathring{V}(T)}$ where, for every $v \in \mathring{V}(T)$, $<_v$ is an ordering on child$(v)$. Let $\mathbf{BinOrdTree}_n$, $n \geq 1$, be the set of bifurcating ordered trees with $n$ leaves. The root join of ordered trees is defined as usual, but with the addition that the ordering $<_\rho$ on the set of the roots of the maximal pending subtrees of the result is given by the order in which the trees are joined.

We can extend a probabilistic model of bifurcating trees, and in particular the $\alpha$-model, to a *probabilistic model of ordered trees* $P^o_{\alpha,n} : \mathbf{BinOrdTree}_n \to [0,1]$ by simply equally splitting the probability of a tree $T$ among all ordered trees $(T, <_T)$ on this shape. There are $2^{n-1-s(T)}$ such ordered trees (from the $2^{n-1}$ possible ways to define the vector of orderings $(<_v)_{v\in \mathring{V}(T)}$, those differing only on the orderings on the children of the $s(T)$ symmetry nodes are actually the same ordered tree), and therefore

$$P^o_{\alpha,n}(T, <_T) = \frac{1}{2^{n-1-s(T)}} P^*_{\alpha,n}(T). \tag{1.6}$$

Ford proves (correctly) that $(P^o_{\alpha,n})_n$ is Markovian, and more specifically that, for every $0 < k < n$ and for every $(T_k, <_{T_k}) \in \mathbf{BinOrdTree}_k$ and $(T_{n-k}, <_{T_{n-k}}) \in \mathbf{BinOrdTree}_{n-k}$,

$$P^o_{\alpha,n}((T_k, <_{T_k}) * (T_{n-k}, <_{T_{n-k}})) = q_\alpha(k, n-k) P^o_{\alpha,k}(T_k, <_{T_k}) P^o_{\alpha,n-k}(T_{n-k}, <_{T_{n-k}})$$

with $q_\alpha$ defined as in (1.4), from where he deduces that, for every $(T_n, <_{T_n}) \in \mathbf{BinOrdTree}_n$,

$$P^o_{\alpha,n}(T_n, <_{T_n}) = \prod_{(a,b)\in \mathrm{NS}(T)} q_\alpha(a,b). \tag{1.7}$$

For the proofs of these two facts, see Lemma 27 and Proposition 28 in [43], respectively.

Now, given $(T, \lambda) \in \mathbf{BinPhyloTree}_n$, consider an ordered tree $(T, <_T)$ obtained by forgetting about the labels in $T$ and adding an ordering on each child$(v)$, $v \in \mathring{V}(T)$. Then, by the shape invariance and Equations (1.6) and (1.7),

$$P_{\alpha,n}(T, \lambda) = \frac{2^{s(T)}}{n!} \cdot P^*_{\alpha,n}(T) = \frac{2^{s(T)}}{n!} \cdot 2^{n-s(T)-1} \cdot P^o_{\alpha,n}(T, <_T) = \frac{2^{n-1}}{n!} \prod_{(a,b)\in \mathrm{NS}(T)} q_\alpha(a,b).$$

It remains to simplify this product. If, for every $v \in \mathring{V}(T)$, we denote its children by $v_1$ and $v_2$, then

$$\prod_{(a,b)\in \mathrm{NS}(T)} q_\alpha(a,b) = \prod_{v\in \mathring{V}(T)} \frac{\Gamma_\alpha(\kappa_T(v_1))\Gamma_\alpha(\kappa_T(v_2))}{\Gamma_\alpha(\kappa_T(v))} \varphi_\alpha(\mathrm{NS}(v)).$$

For every $v \in \mathring{V}(T) \setminus \{\rho_T\}$, the term $\Gamma_\alpha(\kappa_T(v))$ appears twice in this product: in the denominator of the factor corresponding to $v$ itself and in the numerator of the factor corresponding to its parent. Therefore, all terms $\Gamma_\alpha(\kappa_T(v))$ in this product vanish except $\Gamma_\alpha(\kappa_T(\rho_T)) = \Gamma_\alpha(n)$ (that appears in the denominator of its factor) and every $\Gamma_\alpha(\kappa_T(v)) = \Gamma_\alpha(1) = 1$ with $v \in L(T)$. Thus,

$$P_{\alpha,n}(T) = \frac{2^{n-1}}{n!} \cdot \frac{1}{\Gamma_\alpha(n)} \cdot \prod_{v\in \mathring{V}(T)} \varphi_\alpha(\mathrm{NS}(v))$$

as we claimed. □

> **Example:**
> Since $\mathrm{NS}(T_4^{\mathrm{bal}}) = \big\{(1,1),(1,1),(2,2)\big\}$ and $\mathrm{NS}(T_4^{\mathrm{cat}}) = \big\{(1,1),(2,1),(3,1)\big\}$,
>
> $$P^*_{\alpha,4}(T_4^{\mathrm{bal}}) = 3\cdot\frac{2^3}{4!}\cdot\frac{\varphi_\alpha(1,1)^2\varphi_\alpha(2,2)}{\Gamma_\alpha(4)} = \frac{(1-\alpha)^2(2-\alpha)}{(3-\alpha)(2-\alpha)(1-\alpha)} = \frac{1-\alpha}{3-\alpha}$$
>
> $$P^*_{\alpha,4}(T_4^{\mathrm{cat}}) = 12\cdot\frac{2^3}{4!}\cdot\frac{\varphi_\alpha(1,1)\varphi_\alpha(2,1)\varphi_\alpha(3,1)}{\Gamma_\alpha(4)} = 4\cdot\frac{(1-\alpha)(1-\frac{1}{2}\alpha)\cdot 1}{(3-\alpha)(2-\alpha)(1-\alpha)} = \frac{2}{3-\alpha}$$
>
> in agreement with Figure 1.8.

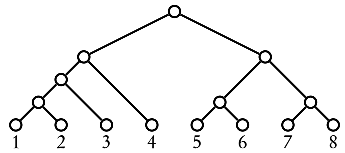


Figure 1.9: The phylogenetic tree $\widetilde{T}$ used in Remark 1.28.

**Remark 1.28.** Ford states (see [43, Prop. 32 and page 30]) that if $(T,\lambda)\in\mathbf{BinPhyloTree}_n$, then

$$P_{\alpha,n}(T,\lambda) = \frac{2^{s}(T)}{n!}\prod_{(a,b)\in\mathrm{NS}(T)}\widetilde{q}_\alpha(a,b)$$

where

$$\widetilde{q}_\alpha(a,b) = \begin{cases} 2q_\alpha(a,b) & \text{if } a\neq b\\ q_\alpha(a,b) & \text{if } a=b\end{cases}$$

If we simplify $\prod_{(a,b)\in\mathrm{NS}(T)}\widetilde{q}_\alpha(a,b)$ as in the proof of Lemma 1.27, this formula for $P_{\alpha,n}(T,\lambda)$ becomes

$$P_{\alpha,n}(T,\lambda) = \frac{2^{s(T)+m}}{n!\cdot\Gamma_\alpha(n)}\cdot\prod_{(a,b)\in\mathrm{NS}(T)}\varphi_\alpha(a,b) \tag{1.8}$$

where $m$ is the number of internal nodes whose children have different numbers of descendant leaves. This formula does not agree with the one given in Lemma 1.27 above, because

$$s(T)+m = n-1-\big|\{v\in\mathring{V}(T) : \mathrm{child}(v)=\{v_1,v_2\} \text{ and } \kappa_T(v_1)=\kappa_T(v_2)\\ \text{but } \pi_1(T_{v_1})\neq\pi_1(T_{v_2})\}\big|$$

and, hence, it may happen that $s(T)+m<n-1$. The first example of a phylogenetic tree with this property (and the only one, up to relabelings, with at most eight leaves) is the tree $\widetilde{T}\in\mathbf{BinPhyloTree}_8$ depicted in Fig. 1.9. For it, our formula gives

$$P_{\alpha,8}(\widetilde{T}) = \frac{(1-\alpha)^2(2-\alpha)}{126(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$$

while (1.8) assigns to $\widetilde{T}$ a probability of half this value:

$$\frac{(1-\alpha)^2(2-\alpha)}{252(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}. \tag{1.9}$$

This last figure cannot be right. For one reason, as we shall see anon, when $\alpha = \frac{1}{2}$, Ford's model is equivalent to the Uniform model, where every phylogenetic tree in $\mathbf{BinPhyloTree}_n$ has the same probability

$$\frac{1}{|\mathbf{BinPhyloTree}_n|} = \frac{1}{(2n-3)!!}.$$

In particular, $P_{\frac{1}{2},8}(\widetilde{T})$ should be equal to $1/135135$. This figure is consistent with our formula, while expression (1.9) yields half this value.

And now we can prove the Markovianity of the $\alpha$-model of phylogenetic trees.

**Theorem 1.29.** *Let $n \geq 2$, $1 \leq k \leq n-1$, $T_k \in \mathbf{BinTree}_k$, $T_{n-k} \in \mathbf{BinTree}_{n-k}$, and $(T_k * T_{n-k}, \lambda) \in \mathbf{BinPhyloTree}_n$. Then,*

$$P_{\alpha,n}(T_k * T_{n-k}, \lambda) = 2\frac{q_\alpha(k,n-k)}{\binom{n}{k}} P_{\alpha,k}(T_k, \lambda|_{T_k}) P_{\alpha,n-k}(T_{n-k}, \lambda|_{T_{n-k}}).$$

*Proof.* If $T_k \in \mathbf{BinTree}_k$ and $T_{n-k} \in \mathbf{BinTree}_{n-k}$ and we set $\lambda_k = \lambda|_{T_k}$ and $\lambda_{n-k} = \lambda|_{T_{n-k}}$, then

$$\begin{aligned}
P_{\alpha,k}(T_k, \lambda_k) &= \frac{2^{k-1}}{k!\Gamma_\alpha(k)} \prod_{(a,b)\in \mathrm{NS}(T_k)} \varphi_\alpha(a,b) \\
P_{\alpha,n-k}(T_{n-k}, \lambda_{n-k}) &= \frac{2^{n-k-1}}{(n-k)!\Gamma_\alpha(n-k)} \prod_{(a,b)\in \mathrm{NS}(T_{n-k})} \varphi_\alpha(a,b)
\end{aligned}$$

and

$$\begin{aligned}
P_{\alpha,n}(T_k * T_{n-k}, \lambda) &= \frac{2^{n-1}}{n!\Gamma_\alpha(n)} \prod_{(a,b)\in \mathrm{NS}(T_k * T_{n-k})} \varphi_\alpha(a,b) \\
&= \frac{2^{n-1}}{n!\Gamma_\alpha(n)} \varphi_\alpha(k,n-k) \Big( \prod_{(a,b)\in \mathrm{NS}(T_k)} \varphi_\alpha(a,b) \Big) \Big( \prod_{(a,b)\in \mathrm{NS}(T_{n-k})} \varphi_\alpha(a,b) \Big) \\
&= \frac{2^{n-1}}{n!\Gamma_\alpha(n)} \varphi_\alpha(k,n-k) \frac{k!\Gamma_\alpha(k)}{2^{k-1}} P_{\alpha,k}(T_k, \lambda_k) \\
&\qquad \cdot \frac{(n-k)!\Gamma_\alpha(n-k)}{2^{n-k-1}} P_{\alpha,n-k}(T_{n-k}, \lambda_{n-k}) \\
&= \frac{2q_\alpha(k,n-k)}{\binom{n}{k}} P_{\alpha,k}(T_k, \lambda_k) P_{\alpha,n-k}(T_{n-k}, \lambda_{n-k})
\end{aligned}$$

as we claimed. □

### The Yule model

The *Yule model* (also called *Yule-Harding model* or *Equal Rates Markov (ERM) model*) is probably the oldest probabilistic model of phylogenetic trees found in the literature, and it dates back to the original Yule paper in 1922 [127]. It can be considered to be an instance of Ford's $\alpha$-model by setting $\alpha = 0$, which makes all new leaves to be added at pendant edges (because in the $\alpha$-model the probability of chosing an internal edge of an intermediate tree $T_m$ is $\alpha/(m-\alpha)$). This model has a variety of interesting properties, and has been widely studied [2, 12, 57, 109]. Intuitively, in biological terms it expresses that in a certain phase of the evolutionary process, when either an speciation or an extinction occurs, it is equally likely to occur to any of the species extant at that moment [2].

Being a special case of the $\alpha$-model, the Yule model is both shape invariant and sampling consistent (Prop. 42 in [43]). By Lemma 1.27 with $\alpha = 0$,

$$\begin{aligned}
P_{\text{Yule},n}(T,\lambda) &= \frac{2^{n-1}}{n!\cdot\Gamma_0(n)}\prod_{(a,b)\in\text{NS}(T)}\varphi_0(a,b) = \frac{2^{n-1}}{n!(n-1)!}\prod_{(a,b)\in\text{NS}(T)}\binom{a+b-2}{a-1}\\
&= \frac{2^{n-1}}{n!(n-1)!}\prod_{v\in\mathring{V}(T)}\frac{(\kappa_T(v)-2)!}{(\kappa_T(v_1)-1)!(\kappa_T(v_2)-1)!}\\
&= \frac{2^{n-1}}{n!(n-1)!}\prod_{v\in\mathring{V}(T)}\frac{(\kappa_T(v)-1)!}{(\kappa_T(v)-1)(\kappa_T(v_1)-1)!(\kappa_T(v_2)-1)!}\\
&\text{(where, for every } v\in\mathring{V}(T),\ \text{child}(v)=\{v_1,v_2\})\\
&= \frac{2^{n-1}}{n!(n-1)!}\cdot(n-1)!\prod_{v\in\mathring{V}(T)}\frac{1}{\kappa_T(v)-1} = \frac{2^{n-1}}{n!}\prod_{v\in\mathring{V}(T)}\frac{1}{\kappa_T(v)-1}
\end{aligned} \tag{1.10}$$

(where the second last equality is obtained by reasoning as in the last paragraph of the proof of Lemma 1.27). And then, by the shape invariance,

$$\begin{aligned}
P^*_{\text{Yule},n}(T) &= \frac{n!}{2^{s(T)}}\cdot P_{\text{Yule},n}(T,\lambda) = \frac{n!}{2^{s(T)}}\cdot\frac{2^{n-1}}{n!}\prod_{v\in\mathring{V}(T)}\frac{1}{\kappa_T(v)-1}\\
&= 2^{n-1-s(T)}\prod_{v\in\mathring{V}(T)}\frac{1}{\kappa_T(v)-1}.
\end{aligned} \tag{1.11}$$

Finally, since

$$q_0(a,b) = \frac{\Gamma_0(a)\Gamma_0(b)}{\Gamma_0(a+b)}\cdot\varphi_0(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}\cdot\binom{a+b-2}{a-1} = \frac{1}{a+b-1},$$

Theorem 1.29 says, in the Yule model, that, for every $n\geq 2$, $1\leq k\leq n-1$, $T_k\in$ **BinTree**$_k$, $T_{n-k}\in$ **BinTree**$_{n-k}$, and $(T_k * T_{n-k},\lambda)\in$ **BinPhyloTree**$_n$,

$$P_{\text{Yule},n}(T_k * T_{n-k},\lambda) = \frac{2}{(n-1)\binom{n}{k}}P_{\text{Yule},k}(T_k,\lambda|_{T_k})P_{\text{Yule},n-k}(T_{n-k},\lambda|_{T_{n-k}}). \tag{1.12}$$

### The Uniform model

The *Uniform model* (also called *Proportional-to-Distinguishable Arrangements (PDA) model* in biological studies) is easier to define: in this model, each phylogenetic tree $(T, \lambda) \in$ $\mathbf{PhyloTree}_n$ is considered to be equiprobable. In the bifurcating case, which is the one that concerns us here, this says that every $(T, \lambda) \in \mathbf{BinPhyloTree}_n$ has probability

$$P_{\mathrm{unif},n}(T, \lambda) = \frac{1}{(2n-3)!!}.$$

It turns out that this is the probability of a phylogenetic tree under the $\alpha$-model when $\alpha = \frac{1}{2}$ [43, §3.2]. Indeed,

$$\varphi_{\frac{1}{2}}(a, b) = \frac{1}{4}\binom{a+b}{a}$$

and

$$\Gamma_{1/2}(n) = \Big(n-1-\frac{1}{2}\Big)\Big(n-2-\frac{1}{2}\Big)\cdots\frac{1}{2} = \frac{(2n-3)!!}{2^{n-1}}$$

and hence, by Lemma 1.27,

$$\begin{aligned}
P_{\frac{1}{2},n}(T, \lambda) &= \frac{2^{n-1}}{n! \cdot \Gamma_{1/2}(n)} \prod_{(a,b)\in \mathrm{NS}(T)} \varphi_{\frac{1}{2}}(a, b) = \\
&= \frac{2^{n-1} \cdot 2^{n-1}}{n! \cdot (2n-3)!!} \prod_{(a,b)\in \mathrm{NS}(T)} \frac{1}{4}\binom{a+b}{a} = \frac{1}{n! \cdot (2n-3)!!} \cdot \prod_{v \in \mathring{V}(T)} \frac{\kappa_T(v)!}{\kappa_T(v_1)!\kappa_T(v_2)!} \\
&\quad \text{(where, for every } v \in \mathring{V}(T), \text{child}(v) = \{v_1, v_2\}) \\
&= \frac{1}{n! \cdot (2n-3)!!} \cdot n! = \frac{1}{(2n-3)!!}.
\end{aligned}$$

Notice that if in Algorithm 2 we take $\alpha = \frac{1}{2}$, all possible places to add the new leaf in each repetition of the **while** loop have the same probability. Therefore, while in the Yule model at each step a pendant edge is chosen equiprobably as the place to add the new leaf, we can understand the Uniform model as an algorithmic model where at each step some edge on any type (or a new root) is chosen equiprobably as the place to add the new leaf.

With respect to the Markovianity of the Uniform model, since

$$\begin{aligned}
q_{\frac{1}{2}}(a, b) &= \frac{\Gamma_{1/2}(a)\Gamma_{1/2}(b)}{\Gamma_{1/2}(a+b)} \cdot \varphi_{\frac{1}{2}}(a, b) = \frac{(2a-3)!!(2b-3)!!2^{a+b-1}}{(2(a+b)-3)!!2^{a-1}2^{b-1}} \cdot \frac{1}{4}\binom{a+b}{a} \\
&= \frac{(2a-3)!!(2b-3)!!}{2 \cdot (2(a+b)-3)!!} \cdot \binom{a+b}{a},
\end{aligned}$$

Theorem 1.29 says, when $\alpha = \frac{1}{2}$, that, for every $n \geq 2$, $1 \leq k \leq n-1$, $T_k \in \mathbf{BinTree}_k$, $T_{n-k} \in \mathbf{BinTree}_{n-k}$, and $(T_k * T_{n-k}, \lambda) \in \mathbf{BinPhyloTree}_n$,

$$P_{\mathrm{unif},n}(T_k * T_{n-k}, \lambda) = \frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!} P_{\mathrm{unif},k}(T_k, \lambda|_{T_k}) P_{\mathrm{unif},n-k}(T_{n-k}, \lambda|_{T_{n-k}}), \tag{1.13}$$

which, of course, is also consequence of the uniform probabilities

$$P_{\mathrm{unif},n}(T_k * T_{n-k}, \lambda) = \frac{1}{(2n-3)!!},$$
$$P_{\mathrm{unif},k}(T_k, \lambda|_{T_k}) = \frac{1}{(2k-3)!!},$$
$$P_{\mathrm{unif},n-k}(T_{n-k}, \lambda_{n-k}) = \frac{1}{(2(n-k)-3)!!}.$$

### Aldous' $\beta$-model

The *$\beta$-splitting model* (or *$\beta$-model*, for short) $(P^A_{\beta,n})_n$ [2, 3] is a probabilistic model of bifurcating phylogenetic trees due to D. Aldous that depends on one parameter $\beta \in (-2,\infty)$, and although we present it last, it preceded the $\alpha$-model for ten years. Let us recall its definition. For every $m \geq 2$ and $a = 1,\dots,m-1$, let

$$q_{m,\beta}(a) = \frac{1}{a_m(\beta)} \cdot \frac{\Gamma(\beta+a+1)\Gamma(\beta+m-a+1)}{\Gamma(a+1)\Gamma(m-a+1)},$$

where $\Gamma$ stands for the usual Gamma function defined on $\mathbb{R}_{>0}$,

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

and $a_m(\beta)$ is a suitable normalizing constant so that $\sum\limits_{a=1}^{m-1} q_{m,\beta}(a) = 1$. Recall that $\Gamma$ satisfies that $\Gamma(x+1) = x\Gamma(x)$ and that, for every $n \in \mathbb{N}_{\geq 1}$, $\Gamma(n+1) = n!$.

For every $m \geq 2$ and $a \in \{1,\dots,\lfloor m/2 \rfloor\}$, let

$$\widehat{q}_{m,\beta}(a) = \begin{cases} q_{m,\beta}(a) + q_{m,\beta}(m-a) = 2q_{m,\beta}(a) & \text{if } a \neq m/2 \\ q_{m,\beta}(a) & \text{if } a = m/2 \end{cases}$$

With these notations, the probabilities under this model are computed by means of Algorithm 3.

The last step in the definition of $P^A_{\beta,n}$ makes it shape invariant by construction, and in [2] it is proved that it is sampling consistent. Hence, the $\beta$-model of trees $P^{A,*}_{\beta,n}$ is also sampling consistent. Moreover, $(P^A_{\beta,n})_n$ is also Markovian, with split distribution $\widehat{q}_{n,\beta}$ [2, 77]. This $\beta$-model includes as specific cases the Yule model (when $\beta = 0$) and the Uniform model (when $\beta = -3/2$), and these are the only values of $\beta$ for which the $\beta$-model is equal to some $\alpha$-model [43, Thm. 43].

In Chapter 5 we shall need to know the probability under this model of the maximally balanced tree with four leaves, $P^{A,*}_{\beta,4}(T^{\mathrm{bal}}_4)$. We compute this probability in the following lemma, and in this way we provide a detailed example of how this model associates probabilities to trees through their construction.

**Lemma 1.30.** *For every $\beta \in (-2,\infty)$,*

$$P^{A,*}_{\beta,4}(T^{\mathrm{bal}}_4) = \frac{3\beta+6}{7\beta+18}.$$

**Algorithm 3:** $\beta$-model

```
Input  : $n \in \mathbb{N}_{\geq 1}$
Output: $T_n \in$ BinPhyloTree$_n$ and its probability $P^A_{\beta,n}(T_n)$
$j = 1$;
start with $T'_1$ a single node labelled with $n$ and $P'_{\beta,1}(T'_1) = 1$;
while $j < n$ do
    from $T'_j$, choose equiprobably a leaf with a label $m > 1$;
    choose a number $a \in \{1, \ldots, \lfloor m/2 \rfloor\}$ with probability distribution $\widehat{q}_{m,\beta}(a)$;
    split the leaf into a cherry with the leaves labelled $a$ and $m - a$;
    let $T_{j+1} \in$ BinPhyloTree$_{j+1}$ be the resulting tree and set
      $P'_{\beta,j+1}(T'_{j+1}) = \frac{\widehat{q}_{m,\beta}(a)}{|\{\text{leaves in } T'_j \text{ labeled} > 1\}|} \cdot P'_{\beta,j}(T'_j)$;
    $j = j + 1$;
end
set $P^{A,*}_{\beta,n}(\pi_1(T_n)) = \sum_{T'_n \in \pi_1^{-1}(\pi_1(T_n))} P'_{\beta,n}(T'_n)$;
set $P^A_{\beta,n}(T_n) = \frac{2^{s(T_n)}}{n!} P^{A,*}_{\beta,n}(\pi_1(T_n))$;
return $T_n$ and $P^A_{\beta,n}(T_n)$;
```

*Proof.* We start with a single node labeled 4. In order to obtain a maximally balanced tree $((1,1),(1,1))$ using Algorithm 3, in the first iteration we must split this node into a cherry with both leaves labeled 2. The probability of choosing this split is

$$\widehat{q}_{4,\beta}(2) = q_{4,\beta}(2) = \frac{1}{a_4(\beta)} \cdot \frac{\Gamma(\beta+3)\Gamma(\beta+3)}{\Gamma(3)\Gamma(3)}.$$

Let us compute the normalizing constant $a_4(\beta)$: since

$$q_{4,\beta}(1) = q_{4,\beta}(3) = \frac{1}{a_4(\beta)} \cdot \frac{\Gamma(\beta+2)\Gamma(\beta+4)}{\Gamma(2)\Gamma(4)}$$
$$q_{4,\beta}(2) = \frac{1}{a_4(\beta)} \cdot \frac{\Gamma(\beta+3)\Gamma(\beta+3)}{\Gamma(3)\Gamma(3)}$$

imposing that $q_{4,\beta}(1) + q_{4,\beta}(2) + q_{4,\beta}(3) = 1$ we obtain

$$a_4(\beta) = \frac{2\Gamma(\beta+2)\Gamma(\beta+4)}{6} + \frac{\Gamma(\beta+3)^2}{4} = \frac{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}{12}.$$

Therefore,

$$P'_{\beta,2}((2,2)) = q_{4,\beta}(2) = \frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2}$$

In the second iteration, we choose one of the leaves with probability $1/2$ and we split it into a cherry $(1,1)$. Since there is only one way of splitting a leaf labeled 2, $q_{2,\beta}(1) = 1$. So, the probability of the tree obtained in this step is

$$P'_{\beta,2}(((1,1),2)) = \frac{1}{2} P'_{\beta,2}((2,2)) = \frac{3\Gamma(\beta+3)^2}{2(4\Gamma(\beta+2)\Gamma(\beta+4) + 3\Gamma(\beta+3)^2)}$$

Then, in the third step, we are forced to choose the other leaf labeled 2 and to split it into a cherry $(1, 1)$. We obtain a maximally balanced tree with all its leaves labeled 1 and its probability is still

$$P'_{\beta,2}(((1,1),(1,1))) = P'_{\beta,2}(((1,1),2)) = \frac{3\Gamma(\beta+3)^2}{2(4\Gamma(\beta+2)\Gamma(\beta+4)+3\Gamma(\beta+3)^2)}$$

Now, there are two ways of obtaining the tree $((1,1),(1,1))$ with this construction, depending on which leaf of the cherry $(2,2)$ we choose to split first. So, the probability of the tree $T_4^{\mathrm{bal}}$ is

$$P_{\beta,4}^{A,*}(T_4^{\mathrm{bal}}) = 2P'_{\beta,2}(((1,1),(1,1))) = \frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4)+3\Gamma(\beta+3)^2}$$

Finally, using that $\Gamma(x+1) = x\Gamma(x)$, we have that

$$\begin{aligned}&\frac{3\Gamma(\beta+3)^2}{4\Gamma(\beta+2)\Gamma(\beta+4)+3\Gamma(\beta+3)^2}\\ &\qquad = \frac{3(\beta+2)^2\Gamma(\beta+2)^2}{4(\beta+3)(\beta+2)\Gamma(\beta+2)^2+3(\beta+2)^2\Gamma(\beta+2)^2} = \frac{3\beta+6}{7\beta+18}\end{aligned}$$

as we claimed. □

### 1.3.4 Binary recursive shape indices, revisited

Consider the random variables $C_n$, $S_n$ and $\Phi_n$ that take a binary phylogenetic tree $T \in \mathbf{BinPhyloTree}_n$ and compute $S(T)$, $C(T)$ and $\Phi(T)$, respectively. Previously to the contributions of our Thesis, the following facts on their expected values under the Yule model (denoted by $E_{\mathrm{Yule}}$) and the Uniform model (denoted by $E_{\mathrm{unif}}$), as well as on their variance under these models (denoted by $\sigma^2_{\mathrm{Yule}}$ and $\sigma^2_{\mathrm{unif}}$, respectively) were known:

- For the Colless index:

$$\begin{aligned}&E_{\mathrm{Yule}}(C_n) = n(H_{\lfloor n/2\rfloor}-1)+\lceil n/2\rceil-\lfloor n/2\rfloor && [60]\\ &E_{\mathrm{unif}}(C_n) \sim \sqrt{\pi}n^{3/2} && [8]\\ &\sigma^2_{\mathrm{Yule}}(C_n) = (5n^2+7n)/2+(6n+1)\lfloor n/2\rfloor-4\lfloor n/2\rfloor^2+8\lfloor (n+2)/4\rfloor^2\\ &\qquad +(2\lfloor n/2\rfloor-n(n-3))H_{\lfloor n/2\rfloor}-8(n+1)\lfloor (n+2)/4\rfloor-6nH_n && [13]\\ &\sigma^2_{\mathrm{unif}}(C_n) \sim \left(\frac{10-3\pi}{3}\right)n^3 && [8]\end{aligned}$$

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

- For the Sackin index:

$$\begin{aligned}
E_{\text{Yule}}(S_n) &= 2n(H_n - 1) && [69]\\
E_{\text{unif}}(S_n) &= n\left(\frac{(2n-2)!!}{(2n-3)!!} - 1\right) && [85]\\
\sigma^2_{\text{Yule}}(S_n) &= 7n^2 - n - 2nH_n - 4n^2H_n^{(2)} && [13]\\
\sigma^2_{\text{unif}}(S_n) &\sim \left(\frac{10-3\pi}{3}\right)n^3 && [8]
\end{aligned}$$

where $H_n^{(2)} = \sum_{i=1}^{n} 1/i^2$, the *second order n-th harmonic number*.

- For the cophenetic index

$$\begin{aligned}
E_{\text{Yule}}(\Phi_n) &= n(n+1) - 2nH_n && [85]\\
E_{\text{unif}}(\Phi_n) &= \frac{1}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!} - 2\right) && [85]\\
\sigma^2_{\text{Yule}}(\Phi_n) &= \frac{n^4 - 10n^3 + 131n^2 - 2n}{12} - 6nH_n - 4n^2H_n^{(2)} && [13]
\end{aligned}$$

As a byproduct of our computations, we shall be able to find closed formulæ for $\sigma^2_{\text{unif}}(S_n)$ and $\sigma^2_{\text{unif}}(\Phi_n)$ (see Chaper 4).

One of the key ingredients in the derivation of these, and other, formulæ will be the fact that the combination of a shape invariant Markovian probabilistic model of bifurcating phylogenetic trees $(P_n)_n$ and a binary recursive shape index $I$ allows the obtention of recurrences for several moments of the latter under the former through the following lemma.

**Lemma 1.31.** *Let $I, J$ : **BinPhyloTree** $\to \mathbb{R}$ be binary recursive shape indices and $(P_n)_n$ a shape invariant Markovian probabilistic model of phylogenetic trees, with conditional split distribution $q_P : \mathbb{N}_{\geq 1} \times \mathbb{N}_{\geq 1} \to \mathbb{R}$. Set*

$$Q_P(k, n-k) = \frac{1}{2}\binom{n}{k} q_P(k, n-k).$$

*For every $n \geq 1$, let $I_n$ and $J_n$ be the random variables that choose a tree $T \in$ **BinPhyloTree**$_n$ with probability $P_n(T)$ and compute $I(T)$ and $J(T)$, respectively. Then, for every $n \geq 2$,* *the expected values of $I_n$, $I_nJ_n$, and $I_n^2$ under $(P_n)_n$ satisfy, respectively:*

$$E_P(I_n) = \sum_{k=1}^{n-1} Q_P(k, n-k)\big(2E_P(I_k) + f_I(k, n-k)\big) \tag{1.14}$$

$$\begin{aligned} E_P(I_nJ_n) = \sum_{k=1}^{n-1} Q_P(k, n-k)\Big(&2E_P(I_kJ_k) + 2E_P(I_k)E_P(J_{n-k}) \\ &+ 2f_I(k, n-k)E_P(J_k) + 2f_J(k, n-k)E_P(I_k) \\ &+ f_I(k, n-k)f_J(k, n-k)\Big) \end{aligned} \tag{1.15}$$

$$\begin{aligned} E_P(I_n^2) = \sum_{k=1}^{n-1} Q_P(k, n-k)\Big(&2E_P(I_k^2) + 2E_P(I_k)E_P(I_{n-k}) \\ &+4f_I(k, n-k)E_P(I_k) + f_I(k, n-k)^2\Big). \end{aligned} \tag{1.16}$$

*Proof.* We shall prove the first two equations, because the third one is a particular case of the second, taking $I = J$. The key idea in these recurrences is the fact that every $T \in \mathbf{BinPhyloTree}_n$ can be obtained by choosing a number $k$ of leaves between 1 and $n-1$, a subset $\Lambda_k \in \mathrm{Part}_k([n])$, a phylogenetic tree $T_k \in \mathbf{BinPhyloTree}(\Lambda_k)$ and a phylogenetic tree $T_{n-k} \in \mathbf{BinPhyloTree}(\Lambda_k^c)$, where $\Lambda_k^c = [n] \setminus \Lambda_k$, and then taking their root join $T_k * T'_{n-k}$. Actually, every $T \in \mathbf{BinPhyloTree}_n$ is obtained twice in this way, depending on whether the result of our first choice of set of labels turns out to be $\Lambda_k$ or $\Lambda_k^c$.

So, to prove (1.14), we develop $E_P(I_n)$, for $n \geq 2$, as follows:

$$\begin{aligned} E_P(I_n) &= \sum_{T \in \mathbf{BinPhyloTree}_n} I(T) \cdot P_n(T) \\ &= \frac{1}{2}\sum_{k=1}^{n-1} \sum_{\Lambda_k \in \mathrm{Part}_k([n])} \sum_{T_k \in \mathbf{BinPhyloTree}(\Lambda_k)} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}(\Lambda_k^c)} I(T_k * T_{n-k}) \cdot P_n(T_k * T_{n-k}) \\ &= \frac{1}{2}\sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} I(T_k * T_{n-k}) \cdot P_n(T_k * T_{n-k}) \\ &\quad \text{(by the shape invariance of } I \text{ and } (P_n)_n) \\ &= \frac{1}{2}\sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} \Big(I(T_k) + I(T_{n-k}) \\ &\qquad + f_I(k, n-k)\Big) q_P(k, n-k) P_k(T_k) P_{n-k}(T_{n-k}) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n-1} Q_P(k, n-k)\Big( \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} I(T_k) P_k(T_k) P_{n-k}(T_{n-k}) \\
&\qquad + \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} I(T_{n-k}) P_k(T_k) P_{n-k}(T_{n-k}) \\
&\qquad + \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} f_I(k, n-k) P_k(T_k) P_{n-k}(T_{n-k}) \Big) \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k)\Bigg[ \Big( \sum_{T_k \in \mathbf{BinPhyloTree}_k} I(T_k) P_k(T_k) \Big) \Big( \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} P_{n-k}(T_{n-k}) \Big) \\
&\qquad + \Big( \sum_{T_k \in \mathbf{BinPhyloTree}_k} P_k(T_k) \Big) \Big( \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} I(T_{n-k}) P_{n-k}(T_{n-k}) \Big) \\
&\qquad + f_I(k, n-k) \Big( \sum_{T_k \in \mathbf{BinPhyloTree}_k} P_k(T_k) \Big) \Big( \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} P_{n-k}(T_{n-k}) \Big) \Bigg] \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \big( E_P(I_k) + E_P(I_{n-k}) + f_I(k, n-k) \big) \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \big( 2E_P(I_k) + f_I(k, n-k) \big),
\end{aligned}
$$

where in the last step we have used the symmetry of $Q_P$.

The proof of (1.15) is similar:

$$
\begin{aligned}
E_P(I_n J_n) &= \sum_{T \in \mathbf{BinPhyloTree}_n} I(T) J(T) \cdot P_n(T) \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \sum_{\Lambda_k \in \mathrm{Part}_k([n])} \sum_{T_k \in \mathbf{BinPhyloTree}(\Lambda_k)} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}(\Lambda_k^c)} I(T_k * T_{n-k}) J(T_k * T_{n-k}) P_n(T_k * T_{n-k}) \\
&= \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} \big( I(T_k) + I(T_{n-k}) + f_I(k, n-k) \big) \\
&\qquad \cdot \big( J(T_k) + J(T_{n-k}) + f_J(k, n-k) \big) q_P(k, n-k) P_k(T_k) P_{n-k}(T_{n-k}) \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} \big( I(T_k) + I(T_{n-k}) + f_I(k, n-k) \big) \\
&\qquad \cdot \big( J(T_k) + J(T_{n-k}) + f_J(k, n-k) \big) P_k(T_k) P_{n-k}(T_{n-k}) \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} \Big( I(T_k) J(T_k) + I(T_{n-k}) J(T_{n-k}) \\
&\qquad + I(T_k) J(T_{n-k}) + I(T_{n-k}) J(T_k) + f_J(k, n-k) I(T_k) \\
&\qquad + f_J(k, n-k) I(T_{n-k}) + f_I(k, n-k) J(T_k) + f_I(k, n-k) J(T_{n-k}) \\
&\qquad + f_I(k, n-k) f_J(k, n-k) \Big) P_k(T_k) P_{n-k}(T_{n-k})
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \Bigg( \sum_{T_k \in \mathbf{BinPhyloTree}_k} \sum_{T_{n-k} \in \mathbf{BinPhyloTree}_{n-k}} \Big( I(T_k)J(T_k)P_k(T_k)P_{n-k}(T_{n-k}) \\
&\quad + I(T_{n-k})J(T_{n-k})P_k(T_k)P_{n-k}(T_{n-k}) + I(T_k)J(T_{n-k})P_k(T_k)P_{n-k}(T_{n-k}) \\
&\quad + I(T_{n-k})J(T_k)P_k(T_k)P_{n-k}(T_{n-k}) + f_J(k,n-k)I(T_k)P_k(T_k)P_{n-k}(T_{n-k}) \\
&\quad + f_J(k,n-k)I(T_{n-k})P_k(T_k)P_{n-k}(T_{n-k}) + f_I(k,n-k)J(T_k)P_k(T_k)P_{n-k}(T_{n-k}) \\
&\quad + f_I(k,n-k)J(T_{n-k})P_k(T_k)P_{n-k}(T_{n-k}) + f_I(k,n-k)f_J(k,n-k)P_k(T_k)P_{n-k}(T_{n-k}) \Big) \Bigg) \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \Big( E_P(I_k J_k) + E_P(I_{n-k} J_{n-k}) + E_P(I_k)E_P(J_{n-k}) \\
&\quad + E_P(I_{n-k})E_P(J_k) + f_J(k,n-k)E_P(I_k) + f_J(k,n-k)E_P(I_{n-k}) \\
&\quad + f_I(k,n-k)E_P(J_k) + f_I(k,n-k)E_P(J_{n-k}) + f_I(k,n-k)f_J(k,n-k) \Big) \\
&= \sum_{k=1}^{n-1} Q_P(k, n-k) \Big( 2E_P(I_k J_k) + 2E_P(I_k)E_P(J_{n-k}) \\
&\quad + 2f_J(k,n-k)E_P(I_k) + 2f_I(k,n-k)E_P(J_k) + f_I(k,n-k)f_J(k,n-k) \Big)
\end{aligned}
$$

using the symmetry of $Q_P(k, n-k)$, $f_I(k, n-k)$, and $f_J(k, n-k)$. □

Under the Yule model, by Equation (1.12)

$$Q_{\text{Yule}}(k, n-k) = \frac{1}{n-1} \tag{1.17}$$

and under the Uniform model, by Equation (1.13)

$$Q_{\text{unif}}(k, n-k) = \frac{1}{2}\binom{n}{k}\frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}. \tag{1.18}$$

To simplify the notations, we shall denote henceforth $Q_{\text{unif}}(k, n-k)$ by $C_{k,n-k}$: so,

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

## 1.4 Hypergeometric series

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

$$\sum_{k\geq 0} t_k z^k \in \mathbb{R}[[z]]$$

such that $t_0 = 1$ and there exist polynomials $P, Q \in \mathbb{R}[x]$ such that

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

In this case, the polynomials $P, Q$ are called the *hypergeometric polynomials* of the series.

**Example:**

The following series are all hypergeometric:

- $\sum_{k\geq 0} z^k$. In this case, $t_k = 1$, and
$$\frac{t_{k+1}}{t_k} = \frac{1}{1} = \frac{P(k)}{Q(k)}$$
with $P(k) = Q(k) = 1$.
- $\sum_{k\geq 0} k!z^k$. Here, $t_k = k!$ and thus
$$\frac{t_{k+1}}{t_k} = \frac{(k+1)!}{k!} = \frac{k}{1} = \frac{P(k)}{Q(k)}$$
with $P(k) = k$ and $Q(k) = 1$.

A *hypergeometric function* is a function represented by a hypergeometric series. The theory of hypergeometric functions is powerful and has been thoroughly developed ever since the first systematic study by Gauss in 1813.

Let $\sum_{k\geq 0} t_k z^k$ be a hypergeometric series and suppose that its hypergeometric polynomials $P, Q \in \mathbb{R}[x]$ are monic and with all their roots in $\mathbb{R}$, and that we are given them factorized, so that
$$\frac{P(k)}{Q(k)} = \frac{(k+a_1)\cdots(k+a_p)}{(k+b_1)\cdots(k+b_q)(k+1)}$$
with $a_1, \dots, a_p, b_1, \dots, b_q \in \mathbb{R}$ (notice that, if need be, we can always impose a term $k+1$ in both the denominator *and* the numerator). Then, we denote this hypergeometric series $\sum_{k\geq 0} t_k z^k$ by means of the *hypergeometric function* [92]
$${}_pF_q\left[\begin{matrix} a_1, & \dots, & a_p \\ b_1, & \dots, & b_q \end{matrix}; z\right].$$
If $P, Q$ are not monic, and, say,
$$\frac{P(k)}{Q(k)} = \frac{c(k+a_1)\cdots(k+a_p)}{d(k+b_1)\cdots(k+b_q)(k+1)}$$
with $a_1, \dots, a_p, b_1, \dots, b_q, c, d \in \mathbb{R}$, then
$$\sum_{k\geq 0} t_k z^k = {}_pF_q\left[\begin{matrix} a_1, & \dots, & a_p \\ b_1, & \dots, & b_q \end{matrix}; \frac{cz}{d}\right].$$
These functions help us establish a standard representation for the hypergeometric series. The function above is showed to be equal to [92]
$${}_pF_q\left[\begin{matrix} a_1, & \dots, & a_p \\ b_1, & \dots, & b_q \end{matrix}; z\right] = \sum_{k\geq 0} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_q)_k k!} z^k,$$
where $(a)_k$ is the already defined *Pochhammer symbol* (see (1.5) in page 35). The domain of this function is the set of values of $z$ for which the series converges:

- If $p < q + 1$, this domain is the whole $\mathbb{R}$.
- If $p = q + 1$, this domain lies between the open interval $(-1, 1)$ and its closure $[-1, 1]$; the cases $z = 1, -1$ must be treated separately.
- If $p > q + 1$, the series only converges when $z = 0$.

Therefore, in the sequel we shall always assume that $p \leq q + 1$.

This procedure of standarization of hypergeometric series (combined, if necessary, with any necessary modification to ensure that $t_0 = 1$) is referred to as the *lookup algorithm* in [92]. Broadly speaking, there are many results proved for hypergeometric functions with small $p$ and $q$, and that is the reason why it is important to identify and, if possible, relate a given hypergeometric function to a series.

**Example:**

Let us consider the series

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

To understand it as a hypergeometric series, we must shift the summation index so that it starts with $k = 0$ and then to extract $t_0$ as a common factor if it is different from 1:

$$\sum_{k\geq 1} \frac{(2k+1)!}{(3k-2)!} = \sum_{k\geq 0} \frac{(2k+3)!}{(3k+1)!} = 6 \sum_{k\geq 0} \frac{(2k+3)!}{6\cdot(3k+1)!}$$

We focus now on the series

$$\sum_{k\geq 0} \frac{(2k+3)!}{6\cdot(3k+1)!}$$

with $t_k = \frac{(2k+3)!}{6\cdot(3k+1)!}$. It satisfies that $t_0 = 1$ and

$$\begin{aligned}\frac{t_{k+1}}{t_k} &= \frac{\frac{(2k+5)!}{6\cdot(3k+4)!}}{\frac{(2k+3)!}{6\cdot(3k+1)!}} = \frac{(2k+5)!(3k+1)!}{(3k+4)!(2k+3)!} = \frac{(2k+5)(2k+4)}{(3k+4)(3k+3)(3k+2)}\\ &= \frac{4(k+\frac{5}{2})(k+2)}{27(k+\frac{4}{3})(k+\frac{2}{3})(k+1)}\end{aligned}$$

Therefore

$$\sum_{k\geq 0} \frac{(2k+3)!}{6\cdot(3k+1)!} = {}_2F_2\left[\begin{matrix}5/2,\ 2\\ 4/3,\ 2/3\end{matrix}; \frac{4}{27}\cdot 1\right]$$

and finally

$$\sum_{k\geq 1} \frac{(2k+1)!}{(3k-2)!} = 6\cdot {}_2F_2\left[\begin{matrix}5/2,\ 2\\ 4/3,\ 2/3\end{matrix}; \frac{4}{27}\right].$$

What is the interest of finding an identity like this? Unfortunately, in this case none whatsoever; but had we been lucky, we could have applied one of the more than 31,000 formulæ on ${}_2F_2$ known to the *Wolfram Function Site* (`https://functions.wolfram.com`) to compute the value on the right-hand side.

**Remark 1.32.** The inclusion of $k+1$ as a concrete factor is of no further mathematical relevance but, as it is often the case, is due to the *mores* of the mathematical community. In regard to this aspect, notice that in the second expression for the function ${}_pF_q$, the $k!$ term in the denominator is just $(1)_k$.

> **Example:**
>
> Many well-known functions of classical analysis have a hypergeometric expansion. Take, for instance, the exponential function *par excellence*: $e^x$. We know that its series expansion is
>
> $$e^x = \sum_{k\geq 0} \frac{x^k}{k!}.$$
>
> Thus, $t_k = \frac{1}{k!}$, and therefore $\frac{t_{k+1}}{t_k} = \frac{1}{k+1}$; thence
>
> $$e^x = {}_0F_0\left[\begin{matrix}-\\-\end{matrix}; x\right].$$

Imagine that we wish to compute $(-m)_k$ for some $m \in \mathbb{N}_{\geq 1}$; in this case

$$(-m)_k = (-m)(-m+1)\cdots(-m+k-1) = \begin{cases} \frac{(-1)^k m!}{(m-k)!} & \text{if } k \leq m \\ 0 & \text{if } k > m \end{cases}$$

and so there cannot be any hypergeometric function whose parameters $b_1, \ldots, b_q$ contain a negative integer, while any hypergeometric function with some parameter $a_1, \ldots, a_p$ negative is the sum of a finite number of terms.

### 1.4.1 Using hypergeometric series to solve a specific family of recurrences

In this segment we are going to present and prove a rather long theorem —that is the reason why we considered its proof to be presented in this *aparte*. It will be instrumental in some proofs presented in Chapters 3 and 4. As far as we know, its proof is new in the literature, and we have opted to introduce it in the Preliminaries in order not to weighten the aforementioned chapters.

Before proceeding to it, we would like to point out to the reader that, for any $m \in \mathbb{N}$,

$$(2m)!! = (2m)\cdot(2m-2)\cdots 4\cdot 2 = 2^m \cdot m!$$
$$(2m+1)!! = (2m+1)\cdot(2m-1)\cdots 3\cdot 1 = \frac{(2m+2)!}{(2m+2)!!} = \frac{(2m+2)!}{2^{m+1}\cdot(m+1)!},$$

a fact that will be used without further notice in the following proofs. Recall also from page 47 that, for every $n \in \mathbb{N}_{\geq 2}$ and for every $k \in \{1, \ldots, n-1\}$,

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

**Lemma 1.33.** *Let $n \in \mathbb{N}_{\geq 2}$. Then,*

*(i)* $\displaystyle\sum_{k=1}^{n-1} C_{k,n-k} = 1.$

*(ii) For every $m \geq 1$,*

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

*Proof.* We shall begin by proving *(ii)*:

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

by setting $k \mapsto k+m$. This last sum is, in fact, part of a hypergeometric series. Indeed, let

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

then,

$$t_0 = \frac{(2m-2)!(2n-2m-2)!}{(m-1)!(n-m-1)!(n-m)!}, \qquad \frac{t_{j+1}}{t_j} = \frac{\left(j+m-\frac{1}{2}\right)(j+m-n)}{\left(j+m+\frac{3}{2}-n\right)(j+1)}.$$

Now, we would be naturally inclined to deduce that

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

but that is not quite correct. Indeed, since $(m-n)_k = 0$ for $k > n-m$, but

$$(m-n)_{n-m} = (m-n)(m-n+1)\cdots(-1) = (-1)^{n-m}(n-m)! \neq 0$$

and, hence,

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

while we are only interested in the sum up to $n-m-1$. Therefore, what we actually have is that

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

Let us now compute the substrahend in this expression. In order to do that, we shall use the identities:

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

as well as the already stated $(m-n)_{n-m}=(-1)^{n-m}(n-m)!$. Then,

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

It remains to compute the value of the hypergeometric function above. Since $m\le n$, we can apply the identity https://functions.wolfram.com/07.23.03.0003.01 and thus obtain

$${}_2F_1\left[\begin{matrix} m-\frac{1}{2},\ m-n \\ m+\frac{3}{2}-n\end{matrix};1\right] = \frac{(2-n)_{n-m}}{\left(m+\frac{3}{2}-n\right)_{n-m}}.$$

Now, the numerator of this expression is

$$\begin{aligned}(2-n)_{n-m} = (2-n)(3-n)\cdots(1-m) &= \begin{cases} 0 & \text{if } m=1\\ (-1)^{n-m}\frac{(n-2)!}{(m-2)!} & \text{if } m>1\end{cases}\\ &= (-1)^{n-m}(m-1)\frac{(n-2)!}{(m-1)!};\end{aligned}$$

as far as the denominator goes, as we have already seen, $\left(m+\frac{3}{2}-n\right)_{n-m}=(-1)^{n-m-1}2^{m-n}(2n-2m-3)!!$ (as we have already seen). So

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

Therefore, combining all we know,

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

This proves *(ii)* when $1 \leq m \leq n-1$. Now, if $m = n \geq 2$,

$$\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{n} = 0 = \frac{1}{2}\binom{n}{n}\left(1 - \frac{n-1}{n-1}\frac{(2n-3)!!}{(2n-2)!!}\frac{(2n-2)!!}{(2n-3)!!}\right),$$

and when $m > n$

$$\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{m} = 0 = \binom{n}{m}.$$

Thus *(ii)* holds. Finally, let us prove *(i)*. By the symmetry of $C_{k,n-k}$,

$$\sum_{k=1}^{n-1} C_{k,n-k}k = \sum_{k=1}^{n-1} C_{k,n-k}(n-k),$$

and therefore

$$\sum_{k=1}^{n-1} C_{k,n-k} = \frac{2}{n}\sum_{k=1}^{n-1} C_{k,n-k}k = \frac{2}{n}\frac{n}{2} = 1,$$

where the second equality is due to the formula of *(ii)* when $m = 1$. □

**Lemma 1.34.** *Let $n \in \mathbb{N}_{\geq 2}$. Then,*

*(i)* $\displaystyle\sum_{k=1}^{n-1} C_{k,n-k}\frac{(2k-2)!!}{(2k-3)!!} = \frac{1}{2}\frac{(2n-2)!!}{(2n-3)!!} + \frac{1}{4}(2H_{2n-2} - H_{n-1} - 2).$

*(ii)* *For every $m \geq 1$,*

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

*Proof.* As we did before, we begin by first showing *(ii)*. Let us develop the sum:

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

by, again, setting $k \mapsto k+m$. This is also part of a hypergeometric series: indeed, take

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

so that

$$t_0 = \frac{(2n-2m-2)!(m-1)!}{(n-m-1)!(n-m)!}, \qquad \frac{t_{j+1}}{t_j} = \frac{(j+m)(j+m-n)}{\left(j+m-n+\frac{3}{2}\right)(j+1)}.$$

Again, by an analogous argument as that in the proof of Lemma I.33, we get

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

By identity https://functions.wolfram.com/07.23.03.0003.01,

$${}_2F_1\left[\begin{matrix} m,\ m-n \\ m+\frac{3}{2}-n\end{matrix};1\right] = \frac{\left(\frac{3}{2}-n\right)_{n-m}}{\left(\frac{3}{2}+m-n\right)_{n-m}}.$$

The numerator in this expression is

$$\left(\frac{3}{2}-n\right)_{n-m} = \left(\frac{3}{2}-n\right)\left(\frac{5}{2}-n\right)\cdots\left(\frac{3}{2}-m-1\right) = \frac{(-1)^{n-m}(2n-3)!!}{2^{n-m}(2m-3)!!}$$

and, as we have already seen (cf. the proof of Lemma I.33), $\left(m+\frac{3}{2}-n\right)_{n-m} = (-1)^{n-m-1}2^{m-n}(2n-2m-3)!!$, and therefore

$${}_2F_1\left[\begin{matrix} m,\ m-n \\ m+\frac{3}{2}-n\end{matrix};1\right] = -\frac{(2n-3)!!}{(2m-3)!!(2n-2m-3)!!}.$$

As for the substrahend,

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

So, all in all, we have

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

and thus finishes the proof of *(ii)* whenever $1 \leq m \leq n-1$. Now, whenever $m \geq n$,

$$\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{m}\frac{(2k-2)!!}{(2k-3)!!} = 0 = \frac{1}{2}\binom{n}{m}\left(\frac{(2n-2)!!}{(2n-3)!!} - \frac{(2m-2)!!}{(2m-3)!!}\right),$$

since $\binom{k}{m} = 0$ because $k \leq n-1$. So, *(ii)* holds for every $m \geq 1$. Now the only remaining case is *(i)*; i.e., $m = 0$. In this case we deal with the sum

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

by setting $k \mapsto k+1$. This is again part of a hypergeometric series, and thus we repeat the previous procedure. By taking

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

we get

$$t_0 = \frac{(2n-4)!}{(n-2)!(n-1)!}, \qquad \frac{t_{j+1}}{t_j} = \frac{(j+1)^2(j+1-n)}{(j+2)\left(j-n+\frac{5}{2}\right)(j+1)}.$$

Notice, now, that $(1-n)_k = 0$ if $k \geq n$ but that $(1-n)_{n-1} = (-1)^{n-1}(n-1)!$, and so we have that

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

Now, since

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

the substrahend in the expression above is

$$\frac{(1)_{n-2}^2(1-n)_{n-1}}{(2)_{n-1}\left(\frac{5}{2}-n\right)_{n-1}(n-1)!} = \frac{(-1)^{n-1}(n-1)!^3 2^{n-1}}{n!(-1)^{n-2}(2n-5)!!(n-1)!} = -\frac{(2n-2)!!}{n(2n-5)!!}.$$

As for the hypergeometric function, applying transformation (3.1.2) in [48], we obtain

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

where, by identity http://functions.wolfram.com/07.27.03.0017.01,

$$\begin{aligned}
&{}_3F_2\left[\begin{matrix}\frac{3}{2},\ 1,\ \frac{3}{2}-n\\ \frac{5}{2},\ \frac{5}{2}-n\end{matrix};1\right]\\
&\quad= \frac{\left(\frac{3}{2}-n\right)\left(-\frac{1}{2}\right)_n(n-1)!\Gamma\left(\frac{5}{2}\right)\Gamma(1)}{\frac{1}{2}\left(-\frac{1}{2}\right)_n(1)_n\Gamma(1)\Gamma\left(\frac{3}{2}\right)}\sum_{k=0}^{n-1}\frac{\left(-\frac{1}{2}\right)_k(1)_k}{\left(\frac{1}{2}\right)_k k!}\\
&\quad= -\frac{9-6n}{2n}\sum_{k=0}^{n-1}\frac{1}{2k-1} = -\frac{9-6n}{2n}\left(-1+\sum_{j=1}^{2n-2}\frac{1}{j}-\frac{1}{2}\sum_{j=1}^{n-1}\frac{1}{j}\right)\\
&\quad= -\frac{9-6n}{2n}\left(H_{2n-2}-\frac{1}{2}H_{n-1}-1\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
{}_3F_2\left[\begin{matrix}1,\ 1,\ 1-n\\ 2,\ \frac{5}{2}-n\end{matrix};1\right] &= -\frac{2}{3}\frac{9-6n}{2n}\left(H_{2n-2}-\frac{1}{2}H_{n-1}-1\right)\\
&= \frac{3-2n}{2n}(2H_{2n-2}-H_{n-1}-1).
\end{aligned}$$

Finally,

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

as we, sigh, claimed. □

Finally, we state and give proof to the main result of this section: the solution to the following family of recurrences (given by combining the equations in Lemmata 1.33 and 1.34).

**Theorem 1.35.** *The solution to the recurrence*

$$x_n = 2\sum_{k=1}^{n-1} C_{k,n-k} x_k + \sum_{l=1}^{r} a_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{l=1}^{s} b_l \binom{n}{l},$$

*with* $(a_1, \dots, a_r, b_1, \dots, b_s) \in \mathbb{R}^{r+s}$, *is*

$$x_n = \sum_{i=1}^{s+1} \hat{a}_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{l=1}^{r} \hat{b}_l \binom{n}{l},$$

*where*

$$\begin{aligned}
\widehat{a}_1 &= x_1 - a_1 \\
\widehat{a}_l &= \frac{l \cdot (2l-2)!!}{(2l-3)!!}\left(\frac{b_l}{l} + \frac{b_{l-1}}{l-1}\right), \quad l \in \{2, \dots, s\} \\
\widehat{a}_{s+1} &= \frac{(s+1) \cdot (2s)!!}{(2s-1)!!} \cdot \frac{b_s}{s} \\
\widehat{b}_l &= \frac{(2l-3)!!}{(2l-2)!!} \cdot a_l, \quad l \in \{1, \dots, r\}
\end{aligned}$$

*Proof.* Consider the following sequence:

$$x_n = \sum_{i=1}^{s+1} \hat{a}_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{l=1}^{r} \hat{b}_l \binom{n}{l},$$

with $(\hat{a}_1, \dots, \hat{a}_{s+1}, \hat{b}_1, \dots, \hat{b}_r) \in \mathbb{R}^{r+s+1}$ not known. We are going to show that they must be equal to the ones given in the statement of this theorem. Then,

$$\begin{aligned}
x_n &- 2\sum_{k=1}^{n-1} C_{k,n-k} x_k \\
&= \sum_{l=1}^{s+1} \hat{a}_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{l=1}^{r} \hat{b}_l \binom{n}{l} \\
&\quad - 2\sum_{k=1}^{n-1} C_{k,n-k} \left( \sum_{l=1}^{s+1} \hat{a}_l \binom{k}{l} + \sum_{l=1}^{r} \binom{k}{l} \hat{b}_l \frac{(2k-2)!!}{(2k-3)!!} \right) \\
&= \sum_{l=1}^{s+1} \hat{a}_l \left( \binom{n}{l} - 2\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{l} \right) \\
&\quad + \sum_{l=1}^{r} \hat{b}_l \left( \binom{n}{l} \frac{(2n-2)!!}{(2n-3)!!} - 2\sum_{k=1}^{n-1} C_{k,n-k} \binom{k}{l} \frac{(2k-2)!!}{(2k-3)!!} \right) \\
&= \sum_{l=1}^{s+1} \hat{a}_l \frac{l-1}{n-1} \frac{(2l-3)!!}{(2l-2)!!} \binom{n}{l} \frac{(2n-2)!!}{(2n-3)!!} + \sum_{l=1}^{r} \hat{b}_l \binom{n}{l} \frac{(2l-2)!!}{(2l-3)!!}
\end{aligned}$$

by the *(ii)* part of Lemmata 1.33 and 1.34. Then, by writing the equation in the statement of this result, we will prove that

$$\frac{(2n-2)!!}{(2n-3)!!}\sum_{l=1}^{s+1}\frac{(l-1)(2l-3)!!}{(2l-2)!!}\hat{a}_l\binom{n}{l}\frac{1}{n-1}+\sum_{l=1}^{r}\hat{b}_l\frac{(2l-2)!!}{(2l-3)!!}\binom{n}{l}$$
$$=\frac{(2n-2)!!}{(2n-3)!!}\sum_{l=1}^{s}b_l\binom{n}{l}+\sum_{l=1}^{r}a_l\binom{n}{l},$$

by means of satisfying the following conditions:

$$\sum_{l=1}^{r}\hat{b}_l\frac{(2l-2)!!}{(2l-3)!!}\binom{n}{l}=\sum_{l=1}^{r}a_l\binom{n}{l} \tag{1.19}$$

and

$$\sum_{l=1}^{s+1}\frac{(l-1)(2l-3)!!}{(2l-2)!!}\hat{a}_l\binom{n}{l}\frac{1}{n-1}=\sum_{l=1}^{s}b_l\binom{n}{l}. \tag{1.20}$$

Now, by setting

$$\hat{b}_l=\frac{(2l-3)!!}{(2l-2)!!}a_l,\quad l\in\{1,\ldots,d\},$$

we clearly solve 1.19. As for Equation 1.20, if $l\geq 1$ we can easily check that

$$\binom{n}{l}\frac{1}{n-1}=\frac{1}{l}\binom{n}{l-1}-\frac{l-2}{l}\binom{n}{l-1}\frac{1}{n-1}.$$

Now, we shall prove by induction on $l\geq 2$ that

$$\binom{n}{l}\frac{1}{n-1}=\sum_{j=1}^{l-1}(-1)^{j+1}\frac{l-j}{l(l-1)}\binom{n}{l-j}. \tag{1.21}$$

Indeed: suppose that $l=2$; then

$$\binom{n}{2}\frac{1}{n-1}=\frac{1}{2}\binom{n}{1}=\sum_{j=1}(-1)^{j+1}\frac{(l-j)}{l(l-1)}\binom{n}{l-j}.$$

Now suppose that this proposition holds up to $l-1$, for $l\geq 3$. Then,

$$\binom{n}{l-1}\frac{1}{n-1}=\sum_{j=1}^{l-2}(-1)^{j+1}\frac{l-j-1}{(l-1)(l-2)}\binom{n}{l-j-1};$$

we therefore have that

$$\begin{aligned}
\binom{n}{l}\frac{1}{n-1} &= \frac{1}{l}\binom{n}{l-1} - \frac{l-2}{l}\binom{n}{l-1}\frac{1}{n-1}\\
&= \frac{1}{l}\binom{n}{l-1} - \frac{l-2}{l}\sum_{j=1}^{l-2}(-1)^{j+1}\frac{l-1-j}{(l-1)(l-2)}\binom{n}{l-1-j}\\
&\text{(by the induction hypothesis)}\\
&= \frac{l-1}{l(l-1)}\binom{n}{l-1} - \sum_{j=1}^{l-2}(-1)^{j+1}\frac{l-1-j}{l(l-1)}\binom{n}{l-1-j}\\
&= \frac{l-1}{l(l-1)}\binom{n}{l-1} + \sum_{j=2}^{l-1}(-1)^{j+1}\frac{l-j}{l(l-1)}\binom{n}{l-j}\\
&\text{(by setting } l \mapsto l-1)\\
&= \sum_{j=1}^{l-1}(-1)^{j+1}\frac{(l-j)}{l(l-1)}\binom{n}{l-j},
\end{aligned}$$

and we have then established the truth of Equation 1.21. Finally, we can proceed to the end of the proof: Equation 1.20 is then

$$\begin{aligned}
\sum_{l=1}^{s} b_l\binom{n}{l} &= \sum_{l=1}^{s+1}\frac{(l-1)\cdot(2l-3)!!}{(2l-2)!!}\cdot\hat{a}_l\binom{n}{l}\frac{1}{n-1}\\
&= \sum_{l=2}^{s+1}\frac{(l-1)\cdot(2l-3)!!}{(2l-2)!!}\cdot\hat{a}_l\binom{n}{l}\frac{1}{n-1}\\
&= \sum_{l=2}^{s+1}\left(\frac{(l-1)\cdot(2l-3)!!}{(2l-2)!!}\cdot\hat{a}_l\sum_{j=1}^{l-1}(-1)^{j+1}\frac{(l-j)}{l(l-1)}\cdot\binom{n}{l-j}\right)\\
&= \sum_{l=2}^{s+1}\left(\frac{(2l-3)!!}{l\cdot(2l-2)!!}\cdot\hat{a}_l\sum_{h=1}^{l-1}(-1)^{l-h+1}h\binom{n}{h}\right)\\
&\text{(by setting } l-j \mapsto h)\\
&= \sum_{h=1}^{s}\left(\sum_{l=h+1}^{s+1}(-1)^{l-h+1}\frac{h\cdot(2l-3)!!}{l\cdot(2l-2)!!}\cdot\hat{a}_l\right)\binom{n}{h}
\end{aligned}$$

and thus, it is satisfied if

$$\sum_{l=h+1}^{s+1}(-1)^{l-h+1}\frac{h\cdot(2l-3)!!}{l\cdot(2l-2)!!}\cdot\hat{a}_l = b_h, \quad h\in\{1,\dots,s\}.$$

Notice that the system of linear equations in $\hat{a}_2,\dots,\hat{a}_{s+1}$ gives rise to a triangular matrix of the coefficients, and thus it has only one solution. This solution must satisfy that

$$\begin{aligned}
\hat{a}_{s+1} &= \frac{(s+1)\cdot(2s)!!}{s\cdot(2s-1)!!}\cdot b_s \\
\hat{a}_h &= \frac{h\cdot(2h-2)!!}{(h-1)\cdot(2h-3)!!}\cdot b_{h-1} \\
&\qquad - \sum_{l=h+1}^{s+1}(-1)^{l-h}\frac{h\cdot(2h-2)!!(2l-3)!!}{l\cdot(2l-2)!!(2h-3)!!}\cdot\hat{a}_l \\
&= \frac{h\cdot(2h-2)!!}{(2h-3)!!}\Big(\frac{b_{h-1}}{h-1} - \sum_{l=h+1}^{s+1}\frac{(-1)^{l-h}(2l-3)!!}{l\cdot(2l-2)!!}\cdot\hat{a}_l\Big), \quad h\in\{2,\dots,s\}
\end{aligned}$$

Now, let

$$\widetilde{a}_l = \frac{(2l-3)!!}{l\cdot(2l-2)!!}\widehat{a}_l.$$

Then, the previous formulæ can be rewritten as

$$\widetilde{a}_{s+1} = \frac{b_s}{s} \qquad\qquad \widetilde{a}_l = \frac{b_{l-1}}{l-1} + \sum_{h=l+1}^{s+1}(-1)^{h-l-1}\widetilde{a}_h, \quad l\in\{2,\dots,s\}$$

and the solution of the last recurrence is

$$\widetilde{a}_l = \frac{b_{l-1}}{l-1} + \frac{b_l}{l}. \tag{1.22}$$

Indeed:

$$\widetilde{a}_s = \frac{b_{s-1}}{s-1} + \widetilde{a}_{s+1} = \frac{b_{s-1}}{s-1} + \frac{b_s}{s}$$

and if (1.22) holds for every $h\in\{l+1,\dots,s+1\}$, then

$$\widetilde{a}_l = \frac{b_{l-1}}{l-1} + \sum_{h=l+1}^{s}(-1)^{h-l-1}\Big(\frac{b_{h-1}}{h-1} + \frac{b_h}{h}\Big) + (-1)^{s-l}\cdot\frac{b_s}{s} = \frac{b_{l-1}}{l-1} + \frac{b_l}{l}.$$

Then, finally, for every $l\in\{2,\dots,s\}$,

$$\widehat{a}_l = \frac{l\cdot(2l-2)!!}{(2l-3)!!}\widetilde{a}_l = \frac{l\cdot(2l-2)!!}{(2l-3)!!}\Big(\frac{b_{l-1}}{l-1} + \frac{b_l}{l}\Big)$$

as we claimed.

Finally, to obtain $\hat{a}_1$ we impose the initial condition

$$\begin{aligned}
x_1 = \sum_{l=1}^{s+1}\hat{a}_l\binom{1}{l} + \frac{(2-2)!!}{(2-3)!!}\sum_{l=1}^{r}\hat{b}_l\binom{1}{l} = \hat{a}_1 + \hat{b}_1 \\
\Rightarrow \hat{a}_1 = x_1 - \hat{b}_1 = x_1 - \frac{(2-3)!!}{(2-2)!!}\cdot a_1 = x_1 - a_1.
\end{aligned}$$

□

We end with an example in order to portray the use of Theorem 1.35. It comes directly from the proof of Theorem 4.29 in Chapter 4.

**Example:**

Suppose we are given the following recurrent equation

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

with initial condition $x_1 = 0$. We can readily see that, with the notations of Theorem 1.35,

$$\begin{aligned} a_1 &= -3, \quad r = 1 \\ b_1 &= \ \ 2, \quad s = 1. \end{aligned}$$

Then,

$$\begin{aligned} \hat{a}_1 &= x_1 - a_1 = 3 \\ \hat{a}_2 &= \hat{a}_{s+1} = \frac{(s+1)(2s)!!}{(2s-1)!!}\frac{b_s}{s} = 8 \\ \hat{b}_1 &= \frac{(2-3)!!}{(2-2)!!}a_1 = -3. \end{aligned}$$

Thus, the solution to the above equation with $x_1 = 0$ is

$$x_n = 3n + 8\binom{n}{2} - 3n\frac{(2n-2)!!}{(2n-3)!!} = 4n^2 - n - 3n\frac{(2n-2)!!}{(2n-3)!!}.$$

CHAPTER 2

# A minimum for the Colless index

Common sense notions of tree balance lead to the recognition of balance as indicating equal numbers of included terminal nodes for both branches of the various furcations (interior nodes) of a dendogram.

K.T. Shao and R. R. Sokal, *Tree Balance* [107], 1990

NO BALANCE index captures the intuitive concept of balance as well as the Colless index does. Recall that, given a bifurcating tree $T$, if we define, for each $u \in \mathring{V}(T)$, its *balance* as $\mathrm{bal}(u) = |\kappa(u_1) - \kappa(u_2)|$, where $u_1$ and $u_2$ are the children of $u$, then the *Colless index* of $T$ is

$$C(T) = \sum_{u \in \mathring{V}(T)} \mathrm{bal}(u). \qquad (2.1)$$

This index, defined by Colless in 1982 [19], is one of the first balance indices for phylogenetic trees introduced in the literature, and it is very popular because it captures in an intuitive way the notion of "global imbalance" of a tree. But despite its age and popularity, several basic questions on it remained unanswered since its inception. The goal of this chapter is to answer one of these questions: what is its minimum possible value for a given number of leaves $n \in \mathbb{N}_{\geq 1}$, and what trees achieve this minimun value, thus becoming the "most balanced" (at least, according to the Colless measure) among all bifurcating trees with their number of leaves?

The intuition that the maximally balanced trees achieve the minimum Colless index in **BinTree**$_n$ has permeated the phylogenetics community and become part of its folklore knowledge. This intuition was ever so strong due to the easy, yet powerful fact (already spotted in [60, 69, 88]) that, for any power of 2, the only tree with minimum

Colless index is the fully symmetric tree —indeed, since it is the only tree whose Colless index is 0. But, in general, even the fact that minimum value of the Colless index is always reached at a maximally balanced tree had not been proved, much less found what other trees, if any, reach this minimum. That is, until independent work carried out by the group of M. Fischer [40] and our group [27] that eventually gave rise to our joint paper [22], filled this gap in the literature, by characterizing the trees with minimum Colless index (which indeed include, but almost never consist solely of, the maximally balanced trees) and providing two alternative closed formulæ for this minimum value.

This chapter is entirely devoted to explain our contribution to this work, and it is organized as follows. First of all, we shall prove that the minimum value of the Colless index is attained at the maximally balanced trees, thus giving us an explicit way to compute that minimum value: namely, that of the Colless index of a maximally balanced tree. We shall present then a new recursive formula that will allow us to give a closed formula for this value, which we call $c(n)$. Thus reasoning, we shall explicitly state the relationship of $c(n)$ with the binary representation of $n$. To end this first section, we shall present the relationship of this value and the so-called Takagi, or Blancmange, fractal curve.

Secondly, we will be interested in characterizing which trees, given a fixed number of leaves $n$, are *minimal Colless* (in the sense that they attain the aforementioned lower bound). In order to do that, we shall first characterize the pairs $(n_1, n_2)$ such that if $T = T_1 * T_2 \in \mathbf{BinTree}_n$, with $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$ and $T_1, T_2$ are minimal Colless, then $T$ is minimal Colless, too. We shall then re-write that characterization in terms of the binary decomposition of $n$, thus allowing us to easily find the pairs satisfying the aforementioned property. This will be the basis of an algorithm that produces all minimal Colless trees for any given number of leaves. We shall end with a brief discussion of some other interesting results included in [22].

## 2.1 The minimum Colless index

We will begin by establishing several basic results, starting with a recurrence that is a direct consequence of Equation (2.1) (see page 21).

**Lemma 2.1.** *If $T = T_1 * T_2$, with $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$, then*

$$C(T) = C(T_1) + C(T_2) + |n_1 - n_2|.$$

**Lemma 2.2.** *Let $T = T_1 * T_2 \in \mathbf{BinTree}_n$ be a bifurcating tree with n leaves, with $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$. If $T$ has minimum Colless index on $\mathbf{BinTree}_n$, then $T_1$ and $T_2$ have minimum Colless indices on $\mathbf{BinTree}_{n_1}$ and $\mathbf{BinTree}_{n_2}$, respectively.*

*Proof.* We will proceed by *modus tollens*. Assume that $C(T_1)$ is not minimal (the argument in the case that $C(T_2)$ is not minimal is analogous), and let $T_1' \in \mathbf{BinTree}_{n_1}$ be a tree with $n_1$ leaves that achieves the minimum value of the Colless index, so that $C(T_1') < C(T_1)$. Then, let $T' = T_1' * T_2 \in \mathbf{BinTree}_n$ and

$$C(T') = C(T_1') + C(T_2) + |n_1 - n_2| < C(T_1) + C(T_2) + |n_1 - n_2| = C(T)$$

thus negating that $C(T)$ is the minimum Colless index. □

**Corollary 2.3.** *Let $T \in \mathbf{BinTree}_n$ be a bifurcating tree with n leaves, and $T' \in \mathbf{BinTree}_k$ a rooted subtree of $T$ with k leaves. If $C(T)$ is minimum in $\mathbf{BinTree}_n$, then so is $C(T')$ in $\mathbf{BinTree}_k$.*

*Proof.* We proceed by induction over the depth $\delta$ of the root $x$ of $T'$ in $T$. If $\delta = 0$ the result is obviously true, since $T = T'$; if $\delta = 1$ the result holds by the previous lemma. Now suppose that $\delta \geq 2$ and that the result is true for every depth up to $\delta - 1$. Let $y$ be the parent of $x$. Since the depth of $y$ is $\delta - 1$, by the induction hypothesis we have that $T_y$ has minimum Colless index in $\mathbf{BinTree}_{\kappa_T(y)}$, and then, since $x$ has depth 1 in $T_y$, $T'$ has minimum Colless index in $\mathbf{BinTree}_k$ by the case $\delta = 1$. □

**Corollary 2.4.** *For every $n \geq 1$ and for every $T \in \mathbf{BinTree}_n$, $C(T) = 0$ if, and only if, $n$ is a power of 2 and $T$ is fully symmetric.*

*Proof.* The "if" implication is a direct consequence of the fact that, in a fully symmetric tree, both children of each internal node $v$ have the same number of descendant leaves and thus $\mathrm{bal}(v) = 0$. We prove now the "only if" implication by induction over $n$. The base case $n = 1$ being obvious, let $n \geq 2$ and let us assume that the assertion is true for every number of leaves up to $n - 1$. Let $T \in \mathbf{BinTree}_n$ be such that $C(T) = 0$, and let $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$ be its subtrees rooted at the children of its root, so that $T = T_1 * T_2$. Then, by Lemma 2.1, $C(T) = 0$ is equivalent to $n_1 = n_2$ and $C(T_1) = C(T_2) = 0$. By the induction hypothesis, this implies that $n_1 = n_2$ is a power of 2, and hence that $n = n_1 + n_1$ is also a power of 2, and that both $T_1$ and $T_2$ are fully symmetric, and hence that $T = T_1 * T_2$ is fully symmetric, too. □

### 2.1.1 The maximally balanced trees are minimal Colless

Maximally balanced trees present the minimum value of the Colless index. The fact is obvious when the number of leaves is a power of 2, since the Colless index of a fully symmetric tree is equal to 0. In the remaining of this section we will prove the result for all $n \in \mathbb{N}$.

Let $c : \mathbb{N} \to \mathbb{N}$ be the function that assigns, to each $n$, the Colless index of a maximally balanced tree with $n$ leaves. By Theorem 1.8 and Lemma 2.1, it is easy to see that $c(n)$ can be computed recurrently as

$$c(n) = \begin{cases} 0 & \text{if } n = 1 \\ c\left(\left\lceil \frac{n}{2} \right\rceil\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \left\lceil \frac{n}{2} \right\rceil - \left\lfloor \frac{n}{2} \right\rfloor & \text{if } n \geq 2 \end{cases} \tag{2.2}$$

Notice that this recurrence says that, for every $n \geq 2$,

- if $n \in 2\mathbb{N}$, then $c(n) = 2c\left(\frac{n}{2}\right)$
- if $n \notin 2\mathbb{N}$, then $c(n) = c\left(\left\lceil \frac{n}{2} \right\rceil\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + 1 = c\left(\frac{n+1}{2}\right) + c\left(\frac{n-1}{2}\right) + 1$

We will proceed by proving the next lemma, which will give us a straightforward argument for the main proposition.

**Lemma 2.5.** *Let $n \in \mathbb{N}_{\geq 1}$ and $s \in \mathbb{N}$. Then,*

$$c(n + s) + c(n) + s \geq c(2n + s).$$

*Proof.* We will proceed by induction over $n$. First of all, we shall see that the thesis is true when $n = 1$; that is (since $C(1) = 0$),

$$c(1 + s) + s \geq c(2 + s) \tag{2.3}$$

for every $s \in \mathbb{N}$. In order to prove this base case, we also proceed by induction, now over $s$.

Suppose $s = 0$; then, we want to show that $c(1) + 0 = 0 \geq c(2) = 0$, which is obviously true. Now we shall suppose that $s > 0$ and that $c(1 + s') + s' \geq c(2 + s')$ for any number $s'$ up to $s - 1$, and will prove that it entails that $c(1 + s) + s \geq c(2 + s)$.

Two roads diverged in a yellow wood:

- If $s \in 2\mathbb{N}$,

$$\begin{aligned} &c(1 + s) + s \geq c(2 + s) \\ &\quad\Longleftrightarrow c\left(\left\lceil\frac{1+s}{2}\right\rceil\right) + c\left(\left\lfloor\frac{1+s}{2}\right\rfloor\right) + 1 + s \geq 2c\left(\frac{s+2}{2}\right) \\ &\quad\Longleftrightarrow c\left(\frac{s}{2} + 1\right) + c\left(\frac{s}{2}\right) + 1 + 2\frac{s}{2} \geq 2c\left(1 + \frac{s}{2}\right) \end{aligned}$$

This last inequality is true if

$$c\left(\frac{s}{2} + 1\right) + \frac{s}{2} \geq c\left(\frac{s}{2} + 1\right), \quad c\left(\frac{s}{2}\right) + \frac{s}{2} + 1 \geq c\left(\frac{s}{2} + 1\right),$$

and these inequalities hold (and they are actually strict): the former trivially and the latter by the induction hypothesis (and the fact that $s \in 2\mathbb{N} \setminus \{0\}$ implies that $\frac{s}{2} - 1 \in \mathbb{N}$):

$$\begin{aligned} c\left(\frac{s}{2}\right) + \frac{s}{2} + 1 &= c\left(1 + \left(\frac{s}{2} - 1\right)\right) + \frac{s}{2} - 1 + 2 \\ &\geq c\left(2 + \frac{s}{2} - 1\right) + 2 = c\left(1 + \frac{s}{2}\right) + 2 > c\left(1 + \frac{s}{2}\right). \end{aligned}$$

- If $s \notin 2\mathbb{N}$,

$$\begin{aligned} &c(1 + s) + s \geq c(2 + s) \\ &\quad\Longleftrightarrow 2c\left(\frac{1+s}{2}\right) + \left\lceil\frac{s}{2}\right\rceil + \left\lfloor\frac{s}{2}\right\rfloor \geq c\left(\left\lceil\frac{2+s}{2}\right\rceil\right) + c\left(\left\lfloor\frac{2+s}{2}\right\rfloor\right) + 1 \\ &\quad\Longleftrightarrow 2c\left(\left\lceil\frac{s}{2}\right\rceil\right) + \left\lceil\frac{s}{2}\right\rceil + \left\lfloor\frac{s}{2}\right\rfloor \geq c\left(1 + \left\lceil\frac{s}{2}\right\rceil\right) + c\left(\left\lceil\frac{s}{2}\right\rceil\right) + 1. \end{aligned}$$

This last inequality is true if

$$c\left(\left\lceil\frac{s}{2}\right\rceil\right) + \left\lceil\frac{s}{2}\right\rceil \geq c\left(1 + \left\lceil\frac{s}{2}\right\rceil\right) + 1, \quad c\left(\left\lceil\frac{s}{2}\right\rceil\right) + \left\lfloor\frac{s}{2}\right\rfloor \geq c\left(\left\lceil\frac{s}{2}\right\rceil\right).$$

Now, the second inequality holds trivially, and the first is, again, a consequence of the induction hypothesis (and the fact that $s > 0$ implies that $\lceil\frac{s}{2}\rceil - 1 \in \mathbb{N}$):

$$\begin{aligned} c\left(\left\lceil\frac{s}{2}\right\rceil\right) + \left\lceil\frac{s}{2}\right\rceil &= c\left(1 + \left\lceil\frac{s}{2}\right\rceil - 1\right) + \left\lceil\frac{s}{2}\right\rceil - 1 + 1 \\ &\geq c\left(2 + \left\lceil\frac{s}{2}\right\rceil - 1\right) + 1 = c\left(1 + \left\lceil\frac{s}{2}\right\rceil\right) + 1 \end{aligned}$$

This completes the proof of the base case $n = 1$.

Now assume that $n > 1$ and that $c(k+s)+c(k)+s \geq c(2k+s)$ is true for any $k < n$ and any $s \in \mathbb{N}$; we want to prove that

$$c(n+s)+c(n)+s \geq c(2n+s) \tag{2.4}$$

is true for any $s \in \mathbb{N}$. Four possibilities exist:

- Assume that $n \in 2\mathbb{N}$ and $s \in 2\mathbb{N}$. Then,

$$\begin{aligned} & c(n+s)+c(n)+s \geq c(2n+s) \\ & \Longleftrightarrow 2c\left(\frac{n+s}{2}\right)+2c\left(\frac{n}{2}\right)+2\frac{s}{2} \geq 2c\left(\frac{2n+s}{2}\right) \\ & \Longleftrightarrow c\left(\frac{n}{2}+\frac{s}{2}\right)+c\left(\frac{n}{2}\right)+\frac{s}{2} \geq c\left(n+\frac{s}{2}\right) \end{aligned}$$

and this last inequality holds due to the induction hypothesis.

- Assume that $n \in 2\mathbb{N}$ and $s \notin 2\mathbb{N}$. Then,

$$\begin{aligned} & c(n+s)+c(n)+s \geq c(2n+s) \\ & \Longleftrightarrow c\left(\left\lceil\frac{n+s}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n+s}{2}\right\rfloor\right)+1+2c\left(\frac{n}{2}\right)+\left\lceil\frac{s}{2}\right\rceil+\left\lfloor\frac{s}{2}\right\rfloor \\ & \qquad \geq c\left(\left\lceil\frac{2n+s}{2}\right\rceil\right)+c\left(\left\lfloor\frac{2n+s}{2}\right\rfloor\right)+1 \\ & \Longleftrightarrow c\left(\frac{n}{2}+\left\lceil\frac{s}{2}\right\rceil\right)+c\left(\frac{n}{2}+\left\lfloor\frac{s}{2}\right\rfloor\right)+2c\left(\frac{n}{2}\right)+\left\lceil\frac{s}{2}\right\rceil+\left\lfloor\frac{s}{2}\right\rfloor \\ & \qquad \geq c\left(n+\left\lceil\frac{s}{2}\right\rceil\right)+c\left(n+\left\lfloor\frac{s}{2}\right\rfloor\right) \end{aligned}$$

where in the last equivalence we have used that, since $n \in 2\mathbb{N}$, $\left\lceil\frac{n+s}{2}\right\rceil = \frac{n}{2}+\left\lceil\frac{s}{2}\right\rceil$ and $\left\lfloor\frac{n+s}{2}\right\rfloor = \frac{n}{2}+\left\lfloor\frac{s}{2}\right\rfloor$. Now, the last inequality is true because both inequalities

$$\begin{aligned} c\left(\frac{n}{2}+\left\lceil\frac{s}{2}\right\rceil\right)+c\left(\frac{n}{2}\right)+\left\lceil\frac{s}{2}\right\rceil \geq c\left(n+\left\lceil\frac{s}{2}\right\rceil\right) \\ c\left(\frac{n}{2}+\left\lfloor\frac{s}{2}\right\rfloor\right)+c\left(\frac{n}{2}\right)+\left\lfloor\frac{s}{2}\right\rfloor \geq c\left(n+\left\lfloor\frac{s}{2}\right\rfloor\right) \end{aligned}$$

hold by the induction hypothesis.

- Assume that $n \notin 2\mathbb{N}$ and $s \in 2\mathbb{N}$. Then

$$\begin{aligned} & c(n+s)+c(n)+s \geq c(2n+s) \\ & \Longleftrightarrow c\left(\left\lceil\frac{n+s}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n+s}{2}\right\rfloor\right)+c\left(\left\lceil\frac{n}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n}{2}\right\rfloor\right)+2+2\frac{s}{2} \\ & \qquad \geq 2c\left(\frac{2n+s}{2}\right) \\ & \Longleftrightarrow c\left(\left\lceil\frac{n}{2}\right\rceil+\frac{s}{2}\right)+c\left(\left\lfloor\frac{n}{2}\right\rfloor+\frac{s}{2}\right)+c\left(\left\lceil\frac{n}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n}{2}\right\rfloor\right)+2+2\frac{s}{2} \\ & \qquad \geq 2c\left(n+\frac{s}{2}\right) \end{aligned}$$

where in the last equivalence we have used that, since $s \in 2\mathbb{N}$, $\left\lceil \frac{n+s}{2} \right\rceil = \left\lceil \frac{n}{2} \right\rceil + \frac{s}{2}$ and $\left\lfloor \frac{n+s}{2} \right\rfloor = \left\lfloor \frac{n}{2} \right\rfloor + \frac{s}{2}$. Now, the last inequality is true because, on the one hand, by the induction hypothesis,

$$\begin{aligned}
&c\left(\left\lceil \frac{n}{2} \right\rceil + \frac{s}{2}\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \frac{s}{2} + 1 \\
&\quad = c\left(\left\lfloor \frac{n}{2} \right\rfloor + \frac{s}{2} + 1\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \frac{s}{2} + 1 \\
&\quad \geq c\left(2\left\lfloor \frac{n}{2} \right\rfloor + \frac{s}{2} + 1\right) = c\left(n - 1 + \frac{s}{2} + 1\right) = c\left(n + \frac{s}{2}\right)
\end{aligned}$$

and, on the other hand, the following inequality also holds:

$$c\left(\left\lfloor \frac{n}{2} \right\rfloor + \frac{s}{2}\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + \frac{s}{2} + 1 \geq c\left(n + \frac{s}{2}\right)$$

but to establish it we must distinguish two cases:

- If $s = 0$, this inequality says

$$c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + 1 \geq c\,(n)$$

and this inequality holds, and it is actually an equality, by Equation (2.2).

- If $s \geq 2$, then

$$\begin{aligned}
&c\left(\left\lfloor \frac{n}{2} \right\rfloor + \frac{s}{2}\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + \frac{s}{2} + 1 \\
&\quad = c\left(\left\lceil \frac{n}{2} \right\rceil + \frac{s}{2} - 1\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + \frac{s}{2} - 1 + 2 \\
&\quad \geq c\left(2\left\lceil \frac{n}{2} \right\rceil + \frac{s}{2} - 1\right) + 2 = c\left(n + 1 + \frac{s}{2} - 1\right) + 2 \\
&\quad = c\left(n + \frac{s}{2}\right) + 2 > c\left(n + \frac{s}{2}\right)
\end{aligned}$$

where the first inequality is due to the induction hypothesis.

- Assume that $n \notin 2\mathbb{N}$ and $s \notin 2\mathbb{N}$. Then

$$\begin{aligned}
&c(n+s) + c(n) + s \geq c(2n+s) \\
&\quad \iff 2c\left(\frac{n+s}{2}\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + 1 + \left\lceil \frac{s}{2} \right\rceil + \left\lfloor \frac{s}{2} \right\rfloor \\
&\qquad\qquad \geq c\left(\left\lceil \frac{2n+s}{2} \right\rceil\right) + c\left(\left\lfloor \frac{2n+s}{2} \right\rfloor\right) + 1 \\
&\quad \iff c\left(\left\lceil \frac{n}{2} \right\rceil + \left\lfloor \frac{s}{2} \right\rfloor\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{s}{2} \right\rceil\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) \\
&\qquad\qquad + \left\lceil \frac{s}{2} \right\rceil + \left\lfloor \frac{s}{2} \right\rfloor \geq c\left(n + \left\lceil \frac{s}{2} \right\rceil\right) + c\left(n + \left\lfloor \frac{s}{2} \right\rfloor\right)
\end{aligned}$$

The last inequality holds because, by the induction hypothesis,

$$\begin{aligned}
&c\left(\left\lceil\frac{n}{2}\right\rceil+\left\lfloor\frac{s}{2}\right\rfloor\right)+c\left(\left\lceil\frac{n}{2}\right\rceil\right)+\left\lfloor\frac{s}{2}\right\rfloor\\
&\qquad\geq c\left(2\left\lceil\frac{n}{2}\right\rceil+\left\lfloor\frac{s}{2}\right\rfloor\right)=c\left(n+1+\left\lfloor\frac{s}{2}\right\rfloor\right)=c\left(n+\left\lceil\frac{s}{2}\right\rceil\right)\\
&c\left(\left\lfloor\frac{n}{2}\right\rfloor+\left\lceil\frac{s}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n}{2}\right\rfloor\right)+\left\lceil\frac{s}{2}\right\rceil\\
&\qquad\geq c\left(2\left\lfloor\frac{n}{2}\right\rfloor+\left\lceil\frac{s}{2}\right\rceil\right)=\left(n-1+\left\lceil\frac{s}{2}\right\rceil\right)=c\left(n+\left\lfloor\frac{s}{2}\right\rfloor\right).
\end{aligned}$$

This completes the proof of the induction step. □

**Corollary 2.6.** *For every $n \notin 2\mathbb{N}$ and $s \in 2\mathbb{N} \setminus \{0\}$,*

$$c(n+s)+c(n)+s > c(2n+s).$$

*Proof.* A close analysis of the proof of the last lemma shows that, on the one hand, the inequality obtained when $s \in 2\mathbb{N}$ in the induction step of the proof of (2.3) is strict, and thus, if $s \in 2\mathbb{N}\setminus\{0\}$, $C(1+s)+s > C(2+s)$, and, on the other hand, the inequality obtained in the induction step of the proof of (2.4) when $n > 1$ is odd and $s \in 2\mathbb{N}\setminus\{0\}$, is strict, and hence, also in this case, $C(n+s)+C(n)+s > C(2n+s)$. □

In particular, if $T = T_1 * T_2 \in \mathbf{BinTree}_n$, with $T_1 \in \mathbf{BinTree}_{n_1}$, $T_2 \in \mathbf{BinTree}_{n_2}$, $n_1, n_2$ odd, and $n_1 > n_2$, then $T$ can *never* have minimal Colless index. Indeed, set $n' = n_1$, which is odd, and $s = n_1 - n_2 > 0$, which is even. Then

$$C(T) = C(T_1) + C(T_2) + n_1 - n_2 \geq c(n'+s) + c(n') + s > c(2n'+s) = c(n).$$

**Theorem 2.7.** *The minimum Colless index in* $\mathbf{BinTree}_n$ *is reached at the maximally balanced trees.*

*Proof.* We proceed by induction over the number $n$ of leaves. The base case $n = 1$ is obvious; suppose it is true up to $n-1$ leaves. Let $T$ be a binary rooted tree with $n$ leaves, and $T_1, T_2$ the children of the root, with $n'+s$ and $n'$ leaves, respectively, for some $n' \in \mathbb{N}_{\geq 1}$ and $s' \in \mathbb{N}$ such that $n = 2n'+s$. Then,

$$\begin{aligned}
C(T) = C(T_1) + C(T_2) + s &\geq C(T^{\mathrm{bal}}_{n'+s}) + C(T^{\mathrm{bal}}_{n'}) + s\\
&= c(n'+s) + c(n') + s \geq c(2n'+s) = C(T^{\mathrm{bal}}_n)
\end{aligned}$$

where the first inequality is due to the induction hypothesis and the second, to Lemma 2.5. □

So, the minimum Colless index in $\mathbf{BinTree}_n$ is $c(n) = C(T^{\mathrm{bal}}_n)$ and in particular it satisfies the recurrence

$$c(n) = c\left(\left\lceil\frac{n}{2}\right\rceil\right) + c\left(\left\lfloor\frac{n}{2}\right\rfloor\right) + \left\lceil\frac{n}{2}\right\rceil - \left\lfloor\frac{n}{2}\right\rfloor. \tag{2.5}$$

Therefore, in what follows we shall refer to $c(n)$ as the *minimum Colless index* in $\mathbf{BinTree}_n$, and to the trees with $n$ leaves that attain it *minimal Colless trees* with $n$ leaves.

**Corollary 2.8.** *Let $n \in \mathbb{N}_{\geq 1}$ and $p \in \mathbb{N}$ be such that $2^p$ divides $n$. Then, $c(n) = 2^p c\left(\frac{n}{2^p}\right)$.*

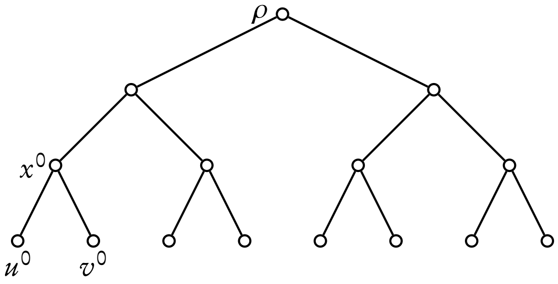
*Proof.* We proceed by induction over $p$. If $p = 0$, then the equation is a tautology. Otherwise, suppose it to be true up to $p - 1$. Let us write $n$ as $n = 2^p n'$. Then,

$$c(n) = c\left(2^p n'\right) = 2c\left(2^{p-1} n'\right) = 2^p c(n')$$

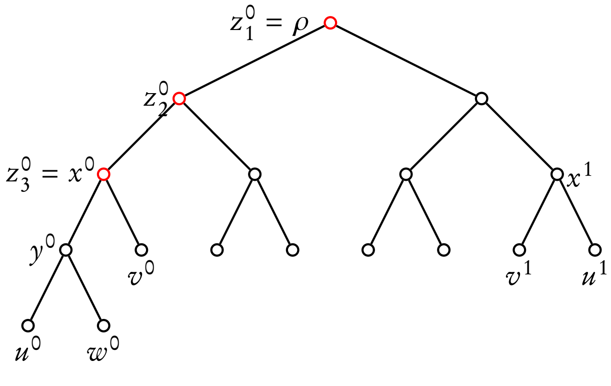
where the second equality is due to Equation (2.5), and the third equality is due to our induction hypothesis. □

### 2.1.2 Another recursive formula for the minimum Colless index

Consider a fully symmetric tree $T^0$ with $n_0 = 2^m$ leaves, all of whose nodes are symmetry nodes and hence $C(T^0) = 0$. Let $\{u^0, v^0\}$ be a cherry in $T^0$, and $x^0$ its parent.

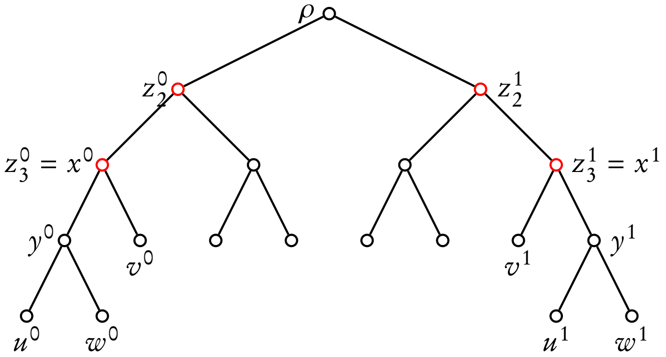


We may now add a sibling $w^0$ to $u^0$, in a way that a new internal node $y^0$ is created, together with an arc that $(x^0, y^0)$, and $T_{y_0} = \{u^0, w^0\}$ is now a cherry. We have thus produced a tree $T^1$ with $n_1 = 2^m + 1$ leaves which is clearly maximally balanced.



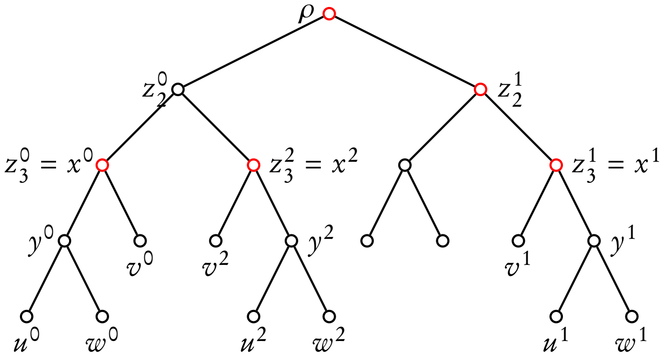
Now, how many unbalanced nodes does it have? The root is plainly unbalanced, since it was balanced before we added $w^0$. Thus arguing, we can conclude that each node in the path $\rho = z_1^0, z_2^0, \ldots, z_m^0 = x^0$ is unbalanced. Furthermore, every internal node that is not in that path remains balanced, since they were so in $T^0$ and no change has occurred to any of its children. Hence, we can conclude that $c(2^m + 1) = C(T^1) = m$.

We can, by picking a cherry $\{u^1, v^1\}$ in $T^1$ and adding a leaf $w^1$ to it in a way analogous to that we have used above, construct a new tree $T^2$. In order for $T^2$ to be maximally balanced, the chosen cherry must lie below a different child of the root than the one in which the cherry $\{u^0, w^0\}$ dwells. Let it be so. Now, the question arises naturally: how many unbalanced nodes does $T^2$ have?



The root is plainly balanced since it is even a symmetry node, both trees rooted at its children having the exact same shape. However, the path $z_2^0, z_3^0, \dots, z_m^0 = x^0$ is still formed by unbalanced nodes, and so is an analogous path $z_2^1, z_3^1, \dots, z_m^1 = x^1$ in the corresponding subtree. It can be readily argued that no other internal node is unbalanced, and hence $c(2^m + 2) = C(T^2) = 2(m-1)$. We could have also argued that, the root being balanced, the Colless index of $T^2$ is the addition of those of both subtrees rooted at the children of the root, which are fully symmetric with $2^{m-1}$ leaves plus an extra cherry and therefore, by the previous discussion, they both have Colless index $m-1$.

Let's pick now a cherry $\{u^2, v^2\}$ in $T^2$ and add a leaf $w^2$ to it forming a new cherry with $u^2$. In order for the resulting tree $T^3$ to be maximally balanced, the chosen cherry must lie below a different grandchild $z_3^2$ of the root than the ancestors $z_3^0$ and $z_3^1$ of the cherries $\{u^0, w^0\}$ and $\{u^1, w^1\}$, respectively.



In $T^3$, the root $\rho$ becomes unbalanced again, its child $z_2^0$ parenting the nodes $z_3^0$ and $z_3^2$ is balanced, the paths $z_3^0, \dots, z_m^0 = x^0$ and $z_2^1, z_3^1, \dots, z_m^1 = x^1$ are still formed by unbalanced nodes, and a new path of unbalanced nodes is added: $z_3^2, \dots, z_m^2 = x^2$. Therefore, $c(2^m + 3) = C(T^3) = 2(m-1) + (m-2) = 3m - 4$.

Now, in general, suppose that $n = 2^m + k$, with $k = 2^r + s < 2^m$ and $0 \leq s < 2^r$, and consider the maximally balanced tree $T^k$ obtained through the procedure explained above. If $s$ were to be 0, it could be argued that every node of imbalance would have depth at least $r$: indeed, each of the $2^r$ trees rooted at nodes of depth $r$ would have $2^{m-r} + 1$ leaves. In this case, $c(2^m + 2^r) = C(T^{2^r}) = 2^r(m-r)$.

Consider now the case in which $s \neq 0$. Of the $2^r$ subtrees mentioned above, $s$ would now have $2^{k-r} + 2$ leaves and hence their Colless index would be $2(m-r-1)$, whereas that of the remaining $2^r - s$ trees would continue to be $m - r$. However, in this situation some nodes of depth less than $s$ may be unbalanced. The swiftest way of deducing how many they can be is to picture the $2^r - s$ subtrees of $2^{k-r} + 1$ leaves as leaves, and those of $2^{k-r} + 2$ leaves as cherries: there are $C(T^k)$ unbalanced nodes.

These considerations lead us to the following alternative recurrence for $c(n)$. Define $\overline{c} : \mathbb{N} \to \mathbb{N}$ as $\overline{c}(2^m) = 0$, for every $m \in \mathbb{N}$, and, if $n = 2^m + k$ with $m \in \mathbb{N}$ and $0 < k < 2^m$,

$$\begin{aligned}\overline{c}(n) &= (2^{\lfloor \log_2(k)\rfloor+1} - k)(m - \lfloor \log_2(k)\rfloor) \\ &\qquad + 2(k - 2^{\lfloor \log_2(k)\rfloor})(m - \lfloor \log_2(k)\rfloor - 1) + \overline{c}(k) \\ &= k(m - \lfloor \log_2(k)\rfloor) - 2(k - 2^{\lfloor \log_2(k)\rfloor}) + \overline{c}(k).\end{aligned}$$

To motivate this recurrence, notice that if $n = 2^m + k$ and $0 < k = 2^r + s < 2^m$ with $0 \le s < 2^r$, so that $r = \lfloor \log_2(k)\rfloor$, then, as we have just discussed, the addend $2(k - 2^{\lfloor \log_2(k)\rfloor})(m - \lfloor \log_2(k)\rfloor - 1) = 2s(m - r - 1)$ is the contribution to $c(n)$ of the $s$ subtrees rooted at nodes of depth $r$ with $2^{k-r} + 2$ leaves, the addend $(2^{\lfloor \log_2(k)\rfloor+1} - k)(m - \lfloor \log_2(k)\rfloor) = 2^r(m - r)$ is the contribution of the remaining $2^r - s$ subtrees rooted at nodes of depth $r$, and $c(k)$ is the contribution of the unbalanced nodes of depth less than $s$.

We now prove now that this recurrence defines $c(n)$.

**Lemma 2.9.** *For every $n \in \mathbb{N}_{\ge 1}$, $c(n) = \overline{c}(n)$.*

*Proof.* We shall prove that $\overline{c}$ satisfies the same recurrence (2.5) as $c$: for every $n \ge 2$,

$$\overline{c}(n) = \overline{c}\left(\left\lceil \frac{n}{2} \right\rceil\right) + \overline{c}\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \left\lceil \frac{n}{2} \right\rceil - \left\lfloor \frac{n}{2} \right\rfloor. \tag{2.6}$$

Since $\overline{c}(1) = \overline{c}(2^0) = 0 = c(1)$, this will imply that $\overline{c}(n) = c(n)$ for every $n$. We proceed by induction over $k$ in the expression of $n$ as $n = 2^m + k$ with $0 \le k < 2^m$.

The base case is $k = 0$, in which case $n = 2^m$ with $m \ge 1$. But then $\lceil \frac{n}{2} \rceil = \lfloor \frac{n}{2} \rfloor = 2^{m-1}$ and both sides of (2.6) are 0.

Let now $0 < k < 2^m$, suppose that (2.6) is true for every number of the form $2^{m'} + k'$ with $0 \le k' < k$ and let us prove it for $2^m + k$: i.e., we want to prove that

$$\overline{c}(2^m + k) = \overline{c}\left(2^{m-1} + \left\lceil \frac{k}{2} \right\rceil\right) + \overline{c}\left(2^{m-1} + \left\lfloor \frac{k}{2} \right\rfloor\right) + \left\lceil \frac{k}{2} \right\rceil - \left\lfloor \frac{k}{2} \right\rfloor. \tag{2.7}$$

Now, two cases arise, depending on whether $k = 2^{\lfloor \log_2(k)\rfloor+1} - 1$ or not.

- If $0 < k = 2^{\lfloor \log_2(k)\rfloor+1} - 1$, then $\lceil \frac{k}{2} \rceil = 2^{\lfloor \log_2(k)\rfloor}$ and $\lfloor \frac{k}{2} \rfloor = 2^{\lfloor \log_2(k)\rfloor} - 1$, and therefore

$\left\lfloor \log_2\left\lceil \frac{k}{2}\right\rceil\right\rfloor = \lfloor \log_2(k)\rfloor$, and $\left\lfloor \log_2\left\lfloor \frac{k}{2}\right\rfloor\right\rfloor = \lfloor \log_2(k)\rfloor - 1$. Thus:

$$\begin{aligned}
&\overline{c}\left(2^{m-1}+\left\lceil\frac{k}{2}\right\rceil\right)\\
&\quad=\left\lceil\frac{k}{2}\right\rceil\left(m-1-\left\lfloor\log_2\left\lceil\frac{k}{2}\right\rceil\right\rfloor\right)-2\left(\left\lceil\frac{k}{2}\right\rceil-2^{\left\lfloor\log_2\left\lceil\frac{k}{2}\right\rceil\right\rfloor}\right)+\overline{c}\left(\left\lceil\frac{k}{2}\right\rceil\right)\\
&\quad=\left\lceil\frac{k}{2}\right\rceil\left(m-1-\lfloor\log_2(k)\rfloor\right)+\overline{c}\left(\left\lceil\frac{k}{2}\right\rceil\right)\\
&\quad=\left\lceil\frac{k}{2}\right\rceil\left(m-\lfloor\log_2(k)\rfloor\right)+\overline{c}\left(\left\lceil\frac{k}{2}\right\rceil\right)-2^{\lfloor\log_2(k)\rfloor}
\end{aligned}$$

and

$$\begin{aligned}
&\overline{c}\left(2^{m-1}+\left\lfloor\frac{k}{2}\right\rfloor\right)\\
&\quad=\left\lfloor\frac{k}{2}\right\rfloor\left(m-1-\left\lfloor\log_2\left\lfloor\frac{k}{2}\right\rfloor\right\rfloor\right)-2\left(\left\lfloor\frac{k}{2}\right\rfloor-2^{\left\lfloor\log_2\left\lfloor\frac{k}{2}\right\rfloor\right\rfloor}\right)+\overline{c}\left(\left\lfloor\frac{k}{2}\right\rfloor\right)\\
&\quad=\left\lfloor\frac{k}{2}\right\rfloor\left(m-\lfloor\log_2(k)\rfloor\right)-2\left(\left\lfloor\frac{k}{2}\right\rfloor-2^{\lfloor\log_2(k)\rfloor-1}\right)+\overline{c}\left(\left\lfloor\frac{k}{2}\right\rfloor\right)\\
&\quad=\left\lfloor\frac{k}{2}\right\rfloor\left(m-\lfloor\log_2(k)\rfloor\right)-2\left(2^{\lfloor\log_2(k)\rfloor-1}-1\right)+\overline{c}\left(\left\lfloor\frac{k}{2}\right\rfloor\right)\\
&\quad=\left\lfloor\frac{k}{2}\right\rfloor\left(m-\lfloor\log_2(k)\rfloor\right)-2^{\lfloor\log_2(k)\rfloor}+2+\overline{c}\left(\left\lfloor\frac{k}{2}\right\rfloor\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\overline{c}\left(2^{m-1}+\left\lfloor\frac{k}{2}\right\rfloor\right)+\overline{c}\left(2^{m-1}+\left\lfloor\frac{k}{2}\right\rfloor\right)+\left\lceil\frac{k}{2}\right\rceil-\left\lfloor\frac{k}{2}\right\rfloor\\
&\quad=\left\lceil\frac{k}{2}\right\rceil\left(m-\lfloor\log_2(k)\rfloor\right)+\overline{c}\left(\left\lceil\frac{k}{2}\right\rceil\right)-2^{\lfloor\log_2(k)\rfloor}\\
&\qquad+\left\lfloor\frac{k}{2}\right\rfloor\left(m-\lfloor\log_2(k)\rfloor\right)-2^{\lfloor\log_2(k)\rfloor}+2+\overline{c}\left(\left\lfloor\frac{k}{2}\right\rfloor\right)+\left\lceil\frac{k}{2}\right\rceil-\left\lfloor\frac{k}{2}\right\rfloor\\
&\quad=k\left(m-\lfloor\log_2(k)\rfloor\right)-2^{\lfloor\log_2(k)\rfloor+1}+2+\overline{c}\left(\left\lceil\frac{k}{2}\right\rceil\right)+\overline{c}\left(\left\lfloor\frac{k}{2}\right\rfloor\right)\\
&\qquad+\left\lceil\frac{k}{2}\right\rceil-\left\lfloor\frac{k}{2}\right\rfloor
\end{aligned}$$

whereas,

$$\begin{aligned}
\overline{c}\left(2^m+k\right)&=k(m-\lfloor\log_2(k)\rfloor)-2(2^{\lfloor\log_2(k)\rfloor+1}-1-2^{\lfloor\log_2(k)\rfloor})+\overline{c}(k)\\
&=k\left(m-\lfloor\log_2(k)\rfloor\right)-2^{\lfloor\log_2(k)\rfloor+1}+2+\overline{c}\left(k\right)
\end{aligned}$$

yielding the desired result by applying our induction hypothesis to $k=2^{m'}+k'$ with $m'=\lfloor\log_2(k)\rfloor$ and $k'=2^{\lfloor\log_2(k)\rfloor}-1<k$.

- If $0 < k < 2^{\lfloor \log_2(k) \rfloor + 1} - 1$, then $\left\lfloor \log_2 \left\lceil \frac{k}{2} \right\rceil \right\rfloor = \left\lfloor \log_2 \left\lfloor \frac{k}{2} \right\rfloor \right\rfloor = \lfloor \log_2(k) \rfloor - 1$. Thus,

$$\begin{aligned}
&\overline{c}\left(2^{m-1} + \left\lceil \frac{k}{2} \right\rceil\right) \\
&\quad = \left\lceil \frac{k}{2} \right\rceil \left(m - 1 - \left\lfloor \log_2 \left\lceil \frac{k}{2} \right\rceil \right\rfloor\right) - 2\left(\left\lceil \frac{k}{2} \right\rceil - 2^{\left\lfloor \log_2 \left\lceil \frac{k}{2} \right\rceil \right\rfloor}\right) + \overline{c}\left(\left\lceil \frac{k}{2} \right\rceil\right) \\
&\quad = \left\lceil \frac{k}{2} \right\rceil (m - \lfloor \log_2(k) \rfloor) - 2\left(\left\lceil \frac{k}{2} \right\rceil - 2^{\lfloor \log_2(k) \rfloor - 1}\right) + \overline{c}\left(\left\lceil \frac{k}{2} \right\rceil\right)
\end{aligned}$$

and

$$\begin{aligned}
&\overline{c}\left(2^{m-1} + \left\lfloor \frac{k}{2} \right\rfloor\right) \\
&\quad = \left\lfloor \frac{k}{2} \right\rfloor \left(m - 1 - \left\lfloor \log_2 \left\lfloor \frac{k}{2} \right\rfloor \right\rfloor\right) - 2\left(\left\lceil \frac{k}{2} \right\rceil - 2^{\left\lfloor \log_2 \left\lfloor \frac{k}{2} \right\rfloor \right\rfloor}\right) + \overline{c}\left(\left\lfloor \frac{k}{2} \right\rfloor\right) \\
&\quad = \left\lfloor \frac{k}{2} \right\rfloor (m - \lfloor \log_2(k) \rfloor) - 2\left(\left\lfloor \frac{k}{2} \right\rfloor - 2^{\lfloor \log_2(k) \rfloor - 1}\right) + \overline{c}\left(\left\lfloor \frac{k}{2} \right\rfloor\right)
\end{aligned}$$

and hence,

$$\begin{aligned}
&\overline{c}\left(2^{m-1} + \left\lfloor \frac{k}{2} \right\rfloor\right) + \overline{c}\left(2^{m-1} + \left\lfloor \frac{k}{2} \right\rfloor\right) + \left\lceil \frac{k}{2} \right\rceil - \left\lfloor \frac{k}{2} \right\rfloor \\
&\quad = \left\lceil \frac{k}{2} \right\rceil (m - \lfloor \log_2(k) \rfloor) - 2\left(\left\lceil \frac{k}{2} \right\rceil - 2^{\lfloor \log_2(k) \rfloor - 1}\right) + \overline{c}\left(\left\lceil \frac{k}{2} \right\rceil\right) \\
&\qquad + \left\lfloor \frac{k}{2} \right\rfloor (m - \lfloor \log_2(k) \rfloor) - 2\left(\left\lfloor \frac{k}{2} \right\rfloor - 2^{\lfloor \log_2(k) \rfloor - 1}\right) + \overline{c}\left(\left\lfloor \frac{k}{2} \right\rfloor\right) \\
&\qquad + \left\lceil \frac{k}{2} \right\rceil - \left\lfloor \frac{k}{2} \right\rfloor \\
&\quad = k\,(m - \lfloor \log_2(k) \rfloor) - 2\left(k - 2^{\lfloor \log_2(k) \rfloor}\right) + \overline{c}\left(\left\lceil \frac{k}{2} \right\rceil\right) + \overline{c}\left(\left\lfloor \frac{k}{2} \right\rfloor\right) \\
&\qquad + \left\lceil \frac{k}{2} \right\rceil - \left\lfloor \frac{k}{2} \right\rfloor
\end{aligned}$$

whereas,

$$\overline{c}\,(2^m + k) = k\,(m - \lfloor \log_2(k) \rfloor) - 2\left(k - 2^{\lfloor \log_2(k) \rfloor}\right) + \overline{c}\,(k)$$

yielding the desired result by applying our induction hypothesis to $k = 2^{m'} + k'$ with $m' = \lfloor \log_2(k) \rfloor$ and $k' < k$.

This completes the proof of the inductive step. □

Thus, we have derived a new recursive formula for the minimum Colless index:

$$c(n) = \begin{cases} 0 & \text{if } n = 2^m \\ k(m - \lfloor \log_2(k) \rfloor) - 2(k - 2^{\lfloor \log_2(k) \rfloor}) + c(k) & \text{if } n = 2^m + k, 0 < k < 2^m \end{cases} \tag{2.8}$$

This recurrence will be useful anon.

### 2.1.3 The value of the minimum Colless index

In this section, we will deal with the problem of computing the actual value of the minimum Colless index by means of a closed expression. An easy, albeit rather useless under the computational point of view, way to find the value of the minimum Colless index is presented in the next result, and deals with the number of symmetry nodes of maximally balanced trees. Its main interest lies in the fact that it entails that the sequence $c(n)$ is the sequence A296062 in Sloane's *On-Line Encyclopedia of Integer Sequences* [108].

**Theorem 2.10.** *Let $T_n^{\mathrm{bal}} \in \mathbf{BinTree}_n$ be a maximally balanced tree with n leaves, and $s(T_n^{\mathrm{bal}})$ its number of symmetry nodes. Then, $c(n) = n - 1 - s(T_n^{\mathrm{bal}})$.*

*Proof.* Let $u$ be an internal node of $T_n^{\mathrm{bal}}$. By construction, the subtrees rooted at the children of $u$ will be two maximally balanced trees whose number of leaves differ in at most 1. Thus, if $\mathrm{bal}(u) = 1$, $u$ is not a symmetry node, while if $\mathrm{bal}(u) = 0$, then the subtrees rooted at the children of $u$ will be isomorphic and hence $u$ will be a symmetry node. So, there are $n - 1$ internal nodes in $T_n^{\mathrm{bal}}$, of which the $s(T_n^{\mathrm{bal}})$ symmetry nodes do not contribute to its Colless index and the remaining $n - 1 - s(T_n^{\mathrm{bal}})$ contribute 1 each. □

Finding the value of the minimum Colless index amounts to solve recurrences (2.5) or (2.8). Recurrence (2.8) gives us the hint to a closed expression in terms of binary decompositions. Indeed, let $n = \sum_{i=0}^{\ell} 2^{m_i}$, with $m_0 < m_1 < \cdots < m_{\ell-1} < m_\ell$, be the binary decomposition of $n$. Then, by (2.8)

$$\begin{aligned}
c\Big(\sum_{i=0}^{\ell} 2^{m_i}\Big) &= \Big(\sum_{i=0}^{\ell-1} 2^{m_i}\Big)(m_\ell - m_{\ell-1}) - 2\Big(\sum_{i=0}^{\ell-1} 2^{m_i} - 2^{m_{\ell-1}}\Big) + c\Big(\sum_{i=0}^{\ell-1} 2^{m_i})\\
&= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + \Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)(m_\ell - m_{\ell-1} - 2) + c\Big(\sum_{i=0}^{\ell-1} 2^{m_i}\Big)\\
&= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + \Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)(m_\ell - m_{\ell-1} - 2)\\
&\quad + \Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)(m_{\ell-1} - m_{\ell-2}) - 2\Big(\sum_{i=0}^{\ell-2} 2^{m_i} - 2^{m_{\ell-2}}\Big) + c\Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)\\
&= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + \Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)(m_\ell - m_{\ell-1} - 2)\\
&\quad + 2^{m_{\ell-2}}(m_{\ell-1} - m_{\ell-2}) + \Big(\sum_{i=0}^{\ell-3} 2^{m_i}\Big)(m_{\ell-1} - m_{\ell-2} - 2) + c\Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)\\
&= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + 2^{m_{\ell-2}}(m_\ell - m_{\ell-2} - 2)\\
&\quad + \Big(\sum_{i=0}^{\ell-3} 2^{m_i}\Big)(m_\ell - m_{\ell-2} - 4) + c\Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)
\end{aligned}$$

$$
\begin{aligned}
&= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + 2^{m_{\ell-2}}(m_\ell - m_{\ell-2} - 2)\\
&\quad + \Big(\sum_{i=0}^{\ell-3} 2^{m_i}\Big)(m_\ell - m_{\ell-2} - 4) + \Big(\sum_{i=0}^{\ell-3} 2^{m_i}\Big)(m_{\ell-2} - m_{\ell-3})\\
&\quad - 2\Big(\sum_{i=0}^{\ell-3} 2^{m_i} - 2^{m_{\ell-3}}\Big) + c\Big(\sum_{i=0}^{\ell-3} 2^{m_i}\Big)\\
&= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + 2^{m_{\ell-2}}(m_\ell - m_{\ell-2} - 2) + 2^{m_{\ell-3}}(m_\ell - m_{\ell-3} - 4)\\
&\quad + \Big(\sum_{i=0}^{\ell-4} 2^{m_i}\Big)(m_\ell - m_{\ell-3} - 6) + c\Big(\sum_{i=0}^{\ell-3} 2^{m_i}\Big)
\end{aligned}
$$

and so forth.

This lead us to conjecture that

$$c(n) = \sum_{i=0}^{\ell-1} 2^{m_i}(m_\ell - m_i - 2(\ell - i - 1)).$$

Next theorem shows that our conjecture was right.

**Theorem 2.11.** *Let $n \in \mathbb{N}$ be a natural number, and $n = \sum_{i=0}^{\ell} 2^{m_i}$ its binary decomposition, with $m_i < m_{i+1}$ for every $i \in \{0, \ldots, \ell - 1\}$. Then,*

$$c(n) = \sum_{i=0}^{\ell-1} 2^{m_i}(m_\ell - m_i - 2(\ell - i - 1)).$$

*Proof.* We will prove the thesis in the statement by induction over $n$. The base case when $n = 1 = 2^0$, or, more in general, the case when $n = 2^m$ (that is, $\ell = 1$) is obvious, because in this case the sum in the right-hand side term in the statement's expression is $0 = c(2^m)$.

Let now $n = \sum_{i=0}^{\ell} 2^{m_i} > 1$ with $m_i < m_{i+1}$ for every $i \in \{0, \ldots, \ell - 1\}$, so that, with the notations of Equation (2.8), $m = m_\ell$ and $k = \sum_{i=0}^{\ell-1} 2^{m_i}$; since we have already proved the case when $\ell = 1$, we shall assume that $\ell \geq 2$ and hence that $k > 0$ and $\lfloor \log_2(k) \rfloor = m_{\ell-1}$, and that the equality in the statement holds for every positive integer up to $n - 1$. Then, by (2.8),

$$
\begin{aligned}
c(n) &= \Big(\sum_{i=0}^{\ell-1} 2^{m_i}\Big)(m_\ell - m_{\ell-1}) - 2\Big(\sum_{i=0}^{\ell-1} 2^{m_i} - 2^{m_{\ell-1}}\Big) + c(k)\\
&= 2^{\ell-1}(m_\ell - m_{\ell-1}) + \Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)(m_\ell - m_{\ell-1} - 2) + c(k)\\
&= 2^{\ell-1}(m_\ell - m_{\ell-1}) + \Big(\sum_{i=0}^{\ell-2} 2^{m_i}\Big)(m_\ell - m_{\ell-1} - 2)\\
&\quad + \sum_{i=0}^{\ell-2} 2^{m_i}(m_{\ell-1} - m_i - 2(\ell - 1 - i - 1))\\
&\quad \text{(by the induction hypothesis)}
\end{aligned}
$$

$$= 2^{m_{\ell-1}}(m_\ell - m_{\ell-1}) + \sum_{i=0}^{\ell-2} 2^{m_i}(m_\ell - m_i - 2(\ell - i - 1))$$
$$= \sum_{i=0}^{\ell-1} 2^{m_i}(m_\ell - m_i - 2(\ell - i - 1))$$

as we wanted to prove. □

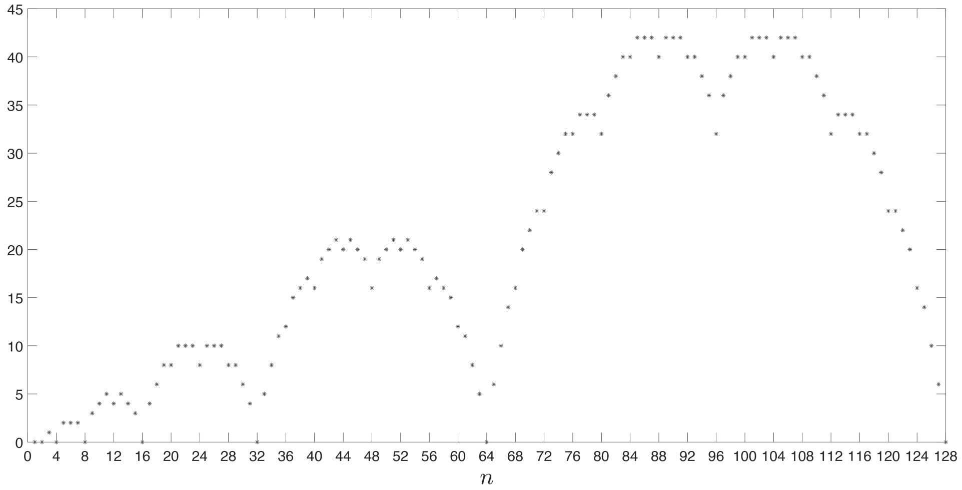


Figure 2.1: Plot of $c(n)$ for $n \in \{1, \ldots, 128\}$.

### 2.1.4 Relationship with the Takagi curve

Figure 2.1 shows the values of the first 128 minimum Colless indices. This figure presents a structure that seems to be fractal, and in particular strongly resembles that of the fractal curve known as the *blancmange*, or *Takagi*, *curve* [116]. This curve, depicted in Figure 2.2, is defined to be the graph of the *Takagi function* $T : [0, 1] \to \mathbb{R}$

$$T(x) = \sum_{i=0}^{\infty} 2^{-i} \cdot s(2^i \cdot x) \tag{2.9}$$

where $s(x) = \min_{z \in \mathbb{Z}} |x - z|$ is the distance from $x$ to the nearest integer (note that $s(x) \in [0, 1/2]$). The following result makes the relationship between $c$ and $T$ explicit.

**Theorem 2.12.** *Let $T(x) : [0, 1] \to \mathbb{R}$ be the Takagi function. Then, for every $n \in \mathbb{N}_{\geq 1}$,*

$$c(n) = 2^{\lfloor \log_2(n) \rfloor} \cdot T\left(\frac{n}{2^{\lfloor \log_2(n) \rfloor}} - 1\right).$$

*Proof.* If $n = 2^m$ for some $m \in \mathbb{N}$, then

$$2^{\lfloor \log_2(n) \rfloor} \cdot T\left(2^{-\lfloor \log_2(n) \rfloor} n - 1\right) = 2^m \cdot T(2^{-m} 2^m - 1)$$
$$= 2^m \cdot T(0) = 0 = c(2^m).$$

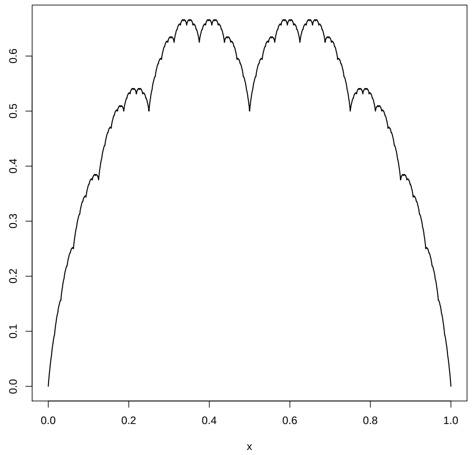


Figure 2.2: The blancmange curve.

The case when $n$ is not a power of 2 can be easily derived from the following reformulation of $T$ given by Tambs-Lyche [117]: if $x = \sum_{j=1}^{\infty} 2^{-l_j}$ with $(l_j)_j$ strictly increasing, then

$$T(x) = \sum_{j=1}^{\infty} \frac{l_j - 2(j-1)}{2^{l_j}}.$$

Now, let $n = \sum_{i=0}^{\ell} 2^{m_i}$ with $m_\ell > m_{\ell-1} > \cdots > m_0$ and $\ell > 0$. Then,

$$\begin{aligned}
2^{\lfloor \log_2(n) \rfloor} \cdot T\left(2^{-\lfloor \log_2(n) \rfloor} n - 1\right) &= 2^{m_\ell} \cdot T\left(2^{-m_\ell} \sum_{i=0}^{\ell} 2^{m_i} - 1\right) \\
&= 2^{m_\ell} \cdot T\left(\sum_{i=0}^{\ell-1} 2^{-(m_\ell - m_i)}\right) = 2^{m_\ell} \cdot T\left(\sum_{j=1}^{\ell} 2^{-(m_\ell - m_{\ell-j})}\right) \\
&= 2^{m_\ell} \sum_{j=1}^{\ell} \frac{m_\ell - m_{\ell-j} - 2(j-1)}{2^{m_\ell - m_{\ell-j}}} \qquad \text{(by Tambs-Lyche identity)} \\
&= \sum_{j=1}^{\ell} 2^{m_{\ell-j}} (m_\ell - m_{\ell-j} - 2(j-1)) \\
&= \sum_{i=0}^{\ell-1} 2^{m_i} (m_\ell - m_i - 2(\ell - i - 1)) = c(n),
\end{aligned}$$

by Theorem 2.11. □

From this, we can deduce the following result, which shall give upper bounds to the value of $c(n)$, as well as some explicit computations and a symmetry property.

**Corollary 2.13.** *The sequence $c(n)$ satisfies the following properties:*

*(i)* *For every* $m \geq 0$, $c(2^m + 1) = m$.

*(ii)* *For every* $n \geq 1$, $c(n) < 2^{\lceil \log_2(n) \rceil}/3$.

*(iii)* *For every* $n \geq 1$, $c(n) < n/2$.

*(iv)* *For every* $m \geq 1$ *and for every* $k \in \{1, \ldots, 2^m - 1\}$, $c(2^m + k) = c(2^{m+1} - k)$.

*Proof.* The first property, *(i)*, was proved in the preamble of Section 2.1.2, and it can also be derived directly from Theorem 2.11.

Property *(ii)* requires Theorem 3.1 in [4], according to which $T(x) \leq 2/3$ for every $x \in [0, 1]$. Using this fact, if $n = 2^m + k$, with $1 \leq k \leq 2^m - 1$, then $\lfloor \log_2(n) \rfloor = m$ and $\lceil \log_2(n) \rceil = m + 1$, and by Theorem 2.12

$$c(n) = 2^m \cdot T\left(2^{-m} n - 1\right) = 2^m \cdot T\left(2^{-m} k\right) \leq 2^m \cdot \frac{2}{3} = \frac{2^{m+1}}{3}.$$

The aforementioned theorem also states that the real numbers $x \in [0, 1]$ that reach the upper bound $T(x) = 2/3$ are exactly those whose coefficients $\varepsilon_i \in \{0, 1\}$ in their binary expansion $x = \sum_{i=1}^{\infty} \frac{\varepsilon_i}{2^i}$ satisfy that $\varepsilon_{2i} + \varepsilon_{2i-1} = 1$ for all $i \in \mathbb{N}_{\geq 1}$. But this is indeed not the case for $x = \frac{k}{2^m}$, since its binary expansion is finite. Finally, the case when $k = 0$, i.e. when $n$ is a power of two, is trivial since then $c(n) = 0$.

We proceed now to the proof of *(iii)*, which will be performed by induction over $n$. The base case clearly holds for $n = 1$, since $c(1) = 0 < 1/2$, and hence assume that the property holds up to $n - 1$ leaves. We distinguish three cases, based on the congruence of $n$ modulo 4:

- If $n$ is even, say $n = 2n_0$, then $c(n) = 2c(n_0)$ by Corollary 2.8. But then, $2c(n_0) < 2n_0/2 = n/2$ by the induction hypothesis.
- If $n = 4n_0 + 1$ for some $n_0 \in \mathbb{N}$, then

$$c(n) = c(2n_0 + 1) + c(2n_0) + 1 \leq n_0 + (n_0 - 1) + 1 = 2n_0 < \frac{n}{2}$$

where the first inequality is due to the induction hypothesis and the fact that $c(2n_0) < n_0$ implies that $c(2n_0) \leq n_0 - 1$.

- If $n = 4n_0 + 3$ for some $n_0 \in \mathbb{N}$, then

$$c(n) = c(2n_0 + 2) + c(2n_0 + 1) + 1 \leq n_0 + n_0 + 1 = 2n_0 + 1 < \frac{n}{2},$$

where the first inequality is due again to the induction hypothesis and the fact that $c(2n_0 + 2) < n_0 + 1$ implies that $c(2n_0 + 2) \leq n_0$.

Thus concludes the proof of *(iii)*.

Finally, *(iv)* is a direct consequence of the symmetry of $T$ around $1/2$. Indeed, suppose that $n = 2^m + k$, for some $k \in \{1, \ldots, 2^m - 1\}$. Then, by Theorem 2.12,

$$\begin{aligned} c(2^m + k) &= 2^m \cdot T\left(2^{-m}(2^m + k) - 1\right) = 2^m \cdot T\left(2^{-m} k\right) \\ c(2^{m+1} - k) &= 2^m \cdot T\left(2^{-m}(2^{m+1} - k) - 1\right) = 2^m \cdot T\left(1 - 2^{-m} k\right) \end{aligned}$$

and $T(x) = T(1-x)$ for every $x \in [0, 1]$. Here concludes the proof of the statement. □

## 2.2 Minimal Colless trees

So far, we have focused on the computation of the *value* of the minimum Colless index, which we have obtained as a consequence of the fact that the maximally balanced trees attain it. Now we turn to the problem of knowing *which trees*, apart from those maximally balanced, are also *minimal Colless*, that is, present the minimum Colless index for their number of leaves. This question arises naturally by observing that, for $n = 6$, there are exactly two trees, $T_3^{\mathrm{bal}} * T_3^{\mathrm{bal}}$ and $T_2^{\mathrm{bal}} * T_4^{\mathrm{bal}}$, with minimum Colless index, 2.

In this section we provide a way of generating all bifurcating trees with minimum Colless index among all bifurcating trees with the same number of leaves. Given $n$, our characterization will be given from the root to the leaves: by first of all finding which pairs of natural numbers $(n_1, n_2) \in \mathbb{N}^2$ are such that a tree $T = T_1 * T_2 \in \mathbf{BinTree}_n$, with $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$, is minimal Colless if $T_1$ and $T_2$ are so, and then repeating recursively this step with $n_1$ and $n_2$. We call the set of all pairs of natural numbers $(n_1, n_2)$ satisfying the aforementioned property $\underset{\sim}{\mathrm{QB}}(n)$. To simplify the notation, throughout this section whenever we write a tree $T$ as the root join $T_1 * T_2$ of $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$, we shall always implicitly assume that $n_1 \geq n_2$, and therefore that all pairs $(n_1, n_2) \in \underset{\sim}{\mathrm{QB}}(n)$ satisfy also this property, as well as $n_1 + n_2 = n$. This allows us to think of them as pairs of the form $(m + s, m)$ with $m \geq 1$ and $s \geq 0$. Since the assertion

> $T = T_1 * T_2 \in \mathbf{BinTree}_m$, with $T_1 \in \mathbf{BinTree}_{m+s}$ and $T_2 \in \mathbf{BinTree}_m$, is minimal Colless if $T_1$ and $T_2$ are so

is equivalent to

$$c(n) = C(T) = C(T_1) + C(T_2) + (m + s) - m = c(m + s) + c(m) + s$$

this leads to the following working definition of the set $\underset{\sim}{\mathrm{QB}}(n)$:

$$\underset{\sim}{\mathrm{QB}}(n) = \big\{(m + s, m) \in \mathbb{N}^2 : m \geq 1, 2m + s = n, c(m + s) + c(m) + s = c(2m + s)\big\}.$$

We call the pairs in $\underset{\sim}{\mathrm{QB}}(n)$ *quasi-balanced*. Notice that, by Equation (2.5), $\underset{\sim}{\mathrm{QB}}(n)$ is always non-empty, because

$$\left(\left\lceil \frac{n}{2} \right\rceil, \left\lfloor \frac{n}{2} \right\rfloor\right) \in \underset{\sim}{\mathrm{QB}}(n).$$

### 2.2.1 Describing the pairs in $\underset{\sim}{\mathrm{QB}}(n)$

The results in this subsection provide a characterization of the pairs $(n_1, n_2) \in \underset{\sim}{\mathrm{QB}}(n)$ which will allow us to compute efficiently this set.

We will begin by constructing, for each $s \in \mathbb{N}$, the succession of all $n_i(s) \in \mathbb{N}$ such that $(n_i(s) + s, n_i(s))$ is a quasi-balanced pair. When $s = 0$, the sequence $n_i(0)$ is the whole $\mathbb{N}_{\geq 1}$, because $(n, n) \in \underset{\sim}{\mathrm{QB}}(2n)$. So, in the next lemmata we shall focus on the case $s > 0$.

The following is a technical lemma that will be extremely useful in the remaining of this subsection, allowing us to decompose a quasi-balanced pair $(n_i(s) + s, n_i(s))$ into two pairs of roughly the same magnitude.

**Lemma 2.14.** *For every $n, s \in \mathbb{N}_{\geq 1}$, the pair $(n+s, n)$ is quasi-balanced if, and only if, both pairs $(\lceil \frac{n}{2} \rceil + \lfloor \frac{s}{2} \rfloor, \lceil \frac{n}{2} \rceil)$ and $(\lfloor \frac{n}{2} \rfloor + \lceil \frac{s}{2} \rceil, \lfloor \frac{n}{2} \rfloor)$ are quasi-balanced and, moreover, $n \in 2\mathbb{N}$ or $s \notin 2\mathbb{N}$.*

*Proof.* Note that, by Corollary 2.6, we already know that $(n + s, n)$ is never quasi-balanced if $n \notin 2\mathbb{N}$ and $s \in 2\mathbb{N} \setminus \{0\}$. So, three cases remain to be discussed:

- If $n \in 2\mathbb{N}, s \in 2\mathbb{N}$, so that we can write $n = 2n'$ and $s = 2s'$, then

$$\begin{aligned} & c(2n' + 2s') + c(2n') + 2s' = c(4n' + 2s') \\ & \quad \iff 2c(n' + s') + 2c(n') + 2s' = 2c(2n' + s') \\ & \quad \iff c(n' + s') + c(n') + s' = c(2n' + s'). \end{aligned}$$

- If $n \in 2\mathbb{N}, s \notin 2\mathbb{N}$, so that we can write $n = 2n'$ and $s = 2s' + 1$, then

$$\begin{aligned} & c(2n' + 2s' + 1) + c(2n') + 2s' + 1 = c(4n' + 2s' + 1) \\ & \iff c(n' + s' + 1) + c(n' + s') + 1 + 2c(n') + 2s' + 1 \\ & \qquad = c(2n' + s' + 1) + c(2n' + s') + 1 \\ & \iff c(n' + s' + 1) + c(n') + s' + 1 = c(2n' + s' + 1) \\ & \qquad \text{and } c(n' + s') + c(n') + s' = c(2n' + s') \end{aligned}$$

since, by Lemma 2.5,

$$\begin{aligned} & c(n' + s' + 1) + c(n') + s' + 1 \geq c(2n' + s' + 1) \\ & c(n' + s') + c(n') + s' \geq c(2n' + s'). \end{aligned}$$

- If $n \notin 2\mathbb{N}, s \notin 2\mathbb{N}$, so that we can write $n = 2n' + 1$ and $s = 2s' + 1$, then

$$\begin{aligned} & c(2n' + 2s' + 2) + c(2n' + 1) + 2s' + 1 = c(4n' + 2s' + 3) \\ & \iff 2c(n' + s' + 1) + c(n' + 1) + c(n') + 1 + 2s' + 1 \\ & \qquad = c(2n' + s' + 2) + c(2n' + s' + 1) + 1 \\ & \iff c(n' + s' + 1) + c(n' + 1) + s' = c(2n' + s' + 2) \\ & \qquad \text{and } c(n' + s' + 1) + c(n') + s' + 1 = c(2n' + s' + 1) \end{aligned}$$

since, by Lemma 2.5,

$$\begin{aligned} & c(n' + s' + 1) + c(n' + 1) + s' \geq c(2n' + s' + 2) \\ & c(n' + s' + 1) + c(n') + s' + 1 \geq c(2n' + s' + 1). \end{aligned}$$

This completes the proof of the equivalence in the statement. □

In fact, as we will see, the sequence $n_i(s)$ will be composed of two disjoint arithmetic subsequences of step $2^{\lceil \log_2(s) \rceil}$ except in the case where $s = 2^k$ for some $k \in \mathbb{N}$, in which case the global sequence $n_i(s)$ is arithmetic. To begin with, the next lemma will allow us to determine the *step* of our subsequences: indeed, it will show that, for any $n_i(s)$, $n_i(s) + 2^{\lceil \log_2(s) \rceil}$ also belongs to this succession.

**Lemma 2.15.** *Let $s \in \mathbb{N}_{\geq 1}$, and let $k = \lceil \log_2(s) \rceil$. If $n \in \mathbb{N}_{\geq 1}$ is such that the pair $(n + s, n)$ is quasi-balanced, $(n + 2^k + s, n + 2^k)$ is also quasi-balanced.*

*Proof.* We proceed by induction on $s$. When $s = 1$, so that $k = 0$, $(n+1, n)$ and $(n+2, n+1)$ are always quasi-balanced. Assume now that $s \geq 2$, so that $k \geq 1$, and that the statement is true for any $s' < s$ and that $(n + s', n)$ is quasi-balanced. Our goal is to prove the equality

$$c(n + 2^k + s) + c(n + 2^k) + s = c(2n + 2^{k+1} + s). \tag{2.10}$$

Notice that, since $k \geq 1$, $n + 2^k$ has the same parity as $n$. Thus, by Lemma 2.14, this equality holds if, and only if, the next two equalities hold:

$$c\left(\left\lfloor \frac{n}{2} \right\rfloor + 2^{k-1} + \left\lceil \frac{s}{2} \right\rceil\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor + 2^{k-1}\right) + \left\lceil \frac{s}{2} \right\rceil = c\left(2\left\lfloor \frac{n}{2} \right\rfloor + 2^{k} + \left\lceil \frac{s}{2} \right\rceil\right) \tag{2.11}$$

$$c\left(\left\lceil \frac{n}{2} \right\rceil + 2^{k-1} + \left\lfloor \frac{s}{2} \right\rfloor\right) + c\left(\left\lceil \frac{n}{2} \right\rceil + 2^{k-1}\right) + \left\lfloor \frac{s}{2} \right\rfloor = c\left(2\left\lceil \frac{n}{2} \right\rceil + 2^{k} + \left\lfloor \frac{s}{2} \right\rfloor\right) \tag{2.12}$$

Now, by Lemma 2.14, $(n + s, n) \in \mathrm{QB}(2n + s)$ implies that

$$c\left(\left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{s}{2} \right\rceil\right) + c\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \left\lceil \frac{s}{2} \right\rceil = c\left(2\left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{s}{2} \right\rceil\right)$$

$$c\left(\left\lceil \frac{n}{2} \right\rceil + \left\lfloor \frac{s}{2} \right\rfloor\right) + c\left(\left\lceil \frac{n}{2} \right\rceil\right) + \left\lfloor \frac{s}{2} \right\rfloor = c\left(2\left\lceil \frac{n}{2} \right\rceil + \left\lfloor \frac{s}{2} \right\rfloor\right)$$

Therefore, on the one hand, if $s \neq 2^{k-1} + 1$, equalities (2.11) and (2.12) are true by our induction hypothesis, since $\lceil \log_2 \lceil \frac{s}{2} \rceil \rceil = \lceil \log_2 \lfloor \frac{s}{2} \rfloor \rceil = \lceil \log_2(s) \rceil - 1 = k - 1$ whenever $s \neq 2^{k-1} + 1$. On the other hand, if $s = 2^{k-1} + 1$, the argument remains the same for (2.11), because $\lceil \log_2 \lceil \frac{s}{2} \rceil \rceil = k - 1$ still holds. As to (2.12), since in this case $\lceil \log_2 \lfloor \frac{s}{2} \rfloor \rceil = k-2$, by the induction assumption we know that $\left(\lceil \frac{n}{2} \rceil + 2^{k-2} + \lfloor \frac{s}{2} \rfloor, \lceil \frac{n}{2} \rceil + 2^{k-2}\right)$ is quasi-balanced, which in turn implies, again by our induction assumption, that

$$\begin{aligned}&\left(\left\lceil \frac{n}{2} \right\rceil + 2^{k-2} + 2^{k-2} + \left\lfloor \frac{s}{2} \right\rfloor, \left\lceil \frac{n}{2} \right\rceil + 2^{k-2} + 2^{k-2}\right)\\ &\quad = \left(\left\lceil \frac{n}{2} \right\rceil + 2^{k-1} + \left\lfloor \frac{s}{2} \right\rfloor, \left\lceil \frac{n}{2} \right\rceil + 2^{k-1}\right)\end{aligned}$$

is also a quasi-balanced pair. □

Given $s \in \mathbb{N}$, we understand the sequence of positive numbers $n_i(s)$ indexed in increasing order by $i \in \mathbb{N}$. That is, given $s \in \mathbb{N}$, $n_i(s)$ is the $(i+1)$-th number such that $(n_i(s) + s, n_i(s))$ is quasi-balanced. So, for instance, $n_i(0) = i + 1$ for every $i \geq 0$. The next lemma, and its corollaries, will allow us to compute the first two members of this succession, $n_0(s)$ and $n_1(s)$.

**Lemma 2.16.** *For every $s \in \mathbb{N}$,*

*i)* $\left\lceil \frac{n_0(s)}{2} \right\rceil \geq n_0\left(\left\lfloor \frac{s}{2} \right\rfloor\right)$ *and* $\left\lfloor \frac{n_0(s)}{2} \right\rfloor \geq n_0\left(\left\lceil \frac{s}{2} \right\rceil\right)$.

*ii) If* $n_1(s) > n_0(s) + 1$, *then* $\left\lceil \frac{n_1(s)}{2} \right\rceil \geq n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right)$ *and* $\left\lfloor \frac{n_1(s)}{2} \right\rfloor \geq n_1\left(\left\lceil \frac{s}{2} \right\rceil\right)$.

*Proof.* By assumption,

$$c(n_i(s) + s) + c(n_i(s)) + s = c(2n_i(s) + s)$$

and therefore, by Lemma 2.14,

$$c\left(\left\lceil\frac{n_i(s)}{2}\right\rceil+\left\lfloor\frac{s}{2}\right\rfloor\right)+c\left(\left\lceil\frac{n_i(s)}{2}\right\rceil\right)+\left\lfloor\frac{s}{2}\right\rfloor=c\left(2\left\lceil\frac{n_i(s)}{2}\right\rceil+\left\lfloor\frac{s}{2}\right\rfloor\right)$$
$$c\left(\left\lfloor\frac{n_i(s)}{2}\right\rfloor+\left\lceil\frac{s}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n_i(s)}{2}\right\rfloor\right)+\left\lceil\frac{s}{2}\right\rceil=c\left(2\left\lfloor\frac{n_i(s)}{2}\right\rfloor+\left\lceil\frac{s}{2}\right\rceil\right).$$

This implies that $\left\lceil\frac{n_0(s)}{2}\right\rceil$, $\left\lceil\frac{n_1(s)}{2}\right\rceil$ belong to the sequence $\left(n_i\left(\left\lfloor\frac{s}{2}\right\rfloor\right)\right)_i$, and that $\left\lfloor\frac{n_0(s)}{2}\right\rfloor$, $\left\lfloor\frac{n_1(s)}{2}\right\rfloor$ belong to the sequence $\left(n_i\left(\left\lceil\frac{s}{2}\right\rceil\right)\right)_i$. Therefore, $\left\lceil\frac{n_0(s)}{2}\right\rceil$ is largest or equal than $n_0\left(\left\lfloor\frac{s}{2}\right\rfloor\right)$, the first member of $\left(n_i\left(\left\lfloor\frac{s}{2}\right\rfloor\right)\right)_i$, and $\left\lfloor\frac{n_0(s)}{2}\right\rfloor$ is largest or equal than $n_0\left(\left\lceil\frac{s}{2}\right\rceil\right)$, the first member of $\left(n_i\left(\left\lceil\frac{s}{2}\right\rceil\right)\right)_i$. This proves *(i)*.

As to *(ii)*, since $n_0(s) < n_1(s)$, it must happen that

$$\left\lceil\frac{n_0(s)}{2}\right\rceil<\left\lceil\frac{n_1(s)}{2}\right\rceil \quad \text{or} \quad \left\lfloor\frac{n_0(s)}{2}\right\rfloor<\left\lfloor\frac{n_1(s)}{2}\right\rfloor.$$

Moreover, if $n_1(s) \geq n_0(s)+2$, then both strict inequalities hold. Now, by *(i)*, if $\left\lceil\frac{n_0(s)}{2}\right\rceil < \left\lceil\frac{n_1(s)}{2}\right\rceil$, then $\left\lceil\frac{n_1(s)}{2}\right\rceil$ will be larger or equal than $n_1\left(\left\lfloor\frac{s}{2}\right\rfloor\right)$, the second member of $\left(n_i\left(\left\lfloor\frac{s}{2}\right\rfloor\right)\right)_i$, and if $\left\lfloor\frac{n_0(s)}{2}\right\rfloor < \left\lfloor\frac{n_1(s)}{2}\right\rfloor$, then $\left\lfloor\frac{n_1(s)}{2}\right\rfloor$ will be larger or equal than $n_1\left(\left\lceil\frac{s}{2}\right\rceil\right)$, the second member of $\left(n_i\left(\left\lceil\frac{s}{2}\right\rceil\right)\right)_i$. This proves *(ii)*. □

**Corollary 2.17.** *Let $s \in \mathbb{N}_{\geq 1}$ be such that $s = 2^k + r$, with $k \in \mathbb{N}$ and $0 \leq r < 2^k$. Then, $n_0(s) = 2^{\lceil \log_2(s)\rceil}$.*

*Proof.* We proceed by induction over $s$. The thesis is true when $s = 1 = 2^0$, because $c(n+1)+c(n)+1 = c(2n+1)$ for every $n \in \mathbb{N}_{\geq 1}$ implies that $n_0(1) = 1 = 2^0$ (and also $n_1(1) = 2 = 2^1$, which will be used in the Corollary 2.19).

Now let $s \geq 2$ and assume the thesis in the statement to be true for any $s' < s$. Since $\lceil \log_2\lceil \frac{s}{2}\rceil\rceil = k-1$ if $r = 0$ and $k$ if $r > 0$, by Lemma 2.16 and the induction hypothesis we have that

$$\left\lfloor\frac{n_0(s)}{2}\right\rfloor \geq n_0\left(\left\lceil\frac{s}{2}\right\rceil\right) = \begin{cases} 2^{k-1} & \text{if } r = 0\\ 2^k & \text{if } r > 0\end{cases}$$
$$\Longrightarrow n_0(s) \geq \begin{cases} 2^k = 2^{\lceil\log_2(s)\rceil} & \text{if } r = 0\\ 2^{k+1} = 2^{\lceil\log_2(s)\rceil} & \text{if } r > 0\end{cases}$$

We want to prove that this last inequality is an equality. To do that, it is enough to prove that, if $r = 0$, $(2^k+s, 2^k) = (2^{k+1}, 2^k)$ is quasi-balanced, and that, if $r > 0$, $(2^{k+1}+s, 2^{k+1})$ is quasi-balanced.

- If $s = 2^k$, then $c(2^{k+1}) + c(2^k) + 2^k = 2^k = c(2^{k+1}+2^k)$, where the last equality is a consequence of Theorem 2.11.
- If $2^k < s < 2^{k+1}$, then, by Equation (2.8):

$$c(2^{k+1}+s)+c(2^{k+1})+s = s(k+1-k)-2(s-2^k)+c(s)+s = 2^{k+1}+c(s)$$
$$c(2^{k+2}+s) = s(k+2-k)-2(s-2^k)+c(s) = 2^{k+1}+c(s).$$

This completes the proof of the identity $n_0(s) = 2^{\lceil \log_2(s) \rceil}$ for every $s \geq 1$. □

**Lemma 2.18.** *If $s \in \mathbb{N}_{\geq 1}$ is not of the form $2^{\lceil \log_2(s) \rceil} - 1$, then $n_1(s) > n_0(s) + 1$.*

*Proof.* Let us write $s$ as $s = 2^k + r$, with $k \in \mathbb{N}$ and $0 \leq r < 2^k$. We must prove that, if $r \neq 2^k - 1$, the pair $(n_0(s) + 1 + s, n_0(s) + 1)$ is not quasi-balanced. We distinguish two cases.

If $r = 0$, so that $n_0(s) = 2^k$, $(n_0(s) + 1 + s, n_0(s) + 1) = (2^{k+1} + 1, 2^k + 1)$ is not quasi-balanced because, by Corollary 2.13.*(i)* and Theorem 2.11,

$$c(2^{k+1} + 1) + c(2^k + 1) + 2^k = 2^k + 2k + 1$$

$$c(2^{k+1} + 2^k + 2) = \begin{cases} 0 & \text{if } k = 1 \\ 2^k + 2(k-2) & \text{if } k \geq 2 \end{cases}$$

Assume now that $1 \leq r \leq 2^k - 2$, so that $n_0(s) = 2^{k+1}$. If $r$ is even, then, by Corollary 2.6, $n_1(s)$ is even, too, and therefore $n_1(s)$ cannot be $2^{k+1} + 1 = n_0(s) + 1$. If $r$ is odd, write it as $r = \sum_{j=0}^{l} 2^{m_j} - 1$, with $k > m_l > \cdots > m_0 \geq 1$. Then, by Theorem 2.11,

$$\begin{aligned}
&c(n_0(s) + 1 + s) + c(n_0(s) + 1) + s \\
&\quad = c\Big(2^{k+1} + 2^k + \sum_{j=0}^{l} 2^{m_j}\Big) + c(2^{k+1} + 1) + 2^k + \sum_{j=0}^{l} 2^{m_j} - 1 \\
&\quad = 2^k + \sum_{j=0}^{l} 2^{m_j}(k + 1 - m_j - 2(l + 2 - j - 1)) + k + 1 + 2^k + \sum_{j=0}^{l} 2^{m_j} - 1 \\
&\quad = 2^{k+1} + \sum_{j=0}^{l} 2^{m_j}(k - m_j - 2l + 2j) + k \\
&c(2n_0(s) + 2 + s) = c\Big(2^{k+2} + 2^k + \sum_{j=0}^{l} 2^{m_j} + 1\Big) \\
&\quad = 2^k \cdot 2 + \sum_{j=0}^{l} 2^{m_j}(k + 2 - m_j - 2(l + 3 - j - 2)) + (k + 2 - 2(l + 2)) \\
&\quad = 2^{k+1} + \sum_{j=0}^{l} 2^{m_j}(k - m_j - 2l + 2j) + k - 2 - 2l \\
&\quad < 2^{k+1} + \sum_{j=0}^{l} 2^{m_j}(k - m_j - 2l + 2j) + k.
\end{aligned}$$

□

**Corollary 2.19.** *Let $s \in \mathbb{N}_{\geq 1}$ be such that $s = 2^k + r$, with $k \in \mathbb{N}$ and $0 \leq r < 2^k$. Then:*

- *If $r = 0$, then $n_1(s) = 2^{k+1}$.*
- *If $r > 0$, then $n_1(s) = 2^{k+1} + 2^k - r$.*

*Proof.* We proceed by induction over $s$. The truth of $n_1(1) = 2$ has been established in the proof of Corollary 2.17. Now let $s \geq 2$ and assume the thesis in the statement to be true for any $s' < s$. By Lemmata 2.16 and 2.18, if $s \neq 2^{k+1} - 1$, then

$$\left\lceil \frac{n_1(s)}{2} \right\rceil \geq n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right) \text{ and } \left\lfloor \frac{n_1(s)}{2} \right\rfloor \geq n_1\left(\left\lceil \frac{s}{2} \right\rceil\right). \tag{2.13}$$

We distinguish four cases.

- Assume that $s = 2^k$. We want to prove that $n_1(s) = 2^{k+1}$. By (2.13) and the induction hypothesis
$$\left\lfloor \frac{n_1(s)}{2} \right\rfloor \geq n_1\left(\frac{s}{2}\right) = 2^k,$$
which implies that $n_1(s) \geq 2^{k+1}$. It remains to prove that $(2^{k+1} + 2^k, 2^{k+1})$ is quasi-balanced. But it is true, because, since $n_0(s) = 2^k$, $(2^{k+1}, 2^k)$ is quasi-balanced and then $(2^{k+1} + 2^k, 2^{k+1})$ is also quasi-balanced by Lemma 2.15.

- Assume that $s = 2^{k+1}-1$. We want to prove that $n_1(s) = 2^{k+1}+1$, and since $n_0(s) = 2^{k+1}$, this amounts to prove that $(2^{k+1} + 1 + s, 2^{k+1} + 1) = (2^{k+2}, 2^{k+1} + 1)$ is quasi-balanced:
$$\begin{aligned} &c(2^{k+2}) + c(2^{k+1} + 1) + 2^{k+1} - 1 = k + 1 + 2^{k+1} - 1 = 2^{k+1} + k \\ &c(2^{k+2} + 2^{k+1} + 1) = 2^{k+1} + (k + 2 - 2) = 2^{k+1} + k. \end{aligned}$$

- Assume that $s = 2^k + 1 < 2^{k+1} - 1$, and in particular that $k \geq 2$. We want to prove that $n_1(s) = 2^{k+1} + 2^k - 1$. By (2.13) and the induction hypothesis,
$$\left\lfloor \frac{n_1(s)}{2} \right\rfloor \geq n_1\left(\left\lceil \frac{s}{2} \right\rceil\right) = n_1(2^{k-1} + 1) = 2^k + 2^{k-1} - 1,$$
which implies that $n_1(s) \geq 2^{k+1} + 2^k - 2$. So, to prove that $n_1(s) = 2^{k+1} + 2^k - 1$, it is enough to check that
$$(2^{k+1} + 2^k - 1 + 2^k + 1, 2^{k+1} + 2^k - 1) = (2^{k+2}, 2^{k+1} + 2^k - 1)$$
is quasi-balanced and that $(2^{k+2} - 1, 2^{k+1} + 2^k - 2)$ is not quasi-balanced if $k \geq 2$. Both assertions can be checked using Corollary 2.13.(iv) and Theorem 2.11:

– $(2^{k+2}, 2^{k+1} + 2^k - 1)$ is quasi-balanced:
$$\begin{aligned} &c(2^{k+2}) + c(2^{k+1} + 2^k - 1) + 2^k + 1 \\ &\quad = c\left(2^{k+1} + 2^k + 1\right) + 2^k + 1 \text{ (by Cor. 2.13.(iv))} \\ &\quad = 2^k + k - 1 + 2^k + 1 = 2^{k+1} + k \\ &c(2^{k+2} + 2^{k+1} + 2^k - 1) = c(2^{k+2} + 2^k + 1) \text{ (again by Cor. 2.13.(iv))} \\ &\quad = 2^{k+1} + k. \end{aligned}$$

– $(2^{k+2}-1, 2^{k+1}+2^k-2)$ is not quasi-balanced if $k \geq 2$:

$$\begin{aligned}
&c(2^{k+2}-1)+c(2^{k+1}+2^k-2)+2^k+1\\
&\quad= c(2^{k+1}+1)+c(2^{k+1}+2^k+2)+2^k+1 \text{ (by Cor. 2.13.(iv))}\\
&\quad= k+1+2^k+2(k+1-1-2)+2^k+1 = 2^{k+1}+3k-2\\
&c(2^{k+2}+2^{k+1}+2^k-3) = c(2^{k+2}+2^k+2+1) \text{ (again by Cor. 2.13.(iv))}\\
&\quad= 2^k\cdot 2+2(k+2-1-2)+(k+2-4) = 2^{k+1}+3k-4.
\end{aligned}$$

- Assume finally that $s = 2^k + r$ for $2 \leq r < 2^k - 1$. We want to prove that $n_1(s) = 2^{k+1} + 2^k - r$. In this case, $\lceil \log_2 \lceil \frac{s}{2} \rceil \rceil = \lceil \log_2 \lfloor \frac{s}{2} \rfloor \rceil = k$. By (2.13) and the induction hypothesis, the following two inequalities hold

$$\begin{aligned}
\left\lceil \frac{n_1(s)}{2} \right\rceil &\geq n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right) = n_1\left(2^{k-1}+\left\lfloor \frac{r}{2} \right\rfloor\right) = 2^k+2^{k-1}-\left\lfloor \frac{r}{2} \right\rfloor\\
\left\lfloor \frac{n_1(s)}{2} \right\rfloor &\geq n_1\left(\left\lceil \frac{s}{2} \right\rceil\right) = n_1\left(2^{k-1}+\left\lceil \frac{r}{2} \right\rceil\right) = 2^k+2^{k-1}-\left\lceil \frac{r}{2} \right\rceil
\end{aligned}$$

and then

$$\begin{aligned}
n_1(s) &= \left\lceil \frac{n_1(s)}{2} \right\rceil + \left\lfloor \frac{n_1(s)}{2} \right\rfloor \geq n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right) + n_1\left(\left\lceil \frac{s}{2} \right\rceil\right)\\
&= 2^k+2^{k-1}-\left\lfloor \frac{r}{2} \right\rfloor + 2^k+2^{k-1}-\left\lceil \frac{r}{2} \right\rceil = 2^{k+1}+2^k-r
\end{aligned}$$

and it is enough to check that

$$(2^{k+1}+2^k-r+2^k+r, 2^{k+1}+2^k-r) = (2^{k+2}, 2^{k+1}+2^k-r)$$

is quasi-balanced. Now, if $r$ is even, this is true because

$$c\left(n_1\left(\frac{s}{2}\right)+\frac{s}{2}\right)+c\left(n_1\left(\frac{s}{2}\right)\right)+\frac{s}{2} = c\left(2n_1\left(\frac{s}{2}\right)+\frac{s}{2}\right)$$

implies

$$c\left(2n_1\left(\frac{s}{2}\right)+s\right)+c\left(2n_1\left(\frac{s}{2}\right)\right)+s = c\left(4n_1\left(\frac{s}{2}\right)+s\right).$$

Let us check the case when $r$ is odd, say $r = 2r_0 + 1$. In this case, we have that

$$\begin{aligned}
&c\left(n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right)+\left\lfloor \frac{s}{2} \right\rfloor\right)+c\left(n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right)\right)+\left\lfloor \frac{s}{2} \right\rfloor = c\left(2n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right)+\left\lfloor \frac{s}{2} \right\rfloor\right)\\
&\quad\Longrightarrow c(2^k+2^{k-1}-r_0+2^{k-1}+r_0)+c(2^k+2^{k-1}-r_0)+2^{k-1}+r_0\\
&\qquad= c(2^{k+1}+2^k-2r_0+2^{k-1}+r_0)\\
&\quad\Longrightarrow c(2^k+2^{k-1}-r_0)+2^{k-1}+r_0 = c(2^{k+1}+2^k+2^{k-1}-r_0)\\
&c\left(n_1\left(\left\lceil \frac{s}{2} \right\rceil\right)+\left\lceil \frac{s}{2} \right\rceil\right)+c\left(n_1\left(\left\lceil \frac{s}{2} \right\rceil\right)\right)+\left\lceil \frac{s}{2} \right\rceil = c\left(2n_1\left(\left\lceil \frac{s}{2} \right\rceil\right)+\left\lceil \frac{s}{2} \right\rceil\right)\\
&\quad\Longrightarrow c(2^k+2^{k-1}-r_0-1+2^{k-1}+r_0+1)+c(2^k+2^{k-1}-r_0-1)\\
&\qquad\quad+2^{k-1}+r_0+1 = c(2^{k+1}+2^k-2r_0-2+2^{k-1}+r_0+1)\\
&\quad\Longrightarrow c(2^k+2^{k-1}-r_0-1)+2^{k-1}+r_0+1 = c(2^{k+1}+2^k+2^{k-1}-r_0-1)
\end{aligned}$$

from which we deduce that

$$\begin{aligned}
&c(2^{k+1}+2^k-r+s)+c(2^{k+1}+2^k-r)+s\\
&\quad=c(2^{k+2})+c(2^{k+1}+2^k-2r_0-1)+2^k+2r_0+1\\
&\quad=c(2^k+2^{k-1}-r_0)+c(2^k+2^{k-1}-r_0-1)+2^k+2r_0+2\\
&\quad=c(2^{k+1}+2^k+2^{k-1}-r_0)+c(2^{k+1}+2^k+2^{k-1}-r_0-1)+1\\
&\quad=c(2^{k+2}+2^{k+1}+2^k-2r_0-1)=c(2^{k+2}+2^{k+1}-2r+s).
\end{aligned}$$

This completes the proof of the induction step. □

It remains to prove that, for every $s$, there are only two maximal arithmetic subsuccessions in $(n_i(s))_i$ whose difference of progression is $2^{\lceil \log_2(s)\rceil}$. In order to do that, we shall use the next two lemmata.

**Lemma 2.20.** *Let $s\in\mathbb{N}_{\geq 1}$, and let $k=\lceil\log_2(s)\rceil$. If $n\in\mathbb{N}_{\geq 1}$ is such that the pair $(n+s,n)$ is quasi-balanced and if $n-2^k\geq n_0(s)$, then $(n-2^k+s,n-2^k)$ is also quasi-balanced.*

*Proof.* We proceed by induction over $s$. The case when $s=1$ is clear, because every pair $(n+1,n)$ is quasi-balanced. Assume now that $s\geq 2$ and that the thesis is true for every $1\leq s'<s$, and let $n$ be such that $(n+s,n)$ is quasi-balanced and $n-2^k\geq n_0(s)$. Our goal is to prove that

$$c(n-2^k+s)+c(n-2^k)+s=c(2n-2^{k+1}+s).$$

By Lemma 2.14, and since $n+2^k$ has the same parity as $n$ because $s\geq 2$ and hence $k\geq 1$, it is enough to prove that

$$\begin{gathered}
c\left(\left\lfloor\frac{n}{2}\right\rfloor-2^{k-1}+\left\lceil\frac{s}{2}\right\rceil\right)+c\left(\left\lfloor\frac{n}{2}\right\rfloor-2^{k-1}\right)+\left\lceil\frac{s}{2}\right\rceil=c\left(2\left\lfloor\frac{n}{2}\right\rfloor-2^k+\left\lceil\frac{s}{2}\right\rceil\right)\\
c\left(\left\lceil\frac{n}{2}\right\rceil-2^{k-1}+\left\lfloor\frac{s}{2}\right\rfloor\right)+c\left(\left\lceil\frac{n}{2}\right\rceil-2^{k-1}\right)+\left\lfloor\frac{s}{2}\right\rfloor=c\left(2\left\lceil\frac{n}{2}\right\rceil-2^k+\left\lfloor\frac{s}{2}\right\rfloor\right)
\end{gathered}$$

Now, these assertions are true, because if $n-2^k\geq n_0(s)$, and using Lemma 2.16.*(i)* and the fact that $n_0(s)$ is a power of 2,

$$\left\lceil\frac{n}{2}\right\rceil-2^{k-1}\geq\left\lfloor\frac{n}{2}\right\rfloor-2^{k-1}\geq\frac{n_0(s)}{2}\geq n_0\left(\left\lceil\frac{s}{2}\right\rceil\right)\geq n_0\left(\left\lfloor\frac{s}{2}\right\rfloor\right)$$

and therefore, by the induction hypothesis, both $(\lfloor\frac{n}{2}\rfloor-2^{k-1}+\lceil\frac{s}{2}\rceil,\lfloor\frac{n}{2}\rfloor-2^{k-1})$ and $(\lceil\frac{n}{2}\rceil-2^{k-1}+\lfloor\frac{s}{2}\rfloor,\lceil\frac{n}{2}\rceil-2^{k-1})$ are quasi-balanced. □

**Lemma 2.21.** *Let $s\in\mathbb{N}_{\geq 1}$, and let $k=\lceil\log_2(s)\rceil$. There does not exist any $m\in\mathbb{N}$, with $n_1(s)<m<n_0(s)+2^k=2^{k+1}$, such that $(m+s,m)$ is quasi-balanced.*

*Proof.* We proceed by induction over $s$. For $s=1$ it is obviously true, since $n_1(s)=2$ and $n_2(s)=3$. Let now $s\geq 2$, assume the thesis to be true for any $1\leq s'<s$. Let $m\in\mathbb{N}$ be such that $n_1(s)<m<2^{k+1}$. We shall assume that $(m+s,m)$ is quasi-balanced and we shall reach a contradiction. Notice that $n_1(s)<m<2^{k+1}$ implies that

$$\left\lceil\frac{n_1(s)}{2}\right\rceil\leq\left\lfloor\frac{m}{2}\right\rfloor\leq\left\lceil\frac{m}{2}\right\rceil\leq 2^k.$$

We distinguish three cases:

- If $s = 2^k$, so that $n_0(s) = 2^k$ and $n_1(s) = 2^{k+1} = n_0(s) + 2^k$, there cannot exist such $m$.
- If $s = 2^{k-1} + 1$, then $n_1(s) = 2^k + 2^{k-1} - 1$. In this case, $n_1(s) < m < 2^{k+1}$ implies that

$$n_1\left(\left\lfloor \frac{s}{2} \right\rfloor\right) = 2^{k-1} < 2^{k-1} + 2^{k-2} = \left\lceil \frac{n_1(s)}{2} \right\rceil \leq \left\lfloor \frac{m}{2} \right\rfloor \leq \left\lceil \frac{m}{2} \right\rceil \leq 2^k$$

$$n_1\left(\left\lceil \frac{s}{2} \right\rceil\right) = 2^{k-1} + 2^{k-2} - 1 = \left\lfloor \frac{n_1(s)}{2} \right\rfloor \leq \left\lfloor \frac{m}{2} \right\rfloor \leq \left\lceil \frac{m}{2} \right\rceil \leq 2^k$$

and by Lemma 2.14, if $(m + s, m)$ is quasi-balanced, then so are $\left(\lfloor \frac{m}{2} \rfloor + \lceil \frac{s}{2} \rceil, \lfloor \frac{m}{2} \rfloor\right)$ and $\left(\lceil \frac{m}{2} \rceil + \lfloor \frac{s}{2} \rfloor, \lceil \frac{m}{2} \rceil\right)$. Now, due to the impossibility of $n_1\left(\lfloor \frac{s}{2} \rfloor\right) = \lceil \frac{m}{2} \rceil$, by the induction hypothesis it must happen that $\lceil \frac{m}{2} \rceil = 2^k$, and thus $m = 2^{k+1} - 1$. But then, $\left(\lfloor \frac{m}{2} \rfloor + \lceil \frac{s}{2} \rceil, \lfloor \frac{m}{2} \rfloor\right)$ could not be quasi-balanced.

- If $s = 2^{k-1} + r$ with $2 \leq r \leq 2^{k-1} - 1$, then $n_1(s) = 2^k + 2^{k-1} - r$ and $n_1(\lfloor \frac{s}{2} \rfloor) = n_1(2^{k-2} + \lfloor \frac{r}{2} \rfloor) = 2^{k-1} + 2^{k-2} - \lfloor \frac{r}{2} \rfloor$. In this case, $n_1(s) < m < 2^{k+1}$ implies that

$$\begin{aligned} n_1\left(\left\lceil \frac{s}{2} \right\rceil\right) &= 2^{k-1} + 2^{k-2} - \left\lceil \frac{r}{2} \right\rceil = \left\lfloor \frac{2^k + 2^{k-1} - r}{2} \right\rfloor = \left\lfloor \frac{n_1(s)}{2} \right\rfloor \\ &\leq \left\lfloor \frac{m}{2} \right\rfloor \leq \left\lceil \frac{m}{2} \right\rceil \leq 2^k \end{aligned}$$

and by Lemma 2.14, if $(m + s, m)$ is quasi-balanced, then $\left(\lfloor \frac{m}{2} \rfloor + \lceil \frac{s}{2} \rceil, \lfloor \frac{m}{2} \rfloor\right)$ is quasi-balanced, too. By the induction hypothesis, all inequalities but one in the sequence above must be equalities. We distinguish two cases:

- If $m$ is odd, so that $\lfloor \frac{m}{2} \rfloor < \lceil \frac{m}{2} \rceil$, we must have

$$2^{k-1} + 2^{k-2} - \left\lceil \frac{r}{2} \right\rceil = \left\lfloor \frac{m}{2} \right\rfloor \text{ and } \left\lceil \frac{m}{2} \right\rceil = 2^k$$

but this is impossible, because it implies that $2^{k-1} + 2^{k-2} - \lceil \frac{r}{2} \rceil + 1 = 2^k$, which cannot happen if $r \geq 2$.

- If $m$ is even, then $m < 2^{k+1}$ implies $\frac{m}{2} < 2^k$, and hence it must happen $2^{k-1}+2^{k-2}-\lceil \frac{r}{2} \rceil = \frac{m}{2}$ and thus

$$m = 2^k + 2^{k-1} - 2\left\lceil \frac{r}{2} \right\rceil > n_1(s) = 2^k + 2^{k-1} - r$$

which is again impossible, because $r \leq 2\lceil \frac{r}{2} \rceil$.

□

**Theorem 2.22.** *Let $s \in \mathbb{N}_{\geq 1}$, and let $k = \lceil \log_2(s) \rceil$. Then, $(n + s, n)$ is quasi-balanced if, and only if, $n = n_0(s) + t \cdot 2^k$ or $n = n_1(s) + t \cdot 2^k$ for some $t \in \mathbb{N}$.*

*Proof.* By Lemma 2.15, all pairs of the forms $(n_0(s) + t \cdot 2^k + s, n_0(s) + t \cdot 2^k)$ or $(n_1(s) + t \cdot 2^k + s, n_1(s) + t \cdot 2^k)$ are quasi-balanced. Let now $m_0 \in \mathbb{N}$ be such that $(m_0 + s, m_0)$ is quasi-balanced, and let $t_0$ be the largest integer such that $n_0(s) \leq m_0 - t_0 \cdot 2^k$. By Lemma 2.20, $(m_0 - t_0 \cdot 2^k + s, m_0 - t_0 \cdot 2^k)$ is also quasi-balanced, and therefore $m_0 - t_0 \cdot 2^k = n_i(s)$ for some $i \in \mathbb{N}$. If $m_0 - t_0 \cdot 2^k = n_0(s)$, then $m_0$ is of the first form described in the statement. If $n_0(s) < m_0 - t_0 \cdot 2^k$, then $i \geq 1$ and hence $m_0 - t_0 \cdot 2^k \geq n_1(s)$. But in this case, since $m_0 - t_0 \cdot 2^k < n_0(s) + 2^k$ (otherwise, we could subtract another $2^k$ to $m_0$), by Lemma 2.21 $m_0 - t_0 \cdot 2^k \leq n_1(s)$ and hence $m_0 - t_0 \cdot 2^k = n_1(s)$ and $m_0$ is of the second form described in the statement. □

**Remark 2.23.** Notice that the theorem above entails that the two subsuccessions of $n_i(s)$ described in it are maximal.

These considerations would be of little help as they are, since they focus on the difference between the number of leaves of the two maximal subtrees hanging from the root, and do not, by themselves, give an answer when the total number of leaves is fixed. We now pursue this answer, by giving a series of results that will characterize the pairs of $\underset{\sim}{\mathrm{QB}}(n)$ for a given $n \in \mathbb{N}_{\geq 1}$.

**Lemma 2.24.** *For every $n, p, n_1, n_2 \in \mathbb{N}_{\geq 1}$ such that $n_1 \geq n_2$ and $n = n_1 + n_2$, $(n_1, n_2) \in \underset{\sim}{\mathrm{QB}}(n)$ if, and only if, $(2^p n_1, 2^p n_2) \in \underset{\sim}{\mathrm{QB}}(2^p n)$.*

*Proof.* This result is a direct consequence of Corollary 2.8. Indeed,

$$\begin{aligned} & c(2^p n_1) + c(2^p n_2) + 2^p n_1 - 2^p n_2 = c(2^p n_1 + 2^p n_2) \\ & \quad \Longleftrightarrow 2^p c(n_1) + 2^p c(n_2) + 2^p (n_1 - n_2) = 2^p c(n_1 + n_2) \\ & \quad \Longleftrightarrow c(n_1) + c(n_2) + n_1 - n_2 = c(n_1 + n_2). \end{aligned}$$

□

So, let $n \in \mathbb{N}_{\geq 2}$, write it as $n = 2n' + s$ for some $n' \in \mathbb{N}_{\geq 1}$ and $s \in \mathbb{N}$. If $s \in \{0, 1\}$, it is clear that $(n' + s, n') \in \underset{\sim}{\mathrm{QB}}(n)$ by Equation 2.5. Assume now that $s \geq 2$, let $k = \lceil \log_2(s) \rceil$ and write $s$ as $s = 2^k - r$ with $0 \leq r < 2^{k-1}$. By Theorem 2.22, $n'$ must be of one of the following two forms, for some $t \in \mathbb{N}$:

$$n' = n_0(s) + t \cdot 2^k = (1 + t)2^k,$$

in which case

$$n' + s = (1 + t)2^k + 2^k - r = (2 + t)2^k - r,$$

or

$$n' = n_1(s) + t \cdot 2^k = \begin{cases} 2^{k+1} + t \cdot 2^k = (2 + t)2^k & (\text{if } r = 0) \\ 2^k + r + t \cdot 2^k = (1 + t)2^k + r & (\text{if } r > 0) \end{cases}$$

in which case

$$n' + s = \begin{cases} (2 + t)2^k + 2^k = (3 + t)2^k & (\text{if } r = 0) \\ (1 + t)2^k + r + 2^k - r = (2 + t)2^k & (\text{if } r > 0) \end{cases}$$

This proves the following theorem, by setting $q = t + 1$:

**Theorem 2.25.** *For every $n \in \mathbb{N}_{\geq 2}$, the pairs of $\underset{\sim}{\mathrm{QB}}(n)$ are exactly those pairs of the following forms:*

*(i)* $(\lceil n/2 \rceil, \lfloor n/2 \rfloor)$

*(ii)* $((q + 1)2^k - r, q \cdot 2^k)$ *for some $q \in \mathbb{N}_{\geq 1}$, $k \in \mathbb{N}_{\geq 1}$, and $0 < r < 2^{k-1}$ such that $n = q \cdot 2^{k+1} + 2^k - r$*

*(iii)* $((q+1)2^k, q \cdot 2^k + r)$ *for some* $q \in \mathbb{N}_{\geq 1}$, $k \in \mathbb{N}_{\geq 1}$ *and* $0 \leq r < 2^{k-1}$ *such that* $n = q \cdot 2^{k+1} + 2^k + r$.

**Corollary 2.26.** *Let* $n \in \mathbb{N}_{\geq 2}$ *and* $p \in \mathbb{N}$ *be such that* $2^p$ *divides* $n$*, and let* $n_0 = n/2^p$*. Then,* $(n_1, n_2) \in QB(n)$ *if, and only if,* $n_1 = n_2$ *or there exists* $(n_1^0, n_2^0) \in \utilde{\mathrm{QB}}(n_0)$ *such that* $(n_1, n_2) = (2^p n_1^0, 2^p n_2^0)$*.*

*Proof.* By Lemma 2.24 and Equation (2.5), it is enough to prove that if $(n_1, n_2) \in QB(n)$ with $n_1 > n_2$, then there exists $(n_1^0, n_2^0) \in \utilde{\mathrm{QB}}(n_0)$ such that $(n_1, n_2) = (2^p n_1^0, 2^p n_2^0)$. We prove this assertion by induction on $p$. The case when $p = 0$ is tautological.

Let us prove now the case $p = 1$, so that $n$ is even: let $n' = n/2$. Let $(n_1, n_2) \in \utilde{\mathrm{QB}}(n)$ with $n_1 > n_2$. By the last theorem, this pair must be of the forms *(ii)* or *(iii)* in its statement, and therefore there must exist some $q \in \mathbb{N}_{\geq 1}$, $k \in \mathbb{N}$, and $0 \leq r < 2^{k-1}$ such that $n = q \cdot 2^{k+1} + 2^k - r$, and then $(n_1, n_2) = ((q+1)2^k - r, q \cdot 2^k)$, or $n = q \cdot 2^{k+1} + 2^k + r$, and then $(n_1, n_2) = ((q+1)2^k, q \cdot 2^k + r)$. But in both cases, since $n$ is even, $k$ must be at least 1 and $r$ must be even, say $r = 2r'$. Then, in the first case, $(n_1, n_2) = (2n_1', 2n_2')$ with $(n_1', n_2') = ((q+1)2^{k-1} - r', q \cdot 2^{k-1}) \in \utilde{\mathrm{QB}}(n')$ because $n' = q \cdot 2^k + 2^{k-1} - r'$, and in the second case $(n_1, n_2) = (2n_1', 2n_2')$ with $(n_1', n_2') = ((q+1)2^{k-1}, q \cdot 2^{k-1} + r') \in \utilde{\mathrm{QB}}(n')$ because $n' = q \cdot 2^k + 2^{k-1} + r'$. So, in summary, we have proved that if $(n_1, n_2) \in \utilde{\mathrm{QB}}(n)$ with $n_1 > n_2$, then there exists $(n_1', n_2') \in \utilde{\mathrm{QB}}(n')$ such that $(n_1, n_2) = (2n_1', 2n_2')$.

Let now $p \geq 2$, assume that the assertion is true for every $n' \in \mathbb{N}_{\geq 2}$ and every $0 \leq p' < p$ such that $2^{p'}$ divides $n'$, and let $n \in \mathbb{N}_{\geq 2}$ be such that $2^p$ divides $n$. Let $n' = n/2$ and $n_0 = n/2^p = n'/2^{p-1}$, and let $(n_1, n_2) \in \mathbb{N}_{\geq 1}^2$ with $n_1 > n_2$. Then

$$
\begin{aligned}
&(n_1, n_2) \in \utilde{\mathrm{QB}}(n) \\
&\quad \Longrightarrow (n_1, n_2) = (2n_1', 2n_2') \ \text{ for some } (n_1', n_2') \in \utilde{\mathrm{QB}}(n') \\
&\quad \Longrightarrow (n_1, n_2) = (2 \cdot 2^{p-1} n_1^0, 2 \cdot 2^{p-1} n_1^0) = (2^p n_1^0, 2^p n_1^0) \ \text{ for some } (n_1^0, n_2^0) \in \utilde{\mathrm{QB}}(n_0)
\end{aligned}
$$

where the first implication is due to the case $p = 1$ and the second, to the induction hypothesis. □

### 2.2.2 Computing the elements of $\utilde{\mathrm{QB}}(n)$

Theorem 2.25 will allow us to obtain a non-redundant description of the set $\utilde{\mathrm{QB}}(n)$ that will yield an algorithmic approach to its computation as well as an expression for its cardinality. We illustrate this with an easy example, before stating and proving the aforementioned description.

> **Example:**
>
> Let $n = 813$, whose binary expression is $1100101101_{(2)}$. By Theorem 2.25 *(i)*, $n$ can be decomposed into $110010111_{(2)}$ and $110010110_{(2)}$ and give rise to a pair of $\utilde{\mathrm{QB}}(n)$.
>
> Both *(ii)* and *(iii)* contain $q \cdot 2^k$ for some $q \in \mathbb{N}_{\geq 1}$ in some coordinate. Let us focus now on *(ii)*; we want to find all decompositions of $n$ of the form $n = q \cdot 2^{k+1} + 2^k - r$, for some $k \in \mathbb{N}$, and $0 < r < 2^{k-1}$. We can write the pair *(ii)* as $(q \cdot 2^k + s, q \cdot 2^k)$, with $s = 2^k - r \leq 2^k$. In terms of the binary decomposition of $n = q \cdot 2^{k+1} + s$, it

means that, since $\lfloor \log_2(s) \rfloor < \left\lfloor \log_2(q \cdot 2^{k+1}) \right\rfloor - 1$, $n$ will be of the form

$$n = 1\dots 0\overbrace{1\dots x}^{s}{}_{(2)}$$

with $x \in \{0, 1\}$. Thus, in principle, $n = 813 = 1100101101_{(2)}$ could be decomposed in pairs of the form *(ii)* in the following manners:

- $s = 101101_{(2)}$, and therefore $q \cdot 2^{k+1} = 1100000000_{(2)}$. We have, thus,

$$q \cdot 2^k + s = 110101101_{(2)} \quad \text{and} \quad q \cdot 2^k = 110000000_{(2)}$$

- $s = 1101_{(2)}$, and therefore $q \cdot 2^{k+1} = 1100100000_{(2)}$;

$$q \cdot 2^k + s = 110001101_{(2)} \quad \text{and} \quad q \cdot 2^k = 110100000_{(2)}$$

- $s = 1_{(2)}$, which corresponds to the pair *(i)*.

Finally, let us discuss the pairs of the form *(iii)*. These account to writing $n$ in the form $n = q \cdot 2^{k+1} + 2^k + r$ and therefore the pair as $((q + 1)2^k, q \cdot 2^k + r)$, where $0 \leq r < 2^{k-1}$. Therefore, as $\lfloor \log_2(r) \rfloor < k - 1$, we will consider $n$ to be of the form

$$n = 1\dots 1\overbrace{0\dots x}^{r}{}_{(2)}$$

with $x \in \{0, 1\}$. Therefore, $n = 813 = 1100101101_{(2)}$ could be decomposed in pairs of the form *(iii)* in the following manners:

- $r = 101101_{(2)}$, and therefore $2^k = 2^9 = 100000000_{(2)}, q = 1$, and so

$$q \cdot 2^k + 2^k = 1000000000_{(2)} \quad \text{and} \quad q \cdot 2^k + r = 100101101_{(2)}$$

- $r = 1101_{(2)}$, and therefore $2^k = 2^5 = 100000_{(2)}, q = 2^4 + 2^3$, and so

$$q \cdot 2^k + 2^k = 110100000_{(2)} \quad \text{and} \quad q \cdot 2^k + r = 110001101_{(2)}$$

- $r = 1_{(2)}$, and therefore $2^k = 2^2 = 100_{(2)}, q = 2^7 + 2^6 + 2^3 + 2$, and so

$$q \cdot 2^k + 2^k = 110000100_{(2)} \quad \text{and} \quad q \cdot 2^k + r = 110101001_{(2)}$$

The previous example is an instance of the following result.

**Corollary 2.27.** *Let $n \in \mathbb{N}_{\geq 2}$, let $2^p$ be the highest power of 2 that divides $n$, let $n_0 = n/2^p$, and let $n_0 = \sum_{i=0}^{\ell} 2^{m_i}$, with $0 = m_0 < \cdots < m_{\ell-1} < m_\ell$, be the binary decomposition of $n_0$. Then,*

*(i)* *If $\ell = 0$, i.e., if $n = 2^p$, then $\underset{\sim}{\mathrm{QB}}(n) = \{(n/2, n/2)\}$.*

*(ii)* *If $\ell > 0$, then*

*(ii.1)* $\underset{\smile}{\mathrm{QB}}(n)$ *always contains the pair*

$$\Big(2^p\Big(\sum_{i=1}^{\ell}2^{m_i-1}+1\Big),2^p\Big(\sum_{i=1}^{\ell}2^{m_i-1}\Big)\Big).$$

*(ii.2) For every* $j\in\{1,\dots,\ell-1\}$ *such that* $m_{j+1}>m_j+1$, $\underset{\smile}{\mathrm{QB}}(n)$ *contains the pair*

$$\Big(n-2^p\sum_{i=j+1}^{\ell}2^{m_i-1},2^p\sum_{i=j+1}^{\ell}2^{m_i-1}\Big).$$

*(ii.3) For every* $j\in\{1,\dots,\ell-1\}$ *such that* $m_j>m_{j-1}+1$, $\underset{\smile}{\mathrm{QB}}(n)$ *contains the pair*

$$\Big(2^p\Big(\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j}\Big),n-2^p\Big(\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j}\Big)\Big).$$

*(ii.4) If* $m_0\geq 1$*; i.e., if* $n\in 2\mathbb{N}$*, then* $\underset{\smile}{\mathrm{QB}}(n)$ *contains the pair* $(n/2,n/2)$.

*And* $\underset{\smile}{\mathrm{QB}}(n)$ *contains no other pair than those described in (ii.1) to (ii.4). Furthermore, they are pairwise different.*

*Proof.* Assertion *(i)* is a consequence of the fact that $c(2^p)=0$. Indeed, if $(n_a,n_b)\in\underset{\smile}{\mathrm{QB}}(2^p)$, then $0=c(2^p)=c(n_a)+c(n_b)+n_a-n_b$ implies that $n_a=n_b=n/2$.

So, assume henceforth that $\ell\geq 1$. Let $(n_1,n_2)\in\mathbb{N}^2$ with $1\leq n_2<n_1$. By Corollary 2.26, $(n_1,n_2)\in\underset{\smile}{\mathrm{QB}}(n)$ if, and only if, there exists some $(n_1^0,n_2^0)\in\underset{\smile}{\mathrm{QB}}(n_0)$ such that $(n_1,n_2)=(2^pn_1^0,2^pn_2^0)$. Moreover, two such pairs in $\underset{\smile}{\mathrm{QB}}(n)$ are different if, and only if, the corresponding pairs in $\underset{\smile}{\mathrm{QB}}(n_0)$ are different. This leads us to find a non-redundant description of the pairs $(n_1^0,n_2^0)\in\underset{\smile}{\mathrm{QB}}(n_0)$ with $n_2^0<n_1^0$ and then multiply them by $2^p$ to obtain a non-redundant description of all pairs $(n_1,n_2)\in\underset{\smile}{\mathrm{QB}}(n)$ with $n_2<n_1$. If $n$ is even, we shall only need to add the pair $(n/2,n/2)$ to those obtained in this way to obtain the whole $\underset{\smile}{\mathrm{QB}}(n)$.

So, in the rest of this proof we shall focus on $n_0$, and more specifically on the pairs $(n_1^0,n_2^0)\in\underset{\smile}{\mathrm{QB}}(n_0)$ with $n_2^0<n_1^0$. By Theorem 2.25 there are three possibilities:

*(ii.1)* Since $n_0$ is odd, $n_0=2n_0'+1$ with $n_0'=\frac{n_0-1}{2}=\sum_{i=1}^{\ell}2^{m_i-1}$. Then

$$\Big(\Big\lceil\frac{n_0}{2}\Big\rceil,\Big\lfloor\frac{n_0}{2}\Big\rfloor\Big)=(n_0'+1,n_0')=\Big(\sum_{i=1}^{\ell}2^{m_i-1}+1,\sum_{i=1}^{\ell}2^{m_i-1}\Big)\in\underset{\smile}{\mathrm{QB}}(n_0).$$

*(ii.2)* Assume $n_0=q\cdot 2^{k+1}+2^k-r$, with $q\geq 1$. Since $n_0$ is odd, there are two possibilities:

- $k=r=0$, so that $n_0=2q+1$. In this case, the pair in $\underset{\smile}{\mathrm{QB}}(n_0)$ correponding to Theorem 2.25 *(ii)* is $(q+1,q)$, the pair described in (ii.1). So, we can omit this case.
- $k\geq 1$ and $0<r<2^{k-1}$ odd. Taking $r'=2^{k-1}-r$, we can write $n_0$ as $n_0=q\cdot 2^{k+1}+2^{k-1}+r'$ with $0<r'<2^{k-1}$. Then, the equality

$$q\cdot 2^{k+1}+2^{k-1}+r'=\sum_{i=0}^{\ell}2^{m_i}$$

will only be satisfied if $k-1=m_j$ for some $j\in\{1,\dots,\ell-1\}$ such that $m_{j+1}\geq k+1=m_j+2$, in which case

$$q=\frac{\sum_{i=j+1}^{\ell}2^{m_i}}{2^{k+1}}=\frac{\sum_{i=j+1}^{\ell}2^{m_i}}{2^{m_j+2}}.$$

So, for each $j\in\{1,\dots,\ell-1\}$ such that $m_{j+1}>m_j+1$, this contributes to $QB(n_0)$ the pair $(n_1^0,n_2^0)$ with

$$n_2^0(j)=q\cdot 2^k=\sum_{i=j+1}^{\ell}2^{m_i-1},\quad n_1^0(j)=n_0-n_2^0=n_0-\sum_{i=j+1}^{\ell}2^{m_i-1}.$$

These pairs are pairwise different because $\sum_{i=j+1}^{\ell}2^{m_i-1}$ is strictly decreasing on $j$ and hence, if $j>j'$, $n_2^0(j)<n_2^0(j')$.

*(ii.3)* Finally, assume $n_0=q\cdot 2^{k+1}+2^k+r$, with $q\geq 1$ and $0\leq r<2^{k-1}$ odd. By equating $n_0$ to its binary representation, this implies that there exists $j\in\{1,\dots,\ell-1\}$ such that $k=m_j$ and $m_j>m_{j-1}+1$. In this case,

$$q=\frac{\sum_{i=j+1}^{\ell}2^{m_i}}{2^{k+1}}=\frac{\sum_{i=j+1}^{\ell}2^{m_i}}{2^{m_j+1}}.$$

So, for each $j\in\{1,\dots,\ell-1\}$ such that $m_j>m_{j-1}+1$, this contributes to $QB(n_0)$ the pair $(n_1^0,n_2^0)$ with

$$\begin{aligned} n_1^0&=(q+1)\cdot 2^k=\Big(\sum_{i=j+1}^{\ell}2^{m_i-m_j-1}+1\Big)2^{m_j}=\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j}\\ n_2^0&=n_0-n_1^0=n_0-\sum_{i=j+1}^{\ell}2^{m_i-1}-2^{m_j}.\end{aligned}$$

These pairs are pairwise different. Indeed,

$$\begin{aligned}\sum_{i=j}^{\ell}2^{m_i-1}+2^{m_{j-1}}&=\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j-1}+2^{m_{j-1}}\\ &\leq\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j-1}+2^{m_j-1}=\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j}\end{aligned}$$

and the inequality is strict if $m_j-1>m_{j-1}$, which is the condition on $j$ that adds a pair $(n_1^0(j),n_2^0(j))$ of this type to $QB(n_0)$. This implies that if $j>j'$,

$$n_1^0(j)=\sum_{i=j+1}^{\ell}2^{m_i-1}+2^{m_j}>\sum_{i=j}^{\ell}2^{m_i-1}+2^{m_{j-1}}\geq\sum_{i=j'+1}^{\ell}2^{m_i-1}+2^{m_{j'}}=n_1^0(j').$$

Cases *(ii.1)* to *(ii.3)* give all pairs $(n_1^0,n_2^0)\in QB(n_0)$ with $n_1^0>n_2^0$. Let us prove that all these pairs in $QB(n_0)$ are pairwise different.

- Along our construction we have already proved that all pairs of the form *(ii.2)* are pairwise different, as well as all pairs of the form *(ii.3)*.
- Pairs of the form *(ii.2)* are different from those of the form *(ii.1)* because their second entry is strictly smaller than that of *(ii.1)* since, in *(ii.2)*, $j \geq 1$.
- Pairs of the form *(ii.3)* are different from those of the form *(ii.1)* because their first entry is strictly larger than that of *(ii.1)*:

$$\begin{aligned}\sum_{i=1}^{\ell} 2^{m_i-1} + 1 - \Big(\sum_{i=j+1}^{\ell} 2^{m_i-1} + 2^{m_j}\Big) &= \sum_{i=1}^{j} 2^{m_i-1} + 1 - 2^{m_j} \\ &= \sum_{i=1}^{j-1} 2^{m_i-1} + 1 - 2^{m_j-1} \leq \sum_{i=0}^{m_{j-1}-1} 2^{i} + 1 - 2^{m_j-1} = 2^{m_{j-1}} - 2^{m_j-1} < 0\end{aligned}$$

  because $m_{j-1} < m_j - 1$.
- Pairs of the form *(ii.2)* are different from those of *(ii.3)* because those of the form *(ii.2)* have their first entry odd, while those of the form *(ii.2)* have first entry even.

So, we have a non-redundant description of all pairs $(n_1^0, n_2^0) \in \mathrm{QB}(n_0)$ with $n_1^0 > n_2^0$ and hence, multiplying them by $2^p$, a non-redundant description of all pairs $(n_1, n_2) \in \mathrm{QB}(n)$ with $n_1 > n_2$; if $n \in 2\mathbb{N}$, we just need to add the pair $(n/2, n/2)$ corresponding to *(ii.4)* to complete $\mathrm{QB}(n)$. This pair is oviously different from those coming from $\mathrm{QB}(n_0)$ with $n_1^0 > n_2^0$. □

The previous result gives us a way to compute $|\mathrm{QB}(n)|$, for any given $n \in \mathbb{N}_{\geq 2}$, in terms of the number $M_0(n)$ of maximal sequences of zeroes in the binary representation $n_{(2)}$ of $n$.

If $n$ is a power of 2, then $M_0(n) = 1$, and that is exactly the cardinality of $\mathrm{QB}(n)$. Assume henceforth that $n$ is not a power of 2. Then:

- The pair of the form *(ii.1)* always exists, independently of the maximal sequences of zeroes.
- A pair of the form *(ii.2)* is added to $\mathrm{QB}(n)$ for each $j \in \{1, \ldots, \ell - 1\}$ such that $m_{j+1} > m_j + 1$, that is, for each maximal sequence of zeroes surrounded by ones (i.e., not ending in the units position) and not ending immediately before the last 1. So, *(ii.2)* contributes a pair for every maximal sequence of zeroes not ending in the units position or immediately before the last 1.
- A pair of the form *(ii.3)* is added to $\mathrm{QB}(n)$ for each $j \in \{1, \ldots, \ell - 1\}$ such that $m_j > m_{j-1} + 1$, that is, for each maximal sequence of zeroes surrounded by ones and not starting immediately after the leading 1. So, *(ii.3)* contributes a pair for every maximal sequence of zeroes not ending in the units position and not starting immediately after the leading 1.
- The pair of the form *(ii.4)* is added to $\mathrm{QB}(n)$ if there is a maximal sequence of zeroes ending in the units position.

So, to compute the cardinality $|QB(n)|$ when $n$ is not a power of 2:

- We count twice the number of maximal sequences of zeroes in $n_{(2)}$ plus 1: $2M_0(n)+1$.
- We subtract 1 if $n_{(2)}$ contains a maximal sequence of zeroes ending immediately before the last 1.
- We subtract 1 if $n_{(2)}$ contains a maximal sequence of zeroes starting immediately after the leading 1.
- We subtract 2 and we add 1 (i.e. we subtract 1) if $n_{(2)}$ contains a maximal sequence of zeroes ending in the units position.

If $n \geq 2$ is a power of 2, in which case $|\mathrm{QB}(n)| = 1$, this procedure also works, because $M_0(n) = 1$ and its only maximal sequence of zeros starts immediately after the leading 1 and ends in the units position, and then we have to subtract it twice from $2M_0(n)+1 = 3$.

Let us call any maximal sequence of zeroes in the binary representation of $n$ that starts immediately after the leading 1 or ends immediately before the last 1 or in the units position *extremal*. Then, the procedure above yields the following expression for $|\mathrm{QB}(n)|$:

$$|\mathrm{QB}(n)| = 2M_0(n) + 1 - \text{ number of extremal maximal sequences of zeroes in } n_{(2)} \tag{2.14}$$

where each extremal maximal sequence is counted as many times as it satisfies an "extremal" property; for example, a maximal sequence of zeroes beginning immediately after the first 1 and ending immediately before the last 1 would be counted twice.

For instance, returning to the example above, we have that $813_{(2)} = 1100101101_{(2)}$. Then $M_0(813) = 3$ and there is only one extremal sequence in $813_{(2)}$, the one ending in the last 1. Therefore $|\mathrm{QB}(813)| = 6$, in agreement with the explicit description of $\mathrm{QB}(813)$ given in that example.

As a consequence of Equation (2.14) we obtain the following lower and upper bounds for $|\mathrm{QB}(n)|$.

**Corollary 2.28.** *For every $n \geq 2$, $|\mathrm{QB}(n)| \geq \max\{2M_0(n) - 2, 1\}$.*

*Proof.* The number of extremal maximal sequences of zeroes in $n_{(2)}$ is at most three: one beginning immediately after the leading 1, one ending immediately before the last 1, and one ending in the units position. Therefore, $|\mathrm{QB}(n)| \geq 2M_0(n) - 2$. And since $\mathrm{QB}(n) \neq \emptyset$, $|\mathrm{QB}(n)| \geq 1$. □

**Corollary 2.29.** *For every $n \geq 2$, $|\mathrm{QB}(n)| = 1$ if, and only if, $n \in \{2^m - 1, 2^m, 2^m + 1\}$ for some $m \in \mathbb{N}_{\geq 1}$.*

*Proof.* By the previous corollary and Equation (2.14), $|\mathrm{QB}(n)| = 1$ if, and only if, $n_{(2)}$ contains no 0, which corresponds to the case $n = 2^m - 1$ for some $m \geq 1$, or it contains only one maximal sequence of zeroes and it is twice extremal, which corresponds to $n_{(2)}$ being either of the form $10\ldots0_{(2)}$, i.e., $n = 2^m$ for some $m \geq 1$, or of the form $10\ldots01_{(2)}$, i.e., $n = 2^m + 1$ for some $m \geq 1$. □

**Corollary 2.30.** *For every $n \geq 2$, $|\mathrm{QB}(n)| \leq \lfloor \log_2(n) \rfloor$.*

*Proof.* If $\lfloor \log_2(n) \rfloor \in 2\mathbb{N}$, then by the pigeonhole principle $M_0(n) \leq \lfloor \log_2(n) \rfloor / 2$. But if the binary representation of $n$ contains no extremal sequence, then it must start and end with 11, in which case $M_0(n) \leq \lfloor \log_2(n) \rfloor / 2 - 1$. In either case, $|\mathrm{QB}(n)| \leq \lfloor \log_2(n) \rfloor$.

On the other hand, if $\lfloor \log_2(n) \rfloor \notin 2\mathbb{N}$, then again by the pigeonhole principle $M_0(n) \leq (\lfloor \log_2(n) \rfloor + 1)/2$. However, if the equality is reached, then the binary representation of $n$ contains at least two extremal sequences. Indeed, if it begins or ends with 11, we have that $M_0(n) \leq (\lfloor \log_2(n) \rfloor - 1)/2$. Therefore, in order for $M_0(n)$ to attain that equality, it must begin with 10 and end with 01, 10 or 00, therefore having at least two extremal sequences. In either case, $|\mathrm{QB}(n)| \leq \lfloor \log_2(n) \rfloor$. □

### 2.2.3 Generating all minimal Colless trees

In this section we provide an algorithm that generates all minimal Colless trees with a given number $n$ of leaves. Its validity relies on the following easy, intuitive result.

**Lemma 2.31.** *Let $T = T_1 * T_2 \in \mathbf{BinTree}_n$, with $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$, where $n_1 \geq n_2 \geq 1$. The following three conditions are equivalent:*

*(i)* *$T$ is a minimal Colless tree.*

*(ii)* *$T_1$ and $T_2$ are minimal Colless trees and $(n_1, n_2) \in \mathrm{QB}(n)$.*

*(iii)* *$(\kappa_T(v_1), \kappa_T(v_2)) \in \mathrm{QB}(\kappa_T(v))$ for every $v \in \mathring{V}(T)$ whose children are $v_1, v_2$, so that $\kappa_T(v_1) \geq \kappa_T(v_2)$.*

*Proof.* The implication *(i) ⇒ (ii)* is easy to see, since by Corollary 2.3, if $T$ is minimal Colless, so are $T_1$ and $T_2$, and by definition of $\mathrm{QB}(n)$, $(n_1, n_2) \in \mathrm{QB}(n)$.

We proceed to the implication *(ii) ⇒ (iii)*. By Corollary 2.3, every rooted subtree of $T_1$ and $T_2$ will be minimal Colless. Suppose that $v \in \mathring{V}(T_1)$. Then, by definition, if $T_1$ is minimal Colless, so is the subtree rooted at $v$, and thus $(\kappa_T(v_1), \kappa_T(v_2)) \in \mathrm{QB}(\kappa_T(v))$ for any $v \in \mathring{V}(T_1)$. The case in which $v \in \mathring{V}(T_2)$ is proved analogously, and so it only remains to prove the case in which $v$ is the root, but this is exactly $(n_1, n_2) \in \mathrm{QB}(n)$.

The proof of *(iii) ⇒ (i)* is a bit more convoluted. We shall prove that, if $T$ satisfies $c(\kappa_T(v_1)) + c(\kappa_T(v_2)) + \kappa_T(v_1) - \kappa_T(v_2) = c(\kappa_T(v))$ for every $v \in \mathring{V}(T)$ whose children are $v_1, v_2$ so that $\kappa_T(v_1) \geq \kappa_T(v_2)$, then $C(T) = c(n)$. We will proceed by induction over $n$, the number of leaves. The cases when $n \in \{1, 2, 3\}$ are obvious since for these values of $n$ there is only one tree in $\mathbf{BinTree}_n$. Assume now that the implication holds up to $n - 1$ leaves, for $n \geq 4$. Let $T \in \mathbf{BinTree}_n$ be a tree such that, for every $v \in \mathring{V}(T)$,

$$c(\kappa_T(v_1)) + c(\kappa_T(v_2)) + \kappa_T(v_1) - \kappa_T(v_2) = c(\kappa_T(v)).$$

Let $x_1$ and $x_2$ be the children of the root $\rho$, with $\kappa_T(x_1) \geq \kappa_T(x_2)$. Now, for every $v \in \mathring{V}(T_{x_1})$, we have that

$$c(\kappa_T(v_1)) + c(\kappa_T(v_2)) + \kappa_T(v_1) - \kappa_T(v_2) = c(\kappa_T(v)),$$

which, by the induction hypothesis, implies that $C(T_{x_1}) = c(\kappa_T(x_1))$. By symmetry, we also have that $C(T_{x_2}) = c(\kappa_T(x_2))$. Finally,

$$\begin{aligned} c(n) = c(\kappa_T(\rho)) &= c(\kappa_T(x_1)) + c(\kappa_T(x_2)) + \kappa_T(x_1) - \kappa_T(x_2) \\ &= C(T_{x_1}) + C(T_{x_2}) + \kappa_T(x_1) - \kappa_T(x_2) = C(T), \end{aligned}$$

which is what we wanted to prove. □

A direct consequence of this lemma is that for almost every number $n$ of leaves there exist more than one minimal Colless tree with $n$ leaves.

**Corollary 2.32.** *For every $n \in \mathbb{N}_{\geq 1}$, there exists only one minimal Colless tree in* $\mathbf{BinTree}_n$ *if, and only if, $n \in \{2^m - 1, 2^m, 2^m + 1\}$ for some $m \geq 1$.*

*Proof.* If there exists only one minimal Colless tree in $\mathbf{BinTree}_n$, then in particular $n = 1$ or $|\underset{\smile}{QB}(n)| = 1$ and therefore, by Corollary 2.29, $n \in \{2^m - 1, 2^m, 2^m + 1\}$ for some $m \geq 1$. We prove the converse implication by induction on $m$. The base case $m = 1$ corresponds to $n \in \{1, 2, 3\}$, in which case the assertion is obvious because there exists only one tree in each $\mathbf{BinTree}_n$. Let now $m \geq 2$ and $n \in \{2^m - 1, 2^m, 2^m + 1\}$ and assume that, for every $m' < m$, if $n' \in \{2^{m'} - 1, 2^{m'}, 2^{m'} + 1\}$, then there exists only one minimal Colless tree in $\mathbf{BinTree}_{n'}$. If $n = 2^m$, by Corollary 2.4 there is only one minimal Colless tree in $\mathbf{BinTree}_n$, namely the fully symmetric tree. If $n = 2^m \pm 1$, let $T \in \mathbf{BinTree}_n$ be a minimal Colless tree with $n$ leaves, and understand it as $T = T_1 * T_2$, with $T_1 \in \mathbf{BinTree}_{n_1}$, $T_2 \in \mathbf{BinTree}_{n_2}$, and $n_1 \geq n_2$. If $n = 2^m - 1$, then, by the previous lemma and Corollary 2.29, $(n_1, n_2) \in \underset{\smile}{QB}(n) = \{(2^{m-1}, 2^{m-1} - 1)\}$ and $T_1, T_2$ are minimal Colless. But then, by the induction hypothesis, $T_1$ and $T_2$ are unique and therefore $T$ is also unique. And if $n = 2^m + 1$, then, by the previous lemma, $(n_1, n_2) \in \underset{\smile}{QB}(n) = \{(2^{m-1} + 1, 2^{m-1})\}$ and $T_1, T_2$ are minimal Colless, and then, again by the induction hypothesis, $T_1$ and $T_2$ are unique and therefore $T$ is also unique. This completes the proof of the inductive step. □

So, for all $n \in \mathbb{N}_{\geq 1}$ other than those in $\bigcup_{m \geq 1}\{2^m - 1, 2^m, 2^m + 1\}$, there is at least one minimal Colless tree in $\mathbf{BinTree}_n$ that is not maximally balanced.

Lemma 2.31, together with Corollary 2.4, prove the correctness of Algorithm 4 to produce minimal Colless trees in $\mathbf{BinTree}_n$. If the algorithm is run non-deterministically for all choices of a labeled leaf in line 3 and of a pair $(m_1, m_2) \in QB(m)$ in line 7 (using Corollary 2.27 to find all these pairs) in all executions of the **while** loop, one obtains all minimal Colless trees in $\mathbf{BinTree}_n$, possibly with repetitions that can be then removed (see the example below). The non-deterministic choice of the leaf in line 3 can be made deterministic by considering *ordered trees* (i.e., adding an orientation left-to-right to the pair of children of each internal node, with the number of descendant leaves decreasing from left to right) and then always choosing the left-most remaining labeled leaf, and at the end suppressing the orientations from the resulting trees.

**Example:**

Let us use this Algorithm MinColless to find all minimal Colless trees with 20 leaves; we describe the trees by means of the usual Newick format with the unlabeled leaves represented by a symbol · and omitting the semicolon ending mark in order not to confuse it with a punctuation mark.

1) We start with a single node labeled 20.

2) Since $\underset{\smile}{QB}(20) = \{(10, 10), (12, 8)\}$, this node can split into the cherries $(10, 10)$ and $(12, 8)$.

3.1) Since $\underset{\smile}{QB}(10) = \{(5, 5), (6, 4)\}$, the different ways of splitting the leaves of

**Algorithm 4:** MinColless

```
Input  : $n \in \mathbb{N}$
Output: $T \in \mathbf{BinTree}_n$ with minimum Colless index
start with a single node labeled $n$;
while the current tree contains labeled leaves do
    choose a leaf with label $m$;
    if $m$ is a power of 2 then
        replace this leaf by a fully symmetric tree with $m$ unlabeled leaves;
    else
        find a pair of integers $(m_1, m_2) \in \underset{\smile}{\mathrm{QB}}(m)$;
        split the leaf labeled $m$ into a cherry with unlabeled root and its leaves
          labeled $m_1$ and $m_2$, respectively;
    end
end
return current tree;
```

the tree $(10, 10)$ produce the trees $((5, 5), (5, 5))$, $((5, 5), (6, 4))$, and $((6, 4), (6, 4))$. Now, since $\underset{\smile}{\mathrm{QB}}(5) = \{(3, 2)\}$, $\underset{\smile}{\mathrm{QB}}(6) = \{(3, 3), (4, 2)\}$, and $\underset{\smile}{\mathrm{QB}}(3) = \{(2, 1)\}$, and 1, 2, and 4 are powers of 2, we have the following derivations from these trees through all possible combinations of splitting the leaves in the trees:

$$\begin{aligned}
&((5, 5), (5, 5)) \Rightarrow (((3, 2), (3, 2)), ((3, 2), (3, 2)))\\
&\quad \Rightarrow ((((2, 1), 2), ((2, 1), 2)), (((2, 1), 2), ((2, 1), 2)))\\
&\quad \Rightarrow (((((\cdot, \cdot), \cdot), (\cdot, \cdot)), (((\cdot, \cdot), \cdot), (\cdot, \cdot))), ((((\cdot, \cdot), \cdot), (\cdot, \cdot)), (((\cdot, \cdot), \cdot), (\cdot, \cdot))))\\
&((5, 5), (6, 4)) \Rightarrow (((3, 2), (3, 2)), ((3, 3), 4))\\
&\quad \Rightarrow ((((2, 1), 2), ((2, 1), 2)), (((2, 1), (2, 1)), 4))\\
&\quad \Rightarrow (((((\cdot, \cdot), \cdot), (\cdot, \cdot)), (((\cdot, \cdot), \cdot), (\cdot, \cdot))), ((((\cdot, \cdot), \cdot), ((\cdot, \cdot), \cdot)), ((\cdot, \cdot), (\cdot, \cdot)))\\
&((5, 5), (6, 4)) \Rightarrow (((3, 2), (3, 2)), ((4, 2), 4))\\
&\quad \Rightarrow ((((2, 1), 2), ((2, 1), 2)), ((4, 2), 4))\\
&\quad \Rightarrow (((((\cdot, \cdot), \cdot), (\cdot, \cdot)), (((\cdot, \cdot), \cdot), (\cdot, \cdot))), ((((\cdot, \cdot), (\cdot, \cdot)), (\cdot, \cdot)), ((\cdot, \cdot), (\cdot, \cdot))))
\end{aligned}$$

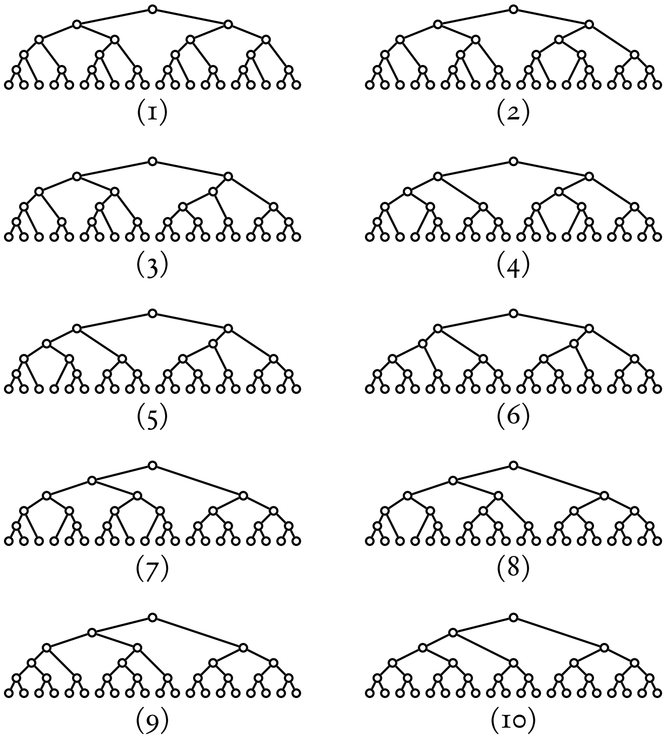
$$\begin{aligned}
&((6, 4), (6, 4)) \Rightarrow (((3, 3), 4), ((3, 3), 4))\\
&\quad \Rightarrow ((((2, 1), (2, 1)), 4), (((2, 1), (2, 1)), 4))\\
&\quad \Rightarrow (((((\cdot, \cdot), \cdot), ((\cdot, \cdot), \cdot)), ((\cdot, \cdot), (\cdot, \cdot))), ((((\cdot, \cdot), \cdot), ((\cdot, \cdot), \cdot)), ((\cdot, \cdot), (\cdot, \cdot))))\\
&((6, 4), (6, 4)) \Rightarrow (((3, 3), 4), ((4, 2), 4))\\
&\quad \Rightarrow ((((2, 1), (2, 1)), 4), ((4, 2), 4))\\
&\quad \Rightarrow (((((\cdot, \cdot), \cdot), ((\cdot, \cdot), \cdot)), ((\cdot, \cdot), (\cdot, \cdot))), ((((\cdot, \cdot), (\cdot, \cdot)), (\cdot, \cdot)), ((\cdot, \cdot), (\cdot, \cdot))))\\
&((6, 4), (6, 4)) \Rightarrow (((4, 2), 4), ((4, 2), 4))\\
&\quad \Rightarrow (((((\cdot, \cdot), (\cdot, \cdot)), (\cdot, \cdot)), ((\cdot, \cdot), (\cdot, \cdot))), ((((\cdot, \cdot), (\cdot, \cdot)), (\cdot, \cdot)), ((\cdot, \cdot), (\cdot, \cdot))))
\end{aligned}$$

3.2) Since $\underset{\smile}{\mathrm{QB}}(12) = \{(6, 6), (8, 4)\}$ and 8 is a power of 2, the tree $(12, 8)$ gives rise to the trees $((6, 6), 8)$ and $((8, 4), 8)$, and then, using $\underset{\smile}{\mathrm{QB}}(6) = \{(3, 3), (4, 2)\}$ and

$$QB(3) = \{(2,1)\},$$

$$\begin{aligned}
&((6,6),8) \Rightarrow (((3,3),(3,3)),8) \Rightarrow ((((2,1),(2,1)),((2,1),(2,1))),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),(((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,6),8) \Rightarrow (((3,3),(4,2)),8) \Rightarrow ((((2,1),(2,1)),(4,2)),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),\cdot),((\cdot,\cdot),\cdot)),(((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((6,6),8) \Rightarrow (((4,2),(4,2)),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot)),(((\cdot,\cdot),(\cdot,\cdot)),(\cdot,\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))\\
&((8,4),8)\\
&\quad \Rightarrow (((((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))),((\cdot,\cdot),(\cdot,\cdot))),(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))
\end{aligned}$$

So, there are 10 different minimal Colless tree shapes with 20 leaves. We depict them in Figure 2.3.



(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

Figure 2.3: The ten trees in **BinTree**$_{20}$ with minimum Colless index, 8. They are enumerated in the same order as they have been produced in the example given in this section.

We have implemented Algorithm MinColless, with the step in line 7 efficiently carried out by means of Corollary 2.27, in a Python script that generates, for every $n$, the Newick description of all minimal Colless trees in **BinTree**$_n$. It is available at the GitHub repository `https://github.com/biocom-uib/Colless`. As a proof of concept, we have computed for every $n$ from 1 to 128 all such minimal Colless trees in **BinTree**$_n$. Figure 2.4 shows their number $\widetilde{c}(n)$ for every $n$. These numbers are in agreement with those provided by the following recurrence for the sequence $\widetilde{c}(n)$, which was established in Proposition 4 of our paper [22]: Starting with $\widetilde{c}(1) = 1$,

$$\tilde{c}(n) = \sum_{(n_1,n_2)\in \underset{\smile}{\mathrm{QB}}(n)} \tilde{c}(n_1) \cdot \tilde{c}(n_2) + \binom{\tilde{c}(n/2)+1}{2} \cdot \chi_{2\mathbb{N}}(n) \tag{2.15}$$

for $n \geq 2$, where $\chi_{2\mathbb{N}}$ is the characteristic function of the set $2\mathbb{N}$.

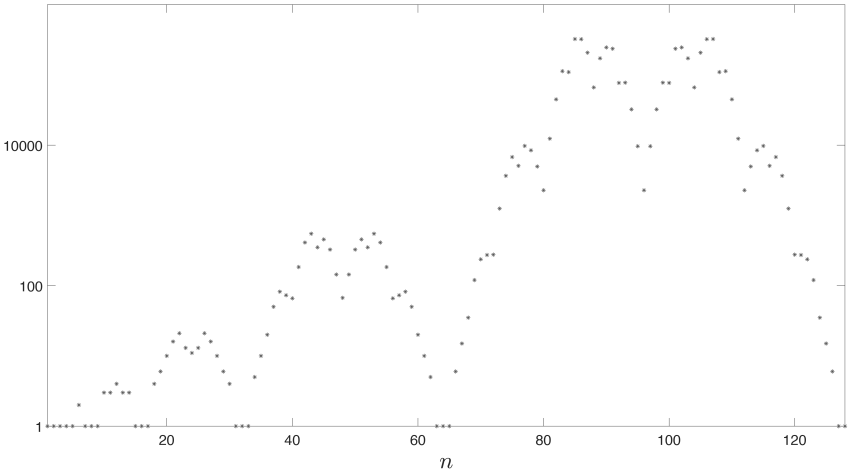


Figure 2.4: Semi-logarithmic plot of the sequence of the number of minimal Colless trees in $\mathbf{BinTree}_n$, for $n \in \{1,\dots,128\}$.

The plot of this sequence $\widetilde{c}(n)$ shows a fractal structure that reminds us of the Takagi curve. But, currently, we do not know of either any closed expression for the computation of $\widetilde{c}(n)$, or any argument supporting its apparent connection with the Takagi curve, should this connection actually exist.

To close this section, we want to point out that in [22] we gave a family of minimal Colless trees, called the *greedy from the bottom* (*GFB*) trees, $T_n^{\mathrm{gfb}} \in \mathbf{BinTree}_n$, such that $T_n^{\mathrm{gfb}} \neq T_n^{\mathrm{bal}}$ for every $n$ such that $\widetilde{c}(n) > 1$. Their name comes from the possibility of recursively building them through a process of root joining trees that we shall not recall here (see Algorithm 2 in [22]). Instead, we want to mention the following alternative characterization of these trees, which is obtained combining Propositions 5 to 7 of [22]: A tree $T = T_1 * T_2 \in \mathbf{BinTree}_n$, with $T_1 \in \mathbf{BinTree}_{n_1}$, $T_2 \in \mathbf{BinTree}_{n_2}$, and $n_1 \geq n_2$, is GFB if, and only if, $(n_1, n_2)$ is the pair in $\underset{\smile}{\mathrm{QB}}(n)$ attaining the maximum difference $n_1 - n_2$ and $T_1, T_2$ are GFB trees.

It can be deduced from the explicit description of all pairs in $\underset{\smile}{\mathrm{QB}}(n)$ given in Corollary 2.27 that the pair $(n_1, n_2)$ in $\underset{\smile}{\mathrm{QB}}(n)$ that attains the maximum difference $n_1 - n_2$ is the following one (see, again, [22, Prop. 7]): If $n = 2^m + s$, with $m = \lfloor \log_2(n) \rfloor$ and $0 \leq s < 2^m$, then

(i) If $0 \leq s \leq 2^{m-1}$, then $n_1 = 2^{m-1} + s$ and $n_2 = 2^{m-1}$.

(ii) If $2^{m-1} \leq s < 2^m$, then $n_1 = 2^m$ and $n_2 = s$.

Notice moreover that, since GFB trees are minimal Colless, their rooted subtrees with a power of 2 number of leaves, being also minimal Colless, must be fully symmetric. Therefore, for every internal node in a GFB tree, one of its two children is the root of a fully symmetric subtree. This is the basis of the following result, which says that the GFB trees are the most symmetrical minimal Colless trees; for a proof, see [22, Appendix A.3].

**Proposition 2.33.** *For every $n \geq 1$, let $n = \sum_{i=0}^{\ell} 2^{m_i}$, with $\ell \geq 0$ and $m_\ell > \cdots > m_0$, be its binary decomposition.*

*(i) The number of symmetry nodes in $T_n^{\mathrm{gfb}}$ is $s(T_n^{\mathrm{gfb}}) = n - 1 - (m_\ell - m_0)$.*

*(ii) For every minimal Colless tree $T \in \mathbf{BinTree}_n$, if $T \neq T_n^{\mathrm{gfb}}$, then $s(T) < s(T_n^{\mathrm{gfb}})$.*

To end this section, let us recall another result from [22], namely Proposition 9 therein.

**Proposition 2.34.** *For every $n \geq 1$, if $T \in \mathbf{BinTree}_n$ is a minimal Colless index, then it has also the minimum Sackin index in $\mathbf{BinTree}_n$.*

In other words, of all the trees with minimum Sackin index, *some* of them are those that also have minimum Colless index. But indeed not all of them, as we can see in Figure 2.5. Thus, we can consider that, in this regard, Colless index is finer than Sackin's is. A consequence of the last result is that, since the trees $T$ attaining the minimum Sackin index satisfy the property that for any two leaves $(u, v) \in L(T)^2$, $|\delta(u) - \delta(v)| \leq 1$ (Theorem 1.19), so do the minimal Colless trees.

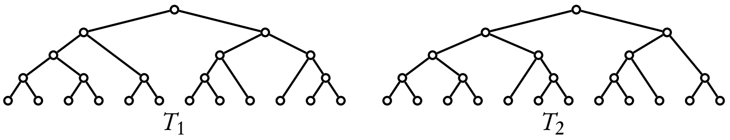


Figure 2.5: Trees $T_1$ and $T_2$ with 12 leaves. Their Sackin indices are $S(T_1) = S(T_2) = 44$, which can be shown to be minimal (cf. Theorem 3 in [39]). However, $C(T_1) = 4 = c(12)$ and $C(T_2) = 6$. Thus, $T_1$ is minimal Colless, while $T_2$ is not.

## 2.3 Discussion

In this chapter, we have focused on the study of the minimum value $c(n)$ of the Colless index for each number of leaves $n \geq 1$, and the *minimal Colless trees* that attain it. This study had not been completely pursued until the independent preprints by Herbst-Fischer-Wicke [40] and Coronado-Rosselló [27] that gave rise to our joint paper [22]. This chapter has followed mainly our preprint [27], although some proofs are different from those published therein.

In this chaper we have shown that, as expected, the maximally balanced bifurcating trees are most balanced according to the Colless index. But if $n$ differs at least 2 from any power of 2, then there are minimal Colless trees with $n$ leaves that are not maximally balanced. In other words, the least global amount of imbalance, which is, at the end of the day, what the Colless index measures, is almost always achieved *also* at trees that do not minimize the local imbalance at each internal node. Then, we have provided a structural characterization of the minimal Colless trees from which we have derived an algorithm to produce all of them for any number $n$ of leaves. It remains an open problem to find a closed formula that, for any $n$, gives their number $\widetilde{c}(n)$, or the number of minimal Colless *phylogenetic* trees with $n$ leaves: a recurrence for $\widetilde{c}(n)$ was given in [22], and we have recalled it in Equation 2.15 above.

Having proved that the maximally balanced trees attain the minimum Colless index, we have been able to find a closed formula for this minimum value $c(n)$ on $\mathbf{BinTree}_n$. Knowing this minimum value, as well as its maximum value, which is reached at the caterpillars and is equal to $\binom{n-1}{2}$ [86], allows one to normalize the Colless index so that its range becomes the unit interval $[0, 1]$, by means of the usual affine transformation:

$$\overline{C}(T) = \frac{C(T) - \min\{C(T') : T' \in \mathbf{BinTree}_n\}}{\max\{C(T') : T' \in \mathbf{BinTree}_n\} - \min\{C(T') : T' \in \mathbf{BinTree}_n\}} = \frac{C(T) - c(n)}{\binom{n-1}{2} - c(n)}$$

thus allowing a sound comparison of the balance of two trees with a different number of leaves. This transformation has the good property of attaining both 0 and 1 when the minimum and the maximum are reached, respectively.

Our formula for $c(n)$ explains the fractal structure of the graph of this sequence related to the fractal Takagi curve (cf. Figure 2.1). We have made explicit this connection with the Takagi function using Tambs-Lyche reformulation of the latter. It turns out that a similar fractal structure seems to appear also in the graph of $(n, \widetilde{c}(n))$ (cf. Figure 2.4), but it is also an open problem to find a reason for it.

Chapter 3

# The Quadratic Colless index

> Twoo multiplications doe make a Cubike nomber. Likewaies .3. multiplications doe giue a square of squares.
>
> R. Recorde, *Whetstone of Witte*, 1557

The Colless index, albeit being widely popular and intuitive, presents some drawbacks. One of them, as we have seen in the previous chapter, is that its minimum value is not reached by a single tree for almost any number of leaves $n$, and the characterization of the set of trees that attain this value is convoluted and, once elucidated, presents no intuitive idea of balance itself. Nevertheless, by Proposition 2.34, every tree that attains the minimum Colless index reaches the minimum Sackin index, too. This gives us the idea that the Colless index is, in this regard, a finer measure of imbalance than the Sackin index is.

Following this lead, our errand in this chapter shall be to present a new balance index that will be finer (in this sense) than the Colless index is. Incidentally, this index shall also have a much wider range of values, and thus, *a priori*, a smaller proportion of draws. This shall be by no means the first such index (see, for example, the Cophenetic index [85] and our Quartet index presented in Chapter 5), but these sacrify, in their definition, what is arguably one of the most interesting advantages of the Colless index: its intuitiveness as a measure of imbalance.

Let $T \in \mathbf{BinTree}_n$ be a bifurcating tree with $n$ leaves. The *Quadratic Colless index* of $T$, $C^{(2)}(T)$, is defined as

$$C^{(2)}(T) = \sum_{u\in \mathring{V}(T)} \mathrm{bal}(u)^2;$$

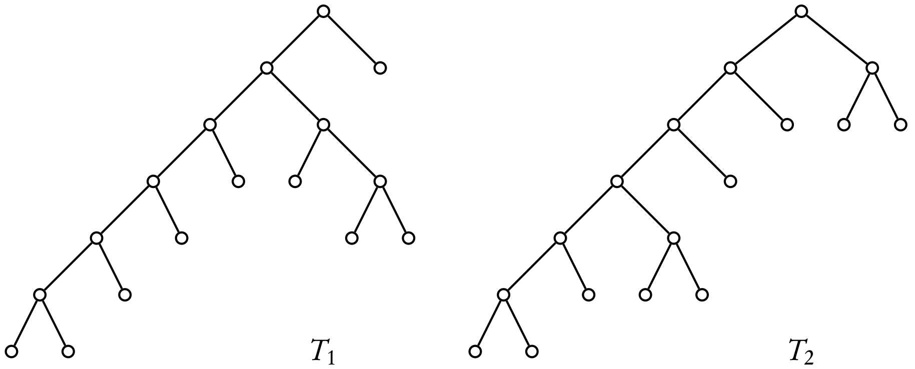
i.e., as the sum, over all internal nodes in $T$, of the squared difference of the number of leaves of the subtrees rooted at their children. It is now straightfowrard to check that it satisfies an analogous relation to that displayed in Lemma 2.1: for any $T = T_1 * T_2 \in$ **BinTree**$_n$, with $T_1 \in$ **BinTree**$_{n_1}$ and $T_2 \in$ **BinTree**$_{n_2}$,

$$C^{(2)}(T) = C^{(2)}(T_1) + C^{(2)}(T_2) + (n_1 - n_2)^2, \tag{3.1}$$

and so is a binary recursive shape index in the sense introduced in the Preliminaries. This definition is not very different from that of the Colless index —substituting the absolute value by the square—, and hence still reflects the intuition behind the latter: to compute a measure of the overall balance of a given tree. However, this definition presents a much better behaviour, as we shall see in this chapter.

**Example:**

Consider the following two trees in **BinTree**$_9$.



It is easy to see that

$$C(T_1) = 7 + 2 + 1 + 3 + 2 + 1 = 16 \quad \text{and} \quad C(T_2) = 6 + 5 + 4 + 1 + 1 = 17$$

while

$$C^{(2)}(T_1) = 7^2 + 2^2 + 1^2 + 3^2 + 2^2 + 1^2 = 68$$
$$C^{(2)}(T_2) = 6^2 + 5^2 + 4^2 + 1^2 + 1^2 = 68$$

This shows that $C^{(2)}(T_1) = C^{(2)}(T_2)$ does not imply $C(T_1) = C(T_2)$. The other implication is not true either: $C(T_n^{\mathrm{bal}}) = C(T_n^{\mathrm{gfb}})$ for every $n \geq 1$, whereas, by Lemma 3.1 below, $C^{(2)}(T_n^{\mathrm{bal}}) < C^{(2)}(T_n^{\mathrm{gfb}})$ for any $n \notin \bigcup_{m\in\mathbb{N}}\{2^m - 1, 2^m, 2^m + 1\}$.

This chapter is organized as follows: in the first section, we will find the extreme values of this new index, as well as the trees attaining them, while drawing a comparison between both the Colless and the Quadratic Colless index. Afterwards, in the second section, we shall find both the expected value and the variance of $C^{(2)}$ under the Yule and Uniform probabilistic models — and recall, here, that this expected value is not known for the Colless index under the Uniform model. The third section presents a series of numerical results on the probability of two different trees having the same $C^{(2)}$ value. We end by discussing this new measure, while pressenting and discussing the possibility on some natural extensions to it.

## 3.1 Extreme values and the trees attaining them

The complexity of the characterization of the minimal Colless trees shall contrast sharply with the readiness with which such question shall be answered in regard to this squared version. Indeed, we shall first present the following lemma, of which that characterization will be but a straightforward corollary.

**Lemma 3.1.** *Let $T \in \mathbf{BinTree}_n$ be a bifurcating tree with $n \geq 1$ leaves. Then,*

$$C^{(2)}(T) \geq C(T)$$

*and the equality is reached if, and only if, $T$ is maximally balanced.*

*Proof.* By definition,

$$C^{(2)}(T) = \sum_{u \in \mathring{V}(T)} \mathrm{bal}(u)^2 \geq \sum_{u \in \mathring{V}(T)} \mathrm{bal}(u) = C(T)$$

since $\mathrm{bal}(u) \in \mathbb{N}$ for all $u \in \mathring{V}(T)$. The equality will be attained if, and only if, for each $u \in \mathring{V}(T)$, $\mathrm{bal}(u)^2 = \mathrm{bal}(u)$, which will happen if and only if $\mathrm{bal}(u) \in \{0, 1\}$. By definition, that only happens in the maximally balanced trees. □

**Theorem 3.2.** *The minimum of the Quadratic Colless index is reached exactly at the maximally balanced trees. Furthermore, this minimum value for $n \geq 1$ leaves is $C^{(2)}(T_n^{\mathrm{bal}}) = c(n)$.*

*Proof.* Let $n \in \mathbb{N}$ and $T \in \mathbf{BinTree}_n$. Then, we know (by Theorem 2.7) that $C(T) \geq C(T_n^{\mathrm{bal}})$, hence

$$C^{(2)}(T) \geq C(T) \geq C(T_n^{\mathrm{bal}}) = C^{(2)}(T_n^{\mathrm{bal}})$$

and therefore $T_n^{\mathrm{bal}}$ presents the minimum $C^{(2)}$ value. Furthermore, the first inequality is strict whenever $T \neq T_n^{\mathrm{bal}}$, and thus $T_n^{\mathrm{bal}}$ is the only tree in $\mathbf{BinTree}_n$ that attains the minimum Quadratic Colless index. □

*Voilà!* In order to give this proof, we only needed to know the fact that the maximally balanced trees are indeed Colless minimal, which was proved in Lemma 2.5 and Theorem 2.7 in the previous chapter. Therefore, the minimum value of the Quadratic Colless index is reached at a unique family of trees, namely the maximally balanced trees, and furthermore its value is exactly that of the Colless index for that family; namely, the one computed in Theorems 2.10, 2.11, and 2.12. But, most importantly, we have established that this minimum value is only attained at the maximally balanced trees, a desirable property that the original Colless index lacked.

We are now concerned with the problem of characterising which trees attain the maximum Quadratic Colless value: as intuition tells us, they should be exactly the caterpillars. This is, of course, the case. Indeed: notice that, for any $T_n^{\mathrm{cat}}$, by Equation (3.1)

$$C^{(2)}(T_n^{\mathrm{cat}}) = C^{(2)}(T_{n-1}^{\mathrm{cat}}) + (n-2)^2 = \sum_{i=1}^{n-2} i^2 = \frac{(n-2)(n-3)(2n-3)}{6} = \binom{n}{3} + \binom{n-1}{3}.$$

Now, we can prove the next theorem.

**Theorem 3.3.** *The maximum of the Quadratic Colless index is reached exactly at the caterpillars. Furthermore, this maximum value for $n \geq 1$ leaves is*

$$C^{(2)}(T_n^{\mathrm{cat}}) = \binom{n}{3} + \binom{n-1}{3}.$$

*Proof.* We proceed by induction on the number of leaves, $n$. For $n \in \{1,2,3\}$ it is obviously true. Suppose now that $n \geq 4$, and that the property holds up to $n-1$ leaves. We want to prove that if $T_1 * T_2 \in \mathbf{BinTree}_n \setminus \{T_n^{\mathrm{cat}}\}$, with $T_1 \in \mathbf{BinTree}_{n_1}$, $T_2 \in \mathbf{BinTree}_{n_2}$ and $n_1, n_2 \in \mathbb{N}$ such that $n = n_1 + n_2$, and $n_1 \geq n_2 \geq 1$,

$$C^{(2)}(T_n^{\mathrm{cat}}) > C^{(2)}(T_1) + C^{(2)}(T_2) + (n_1 - n_2)^2 = C^{(2)}(T_1 * T_2).$$

Since, by the the induction hypothesis $C^{(2)}(T_1) \leq C^{(2)}(T_{n_1}^{\mathrm{cat}})$ and $C^{(2)}(T_2) \leq (T_{n_2}^{\mathrm{cat}})$ and these inequalities are strict unless $T_1 = T_{n_1}^{\mathrm{cat}}$ and $T_2 = T_{n_2}^{\mathrm{cat}}$, it will suffice to show that

$$\begin{aligned} C^{(2)}(T_n^{\mathrm{cat}}) \geq C^{(2)}(T_{n_1}^{\mathrm{cat}}) + C^{(2)}(T_{n_2}^{\mathrm{cat}}) + (n_1 - n_2)^2 &= C^{(2)}(T_{n_1}^{\mathrm{cat}}) + C^{(2)}(T_{n-n_1}^{\mathrm{cat}}) + (2n_1 - n)^2 \\ &= \frac{(n_1-2)(n_1-3)(2n_1-3)}{6} + \frac{(n-n_1-2)(n-n_1-3)(2n-2n_1-3)}{6} + (2n_1-n)^2 \\ &= \frac{(6n-2)n_1^2 - (6n^2-2n)n_1 + 2n^3 - 7n^2 + 27n - 36}{6} \end{aligned}$$

and that the equality only holds when $n_1 = n-1$.

Consider now the function $C_{\mathrm{cat}}^{(2)} : \mathbb{R} \to \mathbb{R}$, defined as

$$C_{\mathrm{cat}}^{(2)}(x) = \frac{(6n-2)x^2 - (6n^2-2n)x + 2n^3 - 7n^2 + 27n - 36}{6}.$$

The curve $y = C_{\mathrm{cat}}^{(2)}(x)$ is a parabola whose leading coefficient is positive, and hence it is concave upward. Its vertex, where the minimum value of $C_{\mathrm{cat}}^{(2)}$ is attained, has first coordinate $x = n/2$. Therefore, the maximum of $C_{\mathrm{cat}}^{(2)}$ in the closed interval $[n/2, 1]$ is reached exactly at the other end of the interval, that is, at $x = n-1$. Since $n_1 \in [n/2, 1]$, this concludes the proof of the result. □

We have thus proven that this new balance index, $C^{(2)}$, has better properties than the Colless index when it comes to its extreme values and those trees that attain them.

To close this section, let us study the range of values that the Quadratic Colless index can attain. As we have seen, for any number of leaves $n \in \mathbb{N}$ we can establish a tight upper bound for $C^{(2)}(T)$, $T \in \mathbf{BinTree}_n$, as

$$C^{(2)}(T) \leq C^{(2)}(T_n^{\mathrm{cat}}) = \binom{n}{3} + \binom{n-1}{3} \sim O(n^3).$$

We also have a tight lower bound for $C^{(2)}(T)$: exactly $c(n)$. Nevertheless, by themselves none of the formulæ given in the previous chapter to compute $c(n)$ are very informative to this regard, leaving aside the obvious fact that $c(2^m) = 0$ for any $m \in \mathbb{N}$. Now, by Corollary 2.13, the lower bound is such that

$$c(n) < \min\{n/2, 2^{\lceil \log_2(n) \rceil}/3\} \sim O(n),$$

and there are values of $n$ for which this upper bound is sharp: for instance, when $n = 11 = 2^3 + 3$, $c_n = 5 = (11-1)/2 = \lfloor 2^4/3 \rfloor$.

So, the range of $C^{(2)}$ grows in $O(n^3)$. In order to give some perspective, this is one order of magnitude higher than the range of values of the Colless and Sackin indices are [22, 39], and it shares this order of magnitude with the Cophenetic index [85], although it is always wider than this last one: cf. Table 3.1. Our Quartet index will have an even wider range of values.

| Index | Minimum | Maximum |
|---|---|---|
| $C$ | $O(n)$ | $\binom{n-1}{2}$ |
| $S$ | $O(n\log(n))$ | $\binom{n+1}{2} - 1$ |
| $\Phi$ | $O(n^2)$ | $\binom{n}{3}$ |
| $C^{(2)}$ | $O(n)$ | $\binom{n}{3} + \binom{n-1}{3}$ |

Table 3.1: Range of values of the Colless index $C$, the Sackin index $S$, the Cophenetic index $\Phi$ and the Quadratic Colless index $C^{(2)}$

## 3.2 The expected value and variance under the Uniform and Yule models

The Quadratic Colless index $C^{(2)}$ can be extended to bifurcating phylogenetic trees in the usual manner: for any $(T, \lambda) \in \mathbf{BinPhyloTree}_n$, $C^{(2)}(T, \lambda) = C^{(2)}(T)$. Given some probabilistic model $P_n$ of bifurcating phylogenetic trees, let $C_n^{(2)}$ be the random variable that chooses a phylogenetic tree $(T, \lambda) \in \mathbf{BinPhyloTree}_n$ with probability $P_n(T, \lambda)$ and computes $C^{(2)}(T)$. In this section we shall compute the expected value and the variance of $C_n^{(2)}$ under the Uniform and Yule models. The computations are a bit convoluted, and shall occupy the remaining of this section. To put the results of this section into perspective, let us recall that no closed formula for the expected value or the variance of the Colless index under the uniform model has been published yet, only their limit behaviour being known so far, and that a closed formula for the variance of the Colless index under the Yule models was not published until very recently [13].

### 3.2.1 The Uniform model

The main goal of this section is to prove the following result.

**Theorem 3.4.** *Let $n \in \mathbb{N}_{\geq 1}$.*

*(i) The expected value of $C_n^{(2)}$ under the Uniform model is*

$$E_{\mathrm{unif}}(C_n^{(2)}) = \binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!} - n(2n-1).$$

*(ii) The variance of $C_n^{(2)}$ under the Uniform model is*

$$\sigma^2_{\mathrm{unif}}(C_n^{(2)}) = \frac{2}{15}(2n-1)(7n^2+9n-1)\binom{n+1}{2} - \frac{1}{8}(5n^2+n+2)\binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!} - \binom{n+1}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.$$

In order to do this, we shall use Lemma 1.31 which, applied to the Uniform model, says that, for any binary recursive shape index $I$,

$$E_{\mathrm{unif}}(I_n) = \sum_{k=1}^{n-1} C_{k,n-k}\left(2E_{\mathrm{unif}}(I_k) + f_I(k,n-k)\right) \tag{3.2}$$

$$E_{\mathrm{unif}}(I_n^2) = \sum_{k=1}^{n-1} C_{k,n-k}\Big(2E_{\mathrm{unif}}(I_k^2) + 4f_I(k,n-k)E_{\mathrm{unif}}(I_k) + 2E_{\mathrm{unif}}(I_k)E_{\mathrm{unif}}(I_{n-k}) + f_I(k,n-k)^2\Big) \tag{3.3}$$

where

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

Since the proof of the theorem will occupy the whole of this section, we will split it into two lemmata, each one for each point in its statement. In their proof, we will invoke Lemmata 1.33 and 1.34, as well as their corollary, Theorem 1.35.

**Lemma 3.5.** *For every $n \in \mathbb{N}_{\geq 1}$,*

$$E_{\mathrm{unif}}(C_n^{(2)}) = \binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!} - n(2n-1).$$

*Proof.* As we have already pointed out, since $C^{(2)}$ is a binary recursive shape index with $f_{C^{(2)}}(k,n-k) = (n-2k)^2$, by Equation (3.2)

$$\begin{aligned}
E_{\mathrm{unif}}(C_n^{(2)}) &= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}(C_k^{(2)}) + \sum_{k=1}^{n-1} C_{k,n-k}(n-2k)^2\\
&= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}(C_k^{(2)}) + n^2\sum_{k=1}^{n-1} C_{k,n-k} - 4(n-1)\sum_{k=1}^{n-1} C_{k,n-k}k + 8\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\\
&= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}(C_k^{(2)}) + n^2 - 2n(n-1) + 4\binom{n}{2}\left(1 - \frac{1}{2(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}(C_k^{(2)}) + 2\binom{n}{2} + n - n\cdot\frac{(2n-2)!!}{(2n-3)!!},
\end{aligned}$$

where we have used, in the second last equality, Lemma 1.33. Thus, by Theorem 1.35, and since $E_{\mathrm{unif}}(C_1^{(2)}) = 0$,

$$E_{\mathrm{unif}}(C_n^{(2)}) = \left(\binom{n}{2} + n\right)\frac{(2n-2)!!}{(2n-3)!!} - \left(4\binom{n}{2} + n\right) = \binom{n+1}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} - n(2n-1),$$

as we claimed. □

**Lemma 3.6.** *For every $n \in \mathbb{N}_{\geq 1}$,*

$$\begin{aligned}\sigma^2_{\mathrm{unif}}(C_n^{(2)}) = &\frac{2}{15}(2n-1)(7n^2+9n-1)\binom{n+1}{2} \\ &- \frac{1}{8}(5n^2+n+2)\binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!} - \binom{n+1}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.\end{aligned}$$

*Proof.* In order to simplify the notations, throughout this proof by $\alpha_n$ we shall denote $(2n-2)!!/(2n-3)!!$. To compute the variance $\sigma^2_{\mathrm{unif}}(C_n^{(2)})$, we will proceed by means of the identity

$$\sigma^2_{\mathrm{unif}}(C_n^{(2)}) = E_{\mathrm{unif}}((C_n^{(2)})^2) - E_{\mathrm{unif}}(C_n^{(2)})^2 \tag{3.4}$$

where the value of $E_{\mathrm{unif}}(C_n^{(2)})$ is given in Lemma 3.5. Therefore, we just need to compute $E_{\mathrm{unif}}((C_n^{(2)})^2)$.

Now, applying Equation (3.3) to $C^{(2)}$, we obtain

$$\begin{aligned}E_{\mathrm{unif}}((C_n^{(2)})^2) &= \sum_{k=1}^{n-1} C_{k,n-k}\Big(2E_{\mathrm{unif}}((C_k^{(2)})^2) + (n-2k)^4 \\ &\qquad +4(n-2k)^2 E_{\mathrm{unif}}(C_k^{(2)}) + 2E_{\mathrm{unif}}(C_k^{(2)})E_{\mathrm{unif}}(C_{n-k}^{(2)})\Big) \\ &= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}((C_k^{(2)})^2) \\ &\quad + \sum_{k=1}^{n-1} C_{k,n-k}\left[(n-2k)^4 + 4(n-2k)^2\left(\binom{k+1}{2}\alpha_k - k(2k-1)\right)\right. \\ &\qquad + 2\left(\binom{k+1}{2}\alpha_k - k(2k-1)\right) \\ &\qquad \left.\cdot\left(\binom{n-k+1}{2}\alpha_{n-k} - (n-k)(2(n-k)-1)\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\mathrm{unif}}((C_k^{(2)})^2) \\
&\quad + \sum_{k=1}^{n-1} C_{k,n-k}\Big((n-2k)^4 - 4(n-2k)^2k(2k-1) \\
&\qquad\qquad + 2k(2k-1)(n-k)(2(n-k)-1)\Big) \\
&\quad + \sum_{k=1}^{n-1} C_{k,n-k}\left[4(n-2k)^2\binom{k+1}{2}\alpha_k - 2\binom{n-k+1}{2}k(2k-1)\alpha_{n-k}\right. \\
&\qquad\qquad \left. -2\binom{k+1}{2}(n-k)(2(n-k)-1)\alpha_k\right] \\
&\quad + 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\alpha_k\alpha_{n-k} \\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\mathrm{unif}}((C_k^{(2)})^2) \\
&\quad - \sum_{k=1}^{n-1} C_{k,n-k}\Big(8k^4 + 16(n-1)k^3 - 2(12n^2-6n-1)k^2 - (2n-8n^3)k - n^4\Big) \\
&\quad + \sum_{k=1}^{n-1} C_{k,n-k}\left[4(n-2k)^2\binom{k+1}{2} - 4\binom{k+1}{2}(n-k)(2(n-k)-1)\right]\alpha_k \\
&\quad + 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\alpha_k\alpha_{n-k} \\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\mathrm{unif}}((C_k^{(2)})^2) \\
&\quad - \sum_{k=1}^{n-1} C_{k,n-k}\left[192\binom{k}{4} + 96(n+2)\binom{k}{3} - 4(12n^2-30n-5)\binom{k}{2}\right. \\
&\qquad\qquad \left. + (8n^3-24n^2+26n-6)k - n^4\right] \\
&\quad + \sum_{k=1}^{n-1} C_{k,n-k}\left[96\binom{k}{4} + 156\binom{k}{3} - 4(n^2-n-16)\binom{k}{2}\right. \\
&\qquad\qquad \left. - 4(n^2-n-1)k\right]\alpha_k \\
&\quad + 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k+1}{2}\binom{n-k+1}{2}\alpha_k\alpha_{n-k}.
\end{aligned}
\tag{3.5}
$$

Let us now compute the independent term in this equation. Its first sum can be computed using Lemma 1.33:

$$\begin{aligned}
&\sum_{k=1}^{n-1} C_{k,n-k}\Big(192\binom{k}{4} + 96(n+2)\binom{k}{3} - 4(12n^2-30n-5)\binom{k}{2}\\
&\qquad\qquad + (8n^3-24n^2+26n-6)k - n^4\Big)\\
&\quad = 96\binom{n}{4}\Big(1 - \frac{3}{n-1}\cdot\frac{5!!}{6!!}\cdot\alpha_n\Big)\\
&\qquad\qquad + 48(n+2)\binom{n}{3}\Big(1-\frac{2}{n-1}\cdot\frac{3!!}{4!!}\cdot\alpha_n\Big)\\
&\qquad\qquad - 2(12n^2-30n-5)\binom{n}{2}\Big(1-\frac{1}{2(n-1)}\cdot\alpha_n\Big)\\
&\qquad\qquad + \frac{1}{2}(8n^3-24n^2+26n-6)n - n^4\\
&\quad = (3n-2)n^3 - \frac{n(15n^2-15n+4)}{4}\cdot\alpha_n.
\end{aligned}$$

The second sum in this independent term can be computed using Lemma 1.34:

$$\begin{aligned}
&\sum_{k=1}^{n-1} C_{k,n-k}\left[96\binom{k}{4} + 156\binom{k}{3} - 4(n^2-n-16)\binom{k}{2} - 4(n^2-n-1)k\right]\alpha_k\\
&\quad = 48\binom{n}{4}\Big(\alpha_n - \frac{6!!}{5!!}\Big) + 78\binom{n}{3}\Big(\alpha_n - \frac{4!!}{3!!}\Big)\\
&\qquad - 2(n^2-n-16)\binom{n}{2}(\alpha_n-2) - 2(n^2-n-1)n(\alpha_n-1)\\
&\quad = n^3(n+1)\alpha_n - \frac{2n(33n^3-13n^2-12n+7)}{15}.
\end{aligned}$$

Finally, as far as the third sum in the independent term goes, it can be computed as follows:

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

Therefore, the independent term of Equation (3.5) is

$$\begin{aligned}
&\frac{n(15n^2-15n+4)}{4}\cdot\alpha_n-(3n-2)n^3\\
&\qquad+n^3(n+1)\alpha_n-\frac{2n(33n^3-13n^2-12n+7)}{15}+\frac{n+6}{8}\cdot\binom{n+1}{3}\cdot\alpha_n\\
&=\frac{n(49n^3+234n^2-181n+42)}{48}\cdot\alpha_n-\frac{n(111n^3-56n^2-24n+14)}{15}\\
&=\left(3n+36\binom{n}{2}+66\binom{n}{3}+\frac{49}{2}\binom{n}{4}\right)\alpha_n-3n-78\binom{n}{2}-244\binom{n}{3}-\frac{888}{5}\binom{n}{4}.
\end{aligned}$$

Thus, Equation (3.5) becomes

$$\begin{aligned}
E_{\mathrm{unif}}((C_n^{(2)})^2)&=2\sum_{k=1}^{n-1}C_{k,n-k}E_{\mathrm{unif}}((C_k^{(2)})^2)-3n-78\binom{n}{2}-244\binom{n}{3}-\frac{888}{5}\binom{n}{4}\\
&\qquad+\left(3n+36\binom{n}{2}+66\binom{n}{3}+\frac{49}{2}\binom{n}{4}\right)\alpha_n.
\end{aligned}$$

This equation can be solved using Theorem 1.35 in the Preliminaries, and the fact that $E_{\mathrm{unif}}((C_1^{(2)})^2)=0$. Its solution is

$$\begin{aligned}
E_{\mathrm{unif}}((C_n^{(2)})^2)&=3n+84\binom{n}{2}+320\binom{n}{3}+360\binom{n}{4}+112\binom{n}{5}\\
&\qquad-\left(3n+39\binom{n}{2}+\frac{183}{2}\binom{n}{3}+\frac{111}{2}\binom{n}{4}\right)\alpha_n\\
&=\frac{n}{15}(14n^4+85n^3-60n^2+5n+1)-\frac{n}{16}(37n^3+22n^2-13n+2)\alpha_n.
\end{aligned}$$

Finally,

$$\begin{aligned}
\sigma^2_{\mathrm{unif}}(C_n^{(2)})&=E_{\mathrm{unif}}((C_n^{(2)})^2)-E_{\mathrm{unif}}(C_n^{(2)})^2=\frac{2}{15}(2n-1)(7n^2+9n-1)\binom{n+1}{2}\\
&\qquad-\frac{1}{8}(5n^2+n+2)\binom{n+1}{2}\frac{(2n-2)!!}{(2n-3)!!}-\binom{n+1}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2,
\end{aligned}$$

as we claimed. □

We shall now briefly discuss the asymptotic behaviour of $E_{\mathrm{unif}}(C_n^{(2)})$ and $\sigma_{\mathrm{unif}}(C_n^{(2)})$. By using Stirling's approximation for large factorials, we have

$$\frac{(2n-2)!!}{(2n-3)!!}=\frac{(2^{n-1}(n-1)!)^2}{(2n-2)!}\sim\frac{\left(2^{n-1}\sqrt{2\pi(n-1)}(n-1)^{n-1}e^{-(n-1)}\right)^2}{\sqrt{2\pi(2n-2)}(2n-2)^{2n-2}e^{-(2n-2)}}\sim\sqrt{\pi n}. \tag{3.6}$$

Then, Theorem 3.4 implies the following limit behaviour:

$$E_{\mathrm{unif}}(C_n^{(2)})\sim\frac{\sqrt{\pi}}{2}n^{5/2},\quad \sigma_{\mathrm{unif}}(C_n^{(2)})\sim\sqrt{\frac{14}{15}}n^{5/2}.$$

So, under the Uniform model, the expected value and the standard deviation of $C_n^{(2)}$ grow with $n$ in the same order. This already happenned with the Colless index, for which it is known that (see [8] for details):

$$E_{\mathrm{unif}}(C_n) \sim \sqrt{\pi} n^{3/2}, \quad \sigma_{\mathrm{unif}}(C_n) \sim \sqrt{\frac{10-3\pi}{3}} n^{3/2}.$$

### 3.2.2 The Yule model

The main goal of this section is to prove the following result. In it, $H_n = \sum_{i=1}^n 1/i$ and $H_n^{(2)} = \sum_{i=1}^n 1/i^2$ are the $n$-th harmonic and second order harmonic numbers, respectively.

**Theorem 3.7.** *Let $n \in \mathbb{N}_{\geq 1}$.*

*(i) The expected value of $C_n^{(2)}$ under the Yule model is*

$$E_{\mathrm{Yule}}(C_n^{(2)}) = n(n+1) - 2nH_n.$$

*(ii) The variance of $C_n^{(2)}$ under the Yule model is*

$$\sigma^2_{\mathrm{Yule}}(C_n^{(2)}) = \frac{1}{3} n \left( n^3 - 8n^2 + 50n - 1 - 40H_n - 12nH_n^{(2)} \right).$$

In its proof we shall use Lemma 1.31 which, applied to the Yule model, says that, for any binary recursive shape index $I$,

$$E_{\mathrm{Yule}}(I_n) = \frac{1}{n-1} \sum_{k=1}^{n-1} \left( 2E_{\mathrm{Yule}}(I_k) + f_I(k, n-k) \right) \tag{3.7}$$

$$\begin{aligned} E_{\mathrm{Yule}}(I_n^2) = \frac{1}{n-1} \sum_{k=1}^{n-1} \Big( & 2E_{\mathrm{Yule}}(I_k^2) + 4 f_I(k, n-k) E_{\mathrm{Yule}}(I_k) + 2E_{\mathrm{Yule}}(I_k) E_{\mathrm{Yule}}(I_{n-k}) \\ & + f_I(k, n-k)^2 \Big). \end{aligned} \tag{3.8}$$

As we did in the last section, and since this proof will occupy the whole section, we shall split it into two lemmata, one for each statement.

**Lemma 3.8.** *For every $n \in \mathbb{N}_{\geq 1}$,*

$$E_{\mathrm{Yule}}(C_n^{(2)}) = n(n+1) - 2nH_n.$$

*Proof.* By Equation (3.7),

$$\begin{aligned}
E_{\text{Yule}}(C_n^{(2)}) &= \frac{2}{n-1}\sum_{k=1}^{n-1} E_{\text{Yule}}(C_k^{(2)}) + \frac{1}{n-1}\sum_{k=1}^{n-1}(n-2k)^2\\
&= \frac{2}{n-1}\sum_{k=1}^{n-1} E_{\text{Yule}}(C_k^{(2)}) + \frac{1}{3}n(n-2)\\
&= \frac{2}{n-1}E_{\text{Yule}}(C_{n-1}^{(2)}) + \frac{n-2}{n-1}\Big(\frac{2}{n-2}\sum_{k=1}^{n-2} E_{\text{Yule}}(C_k^{(2)})\Big) + \frac{1}{3}n(n-2)\\
&= \frac{2}{n-1}E_{\text{Yule}}(C_{n-1}^{(2)}) + \frac{n-2}{n-1}\Big(E_{\text{Yule}}(C_{n-1}^{(2)}) - \frac{1}{3}(n-1)(n-3)\Big) + \frac{1}{3}n(n-2)\\
&= \frac{n}{n-1}E_{\text{Yule}}(C_{n-1}^{(2)}) + n - 2.
\end{aligned}$$

We can now divide this equation by $n$ and by setting $X_n = E_{\text{Yule}}(C_n^{(2)})/n$, we obtain the equation

$$X_n = X_{n-1} + 1 - \frac{2}{n}$$

whose solution, with initial condition $X_1 = E_{\text{Yule}}(C_1^{(2)}) = 0$, is

$$X_n = \sum_{k=2}^{n}\Big(1 - \frac{2}{k}\Big) = n + 1 - 2H_n.$$

Thus, finally,

$$E_{\text{Yule}}(C_n^{(2)}) = nX_n = n(n+1) - 2nH_n,$$

as we claimed. □

**Lemma 3.9.** *For every $n \in \mathbb{N}_{\geq 1}$,*

$$\sigma^2_{\text{Yule}}(C_n^{(2)}) = \frac{1}{3}n\big(n^3 - 8n^2 + 50n - 1 - 30H_n - 12nH_n^{(2)}\big).$$

*Proof.* We will compute the value of $\sigma^2_{\text{Yule}}(C_n^{(2)})$ by means of the identity

$$\sigma^2_{\text{Yule}}(C_n^{(2)}) = E_{\text{Yule}}((C_n^{(2)})^2) - E_{\text{Yule}}(C_n^{(2)})^2, \tag{3.9}$$

where $E_{\text{Yule}}(C_n^{(2)})$ is given by the previous result, Lemma 3.8. Now, we must compute

$E_{\text{Yule}}((C_n^{(2)})^2)$. By Equation (3.8),

$$\begin{aligned}
E_{\text{Yule}}((C_n^{(2)})^2) &= \frac{1}{n-1}\sum_{k=1}^{n-1}\Big(2E_{\text{Yule}}((C_k^{(2)})^2) + (n-2k)^4 \\
&\quad + 4(n-2k)^2 E_{\text{Yule}}(C_k^{(2)}) + 2E_{\text{Yule}}(C_k^{(2)})E_{\text{Yule}}(C_{n-k}^{(2)})\Big) \\
&= \frac{2}{n-1}\sum_{k=1}^{n-1} E_{\text{Yule}}((C_k^{(2)})^2) + \frac{1}{n-1}\sum_{k=1}^{n-1}(n-2k)^4 \\
&\quad + \frac{4}{n-1}\sum_{k=1}^{n-1}(n-2k)^2 k(k+1-2H_k) \\
&\quad + \frac{2}{n-1}\sum_{k=1}^{n-1} k(n-k)(k+1-2H_k)(n-k+1-2H_{n-k})
\end{aligned}$$

Set $T_n$ as the independent term in this equation, so that it can be re-written as

$$\begin{aligned}
E_{\text{Yule}}((C_n^{(2)})^2) &= \frac{2}{n-1}\sum_{k=1}^{n-1} E_{\text{Yule}}((C_k^{(2)})^2) + T_n \\
&= \frac{2}{n-1}E_{\text{Yule}}((C_{n-1}^{(2)})^2) + \frac{n-2}{n-1}\cdot\frac{2}{n-2}\sum_{k=1}^{n-2} E_{\text{Yule}}((C_k^{(2)})^2) + T_n \\
&= \frac{2}{n-1}E_{\text{Yule}}((C_{n-1}^{(2)})^2) + \frac{n-2}{n-1}(E_{\text{Yule}}((C_{n-1}^{(2)})^2) - T_{n-1}) + T_n \\
&= \frac{n}{n-1}E_{\text{Yule}}((C_{n-1}^{(2)})^2) + T_n - \frac{n-2}{n-1}T_{n-1}.
\end{aligned}$$

We divide this equation by $n$ and set $Y_n = E_{\text{Yule}}((C_n^{(2)})^2)/n$. We thus obtain the relation

$$Y_n = Y_{n-1} + \frac{1}{n}\Big(T_n - \frac{n-2}{n-1}T_{n-1}\Big). \tag{3.10}$$

Our next goal is to compute the independent term of this equation as an explicit expression in $n$. In order to achieve that goal, we need first to compute all three sums that form $T_n$. First,

$$\frac{1}{n-1}\sum_{k=1}^{n-1}(n-2k)^4 = \frac{1}{15}n(n-2)(3n^2-6n-4). \tag{3.11}$$

As for the second sum,

$$
\begin{aligned}
&\frac{4}{n-1}\sum_{k=1}^{n-1}(n-2k)^2k(k+1-2H_k)\\
&\quad=\frac{4}{n-1}\left(\sum_{k=1}^{n-1}(n-2k)^2k(k+1)-2(n-2)^2\sum_{k=1}^{n-1}kH_k\right.\\
&\qquad\qquad\left.+16(n-3)\sum_{k=1}^{n-1}\binom{k}{2}H_k-48\sum_{k=1}^{n-1}\binom{k}{3}H_k\right)\\
&\quad=\frac{4}{n-1}\left(\frac{1}{15}(n-1)n(n+1)(2n^2-5n+2)-2(n-2)^2\binom{n}{2}\left(H_n-\frac{1}{2}\right)\right.\\
&\qquad\qquad\left.+16(n-3)\binom{n}{3}\left(H_n-\frac{1}{3}\right)-48\binom{n}{4}\left(H_n-\frac{1}{4}\right)\right)\\
&\quad=\frac{2}{45}n(n-2)(12n^2+16n+9)-\frac{4}{3}n^2(n-2)H_n
\end{aligned}
\tag{3.12}
$$

using, in the second last equality above, that

$$
\sum_{k=1}^{n-1}\binom{k}{m}H_k=\binom{n}{m+1}\left(H_n-\frac{1}{m+1}\right);
\tag{3.13}
$$

see Equation (6.70) in [51].

Finally,

$$
\begin{aligned}
&\frac{2}{n-1}\sum_{k=1}^{n-1}k(n-k)(k+1-2H_k)(n-k+1-2H_{n-k})\\
&\quad=\frac{2}{n-1}\left[\sum_{k=1}^{n-1}k(n-k)(k+1)(n-k+1)-2\sum_{k=1}^{n-1}k(n-k)(n-k+1)H_k\right.\\
&\qquad\qquad\left.-2\sum_{k=1}^{n-1}k(n-k)(k+1)H_{n-k}+4\sum_{k=1}^{n-1}k(n-k)H_kH_{n-k}\right]\\
&\quad=\frac{2}{n-1}\left[\sum_{k=1}^{n-1}k(k+1)(n-k)(n-k+1)-4\sum_{k=1}^{n-1}k(n-k)(n-k+1)H_k\right.\\
&\qquad\qquad\left.+4n\sum_{k=1}^{n-1}kH_kH_{n-k}-4\sum_{k=1}^{n-1}k^2H_kH_{n-k}\right]\\
&\quad=\frac{2}{n-1}\left[\sum_{k=1}^{n-1}k(k+1)(n-k)(n-k+1)-4\sum_{k=1}^{n-1}\left(6\binom{k}{3}-4(n-1)\binom{k}{2}+n(n-1)k\right)H_k\right.\\
&\qquad\left.+4n\sum_{k=1}^{n-1}kH_kH_{n-k}-4\sum_{k=1}^{n-1}k^2H_kH_{n-k}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2}{n-1}\Bigg[4\binom{n+3}{5} - 24\binom{n}{4}\Big(H_n - \frac{1}{4}\Big) + 16(n-1)\binom{n}{3}\Big(H_n - \frac{1}{3}\Big)\\
&\qquad - 4n(n-1)\binom{n}{2}\Big(H_n - \frac{1}{2}\Big) + 4n\binom{n+1}{2}(H_{n+1}^2 - H_{n+1}^{(2)} - 2H_{n+1} + 2)\\
&\qquad - \frac{4}{3}\binom{n+1}{2}\Big((2n+1)(H_{n+1}^2 - H_{n+1}^{(2)}) - \frac{13n+5}{3}H_{n+1} + \frac{71n+37}{18}\Big)\Bigg]\\
&= \frac{1}{270}n(18n^3 + 303n^2 + 1163n + 98) - \frac{2}{9}n(n+1)(3n+16)H_n\\
&\qquad + \frac{4}{3}n(n+1)(H_{n+1}^2 - H_{n+1}^{(2)}) \qquad (3.14)
\end{aligned}
$$

using, in the second last equality above, Equation (3.13) and the identities

$$
\begin{aligned}
\sum_{k=1}^{n-1} kH_kH_{n-k} &= \binom{n+1}{2}(H_{n+1}^2 - H_{n+1}^{(2)} - 2H_{n+1} + 2)\\
\sum_{k=1}^{n-1} k^2H_kH_{n-k} &= \frac{n(n+1)}{6}\Big((2n+1)(H_{n+1}^2 - H_{n+1}^{(2)}) - \frac{13n+5}{3}H_{n+1} + \frac{71n+37}{18}\Big)
\end{aligned}
$$

proved in [125].

Therefore,

$$
\begin{aligned}
T_n &= \frac{1}{15}n(n-2)(3n^2 - 6n - 4) + \frac{2}{45}n(n-2)(12n^2 + 16n + 9) - \frac{4}{3}n^2(n-2)H_n\\
&\quad + \frac{1}{270}n(18n^3 + 303n^2 + 1163n + 98) - \frac{2}{9}n(n+1)(3n+16)H_n\\
&\quad + \frac{4}{3}n(n+1)(H_{n+1}^2 - H_{n+1}^{(2)})\\
&= \frac{1}{270}n(216n^3 - 9n^2 + 1031n + 26) - \frac{2}{9}n(9n^2 + 7n + 16)H_n\\
&\quad + \frac{4}{3}n(n+1)(H_{n+1}^2 - H_{n+1}^{(2)})
\end{aligned}
$$

and, so, the independent term in Equation (3.10) is

$$
\begin{aligned}
&\frac{1}{n}\Big(T_n - \frac{n-2}{n-1}T_{n-1}\Big)\\
&\quad = \frac{1}{n}\Bigg[\frac{1}{270}n(216n^3 - 9n^2 + 1031n + 26) - \frac{2}{9}n(9n^2 + 7n + 16)H_n\\
&\qquad + \frac{4}{3}n(n+1)(H_{n+1}^2 - H_{n+1}^{(2)})\\
&\qquad - \frac{n-2}{n-1}\Bigg(\frac{1}{270}(n-1)(216(n-1)^3 - 9(n-1)^2 + 1031(n-1) + 26)\\
&\qquad - \frac{2}{9}(n-1)(9(n-1)^2 + 7(n-1) + 16)H_{n-1}\\
&\qquad + \frac{4}{3}(n-1)n(H_n^2 - H_n^{(2)})\Bigg)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\Bigg[\frac{1}{270}n(216n^3 - 9n^2 + 1031n + 26)\\
&\qquad - \frac{2}{9}n(9n^2 + 7n + 16)H_{n-1} - \frac{2}{9}(9n^2 + 7n + 16)\\
&\qquad + \frac{4}{3}n(n+1)(H_n^2 - H_n^{(2)}) + \frac{8}{3}nH_{n-1} + \frac{8}{3}\\
&\qquad - \frac{1}{270}(n-2)(216n^3 - 657n^2 + 1697n - 1230)\\
&\qquad + \frac{2}{9}(n-2)(9n^2 - 11n + 18)H_{n-1}\\
&\qquad - \frac{4}{3}(n-2)n(H_n^2 - H_n^{(2)})\Bigg]\\
&= \frac{1}{n}\Big(\frac{1}{3}(12n^3 - 28n^2 + 47n - 30) - 8(n^2 - n + 1)H_{n-1}\\
&\qquad + 4n(H_n^2 - H_n^{(2)})\Big)\\
&= 4n^2 - \frac{28}{3}n + \frac{47}{3} - \frac{10}{n} - 8(n-1)H_{n-1} - \frac{8H_{n-1}}{n} + 4H_n^2 - 4H_n^{(2)}
\end{aligned}
$$

The solution to Equation (3.10) with initial condition $Y_1 = E_{\text{Yule}}(C_1^{(2)}) = 0$ is

$$
\begin{aligned}
Y_n &= \sum_{k=2}^{n} \frac{1}{k}\Big(T_k - \frac{k-2}{k-1}T_{k-1}\Big)\\
&= \sum_{k=2}^{n}\Big(4k^2 - \frac{28}{3}k + \frac{47}{3} - \frac{10}{k} - 8(k-1)H_{k-1} - \frac{8H_{k-1}}{k} + 4H_k^2 - 4H_k^{(2)}\Big)\\
&= \sum_{k=1}^{n-1}\left(4(k+1)^2 - \frac{28}{3}(k+1) + \frac{47}{3} - \frac{10}{k+1} - 8kH_k - \frac{8H_k}{k+1} + 4H_{k+1}^2 - 4H_{k+1}^{(2)}\right)\\
&\overset{(*)}{=} \frac{1}{3}(4n^3 - 8n^2 + 35n - 31) - 10(H_n - 1) - 8\binom{n}{2}\Big(H_n - \frac{1}{2}\Big) - 4(H_n^2 - H_n^{(2)})\\
&\qquad + 4\big((n+1)H_n^2 - (2n+1)H_n + 2n - 1\big) - 4\big((n+1)H_n^{(2)} - H_n - 1\big)\\
&= \frac{1}{3}(4n^3 - 2n^2 + 53n - 1) - 2(2n^2 + 2n + 5)H_n + 4n(H_n^2 - H_n^{(2)})
\end{aligned}
$$

where, in the second last identity (marked with (∗)) we have used Equation (3.13) and the identities

$$\sum_{k=1}^{n-1} \frac{H_k}{k+1} = \frac{1}{2}(H_n^2 - H_n^{(2)})$$

(cf. Equation (6.71) in [51]) and

$$
\begin{aligned}
\sum_{k=1}^{n-1} H_k^2 &= nH_n^2 - (2n+1)H_n + 2n\\
\sum_{k=1}^{n-1} H_k^{(2)} &= nH_n^{(2)} - H_n
\end{aligned}
$$

(see [70, §1.2.7]).

Thus,

$$\begin{aligned} E_{\text{Yule}}((C_n^{(2)})^2) &= nY_n \\ &= \frac{n}{3}(4n^3 - 2n^2 + 53n - 1) - 2n(2n^2 + 2n + 5)H_n + 4n^2(H_n^2 - H_n^{(2)}) \end{aligned}$$

and

$$\begin{aligned} \sigma^2_{\text{Yule}}(C_n^{(2)}) &= E_{\text{Yule}}((C_n^{(2)})^2) - E_{\text{Yule}}(C_n^{(2)})^2 \\ &= \frac{1}{3}n\big(n^3 - 8n^2 + 50n - 1 - 30H_n - 12nH_n^{(2)}\big) \end{aligned}$$

as we claimed. □

Let us determine the asymptotic behaviour of $E_{\text{Yule}}(C_n^{(2)})$ and $\sigma_{\text{Yule}}(C_n^{(2)})$. Using (see, for instance, [51]) that

$$H_n \sim \ln(n), \quad H_n^{(2)} \sim \frac{\pi^2}{6},$$

Theorem 3.7 implies that

$$E_{\text{Yule}}(C_n^{(2)}) \sim n^2, \qquad \sigma_{\text{Yule}}(C_n^{(2)}) \sim \frac{1}{\sqrt{3}}n^2.$$

So, again under the Yule model, the expected value and the standard deviation of $C_n^{(2)}$ grow with $n$ in the same, quadratic, order. This is in contrast with the Colless index, for which the expected value grows faster than the standard deviation (see [8, 13]):

$$E_{\text{Yule}}(C_n) \sim n\log(n), \qquad \sigma_{\text{Yule}}(C_n) \sim \sqrt{\frac{18 - 6\log(2) - \pi^2}{6}}n.$$

## 3.3 Numerical results

Due to the fact that the range of values that the Quadratic Colless index can attain on **BinTree**$_n$, for a fixed number of leaves $n \geq 1$, is an order of magnitude bigger than that of the Colless index and the Sackin index, and it is also slightly bigger than that of the Cophenetic index (see Table 3.1), the question of whether the probability of two trees with the same number of leaves having the exact same $C^{(2)}$ value is smaller than this probability under those other indices reveals itself as pertinent. Indeed, our intuition tells us that this should be, indeed, the case. In order to simplify the language, we shall say that whenever a balance index $I$ takes the same value on a pair of different trees of **BinTree**$_n$ (or **Tree**$_n$, for that matter) we have a *tie*. Of course, for $n \geq 12$, ties are inevitable: from that number on, the range of possible values of $C^{(2)}$ is much narrower than the number of trees in **BinTree**$_n$ (see [38, Table 3.3] for the cardinality of **BinTree**$_n$ for small values of $n$), and so the pigeonhole principle implies that ties will always take place.

In order to check this hypothesis, we have computed the probability of a balance index $I \in \{C, S, \Phi, C^{(2)}, QI\}$ having a tie in **BinTree**$_n$, where QI stands for the Quartet index defined in Chapter 5. We have opted to call this probability $p_n(I)$. Concretely, we have defined $p_n(I)$ as the number of unordered pairs of trees $\{T_1, T_2\}$, with $T_1, T_2 \in$ **BinTree**$_n$, such that $T_1 \neq T_2$ and $I(T_1) = I(T_2)$, divided by $\binom{|\mathbf{BinTree}_n|}{2}$. As it can be seen in Figure 3.1, the Quadratic Colless index has less ties than the Colless, Sackin and Cophenetic indices, although the Quartet index (to which we devote the integrity of Chapter 5) shows a better performance towards $n = 16$ and beyond, as it also should be expected because, as we shall see, the width of its range of values is in $O(n^4)$.

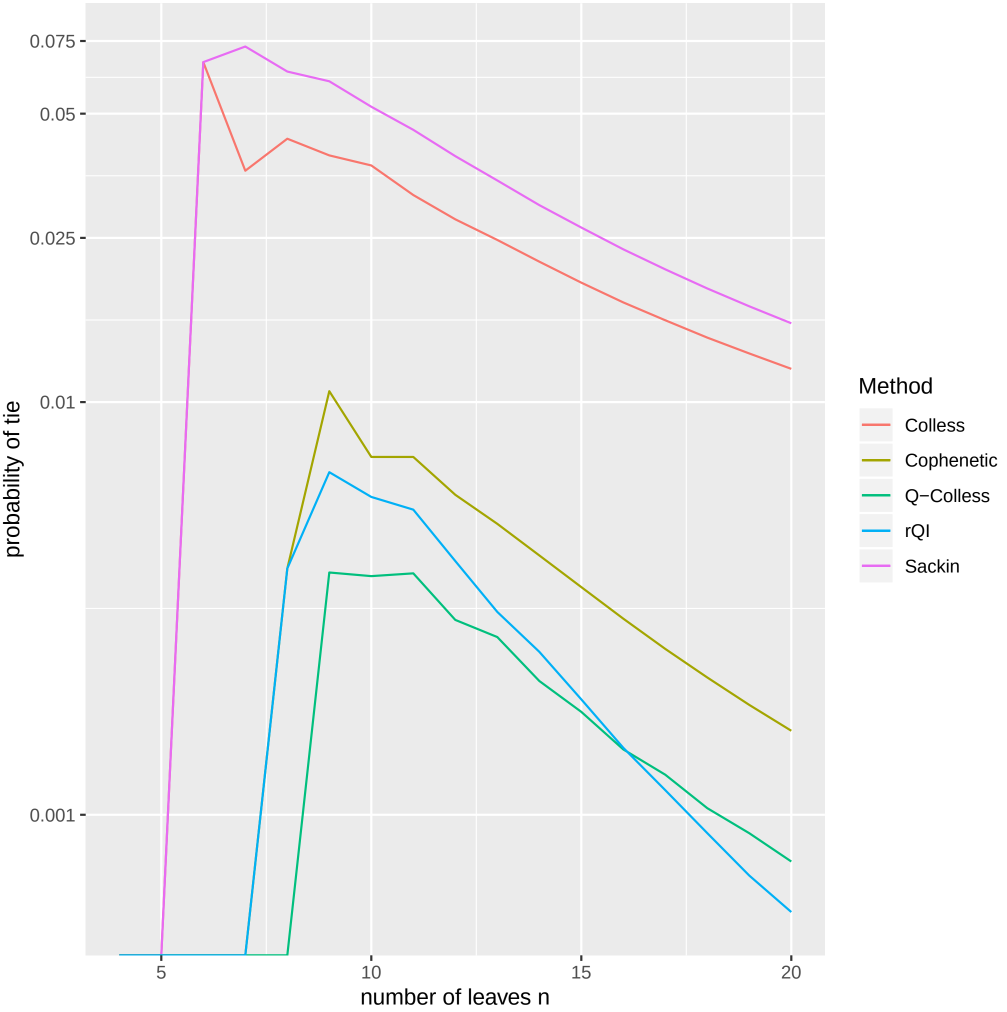


Figure 3.1: Plot of $p_n(I)$ for $I \in \{C, S, \Phi, C^{(2)}, \mathrm{QI}\}$ and $n \in \{1, \ldots, 20\}$.

Another way to assess the discriminating skill of an index is to evaluate its power in statistical tests designed to distinguish between dissimilar trees, and compare it with that of other balance measures. In [59], Hayati, Shadgar and Chindelevitch have developed a new resolution function based on the Laplacian matrix of the tree that seeks to evaluate the power of tree shape statistics discriminating between dissimilar trees. They test this resolution function together with the usual function based on the NNI metric. Thus, they are able to rank some balance indices by their discriminative power on all possible

| $n$ | $C$ | $S$ | $V$ | $I_2$ | $B_1$ | $B_2$ | Saless | $C^{(2)}$ |
|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 0.8157 | 0.8510 | 0.8144 | 0.7611 | 0.7546 | 0.8705 | 0.8315 | 0.8709 |
| 7 | 0.9251 | 0.9303 | 0.9023 | 0.8844 | 0.8649 | 0.9254 | 0.9297 | 0.9360 |
| 8 | 0.9255 | 0.9122 | 0.8753 | 0.8612 | 0.8326 | 0.9113 | 0.9235 | 0.9218 |
| 9 | 0.9184 | 0.9208 | 0.8826 | 0.8539 | 0.8324 | 0.907 | 0.9224 | 0.9302 |
| 10 | 0.941 | 0.9380 | 0.8985 | 0.8545 | 0.8326 | 0.9085 | 0.9426 | 0.9475 |
| 11 | 0.9531 | 0.9514 | 0.9102 | 0.8552 | 0.8375 | 0.9132 | 0.9551 | 0.9604 |
| 12 | 0.9533 | 0.9523 | 0.9086 | 0.8504 | 0.8311 | 0.9045 | 0.9556 | 0.9632 |
| 13 | 0.9541 | 0.9542 | 0.9078 | 0.8416 | 0.8247 | 0.8992 | 0.9567 | 0.9657 |
| 14 | 0.9552 | 0.9548 | 0.9070 | 0.8374 | 0.82 | 0.8902 | 0.9575 | 0.967 |
| 15 | 0.9546 | 0.9544 | 0.9049 | 0.8298 | 0.813 | 0.8826 | 0.9569 | 0.9674 |
| 16 | 0.9543 | 0.9541 | 0.9034 | 0.8265 | 0.8089 | 0.8743 | 0.9564 | 0.9677 |
| 17 | 0.9534 | 0.9534 | 0.9006 | 0.8199 | 0.8024 | 0.8678 | 0.9555 | 0.9679 |

Table 3.2: Scaled resolution scores for shape indices on the NNI distance matrix, for different numbers $n$ of leaves. The value of the resolution is between 0 and 1. Higher values represent more discriminating power.

phylogenetic trees on the same number of leaves.

We have performed the exact same experiment on the exact same data (provided, as well as the code, along with [59]; we want to thank the authors for their readiness to help us understand their code). As it turns out, $C^{(2)}$ performs better than any of the other tested indices do, including the *Saless index*, an *ad hoc* linear combination of the Sackin and Colless indices introduced in the same paper [59]. This Saless index was the best performing one when tested under the NNI metric, although not so under the resolution function proposed in [59]: under this one the Colless index performed better. Here, we present the two tables we have obtained in our experiment: Table 3.2, with the scores under the NNI distance (bigger values represent more power), and Table 3.3 under the Hayati-Shadgar-Chindelevitch resolution function (lower values represent more power). As we see in the first table, $C^{(2)}$ performs best except when $n = 8$, in which it is outperformed by the Saless index. On the other hand, in the second table $C^{(2)}$ performs second until $n = 14$ behind the Sackin ($n = 7$) and Colless indices ($n \in \{8, \dots, 13\}$).

## 3.4 Discussion

Despite being one of the oldest and most popular balance indices in the phylogenetic literature (it dates back to 1982 and its number of cites in Google Scholar doubles that of the second most cited balance measure, the Sackin index[1]), the Colless index has a number of relevant drawbacks. For instance, as we saw in the previous chapter, the characterization of the trees that achieve its minimum value is far from intuitive, and it clashes with the idea that only "the most balanced" trees, i.e. the maximally balanced trees [107], should attain it. Notice that several balance indices existing in the literature

[1] 260 *vs* 131 citations; data retrieved on June 20, 2020.

| $n$ | $C$ | $S$ | $V$ | $I_2$ | $B_1$ | $B_2$ | $C^{(2)}$ |
|---|---|---|---|---|---|---|---|
| 7 | 0.0984 | 0.0937 | 0.1082 | 0.1115 | 0.1178 | 0.0989 | 0.0948 |
| 8 | 0.0808 | 0.0955 | 0.111 | 0.0893 | 0.1164 | 0.0965 | 0.0941 |
| 9 | 0.0507 | 0.0566 | 0.0662 | 0.068 | 0.0797 | 0.0653 | 0.0558 |
| 10 | 0.0327 | 0.0379 | 0.0471 | 0.0535 | 0.0629 | 0.0451 | 0.0357 |
| 11 | 0.0222 | 0.0255 | 0.0326 | 0.0458 | 0.0511 | 0.0348 | 0.0236 |
| 12 | 0.0183 | 0.0217 | 0.0282 | 0.0429 | 0.0473 | 0.0304 | 0.0194 |
| 13 | 0.016 | 0.0185 | 0.0238 | 0.0413 | 0.0441 | 0.0283 | 0.0163 |
| 14 | 0.0147 | 0.0170 | 0.0217 | 0.04 | 0.0421 | 0.0265 | 0.0147 |
| 15 | 0.0137 | 0.0157 | 0.0197 | 0.039 | 0.0404 | 0.0256 | 0.0134 |
| 16 | 0.013 | 0.0148 | 0.0184 | 0.038 | 0.0389 | 0.0247 | 0.0126 |
| 17 | 0.0123 | 0.014 | 0.017 | 0.037 | 0.0375 | 0.0238 | 0.0118 |
| 18 | 0.0117 | 0.0132 | 0.016 | 0.0358 | 0.0361 | 0.0229 | 0.0111 |
| 19 | 0.0112 | 0.0127 | 0.015 | 0.0347 | 0.0349 | 0.0222 | 0.0105 |
| 20 | 0.0107 | 0.012 | 0.0141 | 0.0339 | 0.0338 | 0.0217 | 0.01 |
| 21 | 0.0102 | 0.0114 | 0.0133 | 0.0329 | 0.0327 | 0.0209 | 0.01 |

Table 3.3: Scaled resolution scores for shape indices on the resolution function presented in [59] for different numbers $n$ of leaves. The value of the resolution is between 0 and 1. Lower values represent more discriminating power.

do agree with this condition, such as the Cophenetic index [85] and our Quartet index (Chapter 5). Furthermore, closed formulæ for its expected value and variance under the Uniform model are not yet known.

In this chapter, we have presented an alternative way to capture the intuition behind the Colless index that turns out to avoid the aforementioned disadvantages, by squaring the balance values of the internal nodes, instead of taking their absolute value, in the definition of the index. In the first section, we have shown that, given a number of leaves $n \in \mathbb{N}_{\geq 1}$, the maximum and minimum values of this Quadratic Colless index, $C^{(2)}$, are reached by a single tree each. We have computed these values —showing, in particular, that the minimum value of $C^{(2)}$ is equal to that of the Colless index— and proved that the trees attaining them are exactly the caterpillar and the maximally balanced tree. By computing these values, we have established that the range of the Quadratic Colless index is $O(n^3)$, which is an order of magnitude bigger than that of the Colless index and on pair with that of the Cophenetic index (see Table 3.1).

We have then proceeded to compute both the expected value and the variance of the random variable $C_n^{(2)}$ under the Uniform and the Yule model in the second section. To our knowledge, they are both still unknown in the case of the Colless index under the Uniform model. Hence, we are confident that in this regard the Quadratic Colless index also presents an advantage over the original Colless index.

Finally, in the third section of this chapter, we have presented the results of some numerical experiments aimed to assess the discriminatory power of this new measure. This has been done by, firstly, computing the probability of a tie up to 20 leaves and seeing that $C^{(2)}$ fares way better than the Cophenetic, Colless or Sackin indices do, although slightly worse than the Quartet index does for large values of $n$. Secondly, we have tested under the metrics proposed in [59] the power of the Quadratic Colless index when it comes to discriminate between similar or dissimilar trees. Under both metrics proposed in [59], $C^{(2)}$ has sistematically been one of the best performing measures, being often superior to both the Sackin and Colless indices.

### 3.4.1 Colless to the $2m$?

Now, Theorem 3.2 can be easily stated and proved if we consider $C^{(2m)}$ for some $m \geq 1$. This fact gives rise to the question of whether Theorem 3.3 can also be so converted. Indeed, for then we could define a balance index as, for a given bifurcating tree $T \in$ **BinTree**$_n$, $n \in \mathbb{N}_{\geq 1}$,

$$C^{(2m)}(T) = \sum_{u \in \mathring{V}(T)} \mathrm{bal}(u)^{2m}.$$

Such an index would have an obvious recurrent representation, given, for any bifurcating tree $T = T_1 * T_2$, with $T_1 \in$ **BinTree**$_{n_1}$ and $T_2 \in$ **BinTree**$_{n_2}$, by

$$C^{(2m)}(T_1 * T_2) = C^{(2m)}(T_1) + C^{(2m)}(T_2) + (n_1 - n_2)^{2m}.$$

In this final section we shall prove that the maximum value of such an index is attained exactly by the caterpillar. Therefore, its minimum and maximum values are $c(n)$ and $\sum_{i=1}^{n-2} i^{2m}$ for any number of leaves $n \in \mathbb{N}_{\geq 1}$. This means that the range of values that such an index can attain is $O(n^{2m+1})$, by the Faulhaber's formula. However, the computation of the moments of $C_n^{(2m)}$ becomes more and more complex.

**Theorem 3.10.** *The maximum of* $C^{(2m)}$ *is reached exactly at the caterpillars. Furthermore, this maximum value for* $n \geq 1$ *leaves is*

$$C^{(2m)}(T_n^{\mathrm{cat}}) = \sum_{i=1}^{n-2} i^{2m} = \frac{(n-2)^{2m+1}}{2m+1} + \frac{1}{2}(n-2)^{2m} \frac{B_k}{k!}(2m)_{k-1}(n-2)^{2m-l+1},$$

*where* $B_k$ *is the* $k$*-th Bernoulli number and* $(2m)_{k-1}$ *is a Pochhammer symbol.*

*Proof.* The value of $C^{(2m)}(T_n^{\mathrm{cat}})$ is easily computed using the recursive formula; the second equality is Faulhaber's polynomial. In order to prove that this value is attained only by the caterpillars, we shall proceed by induction. The result obviously holds when $n = 1$, and so suppose it holds up to $n - 1$ leaves.

Let $T_1 * T_2 \in$ **BinTree**$_n$ be such that $T_1 \in$ **BinTree**$_{n_1}$ and $T_2 \in$ **BinTree**$_{n_2}$. We then have that

$$\begin{aligned} C^{(2m)}(T_1 * T_2) &= C^{(2m)}(T_1) + C^{(2m)}(T_2) + (n_1 - n_2)^{2m} \\ &\leq C^{(2m)}(T_{n_1}^{\mathrm{cat}}) + C^{(2m)}(T_{n_2}^{\mathrm{cat}}) + (n_1 - n_2)^{2m} = (*) \end{aligned}$$

and the inequality is an equality if, and only if, $T_1 = T_{n_1}^{\mathrm{cat}}$ and $T_2 = T_{n_2}^{\mathrm{cat}}$, by the induction hypothesis. Now, on the one hand,

$$\begin{aligned} C^{(2m)}(T_{n_1}^{\mathrm{cat}}) + C^{(2m)}(T_{n_2}^{\mathrm{cat}}) &= \sum_{i=1}^{n_1-2} i^{2m} + \sum_{i=1}^{n_2-2} i^{2m} \\ &\leq \sum_{i=1}^{n_1+n_2-3} i^{2m} = \sum_{i=1}^{n-3} i^{2m} = C^{(2m)}(T_{n-1}^{\mathrm{cat}}) + C^{(2m)}(T_1^{\mathrm{cat}}) \end{aligned}$$

and, on the other hand, $(n_1 - n_2)^{2m} \leq (n-2)^{2m}$, for every $n_1, n_2 \in \mathbb{N}_{\geq 1}$ such that $n_1 + n_2 = n$. Notice moreover that the equality in both expressions is attained if, and only if, $n_1 = n-1$ and $n_2 = 1$. Therefore

$$(*) \leq C^{(2m)}(T_{n-1}^{\mathrm{cat}}) + C^{(2m)}(T_1^{\mathrm{cat}}) + (n-2)^{2m} = C^{(2m)}(T_n^{\mathrm{cat}})$$

and the equality $C^{(2m)}(T_1 * T_2) = C^{(2m)}(T_n^{\mathrm{cat}})$ holds only when $T_1 = T_{n_1}^{\mathrm{cat}}$ and $T_2 = T_{n_2}^{\mathrm{cat}}$ and $n_1 = n-1$, that is, when $T_1 * T_2 = T_n^{\mathrm{cat}}$. □

Finally, we end by noting that this last argument could have also proved Theorem 3.3.

Chapter 4

# The Variance of depths

> The symmetrical phenogram looks "better" than the skew one because (1) fewer taxonomic rank categories need be postulated and (2) the cluster sizes at any category level are more constant. In terms of **b** [the vector of depths], (1) states that the highest **b**-value is lower for the symmetrical than for the skew phenogram, and (2) may conveniently be translated into the statement that the variation among the **b**-values is lower for the symmetrical phenogram.
>
> M. J. Sackin, *"Good" and "bad" phenograms* [102], 1972

In his 1972 paper on *"Good" and "bad" phenograms*, Sackin pointed out that more balanced trees tend to have lower (maximum) depth and smaller variation of the leaves' depths. In order to illustrate this observation, he compared the multisets of depths of the fully balanced tree and the caterpillar with 8 leaves. Indeed, since the maximum depth of $T_8^{\mathrm{bal}}$ is the least possible one for a tree with 8 leaves, 3, while that of $T_8^{\mathrm{cat}}$ is the largest possible one for a tree with 8 leaves, 7, he backed his first point; and by observing that all the leaves' depths in $T_8^{\mathrm{bal}}$ are equal —for all leaves have depth 3— while all the leaves' depths in $T_8^{\mathrm{bal}}$, except for the compulsory pair of larger values, are different, he found support for the second. In fact, as we shall prove in Theorem 4.4, the caterpillar with $n$ leaves is always the only tree with the largest variance of the leaves' depths among its peers.

Nevertheless, as we have already seen in this memoir, the index that has later become known as the Sackin index $S$ is not the maximum depth of a tree nor any measure of the variation of its leaves' depths, but their sum. This index was actually later introduced by Sokal in [112], and baptized by Shao and Sokal in [107]. Clearly, the maximum depth is a very coarse index, since the range of values it can attain for a tree with $n$ leaves goes from 1 to only $n-1$, and thus it is easy to understand why it did not crystalize as a balance index. However, Sackin's second proposal, to use a measure of the variation of the leaves' depths, seems to be a fairly reasonable idea. It was implemented by Kirkpatrick and Slatkin in [69] as the variance of the leaves' depths, which we shall henceforth call *the Variance of depths*, $V$, and empirically shown to have power similar and sometimes even higher than that of $S$ in some statistical tests whose alternative hypothesis represented "this tree is not random". Yet, although the Variance of depths was used as a shape index in a few early studies [1, 60, 65, 69] and was even collected by Felsenstein in the section "Measures of overall asymmetry" of his textbook [38], it has now been neglected in favour of other indices such as Colless' or Sackin's, and it seems to survive only in a few studies, for instance in [59].

Now, let us define the *Variance of depths* of a tree $T \in \mathbf{Tree}_n$ as

$$V(T) = \frac{1}{n} \sum_{x \in L(T)} (\delta(x) - \overline{S}(T))^2,$$

where $\overline{S}$ denotes the mean depth of $T$ (see page 17). One can easily see that

$$V(T) = \frac{1}{n} S^{(2)}(T) - \overline{S}(T)^2 = \frac{1}{n} S^{(2)}(T) - \frac{1}{n^2} S(T)^2, \tag{4.1}$$

where

$$S^{(2)}(T) = \sum_{x \in L(T)} \delta(x)^2.$$

The main question that is pursued in this chapter is to study the properties of the Variance of depths as a balance index. As we have already discussed, one of the most important properties that such a measure can present is in regard to its extreme values and the trees that attain them. With respect to the maximum value, we have already advanced that the least balanced tree according to it is exactly the caterpillar —just as it is the case in any other balance index worthy of its name. As it was the case with the Sackin index, two trees $T_1, T_2 \in \mathbf{Tree}_n$ for some $n \in \mathbb{N}_{\geq 1}$ such that $\Delta(T_1) = \Delta(T_2)$ share the same Variance of depths, and thence the minimum might not be unique. And this is indeed the case: for instance, when taking into account multifurcating trees, any tree all of whose leaves have the same depth (i.e., any *taxonomic* tree) has Variance of depths 0. Since the variance is always positive or zero, this is indeed its minimum value.

So far so good, but the problem arises now: when the search of the minimum value of this index among bifurcating trees is attempted, the intuition tells us that it should be reached at the maximally balanced trees, as well as at those depth-equivalent to them. And Theorem 1.19 only reinforces this intuition: indeed, for the maximally balanced trees, and those depth-equivalent to them, are the only trees whose leaves' depths are at distance at most 1. We are naturally inclined to believe that this property minimizes their variance. But, as it turns out, in general, it does not.

This educated guess holds for $n$ up to 183, but not beyond that number. At $n = 184$ there is at least one tree, presented in Figure 4.1, that, having 174 leaves at depth 8, nine at depth 7 and one leaf at depth 2, presents a Variance of depths of 0.2379 versus that of the maximally balanced tree with that same number of leaves, 0.2382. In the first part of this chapter, we study the characterization of the bifurcating trees that attain the maximum Variance of depths. And while we fail to give a complete characterization, we end up presenting quite interesting problems and regularities.

Finally, in the second part we are going to provide closed formulæ for the expected value of $V$ under two models for bifurcating phylogenetic trees: the Uniform and Yule models. Furthermore, closed expressions for the variance, under the Uniform model, of the Sackin and Cophenetic indices will be given, as well as their covariance —whereas, so far, only recursive formulæ for them [100, §2.5–2.7] and the asymptotic behaviour of the variance of the Sackin index [8] were known. We shall use this to find the variance of the distance between two leaves in $\mathbf{BinTree}_n$ under the Uniform model. Our actual contribution is the solution of the aforementioned recursive equations obtained by L. Rotger in her PhD Thesis [100], but, to ease the task of the reader, we also include derivations of those recursive equations based on the tools developed in this memoir (mainly, Lemmata 1.33, 1.34, and 4.25).

This chapter is organized as follows. In the first section, we shall study and find the maximum Variance of depth displayed by any tree with $n$ leaves, as well as the (only) tree that actually achieves it. The second section will be devoted to the quest for the minimum value: we shall, first, find a necessary condition for a tree to attain it, and then show that, contrary to our intuition, almost no number of leaves is such that the maximally balanced tree with that number of leaves minimizes the Variance of depths. However, no characterisation of this minimum value is given, leaving it as an open problem. Then, in the third section, we proceed to the computation of the expected value of $V$ under the Yule and Uniform model. After that, we dedicate one section to the computation of both the variance and covariance of the Sackin and Cophenetic indices under the Uniform model, in order to include some results proven as a by-product of the techniques introduced in the Preliminaries. We will end with a discussion where we shall present a number of some interesting open problems that have eluded the best of our efforts.

## 4.1 The maximum Variance of depths

We begin by showing that, as it was expected, the maximum Variance of depths is attained exactly at the caterpillars: they are the trees that present the widest range of depths and have leaves whose depth attain each value in that range only once, except for the cherry at the bottom.

We have already pointed out in Section 1.2.2 that the caterpillars are exactly the trees with maximum Sackin index and that value is $S(T_n^{\mathrm{cat}}) = \frac{(n-1)(n+2)}{2}$. We can easily

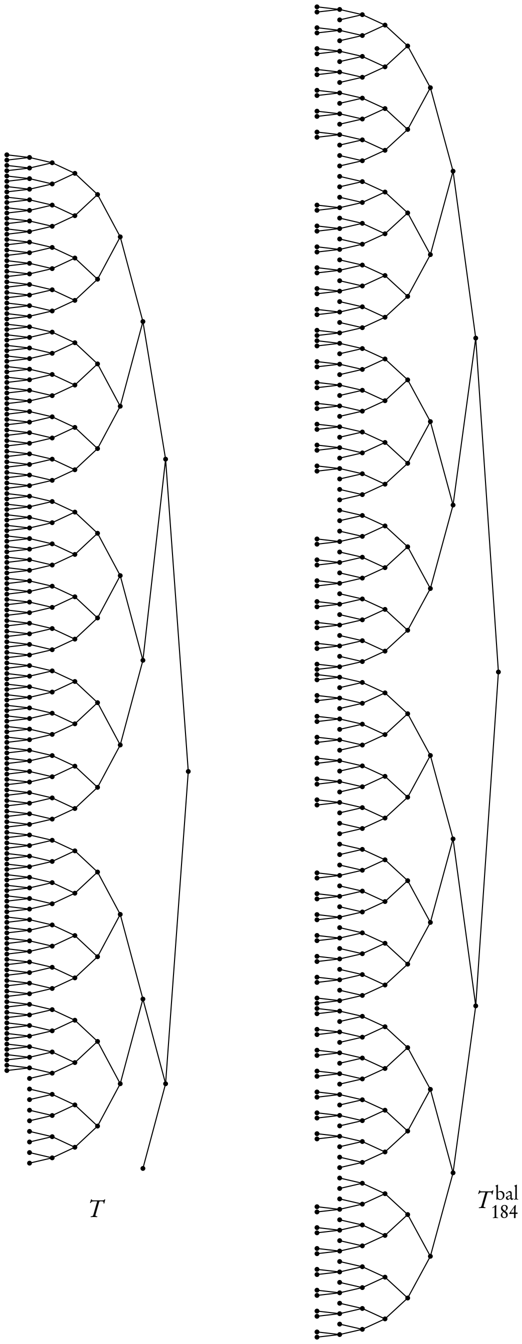


Figure 4.1: The leaves' depths of the left-hand side tree $T \in \mathbf{BinTree}_{184}$ have smaller variance than those of the right-hand side maximally balanced tree $T^{\mathrm{bal}}_{184}$.

see that, for any $n \in \mathbb{N}_{\geq 1}$,

$$\begin{aligned} V(T_n^{\mathrm{cat}}) &= \frac{1}{n} S^{(2)}(T_n^{\mathrm{cat}}) - \frac{1}{n^2} S(T_n^{\mathrm{cat}})^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n-1} i^2 + (n-1)^2\right) - \frac{1}{n^2}\left(\frac{(n-1)(n+2)}{2}\right)^2 \\ &= \frac{(n-1)(2n^2+5n-6)}{6n} - \frac{(n-1)^2(n+2)^2}{4n^2} \\ &= \frac{(n-1)(n-2)(n^2+3n-6)}{12n^2}. \end{aligned}$$

In order to prove that this is indeed the maximum value the Variance of depths can attain, and that it is reached exactly by the caterpillars, we need first to prove a series of lemmata describing the behaviour of $V(T)$ when we remove a deepest leaf from $T$.

Given a tree $T \in \mathbf{Tree}_n$, we shall henceforth denote by $x_1, \ldots x_n$ its leaves ordered in non-decreasing order of depth; i.e., such that $\delta(x_i) \leq \delta(x_{i+1})$ for $i \in \{1, \ldots, n-1\}$; we set $d_i = \delta(x_i)$. Since there is always at least one $k$-fan at maximum depth, for some $k \in \mathbb{N}_{\geq 2}$, it is always true that $d_{n-1} = d_n$, and hence we shall always assume, without loss of generality, that $x_{n-1}$ and $x_n$ are siblings.

**Lemma 4.1.** *Let $T \in \mathbf{Tree}_n$ be a tree with two leaves of maximum depth forming a cherry. Let $T' \in \mathbf{Tree}_{n-1}$ be the tree obtained by removing both leaves in this cherry, so that the root of the cherry becomes a leaf. Then,*

$$n \cdot V(T') = n \cdot V(T) - \frac{n}{n-1}(\delta(T) - \overline{S}(T) + 1)^2 + 2.$$

*Proof.* In this proof, we shall denote $\delta(T)$ by $\delta$, and we shall suppose, without loss of generality, that $x_{n-1}$ and $x_n$ are not only sibling, but form the cherry at the bottom. Let $T'$ be the tree described in the statement of the lemma; we shall still call this new leaf $x_{n-1}$; cf. Fig. 4.2. Thus,

$$\Delta(T') = \{d_1, \ldots, d_{n-2}, d_{n-1} - 1\}.$$

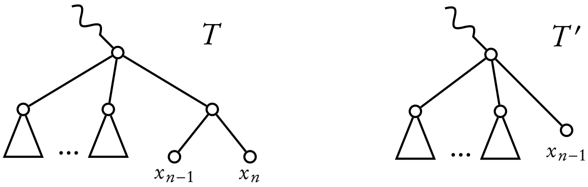


Figure 4.2: A tree $T$ with a cherry at the bottom and the tree $T'$ obtained by removing this cherry.

Then,

$$\overline{S}(T') = \frac{\sum_{i=1}^{n-1} d_i - 1}{n-1} = \frac{n\overline{S}(T) - \delta - 1}{n-1} = \overline{S}(T) - \frac{\delta - \overline{S}(T) + 1}{n-1}.$$

Thus,

$$
\begin{aligned}
n\cdot V(T') &= \sum_{i=1}^{n-2}(d_i-\overline{S}(T'))^2+(d_{n-1}-1-\overline{S}(T'))^2\\
&= \sum_{i=1}^{n-2}(d_i-\overline{S}(T'))^2+(d_{n-1}-\overline{S}(T'))^2-2(d_{n-1}-\overline{S}(T'))+1\\
&= \sum_{i=1}^{n}(d_i-\overline{S}(T'))^2-2(d_{n-1}-\overline{S}(T'))+1-(d_n-\overline{S}(T'))^2\\
&\text{(\textit{porque de ``dame un gato'' a ``toma un gato'' van dos gatos})}\\
&= \sum_{i=1}^{n}(d_i-\overline{S}(T'))^2-2(\delta-\overline{S}(T'))+1-(\delta-\overline{S}(T'))^2\\
&\text{(since } d_{n-1}=d_n=\delta)\\
&= \sum_{i=1}^{n}(d_i-\overline{S}(T'))^2-(\delta-\overline{S}(T')+1)^2+2\\
&= \sum_{i=1}^{n}\left(d_i-\overline{S}(T)+\frac{\delta-\overline{S}(T)+1}{n-1}\right)^2-\left(\delta-\overline{S}(T)+\frac{\delta-\overline{S}(T)+1}{n-1}+1\right)^2+2\\
&= \sum_{i=1}^{n}(d_i-\overline{S}(T))^2+2\left(\frac{\delta-\overline{S}(T)+1}{n-1}\right)\sum_{i=1}^{n}(d_i-\overline{S}(T))\\
&\qquad +n\left(\frac{\delta-\overline{S}(T)+1}{n-1}\right)^2-\left(\frac{n(\delta-\overline{S}(T)+1)}{n-1}\right)^2+2\\
&= n\cdot V(T)-\frac{n}{n-1}(\delta-\overline{S}(T)+1)^2+2,
\end{aligned}
$$

because $\sum_{i=1}^{n}(d_i-\overline{S}(T))=0$. □

**Lemma 4.2.** *Let $T\in\mathbf{Tree}_n$ be a tree with $k$ leaves, $k\geq 3$, forming a $k$-fan at the bottom. Let $T'\in\mathbf{Tree}_{n-1}$ be the tree obtained by removing one leaf from this $k$-fan. Then,*

$$n\cdot V(T')=n\cdot V(T)-\frac{n}{n-1}(\delta(T)-\overline{S}(T))^2.$$

*Proof.* Again, we shall denote $\delta(T)$ by $\delta$. Suppose that $(x_{n-k+1},\dots,x_{n-1},x_n)\in L(T)^k$ are the leaves forming the $k$-fan at the bottom; since all of them have depth $\delta$, and since $k\geq 3$, then the removal of, say, $x_n$, does not alter the depth of the other leaves, and therefore

$$\Delta(T')=\{d_1,\dots,d_{n-2},d_{n-1}\}.$$

Hence,

$$\overline{S}(T')=\frac{\sum_{i=1}^{n}d_i-d_n}{n-1}=\frac{n\overline{S}(T)-\delta}{n-1}=\overline{S}(T)-\frac{\delta-\overline{S}(T)}{n-1}.$$

Computing $n\cdot V(T')$ in terms of $n\cdot V(T)$, as we did in the proof of the previous lemma, we get

$$\begin{aligned}
n\cdot V(T') &= \sum_{i=1}^{n-1}(d_i-\overline{S}(T'))^2 = \sum_{i=1}^{n}(d_i-\overline{S}(T'))^2-(d_n-\overline{S}(T'))^2\\
&= \sum_{i=1}^{n}\left(d_i-\overline{S}(T)+\frac{\delta-\overline{S}(T)}{n-1}\right)^2-\left(\delta-\overline{S}(T)+\frac{\delta-\overline{S}(T)}{n-1}\right)^2\\
&= \sum_{i=1}^{n}(d_i-\overline{S}(T))^2+2\left(\frac{\delta-\overline{S}(T)}{n-1}\right)\sum_{i=1}^{n}(d_i-\overline{S}(T))+n\left(\frac{\delta-\overline{S}(T)}{n-1}\right)^2\\
&\quad-\left(\frac{n(\delta-\overline{S}(T))}{n-1}\right)^2\\
&= n\cdot V(T)-\frac{n}{n-1}(\delta-\overline{S}(T))^2.
\end{aligned}$$

□

**Lemma 4.3.** *Let $T\in\mathbf{Tree}_n$ be a tree with two leaves of maximum depth forming a cherry. Then,*

$$\delta(T)-\overline{S}(T)\leq\frac{(n-1)(n-2)}{2n},$$

*and the equality holds if, and only if, $T=T_n^{\text{cat}}$.*

*Proof.* The fact that the equality holds when $T=T_n^{\text{cat}}$ is already known to us, since we have already computed $S(T_n^{\text{cat}})$ and we know $\delta(T_n^{\text{cat}})$ and then it is simply a matter of doing the computations. Now, we need to see that, for any $T\in\mathbf{Tree}_n\setminus\{T_n^{\text{cat}}\}$ with two leaves of maximum depth forming a cherry,

$$\delta(T)-\overline{S}(T)<\frac{(n-1)(n-2)}{2n}.$$

We proceed by induction on $n$. The cases when $n\in\{1,2\}$ are obvious, since $|\mathbf{Tree}_1|=|\mathbf{Tree}_2|=1$, and for $n=3$ there are only two trees, the caterpillar and the star, and $\delta(T_n^{\text{star}})-\overline{S}(T_n^{\text{star}})$ is always 0. Let us assume now that $n\geq 4$ and the result holds up to $n-1$ leaves. In order to ease the notations, we shall set

$$\Psi(T)=\delta(T)-\overline{S}(T)+1.$$

Let $T$ be such a tree, and assume that leaves $x_{n-1}$ and $x_n$ form a cherry of maximum depth. Let $T'\in\mathbf{Tree}_{n-1}$ be the tree obtained by removing both leaves in this cherry and replacing them by their parent, which, again, we shall call $x_{n-1}$. Then, either $x_{n-1},x_n$ were the only leaves at depth $\delta(T)$, and then $\delta(T')=\delta(T)-1$, or they were not, and then $\delta(T')=\delta(T)$. On the other hand, in the proof of Lemma 4.1, we have proved that

$$\delta(T)-\overline{S}(T')+1=\delta(T)-\overline{S}(T)+\frac{\delta(T)-\overline{S}(T)+1}{n-1}+1=\frac{n}{n-1}\Psi(T).$$

Hence,

$$\begin{aligned}\Psi(T) &= \frac{n-1}{n}(\delta(T) - \overline{S}(T') + 1)\\ &\leq \frac{n-1}{n}(\delta(T') + 1 - \overline{S}(T') + 1) = \frac{n-1}{n}(\Psi(T') + 1),\end{aligned}$$

and the equality holds only when $x_{n-1}, x_n$ are the only leaves at depth $\delta(T)$. Now, two cases arise:

(a) If $T'$ contains a cherry at maximum depth, we lie under the induction hypothesis, and hence we obtain that

$$\Psi(T') = \delta(T') - \overline{S}(T) + 1 \leq \frac{(n-2)(n-3)}{2(n-1)} + 1,$$

with equality if, and only if, $T' = T^{\mathrm{cat}}_{n-1}$. Then,

$$\begin{aligned}\delta(T) - \overline{S}(T) = \Psi(T) - 1 &\leq \frac{n-1}{n}(\Psi(T') + 1) - 1\\ &\leq \frac{n-1}{n}\left(\frac{(n-2)(n-3)}{2(n-1)} + 2\right) - 1 = \frac{(n-1)(n-2)}{2n}.\end{aligned}$$

Notice that the equality is reached if, and only if, $T' = T^{\mathrm{cat}}_{n-1}$ and $x_{n-1}, x_n$ are the only leaves at depth $\delta(T)$. Therefore, since $T$ is obtained by adding a cherry to a leaf of depth $\delta(T')$ in $T'$, in this case we would have that $T = T^{\mathrm{cat}}_n$.

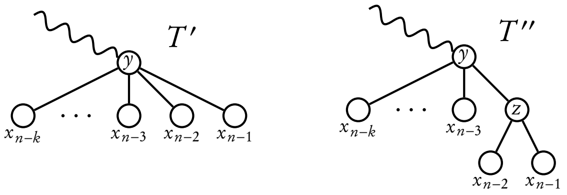


Figure 4.3: The trees $T'$ and $T''$ in case (b) of the proof of Lemma 4.3.

(b) Now, assume that $T'$ has no cherry at maximum depth, and hence that $x_{n-1}$ is part of a $k$-fan, with $k \geq 3$. Let $y \in \mathring{V}(T')$ be the root of the $k$-fan. Suppose, without loss of generality, that $x_{n-2}$ is a sibling of $x_{n-1}$, and let $T'' \in \mathbf{Tree}_{n-1}$ be the tree obtained by replacing the edges $(y, x_{n-2})$ and $(y, x_{n-1})$ by a new edge $(y, z)$ to a new internal node, $z \in \mathring{V}(T'')$, and let $x_{n-2}, x_{n-1}$ pend from $z$ forming a cherry at maximum depth $\delta(T'') = \delta(T') + 1$; see Fig. 4.3. Since both leaves increase their depth in one unit with respect to $T'$, and all other leaves in $T'$ maintain their depths, $\overline{S}(T'') = \overline{S}(T') + \frac{2}{n-1}$. Hence,

$$\begin{aligned}\Psi(T'') = \delta(T'') - \overline{S}(T'') + 1 &= \delta(T') + 1 - \overline{S}(T') - \frac{2}{n-1} + 1\\ &= \Psi(T') + \frac{n-3}{n-2} > \Psi(T'),\end{aligned}$$

as we are assuming that $n \geq 4$. Then,

$$\Psi(T) \leq \frac{n-1}{n}(\Psi(T')+1) < \frac{n-1}{n}(\Psi(T'')+1).$$

Now, notice that $T''$ has a cherry at maximum depth and $n-1$ leaves, and so we can apply to it our induction hypothesis:

$$\delta(T'') - \overline{S}(T'') \leq \frac{(n-2)(n-3)}{2(n-1)}.$$

Hence, we can proceed as we did in the previouse case:

$$\begin{aligned}\delta(T) - \overline{S}(T) = \Psi(T) - 1 < \frac{n-1}{n}(\Psi(T'')+1) - 1 \leq \frac{n-1}{n}\left(\frac{(n-2)(n-3)}{2(n-1)} + 2\right) - 1\\ = \frac{(n-1)(n-2)}{2n}.\end{aligned}$$

Note that in this case $T$ can never be a caterpillar and that the inequality is strict.

Thus concludes the proof. □

Finally, thanks to these three lemmata, we are in a position to state and prove the main theorem of this section.

**Theorem 4.4.** *For any $n \in \mathbb{N}_{\geq 1}$, the maximum value of $V$ on $\mathbf{Tree}_n$ is attained exactly at the caterpillar $T_n^{\text{cat}}$.*

*Proof.* We shall prove the result for $n \cdot V$, which will trivially imply it for $V$. Hence, we shall prove by induction on $n$, that for any $T \in \mathbf{Tree}_n$, $n \cdot V(T) \leq n \cdot V(T_n^{\text{cat}})$, and that the equality holds if, and only if, $T = T_n^{\text{cat}}$.

The cases when $n \in \{1,2\}$ are obvious, since there is only one tree at each of them; for $n = 3$ the result is also trivial, because there are only two trees: the caterpillar and the star, and the latter has Variance of depths 0 while the former has not. Therefore, suppose that $n \geq 4$ and the result holds up to $n-1$ leaves.

Let $T \in \mathbf{Tree}_n$; we must distinguish two cases.

(a) Suppose that $T$ has a $k$-fan at the bottom, and suppose without loss of generality that it is formed by the leaves $x_{n-k+1}, \ldots, x_{n-1}, x_n$, and let $y$ be their parent. As we did in the proof of Lemma 4.3, let $T' \in \mathbf{Tree}_n$ be the tree obtained by replacing the edges $(y, x_{n-1})$ and $(y, x_n)$ in $T$ by a new edge $(y, z)$ to a new internal node $z \in \mathring{V}(T')$, and then let $x_{n-1}, x_n$ be the cherry rooted at $z$. As we argued in the proof of the aforementioned lemma, $\delta(T') = \delta(T) + 1$ and $\overline{S}(T') = \overline{S}(T) + \frac{2}{n}$.
Now, consider $T'' \in \mathbf{Tree}_{n-1}$ to be a tree obtained by $T$ by removing leaf $x_n$ or, equivalently, from $T'$ by removing the cherry at $(x_{n-1}, x_n)$ and renaming $z = x_{n-1}$. Therefore, by Lemmata 4.1 and 4.2,

$$\begin{aligned}n \cdot V(T'') &= n \cdot V(T') - \frac{n}{n-1}(\delta(T') - \overline{S}(T') + 1)^2 + 2\\ n \cdot V(T'') &= n \cdot V(T) - \frac{n}{n-1}(\delta(T) - \overline{S}(T))^2,\end{aligned}$$

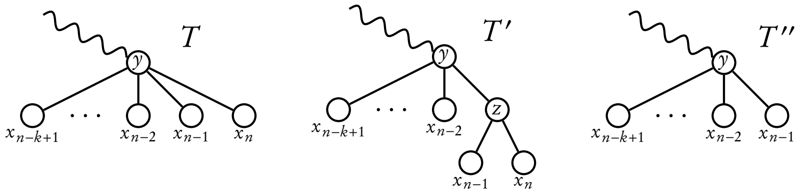


Figure 4.4: The trees $T$, $T'$, and $T''$ in case (a) of the proof of Theorem 4.4.

and so

$$\begin{aligned}
n \cdot V(T) &= n \cdot V(T') - \frac{n}{n-1}(\delta(T') - \overline{S}(T') + 1)^2 + \frac{n}{n-1}(\delta(T) - \overline{S}(T))^2 + 2 \\
&= n \cdot V(T') + 2 - \frac{n}{n-1}\left(\delta(T) - \overline{S}(T) - \frac{2}{n} + 2\right)^2 + \frac{n}{n-1}(\delta(T) - \overline{S}(T))^2 \\
&= n \cdot V(T') + 2 - 4\left(\delta(T) - \overline{S}(T) + \frac{n-1}{n}\right) < n \cdot V(T'),
\end{aligned}$$

where this last inequality holds because $n \geq 4$ and $\delta(T) \geq \overline{S}(T)$. Therefore, if $T$ contains a $k$-fan at the bottom, with $k \geq 3$, it does not present the maximum Variance of depths, and thence we consider the next case.

(b) Suppose now that $T$ contains a cherry at the bottom and assume, without loss of generality, that it is formed by the leaves $x_{n-1}$ and $x_n$. Let $T' \in \mathbf{Tree}_{n-1}$ be the tree obtained by removing this cherry and naming its root $x_{n-1}$. Then, by Lemma 4.1,

$$\begin{aligned}
n \cdot V(T) &= n \cdot V(T') + \frac{n}{n-1}(\delta(T) - \overline{S}(T) + 1)^2 - 2 \\
&\leq n \cdot V(T') + \frac{n}{n-1}\left(\frac{(n-1)(n-2)}{2n} + 1\right)^2 - 2 \\
&\text{(by Lemma 4.3; the equality holds exactly when } T = T_n^{\mathrm{cat}}) \\
&\leq n \cdot V(T_{n-1}^{\mathrm{cat}}) + \frac{n}{n-1}\left(\frac{(n-1)(n-2)}{2n} + 1\right)^2 - 2 \\
&\text{(by the induction hypothesis; the equality holds exactly when } T' = T_{n-1}^{\mathrm{cat}}) \\
&= n \cdot V(T_{n-1}^{\mathrm{cat}}) + \frac{n}{n-1}\left(\delta(T_n^{\mathrm{cat}}) - \overline{S}(T_n^{\mathrm{cat}}) + 1\right)^2 - 2 \\
&= n \cdot V(T_n^{\mathrm{cat}}),
\end{aligned}$$

again by Lemma 4.1. Now, both inequalities hold if, and only if, $T = T_n^{\mathrm{cat}}$; therefore, in any other case, $T$ does not present the maximum Variance of depths. Which is what we wanted to prove.

□

We end this section by reminding the reader that Theorem 4.4 is akin to our intuition, since this is one of the properties that a balance measure is expected to satisfy. Indeed, the unicity of the "least balanced tree" is satisfied by all the balance indices for trees reviewed in this report, and provides a curious insight of the fact that, regardless of how the notion of "balance" is defined, we all agree in what we consider *not* to be balanced.

## 4.2 The minimum Variance of depths

For any tree $T \in \mathbf{Tree}$, $V(T) \geq 0$, and the equality holds if, and only if, all the leaves in $T$ have the exact same depth. This gives a clear and easy answer to the problem of finding the minimum value of the Variance of depths in $\mathbf{Tree}_n$ and all the trees that attain it: the minimum value is 0, and all trees presenting the same depth for every leaf reach it —for example, the star. This, too, satisfies our natural understanding of balance: stars are the trees with the most automorphisms, while caterpillars are those with the least. However, notice that $V$ achieves the value 0 not only at the stars: indeed, any taxonomic tree would, too, present no variance of depths whatsoever. Thus, this index is utterly useless in order to assess the balance of a taxonomic tree, just as its relative the Sackin index was.

However, things get convoluted when the search for the minimum Variance of depths in the domain of bifurcating trees is attempted. As we have already mentioned, there were several reasons to expect this minimum to happen at the maximally balanced trees. To begin with, such trees present the minimum value of the Colless [22], the Sackin [39], the Cophenetic [85], and the Quartet (Chapter 5 in this thesis) indices. Furthermore, the maximally balanced trees, and those depth-equivalent to them, are the only family of trees whose leaves differ in at most one unity, thus presenting a fairly small variation (cf. Theorem 1.19). And finally, the fact that, whenever the number of leaves is a power of 2, the maximally balanced tree presents Variance of depths 0, also contributed to support this idea.

In fact, this is seldom the case. This educated guess holds up to 183 leaves, but not beyond that. For $n = 184$, there exists a tree, presented in Figure 4.1, with lower Variance of depths than that of $T^{\text{bal}}_{184}$. In fact, in this chapter, we shall prove that, as $n$ tends to infinity, the fraction of values for which the minimum Variance of depths is attained by a maximally balanced tree tends to zero.

### 4.2.1 A necessary condition

Now, let us define a family of trees that generalizes the maximally balanced trees. We say that a bifurcating tree $T \in \mathbf{BinTree}_n$ is *of type* $T^{\mathbf{l}}_n$, for some vector $\mathbf{l} = (l_1, \ldots, l_j) \in \mathbb{N}^j_{\geq 2}$, with $j \in \mathbb{N}$ and $2 \leq l_1 < \cdots < l_j \leq \delta(T) - 2$, if it has exactly one leaf at depth $\delta(T) - l_i$ for every $i \in \{1, \ldots, j\}$, and the rest of the leaves at depths $\delta(T) - 1$ and $\delta(T)$. There may be many trees of a given type $T^{\mathbf{l}}_n$, but by the next lemma all of them have the same set of depths —thus, the same variance of depths. Thus, we will often commit the abuse of language of writing $V(T^{\mathbf{l}}_n)$ to mean $V(T)$ for some $T$ of the form $T^{\mathbf{l}}_n$. If $\mathbf{l} = \emptyset$, i.e., $j = 0$, then the trees of type $T^{\emptyset}_n$ are, by Theorem 1.19, $T^{\text{bal}}_n$ and the trees depth-equivalent to it, that is, those trees achieving the minimum Sackin index in $\mathbf{BinTree}_n$.

Given a tree $T \in \mathbf{BinTree}_n$, let us call $p_1(T)$ and $p_0(T)$ its number of leaves of depths $\delta(T) - 1$ and $\delta(T)$, respectively.

**Lemma 4.5.** *Let $n = 2^m + k \in \mathbb{N}$, with $m = \lfloor \log_2(n) \rfloor$. For every tree $T \in$ **BinTree**$_n$ of type $T_n^{\mathbf{l}}$, with $\mathbf{l} = (l_1, \dots, l_j)$, we have:*

*(i) If $k + \sum_{i=1}^{j}(2^{l_i} - 1) = 0$, then $p_1(T) = 0$ and the tree is fully symmetric and $\delta(T) = m$.*

*(ii) If $0 < k + \sum_{i=1}^{j}(2^{l_i} - 1) \leq 2^m$, then $p_1(T) = 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1)$ and $\delta(T) = m + 1$.*

*(iii) If $k + \frac{1}{2}\sum_{i=1}^{j}(2^{l_i} - 2) > 2^m$, then $p_1(T) = 3 \cdot 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1)$ and $\delta(T) = m + 2$.*

*(iv) If $k + \frac{1}{2}\sum_{i=1}^{j}(2^{l_i} - 2) \leq 2^m < k + \sum_{i=1}^{j}(2^{l_i} - 1)$, then there does not exist any tree $T$ of type $T_n^{\mathbf{l}}$.*

*Proof.* Let $T \in$ **BinTree**$_n$ be of type $T_n^{\mathbf{l}}$, with $\mathbf{l} = (l_1, \dots, l_j)$, and set $\delta = \delta(T)$ and $p_1 = p_1(T)$.

Suppose that $j = 0$, and thence that $T$ has only leaves of depth $\delta - 1$ and $\delta$. If $k = 0$, then $n = 2^m$ and hence the tree is fully symmetric, which proves *(i)*. If $k \geq 1$, then $T$ is depth-equivalent to a maximally balanced tree and hence the thesis of *(ii)* holds in this case.

Suppose henceforth that $j \geq 1$. Notice that, in order to "complete" $T$ to a fully symmetric tree with $2^\delta$ leaves, we must append a cherry to each leaf with depth $\delta - 1$, but also a fully symmetric tree of $2^{l_i}$ leaves to any leaf of depth $\delta - l_i$. This implies that

$$2^\delta = n + p_1 + \sum_{i=1}^{j}(2^{l_i} - 1), \tag{4.2}$$

because we add one leaf in the first case and $2^{l_i} - 1$ leaves in the second. Since $j \geq 1$ by assumption, $n < 2^\delta$ and so $m \leq \delta - 1$. On the other hand, as $p_1 < n < 2^{m+1}$,

$$2^\delta = n + p_1 + \sum_{i=1}^{j}(2^{l_i} - 1) < 2n + \sum_{i=2}^{\delta-2}(2^i - 1) < 2^{m+2} + 2^{\delta-1}, \tag{4.3}$$

and hence $2^{\delta-1} < 2^{m+2}$. All in all, this implies that $\delta \in \{m+1, m+2\}$. Therefore, we distinguish these two cases:

- Suppose that $\delta = m + 1$. Then, by Equation (4.2),

$$p_1 = 2^{m+1} - 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1) = 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1) \geq 0.$$

Hence, in this case, $k + \sum_{i=1}^{j}(2^{l-1} - 1) \leq 2^m$.

- Now suppose that $\delta = m + 2$; in this case we have

$$p_1 = 3 \cdot 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1),$$

and since we know that $T$ contains at least two leaves of depth $\delta$ and exactly $j$ leaves of depth smaller than $\delta - 1$, $p_1 \leq n - j - 2$:

$$p_1 = 3 \cdot 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1) \leq 2^m + k - j - 2,$$

or, equivalently,

$$2^{m+1} \leq 2k + \sum_{i=1}^{j}(2^{l_i} - 2) - 2,$$

which is to say that $k + \frac{1}{2}\sum_{i=1}^{j}(2^{l_i} - 2) > 2^m$.

This completes the proof of *(ii)* and *(iii)*. Finally, since we have covered all possible cases, $k + \frac{1}{2}\sum_{i=1}^{j}(2^{l_i} - 2) \leq 2^m < k + \sum_{i=1}^{j}(2^{l_i} - 1)$ can never be the case, which proves *(iv)*. □

**Remark 4.6.** Notice that if we allowed $l_j$ to be $\delta(T) - 1$, then Equation (4.3) could be

$$2^\delta = n + p_1 + \sum_{i=1}^{j}(2^{l_i} - 1) < 2n + \sum_{i=1}^{\delta-1}(2^i - 1) < 2^{m+2} + 2^\delta,$$

and no information would be gathered from it. In particular, notice that in this case the caterpillar could also be considered a tree of type $T_n^1$, but it does not satisfy the thesis in the previous lemma since its depth is larger than $m + 2$ when $n \geq 6$.

**Lemma 4.7.** *If $T$ is a tree of type $T_n^1$, then*

$$V(T) = \frac{1}{n^2}\left(n\Big(p_1(T) + \sum_{i=1}^{j} l_i^2\Big) - \Big(p_1(T) + \sum_{i=1}^{j} l_i\Big)^2\right).$$

*Proof.* Set $\delta = \delta(T)$, $p_0 = p_0(T)$, and $p_1 = p_1(T)$; since $n = p_0 + p_1 + j$,

$$\overline{S}(T) = \frac{p_0\delta + p_1(\delta - 1) + \sum_{i=1}^{j}(\delta - l_i)}{n} = \delta - \frac{p_1 + \sum_{i=1}^{j} l_i}{n}.$$

Thus,

$$\begin{aligned}
n \cdot V(T) &= p_0(\delta - \overline{S}(T))^2 + p_1(\delta - 1 - \overline{S}(T))^2 + \sum_{i=1}^{j}(\delta - l_i - \overline{S}(T))^2 \\
&= p_0\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\right)^2 + p_1\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n} - 1\right)^2 + \sum_{i=1}^{j}\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n} - l_i\right)^2 \\
&= p_0\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\right)^2 + p_1\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\right)^2 - 2p_1\frac{p_1 + \sum_{i=1}^{j} l_i}{n} + p_1 \\
&\quad + j\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\right)^2 - 2\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\sum_{i=1}^{j} l_i + \sum_{i=1}^{j} l_i^2 \\
&= n\left(\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\right)^2 - 2\frac{p_1 + \sum_{i=1}^{j} l_i}{n}\left(p_1 + \sum_{i=1}^{j} l_i\right) + p_1 + \sum_{i=1}^{j} l_i^2 \\
&= p_1 + \sum_{i=1}^{j} l_i^2 - \frac{(p_1 + \sum_{i=1}^{j} l_i)^2}{n},
\end{aligned}$$

as we claimed. □

Combining the last two lemmata, we obtain that, if $n = 2^m + k$ with $m = \lfloor \log_2(n) \rfloor$, then, for every tree $T$ of type $T_n^{\mathbf{l}}$:

- If $\sum_{i=1}^{j}(2^{l_i} - 1) \leq 2^m - k$

$$V(T) = \frac{2^m - k - \sum\limits_{i=1}^{j}(2^{l_i} - l_i^2 - 1)}{n} - \frac{\left(2^m - k - \sum\limits_{i=1}^{j}(2^{l_i} - l_i - 1)\right)^2}{n^2}. \tag{4.4}$$

- If $\sum_{i=1}^{j}(2^{l_i - 1} - 1) > 2^m - k$

$$V(T) = \frac{3 \cdot 2^m - k - \sum\limits_{i=1}^{j}(2^{l_i} - l_i^2 - 1)}{n} - \frac{\left(3 \cdot 2^m - k - \sum\limits_{i=1}^{j}(2^{l_i} - l_i - 1)\right)^2}{n^2}. \tag{4.5}$$

In particular, when $j = 0$, the formula (4.4) applies and we obtain

$$V(T_n^{\mathrm{bal}}) = V(T_n^{\emptyset}) = \frac{2^m - k}{n} - \frac{(2^m - k)^2}{n^2} = \frac{2k(2^m - k)}{n^2}. \tag{4.6}$$

This identity could have also be obtained directly from the fact that $T_n^{\mathrm{bal}}$ has $2^m - k$ leaves of depth $m$ and $2k$ leaves of depth $m + 1$ by Theorems 1.18 and 1.19.

Let us return to the problem of finding the trees in $\mathbf{BinTree}_n$ with minimum Variance of depths. Since $|\mathbf{BinTree}_n| = 1$ for $n \in \{1, 2, 3\}$, it will suffice to consider $n \geq 4$; but, since in $\mathbf{BinTree}_4$ there are only two trees and one of them has Variance of depths 0 while the other has not, it will actually suffice to consider $n \geq 5$. We begin by showing, as a necessary condition, that any tree $T \in \mathbf{BinTree}_n$ must be of some type $T_n^{\mathbf{l}}$, with some further restrictions on $\mathbf{l}$, in order for the minimum of the Variance of depths to be reached by it. The proof is quite long, and we have opted to present it as the sum of some partial results, which we shall now prove.

**Lemma 4.8.** *Let $n \geq 5$. If $T \in \mathbf{BinTree}_n$ has a leaf of depth 1, then $V(T)$ is not minimum in $\mathbf{BinTree}_n$.*

*Proof.* Let $T \in \mathbf{BinTree}_n$ be a tree with a leaf of depth 1, so that $T$ is the root join of a tree $T_0 \in \mathbf{BinTree}_{n-1}$ and a leaf $x_1$. Consider the sequence of depths of $T$ to be $d_1 = 1, d_2, \ldots, d_{n-1}, d_n$ in non-decreasing order, so that the sequence of depths of $T_0$ in non-decreasing order is $d_2 - 1, \ldots, d_{n-1} - 1, d_n - 1$. Let $T' \in \mathbf{BinTree}_n$ be a tree obtained from $T_0$ by replacing a leaf of the smallest depth, $d_2 - 1$, in $T_0$ by a cherry of depth $d_2$, so that the sequence of depths of $T'$ is $d_2, d_2, d_3 - 1, \ldots, d_{n-1} - 1, d_n - 1$. Then,

$$\begin{aligned}
\overline{S}(T') &= \frac{\sum_{i=2}^{n} d_i + d_2 - (n-2)}{n} = \frac{\left(1 + \sum_{i=2}^{n} d_i\right) + d_2 - n + 1}{n} \\
&= \frac{n\overline{S}(T) + d_2 - n + 1}{n} = \overline{S}(T) - 1 + \frac{d_2 + 1}{n},
\end{aligned}$$

and hence

$$\begin{aligned}
n \cdot V(T') &= 2(d_2 - \overline{S}(T'))^2 + \sum_{i=3}^{n}(d_i - 1 - \overline{S}(T'))^2 \\
&= 2\left(d_2 - \overline{S}(T) + 1 - \frac{d_2+1}{n}\right)^2 + \sum_{i=3}^{n}\left(d_i - \overline{S}(T) - \frac{d_2+1}{n}\right)^2 \\
&= 2(d_2 - \overline{S}(T))^2 + 4(d_2 - \overline{S}(T))\left(1 - \frac{d_2+1}{n}\right) + 2\left(1 - \frac{d_2+1}{n}\right)^2 \\
&\quad + \sum_{i=3}^{n}(d_i - \overline{S}(T))^2 - 2\left(\frac{d_2+1}{n}\right)\sum_{i=3}^{n}(d_i - \overline{S}(T)) + (n-2)\left(\frac{d_2+1}{n}\right)^2 \\
&= n \cdot V(T) - (1 - \overline{S}(T))^2 + (d_2 - \overline{S}(T))^2 + 4(d_2 - \overline{S}(T)) \\
&\quad - 2(d_2 - \overline{S}(T))\left(\frac{d_2+1}{n}\right) - 2\left(\frac{d_2+1}{n}\right)\sum_{i=2}^{n}(d_i - \overline{S}(T)) \\
&\quad + 2\left(1 - \frac{d_2+1}{n}\right)^2 + (n-2)\left(\frac{d_2+1}{n}\right)^2 \\
&\quad \text{(since } n \cdot V(T) = (1 - \overline{S}(T))^2 + \sum_{i=2}^{n}(d_i - \overline{S}(T))^2\text{)} \\
&= n \cdot V(T) + (d_2 - \overline{S}(T) + 2)^2 - 4 - (1 - \overline{S}(T))^2 \\
&\quad - 2(d_2 - \overline{S}(T))\left(\frac{d_2+1}{n}\right) + 2\left(\frac{d_2+1}{n}\right)(1 - \overline{S}(T)) \\
&\quad + 2 - 4\left(\frac{d_2+1}{n}\right) + n\left(\frac{d_2+1}{n}\right)^2 \\
&\quad \text{(as } 1 + \sum_{i=2}^{n} d_i = n\overline{S}(T)\text{)} \\
&= n \cdot V(T) + (d_2 - \overline{S}(T) + 2)^2 - (1 - \overline{S}(T))^2 - 2 - \frac{(d_2+1)^2}{n} \\
&= n \cdot V(T) + (d_2+1)\left(d_2 + 3 - 2\overline{S}(T) - \frac{d_2+1}{n}\right) - 2.
\end{aligned}$$

Therefore, if

$$(d_2+1)\left(d_2 + 3 - 2\overline{S}(T) - \frac{d_2+1}{n}\right) - 2 < 0$$

then $n \cdot V(T') < n \cdot V(T)$, and so it cannot be minimum in $\mathbf{BinTree}_n$. Let us prove that this is indeed the case whenever $n \geq 5$. Let us rephrase it in terms of $T_0$:

$$\overline{S}(T) = \frac{1 + \sum_{i=2}^{n} d_i}{n} = \frac{n + \sum_{i=2}^{n}(d_i - 1)}{n} = 1 + \frac{(n-1)\overline{S}(T_0)}{n}.$$

Therefore,

$$\begin{aligned}(d_2+1)\left(d_2+3-2\overline{S}(T)-\frac{d_2+1}{n}\right)-2\\ &= (d_2+1)\left(d_2+1-\frac{2(n-1)\overline{S}(T_0)}{n}-\frac{d_2+1}{n}\right)-2\\ &= \frac{n-1}{n}(d_2+1)(d_2+1-2\overline{S}(T_0))-2.\end{aligned}$$

Now, notice that $\overline{S}(T_0) \geq d_2 - 1$, since $d_2 - 1$ is the smallest depth of a leaf in $T_0$; now, if $\overline{S}(T_0) \geq 2$, then the whole expression would be negative, thus proving the result. But, since $n \geq 5$, we know that all the leaves of $T_0$ but at most two of them have depth greater or equal than 3; and if it contains two leaves of depth smaller than 3, they have depths 1 and 2 or depths 2 and 2. Thus,

$$\overline{S}(T_0) \geq \frac{1+2+3(n-3)}{n-1} = \frac{3n-6}{n-1} > 2 \quad \text{if } n \geq 5.$$

Therefore, the result holds. □

**Lemma 4.9.** *Let $T \in$ **BinTree**$_n$ be a bifurcating tree containing a leaf of depth $d < \delta(T)$, and $T_d \in$ **BinTree**$_n$ be the tree obtained by removing a cherry of depth $\delta(T)$ and replacing a leaf of depth $d$ by a cherry of depth $d+1$. Then,*

$$n \cdot V(T_d) = n \cdot V(T) - \left(\frac{\delta(T)-d-1}{n}\right)(n(\delta(T)+d+3-2\overline{S}(T))+\delta(T)-d-1).$$

*Proof.* Let $T \in$ **BinTree**$_n$ be such a tree. With the usual notations, given before Lemma 4.1 (page 129), consider that $d = d_j$ for some $j \leq n-2$, and without loss of generality that the pair of leaves removed are $x_{n-1}$ and $x_n$. Then, the possibly unordered set of depths of $T_d$ is

$$\Delta(T_d) = \{d_1, \ldots, d_{j-1}, d_j+1, d_j+1, d_{j+1}, \ldots, d_{n-2}, d_{n-1}\};$$

we can thence compute the mean depth of $T_d$ as follows:

$$\overline{S}(T_d) = \frac{\sum\limits_{i=1}^{n-2} d_i + d_j + 2 + d_{n-1} - 1}{n} = \frac{\sum\limits_{i=1}^{n} d_i + d + 1 - d_n}{n} = \overline{S}(T) - \frac{\delta - d - 1}{n} \tag{4.7}$$

and therefore

$$
\begin{aligned}
n\cdot V(T_d) &= \sum_{i\in\{1,\dots,n-2\}\setminus\{j\}} (d_i-\overline{S}(T_d))^2+2(d+1-\overline{S}(T_d))^2+(d_{n-1}-1-\overline{S}(T_d))^2\\
&=\sum_{i=1}^{n}(d_i-\overline{S}(T_d))^2+(d+1-\overline{S}(T_d))^2-(d_n-\overline{S}(T_d))^2\\
&\qquad+2(d-\overline{S}(T_d))+1-2(d_{n-1}-\overline{S}(T_d))+1\\
&=\sum_{i=1}^{n}(d_i-\overline{S}(T_d))^2+(d+2-\overline{S}(T_d))^2-(\delta+1-\overline{S}(T_d))^2\\
&\text{(since } d_{n-1}=d_n=\delta)\\
&=\sum_{i=1}^{n}(d_i-\overline{S}(T_d))^2-(\delta+d+3-2\overline{S}(T_d))(\delta-d-1)\\
&=\sum_{i=1}^{n}\left(d_i-\overline{S}(T)+\frac{\delta-d-1}{n}\right)^2\\
&\qquad-\left(\delta+d+3-2\overline{S}(T)+2\frac{\delta-d-1}{n}\right)(\delta-d-1)\\
&=\sum_{i=1}^{n}(d_i-\overline{S}(T))^2+2\frac{\delta-d-1}{n}\sum_{i=1}^{n}(d_i-\overline{S}(T))+n\left(\frac{\delta-d-1}{n}\right)^2\\
&\qquad-(\delta+d+3-2\overline{S}(T))(\delta-d-1)-2\frac{(\delta-d-1)^2}{n}\\
&=n\cdot V(T)-\frac{(\delta-d-1)^2}{n}-(\delta-d-1)(\delta+d+3-2\overline{S}(T))\\
&=n\cdot V(T)-\frac{\delta-d-1}{n}(n(\delta+d+3-2\overline{S}(T))+\delta-d-1),
\end{aligned}
$$

as we claimed. □

**Corollary 4.10.** *If $T\in$ **BinTree**$_n$ has the minimum value of $V$ and it contains some leaf of depth $\delta(T)-l$, where $l>1$, then*

$$l\geq 3+2\frac{n(\delta(T)-\overline{S}(T))+1}{n-1}.$$

*In particular, $T$ does not contain leaves of depths $\delta(T)-2$ or $\delta(T)-3$.*

*Proof.* If $T\in$ **BinTree**$_n$ is such that $V(T)$ is minimum on **BinTree**$_n$ and it contains some leaf of depth $d=\delta(T)-l<\delta(T)-1$, then, by the last lemma,

$$(\delta(T)-d-1)\left(n(\delta(T)+d+3-2\overline{S}(T))+\delta(T)-d-1\right)\leq 0.$$

Since $\delta(T)-d-1=l-1>0$, this is equivalent to

$$n(\delta(T)+d+3-2\overline{S}(T))+\delta(T)-d-1\leq 0.$$

Therefore, by replacing $d$ by $\delta(T) - l$, and then solving for $l$,

$$l \geq \frac{2n(\delta(T) - \overline{S}(T)) + 3n - 1}{n - 1} = 3 + 2\frac{n(\delta(T) - \overline{S}(T)) + 1}{n - 1} > 3,$$

as we claimed. □

**Corollary 4.11.** *Let $T \in$ **BinTree**$_n$ be a bifurcating tree that has a cherry of depth $d < \delta(T)$, and $T_d^* \in$ **BinTree**$_n$ the tree obtained by removing such a cherry, leaving in its place its root as a leaf of depth $d - 1$, and then replacing a leaf of depth $\delta(T)$ by a cherry of depth $\delta(T) + 1$. Then,*

$$n \cdot V(T_d^*) = n \cdot V(T) + \frac{\delta(T) - d + 1}{n}(n(\delta(T) + d + 3 - 2\overline{S}(T)) - (\delta(T) - d + 1)).$$

*Proof.* Using the same notations as in Lemma 4.9, we have that $T = (T_d^*)_{d-1}$. Therefore, by Equation (4.7),

$$\overline{S}(T) = \overline{S}((T_d^*)_{d-1}) = \overline{S}(T_d^*) - \frac{\delta(T_d^*) - (d-1) - 1}{n} = \overline{S}(T_d^*) - \frac{\delta(T) + 1 - d}{n},$$

and, again by Lemma 4.9,

$$\begin{aligned}
n \cdot V(T) &= n \cdot V((T_d^*)_{d-1}) \\
&= n \cdot V(T_d^*) - \frac{\delta(T_d^*) - (d-1) - 1}{n}\Bigg(n(\delta(T_d^*) + (d-1) + 3 - 2\overline{S}(T_d^*)) \\
&\quad + \delta(T_d^*) - (d-1) - 1\Bigg) \\
&= n \cdot V(T_d^*) - \frac{\delta(T) + 1 - d}{n}\Bigg(n\Big(\delta(T) + 1 + d + 2 \\
&\quad - 2\overline{S}(T) - 2\frac{\delta(T) + 1 - d}{n}\Big) + \delta(T) + 1 - d\Bigg) \\
&= n \cdot V(T_d^*) - \frac{\delta(T) - d + 1}{n}\Bigg(n(\delta(T) + d + 3 - 2\overline{S}(T)) - (\delta(T) - d + 1)\Bigg),
\end{aligned}$$

and hence the result holds. □

The previous lemmata have an interesting corollary that will now be proved. It says, roughly, that any tree $T \in$ **BinTree**$_n$ that attains the minimum Variance of depths must be such that it has at most one leaf at each depth $d \in \{2, \ldots, \delta(T) - 4\}$ —and, as we shall see anon, even $\delta(T) - 4$ is not an allowed depth. The rest of the leaves must have depths $\delta(T) - 1$ or $\delta(T)$: this is, indeed, a necessary condition that all trees reaching the minimum of the Variance of depths must comply.

**Corollary 4.12.** *If $T \in$ **BinTree**$_n$ contains two leaves of the same depth $d \leq \delta(T) - 2$, then $V(T)$ is not minimum in* **BinTree**$_n$.

*Proof.* Let $T \in \mathbf{BinTree}_n$ and assume it has two leaves $y_0, y_1 \in L(T)$ at the same depth, $d < \delta(T) - 1$. We distinguish two cases.

- If $\delta(T) + d + 3 - 2\overline{S}(T) \geq 0$ then, with the notations of Lemma 4.9,

$$\begin{aligned} n \cdot V(T_d) &= n \cdot V(T) - (\delta(T) - d - 1)\left(\delta(T) + d + 3 - 2\overline{S}(T) + \frac{\delta(T) - d - 1}{n}\right) \\ &< n \cdot V(T), \end{aligned}$$

  and therefore we have found a tree with $n$ leaves and strictly less variance of the leaves' depths.

- Assume now that $\delta(T) + d + 3 - 2\overline{S}(T) < 0$. If both leaves $y_0$ and $y_1$ belong to a cherry, then, by Corollary 4.11:

$$\begin{aligned} n \cdot V(T_d^*) &= n \cdot V(T) + (\delta(T) - d + 1)\left(\delta(T) + d + 3 - 2\overline{S}(T) - \frac{\delta(T) - d + 1}{n}\right) \\ &< n \cdot V(T), \end{aligned}$$

  and thus $V(T)$ cannot be minimal.
  Finally, suppose that $y_0$ and $y_1$ belong not to any cherry, and let $v_0, v_1 \in \mathring{V}(T)$ be their respective parents and $z_0, z_1 \in \mathring{V}(T)$ their respective siblings, which we suppose not to be leaves (for otherwise, if $z_0 \in L(T)$ we could consider $y_1 = z_0$); see Figure 4.5. Let $T'$ be the tree obtained by interchanging $z_0$ with $y_1$; i.e., the tree obtained by removing the edges $(v_0, z_0)$ and $(v_1, y_1)$ and replacing them by edges $(v_0, y_1)$ and $(v_1, z_0)$. Therefore, $\Delta(T) = \Delta(T')$, and so their respective Variances of depths are the same. Now, $T'$ has a cherry at depth $d$ and thus, as we have just seen, $V(T) = V(T')$ cannot be minimal, either.

□

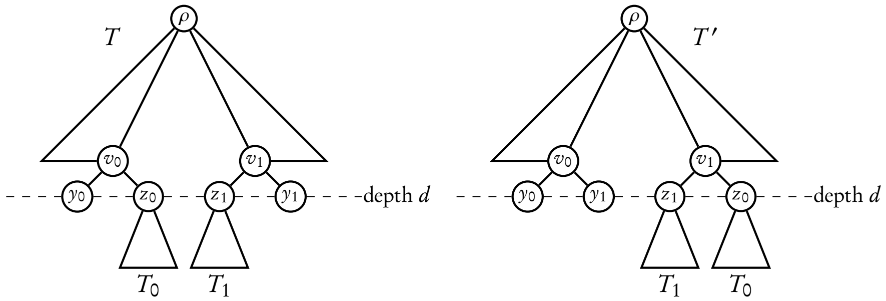


Figure 4.5: The depth-equivalent trees $T$ and $T'$ appearing in the last paragraph of the proof of Corollary 4.12.

In summary, thus far we have proved that if $T$ achieves the minimum Variance of depths in $\mathbf{BinTree}_n$, then it must be of some type $T_n^{\mathbf{l}}$ with $4 \leq l_1 < \cdots < l_j$. Finally, we state another result that will allow us to completely draw the necessary condition that the trees attaining the minimum Variance of depths must satisfy.

**Lemma 4.13.** *Let $T \in$ **BinTree**$_n$ such that $V(T)$ is minimum for that number of leaves. Then, it has no leaf at depth $\delta(T) - 4$.*

*Proof.* Let $T \in$ **BinTree**$_n$ be a tree such that $V(T)$ is minimum in **BinTree**$_n$. Set $\delta = \delta(T)$, $p_0 = p_0(T)$, $p_1 = p_1(T)$, and $n = 2^m + k$, for some $m, k \in \mathbb{N}$ such that $k < 2^m$. As we have just mentioned, by the previous lemmata, we already know that $T$ is of some type $T_n^{\mathbf{l}}$, for some $\mathbf{l} = (l_1, \dots, l_j) \in \mathbb{N}^j$ with $j \geq 0$ and $4 \leq l_1 < \cdots < l_j \leq \delta - 2$. We want to prove that it has no leaf at depth $\delta(T) - 4$, that is, that $l_1 \geq 5$.

By definition,

$$\overline{S}(T) = \delta - \frac{p_1 + \sum_{i=1}^{j} l_i}{n}.$$

Therefore, if $p_0 \leq \frac{n}{2}$, by Corollary 4.10,

$$\begin{aligned} l_1 &\geq 3 + 2\frac{n(\delta - \overline{S}(T)) + 1}{n-1} = 3 + 2\frac{p_1 + \sum_{i=1}^{j} l_i + 1}{n-1} \\ &> 3 + 2\frac{p_1 + j}{n-1} = 3 + 2\frac{n - p_0}{n-1} \geq 3 + \frac{n}{n-1} > 4, \end{aligned}$$

and hence $l_1 \geq 5$.

Suppose now that $p_0 > \frac{n}{2}$. In this case, we can also assume that $n \geq 32$. Indeed, for if $n < 32$ then, by Lemma 4.5, $\delta \leq 6$, and since $4 \leq l_1 \leq \delta - 2 \leq 4$, it must happen precisely that $\delta = 6$, $m = 4$, $l_1 = 4$ and $j = 1$. As $n = 2^4 + k$ with $k \leq 15$ and since $\delta = 4 + 2$, we are under the assumption *(iii)* in the aforementioned lemma. Thus, $p_1 = 3 \cdot 2^4 - k - (2^4 - 1) = 33 - k$. Therefore,

$$p_0 = n - p_1 - 1 = 2^4 + k - 33 + k - 1 = -18 + 2k \leq 8 + \frac{k}{2} = \frac{n}{2},$$

because $k \leq 15$, which contradicts the assumption that $p_0 > \frac{n}{2}$.

Therefore, $n \geq 32$, and hence $T$ contains at least 16 leaves of depth $\delta$. Assume now that $l_1 = 4$, that is, that $T$ has a leaf $x \in L(T)$ of depth $\delta - 4$. We shall prove that in this case, $V(T)$ is not minimal, either. Let $y \in \mathring{V}(T)$ be the sibling of $x$ (which cannot be a leaf by Corollary 4.12) and $z \in \mathring{V}(T)$ their common parent. Since $T$ does not contain leaves of depths $\delta - 3$ nor $\delta - 2$, all the leaves that descend from $y$ must be of depth $\delta - 1$ or $\delta$. By pruning and regrafting cherries at maximum depth, in a similar way as we did in the proof of Corollary 4.12, we can ensure that $T_y = T_{16}^{\mathrm{bal}}$.

Then, by, first, removing both arcs $(z, x)$ and $(z, y)$ and rooting $T_y$ at $z$ and, then, placing $x$ as sibling of a leaf in $T_z$ (as shown in Figure 4.6), we define a tree $T'$.

Now, it is easy to see that $T' = T_n^{\mathbf{l}'}$, with $\mathbf{l}' = (l_2, \dots, l_j)$: $\delta(T') = \delta$; it has two leaves of depth $\delta$ (the new cherry in $T_z'$); fifteen leaves of depth $\delta - 1$ in $T_z'$ plus the remaining $p_1$ leaves of depth $\delta - 1$ that survive from $T$; and then one leaf of each depth $\delta - l_2, \dots, \delta - l_j$. A simple computation shows that that $T'$ has lesser variance than $T$

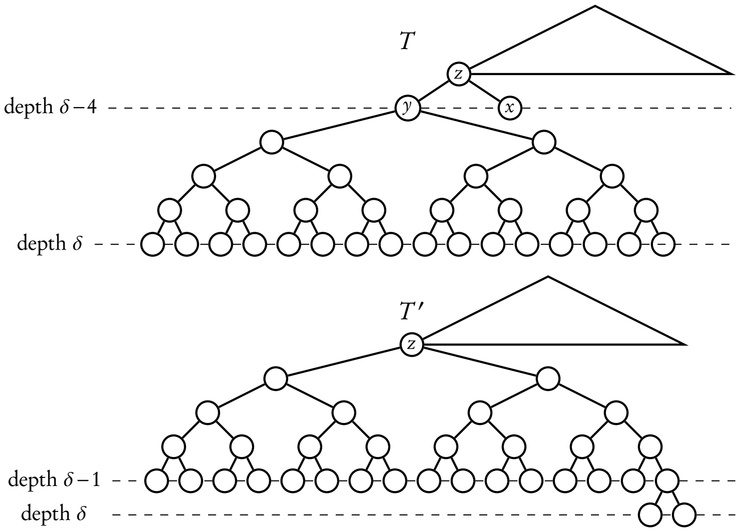


Figure 4.6: The trees $T$ and $T'$ appearing in the last part of the proof of Lemma 4.13.

has. Indeed, by Lemma 4.7 (and recall that we are assuming that $l_1 = 4$),

$$n^2V(T) = n\left(p_1 + 16 + \sum_{i=2}^{j} l_i^2\right) - \left(p_1 + 4 + \sum_{i=2}^{j} l_i\right)^2$$
$$n^2V(T') = n\left(p_1 + 15 + \sum_{i=2}^{j} l_i^2\right) - \left(p_1 + 15 + \sum_{i=2}^{j} l_i\right)^2,$$

and then $V(T') < V(T)$. So, when $p_0 > \frac{n}{2}$ we also conclude that if $T$ has the minimum value of $V$ in $\mathbf{BinTree}_n$, then $l_1 \geq 5$. □

Thus ends the proof of the necessary condition we have found every tree achieving the minimum Variance of depths must satisfy.

**Theorem 4.14.** *If $T \in \mathbf{BinTree}_n$ has the minimum value of $V$, then $T$ is of type $T_n^1$ with $5 \leq l_1 < \cdots < l_j \leq \delta(T) - 2$.*

This result allows us to give Algorithm 5 below that, given $n \in \mathbb{N}_{\geq 2}$, finds in time $O(n\log(n))$ all types of trees $T_n^1$ with minimum Variance of depths, by simply searching for them in the space of all trees satisfying this necessary condition, combined with Lemma 4.5 *(iv)* to discard vectors $\mathbf{l}$, and computing efficiently their Variances of depths using Equations 4.4 or 4.5.

This algorithm runs in time $O(n\log_2(n))$. Indeed, for it *parcours* for each $j \in \{0,\ldots,m-4\}$ the set

$$\left\{(l_1,\ldots,l_j) \in \mathbb{N}^j : 5 \leq l_1 < \cdots < l_j \leq m\right\}.$$

This set has cardinality $\binom{m-4}{j}$, where $m = \lfloor \log_2(n) \rfloor$, and on each vector $(l_1,\ldots,l_j)$ it performs $O(j)$ operations to check the conditions in line 7 and to compute the corresponding Variance of depths in line 8, and hence the total number of operations is of

**Algorithm 5:** MinVarDepths

```
Input   : $n \in \mathbb{N}_{\geq 1}$
Output: $V(n)$ minimum value of $V$ in $\mathbf{BinTree}_n$, and $\mathbf{L}$ the set of vectors $\mathbf{l}$ of
        depths such that $V(T_n^{\mathbf{l}}) = V(n)$.
compute $m = \lfloor \log_2(n) \rfloor$ and $k = n - 2^m$;
if $k = 0$ then
    $V(n) = 0$ and $\mathbf{L} = \{\emptyset\}$;
else
    $V(n) = \frac{2k(2^m - k)}{n^2}$ and $\mathbf{L} = \{\emptyset\}$;
    for $\mathbf{l}_0 \in \mathbb{N}^j$ with $1 \leq j \leq m$ and $5 \leq l_1 < \cdots < l_j \leq m$ do
        if $\left(\sum_{i=1}^{j}(2^{l_i} - 1) \leq 2^m - k\right)$ or $\left(\sum_{i=1}^{j}(2^{l_i - 1} - 1) > 2^m - k\right)$ then
            compute $v_0 = V(T_n^{\mathbf{l}_0})$ using Equations (4.4) or (4.5);
            if $v_0 < V(n)$ then
                $V(n) = v_0$ and $\mathbf{L} = \{\mathbf{l}_0\}$;
            else
                if $v_0 = V(n)$ then
                    $\mathbf{L} = \mathbf{L} \cup \mathbf{l}_0$;
                end
            end
        end
    end
end
return $(V(n), \mathbf{L})$;
```

the order of

$$O\left(\sum_{j=1}^{m-4}\binom{m-4}{j}j\right) = O\left((m-4)\sum_{j=1}^{m-4}\binom{m-5}{j-1}\right)$$

$$= O\left((m-4)\sum_{j=0}^{m-5}\binom{m-5}{j}\right) = O(2^{m-5}m) = O(n\log_2(n)).$$

Another naïve $O(n\log_2(n))$ algorithm can be given in order to find, given $n \in \mathbb{N}_{\geq 2}$, the multisets of depths of all bifurcating trees with $n$ leaves and minimum Variance of depths. It also relies on Theorem 4.14 and, instead of using Lemma 4.5 to compute the values of $p_0(T_n^{\mathbf{l}})$ and $p_1(T_n^{\mathbf{l}})$, it obtains them through a more direct approach applying Lemma 4.15 below, which is basically due to A. Mir-Fuentes [87, Thm. 1]. The best that can be said about this algorithm is that it is not asymptotically worse than Algorithm 5 is.

Given a tree $T$ of depth $\delta$, for any $i \in \mathbb{N}$ we define $p_i(T)$ to be the number of leaves of depth $\delta - i$, and we shall now drop the $T$ to ease the notations. We can then consider the vector $\mathbf{p}_T = (p_0, p_1, \ldots, p_{\delta-1}) \in \mathbb{N}^\delta$. We say that a given vector $\mathbf{v} \in \mathbb{N}^d$ *represents* a tree when there exists some tree $T$ of depth $\delta$ such that $\mathbf{v} = \mathbf{p}_T$.

**Lemma 4.15.** *A sequence* $\mathbf{p} = (p_0, p_1, \dots, p_{\delta-1}) \in \mathbb{N}^\delta$ *represents a bifurcating tree of depth* $\delta$ *if, and only if, it satisfies the following conditions:*

*i)* $p_0 \neq 0$.

*ii)* $\displaystyle\sum_{i=0}^{\delta-1} p_i 2^i = 2^\delta$.

*iii)* $\displaystyle\frac{\sum_{i=0}^{j} p_i 2^i}{2^j} \in 2\mathbb{N}$ *for all* $j \in \{0, \dots, \delta-1\}$.

*Proof.* We prove it by induction on $\delta$. When $\delta = 1$, the only vector represented by a bifurcating tree of depth $\delta$ is (2), and the conditions *(i)–(iii)* in the statement say exactly that $\mathbf{p} = (2)$. So, assume that $\delta \geq 2$ and that the equivalence is true for vectors of length smaller or equal than $\delta - 1$. We apply induction to both implications separately.

Consider first the "only if" implication. Given a bifurcating tree $T$ with $n$ leaves and depth $\delta$, $p_0 \in 2\mathbb{N}_{\geq 1}$ because it is the number of leaves of depth $\delta$, and these appear in pairs forming cherries. This proves *(i)* as well as the case $j = 0$ of *(iii)*. Now, we can build a bifurcating tree $T'$ of depth $\delta - 1$ by pruning all the leaves of maximum depth. In this tree,

$$\mathbf{p}_{T'} = (p'_0, \dots, p'_{\delta-2})$$

with

$$p'_0 = p_1 + \frac{p_0}{2}$$

and $p'_i = p_{i+1}$ for any $i \in \{1, \dots, \delta-2\}$. By the induction hypothesis, $T'$ satisfies *(ii)*, that is,

$$2^{\delta-1} = \sum_{i=0}^{\delta-2} p'_i 2^i = \frac{p_0}{2} + p_1 + \sum_{j=2}^{\delta-1} p_i 2^{i-1},$$

which implies *(ii)* for $T$,

$$2^\delta = p_0 + 2p_1 + 2\sum_{j=2}^{\delta-1} p_i 2^{i-1} = \sum_{i=0}^{\delta-1} p_i 2^i,$$

and $T'$ satisfies *(iii)*, that is

$$\frac{\sum_{i=0}^{j} p'_i 2^i}{2^j} = \frac{1}{2^j}\left(\frac{p_0}{2} + p_1 + \sum_{j=2}^{j+1} p_i 2^{i-1}\right) = \frac{\sum_{i=0}^{j+1} p_i 2^i}{2^{j+1}} \in 2\mathbb{N}$$

for all $j \in \{0, \dots, \delta-2\}$, which is equivalent to *(iii)* for $T$ and $j \in \{1, \dots, \delta-1\}$.

Now let us pursue the other direction, and let us be given a vector $\mathbf{p} = (p_0, \dots, p_{\delta-1})$ that satisfies *(i)–(iii)*. Let us consider briefly the case when $\delta = 2$: equation *(ii)* is now $p_0 + 2p_1 = 4$. Now, since $p_0 \neq 0$ by *(i)* and is even by *(iii)*, it can be either 2 or 4. If $p_0 = 2$, then $p_1 = 1$ and we have the only bifurcating tree with three leaves; when $p_0 = 4$, $p_1 = 0$ and we have the fully symmetric tree of four leaves. These are the only possible bifurcating trees of depth 2.

Now suppose that $\delta \geq 3$. As in the case $\delta = 2$, *(i)* and *(iii)* imply that $p_0$ is non-zero and even. We consider the vector

$$\mathbf{p}' = (p'_0, p'_1, \ldots, p'_{\delta-2}) = \left(p_1 + \frac{p_0}{2}, p_2 \ldots, p_{\delta-1}\right) \in \mathbb{N}^{\delta-1}.$$

It is clear that $p'_0 \in \mathbb{N}_{\geq 1}$, because $p_0 \in 2\mathbb{N}_{\geq 1}$. We will now show that it satisfies the other two conditions:

$$\frac{\sum_{i=0}^{j} p'_i 2^i}{2^j} = \frac{\sum_{i=1}^{j+1} p_i 2^{i-1} + \frac{p_0}{2}}{2^j} = \frac{\sum_{i=1}^{j+1} p_i 2^i + p_0}{2^{j+1}} = \frac{\sum_{i=0}^{j+1} p_i 2^i}{2^{j+1}} \in 2\mathbb{N}$$

up to $j = \delta - 2$, and

$$\sum_{i=0}^{\delta-1} p_i 2^i = 2^\delta \Longrightarrow 2^{\delta-1} = \sum_{i=0}^{\delta-1} p_i 2^{i-1} = \frac{p_0}{2} + p_1 + \sum_{i=1}^{\delta-2} p_{i+1} 2^i = \sum_{i=0}^{\delta-2} p'_i 2^i.$$

Now, by the induction hypothesis, we know that $\mathbf{p}'$ represents a bifurcating tree $T'$. Therefore, $\mathbf{p}$ represents a bifurcating tree $T$ constructed by choosing $\frac{p_0}{2}$ leaves with depth $\delta - 1$ in $T'$ and replacing them by cherries, which yields $p_0$ leaves at depth $\delta$, the remaining $p_1$ leaves at depth $\delta - 1$ and $p_i(T) = p'_{i-1}(T') = p'_i = p_i$ for every $i \geq 2$ because the depth of $T$ is one edge larger than that of $T'$. This concludes the proof. □

Thanks to this lemma, we can solve our problem using some basic linear-algebraic techniques. First of all, note that, for any $j \in \{0, \ldots, \delta - 1\}$, condition $\frac{\sum_{i=0}^{j} p_i 2^i}{2^j} \in 2\mathbb{N}$ can be re-written as $\sum_{i=0}^{j} 2^i p_i = 2^{j+1} k_j$, with $k_j \in \mathbb{N}$ and $k_{\delta-1} = 1$ since $\sum_{i=0}^{\delta-1} p_i 2^i = 2^\delta$. Since $p_0 \in 2\mathbb{N}$ is implied by $p_0 + 2p_1 \in 2^2\mathbb{N}$ and we can fix the number $n$ of leaves with the identity $\sum_{i=0}^{\delta-1} p_i = n$, this leads to the following equation:

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 2 & 0 & \ldots & 0 & 0 \\ 1 & 2 & 2^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & 2^2 & \ldots & 2^{\delta-2} & 0 \\ 1 & 2 & 2^2 & \ldots & 2^{\delta-2} & 2^{\delta-1} \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{\delta-2} \\ p_{\delta-1} \end{pmatrix} = \begin{pmatrix} n \\ 2^2 k_1 \\ 2^3 k_2 \\ \vdots \\ 2^{\delta-1} k_{\delta-2} \\ 2^\delta \end{pmatrix}.$$

Now, this is equivalent to the following equation

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ 0 & 1 & -1 & \ldots & -1 & -1 \\ 0 & 0 & 2^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2^{\delta-2} & 0 \\ 0 & 0 & 0 & \ldots & 0 & 2^{\delta-1} \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{\delta-2} \\ p_{\delta-1} \end{pmatrix} = \begin{pmatrix} n \\ 2^2 k_1 - n \\ 2^3 k_2 - 2^2 k_1 \\ \vdots \\ 2^{\delta-1} k_{\delta-2} - 2^{\delta-2} k_{\delta-3} \\ 2^\delta - 2^{\delta-1} k_{\delta-2} \end{pmatrix}$$

which ensures the compatibility of the system and the uniqueness of its solution. Thus, for $j \in \{2, \ldots, \delta-1\}$, we have that $2^j p_j = 2^{j+1} k_j - 2^j k_{j-1}$, or equivalently, $p_j = 2k_j - k_{j-1}$. But since we know that any tree of minimum variance must be such that $p_j \in \{0, 1\}$, we know that $k_{j-1} \in \{2k_j - 1, 2k_j\}$.

This gives rise to $O(2^{\delta})$ systems, but since we also know that $\delta = O(\log(n))$, we have in fact a linear number of systems. We will argue that all of these can be solved in $O(\log(n))$.

Once all the $k_j$ are settled (and, with them, $p_2, \dots, p_{\delta-1}$ in time $O(\delta) = O(\log(n))$), only $p_0$ and $p_1$ remain to be computed. But these are

$$p_1 = 4k_1 + \textstyle\sum_{j=2}^{\delta-1} p_j - n \quad \text{and} \quad p_0 = n - \textstyle\sum_{j=1}^{\delta-1} p_j$$

which require $O(\delta) = O(\log(n))$ computations each one. Notice that, by construction, the solutions with $p_0 \geq 1$ of these systems represent bifurcating trees of depth $\delta$. Finally, having thus constructed the vector of frequency of leaves' depths, we can easily compute its variance in time $O(\delta) = O(\log(n))$. This gives a total time complexity of $O(n \log n)$.

### 4.2.2 Almost no maximally balanced tree has minimum Variance of depths

We are now in a position to give the main result of this chapter, namely that the fraction of natural numbers $n \geq 1$ smaller than a given $N \in \mathbb{N}$ such that the maximally balanced trees with $n$ leaves happen to achieve the minimum Variance of depths in $\mathbf{BinTree}_n$ tends to zero as $N$ tends to $\infty$. The proof is quite long, and it will take the integrity of this section. It relies on several claims that we will present as separate lemmata in the remaining of the section.

Consider $n \in \mathbb{N}_{\geq 1}$ written as $n = 2^m + k$, where $m = \lfloor \log_2(n) \rfloor$ and $0 \leq k < 2^m$. In this section, we shall only consider sequences $\mathbf{l}$ of the form $5 \leq l_1 < \cdots < l_j \leq m-1$ with $j \geq 1$ and $k \leq 2^m - \sum_{i=1}^{j}(2^{l_i} - 1)$, so that $V(T_n^{\mathbf{l}})$ satisfies Equation (4.4). Given any such sequence $\mathbf{l} \in \mathbb{N}^j$, set $A(\mathbf{l})$ and $B(\mathbf{l})$ to be

$$A(\mathbf{l}) = \textstyle\sum_{i=1}^{j}(2^{l_i} - l_i^2 - 1) \quad \text{and} \quad B(\mathbf{l}) = \textstyle\sum_{i=1}^{j}(2^{l_i} - l_i - 1).$$

With these notations, Equation (4.4) says that

$$V(T_n^{\mathbf{l}}) = \frac{1}{n^2}\left((2^m + k)(2^m - k - A(\mathbf{l})) - (2^m - k - B(\mathbf{l}))^2\right),$$

and we want to know whether the next expression is, or is not, greater than zero:

$$\begin{aligned} &n^2(V(T_n^{\mathrm{bal}}) - V(T_n^{\mathbf{l}})) \\ &\quad = 2k(2^m - k) - (2^m + k)(2^m - k - A(\mathbf{l})) + (2^m - k - B(\mathbf{l}))^2 \\ &\quad = k(A(\mathbf{l}) + 2B(\mathbf{l})) + 2^m(A(\mathbf{l}) - 2B(\mathbf{l})) + B(\mathbf{l})^2. \end{aligned}$$

Now, since $l_i \geq 5$ for all $i \in \{1, \dots, j\}$, $A(\mathbf{l}) + 2B(\mathbf{l}) = \sum_{i=1}^{j}(3 \cdot 2^{l_i} - l_i^2 - 2l_i - 3)$ is always strictly positive. Thus, $V(T_n^{\mathrm{bal}}) > V(T_n^{\mathbf{l}})$ if, and only if,

$$k > \frac{2^m(2B(\mathbf{l}) - A(\mathbf{l})) - B(\mathbf{l})^2}{A(\mathbf{l}) + 2B(\mathbf{l})}. \tag{4.8}$$

On the other hand, recall that we are assuming $k \leq 2^m - \sum_{i=1}^{j}(2^{l_i} - 1)$.

We are now going to consider first the particular case of tree types $T_n^{\mathbf{l}}$ such that $j = 1$; i.e., such that $\mathbf{l}$ has only one entry. Set $x = l_1 \in \{5, \dots, m-1\}$, so that $A(\mathbf{l}) = 2^x - x^2 - 1$ and $B(\mathbf{l}) = 2^x - x - 1$. Then, whenever $k$ belongs to the set

$$\bigcup_{x=5}^{m-1} \Big( \frac{2^m(2^x + x^2 - 2x - 1) - (2^x - x - 1)^2}{3 \cdot 2^x - x^2 - 2x - 3}, 2^m - 2^x + 1 \Big]$$

we deduce that $V(T_n^{\mathrm{bal}})$ is not minimal on $\mathbf{BinTree}_n$, since $V(T_n^{\mathrm{bal}}) > V(T_n^{(x)})$. To simplify the notations, we shall set

$$F_1(x) = \frac{2^m(2^x + x^2 - 2x - 1) - (2^x - x - 1)^2}{3 \cdot 2^x - x^2 - 2x - 3}, \quad G_1(x) = 2^m - 2^x + 1,$$

so that the union of intervals is rewritten as $\bigcup_{x=5}^{m-1}(F_1(x), G_1(x)]$. We shall prove that there exists an $m_0 \in \mathbb{N}$ such that, for any $m \geq m_0$,

$$\bigcup_{x=5}^{m-1}(F_1(x), G_1(x)] = (F_1(m-1), G_1(5)].$$

**Lemma 4.16.** *For every $m \in \mathbb{N}_{\geq 7}$ and every $x \in \{5, \dots, m-2\}$, $F_1(x+1) < F_1(x)$ and $G_1(x+1) < G_1(x)$.*

*Proof.* That $G_1$ strictly decreases is clear; let us see that $F_1$ also does: $F_1(x) > F_1(x+1)$ if, and only if,

$$\begin{aligned}
&\big(2^m(2^x + x^2 - 2x - 1) - (2^x - x - 1)^2\big)\big(3 \cdot 2^{x+1} - (x+1)^2 - 2(x+1) - 3\big) \\
&\quad > \big(2^m(2^{x+1} + (x+1)^2 - 2(x+1) - 1) - (2^{x+1} - (x+1) - 1)^2\big)\big(3 \cdot 2^x - x^2 - 2x - 3\big)
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
2^x\big(&6 \cdot 2^{2x} + 2^{m+2}x^2 - 3 \cdot 2^x x^2 - 3 \cdot 2^{m+2}x - 4 \cdot 2^x x - 18 \cdot 2^x \\
&+ 2x^3 + 3x^2 + 8x + 18\big) + 2^{m+2}(x^2 + 3x) - 4x - 6 > 0.
\end{aligned}$$

Now, since $x \leq m-2$,

$$\begin{aligned}
&2^x\big(6 \cdot 2^{2x} + 2^{m+2}x^2 - 3 \cdot 2^x x^2 - 3 \cdot 2^{m+2}x - 4 \cdot 2^x x - 18 \cdot 2^x \\
&\qquad + 2x^3 + 3x^2 + 8x + 18\big) + 2^{m+2}(x^2 + 3x) - 4x - 6 \\
&\geq 2^x\big(6 \cdot 2^{2x} + 2^{m+2}x^2 - 3 \cdot 2^{m-2}x^2 - 3 \cdot 2^{m+2}x - 4 \cdot 2^{m-2}x - 18 \cdot 2^{m-2} \\
&\qquad + 2x^3 + 3x^2 + 8x + 18\big) + 2^{m+2}(x^2 + 3x) - 4x - 6 \\
&= 2^{m-2}(13 \cdot 2^x x^2 - 52 \cdot 2^x x - 18 \cdot 2^x + 16x^2 + 48x) \\
&\qquad + (2^x(6 \cdot 2^{2x} + 2x^3 + 3x^2 + 8x + 18) - 4x - 6).
\end{aligned}$$

Finally, if $x \geq 5$, then both addends in this last expression are positive. □

**Lemma 4.17.** *For every $m \in \mathbb{N}_{\geq 9}$ and for every $x \in \{5, \dots, m-2\}$, $G_1(x+1) > F_1(x)$.*

*Proof.* We have that

$$
\begin{aligned}
&G_1(x+1) > F_1(x)\\
&\quad\iff (2^m - 2^{x+1} + 1)(3\cdot 2^x - x^2 - 2x - 3) > 2^m(2^x + x^2 - 2x - 1) - (2^x - x - 1)^2\\
&\quad\iff 2^{m+1}(2^x - x^2 - 1) - 2^x(5\cdot 2^x - 2x^2 - 2x - 7) - 2 > 0\\
&\quad\iff 2^{m-2}(3\cdot 2^x - 8x^2 - 8) + 2^x(5\cdot 2^{m-2} - 5\cdot 2^x + 2x^2 + 2x + 7) - 2 > 0.
\end{aligned}
$$

This last inequality holds whenever $m \geq 9$ and $5 \leq x \leq m-2$. Indeed, suppose that $m \geq 9$; we distinguish two cases. On the one hand, if $8 \leq x \leq m-2$ (and, hence, $m \geq 10$), then $3\cdot 2^x - 8x^2 - 8 > 0$, and

$$2^x(5\cdot 2^{m-2} - 5\cdot 2^x + 2x^2 + 2x + 7) \geq 2^x(2x^2 + 2x + 7) \geq 38656.$$

On the other hand, if $x \in \{5, 6, 7\}$,

$$
\begin{aligned}
&2^{m-2}(3\cdot 2^5 - 8\cdot 5^2 - 8) + 2^5(5\cdot 2^{m-2} - 5\cdot 2^5 + 2\cdot 5^2 + 2\cdot 5 + 7) - 2 = 12\cdot 2^m - 2978\\
&2^{m-2}(3\cdot 2^6 - 8\cdot 6^2 - 8) + 2^6(5\cdot 2^{m-2} - 5\cdot 2^6 + 2\cdot 6^2 + 2\cdot 6 + 7) - 2 = 6(9\cdot 2^m - 2443)\\
&2^{m-2}(3\cdot 2^7 - 8\cdot 7^2 - 8) + 2^7(5\cdot 2^{m-2} - 5\cdot 2^7 + 2\cdot 7^2 + 2\cdot 7 + 7) - 2 = 78(2\cdot 2^m - 855)
\end{aligned}
$$

and therefore they are all positive whenever $m \geq 9$. □

Thus, if $m \geq 9$, Lemma 4.16 implies that both the left-hand side and the right-hand side of the intervals in the union decrease with $x$, and Lemma 4.17, that the intersection of two consecutive intervals is not empty. Thus, if $m \geq 9$, we can write the union of intervals as

$$\bigcup_{x=5}^{m-1} (F_1(x), G_1(x)] = (F_1(m-1), G_1(5)].$$

In summary, thus far we have proved that:

> If $n = 2^m + k$ with $m \geq 9$ and $k \in \big(F_1(m-1), G_1(5)\big]$, then the minimum of $V$ on $\mathbf{BinTree}_n$ is not attained at $T_n^{\mathrm{bal}}$.

Let us now consider another family of trees of types $T_n^{\mathbf{l}}$: namely, those such that $j = m-5$ and $\mathbf{l} = (5, \ldots, x-1, x+1, \ldots, m-1)$, for some $x \in \{6, \ldots, m-2\}$. In this case,

$$
\begin{aligned}
A(\mathbf{l}) &= \sum_{i=5}^{m-1} (2^i - i^2 - 1) - 2^x + x^2 + 1 = 2^m - 2^x + x^2 - \frac{2m^3 - 3m^2 + 7m - 24}{6}\\
B(\mathbf{l}) &= \sum_{i=5}^{m-1} (2^i - i - 1) - 2^x + x + 1 = 2^m - 2^x + x - \frac{m^2 + m + 32}{2}.
\end{aligned}
$$

Therefore, if we set

$$
\begin{aligned}
F_2(x) &= \frac{2^m(2B(\mathbf{l}) - A(\mathbf{l})) - B(\mathbf{l})^2}{A(\mathbf{l}) + 2B(\mathbf{l})}\\
G_2(x) &= 2^m - \sum_{i=1}^{j} (2^{l_i} - 1)
\end{aligned}
$$

where, if we do the computations,

$$\begin{aligned}2^m(2B(1) - A(1)) - B(1)^2 &= \frac{1}{12}\Big(2^{m+1}(6\cdot 2^x + 2m^3 - 3m^2 + 7m - 6x^2 - 24)\\ &\qquad - 3(2^{x+1} - 2x + m^2 + m + 32)^2\Big)\\ A(1) + 2B(1) &= \frac{1}{6}\left(9\cdot 2^{m+1} - 9\cdot 2^{x+1} - 2m^3 - 3m^2 - 13m + 6x^2 + 12x - 168\right)\end{aligned}$$

and

$$2^m - \sum_{i=1}^{j}(2^{l_i} - 1) = 2^m - \left(\sum_{j=5}^{m-1}(2^j - 1) - (2^x - 1)\right) = 2^x + m + 26,$$

then, by Inequation (4.8), whenever $k$ belongs to

$$\bigcup_{x=6}^{m-2}(F_2(x), G_2(x)],$$

$V(T_n^{\mathrm{bal}})$ will be not minimal on $\mathbf{BinTree}_n$. We shall now prove a result similar in spirit to the one proven above: namely, that there exists an $m_1 \in \mathbb{N}$ such that, if $m \geq m_1$, then

$$\bigcup_{x=\lceil 3\log_2(m)\rceil}^{m-2}(F_2(x), G_2(x)] = \big(F_2(\lceil 3\log_2(m)\rceil), G_2(m-2)\big]. \tag{4.9}$$

The fact that $G_2$ is obviously increasing on $x$ leads us to prove only that function $F_2$ is.

**Lemma 4.18.** *There exists an $m_2 \in \mathbb{N}$ such that, for any $m \geq m_2$ and every $x \in \{6, \ldots, m-3\}$, $F_2(x+1) > F_2(x)$.*

*Proof.* Notice, to begin the proof, that, by grouping the terms with $x$, the denominator of $F_2(x)$ can be written as

$$\begin{aligned}&2\left(9\cdot 2^{m+1} - 9\cdot 2^{x+1} - 2m^3 - 3m^2 - 13m + 6x^2 + 12x - 168\right)\\ &\quad = 2\left(9\cdot 2^{m+1} - 2m^3 - 3m^2 - 13m - 168 - 6(3\cdot 2^x - x^2 - 2x)\right).\end{aligned}$$

and this expression is decreasing for $x \in [6, m-2]$ because the function $x \mapsto 3\cdot 2^x - x^2 - 2x$ is increasing on $[6, \infty)$. Consider now the numerator; its derivative, up to a factor of 12, is

$$2^x\ln(2)(2^m - 2^{x+1} - m^2 - m + 2x - 32) + 2^{x+1} - 2^{m+1}x + m^2 + m - 2x + 32.$$

Assume now that $x \leq m - 2$. Then,

$$\begin{aligned}
&2^x \ln(2)(2^m - 2^{x+1} - m^2 - m + 2x - 32)\\
&\qquad + 2^{x+1} - 2^{m+1}x + m^2 + m - 2x + 32\\
&\geq 2^x \ln(2)(2^m - 2^{m-1} - m^2 - m + 2x - 32)\\
&\qquad + 2^{x+1} - 2^{m+1}x + m^2 + m - 2x + 32\\
&= 2^x \ln(2)(2^{m-1} - m^2 - m + 2x - 32)\\
&\qquad + 2^{x+1} - 2^{m+1}x + m^2 + m - 2x + 32\\
&= 2^x \ln(2)(2^{m-3} - m^2 - m + 2x - 32)\\
&\qquad + 2^{m-3}(3 \cdot 2^x \ln(2) - 16x) + m^2 + m + 2^{x+1} - 2x + 32.
\end{aligned}$$

Now, if $x \geq 6$, then $3 \cdot 2^x \ln(2) - 16x > 0$ and $2^{x+1} - 2x > 0$ and, for $m$ large enough, $2^{m-3} - m^2 - m + 2x - 32 > 0$ too. Thus, if $m$ is large enough, the derivative is positive and so the numerator increases, while the denominator decreases: $F_2$ is increasing for $x \geq 6$ and any $m$ large enough. □

Now, as we did with $F_1$ and $G_1$, we need an analogous relation between $F_2$ and $G_2$ in order to prove (4.9).

**Lemma 4.19.** *There exists an $m_3 \in \mathbb{N}$ such that for every $m \geq m_3$ and for every $3\log_2(m) \leq x \leq m - 3$, $F_2(x + 1) < G_2(x)$.*

*Proof.* The inequality $F_2(x + 1) < G_2(x)$ can be rephrased as

$$F_2^2(x + 1)G_2(x) > F_2^1(x + 1),$$

where $F_2^1$ is the numerator of $F_2$ and $F_2^2$ is its denominator. Thus, this is equivalent to

$$\begin{aligned}
&2\big(9(2^{m+1} - 2^{x+2}) - 2m^3 - 3m^2 - 13m + 6(x + 1)^2 + 12(x + 1) - 168\big)(2^x + m + 26)\\
&\qquad - 2^{m+1}(6 \cdot 2^{x+1} + 2m^3 - 3m^2 + 7m - 6(x + 1)^2 - 24)\\
&\qquad + 3(2^{x+2} - 2(x + 1) + m^2 + m + 32)^2 > 0.
\end{aligned}$$

The left-hand side of this inequality is

$$\begin{aligned}
&2\big(9(2^{m+1} - 2^{x+2}) - 2m^3 - 3m^2 - 13m + 6(x+1)^2 + 12(x+1) - 168\big)(2^x + m + 26)\\
&\qquad - 2^{m+1}(6 \cdot 2^{x+1} + 2m^3 - 3m^2 + 7m - 6(x+1)^2 - 24)\\
&\qquad + 3(2^{x+2} - 2(x+1) + m^2 + m + 32)^2\\
&\quad = 2^{x+1}(3 \cdot 2^{m+1} - 3 \cdot 2^{x+2} - 2m^3 + 9m^2 - 37m + 6x^2 - 726)\\
&\qquad + 2^{m+1}(-2m^3 + 3m^2 + 11m + 6x^2 + 12x + 498)\\
&\qquad - m^4 - 104m^3 - m^2\,(12x - 1) + m(36x - 796)\\
&\qquad + 12mx^2 + 324x^2 + 888x - 5100\\
&\quad \geq 2^{x+1}(3 \cdot 2^{m+1} - 3 \cdot 2^{m-1} - 2m^3 + 9m^2 - 37m + 6x^2 - 726)\\
&\qquad + 2^{m+1}(-2m^3 + 3m^2 + 11m + 6x^2 + 12x + 498)\\
&\qquad - m^4 - 104m^3 - m^2\,(12x - 1) + m(36x - 796)\\
&\qquad + 12mx^2 + 324x^2 + 888x - 5100\\
&\quad \text{(because } x \leq m - 3)\\
&\quad = 2^{x+1}(9 \cdot 2^{m-1} - 2m^3 + 9m^2 - 37m + 6x^2 - 726)\\
&\qquad + 2^{m+1}(-2m^3 + 3m^2 + 11m + 6x^2 + 12x + 498)\\
&\qquad - m^4 - 104m^3 - m^2(12x - 1) + m(36x - 796)\\
&\qquad + 12mx^2 + 324x^2 + 888x - 5100\\
&\quad = 2^{x+1}(5 \cdot 2^{m-1} - 2m^3 + 9m^2 - 37m + 6x^2 - 726)\\
&\qquad + 2^{m+1}(2^{x+1} - 2m^3 + 3m^2 + 11m + 6x^2 + 12x + 498)\\
&\qquad - m^4 - 104m^3 - m^2(12x - 1) + m(36x - 796)\\
&\qquad + 12mx^2 + 324x^2 + 888x - 5100. \qquad (4.10)
\end{aligned}$$

Now, on the one hand,

$$5 \cdot 2^{m-1} - 2m^3 + 9m^2 - 37m + 6x^2 - 726 \geq 5 \cdot 2^{m-1} - 2m^3 + 9m^2 - 37m - 726$$

and, if $m$ is large enough, the expression on the right-hand side of this inequality is positive. On the other hand, if $3\log_2(m) \leq x \leq m - 3$, then

$$\begin{aligned}
&2^{m+1}(2^{x+1} - 2m^3 + 3m^2 + 11m + 6x^2 + 12x + 498)\\
&\qquad - m^4 - 104m^3 - m^2(12x - 1) + m(36x - 796)\\
&\qquad + 12mx^2 + 324x^2 + 888x - 5100\\
&\quad \geq 2^{m+1}(2m^3 - 2m^3 + 3m^2 + 11m + 54\log_2(m)^2 + 36\log_2(m) + 498)\\
&\qquad - m^4 - 104m^3 - m^2(12(m-3) - 1) + m(108\log_2(m) - 796)\\
&\qquad + 108m\log_2(m)^2 + 2664\log_2(m) - 5100\\
&\quad = 2^{m+1}(3m^2 + 11m + 54\log_2(m)^2 + 36\log_2(m) + 498)\\
&\qquad - (m^4 + 116m^3 - 37m^2 + 796m + 5100)\\
&\qquad + 108m\log_2(m)(\log_2(m) + 1) + 2664\log_2(m)
\end{aligned}$$

and, again, if $m$ is large enough, this expression (dominated by $3 \cdot 2^{m+1}m^2$) will also be positive. Therefore, if $m$ is large enough the right-hand side expression in (4.10) is positive, and the inequality in the statement of this lemma holds. □

Lemmata 4.18 and 4.19 jointly imply, following an argument analogous to that used above, that, for any $m \geq m_1 = \max\{m_2, m_3\}$,

$$\bigcup_{x=\lceil 3\log_2(m)\rceil}^{m-2} (F_2(x), G_2(x)] = \left(F_2\left(\lceil 3\log_2(m)\rceil\right), G_2(m-2)\right].$$

Therefore, as a consequence of Equation 4.8, we obtain the following result

> There exists $m_1 \in \mathbb{N}$ such that whenever $n = 2^m + k$ with $m \geq m_1$ and $k \in \left(F_2(\lceil 3\log_2(m)\rceil), G_2(m-2)\right]$, the minimum of $V$ on **BinTree**$_n$ is not attained at $T_n^{\mathrm{bal}}$.

Finally, it turns out that, there exists an $m_4 \in \mathbb{N}$ such that, if $m \geq m_4$, the intervals

$$\left(F_1(m-1), G_1(5)\right], \qquad \left(F_2(\lceil 3\log_2(m)\rceil), G_2(m-2)\right]$$

overlap; more specifically, we have that, if $m$ is large enough,

$$F_2(\lceil 3\log_2(m)\rceil) < F_1(m-1) < G_2(m-2) < 2^{m+1} - 31 = G_1(5).$$

Indeed:

- $F_2(\lceil 3\log_2(m)\rceil) < F_1(m-1)$ holds for $m$ large enough because the right-hand side of this inequality is in $O(2^m)$ while the left-hand side is in $O(m^3)$.

- As for
$$F_1(m-1) < G_2(m-2)$$
it is equivalent to
$$4^{m-1} + 2^m(m-1)(m-2) - m^2 < (2^{m-2} + m + 26)(3 \cdot 2^{m-1} - m^2 - 2)$$
that is, to
$$2^{2m} - 10 \cdot 2^m m^2 + 36 \cdot 2^m m + 292 \cdot 2^m - 8m^3 - 200m^2 - 16m - 416 > 0,$$
which holds for $m$ large enough, since the leading term in the left-hand side expression is $2^{2m}$.

- Finally, $G_2(m-1) < 2^{m+1} - 31$, that is $2^{m-2} + m + 26 < 2^m - 31$, holds for $m \geq 7$.

Therefore, taking $m_0 = \max\{9, m_1, m_4\}$, thus far we have shown the following result:

**Proposition 4.20.** *There exists an $m_0 \in \mathbb{N}$ such that, whenever $m \geq m_0$, if*

$$n \in \left(2^m + F_2(\lceil 3\log_2(m)\rceil), 2^{m+1} - 31\right],$$

*then $V(T_n^{\mathrm{bal}})$ is not minimal on* **BinTree**$_n$.

**Remark 4.21.** Concerning the right-hand side end of the interval in the last proposition, we shall prove in Theorem 4.36 that, in fact, if $m \geq 8$ and $n = 2^{m+1} - 30$, then $T_n^{\mathrm{bal}}$ is not minimal in $\mathbf{BinTree}_n$, but not beyond that bound. That is: that, for $n \in \{2^{m+1}-29, \dots, 2^{m+1}\}$ the minimum Variance of depths it achieved at the maximally balanced trees and the trees depth-equivalent to them.

Now, we have already remarked that $F_2(\lceil 3\log_2(m)\rceil)$ is in $O(m^3)$. Therefore, the cardinality of the set of numbers $n \in [2^m, 2^{m+1})$ such that $T_n^{\mathrm{bal}}$ is minimal is in $O(m^3)$, because it is contained in

$$\big[2^m, 2^m + F_2(\lceil 3\log_2(m)\rceil)\big] \cup \big[2^{m+1} - 30, 2^{m+1}\big].$$

Thus, for every $m \geq m_0$, with $m_0$ the lower bound obtained in Theorem 4.20, the *fraction* of values $n \in [2, 2^{m+1})$ such that $V(T_n^{\mathrm{bal}})$ is minimal on $\mathbf{BinTree}_n$ is bounded from above by

$$O\Big(\frac{2^{M+1} + \sum_{p=M+1}^{m} p^3}{2^{m+1}}\Big) = O\Big(\frac{m^4}{2^{m+1}}\Big)$$

which tends to 0 as $m \to \infty$. This proves the theorem that was the goal of this section:

**Theorem 4.22.** *As $N$ grows to $\infty$, the fraction of values $n \in \big[2, 2^N\big]$ such that $V(T_n^{\mathrm{bal}})$ is minimal on* $\mathbf{BinTree}_n$ *tends to* 0.

## 4.3 Expected value of the Variance of depths

Let $V_n$ be the random variable that chooses a tree $T \in \mathbf{BinPhyloTree}_n$ and then computes $V(T)$. In this section we are going to give closed expressions for the expected value of $V_n$ under the Yule and the Uniform models. More in general, given a probabilistic model of phylogenetic trees $(P_n)_n$, let us denote, as in Section 1.3.4, by $E_P$ the expected value of some random variable under this model. By Equation (4.1) and the linearity of the expectation of a random variable, we have that

$$E_P(V_n) = \frac{1}{n}E_P(S_n^{(2)}) - \frac{1}{n^2}E_P(S_n^2).$$

In this expression, $S_n$ and $S_n^{(2)}$ are the random variables that choose a tree $T \in \mathbf{BinPhyloTree}_n$ and compute its Sackin index $S(T)$ and the sum of the squares of its depths $S^{(2)}(T)$, respectively. In Lemma 1.31 we have given a recurrence that can be applied to compute $E_P(S_n^2)$ when $(P_n)_n$ is shape invariant and Markovian, which is the case of the Yule and the Uniform models. We are interested in a similar recurrence for $E_P(S_n^{(2)})$, but we cannot apply to $S_n^{(2)}$ the aforementioned lemma, because it is not a recursive shape index. Fortunately, it is close enough to being so, as the next lemma shows, to allow us to find such a recurrence.

**Lemma 4.23.** *Let $T \in$* $\mathbf{Tree}_n$. *Then, if $T = T_1 * \dots * T_k$, with $k \geq 2$,*

$$S^{(2)}(T) = \sum_{i=1}^{k} S^{(2)}(T_i) + 2\sum_{i=1}^{k} S(T_i) + n.$$

*Proof.* It is simply a matter of expanding the definition:

$$\begin{aligned}
S^{(2)}(T) &= \sum_{x\in L(T)} \delta_T(x)^2 = \sum_{i=1}^{k} \sum_{x\in L(T_i)} (\delta_{T_i}(x)+1)^2\\
&= \sum_{i=1}^{k} \Big(\delta_{T_i}(x)^2 + 2\delta_{T_i}(x) + 1\Big)\\
&= \sum_{i=1}^{k} \sum_{x\in L(T_i)} \delta_{T_i}(x)^2 + 2\sum_{i=1}^{k} \sum_{x\in L(T_i)} \delta_{T_i}(x) + \sum_{i=1}^{k} \sum_{x\in L(T_i)} 1\\
&= \sum_{i=1}^{k} S^{(2)}(T_i) + 2\sum_{i=1}^{k} S(T_i) + n.
\end{aligned}$$

□

Now we have the following extension of Lemma 1.31 to $S_n^{(2)}$:

**Lemma 4.24.** *Let $(P_n)_n$ be a shape invariant Markovian probabilistic model of phylogenetic trees, with conditional split distribution $q_P : \mathbb{N}_{\geq 1} \times \mathbb{N}_{\geq 1} \longrightarrow \mathbb{R}$, and set*

$$Q_P(k, n-k) = \frac{1}{2}\binom{n}{k} q_P(k, n-k).$$

*Then, for any $n \geq 2$,*

$$E_P(S_n^{(2)}) = \sum_{k=1}^{n-1} Q_P(k, n-k)\big(2E_P(S_k^{(2)}) + 4E_P(S_k) + n\big).$$

*Proof.* We develop $E_P(S_n^{(2)})$ following the spirit of the proof of Lemma 1.31 *(i)*, using Lemma 4.23:

$$\begin{aligned}
E_P(S_n^{(2)}) &= \sum_{T\in\textbf{BinPhyloTree}_n} S^{(2)}(T)\cdot P_n(T)\\
&= \frac{1}{2}\sum_{k=1}^{n-1} \sum_{\Lambda_k\in\mathrm{Part}_k([n])} \sum_{T_k\in\textbf{BinPhyloTree}(\Lambda_k)} \sum_{T_{n-k}\in\textbf{BinPhyloTree}(\Lambda_k^c)} S^{(2)}(T_k * T_{n-k})\cdot P_n(T_k * T_{n-k})\\
&= \frac{1}{2}\sum_{k=1}^{n-1} \sum_{\Lambda_k\in\mathrm{Part}_k([n])} \sum_{T_k\in\textbf{BinPhyloTree}(\Lambda_k)} \sum_{T_{n-k}\in\textbf{BinPhyloTree}(\Lambda_k^c)} \big(S^{(2)}(T_k) + S^{(2)}(T_{n-k})\\
&\quad + 2S(T_k) + 2S(T'_{n-k}) + n\big) q_P(k, n-k) P_k(T_k)\cdot P_{n-k}(T_{n-k})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}\sum_{T_k\in\mathbf{BinPhyloTree}_k}\ \sum_{T_{n-k}\in\mathbf{BinPhyloTree}_{n-k}}\big(S^{(2)}(T_k)+S^{(2)}(T_{n-k})\\
&\quad + 2S(T_k)+2S(T'_{n-k})+n\big)q_P(k,n-k)P_k(T_k)\cdot P_{n-k}(T_{n-k})
\end{aligned}
$$

(by the shape invariance of $S$ and $S^{(2)}$)

$$
\begin{aligned}
&= \sum_{k=1}^{n-1} Q_P(k,n-k)\sum_{T_k}\sum_{T_{n-k}}\big(S^{(2)}(T_k)+S^{(2)}(T_{n-k})+2S(T_k)+2S(T'_{n-k})+n\big)\\
&\qquad\qquad\cdot P_k(T_k)\cdot P_{n-k}(T_{n-k})\\
&= \sum_{k=1}^{n-1} Q_P(k,n-k)\Big(\sum_{T_k}\sum_{T_{n-k}} S^{(2)}(T_k)P_k(T_k)P_{n-k}(T_{n-k})\\
&\quad + \sum_{T_k}\sum_{T_{n-k}} S^{(2)}(T_{n-k})P_k(T_k)P_{n-k}(T_{n-k}) + 2\sum_{T_k}\sum_{T_{n-k}} S(T_k)P_k(T_k)P_{n-k}(T_{n-k})\\
&\quad + 2\sum_{T_k}\sum_{T_{n-k}} S(T_{n-k})P_k(T_k)P_{n-k}(T_{n-k}) + \sum_{T_k}\sum_{T_{n-k}} nP_k(T_k)P_{n-k}(T_{n-k})\Big)\\
&= \sum_{k=1}^{n-1} Q_P(k,n-k)\Bigg[\Big(\sum_{T_k} S^{(2)}(T_k)P_k(T_k)\Big)\Big(\sum_{T_{n-k}} P_{n-k}(T_{n-k})\Big)\\
&\quad + \Big(\sum_{T_k} P_k(T_k)\Big)\Big(\sum_{T_{n-k}} S^{(2)}(T_{n-k})P_{n-k}(T_{n-k})\Big)\\
&\quad + 2\Big(\sum_{T_k} S(T_k)P_k(T_k)\Big)\Big(\sum_{T_{n-k}} P_{n-k}(T_{n-k})\Big)\\
&\quad + 2\Big(\sum_{T_k} P_k(T_k)\Big)\Big(\sum_{T_{n-k}} S(T_{n-k})P_{n-k}(T_{n-k})\Big)\\
&\quad + n\Big(\sum_{T_k} P_k(T_k)\Big)\Big(\sum_{T_{n-k}} P_{n-k}(T_{n-k})\Big)\Bigg]\\
&= \sum_{k=1}^{n-1} Q_P(k,n-k)\Big(\sum_{T_k} S^{(2)}(T_k)P_k(T_k) + \sum_{T_{n-k}} S^{(2)}(T_{n-k})P_{n-k}(T_{n-k})\\
&\quad + 2\sum_{T_k} S(T_k)P_k(T_k) + 2\sum_{T_{n-k}} S(T_{n-k})P_{n-k}(T_{n-k}) + n\Big)\\
&= \sum_{k=1}^{n-1} Q_P(k,n-k)\Big(E_P(S_k^{(2)})+E_P(S_{n-k}^{(2)})+2E_P(S_k)+2E_P(S_{n-k})+n\Big)\\
&= 2\sum_{k=1}^{n-1} Q_P(k,n-k)E_P(S_k^{(2)}) + 4\sum_{k=1}^{n-1} Q_P(k,n-k)E_P(S_k) + n\sum_{k=1}^{n-1} Q_P(k,n-k)
\end{aligned}
$$

by the symmetry of $Q_P(k,n-k)$. □

In our computations under the uniform model we shall also make use of the following lemma:

**Lemma 4.25.** *For every $n \in \mathbb{N}_{\geq 2}$ and for every $s,t \in \{1,2,\dots,n-1\}$,*

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

*Proof.* It is enough to carefully develop the sum:

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

□

Note that $\binom{k-1}{s-1}\binom{n-k-1}{t-1}$ is a polynomial of degree $s+t-2$ in $k$ and therefore

$$\sum_{k=1}^{n-1}\binom{k-1}{s-1}\binom{n-k-1}{t-1}$$

is a polynomial of degree $s+t-1$ in $n$.

### 4.3.1 Expected value under the Yule model

We are interested in the computation of

$$E_{\mathrm{Yule}}(V_n)=\frac{1}{n}E_{\mathrm{Yule}}(S_n^{(2)})-\frac{1}{n^2}E_{\mathrm{Yule}}(S_n^2).$$

In this expression, the expected value of $S_n^2$ was already computed in Theorem 2 in [13]:

$$E_{\mathrm{Yule}}(S_n^2)=4n^2(H_n^2-H_n^{(2)}-2H_n)-2nH_n+11n^2-n, \qquad (4.11)$$

where, as it is usual, $H_n=\sum_{i=1}^n\frac{1}{i}$ and $H_n^{(2)}=\sum_{i=1}^n\frac{1}{i^2}$.

With respect to $E_{\mathrm{Yule}}(S_n^{(2)})$, we have to the following proposition.

**Proposition 4.26.** *For every $n\in\mathbb{N}_{\geq 1}$,*

$$E_{\mathrm{Yule}}(S_n^{(2)})=2n(2H_n^2-3H_n-2H_n^{(2)}+3).$$

*Proof.* Lemma 4.24 gives us, in the Yule case, for which $Q_{\text{Yule}}(k, n-k) = \frac{1}{n-1}$, the recurrence

$$E_{\text{Yule}}(S_n^{(2)}) = \frac{2}{n-1}\sum_{k=1}^{n-1} E_{\text{Yule}}(S_k^{(2)}) + \frac{4}{n-1}\sum_{k=1}^{n-1} E_{\text{Yule}}(S_k) + n \tag{4.12}$$

which allows us to write

$$\begin{aligned}
E_{\text{Yule}}(S_n^{(2)}) &= \frac{2}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + \frac{2}{n-1}\sum_{k=1}^{n-2} E_{\text{Yule}}(S_k^{(2)}) + \frac{4}{n-1}E_{\text{Yule}}(S_{n-1})\\
&\quad + \frac{4}{n-1}\sum_{k=1}^{n-2} E_{\text{Yule}}(S_k) + n\\
&= \frac{2}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + \frac{n-2}{n-1}\frac{2}{n-2}\sum_{k=1}^{n-2} E_{\text{Yule}}(S_k^{(2)}) + \frac{4}{n-1}E_{\text{Yule}}(S_{n-1})\\
&\quad + \frac{n-2}{n-1}\frac{4}{n-2}\sum_{k=1}^{n-2} E_{\text{Yule}}(S_k) + \frac{n-2}{n-1}(n-1) + 2\\
&= \frac{2}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + \frac{4}{n-1}E_{\text{Yule}}(S_{n-1}) + 2\\
&\quad + \frac{n-2}{n-1}\Big(\frac{2}{n-2}\sum_{k=1}^{n-2} E_{\text{Yule}}(S_k^{(2)}) + \frac{4}{n-2}\sum_{k=1}^{n-2} E_{\text{Yule}}(S_k) + n - 1\Big)\\
&= \frac{2}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + \frac{4}{n-1}E_{\text{Yule}}(S_{n-1}) + 2 + \frac{n-2}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)})\\
&\quad \text{(by (4.12) for } E_{\text{Yule}}(S_{n-1}^{(2)}))\\
&= \frac{n}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + \frac{4}{n-1}E_{\text{Yule}}(S_{n-1}) + 2\\
&= \frac{n}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + \frac{4}{n-1}2(n-1)(H_{n-1}-1) + 2\\
&= \frac{n}{n-1}E_{\text{Yule}}(S_{n-1}^{(2)}) + 8H_{n-1} - 6
\end{aligned}$$

(we recalled the value of $E_{\text{Yule}}(S_n)$ in page 44).

Now, if we set $x_n = \frac{1}{n}E_{\text{Yule}}(S_n^{(2)})$, then the last expression is equivalent to

$$x_n = x_{n-1} + \frac{8}{n}H_{n-1} - \frac{6}{n}.$$

And, as $x_1 = 0$ and, by Equation (6.71) in [50],

$$\sum_{k=1}^{n-1} \frac{H_k}{k+1} = \frac{1}{2}(H_n^2 - H_n^{(2)}),$$

we have that

$$\begin{aligned}
x_n &= \sum_{k=2}^{n} \frac{8H_{k-1}}{k} - \sum_{k=2}^{n}\frac{6}{k} = 8\sum_{k=1}^{n-1}\frac{H_k}{k+1} - 6(H_n - 1)\\
&= 4(H_n^2 - H_n^{(2)}) - 6H_n + 6.
\end{aligned}$$

Finally,

$$E_{\mathrm{Yule}}(S_n^{(2)}) = nx_n = 2n(2H_n^2 - 2H_n^{(2)} - 3H_n + 3)$$

as we claimed. □

This leads to the main result in this subsection.

**Theorem 4.27.** *Let $n \in \mathbb{N}_{\geq 1}$. Then,*

$$E_{\mathrm{Yule}}(V_n) = 2\Big(\frac{n+1}{n}\Big)H_n + \frac{1}{n} - 5.$$

*Proof.* As we have already mentioned, we have

$$\begin{aligned} E_{\mathrm{Yule}}(V_n) &= \frac{1}{n}E_{\mathrm{Yule}}(S_n^{(2)}) - \frac{1}{n^2}E_{\mathrm{Yule}}(S_n^2) \\ &= 4H_n^2 - 4H_n^{(2)} - 6H_n + 6 - 4H_n^2 + 4H_n^{(2)} + 8H_n + \frac{2}{n}H_n - 11 + \frac{1}{n} \\ &= \left(2 + \frac{2}{n}\right)H_n + \frac{1}{n} - 5. \end{aligned}$$

□

Finally, notice that, as $H_n \sim \ln n + O(1)$ (cf. p. 264 in [51]) and $E_{\mathrm{Yule}}(S_n) = 2n(H_n - 1)$, then

$$\begin{aligned} E_{\mathrm{Yule}}(\overline{S}_n) &= 2H_n - 2 \sim 2\ln n \\ E_{\mathrm{Yule}}(V_n) &= 2H_n + \frac{2}{n}H_n + \frac{1}{n} - 5 \sim 2\ln n, \end{aligned}$$

and therefore both the expected value of the mean and the variance of the leaves' depths of a bifurcating phylogenetic tree generated under the Yule model grow asymptotically as $2\ln n$.

### 4.3.2 Expected value under the Uniform model

We proceed now to the computation of the expected value of the Variance of depths under the Uniform model for binary phylogenetic trees. The line of reasoning is the same as in the Yule case. First, we have that

$$E_{\mathrm{unif}}(V_n) = \frac{1}{n}E_{\mathrm{unif}}(S_n^{(2)}) - \frac{1}{n^2}E_{\mathrm{unif}}(S_n^2).$$

Now, Lemma 4.24 gives us a recursive expression for $E_{\mathrm{unif}}(S_n^{(2)})$ and Lemma 1.31 a recursive expression for $E_{\mathrm{unif}}(S_n^2)$ —recursive expressions that we shall then solve with the aid of the results established in Section 1.4.1. The recursive expression for $E_{\mathrm{unif}}(S_n^2)$ was already given in L. Rotger's PhD Thesis (see Prop. 2.36 in [100]).

Recall that, with the notations of Lemmata 1.31 and 4.24, in the Uniform model we have

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

**Lemma 4.28.** *Let $n \in \mathbb{N}_{\geq 2}$. Then,*

*(i)* $$E_{\text{unif}}(S_n^{(2)}) = 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k^{(2)}) + 2n\frac{(2n-2)!!}{(2n-3)!!} - 3n.$$

*(ii)* $$E_{\text{unif}}(S_n^2) = 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k^2) + \frac{5n^2}{2}\frac{(2n-2)!!}{(2n-3)!!} - n(5n-2).$$

*Proof.* We shall begin by the proof of *(i)*. By Lemma 4.24,

$$\begin{aligned}
E_{\text{unif}}(S_n^{(2)}) &= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k^{(2)}) + 4\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k) + n\sum_{k=1}^{n-1} C_{k,n-k} \\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k^{(2)}) + 4\sum_{k=1}^{n-1} C_{k,n-k} \cdot k \cdot \frac{(2k-2)!!}{(2k-3)!!} \\
&\qquad - 4\sum_{k=1}^{n-1} C_{k,n-k} k + n\sum_{k=1}^{n-1} C_{k,n-k} \\
&\text{(by using the value of } E_{\text{unif}}(S_n) \text{ recalled in page 44)} \\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k^{(2)}) + 2n\Big(\frac{(2n-2)!!}{(2n-3)!!} - 1\Big) - 2n + n \\
&\text{(by Lemmata 1.33 and 1.34)} \\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\text{unif}}(S_k^{(2)}) + 2n \cdot \frac{(2n-2)!!}{(2n-3)!!} - 3n
\end{aligned}$$

as we claimed.

Let us now proceed to the proof of *(ii)*. In order to do that, we shall use Lemma 1.31, which tells us (recall that $f_S(k, n-k) = n$; cf Section 1.2.4) that

$$\begin{aligned}
E_{\text{unif}}(S_n^2) &= \sum_{k=1}^{n-1} C_{k,n-k} \big(2E_{\text{unif}}(S_k^2) + 2E_{\text{unif}}(S_k)E_{\text{unif}}(S_{n-k}) + n^2 + 4nE_{\text{unif}}(S_k)\big) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \Bigg(2E_{\text{unif}}(S_k^2) + 2k(n-k)\left(\frac{(2k-2)!!}{(2k-3)!!} - 1\right)\left(\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} - 1\right) \\
&\qquad + n^2 + 4nk\left(\frac{(2k-2)!!}{(2k-3)!!} - 1\right)\Bigg) \\
&\text{(by using again the value of } E_{\text{unif}}(S_n)) \\
&= \sum_{k=1}^{n-1} C_{k,n-k} \Big(2E_{\text{unif}}(S_k^2) + n^2 + 4nk\frac{(2k-2)!!}{(2k-3)!!} - 4nk \\
&\qquad + 2k(n-k)\frac{(2k-2)!!}{(2k-3)!!}\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} + 2k(n-k) \\
&\qquad - 2k(n-k)\frac{(2k-2)!!}{(2k-3)!!} - 2k(n-k)\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!}\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n-1} C_{k,n-k}\Big(2E_{\mathrm{unif}}(S_k^2) + 4nk\frac{(2k-2)!!}{(2k-3)!!} - 2k^2 \\
&\qquad + 2k(n-k)\frac{(2k-2)!!}{(2k-3)!!}\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!} - 4k(n-k)\frac{(2k-2)!!}{(2k-3)!!}\Big) \\
&= (*),
\end{aligned}
$$

since $C_{k,n-k}$ is symmetric and so

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

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

$$
\begin{aligned}
(*) = 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}(S_k^2) - 2\sum_{k=1}^{n-1} C_{k,n-k}k^2 + 4\sum_{k=1}^{n-1} C_{k,n-k}k^2\frac{(2k-2)!!}{(2k-3)!!} \\
+ 2\sum_{k=1}^{n-1} C_{k,n-k}k(n-k)\frac{(2k-2)!!}{(2k-3)!!}\frac{(2(n-k)-2)!!}{(2(n-k)-3)!!}.
\end{aligned}
$$

Now, as for the first two sums appearing in the independent term of this recurrence —namely, the first two sums without terms $E_{\mathrm{unif}}(S_k^2)$—, their value can be computed using Lemmata 1.33 and 1.34:

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

As to the third sum in this independent term, its value can be computed using Lemma 4.25 with $s = t = 1$:

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

Hence, the independent term is

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

Thus concludes the proof. □

In the next theorem we give solutions to the recurrences presented in the previous lemma, and in so doing we will be able to finally give a value for $E_{\mathrm{unif}}(V_n)$. In order to do that, we shall use Theorem 1.35.

**Theorem 4.29.** *Let $n \in \mathbb{N}_{\geq 1}$. Then,*

$$\begin{aligned}E_{\mathrm{unif}}(S_n^{(2)}) &= (4n-1)n - 3n\frac{(2n-2)!!}{(2n-3)!!} \\ E_{\mathrm{unif}}(S_n^2) &= \frac{n(10n^2-1)}{3} - \frac{n(5n+1)}{2}\frac{(2n-2)!!}{(2n-3)!!}.\end{aligned}$$

*Proof.* We know, by Theorem 4.28, that $E_{\mathrm{unif}}(S_n^{(2)})$ is the solution of the recurrent equation

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

with initial condition $x_1 = E_{\mathrm{unif}}(S_1^{(2)}) = 0$. Now, by Theorem 1.35, this solution is

$$E_{\mathrm{unif}}(S_n^{(2)}) = 3n + 8\binom{n}{2} - 3n\frac{(2n-2)!!}{(2n-3)!!} = 4n^2 - n - 3n\frac{(2n-2)!!}{(2n-3)!!}.$$

Finally, by Theorem 4.28, $E_{\mathrm{unif}}(S_n^2)$ is the solution of

$$y_n = 2\sum_{k=1}^{n-1} C_{k,n-k} y_k - 10\binom{n}{2} - 3n + \left(5\binom{n}{2} + \frac{5}{2}n\right)\frac{(2n-2)!!}{(2n-3)!!}$$

such that $y_1 = E_{\text{unif}}(S_1^2) = 0$. Thus, by Theorem 1.35, this solution is

$$\begin{aligned} E_{\text{unif}}(S_n^2) &= 20\binom{n}{3} + 20\binom{n}{2} + 3n - \left(5\binom{n}{2} + 3n\right)\frac{(2n-2)!!}{(2n-3)!!} \\ &= \frac{10n^3 - n}{3} - \frac{5n^2+n}{2}\frac{(2n-2)!!}{(2n-3)!!}. \end{aligned}$$

□

Finally, then, to end this section, the immediate corollary of Theorem 4.29 is the main result of this subsection.

**Theorem 4.30.** *Let $n \in \mathbb{N}_{\geq 1}$. Then,*

$$E_{\text{unif}}(V_n) = \frac{(2n-1)(n-1)}{3n} - \frac{n-1}{2n}\frac{(2n-2)!!}{(2n-3)!!}.$$

We shall now briefly discuss the asymptotic behaviour of $E_{\text{unif}}(\overline{S}_n)$ and $E_{\text{unif}}(V_n)$. As we have already mentioned, Theorem 22 in [85] implies that

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

Now let us fix our attention to that ubiquitous fraction, $\frac{(2n-2)!!}{(2n-3)!!}$. By using Stirling's approximation for large factorials, we have

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

Thus, we now have that

$$\begin{aligned} E_{\text{unif}}(\overline{S}_n) &= \frac{(2n-2)!!}{(2n-3)!!} - 1 \sim \sqrt{\pi n} \\ E_{\text{unif}}(V_n) &= \frac{(2n-1)(n-1)}{3n} - \frac{n-1}{2n}\frac{(2n-2)!!}{(2n-3)!!} \sim \frac{2}{3}n, \end{aligned}$$

contrary to what happened under the Yule model, in which both the expected value of $\overline{S}_n$ and $V_n$ had the same asymptotic behaviour.

## 4.4 Some new results for the Sackin and Cophenetic indices under the Uniform model

In this section we shall present three theorems that were initially obtained as a by-product of the techniques used above and are, to the extent of our knowledge, new. Let $S_n$ and $\Phi_n$ be, respectively, the random variable that take a tree $T \in \mathbf{BinTree}_n$ and compute its Sackin index $S(T)$ and its Cophenetic index $\Phi(T)$. Recall that the

Cophenetic index, as we defined it in the Preliminaries, is the sum, over all pairs of leaves of a given tree, of the depth of their least common ancestor.

We shall begin by computing the variance of $S_n$ under the Uniform model — the variance under the Yule model being already computed in [13].

**Theorem 4.31.** *Let $n \in \mathbb{N}_{\geq 1}$. The variance of $S_n$ under the Uniform model is*

$$\sigma^2_{\text{unif}}(S_n) = \frac{n(10n^2 - 3n - 1)}{3} - \binom{n+1}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} - n^2 \left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.$$

*Proof.* The variance of $S_n$ under the Uniform model is

$$\sigma^2_{\text{unif}}(S_n) = E_{\text{unif}}(S_n^{(2)}) - E_{\text{unif}}(S_n)^2,$$

and in order to compute it we can use the expression for $E_{\text{unif}}(S_n^{(2)})$ obtained in Theorem 4.29. Indeed, we then compute

$$\begin{aligned}
\sigma^2_{\text{unif}}(S_n) &= E_{\text{unif}}(S_n^{(2)}) - E_{\text{unif}}(S_n)^2 \\
&= \frac{10}{3}n^3 - \frac{1}{3}n - \frac{n(5n+1)}{2}\frac{(2n-2)!!}{(2n-3)!!} - n^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2 \\
&= \frac{10}{3}n^3 - \frac{1}{3}n - n^2 - n^2\left(\frac{(2n-2)!!}{(2n-3)!!} - \frac{n(n+1)}{2}\frac{(2n-2)!!}{(2n-3)!!}\right) \\
&= \frac{n(10n^2 - 3n - 1)}{3} - \binom{n+1}{2} \cdot \frac{(2n-2)!!}{(2n-3)!!} - n^2 \left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2,
\end{aligned}$$

as we wanted to prove. □

**Remark 4.32.** Notice the difference between the magnitude computed above and the expected value of the Variance of depths under the Uniform model. The former is

$$\sigma^2_{\text{unif}}(S_n) = E_{\text{unif}}(S_n^{(2)}) - E_{\text{unif}}(S_n)^2$$

whereas the latter is

$$E_{\text{unif}}(V_n) = \frac{1}{n}E_{\text{unif}}(S_n^{(2)}) - \frac{1}{n^2}E_{\text{unif}}(S_n)^2.$$

The previous formula agrees with the already established asymptotic behaviour of $\sigma^2_{\text{unif}}(S_n)$, stated in the Preliminaries. Indeed, since we have already argued that $\frac{(2n-2)!!}{(2n-3)!!} \sim \sqrt{\pi n}$, we have

$$\sigma^2_{\text{unif}}(S_n) \sim \frac{n(10n^2 - 3n - 1}{3} - \binom{n+1}{2}\sqrt{\pi n} - n^2\pi n \sim \frac{10 - 3\pi}{3}n^3. \tag{4.13}$$

The following two proofs are much more long, but they rely again on the results established in Section 1.4.1 and Lemma 4.25.

**Theorem 4.33.** *Let $n \in \mathbb{N}_{\geq 1}$. The variance of $\Phi_n$ under the Uniform model is*

$$\begin{aligned}
\sigma^2_{\text{unif}}(\Phi_n) = \binom{n}{2}\frac{(2n-1)(7n^2-3n-2)}{30} - \binom{n}{2}\frac{5n^2-n-2}{32} \cdot \frac{(2n-2)!!}{(2n-3)!!} \\
- \frac{1}{4}\binom{n}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.
\end{aligned}$$

*Proof.* If we apply Equation 1.16 in Lemma 1.31, by taking $I$ as the Cophenetic index $\Phi$, for which we recall from the Preliminaries that

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

we obtain the following recurrence for $E_{\text{unif}}(\Phi_n^2)$:

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

We shall now simplify this recurrence. The final form we obtain was already given by L. Rotger in Lemma 2.30 of her PhD Thesis [100]. We begin by simplifying the following expression

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

and then, by the symmetry of $C_{k,n-k}$, we have

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

so that the recursion becomes

$$\begin{aligned}
E_{\mathrm{unif}}(\Phi_n^2) &= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\mathrm{unif}}(\Phi_k^2) + 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}^2\frac{(2k-2)!!}{(2k-3)!!}\\
&\quad - 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}^2 + \frac{1}{2}\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\binom{n-k}{2}\frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!}.
\end{aligned}\tag{4.14}$$

Now, by Lemmata 1.33, 1.34, and 4.25

$$\begin{aligned}
\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}^2 &= \sum_{k=1}^{n-1} C_{k,n-k}\left(6\binom{k}{4}+6\binom{k}{3}+\binom{k}{2}\right)\\
&= 3\binom{n}{4}\left(1-\frac{15}{16(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right) + 3\binom{n}{3}\left(1-\frac{3}{4(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad + \frac{1}{2}\binom{n}{2}\left(1-\frac{1}{2(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&= \frac{1}{2}\binom{n}{2}^2 - \frac{n(15n^2-27n+10)}{2^7}\cdot\frac{(2n-2)!!}{(2n-3)!!}
\end{aligned}$$

$$\begin{aligned}
\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}^2\frac{(2k-2)!!}{(2k-3)!!} &= \sum_{k=1}^{n-1} C_{k,n-k}\left(6\binom{k}{4}+6\binom{k}{3}+\binom{k}{2}\right)\frac{(2k-2)!!}{(2k-3)!!}\\
&= 3\binom{n}{4}\left(\frac{(2n-2)!!}{(2n-3)!!}-\frac{16}{5}\right) + 3\binom{n}{3}\left(\frac{(2n-2)!!}{(2n-3)!!}-\frac{8}{3}\right)\\
&\quad + \frac{1}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}-2\right)
\end{aligned}$$

$$= \frac{1}{2}\binom{n}{2}^2\frac{(2n-2)!!}{(2n-3)!!} - \binom{n}{2}\frac{12n^2-20n+7}{15}$$

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

since

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

which can be shown by induction using that $\binom{n-1}{3} - \binom{n-2}{2} = \binom{n-2}{3}$. Now, the independent term of Equation (4.14) will be

$$2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}^2\frac{(2k-2)!!}{(2k-3)!!} - 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}^2$$
$$+ \frac{1}{2}\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\binom{n-k}{2}\frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!}$$
$$= 2\left(\frac{1}{2}\binom{n}{2}^2\frac{(2n-2)!!}{(2n-3)!!} - \binom{n}{2}\frac{12n^2-20n+7}{15}\right)$$
$$- 2\left(\frac{1}{2}\binom{n}{2}^2 - \frac{n(15n^2-27n+10)}{2^7}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right) + \frac{1}{8}\binom{n}{4}\frac{(2n-2)!!}{(2n-3)!!}$$
$$= \frac{n(49n^3-57n^2-22n+24)}{192}\cdot\frac{(2n-2)!!}{(2n-3)!!} - \frac{n(n-1)(63n^2-95n+28)}{60}.$$

Therefore, the sequence $E_{\mathrm{unif}}(\Phi_n^2)$ is the solution of the recurrence

$$x_n = 2\sum_{k=1}^{n-1} C_{k,n-k}x_k - \frac{63n^4-158n^3+123n^2-28n}{60}$$
$$+ \frac{49n^4-57n^3-22n^2+24n}{192}\cdot\frac{(2n-2)!!}{(2n-3)!!}$$
$$= 2\sum_{k=1}^{n-1} C_{k,n-k}x_k - \frac{126}{5}\binom{n}{4} - 22\binom{n}{3} - 3\binom{n}{2}$$
$$+ \left(\frac{49}{8}\binom{n}{4} + \frac{237}{32}\binom{n}{3} + \frac{25}{16}\binom{n}{2} - \frac{1}{32}n\right)\frac{(2n-2)!!}{(2n-3)!!},$$

whose initial condition is $x_1 = E_{\text{unif}}(\Phi_1^2) = 0$. By Theorem 1.35, the solution is

$$\begin{aligned}
E_{\text{unif}}(\Phi_n^2) &= 28\binom{n}{5} + \frac{256}{5}\binom{n}{4} + 26\binom{n}{3} + 3\binom{n}{2} \\
&\quad - \left(\frac{63}{8}\binom{n}{4} + \frac{33}{4}\binom{n}{3} + \frac{3}{2}\binom{n}{2}\right) \cdot \frac{(2n-2)!!}{(2n-3)!!} \\
&= \binom{n}{2}\left(\frac{7n^3 + n^2 - 8n + 1}{15} - \frac{21n^2 - 17n - 2}{32} \cdot \frac{(2n-2)!!}{(2n-3)!!}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\sigma_{\text{unif}}(\Phi_n)^2 &= E_{\text{unif}}(\Phi_n^2) - E_{\text{unif}}(\Phi_n)^2 \\
&= \binom{n}{2}\left(\frac{7n^3 + n^2 - 8n + 1}{15} - \frac{21n^2 - 17n - 2}{32} \cdot \frac{(2n-2)!!}{(2n-3)!!}\right) \\
&\quad - \frac{1}{4}\binom{n}{2}^2 \left(\frac{(2n-2)!!}{(2n-3)!!} - 2\right)^2 \\
&= \binom{n}{2}\frac{(2n-1)(7n^2 - 3n - 2)}{30} - \binom{n}{2}\frac{5n^2 - n - 2}{32} \cdot \frac{(2n-2)!!}{(2n-3)!!} \\
&\quad - \frac{1}{4}\binom{n}{2}^2 \left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.
\end{aligned}$$

This completes the proof. □

Finally, we end by stating and proving the following result, which presents a closed formula to compute the covariance of the Sackin and Cophenetic indices under the Uniform model. Before that, we shall remind the reader that for two random variables $X, Y$, their covariance can be computed as

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y).$$

Thus, already knowing the values of $E_{\text{unif}}(S_n)$ and $E_{\text{unif}}(\Phi_n)$, all that remains is to compute $E_{\text{unif}}(\Phi_n S_n)$, and in order to do so we shall use Lemma 1.31.

**Theorem 4.34.** *Let $n \in \mathbb{N}_{\geq 2}$. The covariance of $S_n$ and $\Phi_n$ under the Uniform model is*

$$\begin{aligned}
\operatorname{cov}_{\text{unif}}(\Phi_n, S_n) &= \binom{n}{2}\frac{26n^2 - 5n - 4}{15} - \frac{3n+2}{8}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!} \\
&\quad - \frac{n}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.
\end{aligned}$$

*Proof.* As in the previous result, we shall begin by obtaining a recurrence for $\operatorname{cov}_{\text{unif}}(\Phi_n, S_n)$ that was already obtained by L. Rotger in her PhD Thesis [100] (see Proposition 2.41 therein). We give the derivation using our lemmata to ease the task of the reader, and then we solve this recurrence.

If we apply Equation (1.15) in Lemma 1.31 setting $I$ and $J$ to be the Cophenetic index $\Phi$ and the Sackin index $S$, for which

$$\begin{aligned}
f_\Phi(k, n-k) &= \binom{k}{2} + \binom{n-k}{2}, & E_{\text{unif}}(\Phi_k) &= \frac{1}{2}\binom{k}{2}\left(\frac{(2k-2)!!}{(2k-3)!!} - 2\right), \\
f_S(k, n-k) &= n, & E_{\text{unif}}(S_k) &= k\left(\frac{(2k-2)!!}{(2k-3)!!} - 1\right),
\end{aligned}$$

we obtain

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

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

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

and

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

Thus, we are able to further simplify $(*)$, obtaining

$$\begin{aligned}
(*) &= 2\sum_{k=1}^{n-1} C_{k,n-k}E_{\text{unif}}(\Phi_k S_k) + \sum_{k=1}^{n-1} C_{k,n-k}\Bigg(3k\binom{k}{2}\frac{(2k-2)!!}{(2k-3)!!}\\
&\quad - 2k\binom{k}{2}+(n-k)\binom{k}{2}\frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!}\Bigg).
\end{aligned}$$

Now, let us compute the independent term in the above recurrence applying Lemmata

1.33, 1.34 and 4.25:

$$\begin{aligned}
&\sum_{k=1}^{n-1} C_{k,n-k} k\binom{k}{2} = 3\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{3} + 2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\\
&\quad = \frac{3}{2}\binom{n}{3}\left(1-\frac{3}{4(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)+\binom{n}{2}\left(1-\frac{1}{2(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad = \frac{n}{2}\binom{n}{2}-\frac{n(3n-2)}{16}\cdot\frac{(2n-2)!!}{(2n-3)!!}\\
&\sum_{k=1}^{n-1} C_{k,n-k} k\binom{k}{2}\frac{(2k-2)!!}{(2k-3)!!}\\
&\quad = 3\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{3}\frac{(2k-2)!!}{(2k-3)!!}+2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\frac{(2k-2)!!}{(2k-3)!!}\\
&\quad = \frac{3}{2}\binom{n}{3}\left(\frac{(2n-2)!!}{(2n-3)!!}-\frac{8}{3}\right)+\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}-2\right)\\
&\quad = \frac{n}{2}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!}-\frac{2(2n-1)}{3}\binom{n}{2}\\
&\sum_{k=1}^{n-1} C_{k,n-k}(n-k)\binom{k}{2}\frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!}\\
&\quad = \frac{2^{n-4}n!}{(2n-3)!!}\sum_{k=1}^{n-1}(k-1) = \frac{n-2}{8}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!}.
\end{aligned}$$

Therefore, the independent term of the equation for $E_{\mathrm{unif}}(\Phi_n S_n)$ turns out to be

$$\begin{aligned}
&3\sum_{k=1}^{n-1} C_{k,n-k} k\binom{k}{2}\frac{(2k-2)!!}{(2k-3)!!}-2\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}k\\
&\qquad +\sum_{k=1}^{n-1} C_{k,n-k}(n-k)\binom{k}{2}\frac{(2k-2)!!(2(n-k)-2)!!}{(2k-3)!!(2(n-k)-3)!!}\\
&\quad = 3\left(\frac{n}{2}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!}-\frac{2(2n-1)}{3}\binom{n}{2}\right)\\
&\qquad -2\left(\frac{n}{2}\binom{n}{2}-\frac{n(3n-2)}{16}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)+\frac{n-2}{8}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!}\\
&\quad = \frac{n(13n^2-9n-2)}{16}\cdot\frac{(2n-2)!!}{(2n-3)!!}-\binom{n}{2}(5n-2).
\end{aligned}$$

Thus far, we have proved that $E_{\text{unif}}(\Phi_n S_n)$ is the solution of the recurrence

$$\begin{aligned} x_n &= 2\sum_{k=1}^{n-1} C_{k,n-k}x_k - (5n-2)\binom{n}{2} + \frac{n(13n^2-9n-2)}{16}\cdot\frac{(2n-2)!!}{(2n-3)!!} \\ &= 2\sum_{k=1}^{n-1} C_{k,n-k}x_k - 15\binom{n}{3} - 8\binom{n}{2} \\ &\qquad + \left(\frac{39}{8}\binom{n}{3} + \frac{15}{4}\binom{n}{2} + \frac{1}{8}\cdot n\right)\frac{(2n-2)!!}{(2n-3)!!}, \end{aligned}$$

whose initial condition is $x_1 = E_{\text{unif}}(\Phi_1 S_1) = 0$. Now, by Theorem 1.35 this solution is

$$\begin{aligned} E_{\text{unif}}(\Phi_n S_n) &= \frac{104}{5}\binom{n}{4} + 28\binom{n}{3} + 8\binom{n}{2} - \left(\frac{45}{8}\binom{n}{3} + 4\binom{n}{2}\right)\frac{(2n-2)!!}{(2n-3)!!} \\ &= \binom{n}{2}\left(\frac{26n^2+10n-4}{15} - \frac{15n+2}{8}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right). \end{aligned}$$

Finally,

$$\begin{aligned} \operatorname{cov}_{\text{unif}}(\Phi_n, S_n) &= E_{\text{unif}}(\Phi_n S_n) - E_{\text{unif}}(\Phi_n)E_{\text{unif}}(S_n) \\ &= \binom{n}{2}\left(\frac{26n^2+10n-4}{15} - \frac{15n+2}{8}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right) \\ &\quad - \frac{1}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!} - 2\right) n\left(\frac{(2n-2)!!}{(2n-3)!!} - 1\right) \\ &= \binom{n}{2}\left(\frac{26n^2-5n-4}{15} - \frac{3n+2}{8}\cdot\frac{(2n-2)!!}{(2n-3)!!} - \frac{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2\right), \end{aligned}$$

and thus the claim is proved. □

As for the asymptotic behaviour of both $\sigma^2_{\text{unif}}(\Phi_n)$ and $\operatorname{cov}_{\text{unif}}(\Phi_n, S_n)$, using that $\frac{(2n-2)!!}{(2n-3)!!} \sim \sqrt{\pi n}$ and the above results we obtain that

$$\sigma^2_{\text{unif}}(\Phi_n) \sim \frac{56-15\pi}{240}n^5,\ \operatorname{cov}_{\text{unif}}(\Phi_n, S_n) \sim \frac{52-15\pi}{60}n^4.$$

Moreover, having at our disposal closed formulæ for $\sigma^2_{\text{unif}}(S_n)$, $\sigma^2_{\text{unif}}(\Phi_n)$, and $\operatorname{cov}_{\text{unif}}(\Phi_n, S_n)$, we can obtain a closed formula for Pearson's correlation coefficient of $S_n$ and $\Phi_n$

$$\rho_{\text{unif}}(\Phi_n, S_n) = \frac{\operatorname{cov}_{\text{unif}}(\Phi_n, S_n)}{\sigma_{\text{unif}}(\Phi_n)\sigma_{\text{unif}}(S_n)}.$$

We shall omit the specific expression because of its length, but its asymptotic behaviour can be obtained from that of its components:

$$\rho_{\text{unif}}(\Phi_n, S_n) \sim \frac{\frac{52-15\pi}{60}}{\sqrt{\frac{10-3\pi}{3}\cdot\frac{56-15\pi}{240}}} \approx 0.965,$$

Notice that under the Yule model this correlation had already been computed [13], and its limit turns out to be around 0.89.

To close this section, we shall reveal an interesting consequence of this last section. For every $T \in \mathbf{BinTree}_n$ and for every $x, y \in L(T)$, the *nodal distance* between $x$ and $y$ is the length $d_T(x, y)$ of the shortest (non directed) path connecting them. The *total tree area* of $T$ [81] is then defined as the sum of the nodal distances between all pairs of different leaves in it:

$$D(T) = \frac{1}{2} \sum_{\substack{(x,y)\in L(T)^2 \\ x\neq y}} d_T(x, y).$$

Let $D_n$ be the random variable that chooses a tree $T \in \mathbf{BinPhyloTree}_n$ and then computes $D(T)$. A lot of information about the behaviour of $D_n$ under the Uniform model is already known: its expected value [85], its mode [83], a limit formula for its median [82], and its limit distribution [84]. But its variance $\sigma^2_{\text{unif}}(D_n)$ under this model was not known so far. We mend this hole in the literature with the following corollary of the results established in the last section.

**Corollary 4.35.** *Let $n \in \mathbb{N}_{\geq 1}$. The variance of $D_n$ under the Uniform model is*

$$\begin{aligned}\sigma^2_{\text{unif}}(D_n) = {} & 2\binom{n}{2}\frac{12n^3 - 16n^2 + 7n - 1}{15} - \binom{n}{2}\frac{n^2 + 3n - 2}{8}\frac{(2n-2)!!}{(2n-3)!!} \\ & - \binom{n}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.\end{aligned}$$

*Proof.* By Lemma 6 in [85], for every $T \in \mathbf{BinPhyloTree}_n$, $D(T) = (n-1)S(T) - 2\Phi(T)$ and hence

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

Therefore,

$$\sigma^2_{\text{unif}}(D_n) = (n-1)^2\sigma^2_{\text{unif}}(S_n) + 4\sigma^2_{\text{unif}}(\Phi_n) - 4(n-1)\text{cov}_{\text{unif}}(S_n, \Phi_n). \tag{4.15}$$

Replacing $\sigma^2_{\text{unif}}(S_n)$, $\sigma^2_{\text{unif}}(\Phi_n)$, and $\text{cov}_{\text{unif}}(S_n, \Phi_n)$ in the right-hand side expression of this equality by their values obtained in the previous section,

$$\sigma^2_{\text{unif}}(S_n) = \frac{n(10n^2 - 3n - 1)}{3} - \binom{n+1}{2}\cdot\frac{(2n-2)!!}{(2n-3)!!} - n^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2$$

$$\begin{aligned}\sigma^2_{\text{unif}}(\Phi_n) = {} & \binom{n}{2}\frac{(2n-1)(7n^2 - 3n - 2)}{30} - \binom{n}{2}\frac{5n^2 - n - 2}{32}\cdot\frac{(2n-2)!!}{(2n-3)!!} \\ & - \frac{1}{4}\binom{n}{2}^2\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2\end{aligned}$$

$$\begin{aligned}\text{cov}_{\text{unif}}(\Phi_n, S_n) = {} & \binom{n}{2}\frac{26n^2 - 5n - 4}{15} - \frac{3n+2}{8}\binom{n}{2}\frac{(2n-2)!!}{(2n-3)!!} \\ & - \frac{n}{2}\binom{n}{2}\left(\frac{(2n-2)!!}{(2n-3)!!}\right)^2.\end{aligned}$$

we obtain the formula in the statement. □

Now, recall that $\frac{(2n-2)!!}{(2n-3)!!} \sim \sqrt{\pi n}$; then, when $n$ tends to $\infty$, the leading coefficient of the variance of $D_n$ under the Uniform model tends to

$$\frac{12}{15} - \frac{\pi}{4} = \frac{48 - 15\pi}{60} \approx 0.876.$$

## 4.5 Discussion

In his seminal paper on the shape of phylogenetic trees [102], Sackin introduced two measures that, he thought, captured the idea of imbalance of a rooted bifurcating tree: the maximum depth and the variation of the leaves' depths. Despite appearing in some initial studies as the Variance of the leaves' depths, denoted here by $V$, [60, 65, 69, 1], the second measure rapidly faded from use, and is now found marginally in some papers like [59], which extend the experiments performed in [69]. Contrary to what happens with the first Sackin's proposal, that of the maximum depth —which, having only a linear range, has a very poor discriminatory power—, there was no clear reason to explain the ostracism this measure has suffered.

Now we know that the phylogenetic community has been wise in prefering other balance indices to the Variance of the leaves' depths. Up to 183 leaves, $V$ "correctly" classifies the extremal trees, if we are to judge by our intuition of what "balance" means. This property, however, does not hold beyond that number: indeed, for almost all natural numbers $n \in \mathbb{N}$, the minimum of the Variance of depths on $\mathbf{BinTree}_n$ is not held at $T_n^{\mathrm{bal}}$ nor at any depth-equivalent tree. Nevertheless, the maximum value is always reached at the caterpillars, which is akin to our intuition. *Dios aprieta pero no ahoga.*

We have provided two quasi-linear algorithms in this chapter that, given $n \in \mathbb{N}$, compute the multisets of depths of the trees $T \in \mathbf{BinTree}_n$ having minimum Variance of depths. They are basically search algorithms that reduce the scope of the search to a space of multisets of depths of cardinality linear in $n$, and then search the minimum there. Once the vector of depths is provided, then it is easy to retrieve trees that are represented by that vector of lengths by just considering a fully symmetric tree with $2^{\lceil \log_2(n) \rceil}$ leaves and prunning the appropriate subtrees. This method, along with others, can be found in our GitHub repository https://github.com/biocom-uib/biotrees.

In this chapter we have also presented closed formulæ for the expected value and variances of several balance indices under the Yule and the Uniform models. We have begun by computing the expected value of the Variance of depths under the Yule model. From there, we have proceeded to the computation of the expected value of $V$ under the Uniform model —an errand which, as it is seen above, is way more convoluted. Nevertheless, the techniques that aid us to give a solution to that problem are, in our opinion, interesting in their own right (Theorem 1.35 and Lemmata 1.33, 1.34, and 4.25) and have allowed us to compute some expressions that were unknown thus far: the variance of the Sackin and Cophenetic indices under the Uniform model, and their covariance. The aforementioned values were already known under the Yule model [85, 13].

These formulæ allow the proper standarization of the above indices relative to the aforementioned probabilistic model. Recall that the standarization of a shape index, say

$I$, relative to a given probabilistic model $(P_n)_n$, is performed by means of the expression

$$\frac{I_n - E_P(I_n)}{\sigma_P(I_n)}$$

i.e., by substracting to the index its expected value under the considered model, and then by dividing the result by its standard deviation (under the considered model). Because of the lack of these formulæ, for instance, the current version of the `R` package `apTreeshape` standardizes the Sackin index under the Uniform model by dividing by the square root of the asymptotic approximation of $\sigma_{\text{unif}}(S_n)$, computed in Equation 4.13 [10].

### 4.5.1 Open problems

> What went we out into this wilderness to find?
>
> Robert Eggers, *The Witch: a New England folktale*

There are questions that, when pursued, unleash more darkness than light. This is true in the History of Science in general, as one could think that all the knowledge acquired during the last thousands of years has served not only to better understand the world, but, *per negationem*, to vividly draw the *contours* of that which remains unknown. An image comes across as appropriate: in the late 18th century, even though the oceans of the world had been thoroughly explored by the likes of Magalhães, Bougainville and Captain Hook, the heart of Africa lay down blank, inmaculated, and not even the wisest man or woman alive at that time in the Western world could say what wonders dwelt there. Endless cabilations and discussions took place, but in the end all that could be said can be reduced to the old Latin sentence: *hic sunt dracones*.

The Dictionary of Obscure Sorrows [71] gives definitions to specific feelings Human beings sometimes have. In it, one finds the following entry:

> **la cuna**
>
> n. a twinge of sadness that there's no frontier left, that as the last explorer trudged with his armies toward a blank spot on the map, he didn't suddenly remember his daughter's upcoming piano recital and turn for home, leaving a new continent unexplored so we could set its mists and mountains aside as a strategic reserve of mystery, if only to answer more of our children's questions with "Nobody knows! Out there, anything is possible."

The question that occupies this chapter is prolific in raising new uncertainties, and as it fails to satisfy our intuition, it, too, questions our knowledge of the subject. Indeed, the computations carried out present intriguing regularities that seem to whisper some hidden structure that has escaped our repeated attempts to unveil it. Take, for instance, Figure 4.7. In it, some kind of fractal structure for the minimum value $V(n)$ of $V$ on $\mathbf{BinTree}_n$ is suggested, as well as a clear tendency to decrease with $n$. We have not been able to find any reason for either one or the other.

We have experimentally found the minima of $V$ on **BinTree**$_n$ for $n$ between 1 and $2^{16}$ using our algorithms, and other such regularities have been observed —of which, to this day we have only been able to prove one. We leave the verification of the rest as open problems.

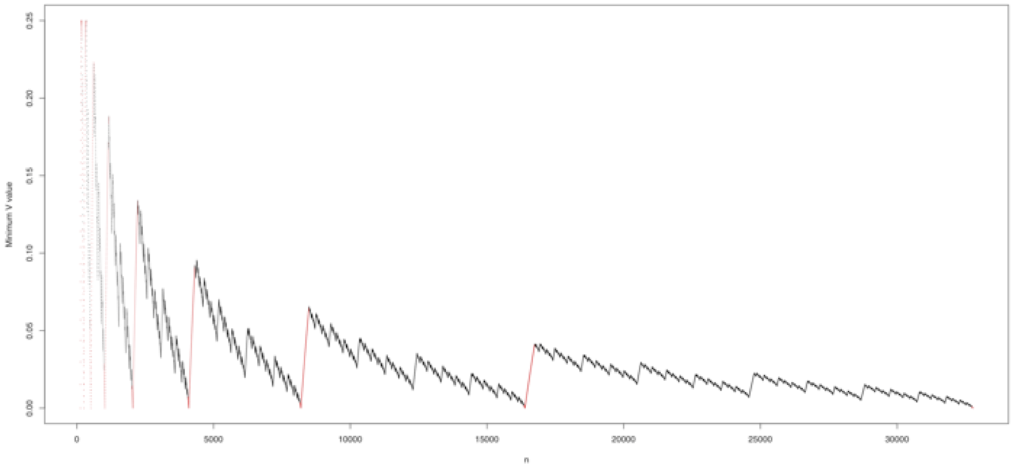


Figure 4.7: Scatter plot of the values of $V(n)$ for $n \in [2^7, 2^{15}]$. The values of $n$ for which this minimum is achieved at the maximally balanced trees are depicted in red.

### "Unicity" of the minima

For all tested $n \in \{1, 2, 3, \dots, 2^{16}\}$, the minimum value of $V$ on **BinTree**$_n$ has been found to occur at just *one* type of tree, modulo depth-equivalence; i.e., it has only been achieved at *only one* type of trees $T_n^{\mathbf{l}}$ for some sequence **l**. We have not been able to prove or disprove this property, although we conjecture it to hold for every $n \in \mathbb{N}$.

### Characterization of the intervals where the maximally balanced trees are minimal

Above, we have already proved that, for any $m$ large enough, the minima of the Variance of depths on **BinTree**$_n$ are never achieved at the maximally balanced trees for $n$ in an interval $[2^m + O(m^3), 2^{m+1} - 31]$. The right-hand side of this bound can be slightly improved to $2^{m+1} - 30$, and its tightness can be proven. This strikes us as capricious, but we now present the result and its proof in order to convince the reader of such property.

**Theorem 4.36.** *Let $n = 2^m + k$ for $n \in \mathbb{N}$ and $m = \lfloor \log_2(n) \rfloor \geq 5$. Then,*

*(i) If $k \geq 2^m - 29$, the minimum value of $V$ on* **BinTree**$_n$ *is attained exactly at the trees depth-equivalent to $T_n^{\text{bal}}$.*

*(ii) If $k = 2^m - 30$, the minimum value of $V$ on* **BinTree**$_n$ *is attained exactly at the trees depth-equivalent to $T_n^{\text{bal}}$ if $m \in \{5, 6, 7\}$, but at the trees of type $T_n^{(6)}$ if $m \geq 8$.*

*Proof.* Rewrite $n$ as $n = 2^{m+1} - x$, so that $k = 2^m - x$, and suppose henceforth that $x \leq 30$. Then, if $j \geq 1$, by Lemma 4.5 two possibilities exist: either

$$k + \sum_{i=1}^{j}(2^{l_i} - 1) \leq 2^m$$

or

$$k + \frac{1}{2}\sum_{i=1}^{j}(2^{l_i} - 2) > 2^m.$$

The first possibility cannot hold; indeed, for

$$k + \sum_{i=1}^{j}(2^{l_i} - 1) \geq 2^m - x + 2^5 - 1 > 2^m.$$

Therefore, if the minimum Variance of depths is achieved at a tree of type $T_n^{\mathbf{l}}$ with $\mathbf{l} \in \mathbb{N}^j$ and $j \geq 1$, then it must happen that

$$p_1 = 3 \cdot 2^m - k - \sum_{i=1}^{j}(2^{l_i} - 1) = 2^{m+1} + x - \sum_{i=1}^{j}(2^{l_i} - 1),$$

and it must have depth $m + 2$. Recall, from that same lemma, that not every such tree exists: indeed, it must also satisfy that $k + \frac{1}{2}\sum_{i=1}^{j}(2^{l_i} - 2) \leq 2^m$.

Recall the notations introduced in the proof of Theorem 4.22:

$$A(\mathbf{l}) = \textstyle\sum_{i=1}^{j}(2^{l_i} - l_i^2 - 1), \quad B(\mathbf{l}) = \sum_{i=1}^{j}(2^{l_i} - l_i - 1).$$

Then, we have that, by Lemma 4.7,

$$\begin{aligned} n \cdot V(T_n^{\mathbf{l}}) &= p_1 + \sum_{i=1}^{j} l_i^2 - \frac{(p_1 + \sum_{i=1}^{j} l_i)^2}{n} \\ &= 2^{m+1} + x - A(\mathbf{l}) - \frac{(2^{m+1} + x - B(\mathbf{l}))^2}{2^{m+1} - x} \\ &= \frac{1}{2^{m+1} - x}\left((2^{m+1} - x)(2^{m+1} + x - A(\mathbf{l})) - (2^{m+1} + x - B(\mathbf{l}))^2\right) \\ &= \frac{1}{2^{m+1} - x}\left(2^{m+1}(2B(\mathbf{l}) - A(\mathbf{l})) - B(\mathbf{l})^2 - 2x(2^{m+1} + x) + x(2B(\mathbf{l}) + A(\mathbf{l}))\right). \end{aligned}$$

Therefore,

$$n \cdot V(T_n^{\mathbf{l}}) \leq n \cdot V(T_n^{\mathrm{bal}}) = \frac{2k(2^m - k)}{2^m + k} = \frac{2x(2^m - x)}{2^{m+1} - x}$$

is equivalent to

$$2^{m+1}(2B(\mathbf{l}) - A(\mathbf{l})) - B(\mathbf{l})^2 - 2x(2^{m+1} + x) + x(2B(\mathbf{l}) + A(\mathbf{l})) \leq 2x(2^m - x),$$

which happens if, and only if,

$$(2^{m+1} - B(\mathbf{l}))(B(\mathbf{l}) - 3x) + (2^{m+1} - x)(B(\mathbf{l}) - A(\mathbf{l})) \leq 0. \tag{4.16}$$

Now, since $j \geq 1$ and $l_1 \geq 5$, $B(\mathbf{l}) > A(\mathbf{l})$ and, since $B(\mathbf{l}) \leq \sum_{i=5}^{m} 2^i < 2^{m+1}$ whereas $x \leq 30$, then Inequality (4.16) would imply that

$$(2^{m+1} - B(\mathbf{l}))(B(\mathbf{l}) - 90) + (2^{m+1} - 30)(B(\mathbf{l}) - A(\mathbf{l})) \leq 0. \tag{4.17}$$

Let us now consider the left-hand side of this inequality. Since $m \geq 5$, $(2^{m+1} - 30)(B(\mathbf{l}) - A(\mathbf{l})) > 0$. Furthermore, since $2^5 - 5 - 1 = 26, 2^6 - 6 - 1 = 57$ and $2^7 - 7 - 1 = 120$, it turns out that if some $l_i$ is larger than 6, then $B(\mathbf{l}) - 90 \geq 0$. Since $B(\mathbf{l}) < 2^{m+1}$, then Inequality (4.17) can only hold when either $j = 1$ and $l_1 \in \{5, 6\}$ or $j = 2$ and $\mathbf{l} = (5, 6)$; in any other case, $V(T_n^{\text{bal}}) < V(T_n^{\mathbf{l}})$. Let us now check these three cases.

- Suppose that $\mathbf{l} = (5)$. In this case, the necessary condition for the existence of $T_n^{(5)}$ is $x < \sum_{i=1}^{j}(2^{l_i - 1} - 1) = 2^4 - 1 = 15$. But then, since $A((5)) = 6$ and $B((5)) = 26$, Inequality (4.16) says

$$\begin{aligned}&(2^{m+1} - 26)(26 - 3x) + 20(2^{m+1} - x)\\ &\qquad \geq (2^6 - 26)(26 - 3x) + 20(2^6 - x) = 2268 - 134x > 0\end{aligned}$$

  whenever $x < 15$. Therefore, for this range of values, $V(T_n^{\text{bal}}) < V(T_n^{(5)})$.

- Suppose now that $\mathbf{l} = (6)$. In this case, the necessary condition for the existence of $T_n^{(6)}$ is $x < 2^5 - 1 = 31$, which is always satisfied since by hypothesis we imposed $x$ to be smaller or equal than 30. And so, Inequality (4.16) becomes

$$(2^{m+1} - 57)(57 - 3x) + 30(2^{m+1} - x) \leq 0,$$

  but it can be checked that if $m \in \{5, 6, 7\}$ and $x \leq 30$ this inequality does not hold, whereas if $m \geq 8$ and $x \leq 29$,

$$(2^{m+1} - 57)(57 - 3x) + 30(2^{m+1} - x) \geq (2^9 - 57)(57 - 87) + 30(2^9 - 29) > 0.$$

  When $x = 30$,

$$(2^{m+1} - 57)(57 - 90) + 30(2^{m+1} - 30) = 981 - 3 \cdot 2^{m+1} < 0,$$

  and thus in this case $V(T_n^{(6)}) < V(T_n^{\text{bal}})$.

- Suppose, finally, that $\mathbf{l} = (5, 6)$. In this case, the necessary condition on $x$ for the existence of $T_n^{(5,6)}$ is $x < 2^4 - 1 + 2^5 - 1 = 46$, and so is always satisfied. However, $A((5, 6)) = 33$ and $B((5, 6)) = 83$, and thus, whenever $m \geq 5$, the left-hand side of Inequality (4.17) is

$$\begin{aligned}&(2^{m+1} - 83)(83 - 3x) + 50(2^{m+1} - x)\\ &\qquad \geq (2^6 - 83)(83 - 3x) + 50(2^6 - x) = 7x + 1623 > 0,\end{aligned}$$

  and so the inequality itself is never satisfied. Hence, $V(T_n^{\text{bal}}) < V(T_n^{(5,6)})$.

Thus concludes the proof of the theorem. □

However, an exact expression for the lower bound is still unknown to us. That is: we have not been able to find a closed formula that, given $m$, ouputs the first $k_m \geq 0$ such that $V(T_{2^m + k_m}^{\text{bal}})$ is not minimum in $\mathbf{BinTree}_{2^m + k_m}$. For the range of tested values, we may approximate this value as $k_m \sim 0.1015m^{3.11}$, but in any case it is shown in the proof of Theorem 4.22 that $k_m$ is at most in $O(m^3)$. So, given the results presented by our tests, we conjecture it to be $\Theta(m^3)$.

### "Persistence" of the minima

For all tested $n = 2^m + k \geq 2^8$, $m = \lfloor \log_2(n) \rfloor$, some sort of "persistence" has been found, in the following sense: that, if the minimum of $V$ is achieved at the trees of the type $T_n^{\mathbf{l}}$ for $\mathbf{l} \in \mathbb{N}^j$ with $j \geq 1$ and $k < 2^m - \sum_{i=1}^{j}(2^{l_i} - 1)$, then the minimum Variance of depths on $\mathbf{BinTree}_{n+1}$ is also achieved at the trees of the type $T_{n+1}^{\mathbf{l}}$, with the same $\mathbf{l}$.

Notice that the fact that $j \geq 1$ is necessary for this property to hold. Indeed, for if $j = 0$, it may very well happen that for some $n$ the minimum of $V$ is achieved at $T_n^{\emptyset}$ on $\mathbf{BinTree}_n$ whereas on $\mathbf{BinTree}_{n+1}$ is not achieved at $T_{n+1}^{\emptyset}$. For example, the case already mentioned in the passage from $\mathbf{BinTree}_{183}$ to $\mathbf{BinTree}_{184}$.

We have not been able to prove this property either, but we conjecture it to hold for every $n \geq 2^8$.

### The Whisperer in Darkness

In relation to this last point, for all tested $n \in [2^m, 2^{m+1})$, this "persistence" presents intriguing regularities each time a series of $T_n^{\mathbf{l}}$ changes; i.e., the sequence formed by the lengths of the segments of consecutive numbers $n$ of leaves such that the minimum is achieved at a tree of type $T_n^{\mathbf{l}}$ for the same $\mathbf{l}$ hints at some hidden structure which remains veiled.

Here, we present the sequences corresponding to $m \in \{12, 13\}$. Take the first one, presented in Table 4.1 in reversed order —so that the pattern is more easily spotted. The interpretation of the numbers presented is that, when $n$ descends from $2^{13} - 1$ to $2^{12}$, for the first 29 values the trees $T_n^{\mathbf{l}}$ achieving the minimum $V$ value on $\mathbf{BinTree}_n$ have the same $\mathbf{l}$ vector (as we have already mentioned (Theorem 4.36), it is $\mathbf{l} = \emptyset$); then, the same happens with the next two values of $n$; afterwards, the same goes for the next 25 values; and so on. As we can see, the sequence ends in 52 each two lines, in 88 each four lines, and in 132 each eight lines.

As $m$ increases (see, for instance, the sequence associated to $m = 13$ in Table 4.2), the different sequences associated with it present the same pattern with practically the same numbers, and only small perturbations as if the sequence for $m = 12$ had "lost some information" due to its "contraction".

We present no conjecture, no hypothesis apart from what is obvious: that the conscience behind these figures did not present them randomly, that these patterns, their disposition, far from being hazardous, have been conceived, measured, by an entity whose whispers have hitherto gone unnoticed, whose will we have barely begun to unveil.

Table 4.1: Sequence, in reverse order, of the numbers of consecutive values $n \in [2^{12}, 2^{13})$ such that the trees $T_n^{\mathbf{l}}$ achieving the minimum $V$ value on $\mathbf{BinTree}_n$ have the same **l**.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 132 | | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 27 | 206 | | | | | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 132 | | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 7 | 442 | | | | | | | | | | | | |

Table 4.2: Sequence, in reverse order, of the numbers of consecutive values $n \in [2^{13}, 2^{14})$ such that the trees $T_n^{\mathbf{l}}$ achieving the minimum $V$ value on **BinTree**$_n$ have the same $\mathbf{l}$.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 1 | 130 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 28 | 204 | | | | | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 1 | 130 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 13 | 302 | | | | | | | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 1 | 130 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 28 | 204 | | | | | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 1 | 130 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 29 | 2 | 7 | 52 |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 21 | 88 | | |
| 29 | 2 | 25 | 12 | 29 | 2 | 18 | 28 | 29 | 2 | 25 | 12 | 26 | 598 | | |

CHAPTER 5

# The Quartet index

> […] It's impossible for a three legged table to wobble. Tables with four legs can, and often do. If the four legs are uneven, or if the floor's not flat, the table will wobble — resting first on one set of three legs, then tipping to another. Put more than four uneven legs on a table, and it can wobble every which way, always seeking to rest on three legs.
>
> A. Boyd, *The Power of Three*, No. 2533

THUS FAR, we can say that we have learned quite a lot about several balance indices. However, none of them is completely satisfying. The Sackin index is meaningless when the depth of the leaves of the tree is fixed, its minimum value is shared among all the trees depth-equivalent to the maximally balanced tree, but at least we know the expected value and variance under the Yule and Uniform models. The Colless index has a smaller number of trees attaining its minimum value, but their characterization is much more complex and their expected value under the Uniform model is not yet known; furthermore, and most importantly, it only has sense on bifurcating trees. And let us not talk about the Variance of depths! We do not even know where does its minimum fall for most numbers of leaves, but we *do* know that it is not attained at the tree we would like it to. On the other hand, the Cophenetic and Quadratic Colless indices *are* an improvement in so that they achieve their minimum value at a single tree, the maximally balanced one, and we are able to compute its expected value and variance under both the Uniform and Yule models and Chapters and; they also have a larger range of values, going from $O(n^2)$ in the case of the Sackin and Colless indices to $O(n^3)$ for any number of leaves $n \in \mathbb{N}$ —although, we should bear in mind, the Quadratic Colless index, too, is only sound when dealing with bifurcating trees.

But as good as these two indices are, we would not be completely Human if we were contented with things as they are, would we? So, can it get any better than that? Can we have a shape index such that, for example, its minimum and maximum values are attained at only one tree each, its first two moments can be computed for a more general probabilistic model of trees (yes! trees, not necessarily bifurcating trees but arbitrary, multifurcating ones) and (why not!) it has an even larger range of values? And can it, too, be such that it has some natural extensions to other directed graphs such as taxonomic or multilabelled trees? To the eternal glory of the Human mind, the answer is *yes*.

The first number of leaves that presents two different bifurcating tree shapes is four. Indeed, there are five different trees with five leaves, namely those in Figure 5.1. Now,

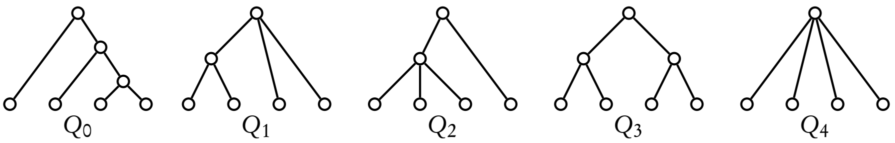


Figure 5.1: The five tree shapes in **Tree**$_4$.

the $Q_0$, $Q_3$ and $Q_4$ are already known to us: indeed, they are $T_4^{\text{cat}}$, $T_4^{\text{bal}}$, and $T_4^{\text{star}}$, respectively; i.e., the caterpillar, the maximally balanced tree and the star with four leaves. Of them, the maximally balanced tree and the caterpillar are the first bifurcating trees that can be distinguished, in the sense that 4 is the first number of leaves $n$ such that $|\mathbf{BinTree}_n| > 1$.

As we shall see, the order in which these trees have been listed is not arbitrary: they are ordered in increasing order of their number of automorphisms, *c'est à dire*, their symmetry. That we already knew when it came to the extreme trees: indeed, we have already proven that the star is the tree such that, the number of leaves being fixed, it has more automorphisms —namely, $4! = 24$—, and that the caterpillar is the tree with the least such number —namely, 2.

As the figure hints, in this chapter, we call the trees presented in the above figure $Q_0, \ldots, Q_4$, respectively, so that $Q_0 = T_4^{\text{cat}}$, $Q_3 = T_4^{\text{bal}}$ and $Q_4 = T_4^{\text{star}}$. Then, we have the following result.

**Lemma 5.1.** *Let $Q_i$ be the trees defined above. Then,*

$$|\text{Aut } Q_0| = 2, \quad |\text{Aut } Q_1| = 4, \quad |\text{Aut } Q_2| = 6, \quad |\text{Aut } Q_3| = 8, \quad |\text{Aut } Q_4| = 24.$$

*Proof.* These figures are easily deduced from Lemma 1.13 and Theorem 1.14, according to which, for every tree $T$,

$$|\text{Aut } T| = \prod_{u \in \mathring{V}(T)} \prod_i n_i(u)!$$

where, for every $u \in \mathring{V}(T)$, $n_1(u), n_2(u), \dots$ denote the cardinalities of the isomorphism classes among the subtrees rooted at the children of $u$. □

This induces an order between the trees in $\mathbf{Tree}_4$. Now suppose we assign to each one of them a number $q_i \in \mathbb{R}_{\geq 0}$, in such a manner that $q_i > q_j$ if, and only if, $|\mathrm{Aut}\, Q_i| > |\mathrm{Aut}\, Q_j|$ for any $(i, j) \in \{0, \dots, 4\}^2$; i.e., by Lemma 5.1, $q_{i+1} > q_i$ for any $i \in \{0, 1, 2, 3\}$. Any such assignment $(Q_i \mapsto q_i)$ is a map $\underset{\sim}{\mathrm{QI}} : \mathbf{Tree}_4 \to \mathbb{R}_{\geq 0}$, with $\underset{\sim}{\mathrm{QI}}(Q_i) = q_i$ for any $i \in \{0, \dots, 4\}$. This map induces a map $\underset{\sim}{\mathrm{QI}} : \mathbf{Tree} \to \mathbb{R}_{\geq 0}$ by means of

$$\begin{aligned} \underset{\sim}{\mathrm{QI}}(T) &= \sum_{Q \in \mathrm{Part}_4(L(T))} \underset{\sim}{\mathrm{QI}}(T(Q)) \\ &= \sum_{i=0}^{4} \left| \{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\} \right| \cdot q_i, \end{aligned}$$

where for every set of four leaves, or *4-tuple*, $Q \in \mathrm{Part}_4(L(T))$, $T(Q)$ denotes the subtree of $T$ induced by $Q$, as defined in page 4 in the Preliminaries. We shall call $T(Q)$ the *quartet* induced by $Q$.

Notice that, in fact, we can always consider $q_0$ to be $0$ without losing any information: indeed, for

$$\begin{aligned} \underset{\sim}{\mathrm{QI}}(T) &= \sum_{i=0}^{4} \left| \{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\} \right| \cdot q_i \\ &= q_0 \binom{|L(T)|}{4} + \sum_{i=1}^{4} \left| \{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\} \right| \cdot (q_i - q_0). \end{aligned}$$

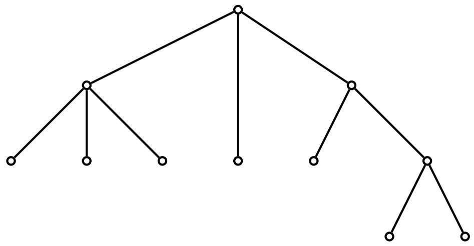
Now, the first term in this expression will remain constant whenever $|L(T)|$ does, and since it does not add any information, it can be circumvented by just imposing $q_0 = 0$. We shall do so in the rest of this chapter, and therefore, in fact, we define

$$\underset{\sim}{\mathrm{QI}}(T) = \sum_{i=1}^{4} \left| \{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\} \right| \cdot q_i.$$

We call such a map $\underset{\sim}{\mathrm{QI}} : \mathbf{Tree} \to \mathbb{R}_{\geq 0}$ a *Quartet index*,[1] and therefore the reasoning above defines a family of *Quartet indices*. We will, however, use the definite article *the* whenever we do not fix any values $q_i, i \in \{1, \dots, 4\}$, and when dealing with trees of a fixed number $n$ of leaves, we shall usually make the abuse of language of writing $\underset{\sim}{\mathrm{QI}} : \mathbf{Tree}_n \to \mathbb{R}_{\geq 0}$ when in fact we are considering a restriction of the $\underset{\sim}{\mathrm{QI}}$ defined above. We shall, also, usually say that a tree $T$ *has*, for some $i \in \{0, \dots, 4\}$, a quartet of the form $Q_i$ when $T(Q) = Q_i$ for some $Q \in \mathrm{Part}_4(L(T))$.

> **Example:**
> Consider the following tree $T \in \mathbf{Tree}_7$:

[1] In our paper [25], and upon request of one of the referees, we used the term *rooted Quartet index*, but we still prefer the name without the adjective "rooted", because all our trees are rooted, and therefore we omit this adjective in this report.



Now, it is easy to see that it has four quartets of shape $Q_0$, eighteen quartets of shape $Q_1$, four of shape $Q_2$, nine of shape $Q_3$ and none of shape $Q_4$. Thus,

$$\mathrm{QI}(T) = 18q_1 + 4q_2 + 9q_3.$$

Oftenly, we will restrict ourselves to bifurcating trees, and since a bifurcating tree can only have quartets of the form $Q_0$ and $Q_3$, and the former adds nought to the value of QI, we define the *Quartet index for bifurcating trees* QIB : **BinTree** ⟶ $\mathbb{N}$ by means of the relation

$$\mathrm{QIB}(T) = \frac{1}{q_3}\mathrm{QI}(T) = \left|\{Q \in \mathrm{Part}_4(L(T)) : Q = Q_3\}\right|.$$

By definition, the Quartet index is a shape index, in the sense that it is a function over tree shapes, although it can easily be extended to **PhyloTree** by precomposition by $\pi_1$. In this chapter we will argue that it is, furthermore, a balance index in the sense discussed in the Preliminaries. The intuition behind the use of QI as a balance index is that a highly balanced evolutive process should give rise to symmetrical evolutive histories of many small subsets of taxa. Then, by means of the values $q_i$, we associate to each 4-tuple of different leaves of the tree $T$ a number that quantifies the symmetry of the joint evolution of the species they represent, and then we add up these values over all 4-tuples of different leaves of $T$, expecting that, the most symmetrical a phylogenetic tree is, the most symmetrical will be its restrictions to subsets of 4 leaves.

**Remark 5.2.** The choice of the actual values of QI can be done following multiple criteria. It is natural, for example, to consider $q_i = i$ for all $i \in \{0, \ldots, 4\}$, or $q_i = 2^i$, or even $q_i = S(Q_0) - S(Q_i)$, where $S$ is the Sackin index, or an analogously derived value from the Cophenetic index. Notice, however, that the Colless index cannot be used to provide such values, since it only makes sense on bifurcating trees. Colless-like indices [86] could, however, be considered.

This chapter is organized as follows. In the first section we will prove several theorems on the computational aspects of QI: a recursive —on the children of the root— formula, and a linear procedure to compute it.

Then, the problem of characterizing, when presented with a number of leaves $n$, the trees that attain the extreme values of QI will be discussed. In the second section, we shall prove that the minimum Quartet index is attained exactly at the caterpillars, whereas the maximum exactly at the star or, if restricted to bifurcating trees, exactly at the maximally balanced trees, as it should. This allows us to consider the QI to be a balance index which, in opposition to the Sackin and Colless indices, increases when it finds a tree to be more balanced.

The third section will be a computation of the expected value and the variance of the Quartet index under any probabilistic model of trees satisfying the shape invariance and sampling consistency conditions. We will end with a discussion on possible extensions of QI to other sets of trees.

## 5.1 Computation of QI

In this section we shall be concerned with the recursive computation of the Quartet index. Let $n \in \mathbb{N}$, and $T \in \mathbf{Tree}_n$ a multifurcating tree. Let $\mathrm{trip}(T)$ be the number of *non-bifurcating triples* of $T$; that is, the number of subtrees of $T$ induced on 3-tuples of leaves that have shape $T_3^{\mathrm{star}}$:

$$\mathrm{trip}(T) = \left|\left\{\{x,y,z\} \in \mathrm{Part}_3(L(T)) : T(\{x,y,z\}) = T_3^{\mathrm{star}}\right\}\right|.$$

Notice, therefore, that if $T = T_1 * \cdots * T_m$, then

$$\mathrm{trip}(T) = \sum_{i=1}^{m} \mathrm{trip}(T_i) + \sum_{1 \le i_1 < i_2 < i_3 \le n} n_{i_1} n_{i_2} n_{i_3}.$$

where $n_i = |L(T_i)|$ for each $i \in \{1, \dots, m\}$. For instance, for the tree $T$ in page 185, $\mathrm{trip}(T) = 10$.

With these notations, we have the following theorem.

**Theorem 5.3.** *Let $T = T_1 * \cdots * T_m \in \mathbf{Tree}_n$, with $T_i \in \mathbf{Tree}_{n_i}$ for every $i \in \{1, \dots, m\}$. Then,*

$$\begin{aligned}
\mathrm{QI}(T) = \sum_{i=1}^{m} \mathrm{QI}(T_i) &+ q_4 \sum_{1 \le i_1 < i_2 < i_3 < i_4 \le m} n_{i_1} n_{i_2} n_{i_3} n_{i_4} \\
&+ q_3 \sum_{1 \le i_1 < i_2 \le m} \binom{n_{i_1}}{2}\binom{n_{i_2}}{2} + q_2 \sum_{1 \le i_1 < i_2 \le n} \left(n_{i_1}\mathrm{trip}(T_{i_2}) + n_{i_2}\mathrm{trip}(T_{i_1})\right) \\
&+ q_1 \sum_{1 \le i_1 < i_2 < i_3 \le m} \left(n_{i_2} n_{i_3}\binom{n_{i_1}}{2} + n_{i_1} n_{i_3}\binom{n_{i_2}}{2} + n_{i_1} n_{i_2}\binom{n_{i_3}}{2}\right).
\end{aligned}$$

*Proof.* For every $Q \in \mathrm{Part}_4(L(T))$ we have the following possibilities:

1. If $Q \in \mathrm{Part}_4(L(T_i))$ for some $i \in \{1, \dots, m\}$, then $T(Q) = T_i(Q)$, and therefore its contribution to $\mathrm{QI}(T)$ will be counted in $\mathrm{QI}(T_i)$.
2. If all four leaves of $Q$ belong to different maximal pending subtrees $T_i$, then $T(Q) = T_4^{\mathrm{star}}$ and it adds $q_4$ to the total QI of $T$. There are exactly $\sum_{1 \le i_1 < i_2 < i_3 < i_4 \le m} n_{i_1} n_{i_2} n_{i_3} n_{i_4}$ such quartets.
3. If two leaves of $Q$ belong to a maximal pending subtree $T_{i_1}$ and the other two to another, say $T_{i_2}$, then $T(Q) = T_4^{\mathrm{bal}}$ and it adds $q_3$ to the total $\mathrm{QI}(T)$. For each $(i_1, i_2) \in [m]^2$ with $i_1 < i_2$ there are $\binom{n_{i_1}}{2}\binom{n_{i_2}}{2}$ quartets of this type, and hence the third term in the sum is justified.
4. If three leaves, say $x$, $y$, and $z$, in $Q$ belong to a maximal pending subtree $T_i$ and the fourth to another, then two cases arise:

- If $T(\{x,y,z\}) = T_i(\{x,y,z\}) = T_3^{\mathrm{star}}$, then $T(Q) = Q_2$ and it adds $q_2$ to $\utilde{\mathrm{QI}}(T)$. There are $\sum_{1\le i_1<i_2\le m}\big(n_{i_1}\mathrm{trip}(T_{i_2}) + n_{i_2}\mathrm{trip}(T_{i_1})\big)$ such quartets.
- If $T(\{x,y,z\}) = T_i(\{x,y,z\}) = T_3^{\mathrm{cat}}$, then $T(Q) = T_4^{\mathrm{cat}}$, and it adds nought to the global $\utilde{\mathrm{QI}}$.

5. Finally, if two leaves of $Q$ belong to a maximal pending subtree $T_{i_1}$, and the other two to two different subtrees, say $T_{i_2}$ and $T_{i_3}$, then $T(Q) = Q_1$, and it contributes $q_1$ to the total addition. There are $\sum_{1\le i_1<i_2<i_3\le m}\Big(n_{i_2}n_{i_3}\binom{n_{i_1}}{2} + n_{i_1}n_{i_3}\binom{n_{i_2}}{2} + n_{i_1}n_{i_2}\binom{n_{i_3}}{2}\Big)$ quartets of this type.

   Thus concludes the proof. □

We deduce the following corollary.

**Corollary 5.4.** *Let $T = T_1 * T_2 \in \mathbf{BinTree}_n$ be a bifurcating tree, with $T_1 \in \mathbf{BinTree}_{n_1}$ and $T_2 \in \mathbf{BinTree}_{n_2}$. Then,*

$$\utilde{\mathrm{QIB}}(T) = \utilde{\mathrm{QIB}}(T_1) + \utilde{\mathrm{QIB}}(T_2) + \binom{n_1}{2}\binom{n_2}{2} = \sum_{v\in \mathring{V}(T)} \binom{\kappa(v_1)}{2}\binom{\kappa(v_2)}{2},$$

*where, for every $v \in \mathring{V}(T)$, $\{v_1, v_2\} = \mathrm{child}(v)$.*

*Proof.* The first equality is just a direct consequence of Theorem 5.3. The second one can be proved as an easy induction on the number of leaves of $T$, using the first equality in the induction step. □

**Remark 5.5.** Notice that Corollary 5.4 ensures that $\utilde{\mathrm{QIB}}$ is a bifurcating recursive shape index, in the sense introduced in the Preliminaries.

To end this section, we present a result that ensures that the $\utilde{\text{Qu}}$artet index is not computationally expensive in time: indeed, for it can be computed in linear time. But before that, we need a previous lemma.

**Lemma 5.6.** *Let $T \in \mathbf{Tree}_n$ and let $f : V(T) \to \mathbb{R}$ be a map such that $f(v)$ can be computed in constant time for any $v \in L(T)$, and in $O(\deg_{\mathrm{out}}(v))$ for any $v \in \mathring{V}(T)$, given its value on $\mathrm{child}(v)$. Then, the vector $(f(v))_{v\in V(T)}$ can be computed in time $O(n)$.*

*Proof.* We traverse the tree in post-order; that is, we begin by computing the value of $f$ on the leaves of the tree, then, over the parents of the leaves, and so forth. It can be computed in constant time over each leaf, and so the computation of $f$ over them takes already time $O(n)$. Then, for each internal node $v \in \mathring{V}(T)$, $f$ can be computed in time $O(\deg_{\mathrm{out}}(v))$ provided that $f(u)$ is known for every $u \in \mathrm{child}(v)$. Let $m_k$ be the number of internal nodes of out-degree $k$; then, the value of $f$ on all $v \in V(T)$ can be computed in time

$$O(n + \sum_k m_k k) = O(n + 2n - 2) = O(n)$$

because the number of edges in $T$ is, on the one hand, equal to $\sum_k m_k k$, and, on the other, at most, $2n - 2$ since $|V(T)| \le 2n - 1$. □

**Remark 5.7.** In particular, the result above ensures us that a linear time function on the vector $(f(v))_{v\in V(T)}$, such as its addition, can be computed in linear time.

Thus, Lemma 5.6 ensures that the computations in Corollary 5.4 can be performed in $O(n)$ time. Let us see that this is indeed the case in general.

**Theorem 5.8.** *If* $T \in \mathbf{Tree}_n$, $\underset{\smile}{\mathrm{QI}}(T)$ *can be computed in time* $O(n)$*.*

*Proof.* By the lemma above, the vector $(\kappa(v))_{v\in\mathring{V}(T)}$ can be computed in linear time. Now, in order to simplify the notations, for any $v \in \mathring{V}(T)$, let

$$\begin{aligned}
E_l(v) &= \sum_{\{v_1,\dots,v_l\}\subseteq \mathrm{child}(v)} \kappa(v_1)\cdots\kappa(v_l) \qquad \text{for } l \in \{2,\dots,\deg_{\mathrm{out}}(v)\}\\
F_1(v) &= \sum_{\{v_1,v_2,v_3\}\subseteq \mathrm{child}(v)} \left(\binom{\kappa(v_1)}{2}\kappa(v_2)\kappa(v_3) + \binom{\kappa(v_2)}{2}\kappa(v_1)\kappa(v_3) + \binom{\kappa(v_3)}{2}\kappa(v_1)\kappa(v_2)\right)\\
F_2(v) &= \sum_{\{v_1,v_2\}\subseteq \mathrm{child}(v)} \left(\kappa(v_1)\cdot \mathrm{trip}(T_{v_2}) + \kappa(v_2)\cdot \mathrm{trip}(T_{v_1})\right)\\
F_3(v) &= \sum_{\{v_1,v_2\}\subseteq \mathrm{child}(v)} \binom{\kappa(v_1)}{2}\binom{\kappa(v_2)}{2}
\end{aligned}$$

in such a way that

$$\begin{aligned}
\mathrm{trip}(T) &= \sum_{v\in\mathring{V}(T)} E_3(v)\\
\underset{\smile}{\mathrm{QI}}(T) &= \sum_{v\in\mathring{V}(T)} (q_1F_1(v) + q_2F_2(v) + q_3F_3(v) + q_4E_4(v)).
\end{aligned}$$

We want to prove that each one of the vectors

$$\big(F_1(v)\big)_{v\in\mathring{V}(T)},\ \big(F_2(v)\big)_{v\in\mathring{V}(T)},\ \big(F_3(v)\big)_{v\in\mathring{V}(T)},\ \big(E_4(v)\big)_{v\in\mathring{V}(T)}$$

can be computed in time $O(n)$. Indeed, if we succeed in proving this, then $\underset{\smile}{\mathrm{QI}}$ will be shown to be computed in linear time.

In order to prove this, we will use the *Newton-Girard formulæ* (see, for instance, [75, §I.2]): given a multiset of numbers $X = \{x_1,\dots,x_k\}$, if we set

$$P_l(X) = \textstyle\sum_{i=1}^k x_i^l \quad \text{and} \quad E_l(X) = \sum_{1\le i_1<\cdots<i_l\le k} x_{i_1}\cdots x_{i_l}$$

then,

$$E_l(X) = \frac{1}{l!}\det\begin{pmatrix}
P_1(X) & 1 & 0 & \cdots & 0 & 0\\
P_2(X) & P_1(X) & 2 & \cdots & 0 & 0\\
P_3(X) & P_2(X) & P_1(X) & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
P_{l-1}(X) & P_{l-2}(X) & P_{l-3}(X) & \cdots & P_1(X) & l-1\\
P_l(X) & P_{l-1}(X) & P_{l-2}(X) & \cdots & P_2(X) & P_1(X)
\end{pmatrix}.$$

If we consider $l$ as a fixed parameter, each $P_l(X)$ can be computed in time $O(k)$, and in this case the expression of $E_l(X)$ as an $l \times l$ determinant can be computed in time $O(l^4 k) = O(k)$ using for instance Bareiss cubic algorithm to compute determinants.

Now consider, for every $v \in \mathring{V}(T)$, the multiset $X_v = \{\kappa(u) : u \in \mathrm{child}(v)\}$. Then, every $E_l(v) = E_l(X_v)$ can be computed in time $O(\deg_{\mathrm{out}}(v))$, and therefore the whole vector $(E_l(v))_{v\in\mathring{V}(T)}$ can be computed in time $O(n)$, by Lemma 5.6. In particular, that is the case for $(E_3(v))_{v\in\mathring{V}(T)}$ and $(E_4(v))_{v\in\mathring{V}(T)}$. And so, again by Lemma 5.6, the recursion

$$\mathrm{trip}(T_v) = \sum_{u\in\mathrm{child}(v)} \mathrm{trip}(T_u) + E_3(v) = \sum_{u\in\mathring{V}(T_v)} E_3(u)$$

can also be computed in linear time, and so does the vector $\big(\mathrm{trip}(T_v)\big)_{v\in\mathring{V}(T)}$. Therefore,

$$\begin{aligned}
F_3(v) &= \sum_{\{v_1,v_2\}\subseteq\mathrm{child}(v)} (\kappa(v_1)\mathrm{trip}(T_{v_2}) + \kappa(v_2)\mathrm{trip}(T_{v_1}))\\
&= \left(\sum_{v_i\in\mathrm{child}(v)} \kappa(v_i)\right) \sum_{v_j\in\mathrm{child}(v)} \mathrm{trip}(T_{v_j}) - \sum_{v_i\in\mathrm{child}(v)} \kappa(v_i)\mathrm{trip}(T_{v_i})\\
&= \kappa(v)(\mathrm{trip}(T_v) - E_3(v)) - \sum_{v_i\in\mathrm{child}(v)} \kappa(v_i)\mathrm{trip}(T_{v_i})
\end{aligned}$$

and this implies that every $F_3(v)$ is computed in time $O(\deg_{\mathrm{out}}(v))$, and hence the whole vector $(F_3(v))_{v\in\mathring{V}(T)}$ can be computed in time $O(n)$.

It remains to show that the vectors $\big(F_1(v)\big)_{v\in\mathring{V}(T)}$ and $\big(F_2(v)\big)_{v\in\mathring{V}(T)}$ can also be computed in linear time. Let us focus first on the latter. By definition,

$$\begin{aligned}
F_2(v) &= \sum_{\{v_1,v_2\}\subseteq\mathrm{child}(v)} \binom{\kappa(v_1)}{2}\binom{\kappa(v_2)}{2}\\
&= \frac{1}{4}\sum_{\{v_1,v_2\}\subseteq\mathrm{child}(v)} \kappa(v_1)\kappa(v_2) + \frac{1}{4}\sum_{\{v_1,v_2\}\subseteq\mathrm{child}(v)} \kappa(v_1)^2\kappa(v_2)^2\\
&\qquad - \frac{1}{4}\sum_{\{v_1,v_2\}\subseteq\mathrm{child}(v)} \Big(\kappa(v_1)^2\kappa(v_2) + \kappa(v_1)\kappa(v_2)^2\Big).
\end{aligned}$$

Now, the two summands in the above expression can be considered to be $E_2(v)$ and $E_2(\{\kappa(u)^2 : u \in \mathrm{child}(v)\})$, and therefore they are computed in time $O(\deg_{\mathrm{out}} v)$. As far as the substrahend goes,

$$\begin{aligned}
&\sum_{\{v_1,v_2\}\subseteq\mathrm{child}(v)} (\kappa(v_1)^2\kappa(v_2) + \kappa(v_1)\kappa(v_2)^2)\\
&\qquad = \left(\sum_{v_i\in\mathrm{child}(v)} \kappa(v_i)^2\right) \sum_{v_j\in\mathrm{child}(v)} \kappa(v_j) - \sum_{v_i\in\mathrm{child}(v)} \kappa(v_i)^3
\end{aligned}$$

and hence, $F_2(v)$ can be computed in time $O(\deg_{\mathrm{out}}(v))$. Thus, by Lemma 5.6, the whole vector $(F_2(v))_{v\in\mathring{V}(T)}$ can be computed in time $O(n)$.

Finally, let us consider $F_1$:

$$\begin{aligned}
F_1(v) &= \sum_{\{v_1,v_2,v_3\}\subseteq \mathrm{child}(v)} \left(\binom{\kappa(v_1)}{2}\kappa(v_2)\kappa(v_3) + \binom{\kappa(v_2)}{2}\kappa(v_1)\kappa(v_3) + \binom{\kappa(v_3)}{2}\kappa(v_1)\kappa(v_2)\right)\\
&= \frac{1}{2}\sum_{\{v_1,v_2,v_3\}\subseteq \mathrm{child}(v)} \kappa(v_1)\kappa(v_2)\kappa(v_3)\big(\kappa(v_1)+\kappa(v_2)+\kappa(v_3)-3\big)\\
&= \frac{1}{2}\left(\sum_{\{v_1,v_2,v_3\}\subseteq \mathrm{child}(v)} \kappa(v_1)\kappa(v_2)\kappa(v_3)\right)\sum_{v_i\in \mathrm{child} v}\kappa(v_i)\\
&\quad -2\sum_{\{v_1,v_2,v_3,v_4\}\subseteq \mathrm{child}(v)} \kappa(v_1)\kappa(v_2)\kappa(v_3)\kappa(v_4) - \frac{3}{2}E_3(v)\\
&= \frac{1}{2}E_3(v)E_1(v) - 2E_4(v) - \frac{3}{2}E_3(v),
\end{aligned}$$

and thus, it can also be computed in time $O(\deg_{\mathrm{out}}(v))$, and therefore the whole vector $(F_1(v))_{v\in\mathring{V}(T)}$ can be computed in time $O(n)$, as we wanted to prove. □

## 5.2 Extreme values and the trees that attain them

As we have already mentioned, a *sine qua non* condition for QI to be considered a proper balance index is that its extreme values must be reached at the trees that are considered to represent the extreme cases of balance: that is, it must classify the stars, in the multifurcating case, and the maximally balanced trees, in the bifurcating case, as most balanced, and the caterpillars as most unbalanced. This shall indeed be the case. As it often happens, this quest for the extreme values will be more difficult when it comes to characterizing the maximum QIB index: i.e., the most balanced (according to the Quartet index) bifurcating tree.

Firstly, let us prove that the minimum QI is attained exactly at the caterpillars; i.e. the least balanced trees according to the Quartet index are the caterpillars.

**Theorem 5.9.** *Let $n \in \mathbb{N}_{\geq 1}$. The minimum Quartet index in* $\mathbf{Tree}_n$ *is reached exactly at the caterpillars, and it is*

$$\mathrm{QI}(T_n^{\mathrm{cat}}) = 0.$$

*Proof.* Since $q_i > 0$ for any $i > 0$, that is, for any $Q_i \neq T_4^{\mathrm{cat}}$, the only way in which

$$\mathrm{QI}(T) = \sum_{i=1}^{4} \left|\{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\}\right| \cdot q_i$$

can be $0$ is having only quartets of the form $Q_0 = T_4^{\mathrm{cat}}$. Let us see that such a tree must be a caterpillar itself. Indeed, let $T \in \mathbf{Tree}_n$ be a tree different from the caterpillar. Then, two possibilities exist:

- either there is some $v \in \mathring{V}(T)$ with $\deg_{\mathrm{out}}(v) \geq 3$, and thus $T$ contains at least one quartet of the forms $Q_1$, $Q_2$ or $Q_4$;
- or $T$ is bifurcating and it contains at least two cherries, which induce a quartet of the form $Q_3$.

Now, it remains to be seen that $QI(T_n^{\mathrm{cat}}) = 0$, but this is obvious because all the subtrees of a caterpillar are caterpillars. □

Secondly, we shall prove that the most balanced of all trees with $n$ leaves are, according to QI, the stars.

**Theorem 5.10.** *Let $n \in \mathbb{N}_{\geq 1}$. The maximum Quartet index in* **Tree**$_n$ *is reached exactly at the stars, and it is*

$$QI(T_n^{\mathrm{star}}) = \binom{n}{4} q_4.$$

*Proof.* The fact that $QI(T_n^{\mathrm{star}}) = \binom{n}{4} q_4$ is obvious, since the only quartets that a star presents are stars with four leaves, and there are exactly $\binom{n}{4}$ quartets in each tree with $n$ leaves. Now, since $q_4 > q_3 > q_2 > q_1 > q_0 = 0$, it is clear that the maximum value of QI restricted to trees with $n$ leaves must be reached at the stars.

Now let us see that no other tree in **Tree**$_n$ can reach the same value. Let $T \in \mathbf{Tree}_n$. If $T \neq T_n^{\mathrm{star}}$, then, by definition, $|\mathring{V}(T)| \geq 2$. But this means that $T$ will present some quartet of shape $Q_i$ different from $T_4^{\mathrm{star}}$, and hence with a lower $q_i$, and thus $QI(T) < \binom{n}{4} q_4$. □

These two results should not surprise us, since they are true almost by definition: this is why the Quartet index was defined as it was to begin with. The next subsection will be dedicated, however, to show that, when restricted to bifurcating trees, the maximally balanced trees are *the only trees* that attain the maximum QIB, and hence the most balanced according to it. Since the caterpillars are bifurcating trees, in particular they will be the least balanced bifurcating trees, too.

Furthermore, these two results set the range of the Quartet index to go from 0 to $\binom{n}{4}$, an order of magnitude higher than that of the Quadratic Colless or the Cophenetic (page 19) indices (the range of this last one, going from 0 to $\binom{n}{3}$, was so far the balance index for multifurcating trees with the widest range [85]).

### 5.2.1 The maximum QIB

In this subsection we shall characterize the maximum values of QIB and the trees that attain them. They turn out to be, as in the case of the Quadratic Colless or the Cophenetic indices, exactly the maximally balanced trees, and the proof is quite similar to that of the characterization of the minimum Cophenetic index given in [85].

We shall begin by proving a series of lemmata that will help us solve the main issue. The first one shows that QIB is local, in the sense that if two bifurcating trees differ only in a rooted subtree, the difference in their Quartet indices is equal to the difference between those of these different subtrees. QIB shares this property with many other shape indices, like the Sackin, the Cophenetic, and the Quadratic and classical Colless indices.

**Lemma 5.11.** *Let $T_0$ be a bifurcating tree, let $z \in L(T_0)$, and let $T, T'$ be two trees obtained by appending to the leaf $z$ in $T_0$ the rooted bifurcating subtrees $T_z$ and $T'_z$, respectively, with $L(T_z) = L(T'_z)$. Then,*

$$QIB(T') - QIB(T) = QIB(T'_z) - QIB(T_z).$$

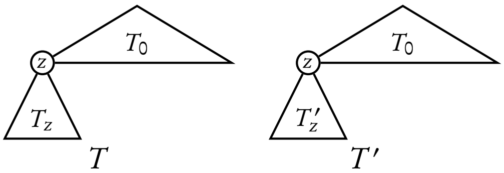


Figure 5.2: The trees $T$ and $T'$ in the statement of Lemma 5.11.

*Proof.* Let $Q = \{a, b, c, d\} \in \mathrm{Part}_4(L(T)) = \mathrm{Part}_4(L(T'))$. Then:

- If $Q \cap L(T_z) = \emptyset$, then $T(Q) = T'(Q) = T_0(Q)$.
- If $Q \cap L(T_z) = \{d\}$ —for instance—, then $T(Q) = T'(Q) = T_0(\{a, b, c, z\})$.
- If $Q \cap L(T_z) = \{c, d\}$ —for instance—, then two cases arise: either $T_0(\{a, b, z\}) = (a, (b, z))$, and in this case $T(Q) = T'(Q) = (a, (b, (c, d)))$; or $T_0(\{a, b, z\}) = ((a, b), z)$, and so $T(Q) = T'(Q) = ((a, b), (c, d))$.
- If $Q \cap L(T_z) = \{b, c, d\}$ —for instance—, then $T(Q) = T'(Q)$ since they are both caterpillars ($T_z$ and $T_z'$ are both bifurcating).

Thus, the only 4-tuples of leaves $Q$ that may define different quartets in $T$ and $T'$ are those contained in $L(T_z) = L(T_z')$, in which case $T(Q) = T_z(Q)$ and $T'(Q) = T_z'(Q)$, and hence

$$\utilde{\mathrm{QIB}}(T') - \utilde{\mathrm{QIB}}(T) = \utilde{\mathrm{QIB}}(T_z') - \utilde{\mathrm{QIB}}(T_z)$$

as we claimed. □

**Lemma 5.12.** *Let $T \in$* **BinTree**$_n$ *be the tree depicted in Figure 5.3 (a). For every $i \in \{1, 2, 3, 4\}$, let $n_1 = |L(T_i)|$, and assume that $n_1 > n_3$ and $n_2 > n_4$. Then, $\utilde{\mathrm{QIB}}(T)$ is not maximum in* **BinTree**$_n$.

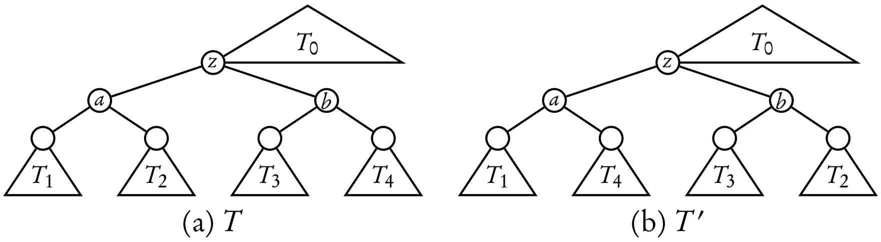


Figure 5.3: (a) The tree $T$ in the statement of Lemma 5.12. (b) The tree $T'$ in the proof of Lemma 5.12.

*Proof.* Let $T'$ be the tree depicted in Figure 5.3 (b): it is constructed by interchanging, in $T$, the subtrees $T_2$ and $T_4$. We shall prove that $\utilde{\mathrm{QIB}}(T') > \utilde{\mathrm{QIB}}(T)$. By the previous lemma,

$$\utilde{\mathrm{QIB}}(T') - \utilde{\mathrm{QIB}}(T) = \utilde{\mathrm{QIB}}(T_z') - \utilde{\mathrm{QIB}}(T_z).$$

Let us now compute this difference. By applying Corollary 5.4,

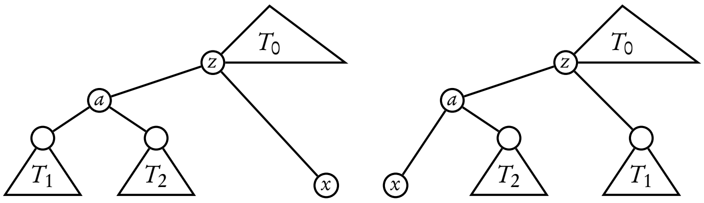
$$\begin{aligned}\utilde{\mathrm{QIB}}(T_z) &= \utilde{\mathrm{QIB}}(T_1) + \utilde{\mathrm{QIB}}(T_2) + \utilde{\mathrm{QIB}}(T_3) + \utilde{\mathrm{QIB}}(T_4)\\ &\quad + \binom{n_1+n_2}{2}\binom{n_3+n_4}{2} + \binom{n_1}{2}\binom{n_2}{2} + \binom{n_3}{2}\binom{n_4}{2}\\ \utilde{\mathrm{QIB}}(T_z') &= \utilde{\mathrm{QIB}}(T_1) + \utilde{\mathrm{QIB}}(T_2) + \utilde{\mathrm{QIB}}(T_3) + \utilde{\mathrm{QIB}}(T_4)\\ &\quad + \binom{n_1+n_3}{2}\binom{n_2+n_4}{2} + \binom{n_1}{2}\binom{n_3}{2} + \binom{n_2}{2}\binom{n_4}{2}\end{aligned}$$

and therefore, their difference is

$$\utilde{\mathrm{QIB}}(T_z') - \utilde{\mathrm{QIB}}(T_z) = \frac{1}{2}(n_1 - n_3)(n_2 - n_4)(n_1n_3 + n_2n_4) > 0,$$

since $n_1 > n_3$ and $n_2 > n_4$. □

**Lemma 5.13.** *Let $T \in \mathbf{BinTree}_n$ be a tree containing a leaf $x$ whose sibling has at least three descendant leaves. Then, $\utilde{\mathrm{QIB}}(T)$ is not maximum in $\mathbf{BinTree}_n$.*



(a) $T$ (b) $T'$

Figure 5.4: (a) The tree $T$ in the statement of Lemma 5.13, where $|L(T_1)| + |L(T_2)| \geq 3$. (b) The tree $T'$ in the proof of Lemma 5.13.

*Proof.* Let $T \in \mathbf{BinTree}_n$ be the tree depicted in Figure 5.4 (a), let $n_1 = |L(T_1)|$ and $n_2 = |L(T_2)|$, so that $n_1 + n_2 \geq 3$, and suppose $n_1 \geq n_2$, which in particular implies that $n_1 \geq 2$. Let $T' \in \mathbf{BinTree}_n$ be the tree depicted in Figure 5.4 (b), obtained from $T$ by interchanging the leaf $x$ and the rooted subtree $T_1$. We shall prove that $\utilde{\mathrm{QIB}}(T') > \utilde{\mathrm{QIB}}(T)$. Now, by Lemma 5.11,

$$\utilde{\mathrm{QIB}}(T') - \utilde{\mathrm{QIB}}(T) = \utilde{\mathrm{QIB}}(T_z') - \utilde{\mathrm{QIB}}(T_z)$$

and, again, by Theorem 5.3,

$$\begin{aligned}\utilde{\mathrm{QIB}}(T_z) &= \utilde{\mathrm{QIB}}(T_1) + \utilde{\mathrm{QIB}}(T_2) + \binom{n_1}{2}\binom{n_2}{2}\\ \utilde{\mathrm{QIB}}(T_z') &= \utilde{\mathrm{QIB}}(T_1) + \utilde{\mathrm{QIB}}(T_2) + \binom{n_1}{2}\binom{n_2+1}{2}\end{aligned}$$

so that

$$\utilde{\mathrm{QIB}}(T_z') - \utilde{\mathrm{QIB}}(T_z) = n_2\binom{n_1}{2} > 0$$

because $n_1 \geq 2$. □

Finally, thanks to these last two lemmata, we can prove the following theorem, which is the main result of this section.

**Theorem 5.14.** *Let $n \in \mathbb{N}_{\geq 1}$. The maximum Quartet index in* $\mathbf{BinTree}_n$ *is reached exactly at the maximally balanced trees.*

*Proof.* Let $T \in \mathbf{BinTree}_n$ be a tree that is not maximally balanced: we shall prove that it cannot present the maximum Quartet index. Since $T$ is not maximally balanced, there exists an internal node, $v \in \mathring{V}(T)$, such that $|\kappa(v_1) - \kappa(v_2)| \geq 2$ for $\{v_1, v_2\} = \text{child}(v)$. Furthermore, suppose that $v$ is such that all its proper descendant nodes are balanced (which can be supposed since $T$ is finite).

If $v_2$ is a leaf, then by Lemma 5.13, since $T_{v_1}$ must have more than three leaves by assumption, $T$ cannot present maximum QIB; the argument is analogous if we consider $v_1$ to be a leaf. Therefore, suppose that neither $v_1$ nor $v_2$ are leaves, and let $T_1, T_2$ be the trees rooted at the children of $v_1$ and $T_3, T_4$ those rooted at the children of $v_2$, and set $n_i = |L(T_i)|$ for $i \in \{1, 2, 3, 4\}$.

Assume, without loss of generality, that $n_1 \geq n_2$, that $n_3 \geq n_4$, and that $n_1 + n_2 \geq n_3 + n_4$. Then, since by assumption $v$ is not balanced, we have that

$$n_1 + n_2 \geq n_3 + n_4 + 2.$$

Since $v_1$ and $v_2$ are balanced, $n_1 \in \{n_2, n_2 + 1\}$ and $n_3 \in \{n_4, n_4 + 1\}$, and therefore $n_1 > n_3$, Indeed, for suppose $n_1 \leq n_3$: then $n_1 + n_2 \geq n_3 + n_4 + 2$ would imply $n_2 \geq n_4 + 2$ and then $n_3 - n_4 \geq n_1 - n_2 + 2 \geq 2$, against the assumption that $v_2$ is balanced. Therefore, $n_1 \geq n_3 + 1$ and hence $n_2 \geq n_1 - 1 \geq n_3 \geq n_4$. But if $n_2 = n_4$, then $n_1 - 1 = n_2 = n_3 = n_4$, contradicting the fact that $v$ is not balanced.

So, finally, we conclude that $n_1 > n_3$ and $n_2 > n_4$, but then, by Lemma 5.12, $\text{QIB}(T)$ cannot be maximum. □

For every $n \in \mathbb{N}$, let $\text{qib}(n)$ be the maximum value of $\text{QIB}(T)$ for $T \in \mathbf{BinTree}_n$. Since $T_n^{\text{bal}} = T_{\lceil n/2 \rceil}^{\text{bal}} * T_{\lfloor n/2 \rfloor}^{\text{bal}}$, the last theorem and Corollary 5.4 imply the following recurrence for the sequence $\text{qib}(n)$.

**Corollary 5.15.** $\text{qib}(1) = 0$, *and, for $n \geq 2$,*

$$\text{qib}(n) = \text{qib}\left(\left\lceil \frac{n}{2} \right\rceil\right) + \text{qib}\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \binom{\lceil \frac{n}{2} \rceil}{2}\binom{\lfloor \frac{n}{2} \rfloor}{2}.$$

We have not been able to solve this recurrence, but we can easily deduce from it the order of growth of $\text{qib}(n)$.

**Corollary 5.16.** $\text{qib}(n)$ *is in* $\Theta(n^4)$.

*Proof.* With the convention of the statement of the Master Theorem for solving recurrences as given in [21, Thm. 4.1], the last corollary implies that the sequence $\text{qib}(n)$ satisfies a recurrence of the form

$$\text{qib}(n) = 2 \cdot \text{qib}(n/2) + F(n)$$

with $F(n) = \binom{\lceil n/2 \rceil}{2}\binom{\lfloor n/2 \rfloor}{2}$ in $\Omega(n^4) = \Omega(n^{\log_2(2)+3})$ and satisfying that

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

Therefore, by case (3) in that theorem, $\text{qib}(n)$ is in $\Theta(F(n))$, i.e., in $\Theta(n^4)$. □

The sequence qib($n$) was not contained in the *On-Line Encyclopedia of Integer Sequences* [108] until we submitted it, being currently sequence A300445 in it. Its values for $n \in \{4, \dots, 20\}$ are

$$1, 3, 9, 19, 38, 64, 106, 162, 243, 343, 479, 645, 860, 1110, 1424, 1790, 2237.$$

An easy induction exercise shows that

$$\mathrm{qib}(2^n) = \left(\frac{4}{7(2^n - 3)} + \frac{3}{7}\right)\binom{2^n}{4}$$

and hence, in particular, $\mathrm{qib}(2^n)/\binom{2^n}{4}$ tends to $\frac{3}{7}$ as $n \to \infty$.

As we have already said, the comparison of tree shapes between trees with different numbers of leaves can only be performed when the index is normalized. In this case, being $\min\{\utilde{\mathrm{QI}}(T) : T \in \mathbf{Tree}_n\} = \min\{\utilde{\mathrm{QIB}}(T) : T \in \mathbf{BinTree}_n\} = 0$ for all $n$, we derive two normalized indices, one for multifurcating trees and the other for bifurcating trees:

$$\overline{\utilde{\mathrm{QI}}}(T) = \frac{\utilde{\mathrm{QI}}(T)}{q_4\binom{n}{4}}, \quad \text{and} \quad \overline{\utilde{\mathrm{QIB}}}(T) = \frac{\utilde{\mathrm{QIB}}(T)}{\mathrm{qib}(n)}$$

where qib($n$) is computed by means of the aforementioned recursion.

## 5.3 The expected value and variance under sampling consistent probabilistic models

In this section we are going to compute the expected value and the variance of $\utilde{\mathrm{QI}}$ under probabilistic models for phylogenetic trees satisfying certain conditions. Notice that, by Remark 5.5, we could also compute the expected value and variance of $\utilde{\mathrm{QIB}}$ by means of the results given in Section 1.3.4, as $\utilde{\mathrm{QIB}}$ is a recursive shape index; however, we shall now give other proofs, and then in Section 5.3.1 we will give the proofs given *via* the results given in the Preliminaries. Now, we can readily extend the definition of $\utilde{\mathrm{QI}}$ to phylogenetic trees by simply setting $\utilde{\mathrm{QI}}(T, \lambda) = \utilde{\mathrm{QI}}(T)$. Let $P_n : \mathbf{PhyloTree}_n \to [0, 1]$ be a probabilistic model for phylogenetic trees, whose induced model for tree shapes is $P_n^* : \mathbf{Tree}_n \to [0, 1]$ where

$$P_n^*(T) = \sum_{(T,\lambda) \in \mathbf{PhyloTree}_n} P_n(T, \lambda).$$

Then, we can consider the random variables $\utilde{\mathrm{QI}}_n : \mathbf{PhyloTree}_n \to [0, 1]$ and $\utilde{\mathrm{QI}}_n^* : \mathbf{Tree}_n \to [0, 1]$, defined by choosing a tree in their respective codomains with probability $P_n$ or $P_n^*$, respectively, and then computing its $\utilde{\mathrm{QI}}$.

**Lemma 5.17.** *For every $n \geq 1$, the distributions of $\utilde{\mathrm{QI}}_n$ and $\utilde{\mathrm{QI}}_n^*$ are the same. In particular, their expected values and variances are the same.*

*Proof.* Let $f_{\underset{\sim}{\mathrm{QI}}_n}$ and $f_{\underset{\sim}{\mathrm{QI}}^*_n}$ be the probability density functions of $\underset{\sim}{\mathrm{QI}}_n$ and $\underset{\sim}{\mathrm{QI}}^*_n$, respectively. Then, for any $x_0 \in \mathbb{R}$,

$$\begin{aligned} f_{\underset{\sim}{\mathrm{QI}}_n}(x_0) &= \sum_{\substack{(T,\lambda)\in\mathbf{PhyloTree}_n\\ \underset{\sim}{\mathrm{QI}}(T,\lambda)=x_0}} P_n(T,\lambda) = \sum_{\substack{T\in\mathbf{Tree}_n\\ \underset{\sim}{\mathrm{QI}}(T)=x_0}} \sum_{(T,\lambda)\in\mathbf{PhyloTree}_n} P_n(T,\lambda)\\ &= \sum_{\substack{T\in\mathbf{Tree}_n\\ \underset{\sim}{\mathrm{QI}}(T)=x_0}} P^*_n(T) = f_{\underset{\sim}{\mathrm{QI}}^*_n}(x_0). \end{aligned}$$

Therefore, $f_{\underset{\sim}{\mathrm{QI}}_n} = f_{\underset{\sim}{\mathrm{QI}}^*_n}$. □

Notice that, then, both the expected value and the variance of $\underset{\sim}{\mathrm{QI}}$ coincide with those of $\underset{\sim}{\mathrm{QI}}^*$. Next theorem computes the expected value of $\underset{\sim}{\mathrm{QI}}_n$ under any probabilistic model for trees that is sampling consistent.

**Theorem 5.18.** *Let $P_n : \mathbf{PhyloTree}_n \to [0,1]$ be a probabilistic model for phylogenetic trees such that $P^*_n$ is sampling consistent. Then,*

$$E_P(\underset{\sim}{\mathrm{QI}}_n) = E_{P^*}(\underset{\sim}{\mathrm{QI}}^*_n) = \binom{n}{4}\sum_{i=1}^{4} P^*_4(Q_i)q_i.$$

*Proof.* The first equality is a direct consequence of the previous lemma. The second equality can be computed as follows:

$$\begin{aligned} E_{P^*}(\underset{\sim}{\mathrm{QI}}^*) &= \sum_{T\in\mathbf{Tree}_n} \underset{\sim}{\mathrm{QI}}(T)P^*_n(T)\\ &= \sum_{T\in\mathbf{Tree}_n}\sum_{i=1}^{4} \big|\{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\}\big| \cdot q_i \cdot P^*_n(T)\\ &= \binom{n}{4}\sum_{i=1}^{4}\sum_{T\in\mathbf{Tree}_n} \frac{\big|\{Q \in \mathrm{Part}_4(L(T)) : T(Q) = Q_i\}\big|}{\binom{n}{4}} P^*_n(T)q_i\\ &= \binom{n}{4}\sum_{i=1}^{4} P^*_4(Q_i)q_i, \end{aligned}$$

by the sampling consistency of $P^*_n$. □

Notice that, intuitively, since $P^*_n$ is sampling consistent, the expected number of quartets of shape $Q_i$ in a tree $T$ is $\binom{n}{4}P^*_4(Q_i)$, and all of them contribute $q_i$ to the $\underset{\sim}{\mathrm{QI}}$ value of the tree. Therefore, the expression found in the theorem above is not surprising.

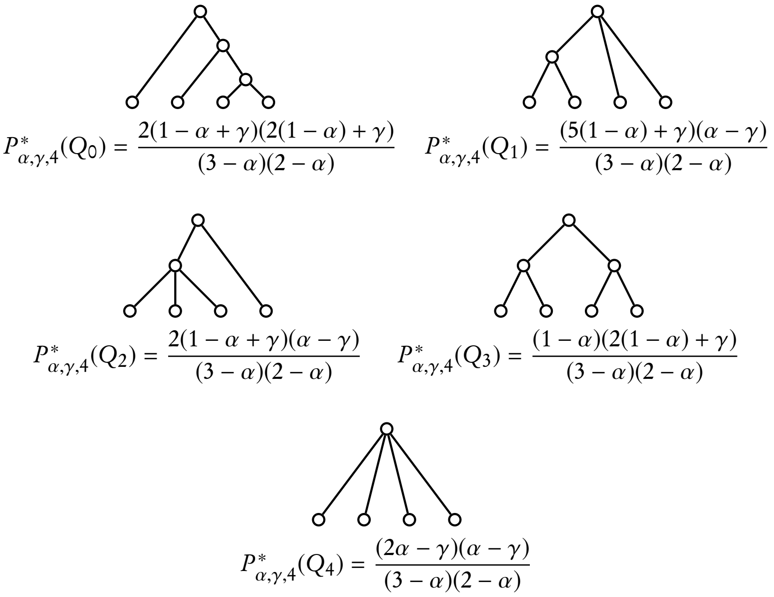
If $P_n$ is a probabilistic model of bifurcating phylogenetic trees, so that $P^*_4(Q_1) = P^*_4(Q_2) = P^*_4(Q_4) = 0$, then the expression in the last theorem becomes

$$E_P(\underset{\sim}{\mathrm{QI}}_n) = \binom{n}{4}P^*_4(Q_3)q_3.$$

Taking $q_3 = 1$, we obtain the following result.

**Corollary 5.19.** *Let $P_n$ : **BinPhyloTree**$_n \to [0,1]$ be a probabilistic model of bifurcating phylogenetic trees such that $P_n^*$ is sampling consistent. Then*

$$E_P(\utilde{\mathrm{QIB}}_n) = E_{P^*}(\utilde{\mathrm{QIB}}^*_n) = \binom{n}{4} P_4^*(Q_3).$$



$$P^*_{\alpha,\gamma,4}(Q_0) = \frac{2(1-\alpha+\gamma)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)} \qquad P^*_{\alpha,\gamma,4}(Q_1) = \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$

$$P^*_{\alpha,\gamma,4}(Q_2) = \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \qquad P^*_{\alpha,\gamma,4}(Q_3) = \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)}$$

$$P^*_{\alpha,\gamma,4}(Q_4) = \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)}$$

Figure 5.5: The five tree shapes in **Tree**$_4$ and their probabilities under the $\alpha$-$\gamma$-model.

Recall that the $\alpha$-$\gamma$-model is sampling consistent for tree shapes, and hence lies under the hypothesis of Theorem 5.18. Using the complete knowledge description of $P^*_{\alpha,\gamma,4}$ on **Tree**$_4$ obtained in Lemma 1.26 and recalled in Figure 5.5, we get the following result.

**Corollary 5.20.** *Let $(P_{\alpha,\gamma,n})_n$ be Chen-Winkel-Ford's $\alpha$-$\gamma$-model of phylogenetic trees, with $0 \leq \gamma \leq \alpha \leq 1$. Then,*

$$E_{P_{\alpha,\gamma}}(\utilde{\mathrm{QI}}_n) = \left( \frac{(2\alpha-\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_4 + \frac{(1-\alpha)(2(1-\alpha)+\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_3 \right.$$
$$\left. + \frac{2(1-\alpha+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_2 + \frac{(5(1-\alpha)+\gamma)(\alpha-\gamma)}{(3-\alpha)(2-\alpha)} \cdot q_1 \right) \binom{n}{4}.$$

For Ford's $\alpha$-model, which corresponds to $\alpha = \gamma$, we get the following result on the expected value of $\utilde{\mathrm{QIB}}_n$ from Corollary 5.19 and the value of $P_{\alpha,4}(Q_3)$ given in the Preliminaries.

**Corollary 5.21.** *Let $(P_{\alpha,n})_n$ be Ford's $\alpha$-model for bifurcating phylogenetic trees, with $\alpha \in [0,1]$. Then,*

$$E_{P_\alpha}(\utilde{\mathrm{QIB}}_n) = \frac{1-\alpha}{3-\alpha}\binom{n}{4}.$$

It is straightforward to check that $E_{P_\alpha}(\underset{\sim}{\mathrm{QIB}}_n)$ agrees with $E_{P_{\alpha,\gamma}}(\underset{\sim}{\mathrm{QI}}_n)$ (up to the factor $q_3$) when $\alpha = \gamma$.

Thus, we can easily compute the expected value of $\underset{\sim}{\mathrm{QIB}}_n$ under the Yule model ($\alpha = 0$) and the Uniform model ($\alpha = \frac{1}{2}$):

$$E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}_n) = \frac{1}{3}\binom{n}{4}, \tag{5.1}$$

$$E_{\mathrm{unif}}(\underset{\sim}{\mathrm{QIB}}_n) = \frac{1}{5}\binom{n}{4}. \tag{5.2}$$

Aldous' $\beta$-model is also sampling consistent for tree shapes, and therefore, from Corollary 5.19 and the value of $P^A_{\beta,4}(Q_3)$ given in the Preliminaries, we obtain the following result.

**Corollary 5.22.** *Let $(P^A_{\beta,n})_n$ be Aldous' $\beta$-model for bifurcating phylogenetic trees, with $\beta \in (-2, \infty)$. Then:*

$$E_{P^A_\beta}(\underset{\sim}{\mathrm{QIB}}_n) = \frac{3\beta + 6}{7\beta + 18}\binom{n}{4}.$$

For this model, the Yule model corresponds to $\beta = 0$ and the Uniform model to $\beta = -3/2$. It is straightforward to check the expected value of $\underset{\sim}{\mathrm{QIB}}_n$ for the Yule and the Uniform model obtained from the last corollary agree with those given in Equations (5.1) and (5.2).

Now, we are in a position to attempt the computation of the variance of $\underset{\sim}{\mathrm{QI}}$ under the same hypothesis as in Theorem 5.18. To try to ease the notations, let us consider, for every $k \in \{5, \dots, 8\}$ and for every $T \in \mathbf{Tree}_k$ and every $(i, j) \in \{1, \dots, 4\}^2$,

$$\begin{aligned}\Theta_{i,j}(T) &= \left|\{(Q, Q') \in \mathrm{Part}_4(L(T))^2 : Q \cup Q' = L(T), T(Q) = Q_i, T(Q') = Q_j\}\right| \\ &= \left|\{(Q, Q') \in \mathrm{Part}_4(L(T))^2 : |Q \cap Q'| = 8 - k, T(Q) = Q_i, T(Q') = Q_j\}\right|.\end{aligned}$$

**Theorem 5.23.** *Let $P_n : \mathbf{PhyloTree}_n \to [0, 1]$ be a probabilistic model for phylogenetic trees such that $P^*_n$ is sampling consistent. Then,*

$$\begin{aligned}\sigma^2_P(\underset{\sim}{\mathrm{QI}}_n) = \sigma^2_P(\underset{\sim}{\mathrm{QI}}^*_n) = \binom{n}{4}\sum_{i=1}^{4} q_i^2 P^*_4(Q_i) - \binom{n}{4}^2\left(\sum_{i=1}^{4} q_i P^*_4(Q_i)\right)^2 \\ + \sum_{i=1}^{4}\sum_{j=1}^{4} q_i q_j \left(\sum_{k=5}^{8}\binom{n}{k}\sum_{T \in \mathbf{Tree}_k}\Theta_{i,j}(T) P^*_k(T)\right).\end{aligned}$$

*Proof.* As a consequence of Lemma 5.17, $\sigma^2_P(\underset{\sim}{\mathrm{QI}}_n) = \sigma^2_{P^*}(\underset{\sim}{\mathrm{QI}}^*_n)$. We shall compute the latter by using the identity $\sigma^2_{P^*}(\underset{\sim}{\mathrm{QI}}^*_n) = E_{P^*}((\underset{\sim}{\mathrm{QI}}^*)^2_n) - E_{P^*}(\underset{\sim}{\mathrm{QI}}^*_n)^2$; hence, we need to compute $E_{P^*}((\underset{\sim}{\mathrm{QI}}^*)^2_n)$.

For every $T \in \mathbf{Tree}_n$, every $Q_i \in \mathbf{Tree}_4$ and every $Q \in \mathrm{Part}_4(L(T))$, set

$$\delta(Q, Q_i, T) = \begin{cases} 1 & \text{if } T(Q) = Q_i \\ 0 & \text{otherwise.}\end{cases}$$

Then,

$$\begin{aligned}
E_{P^*}((\underset{\smile}{\mathrm{QI}}^*)^2_n) &= \sum_{T\in\mathbf{Tree}_n} \underset{\smile}{\mathrm{QI}}(T)^2 P^*_n(T)\\
&= \sum_{T\in\mathbf{Tree}_n}\left(\sum_{Q\in\mathrm{Part}_4(L(T))}\sum_{i=1}^{4} q_i\delta(Q,Q_i,T)\right)^2 P^*_n(T)\\
&= \sum_{T\in\mathbf{Tree}_n}\left(\sum_{Q\in\mathrm{Part}_4(L(T))}\sum_{i=1}^{4} q_i^2\delta(Q,Q_i,T)^2\right)P^*_n(T)\\
&\quad+\sum_{T\in\mathbf{Tree}_n}\left(\sum_{\substack{(Q,Q')\in\mathrm{Part}_4(L(T))^2\\ Q\neq Q'}}\ \sum_{(i,j)\in[4]^2} q_iq_j\delta(Q,Q_i,T)\delta(Q,Q_j,T)\right)P^*_n(T)
\end{aligned}$$

but, since $\delta(Q,Q_i,T)\in\{0,1\}$ by definition, $\delta(Q,Q_i,T)^2=\delta(Q,Q_i,T)$. We define

$$\begin{aligned}
S_1 &= \sum_{T\in\mathbf{Tree}_n}\left(\sum_{Q\in\mathrm{Part}_4(L(T))}\sum_{i=1}^{4} q_i^2\delta(Q,Q_i,T)\right)P^*_n(T)\\
&= \sum_{i=1}^{4}\left(q_i^2\sum_{T\in\mathbf{Tree}_n}\big|\{Q\in\mathrm{Part}_4(L(T)) : T(Q)=Q_i\}\big|P^*_n(T)\right)\\
&= \binom{n}{4}\sum_{i=1}^{4}\left(q_i^2\sum_{T\in\mathbf{Tree}_n}\frac{\big|\{Q\in\mathrm{Part}_4(L(T)) : T(Q)=Q_i\}\big|}{\binom{n}{4}}P^*_n(T)\right)\\
&= \binom{n}{4}\sum_{i=1}^{4} q_i^2P^*_4(Q_i),
\end{aligned}$$

this last equality being a consequence of the sampling consistency of $P^*_n$.

As far as the second addend in the expression of $E_{P^*}((\underset{\smile}{\mathrm{QI}}^*)^2_n)$ goes, we have

$$\begin{aligned}
S_2 &= \sum_{T\in\mathbf{Tree}_n}\ \sum_{\substack{(Q,Q')\in\mathrm{Part}_4(L(T))^2\\ Q\neq Q'}}\left(\sum_{(i,j)\in[4]^2} q_iq_j\delta(Q,Qi,T)\delta(Q',Q_j,T)\right)P^*_n(T)\\
&= \sum_{(i,j)\in[4]^2} q_iq_j\left(\sum_{T\in\mathbf{Tree}_n}\left(\sum_{k=0}^{3}\ \sum_{\substack{(Q,Q')\in\mathrm{Part}_4(L(T))^2\\ |Q\cap Q'|=k}}\delta(Q,Q_i,T)\delta(Q',Q_j,T)\right)P^*_n(T)\right)\\
&= \sum_{(i,j)\in[4]^2}\left(\sum_{k=0}^{3}\sum_{T\in\mathbf{Tree}_n}\big|\{(Q,Q')\in\mathrm{Part}_4(L(T))^2 : |Q\cap Q'|=k, T(Q)=Q_i, T(Q')=Q_j\}\big|P^*_n(T)\right)
\end{aligned}$$

Now, notice that for every $k\in\{0,1,2,3\}$, if $|Q\cap Q'|=k$, then $T(Q\cup Q')$ has $8-k$

leaves, and therefore we can simplify the last sum by considering trees $T_{8-k} \in \mathbf{Tree}_{8-k}$:

$$\begin{aligned}
&\sum_{T\in\mathbf{Tree}_n} \left|\{(Q,Q') \in \mathrm{Part}_4(L(T))^2 : |Q\cap Q'| = k, T(Q) = Q_i, T(Q') = Q_j\}\right| P_n^*(T)\\
&= \sum_{T\in\mathbf{Tree}_n} \Bigg( \sum_{T_{8-k}\in\mathbf{Tree}_{8-k}} \left|\{X \in \mathrm{Part}_{8-k}(L(T)) : T(X) = T_{8-k}\}\right|\\
&\quad \cdot \left|\{(Q,Q') \in \mathrm{Part}_4(L(T_{8-k}))^2 : |Q\cap Q'| = k, T_{8-k}(Q) = Q_i, T_{8-k}(Q') = Q_j\}\right|\Bigg) P_n^*(T)\\
&= \sum_{T_{8-k}\in\mathbf{Tree}_{8-k}} \left|\{(Q,Q') \in \mathrm{Part}_4(L(T_{8-k}))^2 : |Q\cap Q'| = k, T_{8-k}(Q) = Q_i, T_{8-k}(Q') = Q_j\}\right|\\
&\quad \cdot \binom{n}{8-k} \sum_{T\in\mathbf{Tree}_n} \frac{\left|\{X \in \mathrm{Part}_{8-k}(L(T)) : T(X) = T_{8-k}\}\right|}{\binom{n}{8-k}} P_n^*(T)\\
&= \binom{n}{8-k} \sum_{T_{8-k}\in\mathbf{Tree}_{8-k}} \left|\{(Q,Q') \in \mathrm{Part}_4(L(T_{8-k}))^2 : |Q\cap Q'| = k, T_{8-k}(Q) = Q_i, T_{8-k}(Q') = Q_j\}\right|\\
&\quad \cdot P_{8-k}^*(T_{8-k})
\end{aligned}$$

(by the sampling consistency of $(P_n^*)_n$)

$$\begin{aligned}
&= \binom{n}{8-k} \sum_{T_{8-k}\in\mathbf{Tree}_{8-k}} \left|\{(Q,Q') \in \mathrm{Part}_{8-k}(L(T))^2 : Q\cup Q' = L(T_{8-k}), T_{8-k}(Q) = Q_i, T_{8-k}(Q') = Q_j\}\right|\\
&\quad \cdot P_{8-k}^*(T_{8-k})\\
&= \binom{n}{8-k} \sum_{T_{8-k}\in\mathbf{Tree}_{8-k}} \Theta_{i,j}(T_{8-k}) P_{8-k}^*(T_{8-k}).
\end{aligned}$$

Thus,

$$\begin{aligned}
S_2 &= \sum_{(i,j)\in[4]^2} q_i q_j \left( \sum_{k=0}^{3} \binom{n}{8-k} \sum_{T_{8-k}\in\mathbf{Tree}_{8-k}} \Theta_{i,j}(T_{8-k}) P_{8-k}^*(T_{8-k}) \right)\\
&= \sum_{(i,j)\in[4]^2} q_i q_j \left( \sum_{k=5}^{8} \binom{n}{k} \sum_{T\in\mathbf{Tree}_k} \Theta_{i,j}(T) P_k^*(T) \right).
\end{aligned}$$

Therefore, as $\sigma^2_{P^*}(\underline{\mathrm{QI}}^*_n) = S_1 + S_2 - E_{P^*}(\underline{\mathrm{QI}}^*_n)^2$, the proposition is proven. □

If $(P_n)_n$ is a probabilistic model for bifurcating phylogenetic trees, then the only possible quartets are $Q_0$ and $Q_3$, and therefore the formula in the previous theorem becomes (after taking $q_3 = 1$)

$$\sigma_P^2(\underline{\mathrm{QIB}}_n) = \binom{n}{4} P_4^*(Q_3) - \binom{n}{4}^2 P_4^*(Q_3)^2 + \sum_{k=5}^{8} \sum_{T\in\mathbf{BinTree}_k} \Theta_{3,3}(T) P_k^*(T),$$

where the coefficients $\Theta_{3,3}(T)$ can be easily computed for all $T \in \mathbf{BinTree}_k$, for $k \in \{5,6,7,8\}$.These coefficients are provided in the Table 5.1 and they yield the following result.

**Corollary 5.24.** *If $(P_n)_n$ is a probabilistic model of bifurcating phylogenetic trees such that $(P_n^*)_n$ is sampling consistent, then, with the notations given in the Table 5.1,*

$$\begin{aligned}\sigma_P^2(\underset{\sim}{\mathrm{QIB}}_n) = {} & \binom{n}{4}P_4^*(Q_3) - \binom{n}{4}^2 P_4^*(Q_3)^2 \\ & + 6\binom{n}{5}P_5^*(B_{5,3}) + \binom{n}{6}\big(18P_6^*(B_{6,4}) + 6P_6^*(B_{6,5}) + 36P_6^*(B_{6,6})\big) \\ & + \binom{n}{7}\big(8P_7^*(B_{7,8}) + 24P_7^*(B_{7,9}) + 36P_7^*(B_{7,10}) + 36P_7^*(B_{7,11})\big) \\ & + \binom{n}{8}\big(2P_8^*(B_{8,13}) + 6P_8^*(B_{8,14}) + 12P_8^*(B_{8,15}) + 14P_8^*(B_{8,16}) \\ & + 18P_8^*(B_{8,17}) + 36P_8^*(B_{8,21}) + 36P_8^*(B_{8,22}) + 38P_8^*(B_{8,23})\big).\end{aligned}$$

Theorem 5.23 and Corollary 5.24 reduce the computation of the variance of $\underset{\sim}{\mathrm{QI}}_n$ and $\underset{\sim}{\mathrm{QIB}}_n$ under sampling consistent probabilistic models to the knowledge of a finite, fixed number of probabilities $P_k^*$ for $k \in \{4,5,6,7,8\}$. Therefore, we are able to find closed formulæ for the variance of $\underset{\sim}{\mathrm{QI}}_n$ under the $\alpha$-$\gamma$-model and for the variance of $\underset{\sim}{\mathrm{QIB}}_n$ under Ford's $\alpha$-model and Aldous' $\beta$-model, and hence under the Yule and Uniform models as well. We begin with the bifurcating case.

Concerning Ford's $\alpha$-model and Aldous' $\beta$-model, computing all the necessary probabilities, which are given in the Tables 5.2 and 5.3, we obtain the following results.

**Corollary 5.25.** *Under the $\alpha$-model,*

$$\begin{aligned}\sigma_{P_\alpha}^2(\underset{\sim}{\mathrm{QIB}}_n) = {} & \binom{n}{4}\frac{1-\alpha}{3-\alpha} - \binom{n}{4}^2\frac{(1-\alpha)^2}{(3-\alpha)^2} + 12\binom{n}{5}\frac{1-\alpha}{4-\alpha} \\ & + \binom{n}{6}\frac{6(1-\alpha)(112-89\alpha+15\alpha^2)}{(5-\alpha)(4-\alpha)(3-\alpha)} + \binom{n}{7}\frac{20(1-\alpha)(74-63\alpha+7\alpha^2)}{(6-\alpha)(5-\alpha)(3-\alpha)} \\ & + \binom{n}{8}\frac{10(1-\alpha)(506-539\alpha+112\alpha^2-7\alpha^3)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}.\end{aligned}$$

**Corollary 5.26.** *Under the $\beta$-model,*

$$\begin{aligned}\sigma_{P_\beta}^2(\underset{\sim}{\mathrm{QIB}}_n) = {} & \binom{n}{4}\frac{3(\beta+2)}{7\beta+18} - \binom{n}{4}^2\frac{9(\beta+2)^2}{(7\beta+18)^2} + 12\binom{n}{5}\frac{\beta+2}{3\beta+8} \\ & + 90\binom{n}{6}\frac{(\beta+2)(41\beta^2+238\beta+336)}{(31\beta^2+194\beta+300)(7\beta+18)} + 60\binom{n}{7}\frac{(\beta+2)(9\beta^2+53\beta+74)}{(\beta+3)(3\beta+10)(7\beta+18)} \\ & + 630\binom{n}{8}\frac{(\beta+2)(127\beta^4+1637\beta^3+7788\beta^2+16084\beta+12144)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}.\end{aligned}$$

When $\alpha = 0$, or $\beta = 0$, we are presented with the Yule model. In this case, from the previous corollaries we get

$$\sigma_{\mathrm{Yule}}^2(\underset{\sim}{\mathrm{QIB}}_n) = \binom{n}{4}\frac{5n^4+30n^3+118n^2+408n+630}{33075}. \tag{5.3}$$

| Name | Shape | $\Theta_{3,3}(T)$ |
|---|---|---|
| $B_{5,1}$ | $(\cdot,(\cdot,(\cdot,(\cdot,\cdot))))$ | 0 |
| $B_{5,2}$ | $(\cdot,((\cdot,\cdot),(\cdot,\cdot)))$ | 0 |
| $B_{5,3}$ | $((\cdot,\cdot),(\cdot,(\cdot,\cdot)))$ | 6 |
| $B_{6,1}$ | $(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{6,2}$ | $(\cdot,(\cdot,((\cdot,\cdot),(\cdot,\cdot))))$ | 0 |
| $B_{6,3}$ | $(\cdot,((\cdot,\cdot),(\cdot,(\cdot,\cdot))))$ | 0 |
| $B_{6,4}$ | $((\cdot,\cdot),((\cdot,\cdot),(\cdot,\cdot)))$ | 18 |
| $B_{6,5}$ | $((\cdot,\cdot),(\cdot,(\cdot,(\cdot,\cdot))))$ | 6 |
| $B_{6,6}$ | $((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,\cdot)))$ | 36 |
| $B_{7,1}$ | $(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,\cdot))))))$ | 0 |
| $B_{7,2}$ | $(\cdot,(\cdot,(\cdot,((\cdot,\cdot),(\cdot,\cdot)))))$ | 0 |
| $B_{7,3}$ | $(\cdot,(\cdot,((\cdot,\cdot),(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{7,4}$ | $(\cdot,((\cdot,\cdot),((\cdot,\cdot),(\cdot,\cdot))))$ | 0 |
| $B_{7,5}$ | $(\cdot,((\cdot,\cdot),(\cdot,(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{7,6}$ | $(\cdot,((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,\cdot))))$ | 0 |
| $B_{7,7}$ | $((\cdot,\cdot),(\cdot,(\cdot,(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{7,8}$ | $((\cdot,\cdot),(\cdot,((\cdot,\cdot),(\cdot,\cdot))))$ | 8 |
| $B_{7,9}$ | $((\cdot,\cdot),((\cdot,\cdot),(\cdot,(\cdot,\cdot))))$ | 24 |
| $B_{7,10}$ | $((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,(\cdot,\cdot))))$ | 36 |
| $B_{7,11}$ | $((\cdot,(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot)))$ | 36 |
| $B_{8,1}$ | $(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,\cdot)))))))$ | 0 |
| $B_{8,2}$ | $(\cdot,(\cdot,(\cdot,(\cdot,((\cdot,\cdot),(\cdot,\cdot))))))$ | 0 |
| $B_{8,3}$ | $(\cdot,(\cdot,(\cdot,((\cdot,\cdot),(\cdot,(\cdot,\cdot))))))$ | 0 |
| $B_{8,4}$ | $(\cdot,(\cdot,((\cdot,\cdot),((\cdot,\cdot),(\cdot,\cdot)))))$ | 0 |
| $B_{8,5}$ | $(\cdot,(\cdot,((\cdot,\cdot),(\cdot,(\cdot,(\cdot,\cdot))))))$ | 0 |
| $B_{8,6}$ | $(\cdot,(\cdot,((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{8,7}$ | $(\cdot,((\cdot,\cdot),(\cdot,(\cdot,(\cdot,(\cdot,\cdot))))))$ | 0 |
| $B_{8,8}$ | $(\cdot,((\cdot,\cdot),(\cdot,\cdot),(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{8,9}$ | $(\cdot,((\cdot,\cdot),(\cdot,((\cdot,\cdot),(\cdot,\cdot)))))$ | 0 |
| $B_{8,10}$ | $(\cdot,((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{8,11}$ | $(\cdot,((\cdot,(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot))))$ | 0 |
| $B_{8,12}$ | $((\cdot,\cdot),(\cdot,(\cdot,(\cdot,(\cdot,(\cdot,\cdot))))))$ | 0 |
| $B_{8,13}$ | $((\cdot,\cdot),(\cdot,(\cdot,(\cdot,\cdot),(\cdot,\cdot))))$ | 2 |
| $B_{8,14}$ | $((\cdot,\cdot),(\cdot,((\cdot,\cdot),(\cdot,(\cdot,\cdot)))))$ | 6 |
| $B_{8,15}$ | $((\cdot,\cdot),((\cdot,\cdot),(\cdot,(\cdot,(\cdot,\cdot))))))$ | 12 |
| $B_{8,16}$ | $((\cdot,\cdot),((\cdot,\cdot),((\cdot,\cdot),(\cdot,\cdot))))$ | 14 |
| $B_{8,17}$ | $((\cdot,\cdot),((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,\cdot))))$ | 18 |
| $B_{8,18}$ | $((\cdot,(\cdot,\cdot)),(\cdot,(\cdot,(\cdot,(\cdot,\cdot)))))$ | 0 |
| $B_{8,19}$ | $((\cdot,(\cdot,\cdot)),(\cdot,((\cdot,\cdot),(\cdot,\cdot)))))$ | 0 |
| $B_{8,20}$ | $((\cdot,(\cdot,\cdot)),((\cdot,\cdot),(\cdot,(\cdot,\cdot))))$ | 0 |
| $B_{8,21}$ | $((\cdot,(\cdot,(\cdot,\cdot))),(\cdot,(\cdot,(\cdot,\cdot))))$ | 36 |
| $B_{8,22}$ | $((\cdot,(\cdot,(\cdot,\cdot))),((\cdot,\cdot),(\cdot,\cdot)))$ | 36 |
| $B_{8,23}$ | $(((\cdot,\cdot),(\cdot,\cdot)),((\cdot,\cdot),(\cdot,\cdot)))$ | 38 |

Table 5.1: Coefficients of the probabilities of the trees in $\mathbf{BinTree}_k$, for $k \in \{5,6,7,8\}$, in the formula for the variance of $\underset{\sim}{\mathrm{QIB}}_n$.

| Tree | $P_{\alpha,n}^{A,*}$ |
|---|---|
| $Q_3$ | $\frac{1-\alpha}{3-\alpha}$ |
| $B_{5,3}$ | $\frac{2(1-\alpha)}{4-\alpha}$ |
| $B_{6,4}$ | $\frac{(1-\alpha)^2(8-\alpha)}{(5-\alpha)(4-\alpha)(3-\alpha)}$ |
| $B_{6,5}$ | $\frac{2(1-\alpha)(8-\alpha)}{(5-\alpha)(4-\alpha)(3-\alpha)}$ |
| $B_{6,6}$ | $\frac{2(1-\alpha)(2-\alpha)}{(5-\alpha)(4-\alpha)}$ |
| $B_{7,8}$ | $\frac{(1-\alpha)^2(2+\alpha)(10+\alpha)}{(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$ |
| $B_{7,9}$ | $\frac{2(1-\alpha)^2(10+\alpha)}{(6-\alpha)(5-\alpha)(4-\alpha)}$ |
| $B_{7,10}$ | $\frac{10(1-\alpha)(2-\alpha)}{(6-\alpha)(5-\alpha)(3-\alpha)}$ |
| $B_{7,11}$ | $\frac{5(1-\alpha)^2(2-\alpha)}{(6-\alpha)(5-\alpha)(3-\alpha)}$ |
| $B_{8,13}$ | $\frac{8(1-\alpha)^2(1+\alpha)(2+\alpha)(3+\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$ |
| $B_{8,14}$ | $\frac{16(1-\alpha)^2(1+\alpha)(3+\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)}$ |
| $B_{8,15}$ | $\frac{8(1-\alpha)^2(3+\alpha)(8-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$ |
| $B_{8,16}$ | $\frac{4(1-\alpha)^3(3+\alpha)(8-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)(3-\alpha)}$ |
| $B_{8,17}$ | $\frac{8(1-\alpha)^2(2-\alpha)(3+\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(4-\alpha)}$ |
| $B_{8,21}$ | $\frac{20(1-\alpha)(2-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$ |
| $B_{8,22}$ | $\frac{20(1-\alpha)^2(2-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$ |
| $B_{8,23}$ | $\frac{5(1-\alpha)^3(2-\alpha)}{(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}$ |

Table 5.2: Probabilities under the $\alpha$-model of the trees involved in the formula for the variance of $\utilde{\mathrm{QIB}}_n$.

As for the Uniform model, setting $\alpha = \frac{1}{2}$ or $\beta = -\frac{3}{2}$, the formulæ in the last corollaries yield

$$\sigma^2_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_n) = \binom{n}{4}\frac{4(2n-1)(2n+1)(2n+3)(2n+5)}{225225} \tag{5.4}$$

In order to double-check our computations, we have produced independent derivations of these two last formulæ, as well as of (5.1) and (5.2), using the approach used in Section 3.2 to compute the expected values and variances of $C^{(2)}$. We give these alternative proofs in the Subsection 5.3.1.

Notice that the leading term of the variance under the $\alpha$-model is

$$\frac{(1-\alpha)(2\alpha+1)}{84(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)^2}\cdot n^8,$$

| Tree | $P^{B,*}_{\beta,n}$ |
|---|---|
| $Q_3$ | $\frac{3(\beta+2)}{7\beta+18}$ |
| $B_{5,3}$ | $\frac{2(\beta+2)}{3\beta+8}$ |
| $B_{6,4}$ | $\frac{45(\beta+2)^2(\beta+4)}{(31\beta^2+194\beta+300)(7\beta+18)}$ |
| $B_{6,5}$ | $\frac{60(\beta+2)(\beta+3)(\beta+4)}{(31\beta^2+194\beta+300)(7\beta+18)}$ |
| $B_{6,6}$ | $\frac{10(\beta+2)(\beta+3)}{31\beta^2+194\beta+300}$ |
| $B_{7,8}$ | $\frac{3(\beta+2)^2(\beta+4)(\beta+5)}{(\beta+3)(3\beta+8)(3\beta+10)(7\beta+18)}$ |
| $B_{7,9}$ | $\frac{2(\beta+2)^2(\beta+5)}{(\beta+3)(3\beta+8)(3\beta+10)}$ |
| $B_{7,10}$ | $\frac{20(\beta+2)(\beta+3)}{3(3\beta+10)(7\beta+18)}$ |
| $B_{7,11}$ | $\frac{5(\beta+2)^2}{(3\beta+10)(7\beta+18)}$ |
| $B_{8,13}$ | $\frac{504(\beta+2)^2(\beta+4)^2(\beta+5)^2(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(3\beta+8)(7\beta+18)}$ |
| $B_{8,14}$ | $\frac{336(\beta+2)^2(\beta+4)(\beta+5)^2(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(3\beta+8)}$ |
| $B_{8,15}$ | $\frac{1680(\beta+2)^2(\beta+3)(\beta+4)(\beta+5)(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(7\beta+18)}$ |
| $B_{8,16}$ | $\frac{1260(\beta+2)^3(\beta+4)(\beta+5)(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)(7\beta+18)}$ |
| $B_{8,17}$ | $\frac{280(\beta+2)^2(\beta+3)(\beta+5)(\beta+6)}{(127\beta^3+1383\beta^2+4958\beta+5880)(31\beta^2+194\beta+300)}$ |
| $B_{8,21}$ | $\frac{560(\beta+2)(\beta+3)^3(\beta+4)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}$ |
| $B_{8,22}$ | $\frac{840(\beta+2)^2(\beta+3)^2(\beta+4)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}$ |
| $B_{8,23}$ | $\frac{315(\beta+2)^3(\beta+3)(\beta+4)}{(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}$ |

Table 5.3: Probabilities under the $\beta$-model of the trees involved in the formula for the variance of $\underset{\sim}{\mathrm{QIB}}_n$.

and for the variance under the $\beta$-model is

$$\frac{(\beta+2)(2\beta^2+9\beta+12)}{2(127\beta^3+1383\beta^2+4958\beta+5880)(7\beta+18)^2}\cdot n^8.$$

So, the variance of $\underset{\sim}{\mathrm{QIB}}_n$ grows under both models in $O(n^8)$

Finally, as far as the $\alpha$-$\gamma$-model goes, we have written a set of Python scripts that compute all $\Theta_{i,j}(T)$, $(i,j)\in\{1,2,3,4\}^2$, as well as $P^*_{\alpha,\gamma,k}(T)$ for every $T\in\mathbf{Tree}_k$, $k=5,6,7,8$, and combine all these data into an explicit formula for $\sigma^2_{P_{\alpha,\gamma}}(\underset{\sim}{\mathrm{QI}}_n)$. These Python scripts can be found in the GitHub page `https://github.com/biocom-uib/biotrees`[2]. It can be easily checked (using a symbolic computation program) that

[2] There is also an implementation in the GitHub page `https://github.com/biocom-uib/Quartet_Index`, but it is not due to me but to G. Valiente [25]. In particular, the plain text formula (which is too long and uninformative to be reproduced here) is given in the document `variance_table.txt` therein.

when $\alpha = \gamma$ it agrees with the variance under the $\alpha$-model given in Corollary 5.25.

### 5.3.1 Alternative proofs for the Yule and the Uniform models

As we promised, in this section we prove Equations (5.1) to (5.4) through the direct approach provided by Lemma 1.31.

**The Yule case**

By Lemma 1.31 and Corollary 5.4,

$$\begin{aligned} E_{\mathrm{Yule}}(\mathrm{QIB}_n) &= \frac{1}{n-1}\sum_{k=1}^{n-1}\left(2E_{\mathrm{Yule}}(\mathrm{QIB}_k) + \binom{k}{2}\binom{n-k}{2}\right) \\ &= \frac{2}{n-1}\sum_{k=1}^{n-1} E_{\mathrm{Yule}}(\mathrm{QIB}_k) + \frac{1}{n-1}\binom{n+1}{5}, && (5.5) \\ E_{\mathrm{Yule}}(\mathrm{QIB}_n^2) &= \frac{1}{n-1}\sum_{k=1}^{n-1}\left(2E_{\mathrm{Yule}}(\mathrm{QIB}_k^2) + 4\binom{k}{2}\binom{n-k}{2}E_{\mathrm{Yule}}(\mathrm{QIB}_k)\right. \\ &\qquad \left. +2E_{\mathrm{Yule}}(\mathrm{QIB}_k)E_{\mathrm{Yule}}(\mathrm{QIB}_{n-k}) + \binom{k}{2}^2\binom{n-k}{2}^2\right). && (5.6) \end{aligned}$$

Then, regarding the expected value, by (5.5):

$$\begin{aligned} E_{\mathrm{Yule}}(\mathrm{QIB}_n) &= \frac{1}{n-1}\left(2\sum_{k=1}^{n-1} E_{\mathrm{Yule}}(\mathrm{QIB}_k) + \binom{n+1}{5}\right) \\ &= \frac{1}{n-1}\left(2\sum_{k=1}^{n-2} E_{\mathrm{Yule}}(QIB_k) + \binom{n}{5} + 2E_{\mathrm{Yule}}(QIB_{n-1}) + \binom{n+1}{5} - \binom{n}{5}\right) \\ &= \frac{n-2}{n-1}\cdot\frac{1}{n-2}\left(2\sum_{k=1}^{n-2} E_{\mathrm{Yule}}(QIB_k) + \binom{n}{5}\right) + \frac{2}{n-1}E_{\mathrm{Yule}}(QIB_{n-1}) + \frac{1}{n-1}\binom{n}{4} \\ &= \frac{n-2}{n-1}E_{\mathrm{Yule}}(QIB_{n-1}) + \frac{2}{n-1}E_{\mathrm{Yule}}(QIB_{n-1}) + \frac{1}{n-1}\binom{n}{4} \\ &= \frac{n}{n-1}E_{\mathrm{Yule}}(QIB_{n-1}) + \frac{1}{24}n(n-2)(n-3). \end{aligned}$$

Dividing by $n$ both sides of this expression for $E_{\mathrm{Yule}}(\mathrm{QIB}_n)$ and setting $x_n = E_{\mathrm{Yule}}(\mathrm{QIB}_n)/n$, we obtain the recurrence

$$x_n = x_{n-1} + \frac{1}{12}\binom{n-2}{2}.$$

Since $x_1 = 0$, its solution is

$$x_n = \frac{1}{12}\sum_{k=2}^{n}\binom{k-2}{2} = \frac{1}{12}\sum_{k=0}^{n-2}\binom{k}{2} = \frac{1}{12}\binom{n-1}{3},$$

from where we finally obtain

$$E_{\mathrm{Yule}}(QIB_n) = n \cdot x_n = \frac{1}{12}n\binom{n-1}{3} = \frac{1}{3}\binom{n}{4}$$

in agreement with Equation (5.1).

As to the variance under the Yule model, by Equations (5.1) and (5.6),

$$\begin{aligned}
E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_n) &= \frac{1}{n-1}\sum_{k=1}^{n-1}\biggl(2E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_k) + \frac{4}{3}\binom{k}{2}\binom{n-k}{2}\binom{k}{4} \\
&\quad + \frac{2}{9}\binom{k}{4}\binom{n-k}{4} + \binom{k}{2}^2\binom{n-k}{2}^2\biggr) \\
&= \frac{n-2}{n-1}\cdot\frac{1}{n-2}\sum_{k=1}^{n-2}\biggl(2E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_k) + \frac{4}{3}\binom{k}{2}\binom{n-1-k}{2}\binom{k}{4} \\
&\quad + \frac{2}{9}\binom{k}{4}\binom{n-1-k}{4} + \binom{k}{2}^2\binom{n-1-k}{2}^2\biggr) + \frac{2}{n-1}E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_{n-1}) \\
&\quad + \frac{1}{n-1}\sum_{k=1}^{n-2}\Biggl[\biggl(\frac{4}{3}\binom{k}{2}\binom{n-k}{2}\binom{k}{4} + \frac{2}{9}\binom{k}{4}\binom{n-k}{4} + \binom{k}{2}^2\binom{n-k}{2}^2\biggr) \\
&\quad - \biggl(\frac{4}{3}\binom{k}{2}\binom{n-1-k}{2}\binom{k}{4} + \frac{2}{9}\binom{k}{4}\binom{n-1-k}{4} + \binom{k}{2}^2\binom{n-1-k}{2}^2\biggr)\Biggr] \\
&= \frac{n-2}{n-1}E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_{n-1}) + \frac{2}{n-1}E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_{n-1}) + \frac{4}{3(n-1)}\sum_{k=1}^{n-2}(n-k-1)\binom{k}{2}\binom{k}{4} \\
&\quad + \frac{2}{9(n-1)}\sum_{k=1}^{n-2}\binom{n-k-1}{3}\binom{k}{4} + \frac{1}{n-1}\sum_{k=1}^{n-2}\binom{k}{2}^2(n-k-1)^3 \\
&= \frac{n}{n-1}E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_{n-1}) + \frac{n}{3}\binom{n-2}{4}\frac{15n^2-35n+6}{420} \\
&\quad + \frac{n}{9}\binom{n-2}{4}\frac{n^2-13n+42}{840} + n\binom{n-2}{2}\frac{3n^4-18n^3+41n^2-42n+36}{1680} \\
&= \frac{n}{n-1}E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_{n-1}) + \frac{n(n-2)(n-3)(253n^4-2014n^3+6119n^2-7430n+3504)}{181440}.
\end{aligned}$$

Dividing by $n$ both sides of this expression for $E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_n)$ and setting $y_n = E_{\mathrm{Yule}}(\underset{\sim}{\mathrm{QIB}}^2_n)/n$, we obtain the recurrence

$$y_n = y_{n-1} + \frac{(n-2)(n-3)(253n^4-2014n^3+6119n^2-7430n+3504)}{181440}.$$

Since $y_1 = 0$, its solution is

$$\begin{aligned} y_n &= \sum_{k=2}^{n} \frac{(k-2)(k-3)(253k^4 - 2014k^3 + 6119k^2 - 7430k + 3504)}{181440} \\ &= \frac{(n-3)(n-2)(n-1)(1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040)}{6350400} \end{aligned}$$

from where we obtain

$$\begin{aligned} E_{\mathrm{Yule}}(\underset{\smile}{\mathrm{QIB}}{}_n^2) &= ny_n \\ &= \binom{n}{4} \frac{1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040}{264600}. \end{aligned}$$

Finally

$$\begin{aligned} \sigma^2_{\mathrm{Yule}}(\underset{\smile}{\mathrm{QIB}}{}_n) &= E_{\mathrm{Yule}}(\underset{\smile}{\mathrm{QIB}}{}_n^2) - E_{\mathrm{Yule}}(\underset{\smile}{\mathrm{QIB}}{}_n)^2 \\ &= \binom{n}{4} \frac{1265n^4 - 7110n^3 + 14419n^2 - 4086n + 5040}{264600} - \frac{1}{9}\binom{n}{4}^2 \\ &= \binom{n}{4} \frac{5n^4 + 30n^3 + 118n^2 + 408n + 630}{33075}, \end{aligned}$$

thus proving Equation (5.3).

### The Uniform case

By Lemma 1.31 and Corollary 5.4

$$E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_n) = \sum_{k=1}^{n-1} C_{k,n-k} \left( 2E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_k) + \binom{k}{2}\binom{n-k}{2} \right) \tag{5.7}$$

$$\begin{aligned} E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_n^2) = \sum_{k=1}^{n-1} C_{k,n-k} \Bigg( &2E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_k^2) + 4\binom{k}{2}\binom{n-k}{2} E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_k) \\ &+ 2E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_k) E_{\mathrm{unif}}(\underset{\smile}{\mathrm{QIB}}{}_{n-k}) + \binom{k}{2}^2 \binom{n-k}{2}^2 \Bigg) \end{aligned} \tag{5.8}$$

where

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

Now, as far as the expected value goes, and since

$$\binom{k}{2}\binom{n-k}{2} = 6\binom{k}{4} - 3(n-3)\binom{k}{3} + \binom{n-2}{2}\binom{k}{2}$$

we have, by Lemma 1.33, that

$$\begin{aligned}
&\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\binom{n-k}{2}\\
&\quad = 6\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{4} - 3(n-3)\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{3} + \binom{n-2}{2}\sum_{k=1}^{n-1} C_{k,n-k}\binom{k}{2}\\
&\quad = 3\binom{n}{4}\left(1-\frac{15}{16(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right) - \frac{3}{2}(n-3)\binom{n}{3}\left(1-\frac{3}{4(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\qquad + \frac{1}{2}\binom{n-2}{2}\binom{n}{2}\left(1-\frac{1}{2(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad = \frac{3}{16(n-1)}\binom{n}{4}\frac{(2n-2)!!}{(2n-3)!!} = \frac{1}{64}\left(3\binom{n}{3}-2\binom{n}{2}+n\right)\frac{(2n-2)!!}{(2n-3)!!}.
\end{aligned}$$

Then, by Equation (5.7), $E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_n)$ is the solution of the equation

$$x_n = 2\sum_{k=1}^{n-1} C_{k,n-k}x_k + \frac{1}{64}\left(3\binom{n}{3}-2\binom{n}{2}+n\right)\frac{(2n-2)!!}{(2n-3)!!}$$

with initial condition $x_1 = 0$. By Theorem 1.35, this solution is

$$E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_n) = \frac{1}{5}\binom{n}{4},$$

thus proving Equation (5.2).

Consider finally Equation (5.4). By Equations (5.8) and (5.2),

$$\begin{aligned}
E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_n^2) &= \sum_{k=1}^{n-1} C_{k,n-k}\left(2E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_k^2) + \frac{4}{5}\binom{k}{2}\binom{n-k}{2}\binom{k}{4}\right.\\
&\qquad\left. + \frac{2}{25}\binom{k}{4}\binom{n-k}{4} + \binom{k}{2}^2\binom{n-k}{2}^2\right)\\
&= \sum_{k=1}^{n-1} C_{k,n-k}\left(2E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_k^2) + \binom{n-2}{2}^2\binom{k}{2} + \frac{3}{2}(n-3)^2(n^2-10n+22)\binom{k}{3}\right.\\
&\qquad + \frac{1}{300}(451n^4 - 15322n^3 + 147149n^2 - 552518n + 722640)\binom{k}{4}\\
&\qquad - \frac{1}{15}(n-5)(451n^2 - 7333n + 24072)\binom{k}{5}\\
&\qquad + \frac{3}{5}(461n^2 - 6453n + 21342)\binom{k}{6}\\
&\qquad \left. - \frac{14}{5}(481n - 3487)\binom{k}{7} + \frac{14308}{5}\binom{k}{8}\right)
\end{aligned}$$

$$
\begin{aligned}
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\mathrm{unif}}(\underset{\sim}{\mathrm{QIB}}{}_k^2)\\
&\quad + \frac{1}{2}\binom{n-2}{2}^2\binom{n}{2}\left(1-\frac{1}{2(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad + \frac{3}{4}(n-3)^2(n^2-10n+22)\binom{n}{3}\left(1-\frac{3}{4(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad + \frac{1}{600}(451n^4-15322n^3+147149n^2-552518n+722640)\binom{n}{4}\left(1-\frac{15}{16(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad - \frac{1}{30}(n-5)(451n^2-7333n+24072)\binom{n}{5}\left(1-\frac{35}{32(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad + \frac{3}{10}(461n^2-6453n+21342)\binom{n}{6}\left(1-\frac{315}{256(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad - \frac{7}{5}(481n-3487)\binom{n}{7}\left(1-\frac{693}{512(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\quad + \frac{7154}{5}\binom{n}{8}\left(1-\frac{3003}{2048(n-1)}\cdot\frac{(2n-2)!!}{(2n-3)!!}\right)\\
&\text{(by Lemma 1.33)}\\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\mathrm{unif}}(\underset{\sim}{\mathrm{QIB}}{}_k^2)\\
&\quad + \frac{703n^7-7653n^6+35545n^5-88575n^4+119632n^3-78372n^2+18000n}{5898240}\cdot\frac{(2n-2)!!}{(2n-3)!!}\\
&= 2\sum_{k=1}^{n-1} C_{k,n-k} E_{\mathrm{unif}}(\underset{\sim}{\mathrm{QIB}}{}_k^2) + \left[\frac{4921}{8192}\binom{n}{7}+\frac{7110}{8192}\binom{n}{6}+\frac{3195}{8192}\binom{n}{5}\right.\\
&\quad \left. + \frac{516}{8192}\binom{n}{4}-\frac{3}{8192}\binom{n}{3}+\frac{2}{8192}\binom{n}{2}-\frac{1}{8192}n\right]\frac{(2n-2)!!}{(2n-3)!!}.
\end{aligned}
$$

So, $E_{\mathrm{unif}}(\underset{\sim}{\mathrm{QIB}}{}_n^2)$ is the solution of the equation

$$
\begin{aligned}
x_n = 2\sum_{k=1}^{n-1} C_{k,n-k}x_k &+ \left[\frac{4921}{8192}\binom{n}{7}+\frac{7110}{8192}\binom{n}{6}+\frac{3195}{8192}\binom{n}{5}\right.\\
&\left. + \frac{516}{8192}\binom{n}{4}-\frac{3}{8192}\binom{n}{3}+\frac{2}{8192}\binom{n}{2}-\frac{1}{8192}n\right]\frac{(2n-2)!!}{(2n-3)!!}
\end{aligned}
$$

with initial condition $x_1 = 0$. By Theorem 1.35, this solution is

$$
\begin{aligned}
E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}^2_n) = {} & \frac{2\cdot 2!!}{1!!}\left(\frac{\frac{2}{8192}}{2}-\frac{\frac{1}{8192}}{1}\right)\binom{n}{2}+\frac{3\cdot 4!!}{3!!}\left(-\frac{\frac{3}{8192}}{3}+\frac{\frac{2}{8192}}{2}\right)\binom{n}{3}\\
& +\frac{4\cdot 6!!}{5!!}\left(\frac{\frac{516}{8192}}{4}-\frac{\frac{3}{8192}}{3}\right)\binom{n}{4}+\frac{5\cdot 8!!}{7!!}\left(\frac{\frac{3195}{8192}}{5}+\frac{\frac{516}{8192}}{4}\right)\binom{n}{5}\\
& +\frac{6\cdot 10!!}{9!!}\left(\frac{\frac{7110}{8192}}{6}+\frac{\frac{3195}{8192}}{5}\right)\binom{n}{6}+\frac{7\cdot 12!!}{11!!}\left(\frac{\frac{4921}{8192}}{7}+\frac{\frac{7110}{8192}}{6}\right)\binom{n}{7}+\frac{8\cdot 14!!}{13!!}\cdot\left(\frac{\frac{4921}{8192}}{7}\right)\binom{n}{8}\\
= {} & \frac{1}{5}\binom{n}{4}+\frac{12}{7}\binom{n}{5}+\frac{38}{7}\binom{n}{6}+\frac{236}{33}\binom{n}{7}+\frac{1406}{429}\binom{n}{8}
\end{aligned}
$$

Then, finally

$$
\begin{aligned}
\sigma^2_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_n) &= E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}^2_n) - E_{\mathrm{unif}}(\utilde{\mathrm{QIB}}_n)^2\\
&= \binom{n}{4}\Big(\frac{703n^4-3194n^3+6965n^2-3706n-96}{360360}-\frac{1}{25}\binom{n}{4}\Big)\\
&= \binom{n}{4}\frac{4(16n^4+64n^3+56n^2-16n-15)}{225225}\\
&= \binom{n}{4}\frac{4(2n-1)(2n+1)(2n+3)(2n+5)}{225225}.
\end{aligned}
$$

*Et voilà!* Equation (5.4) pops up.

## 5.4 Extensions

A perquisite presented by the Quartet index is that it admits natural generalizations to other sets of trees, such as multilabelled or taxonomic trees, or even to specific types of rooted phylogenetic networks, by simply counting and weighting suitable types of quartets. In this section we expose the general framework of these generalizations, and then we provide some extra detail for the extension that we propose to multilabelled trees. It will be the first balance index proposed in the literature for this type of tree.

### 5.4.1 General framework

Throughout this subsection, by a *subtree* we mean a subtree induced on some subset of leaves. Let $\widehat{\mathbf{Tree}}$ be a set of trees of some kind —for instance, phylogenetic, phylogenetic with nested taxa, taxonomic, multilabelled...— that is closed under subtrees; for every $n$, let $\widehat{\mathbf{Tree}}_n$ be the subset of trees in $\widehat{\mathbf{Tree}}$ with $n$ leaves. We want to emphasize here that although, in order to simplify the language, in this subsection we shall only speak about trees, one can replace everywhere "tree" by "rooted directed acyclic graph" and all assertions and results remain true provided that, for every $n$, the number of rooted directed acyclic graphs with $n$ leaves of the considered type is (up to isomorphisms) finite.

Let $\ell$ be a fixed number of leaves and set $\mathbf{Q} = \widehat{\mathbf{Tree}}_\ell$. Let $N = |\mathbf{Q}| - 1$, and enumerate the elements of $\mathbf{Q}$ as $Q_0, Q_1, \ldots, Q_N$ in such a way that $|\mathrm{Aut}(Q_i)| \leq |\mathrm{Aut}(Q_{i+1})|$. Let $\widehat{QI} : \mathbf{Q} \to \mathbb{R}_{\geq 0}$ be a function such that $\widehat{QI}(Q_i) > \widehat{QI}(Q_j)$ if $|\mathrm{Aut}(Q_i)| > |\mathrm{Aut}(Q_j)|$, and such that the minimum value of $QI$ on $\mathbf{Q}$ is $0$. We shall denote each $\widehat{QI}(Q_i)$ by $\hat{q}_i$, so that (permuting the indices of the elements of $\mathbf{Q}$ with the same number of automorphisms, if necessary) $0 = \hat{q}_0 \leq \hat{q}_1 \leq \cdots \leq \hat{q}_N$, with these inequalities strict when the number of automorphisms jumps.

Then, we can define a function $\widehat{QI} : \widehat{\mathbf{Tree}} \to \mathbb{R}$ by means of

$$\begin{aligned}
\widehat{QI}(T) &= \sum_{\{Q \in \mathbf{Q},\ Q \text{ is a subtree of } T\}} \widehat{QI}(Q) \\
&= \sum_{i=1}^{N} \big|\{\text{subtrees } Q \text{ of } T \text{ such that } Q = Q_i\}\big| \hat{q}_i \\
&= \sum_{i=1}^{N} \big|\{X \in \mathrm{Part}_\ell(L(T)) : T(X) = Q_i\}\big| \hat{q}_i
\end{aligned}$$

and then we can prove the analogous versions of Theorems 5.18 and 5.23.

Let $(\widehat{P}_n)_{n \geq 1}$, be a probability model on $\widehat{\mathbf{Tree}}$: that is, a family of probability mappings $\widehat{P}_n : \widehat{\mathbf{Tree}}_n \longrightarrow [0,1]$, for $n \geq 1$. We shall say that $(\widehat{P}_n)_n$ is *sampling consistent* when, for every $1 \leq m \leq n$ and for every $T_0 \in \widehat{\mathbf{Tree}}_m$,

$$\widehat{P}_m(T_0) = \sum_{T \in \widehat{\mathbf{Tree}}_n} \frac{|\{X \in \mathrm{Part}_m(L(T)) : T(X) = T_0\}|}{\binom{n}{m}} \widehat{P}_n(T).$$

Let $\widehat{QI}_n$ be the random variable that chooses a tree $T \in \widehat{\mathbf{Tree}}_n$ with probability $\widehat{P}_n(T)$ and computes $QI(T)$.

**Theorem 5.27.** *Let $(\widehat{P}_n)_n$ a sampling consistent probabilistic model on $\widehat{\mathbf{Tree}}$. Then, the expected value of $\widehat{QI}_n$ under this model is*

$$E_{\widehat{P}}(\widehat{QI}_n) = \binom{n}{\ell} \sum_{i=1}^{N} \widehat{P}_\ell(Q_i) \cdot \hat{q}_i.$$

*Proof.* We proceed as in the proof of Theorem 5.18,

$$\begin{aligned}
E_{\widehat{P}}(\widehat{QI}) &= \sum_{T \in \widehat{\mathbf{Tree}}_n} \widehat{QI}(T) \widehat{P}_n(T) \\
&= \sum_{T \in \widehat{\mathbf{Tree}}_n} \sum_{i=1}^{N} \big|\{X \in \mathrm{Part}_\ell(L(T)) : T(X) = Q_i\}\big| \hat{q}_i \widehat{P}_n(T) \\
&= \binom{n}{\ell} \sum_{i=1}^{N} \hat{q}_i \sum_{T \in \widehat{\mathbf{Tree}}_n} \frac{\big|\{X \in \mathrm{Part}_\ell(L(T)) : T(X) = Q_i\}\big|}{\binom{n}{\ell}} \widehat{P}_n(T) \\
&= \binom{n}{\ell} \sum_{i=1}^{N} \hat{q}_i \widehat{P}_\ell(Q_i),
\end{aligned}$$

by the sampling consistency of $\widehat{P}_n$. □

**Theorem 5.28.** *Let $(\widehat{P}_n)_n$ a sampling consistent probabilistic model on $\widehat{\mathbf{Tree}}$. Then, the variance of $\widehat{\mathrm{QI}}_n$ under this model is*

$$\begin{aligned}\sigma^2_{\widehat{P}}(\widehat{\mathrm{QI}}_n) &= \binom{n}{\ell}\sum_{i=1}^{N}\hat{q}_i^2\widehat{P}_\ell(Q_i) - \binom{n}{\ell}^2\left(\sum_{i=1}^{N} q_i\widehat{P}_\ell(Q_i)\right)^2\\ &\quad+\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{q}_i\hat{q}_j\left(\sum_{k=\ell+1}^{2\ell}\binom{n}{k}\sum_{T\in\widehat{\mathbf{Tree}}_k}\Theta_{i,j}(T)P^*_k(T)\right)\end{aligned}$$

*where, for every $k\in\{\ell+1,\dots,2\ell\}$, for every $T\in\widehat{\mathbf{Tree}}_k$ and for every $(i,j)\in\{1,\dots,N\}^2$,*

$$\begin{aligned}\Theta_{i,j}(T) &= \left|\{(X,X')\in \mathrm{Part}_\ell(L(T))^2 : X\cup X' = L(T), T(X)=Q_i, T(X')=Q_j\}\right|\\ &= \left|\{(X,X')\in \mathrm{Part}_\ell(L(T))^2 : |X\cap X'| = 2\ell-k, T(X)=Q_i, T(X')=Q_j\}\right|.\end{aligned}$$

*Proof.* We proceed as in the proof of Theorem 5.23. We shall use the identity $\sigma^2_{\widehat{P}}(\widehat{\mathrm{QI}}_n) = E_{\widehat{P}}(\widehat{\mathrm{QI}}_n^2) - E_{\widehat{P}}(\widehat{\mathrm{QI}}_n)^2$; hence, we need to compute $E_{\widehat{P}}(\widehat{\mathrm{QI}}_n^2)$. For every $T\in\widehat{\mathbf{Tree}}_n$, every $Q_i\in\mathbf{Q}$ and every $X\in\mathrm{Part}_\ell(L(T))$, set

$$\delta(X,Q_i,T) = \begin{cases}1 & \text{if } T(X)=Q_i\\ 0 & \text{otherwise.}\end{cases}$$

Then,

$$\begin{aligned}E_{\widehat{P}_n} &= \sum_{T\in\widehat{\mathbf{Tree}}_n}\widehat{\mathrm{QI}}(T)^2\widehat{P}_n(T)\\ &= \sum_{T\in\widehat{\mathbf{Tree}}_n}\left(\sum_{X\in\mathrm{Part}_\ell(L(T))}\sum_{i=1}^{N}\hat{q}_i\delta(X,Q_i,T)\right)^2\widehat{P}_n(T)\\ &= \sum_{T\in\widehat{\mathbf{Tree}}_n}\left(\sum_{X\in\mathrm{Part}_\ell(L(T))}\sum_{i=1}^{N}\hat{q}_i^2\delta(X,Q_i,T)^2\right)\widehat{P}_n(T)\\ &\quad+\sum_{T\in\widehat{\mathbf{Tree}}_n}\left(\sum_{\substack{(X,X')\in\mathrm{Part}_\ell(L(T))^2\\ X\neq X'}}\sum_{(i,j)\in[q]^2}\hat{q}_i\hat{q}_j\delta(X,Q_i,T)\delta(X,Q_j,T)\right)\widehat{P}_n(T)\end{aligned}$$

Now, let

$$\begin{aligned}
S_1 &= \sum_{T\in\widehat{\mathbf{Tree}}_n}\left(\sum_{X\in\mathrm{Part}_\ell(L(T))}\sum_{i=1}^{N}\hat{q}_i\delta(X,Q_i,T)^2\right)\widehat{P}_n(T)\\
&= \sum_{T\in\widehat{\mathbf{Tree}}_n}\left(\sum_{X\in\mathrm{Part}_\ell(L(T))}\sum_{i=1}^{N}\hat{q}_i\delta(X,Q_i,T)\right)\widehat{P}_n(T)\\
&= \sum_{i=1}^{N}\left(\hat{q}_i^2\sum_{T\in\widehat{\mathbf{Tree}}_n}\left|\{X\in\mathrm{Part}_\ell(L(T)):T(X)=Q_i\}\right|\widehat{P}_n(T)\right)\\
&= \binom{n}{\ell}\sum_{i=1}^{N}\left(\hat{q}_i^2\sum_{T\in\widehat{\mathbf{Tree}}_n}\frac{\left|\{X\in\mathrm{Part}_\ell(L(T)):T(X)=Q_i\}\right|}{\binom{n}{\ell}}\widehat{P}_n(T)\right)\\
&= \binom{n}{\ell}\sum_{i=1}^{N}\hat{q}_i^2\widehat{P}_\ell(Q_i),
\end{aligned}$$

(by the sampling consistency of $(\widehat{P}_n)_n$) and

$$\begin{aligned}
S_2 = &\sum_{T\in\widehat{\mathbf{Tree}}_n}\ \sum_{\substack{(X,X')\in\mathrm{Part}_\ell(L(T))^2\\ X\neq X'}}\left(\sum_{(i,j)\in[q]^2}\hat{q}_i\hat{q}_j\delta(X,Qi,T)\delta(X',Q_j,T)\right)\widehat{P}_n(T)\\
&= \sum_{(i,j)\in[q]^2}\hat{q}_i\hat{q}_j\left(\sum_{T\in\widehat{\mathbf{Tree}}_n}\left(\sum_{k=0}^{\ell-1}\sum_{\substack{(X,X')\in\mathrm{Part}_\ell(L(T))^2\\ |X\cap X'|=k}}\delta(X,Q_i,T)\delta(X',Q_j,T)\right)\widehat{P}_n\right)\\
&= \sum_{(i,j)\in[q]^2}\left(\sum_{k=0}^{\ell-1}\sum_{T\in\widehat{\mathbf{Tree}}_n}\left|\{(X,X')\in\mathrm{Part}_\ell(L(T))^2:|X\cap X'|=k,T(X)=Q_i,T(X')=Q_j\}\right|\widehat{P}_n(T)\right).
\end{aligned}$$

Now, notice that for every $k\in\{0,1,\ldots,\ell-1\}$, we are only considering leaves in a tree $T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}$ since we demand that $|X\cap X'|=k$. Therefore,

$$\begin{aligned}
&\sum_{T\in\widehat{\mathbf{Tree}}_n}\left|\{(X,X')\in\mathrm{Part}_\ell(L(T))^2:|X\cap X'|=k,T(X)=Q_i,T(X')=Q_j\}\right|\widehat{P}_n(T)\\
&\quad=\sum_{T\in\widehat{\mathbf{Tree}}_n}\Bigg(\sum_{T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}}\left|\{X\in\mathrm{Part}_{2\ell-k}(L(T)):T(X)=T_{2\ell-k}\}\right|\\
&\qquad\cdot\left|\{(X,X')\in\mathrm{Part}_\ell(L(T_{2\ell-k}))^2:|X\cap X'|=k,T_{2\ell-k}(X)=Q_i,T_{2\ell-k}(X')=Q_j\}\right|\Bigg)\widehat{P}_n(T)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}} \left|\{(X,X')\in \mathrm{Part}_\ell(L(T_{2\ell-k}))^2 : |X\cap X'| = k, T_{2\ell-k}(X)=Q_i, T_{2\ell-k}(X')=Q_j\}\right| \\
&\quad \cdot \binom{n}{2\ell-k} \sum_{T\in\widehat{\mathbf{Tree}}_n} \frac{\left|\{(X,X')\in \mathrm{Part}_\ell(L(T_{2\ell-k}))^2 : |X\cap X'| = k, T_{2\ell-k}(X)=Q_i, T_{2\ell-k}(X')=Q_j\}\right|}{\binom{n}{2\ell-k}} \\
&\quad \cdot \widehat{P}_n(T) \\
&= \binom{n}{2\ell-k} \sum_{T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}} \left|\{(X,X')\in \mathrm{Part}_\ell(L(T_{2\ell-k}))^2 : |X\cap X'| = k, T_{2\ell-k}(X)=Q_i,\right. \\
&\quad \left. T_{2\ell-k}(X')=Q_j\}\right|\widehat{P}_{2\ell-k}(T_{2\ell-k})
\end{aligned}
$$

(by the sampling consistency of $(\widehat{P}_n)_n$)

$$
\begin{aligned}
&= \binom{n}{2\ell-k} \sum_{T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}} \left|\{(X,X')\in \mathrm{Part}_{2\ell-k}(L(T))^2 : X\cup X' = L(T_{2\ell-k}), T_{2\ell-k}(X)=Q_i,\right. \\
&\quad \left. T_{2\ell-k}(X')=Q_j\}\right|\widehat{P}_{2\ell-k}(T_{2\ell-k}) \\
&= \binom{n}{2\ell-k} \sum_{T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}} \Theta_{i,j}(T_{2\ell-k})\widehat{P}_{2\ell-k}(T_{2\ell-k}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
S_2 &= \sum_{(i,j)\in[q]^2} q_iq_j\left(\sum_{k=0}^{\ell-1}\binom{n}{2\ell-k}\sum_{T_{2\ell-k}\in\widehat{\mathbf{Tree}}_{2\ell-k}} \Theta_{i,j}(T_{2\ell-k})\widehat{P}_{2\ell-k}(T_{2\ell-k})\right) \\
&= \sum_{(i,j)\in[q]^2} q_iq_j\left(\sum_{k=\ell+1}^{2\ell}\binom{n}{k}\sum_{T\in\widehat{\mathbf{Tree}}_k} \Theta_{i,j}(T_{2\ell-k})P^*_k(T)\right).
\end{aligned}
$$

Therefore, as $\sigma^2_{\widehat{P}}(\widehat{QI_n}) = S_1 + S_2 - E_{\widehat{P}}(\widehat{QI_n})^2$, we have proven the proposition. □

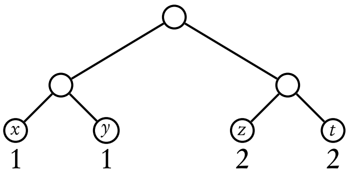
### 5.4.2 A quartet index for multilabelled tree shapes

Remember the discussion in Section 1.1.2 about the elements of **MulTree**. These are pairs $(T,\lambda)$, with $T\in\mathbf{Tree}_n$ and $\lambda : L(T)\to[n]$ for some $n\in\mathbb{N}$. The difference between **MulTree** and **PhyloTree** is that, in the former, we do not demand that $\lambda$ must be bijective. Let us recall the second definition of isomorphism between multilabelled trees given in the aforementioned subsection. Given $(T_1,\lambda_1),(T_2,\lambda_2)\in\mathbf{MulShTree}_n$, a *shape-isomorphism* between them is a pair $(\varphi,\varphi_{[n]})$ such that $\varphi$ is an isomorphism of trees between $T_1$ and $T_2$, and $\varphi_{[n]} : [n]\to[n]$ is bijective and such that the diagram

$$
\begin{tikzcd}
L(T_1) \arrow[r, "\varphi_V|_L"] \arrow[d, "\lambda_1"'] & L(T_2) \arrow[d, "\lambda_2"] \\
{[n]} \arrow[r, "\varphi_{[n]}"] & {[n]}
\end{tikzcd}
$$

commutes. Notice that the restriction of this notion of isomorphism to phylogenetic trees does not yield the usual isomorphism of phylogenetic trees, where $\varphi_{[n]}$ is imposed to be the identity on $[n]$, but rather the isomorphism of their shapes. Indeed, for a shape-isomorphism of multilabelled trees as defined above is an isomorphism $\varphi$ of the underlying tree shapes such that a pair of leaves in $T_1$ has the same label if, and only if, the image of the leaves in $T_2$ have the same label, and this adds no extra restriction on $\varphi$ if $T_1$ and $T_2$ are phylogenetic, because then they contain no pair of leaves with the same label. Recall from Section 1.1.2 that we call *multilabelled tree shapes* the shape-isomorphism classes of multilabelled trees, and we denote by $\mathbf{MulShTree}_n$ the set of multilabelled tree shapes with $n$ leaves and labelled in $[n]$.

**Example:**
Let $T \in \mathbf{MulShTree}_4$ be the tree depicted below.



Its shape-automorphisms are

$$\begin{array}{cccc}
\varphi = \mathrm{id} & \varphi : \begin{array}{l} x \mapsto x \\ y \mapsto y \\ z \mapsto t \\ t \mapsto z \end{array} & \varphi : \begin{array}{l} x \mapsto y \\ y \mapsto x \\ z \mapsto t \\ t \mapsto z \end{array} & \varphi : \begin{array}{l} x \mapsto y \\ y \mapsto x \\ z \mapsto t \\ t \mapsto z \end{array} \\
\\
\varphi : \begin{array}{l} x \mapsto z \\ y \mapsto t \\ z \mapsto x \\ t \mapsto y \end{array} & \varphi : \begin{array}{l} x \mapsto z \\ y \mapsto t \\ z \mapsto y \\ t \mapsto x \end{array} & \varphi : \begin{array}{l} x \mapsto t \\ y \mapsto z \\ z \mapsto x \\ t \mapsto y \end{array} & \varphi : \begin{array}{l} x \mapsto t \\ y \mapsto z \\ z \mapsto y \\ t \mapsto x \end{array}
\end{array}$$

There are 39 different (i.e., non-shape-isomorphic) multilabelled tree shapes with four leaves; 18 of them are bifurcating. These trees, along with their number of shape-automorphisms, are described in the Table 5.4. As it can be seen, between them they only amount to six different numbers of shape-automorphisms. We must assign now to each member of $\mathbf{MulShTree}_4$ a $\widehat{\mathrm{QI}}$ value that increases with the number of shape-automorphisms. We do it here with the following extra requirement: If $T_1$ has the same number shape-automorphisms than $T_2$, but the (unlabelled) shape of $T_1$ has more automorphisms than that of $T_2$, then $\widehat{\mathrm{QI}}(T_1) > \widehat{\mathrm{QI}}(T_2)$. In this way, we can assign to each tree in $\mathbf{MulShTree}_4$ a $\widehat{\mathrm{QI}}$ value $q_i$, with $i \in \{0, \dots, 13\}$ and $0 = q_0 < \cdots < q_{13}$, which is also indicated in the Table 5.4. And then we can define, for every $(T, \lambda) \in \mathbf{MulShTree}$,

$$\widehat{\mathrm{QI}}(T, \lambda) = \sum_{\{(Q, \lambda|_Q) : Q \in \mathrm{Part}_4(L(T))\}} \widehat{\mathrm{QI}}(T(Q), \lambda|_Q).$$

Of course, it would be possible to make a finer association $(T, \lambda) \in \mathbf{MulShTree}_4 \mapsto$

| Multifurcating | | | Bifurcating | | |
|---|---|---|---|---|---|
| Shape $(\cdot,\cdot,\cdot,\cdot)$ | | | Shape $((\cdot,\cdot),(\cdot,\cdot))$ | | |
| (1,2,3,4) | 24 | $q_{13}$ | ((1,2),(3,4)) | 8 | $q_{11}$ |
| (1,1,2,3) | 4 | $q_8$ | ((1,1),(2,3)) | 4 | $q_7$ |
| (1,1,2,2) | 8 | $q_{12}$ | ((1,2),(1,3)) | 2 | $q_5$ |
| (1,1,1,2) | 6 | $q_{10}$ | ((1,1),(2,2)) | 8 | $q_{11}$ |
| (1,1,1,1) | 24 | $q_{13}$ | ((1,1),(1,2)) | 2 | $q_5$ |
| Shape $((\cdot,\cdot,\cdot),\cdot)$ | | | ((1,2),(1,2)) | 4 | $q_7$ |
| ((1,2,3),4) | 6 | $q_9$ | ((1,1),(1,1)) | 8 | $q_{11}$ |
| ((1,2,3),1) | 2 | $q_4$ | Shape $(\cdot,(\cdot,(\cdot,\cdot)))$ | | |
| ((1,1,2),3) | 2 | $q_4$ | (1,(2,(3,4))) | 2 | $q_2$ |
| ((1,1,2),2) | 2 | $q_4$ | (1,(2,(3,3))) | 2 | $q_2$ |
| ((1,1,1),2) | 6 | $q_9$ | (1,(2,(2,3))) | 1 | 0 |
| ((1,1,2),1) | 2 | $q_4$ | (1,(2,(1,3))) | 1 | 0 |
| ((1,1,1),1) | 6 | $q_9$ | (1,(1,(2,3))) | 2 | $q_2$ |
| Shape $((\cdot,\cdot),\cdot,\cdot)$ | | | (1,(1,(2,2))) | 2 | $q_2$ |
| ((1,2),3,4) | 4 | $q_6$ | (1,(2,(1,2))) | 1 | 0 |
| ((1,1),2,3) | 4 | $q_6$ | (1,(1,(1,2))) | 1 | 0 |
| ((1,2),1,3) | 1 | $q_1$ | (1,(2,(1,1))) | 2 | $q_2$ |
| ((1,2),3,3) | 4 | $q_6$ | (1,(2,(2,2))) | 2 | $q_2$ |
| ((1,1),2,2) | 4 | $q_6$ | (1,(1,(1,1))) | 2 | $q_2$ |
| ((1,2),1,2) | 2 | $q_3$ | | | |
| ((1,1),1,2) | 2 | $q_3$ | | | |
| ((1,2),1,1) | 2 | $q_3$ | | | |
| ((1,1),1,1) | 4 | $q_6$ | | | |

Table 5.4: The 39 multilabelled tree shapes in **MulShTree**$_4$, their numbers of shape-automorphisms (second column in each block) and their $\widehat{\mathrm{QI}}$ value according to our schema (third column in each block).

$\widehat{\mathrm{QI}}((T,\lambda))$ by taking into account more information related to the balance than simply the number of automorphisms with and without labels.

With the schema explained above, we have the following results.

**Proposition 5.29.** *Let $n \in \mathbb{N}_{\geq 4}$. The maximum $\widehat{\mathrm{QI}}$ value on* **MulShTree**$_n$ *is reached exactly at two multilabelled tree shapes: the star with all leaves labelled with the same label, and the star with all leaves labelled with different labels.*

*Proof.* The maximum possible value of $\widehat{\mathrm{QI}}$ on **MulShTree**$_n$ is $\binom{n}{4}q_{13}$, and it is reached on the trees all whose quartets are of type $(1,1,1,1)$ or $(1,2,3,4)$. The only trees in **MulShTree**$_n$ satisfying this condition are the star with all leaves labelled with 1, and the star with all leaves labelled with different labels. □

**Theorem 5.30.** *Let $n \in \mathbb{N}_{\geq 4}$. The maximum value of the $\widehat{\mathrm{QI}}$ index for the trees in* **BinMulShTree**$_n$ *is reached exactly at the multilabelled tree shapes $M$ whose underlying tree shape is maximally balanced and their labels satisfy one of the following two conditions:*

*(i) All the leaves of $M$ are labelled differently.*

*(ii) If $M = M_1 * M_2$, then all the leaves in $M_i$, $i \in \{1,2\}$, are labelled equally.*

*Proof.* To begin with, notice that if we take a bifurcating multilabelled tree shape $M$ and we relabel all its leaves with different labels, obtaining a phylogenetic tree $M_P$, then its $\widetilde{\mathrm{QI}}$ value does not decrease, because the quartets of types $((1,2),(3,4))$ and $(1,(2,(3,4)))$ have $\widetilde{\mathrm{QI}}$ value greater or equal than all other quartets with their same unlabellled shape. Therefore, for every $M \in \mathbf{BinMulShTree}_n$,

$$\widetilde{\mathrm{QI}}(M) \leq \widetilde{\mathrm{QI}}(M_P) = \mathrm{QI}(M_P)q_{11} + \left(\binom{n}{4} - \mathrm{QI}(M_P)\right) q_2.$$

Since $q_{11} > q_2$, the maximum value of the right-hand side expression will be reached when $M_P$ has maximum QI value, that is, when its underlying tree shape $\pi_1(M_P) = \pi_1(M)$ is maximally balanced. We conclude that the maximum $\widetilde{\mathrm{QI}}$ value on $\mathbf{BinMulShTree}_n$ is

$$\mathrm{qib}(n) \cdot q_{11} + \left(\binom{n}{4} - \mathrm{qib}(n)\right) q_2$$

and it is reached, for the moment, at the maximally balanced phylogenetic trees, that is, at the multilabelled tree shapes under case *(i)* in the statement.

To complete the proof of the statement, we must prove that any other bifurcating multilabelled tree shape $M$ with $\widetilde{\mathrm{QI}}(M) = \mathrm{qib}(n)\cdot q_{11} + \left(\binom{n}{4} - \mathrm{qib}(n)\right) q_2$ must fall under case *(ii)* in the statement. Notice that such a $M \in \mathbf{BinMulShTree}_n$ must satisfy that:

- $\pi_1(M)$ must be maximally balanced, because otherwise

$$\widetilde{\mathrm{QI}}(M) \leq \mathrm{QI}(M_P)q_{11} + \left(\binom{n}{4} - \mathrm{QI}(M_P)\right) q_2 < \mathrm{qib}(n) \cdot q_{11} + \left(\binom{n}{4} - \mathrm{qib}(n)\right) q_2.$$

- All its fully symmetric quartets must be of types $((1,2),(3,4))$, $((1,1),(2,2))$ or $((1,1),(1,1))$, because otherwise the contribution to $\widetilde{\mathrm{QI}}(M)$ of its fully symmetric quartets would be smaller than $\mathrm{qib}(n) \cdot q_{11}$.

We begin by proving that if $M \in \mathbf{BinMulShTree}_n$ has $\pi_1(M) = T_n^{\mathrm{bal}}$ and $M = M_1 * M_2$ with each $M_i$ such that all its leaves have the same label, then $\widetilde{\mathrm{QI}}(M) = \mathrm{qib}(n) \cdot q_{11} + \left(\binom{n}{4} - \mathrm{qib}(n)\right) q_2$. Indeed, for let $Q$ be a quartet of such a bifurcating multilabelled tree shape $M$. Two possibilities arise:

- If $Q$ has fully symmetric shape, either it is all contained in $M_i$ for $i \in \{1,2\}$, or it has a cherry in $M_1$ and one in $M_2$. In the first case, $Q$ is of type $((1,1),(1,1))$, and in the second case it is of type $((1,1),(1,1))$ (if the labels of the leaves in $M_1$ and $M_2$ are the same) or $((1,1),(2,2))$ (if they are different). Both possibilities have the largest possible $\widetilde{\mathrm{QI}}$ value, $q_{11}$.
- If $Q$ has the shape of a caterpillar, then, again, it can be all contained in $M_i$ for some $i \in \{1,2\}$, or it can have one leaf in $M_1$ and three in $M_2$, or *vice versa*. The first possibility is of type $(1,(1,(1,1)))$, and the second either $(1,(1,(1,1)))$ or $(2,(1,(1,1)))$, depending on whether the labels of the leaves in $M_1$ and $M_2$ are equal or different. Both possibilities have $\widetilde{\mathrm{QI}}$ value $q_2$.

We prove now that if $M = M_1 * M_2 \in \textbf{BinMulShTree}_n$ is such that $\pi_1(M) = T_n^{\mathrm{bal}}$ and it has some pair of leaves with the same label but there is some pair of leaves with different label in some $M_i$, then $\widehat{\mathrm{QI}}(M)$ is smaller than $\mathrm{qib}(n) \cdot q_{11} + \left(\binom{n}{4} - \mathrm{qib}(n)\right) q_2$. To fix ideas, assume that there are two leaves in $M_1$ with different labels, say 1 and 2. Then, for every pair of different leaves $x, y$ in $M_2$ (and there exist at least two of them, because the shape of $M$ is maximally balanced and $n \geq 4$), the quartet $((1,2),(x,y))$ must be of type $((1,2),(3,4))$ in order to have $\widehat{\mathrm{QI}}$ value $q_{11}$. This implies that all leaves in $M_2$ have pairwise different labels and moreover different from 1 and 2. And since 1 and 2 stood for any pair of different leaves' labels in $M_1$, we conclude that all leaves in $M_2$ have pairwise different labels and that the sets of labels of $M_1$ and $M_2$ are disjoint. Now, since $M$ contains some pair of repeated labels, this repetition must appear in $M_1$: say that $M_1$ contains two copies of the label 1. But then, $M$ contains a quartet of type $((1,1),(3,4))$, whose $\widehat{\mathrm{QI}}$ value is $q_7 < q_{11}$.

This completes the proof of the statement. □

As for the minimum value of $\widehat{\mathrm{QI}}$, unfortunately we have to leave its characterization as an open problem. Intuitively, we would conjecture it to be attained at some caterpillar with a specific labelling. But this need not be the case even for bifurcating trees, at least when we restrict the set of labels. To end this section, we shall show this fact. More specifically, we shall prove that, for any values of the $q_i$'s in Table 5.4, the minimum value of $\widehat{\mathrm{QI}}$ on $\textbf{BinMulShTree}_n([2])$, the set of bifurcating multilabelled tree shapes with $n$ leaves labelled on [2] (or bifurcating *bilabelled* tree shapes with $n$ leaves), is not reached at any caterpillar if $n$ is large enough,

If we look in Table 5.4, we will see that the caterpillars in $\textbf{BinMulShTree}_4([2])$ with non-zero $\widehat{\mathrm{QIB}}$ value are exactly those whose cherry at the bottom is labelled with a single label. Let $\mathrm{trip}(M)$ denote the number of triples of $M \in \textbf{BinMulShTree}_n([2])$ of the form $(x,(y,z))$ such that $y = z$.

We shall now present our candidate to minimize $\widehat{\mathrm{QIB}}$ over the bilabelled caterpillars. Let $\widehat{M}_n^{\mathrm{cat}} \in \textbf{BinMulShTree}_n([2])$, $n \geq 2$, be the bilabelled caterpillar whose labelling is defined recurrently as follows:

$$\widehat{M}_2^{\mathrm{cat}} = (1,2),$$

and

$$\widehat{M}_{2k}^{\mathrm{cat}} = 1 * \widehat{M}_{2k-1}^{\mathrm{cat}}, \quad \widehat{M}_{2k+1}^{\mathrm{cat}} = 2 * \widehat{M}_{2k}^{\mathrm{cat}}.$$

In other words,

$$\begin{aligned}
\widehat{M}_{2k}^{\mathrm{cat}} &= (1,(2,(1,(2,(1,\dots,(1,2)\dots))))) \in \textbf{BinMulShTree}_{2k}([2]),\\
\widehat{M}_{2k+1}^{\mathrm{cat}} &= (2,(1,(2,(1,(2,\dots,(1,2)\dots))))) \in \textbf{BinMulShTree}_{2k+1}([2]).
\end{aligned}$$

Recall that, since we are dealing with bilabelled tree *shapes*, the labels 1 and 2 can be interchanged without changing the tree.

**Lemma 5.31.** *If $n \geq 4$, $\widehat{M}_n^{\mathrm{cat}}$ minimizes* trip *among the multilabelled caterpillars in* $\textbf{BinMulShTree}_n([2])$.

*Proof.* We proceed by induction over $n$. When $n = 4$, we have that

$$\mathrm{trip}((1,(2,(1,2)))) = \mathrm{trip}((2,(1,(2,1)))) = 1.$$

Let now $(x,(y,(z,t)))$ be any caterpillar in $\mathbf{BinMulShTree}_4([2])$. If $z = t$, then $(x,(y,(z,t)))$ contains two triples ending in the cherry with repeated labels $(z,t)$ and hence $\mathrm{trip}((x,(y,(z,t)))) \geq 2$. If $z \neq t$, then $y$ will be equal either to $z$ or to $t$ and hence either $(x,(y,z))$ or $(x,(y,t))$ end in a cherry with repeated labels, and thus $\mathrm{trip}((x,(y,(z,t)))) \geq 1$. So, 1 is the minimum possible value for trip on $\mathbf{BinMulShTree}_4([2])$.

Suppose now that the statement holds up to $n$ leaves, $n \geq 4$. Consider $M_{n+1}^{\mathrm{cat}} \in \mathbf{BinMulShTree}_{n+1}([2])$ to be any bilabelled caterpillar. Let $\ell$ be its leaf pending from the root; that is, $M_{n+1}^{\mathrm{cat}} = \ell * M_n^{\mathrm{cat}}$ for some bilabelled caterpillar with $n$ leaves $M_n^{\mathrm{cat}}$. After interchanging the labels 1 and 2 if necessary, we shall assume that $\ell = 1$ if $n+1$ is even and $\ell = 2$ if $n+1$ is odd. Assume that $M_n^{\mathrm{cat}}$ has $k$ leaves labelled 1 and the remaining $n-k$ leaves labelled 2. Then,

$$\mathrm{trip}(M_{n+1}^{\mathrm{cat}}) = \mathrm{trip}(M_n^{\mathrm{cat}}) + \binom{k}{2} + \binom{n-k}{2}.$$

Now, $f(x) = \binom{x}{2} + \binom{n-x}{2}$ is a parabola with vertex at $n/2$, and so $\widehat{M}_n^{\mathrm{cat}}$ would be such that this function is minimum. By the inductive hypothesis, it also minimizes trip in the second member of the equation above. This shows that $\mathrm{trip}(\widehat{M}_{n+1}^{\mathrm{cat}})$ is minimum. □

Notice that we are not saying that $\widehat{M}_n^{\mathrm{cat}}$ is the *only* bilabelled caterpillar that minimizes trip, just as we will not show it to be the *only* bilabelled caterpillar that minimizes $\widehat{\mathrm{QIB}}$.

**Corollary 5.32.** *If $n \geq 4$, $\widehat{M}_n^{\mathrm{cat}}$ minimizes $\widehat{\mathrm{QIB}}$ among the multilabelled caterpillars in* $\mathbf{BinMulShTree}_n([2])$.

*Proof.* We proceed by induction over $n$. When $n = 4$, it is simply due to the fact that $\widehat{\mathrm{QIB}}((1,(2,(1,2)))) = 0$. Suppose now that the statement holds up to $n$ leaves.

Consider $M_{n+1}^{\mathrm{cat}} \in \mathbf{BinMulShTree}_{n+1}([2])$ to be a bilabelled caterpillar and write it as $\ell * M_n^{\mathrm{cat}}$. As in the proof of the last lemma, we shall assume without any loss of generality that $\ell = 1$ if $n+1$ is even and $\ell = 2$ if $n+1$ is odd.

The quartets in $M_{n+1}^{\mathrm{cat}}$ not involving the shallowest leaf $\ell$ that contribute to $\widehat{\mathrm{QIB}}(M_{n+1}^{\mathrm{cat}})$ are exactly those defining $\widehat{\mathrm{QIB}}(M_n^{\mathrm{cat}})$, and a quartet $(\ell,(x,(y,z)))$ with the first label corresponding to the shallowest leaf contributes to $\widehat{\mathrm{QIB}}(M_{n+1}^{\mathrm{cat}})$ if, and only if, $(x,(y,z))$ contributes to $\mathrm{trip}(M_n^{\mathrm{cat}})$. Therefore,

$$\widehat{\mathrm{QIB}}(M_{n+1}^{\mathrm{cat}}) = \widehat{\mathrm{QIB}}(M_n^{\mathrm{cat}}) + \mathrm{trip}(M_n^{\mathrm{cat}})q_2.$$

By Lemma 5.31 and the inductive hypothesis, $\widehat{M}_{n+1}^{\mathrm{cat}}$ minimizes $\widehat{\mathrm{QIB}}$. □

So, in particular, $\widehat{\mathrm{QIB}}((1,(2,(1,(2,(1,2))))))$ is minimal among all 6-legged billabeled caterpillars. The next example shows that it not be the case in the whole $\mathbf{BinMulShTree}_6([2])$.

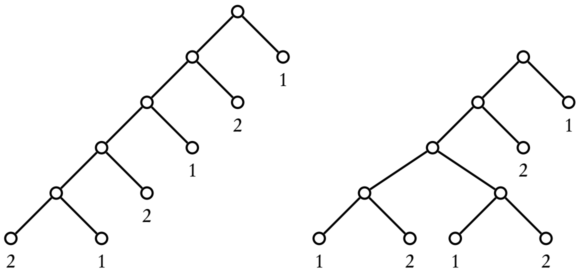


Figure 5.6: $\widehat{M}_6^{\text{cat}}$ and $M$ in $\mathbf{BinMulShTree}_6([2])$.

> **Example:**
> Consider the two multilabelled tree shapes with six leaves depicted in Figure 5.6, $\widehat{M}_6^{\text{cat}}$ and $M$. It can be checked that $\widetilde{\text{QIB}}(\widehat{M}_6^{\text{cat}}) = 6q_2$ while $\widetilde{\text{QIB}}(M) = 4q_2 + q_7$. So, $\widetilde{\text{QIB}}(\widehat{M}_6^{\text{cat}}) \leq \widetilde{\text{QIB}}(M)$ if, and only if, $2q_2 \leq q_7$. This can or cannot be the case.

So, $\widetilde{\text{QIB}}(\widehat{M}_6^{\text{cat}})$ may not be minimum in $\mathbf{BinMulShTree}_6([2])$ in general. The only way it *might* be minimum is that we set $q_7 > 2q_2$, and even then we do not claim it to be so. But, since $\widetilde{\text{QIB}}(\widehat{M}_6^{\text{cat}})$ is minimum among all bilabelled caterpillars in $\mathbf{BinMulShTree}_6([2])$, we conclude that it need not be the minimum in the whole $\mathbf{BinMulShTree}_6([2])$ given values of $q_2$ and $q_7$. The final theorem of this section will show that *for any given values of $q_2$ and $q_7$*, $\widetilde{\text{QIB}}(\widehat{M}_n^{\text{cat}})$ is not minimum in $\mathbf{BinMulShTree}_n([2])$ if $n$ is large enough. In order to prove it, let us establish the following version of Lemma 5.11.

**Lemma 5.33.** *Let $M_0$ be a bifurcating bilabelled tree, $z \in L(M_0)$, and let $M, M'$ be two trees obtained by appending to the leaf $z$ in $M_0$ the rooted bifurcating bilabelled subtrees $M_z, M'_z \in$ $\mathbf{BinMulShTree}_n([2])$, respectively. Assume that $\Lambda(M_z) = \Lambda(M'_z)$ are equal as multisets of labels $1, 2$. Then,*

$$\widetilde{\text{QIB}}(M') - \widetilde{\text{QIB}}(M) = \widetilde{\text{QIB}}(M'_z) - \widetilde{\text{QIB}}(M_z) + (|L(M_0)| - 1)\left(\text{trip}(M'_z) - \text{trip}(M_z)\right) q_2.$$

*Proof.* Let $Q = \{a, b, c, d\} \in \text{Part}_4(L(M)) = \text{Part}_4(L(M'))$. Then:

- If $Q \cap L(M_z) = \emptyset$, then $M(Q) = M'(Q) = M_0(Q)$.
- If $Q \cap L(M_z) = \{d\}$ —for instance—, then $M(Q) = M'(Q) = M_0(\{a, b, c, z\})$.
- If $Q \cap L(M_z) = \{c, d\}$ —for instance—, then two cases arise: either $M_0(\{a, b, z\}) = (a, (b, z))$, and in this case $M(Q) = M'(Q) = (a, (b, (c, d)))$; or $M_0(\{a, b, z\}) = ((a, b), z)$, and so $M(Q) = M'(Q) = ((a, b), (c, d))$.
- If $Q \cap L(M_z) = \{b, c, d\}$ —for instance—, then it may happen that $M(Q) \neq M'(Q)$ since, even if they will be both caterpillars, there are two possibilities:

- if $b, c, d$ are the same label $\ell$, then $M(Q) = M'(Q) = (a, (\ell, (\ell, \ell)))$; but
- if, say, $b, c$ are labelled 1 and $d$ is labelled 2, then it can happen that $M(Q) = (a, (2, (1, 1)))$ and $M'(Q) = (a, (1, (1, 2)))$: the first contributes $q_2$ to $\widetilde{\text{QIB}}(M)$ and the latter 0 to $\widetilde{\text{QIB}}(M')$.

Now, recall that the only caterpillars with four leaves in **BinMulShTree**$_4([2])$ that add $q_2$ to the total $\widehat{\mathrm{QIB}}$ are those whose cherry at the bottom has their leaves equally labelled. Hence, this last case adds $(|L(M_0)| - 1)q_2$ for each triplet whose cherry has its leaves thus labelled. Therefore,

$$\widehat{\mathrm{QIB}}(M') - \widehat{\mathrm{QIB}}(M) = \widehat{\mathrm{QIB}}(M'_z) - \widehat{\mathrm{QIB}}(M_z) + (|L(M_0)| - 1)\left(\mathrm{trip}(M'_z) - \mathrm{trip}(M_z)\right) q_2.$$

□

**Theorem 5.34.** *For every $0 < q_2 < q_7$, there exists $N \in \mathbb{N}$ such that, if $n \geq N$, the bilabelled tree shape that minimizes $\widehat{\mathrm{QIB}}$ over* **BinMulShTree**$_n([2])$ *is not that of a caterpillar.*

*Proof.* Consider $\widehat{M}_n^{\mathrm{cat}}$; we have already shown that this is the bilabelled caterpillar that minimizes $\widehat{\mathrm{QIB}}$ over the bilabelled caterpillars. Consider the bilabelled tree shape $M'$ produced by replacing the subtree formed by the four bottom-most leaves of $\widehat{M}_n^{\mathrm{cat}}$, $(1,(2,(1,2)))$ or $(2,(1,(2,1)))$, by the subtree $\widehat{M}_4^{\mathrm{bal}} = ((1,2),(1,2))$ (cf. Figure 5.6). By Lemma 5.33,

$$\begin{aligned}\widehat{\mathrm{QIB}}(\widehat{M}_n^{\mathrm{cat}}) - \widehat{\mathrm{QIB}}(M') &= \widehat{\mathrm{QIB}}(\widehat{M}_4^{\mathrm{cat}}) - \widehat{\mathrm{QIB}}(\widehat{M}_4^{\mathrm{bal}})\\ &\quad + (n-4)\left(\mathrm{trip}(\widehat{M}_4^{\mathrm{cat}}) - \mathrm{trip}(\widehat{M}_4^{\mathrm{bal}})\right) q_2\\ &= -q_7 + (n-4)q_2\end{aligned}$$

since $\mathrm{trip}(\widehat{M}_4^{\mathrm{cat}}) = 1, \mathrm{trip}(\widehat{M}_4^{\mathrm{bal}}) = 0$, $\widehat{\mathrm{QIB}}(\widehat{M}_4^{\mathrm{cat}}) = 0$, and $\widehat{\mathrm{QIB}}(\widehat{M}_4^{\mathrm{bal}}) = q_7$. Now, the expression above is positive whenever

$$n \geq 4 + \frac{q_7}{q_2}.$$

Thus, for every pair $(q_2, q_7) \in \mathbb{R}_{>0}$ with $q_2 < q_7$, if $n$ is large enough it will happen that $\widehat{\mathrm{QIB}}(M') < \widehat{\mathrm{QIB}}(\widehat{M}_n^{\mathrm{cat}})$. Since $\widehat{\mathrm{QIB}}(\widehat{M}_n^{\mathrm{cat}})$ is minimum among the caterpillars in **BinMulShTree**$_n([2])$, this proves that the minimum $\widehat{\mathrm{QIB}}$ value on **BinMulShTree**$_n([2])$ is not reached at a caterpillar. □

### A sampling consistent probabilistic model for multilabelled trees

If we knew of some sampling consistent probabilistic model for multilabelled trees, then by Theorems 5.27 and 5.28 we would be able to compute the expected value and the variance under it of this Quartet index $\widehat{\mathrm{QI}}$. We begin by proving the following Lemma.

**Lemma 5.35.** *Let $(P_n^*)_n$ be a probabilistic model for mul-shapes such that*

*(i) The probabilistic model of trees $(P_n^{\mathrm{Tree},*})_n$ :* **Tree** $\longrightarrow [0,1]$ *it induces is sampling consistent.*

*(ii) For every $n \geq 2$, for every $M_0 \in$* **MulShTree**$_{n-1}$ *and for every $T \in$* **Tree**$_n$

$$\begin{aligned}&|\{x \in L(T) : T(-x) = \pi_1(M_0)\}| \cdot P_{n-1}^*(M_0 : \pi_1(M_0))\\ &\quad = \sum_{\substack{M \in \mathbf{MulShTree}_n\\ \pi_1(M) = T}} |\{x \in L(M) : M(-x) = M_0\}| \cdot P_n(M|T).\end{aligned}$$

*Then, $(P^*_n)_n$ is sampling consistent.*

*Proof.* With the notations in the statement of this lemma, we begin by considering

$$\begin{aligned}
P^*_{n-1}(M_0) &= P^*_{n-1}(M_0|\pi_1(M_0)) \cdot P^{\mathbf{Tree},*}_{n-1}(\pi_1(M_0))\\
&= P^*_{n-1}(M_0|\pi_1(M_0)) \sum_{T\in\mathbf{Tree}_n} \frac{|\{x\in L(T) : T(-x)=\pi_1(M_0)\}|}{n} P^{\mathbf{Tree},*}_n(T)\\
&\quad \text{(due to the sampling consistency of } (P^{\mathbf{Tree},*}_n)_n)\\
&= \sum_{T\in\mathbf{Tree}_n} \sum_{\substack{M\in\mathbf{MulShTree}_n\\ \pi_1(M)=T}} \frac{|\{x\in L(M) : M(-x)=M_0\}|}{n} P^*_n(M|T)P^{\mathbf{Tree},*}_n(T)\\
&\quad \text{(by the second condition in the statement of this lemma)}\\
&= \sum_{M\in\mathbf{MulShTree}_n} \frac{|\{x\in L(M) : M(-x)=M_0\}|}{n} P^*_n(M),
\end{aligned}$$

which is what we wanted to prove. □

**Theorem 5.36.** *Let $(P^*_n)_n$ be a probabilistic model for mul-shapes such that*

*(i) The probabilistic model of trees $(P^{T,*}_n)_n$ it induces is sampling consistent*

*(ii) There exists an $N\in\mathbb{N}$ such that, for every $M\in\mathbf{MulShTree}_n$ with $\pi_1(M)=T$, $P_n(M|T)$ is the fraction of (arbitrary) labelings $\lambda : L(T)\to[N]$ such that the mul-shape of the mul-tree $(T,\lambda)$ is $M$:*

$$P^*_n(M|T) = \frac{|\{\lambda : L(T)\to[N] : \pi^*(T,\lambda)=M\}|}{N^n}$$

*Then, $(P^*_n)_n$ is sampling consistent.*

*Proof.* In order to prove this result we shall show that all the models satisfying both of the relations stated in this theorem also satisfy the hypotheses of Lemma 5.35.

Let $M_0\in\mathbf{MulShTree}_{n-1}$ and $T\in\mathbf{Tree}_n$. Then, in order to use the previous lemma, we want to check whether

$$\begin{aligned}
&|\{x\in L(T) : T(-x)=\pi_1(M_0)\}| \frac{|\{\lambda_{n-1} : L(\pi_1(M_0))\to[N] : T(-x)=\pi_1(M_0)\}|}{N^{n-1}}\\
&\quad = \sum_{\substack{M\in\mathbf{MulShTree}_n\\ \pi_1(M)=T}} |\{x\in L(T) : M(-x)=M_0\}| \frac{|\{\lambda_n : L(T)\to[N] : \pi^*(T,\lambda_n)=M\}|}{N^n},
\end{aligned}$$

or, equivalently, that

$$\begin{aligned}
&|\{x\in L(T) : T(-x)=\pi_1(M_0)\}|\cdot|\{\lambda_{n-1} : L(\pi_1(M_0))\to[N] : \pi^*(T,\lambda_{n-1})=M_0\}|\\
&\quad = \sum_{\substack{M\in\mathbf{MulShTree}_n\\ \pi_1(M)=T}} |\{x\in L(T) : M(-x)=M_0\}| \frac{|\{\lambda_n : L(T)\to[N] : \pi^*(T,\lambda_n)=M\}|}{N}.
\end{aligned}$$

As for the first member of the equation,

$$\begin{aligned}
&|\{x \in L(T) : T(-x) = \pi_1(M_0)\}| \cdot |\{\lambda_{n-1} : L(\pi_1(M_0)) \to [N] : \pi^*(T, \lambda_{n-1}) = M_0\}| \\
&\quad = |\{(x, \lambda_{n-1}) \in L(T) \times [N]^{L(\pi_1(M_0))} : T(-x) = \pi_1(M_0), \pi^*(\pi_1(M_0), \lambda_{n-1}) = M_0\}| \\
&\quad = \frac{1}{N}|\{(x, \lambda_n) \in L(T) \times [N]^{L(T)} : \pi^*(T, \lambda_n)(-x) = M_0\}| \\
&\quad = \sum_{\substack{M \in \mathbf{MulShTree}_n \\ \pi_1(M)=T}} |\{x \in L(T) : M(-x) = M_0\}| \frac{|\{\lambda_n : L(T) \to [N] : \pi^*(T, \lambda_n) = M\}|}{N},
\end{aligned}$$

since $L(M) = L(\pi_1(M))$ for any $M \in \mathbf{MulShTree}$. □

As a consequence, any probabilistic model for trees that is sampling consistent, endowed with a completely random labelling on $[N]$ for some $N \in \mathbb{N}_{\geq 1}$ is also sampling consistent.

## 5.5 Discussion

In this chapter we have introduced a new balance index that makes sense for bifurcating *and* multifurcating trees. This is worth noting, since the two most widely used balance indices, that is, the Colless and the Sackin indices, have some issues regarding its extension to multifurcating trees. The first one cannot by extended using the definition as it is, since it assumes the existence of two and only two children at each internal node —although some authors have given extensions for multifurcating trees [86]. The second one can be readily extended, but the interpretation of its value might be difficult, since every *taxonomic tree* (that is, a tree such that every leaf has the exact same depth, even if that means that some internal nodes have out-degree 1) would then have the same Sackin index regardless of any intuition of the balance of this shape. In this chapter, we have proven that the Quartet index has a meaningful interpretation in the multifurcating case.

Furthermore, QI has the largest range of values in the literature for a fixed number of leaves $n \in \mathbb{N}$: from 0 to $q_4\binom{n}{4} = O(n^4)$ in the multifurcating case, and from 0 to $O(n^4)$ in the bifurcating one. *A priori*, this reduces the chance of two different trees presenting the same QI (as it was shown in Figure 3.1). As we have shown, this probability is zero when it comes to the extreme values of QI: both its multifurcating and its bifurcating maxima are unique, and its minimum (which happens to be the same in both cases) is also unique. This is not the case for the Colless index (as we saw in the previous chapter), nor it is for the Sackin index [39]; it is, however, the case with the Cophenetic index [85].

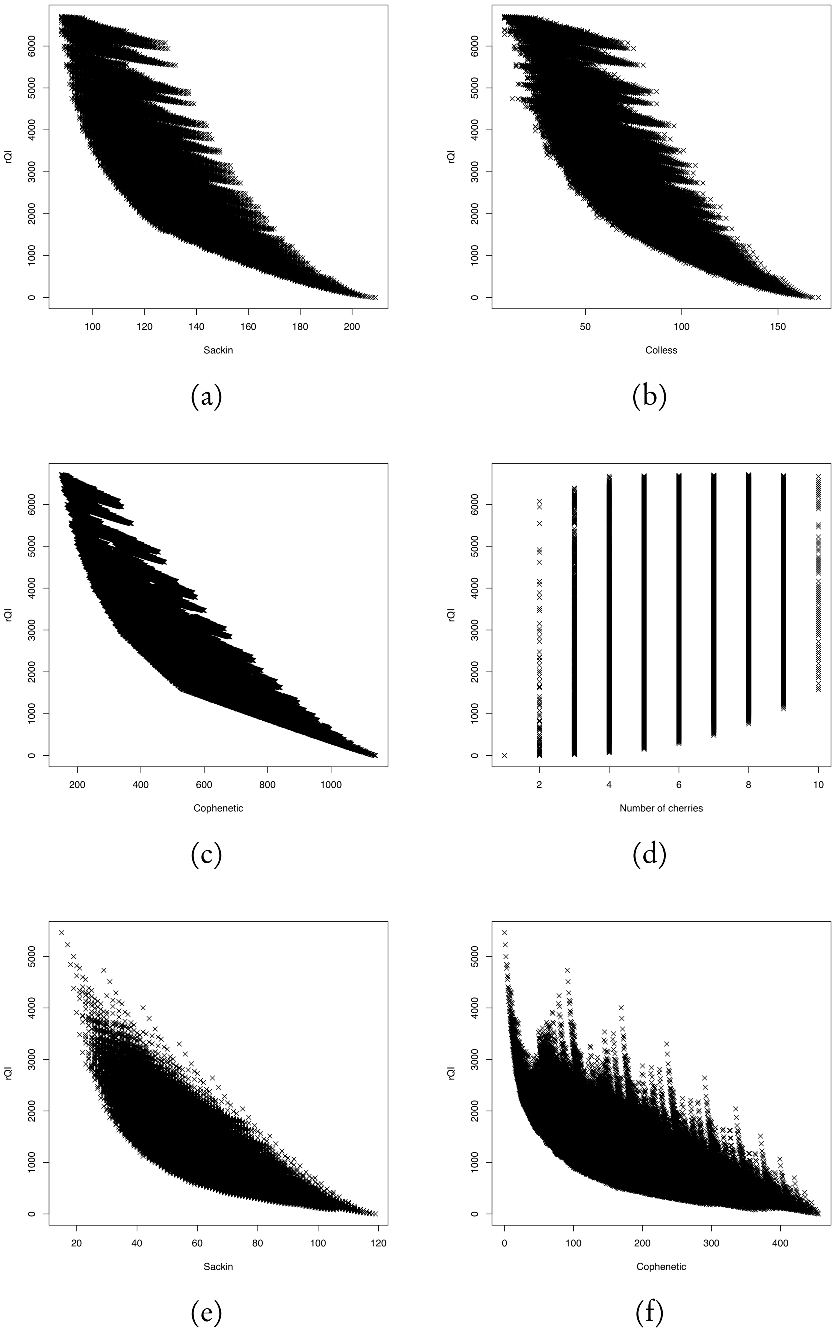
Nevertheless, the Quartet index strongly correlates with both the Colless and Sackin indices, as well as the Cophenetic index. Figure 5.7 shows some scatterplots of the Spearman correlation of QI versus *(a)* the Sackin index, *(b)* the Colless index, *(c)* the Cophenetic index, *(d)* the number of cherries of a given tree, QIB versus *(e)* the Sackin index and *(f)* the Cophenetic index. We point out that the Spearman correlation in the case of the number of cherries (Table 5.5) of a given tree is rather small, which might strike us as counterintuitive, but because of the small range of values of this last index —only about O(log $n$)—, it is not surprising. Notice also that the correlation between

QI and the other indices is negative, since QI increases with balance, while all the other indices decrease with it.

Finally, we want to emphasize that one of the most interesting features of this new index is that it can be easily extended to other sets of directed graphs, thus opening the door to consider the question of how balanced are the graphs in those sets. The case for binary multilabelled trees is only sketched in this memoir, but this will be the direction of future research.

| Correlation | Value |
|---|---|
| QIB *vs* $S$ on **BinTree**$_{20}$ | −0.889 |
| QIB *vs* $C$ on **BinTree**$_{20}$ | −0.893 |
| QIB *vs* $\Phi$ on **BinTree**$_{20}$ | −0.935 |
| QI *vs* number of cherries on **BinTree**$_{20}$ | 0.165 |
| QI *vs* $S$ on **Tree**$_{15}$ | −0.787 |
| QI *vs* $\Phi$ on **Tree**$_{15}$ | −0.827 |

Table 5.5: Spearman's correlations corresponding to the scatterplots in Fig. 5.7.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.7: Scatterplot of $\underset{\sim}{QI}$ *versus*: (a) the Sackin index on **BinTree**$_{20}$; (b) the Colless index on **BinTree**$_{20}$; (c) the total cophenetic index on **BinTree**$_{20}$; (d) the number of cherries on **BinTree**$_{20}$; (e) the Sackin index on **Tree**$_{15}$; (f) the total cophenetic index on **Tree**$_{15}$.

CHAPTER 6

# Conclusions and future work

All shall be well. And all shall be
well. And all manner of things shall
be well.

Julian of Norwich, *Revelations of Divine love*, 14th century

THE MAIN hypothesis motivating the quantitative study of phylogenetic tree shapes is that the branching pattern of a given tree somehow reflects some properties of the evolutionary processes that have given rise to it. In this memoir, and all through the research that has produced it, we have presented some new techniques and results concerning some of its aspects:

- In the Preliminaries Section 1.4.1, we have presented the solution to a family of recurrences involved in the computation of moments of balance indices under the Uniform model for bifurcating phylogenetic trees.

- In Chapter 2, we have solved the problem of the characterization of the bifurcating trees that attain the minimum Colless index, which was implicitly raised by the publication of [19] in 1982, in close collaboration with Mareike Fischer, Lina Herbst and Kristina Wicke. Therein, we have presented closed formulæ that compute this minimum value. We believe that the characterization given of the trees attaining it is quite beautiful, and is directly related to the binary decomposition of the number of leaves $n$.

- As an alternative to the Colless index, we have introduced the Quadratic Colless index (Chapter 3), in which the difference between the number of leaves of the pending subtrees of each internal node is squared. This index captures all the intuitiveness of the Colless index, but it is also better-suited for its analysis (as sums of squares usually behave better than sums of absolute values do). We show that its maximum and minimum values are attained exactly at the caterpillars and the maximally balanced trees, respectively, proving that its range of values is $O(n^3)$ for any number of leaves $n$, an order of magnitude bigger than that of the Colless index. The computation of its expected value and variance under the Yule and Uniform models is also performed —whereas, in the case of the Colless index, the expected value and variance under the Uniform model are still unknown. The chapter ends with some numerical results exposing that the discriminatory power of this index appears to be larger than that of all the other considered balance indices, save for the Quartet index.

- The third chapter is dedicated to the study of one of the original proposals by Sackin [102]: to measure the balance of a tree by assessing the variation of its leaves' depth. This sounds like a fairly reasonable idea, but the phylogenetic community ended up prefering the sum of all the depths, as defined in [107]. In this chapter we take into account Sackin's own proposal, only to find out that, in fact, the phylogenetic community has been wise prefering the sum of all leaves' depths over their variance. Indeed: we have shown that, although the maximum value of this index is always attained at the caterpillars, its minimum over bifurcating trees is almost never reached by a maximally balanced tree —in fact, given a number of leaves $n$ we have not been able to give a characterization of the bifurcating trees with $n$ leaves that attain the minimum value of the Variance of depths, but only two algorithms that run in $O(n \log_2(n))$.

- Finally, we end the central chapters of this thesis by the introduction of the Quartet index. Defined over multifurcating trees, the Quartet index satisfies all the conditions a good balance index should satisfy: it classifies as most and least balanced bifurcating trees the maximally balanced trees and the caterpillars, respectively, and has fairly good statistical properties. Indeed: we have been able to compute both its expected value and its variance under Chen-Ford-Winkel's $\alpha$-$\gamma$-model and Aldous' $\beta$-model, of which the Yule and Uniform models are but instances. Its range of values is $O(n^4)$, which is the largest of the balance indices reviewed so far. Finally, we want to point out that this balance index seems well-fitted to be extended to other sets of directed graphs such as multilabelled trees and some kinds of phylogenetic networks, as we show that its good statistical properties may be preserved in these contexts.

Throughout this process, new questions have been raised that we believe worthy of further study:

- To study the seemingly fractal structure presented by the sequence $\widetilde{c}(n)$ as presented in Figure 2.4, and show whether it has some relationship with the Takagi curve, as it seems to be the case, or not.

- To compute both the expected value and the variance of the Colless index under the Uniform model. To do this, it will probably be necessary an extension of the techniques introduced in Section 1.4.1 to solve recurrences like those considered therein but with independent terms involving floor and ceiling functions.

- To find a characterization of the trees that attain the minimum Variance of depths over bifurcating trees with $n$ leaves in $O(\log_2 n)$ (since only the vector of depths of the trees is needed).

- In relation to this last point, to solve all the weaker problems presented in the Discussion of Chapter 4 concerning the structure of the families of trees attaining the minimum Variance of depths, according to the data we have generated.

- To produce a closed expression for the solution of the recurrence presented in Corollary 5.15, giving the value of the maximum QIB value.

- To further study the natural extensions of the Quartet index to multilabelled trees and phylogenetic networks. In particular, to find the multilabelled tree that minimizes the Quartet index and to study the extreme values of this index for networks, as well as its statistical properties.

He cannot say he has understood all of this. Possibly he's more confused now than ever. But all these moments he's contemplated — something has occurred. The moments feel substantial in his mind, like stones. Kneeling, reaching down toward the closest one, running his hand across it, he finds it smooth, and slightly cold.

He tests the stone's weight; he finds he can lift it, and the others too. He can fit them together to create a foundation, an embankment, a castle.

To build a castle of appropriate size, he will need a great many stones. But what he's got, now, feels like an acceptable start.

Jonathan Blow, *Braid*

# Bibliography

[1] P.-M. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51(6):866–872, 2002.

[2] D. J. Aldous. Probability distributions on cladograms. *Random discrete structures*, pages 1–18, 1996.

[3] D. J. Aldous. Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science*, 16:23–34, 2001.

[4] P. C. Allaart and K. Kawamura. The Takagi function: A survey. *Real Analysis Exchange*, 37(1):1–54, 2012.

[5] M. Avino, T. N. Garway, et al. Tree shape-based approaches for the comparative study of cophylogeny. *Ecology and Evolution*, 9(12):6756–6771, 2019.

[6] K. Bartoszek, T. M. Coronado, A. Mir, and F. Rosselló. Squaring within the Colless index yields a better balance index. *arXiv*, 2020.

[7] M. B. Blum and O. François. On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited. *Mathematical Biosciences*, 195:141–153, 2005.

[8] M. B. Blum, O. François, and S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*, 16(4):2195–2214, 1996.

[9] T. Bonnin and J. Lombard. Situer l'analyse phylogénétique entre les sciences historiques et expérimentales. *Philosophia Scientiae*, 29:131–148, 2019.

[10] N. Bortolussi, E. Durand, M. Blum, and O. François. aptreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*, 22(3):363–364, 2006.

[11] A. V. Z. Brower and E. Rindal. Reality check: A reply to Smith. *Cladistics*, pages 464–645, 2013.

[12] G. Brown. *Language and understanding*. Oxford University Press, 1994.

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

[14] G. Cardona, F. Rosselló, and G. Valiente. Extended newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*, 9(532), 2008.

[15] C. Cathcart. A probabilistic assessment of the Indo-Aryan Inner-Outer Hypothesis. *Journal of Historical Linguistics*, 10(1):42–86, 2020.

[16] L. Chalmandrier, C. Albouy, et al. Comparing spatial diversification and metapopulation models in the Indo-Australian archipelago. *Royal Society Open Science*, 5, 2018.

[17] W. Chang, C. Cathcart, D. Hall, and A. Garrett. Ancestry-constrained phylogenetic analysis supports the Indo-European Steppe Hypothesis. *Language*, 91:194–244, 2015.

[18] B. Chen, D. Ford, and M. Winkel. A new family of Markov branching trees: the alpha-gamma model. *Electron. J. Probab.*, 14:400–430, 2009.

[19] D. H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31:100–104, 1982.

[20] D. H. Colless. Relative symmetry of cladograms and phenograms: An experimental study. *Systematic Biology*, 44:102–108, 19995.

[21] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms (3rd edition)*. The MIT Press, 2009.

[22] T. M. Coronado, M. Fischer, L. Herbst, F. Rosselló, and K. Wicke. On the minimum value of the Colless index and the bifurcating trees that achieve it. *Journal of Mathematical Biology*, 80:1993–2054, 2020.

[23] T. M. Coronado, A. Mir, and F. Rosselló. The probabilities of trees and cladograms under Ford's $\alpha$-model. *Scientific World Journal*, 2018.

[24] T. M. Coronado, A. Mir, F. Rosselló, and L. Rotger. On Sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index. *BMC Bioinformatics*, 21(154), 2020.

[25] T. M. Coronado, A. Mir, F. Rosselló, and G. Valiente. A balance index for phylogenetic trees based on rooted quartets. *Journal of Mathematical Biology*, 79:1105–1148, 2019.

[26] T. M. Coronado, G. Riera, and F. Rosselló. *The Fair Proportion Is a Shapley Value on Phylogenetic Networks Too*, pages 77–87. Springer International Publishing, 2018.

[27] T. M. Coronado and F. Rosselló. The minimum value of the Colless index. *arXiv preprint arXiv:1903.11670*, 2019.

[28] T. Cunha and G. Giribet. A congruent topology for deep gastropod relationships. *Proceedings of the Royal Society B*, 286, 2019.

[29] T. Deli, C. Kiel, and C. D. Schubart. Phylogeographic and evolutionary history analyses of the warty crab *Eriphia verrucosa* (Decapoda, Brachyura, Eriphiidae) unveil genetic imprints of a late Pleistocene vicariant event across the Gibraltar Strait, erased by postglacial expansion and admixture among refugial lineages. *BMC Evolutionary Biology*, 19(105), 2019.

[30] F. Delsuc, H. Brinkmann, and H. Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6:361–375, 2005.

[31] J. d'Huy. Matriarchy and prehistory: a statistical method for testing an old theory. *Les Cahiers de l'ARSS*, 19:159–170, 2017.

[32] J. d'Huy. Du nouveau sur Polyphème. *Mythologie Française*, 277:15–18, 2019.

[33] J. Diniz. Phylogenetic diversity and conservation priorities under distinct models of phenotypic evolution. *Conservation Biology*, 18:698–704, 2004.

[34] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999.

[35] A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambayt. Phylogenetics and Dating with Confidence. *BMC Bioinformatics*, 4(88), 2006.

[36] S. Duchene, R. Bouckaert, D. A. Duchene, T. Stadler, and A. J. Drummond. Phylodynamic model adequacy using posterior predictive simulations. *Systematic Biology*, 68:358–364, 2018.

[37] J. Farris and M. Källersjö. Asymmetry and explanations. *Cladistics*, 14:159–166, 1998.

[38] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2004.

[39] M. Fischer. Extremal values of the Sackin balance index for rooted binary trees. https://arxiv.org/abs/1801.10418, 2018.

[40] M. Fischer, L. Herbst, and K. Wicke. Extremal properties of the Colless balance index for rooted binary trees. *arXiv preprint arXiv:1904.09771*, 2019.

[41] M. Fischer and V. Liebscher. On the balance of unrooted trees. https://arxiv.org/abs/1510.07882, 2018.

[42] M. Fisler, C. Crémière, P. Darlu, and G. Lecointre. The treeness of the tree of historical trees of life. *PlosOne*, 15(1), 2020.

[43] D. J. Ford. Probabilities on cladograms: introduction to the alpha model. https://arxiv.org/abs/math/0511246v1, 2005.

[44] P. Forster, L. Forster, C. Renfrew, and M. Forster. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Science of the United States of America*, 117:9241–9243, 2020.

[45] M. Fuchs and E. Y. Jin. Equality of shapley value and fair proportion index in phylogenetic trees. *Journal of Mathematical Biology*, 71:1133–1147, 2015.

[46] G. Fusco and Q. C. Cronk. A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology*, 175:235–243, 1995.

[47] D. J. Futuyma. Evolution, science and society: Evolutionary biology and the national research agenda, 1999.

[48] G. Gasper and M. Rahman. *Basic Hypergeometric Series*. Cambridge University Press, 1990.

[49] P. A. Goloboff, J. S. Arias, and C. A. Szumik. Comparing tree shapes: beyond symmetry. *Zoologica Scripta*, 46:637–648, 2017.

[50] R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, 1989.

[51] R. Graham, D. Knuth, and O. Patashnik. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*. Addison-Wesley, 1994.

[52] R. D. Gray and Q. D. Atkinson. Language-tree divergence times support the anatolian theory of Indo-European origin. *Nature*, 426:435–439, 2003.

[53] R. D. Gray, A. J. Drummond, and S. J. Greenhill. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323:479–483, 2009.

[54] R. D. Gray and F. M. Jordan. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405:1052–1055, 2000.

[55] W. Gregg, S. Ather, and M. Hahn. Gene-tree reconciliation with mul-trees to resolve polyploidy events. *Systematic Biology*, 66:1007–1018, 2017.

[56] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332, 2004.

[57] E. F. Harding. The probabilities of rooted tree -shapes generated by random bifurcation. *Advances in Appl. Probabilit*, 3:44–77, 1971.

[58] H. Hasegawa. Phylogeny, host-parasite relationship and zoogeography. *Korean J Parasitol*, 37:197–213, 1999.

[59] M. Hayati, B. Shadgar, and L. Chindelevitch. A new resolution function to evaluate tree shape statistics. *PlosOne*, 14(11), 2019.

[60] S. B. Heard. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46:1818–1826, 1992.

[61] S. Y. Her. Bane, loss and phylogeny. http://thephilosophersmeme.com/2015/11/26/bane-loss-and-phylogeny/, 2015.

[62] D. Hillis, J. Bull, and M. White. Experimental phylogenetics: Generation of a known phylogeny. *Science*, 255:589–592, 1992.

[63] T. Holton, M. Wilkinson, and D. Pisani. The shape of modern tree reconstruction methods. *Systematic Biology*, 63:436–441, 2014.

[64] R. R. Hudson et al. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

[65] J. P. Huelsenbeck and M. Kirkpatrick. Do phylogenetic methods produce trees with biased shapes? *Evolution*, 50:1418–1424, 1996.

[66] D. Huson, V. Moulton, M. Steel, and M. Freiberger. Reconstructing the tree of life. https://plus.maths.org/content/reconstructing-tree-life, 2008.

[67] T. Huyse and F. A. M. Volckaert. Comparing host and parasite phylogenies: *Gyrodactylus Flatworms* jumping from Goby to Goby. *Systematic Biology*, 54:710–718, 2005.

[68] H. Kayondo, S. Mwalili, and J. Mango. Inferring multi-type birth-death parameters for a structured host population with application to HIV epidemic in Africa. *Computational Molecular Bioscience*, 9:108–131, 2019.

[69] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47:1171–1181, 1993.

[70] D. Knuth. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*. Addison-Wesley, 1997.

[71] J. Koenig. Dictionary of obscure sorrows. https://www.dictionaryofobscuresorrows.com/, 2006.

[72] T. Kubo and Y. Iwasa. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution*, 49:694–704, 1995.

[73] T. Tsan-Yuk Lam, C.-C. Hon, and J. W. Tang. Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Crit Rev Clin Lab Sci*, 23:5–49, 2010.

[74] J.-M. List, J. Sylvestre Pathmanathan, P. Lopez, and E. Bapteste. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. *Biology Direct*, 11(39), 2016.

[75] I. Grant Macdonald. *Symmetric functions and Hall polynomials (2nd edition)*. Oxford University Press, 1995.

[76] F. A. Matsen. A geometric approach to tree shape statistics. *Systematic Biology*, 55:652–661, 2006.

[77] P. McCullagh, J. Pitman, M. Winkel, et al. Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002, 2008.

[78] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164:81–92, 2000.

[79] C. Metzig, O. Ratmann, D. Bezemer, and C. Colijn. Phylogenies from dynamic networks. *PLoS Computational Biology*, 15, 2019.

[80] D. P. Mindell. The tree of life: metaphor, model, and heuristic device. *Systematic biology*, 62(3):479–489, 2013.

[81] A. Mir and F. Rosselló. The mean value of the squared path-difference distance for rooted phylogenetic trees. *ScienceDirect*, 371(1):168–176, 2010.

[82] A. Mir and F. Rosselló. The median of the distance between two leaves in a phylogenetic tree. In M. P. Rocha, F. Fernández Riverola, H. Shatkay, and J. M. Corchado, editors, *Advances in Bioinformatics*, pages 131–135, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[83] A. Mir and F. Rosselló. The mode of the distance between two leaves in a phylogenetic tree. In *Contributed Talk at the X Jornadas de Bioinformática at Málaga, Spain*, 2010.

[84] A. Mir and F. Rosselló. On the distribution of the distances between pairs of leaves in phylogenetic trees. In *Contributed Talk at the BYOTECHNO 2011: The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*, 2011.

[85] A. Mir, L. Rotger, and F. Rosselló. A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1):125–136, 2013.

[86] A. Mir, L. Rotger, and F. Rosselló. Sound Colless-like balance indices for multifurcating trees. *PlosOne*, 13(9), 2018.

[87] A. Mir-Fuentes. Arbres *k*-furcats amb valors extrems de l'índex de Sackin. Dissertation, Universitat de les Illes Balears, 2019.

[88] A. Ø. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72:31–54, 1997.

[89] J. Mosterín. *Conceptos y teorías en la ciencia*. Alianza Editorial, 1984.

[90] M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8:196–205, 2007.

[91] R.D.M. Page. Parasites, phylogeny and cospeciation. *International Journal for Parasitology*, 23:499–506, 1993.

[92] M. Petkovsek, H. Wilf, and D. Zeilberger. $A = B$. 1997.

[93] A. F. Poon. Phylodynamic inference with kernel ABC and its application to HIV epidemiology. *Molecular Biology and Evolution*, 32:2483–2495, 2015.

[94] A. Purvis. Using interspecies phylogenies to test macroevolutionary hypotheses. *New Uses for New Phylogenies*, pages 153–168, 1996.

[95] A. Purvis, S. Fritz, J. Rodríguez, P. Harvey, and R. Grenyer. The shape of mammalian phylogeny: Patterns, processes and scales. *Philosophical Transactions of The Royal Society B*, 366:2462–2477, 2011.

[96] J.-L. Le Quellec and B. Sergent. *Dictionnaire critique de mythologie*. CNRS, 2017.

[97] E. Rindal and A. V. Z. Brower. Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data. *Cladistics*, 27:331–334, 2011.

[98] J. S. Rogers. Central Moments and Probability Distributions of Colless's Coefficient of Tree Imbalance. *Evolution*, 48:2026–2036, 1994.

[99] J. S. Rogers. Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology*, 45:99–110, 1996.

[100] L. Rotger. New balance indices and metrics for phylogenetic trees. PhD Thesis, Universitat de les Illes Balears, 2020.

[101] G. Rubruquis. *Voyage dans l'Empire Mongol*. Translated from Latin by C. and R. Kappler, 1255.

[102] M. J. Sackin. "Good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972.

[103] E. Saulnier, S. Alizon, and O. Gascuel. Assessing the accuracy of Approximate Bayesian Computation approaches to infer epidemiological parameters from phylogenies. *bioRxiv*.

[104] H. M. Savage. The shape of evolution: systematic tree topology. *Biological Journal of the Linnean Society*, 20:225–244, 1983.

[105] J. Georg Schottel. *Ausführliche Arbeit Von der Teutschen HaubtSprache*. 1663.

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

[107] K.-T. Shao and R. R. Sokal. Tree balance. *Systematic Zoology*, 39(3):266–276, 1990.

[108] N. Sloane. Online Encyclopedia of Integer Sequences. https://oeis.org/, 1964.

[109] J. B. Slowinski. Probabilities of *n*-trees under two models: A demonstration that asymmetrical interior nodes are not improbable. *Systematic Zoology*, 39:89–94, 1990.

[110] E. Sober. *Reconstructing the Past: Parsimony, Evolution, and Inference*. MIT University Press, 1988.

[111] E. Sober. Experimental tests of phylogenetic inference methods. *Systematic Biology*, 42:85–89, 1993.

[112] R. R. Sokal. A phylogenetic analysis of the Caminalcules I: The data base. *Systematic Biology*, 32:159–184, 1983.

[113] E. Stam. Does imbalance in phylogenies reflect only bias? *Evolution*, 56:1292–1295, 2002.

[114] M. Stich and S. C. Manrubia. Topological properties of phylogenetic trees in evolutionary models. *The European Physical Journal*, 70:583–592, 2009.

[115] E. Suárez-Díaz and V. H. Anaya-Muñoz. History, objectivity, and the construction of molecular phylogenies. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 39:451–468, 2008.

[116] T. Takagi. A simple example of the continuous function without derivative. *Proceedings of the Physico-Mathematical Society of Japan*, 1:176–177, 1903.

[117] R. Tambs-Lyche. Une fonction continue sans dérivée. *Enseignement Mathématique*, 38:208–211, 1939.

[118] M. Thuillard, J.-L. Le Quellec, and J. d'Huy. Computational approaches to Myths analysis: Application to the Cosmic Hunt. *Nouvelle Mythologie Comparée*, 4:123–154, 2018.

[119] M. Thuillard, J.-L. Le Quellec, J. d'Huy, and Y. Berezkin. A large-scale study of world myths. *Trames Journal of the Humanities and Social Sciences*, 22:407–424, 2018.

[120] G. Verboom, F. Boucher, D. Ackerly, et al. Species selection regime and phylogenetic tree shape. *Systematic Biology*, 69:774–794, 2019.

[121] J. Villabona-Arenas, W. P. Hanage, and D. C. Tully. Phylogenetic interpretation during outbreaks requires caution. *Nature Microbiology*, 5:876–877, 2020.

[122] F. Villar. *Lenguas y pueblos indoeuropeos*. Ediciones ISTMO, 1971.

[123] E. M. Volz, S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. W. Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183:1421–1430, 2009.

[124] T. Warnow. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences of the United States of America*, 94(13):6585–6590, 1997.

[125] C. Wei, D. Gong, and Q. Wang. Chu-Vandermonde convolution and harmonic number identities Chu–Vaandermonde convolution and harmonic number identities. *Integral Transforms and Special Functions*, 24:324–330, 2013.

[126] T. Wu and K. Choi. On joint subtree distributions under two evolutionary models. *Theoretical Population Biology*, 108:13–23, 2015.

[127] G. U. Yule. A mathematical theory of evolution based on the conclusions of Dr J. C. Willis. *Philosophical Transactions of the Royal Society of London*, Series B 213:21–87, 1924.

[128] K.-K. Zhao, S. Landrein, R. L. Barrett, S. Sakaguchi, M. Maki, W.-X. Mu, T. Yang, Z.-X. Zhu, H. Liu, and H.-F. Wang. Phylogeographic analysis and genetic structure of an endemic Sino-Japanese Disjunctive Genus *Diabelia* (Caprifoliaceae). *Frontiers Plant Science*, 2019.