Enumeration Results on Leaf Labeled Trees

by

Virginia Perkins Johnson

Bachelor of Arts
Antioch College 1971
Master of Science in Math Education
NC A & T State University, 2001
Master of Arts in Mathematics
Wake Forest University, 2007

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosphy in

Mathematics

College of Arts and Sciences

University of South Carolina

2012

Accepted by:

Éva Czabarka Major Professor	Joshua Cooper Committee Member
Linyuan Lu Committee Member	Ognian Trifonov Committee Member
Csilla Farkas External Examiner	Lacy Ford, Vice Provost and Dean of Graduate Studies

Dedication

To Katharine, Patrick, Gregory, Aruno, and Simon: for the joy you bring into my life.

Acknowledgments

I would like to thank the community of people who have helped make this dissertation a reality, and the graduate experience a successful and enjoyable one. First and foremost, my deepest gratitude to Dr. Éva Czabarka, whose patience, encouragement, good humor, and guidance have made this possible. Her guidance, not only with the research and dissertation, but with all aspects of academic life has been invaluable. Thank you Éva. My thanks also to Dr. László A. Székely who always made me feel that he had total confidence in my ability to do the tasks he gave me. I owe much to Dr. Maria Girardi, for without her timely support and encouragement I would have never completed this venture. I am grateful to my dissertation committee (Dr. Joshua Cooper, Dr. Linyuan Lu, Dr. Ognian Trifonov, and Dr. Csilla Farkas) for their time and encouragement. Special thanks go to Dr. Linyuan Lu for providing opportunities for me speak at various math conferences, Dr. Joshua Cooper for his patience in answering questions about Sage and Dr. Francisco Blanco-Silva for helping me unravel the mysteries of Tikz. He is responsible for the programming needed to create Figure 2.1. I am indebted to Dr. Fredric Howard of Wake Forest University for his continuing guidance and advice over the years.

I am grateful to the other graduate students for those many hours of study sessions. Thank you *Dr.* Brett Barwick, *Dr.* Aaron Duttle, *Dr.* Samuel Gross, *Dr.* Andrew Vincent!

I also thank my family for their unwavering support. I am grateful to my parents, Dr. Ken and Margo Perkins for their unshakable belief in my abilities and for their encouragement which has always given me the confidence to step a little outside the

boundaries. Thank you to my sister, Dr. Susan Ashdown for the many hours of phone conversations that helped me keep everything in perspective, and my brother David Perkins, for his support. I am especially indebted to my children and grandchildren for their understanding and tolerance when the role of scholar overshadowed the role of mother or grandmother.

Abstract

In evolutionary biology it is common practice to represent the evolution of species, populations, and organisms with graphs called phylogenetic or species trees [C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, Oxford, (2003)]. Ideally these are rooted leaf-labeled trees where non-root internal vertices have degree at least three and each label is used once. Leaf-multi-labeled trees are a generalization of phylogenetic trees that are used in the study of gene versus species evolution and as the basis for phylogenetic network construction. Unlike a phylogenetic tree, in a leaf-multi-labeled tree it is possible to label more than one leaf by the same element of the underlying label set. In this thesis we first derive formulae for generating functions of leaf-multi-labeled trees and use these to derive recursive functions for counting such trees. In particular, we prove results which generalize previous theorems by Harding [*Advances in Appl. Probability* **3** (1971), 44-77] on so-called tree-shapes, and by Otter [*Ann. of Math.* (2) **49** (1948), 583-599] on relating the number of rooted and unrooted unlabeled trees. We provide some numbers for these trees using a program written using the open-source software program Sage.

Turning our attention to rooted phylogenetic or species trees we show the asypmtotic normality of phylogenetic trees with a fixed number of leaves where the internal number of vertices is allowed to vary. P.L. Erdős and L.A. Székely [*Adv. Appl. Math.* **10** (1989), 488–496] gave a bijection between rooted semi-labeled trees and set partitions. L.H. Harper's results [*Ann. Math. Stat.* **38** (1967), 410–414] on the asymptotic normality of the Stirling numbers of the second kind translate into asymptotic normality of rooted semi-labeled trees with given number of vertices, when the number of

internal vertices varies. The Erdős-Székely bijection specializes to a bijection between phylogenetic trees and set partitions with classes of size at least two. We consider modified Stirling numbers of the second kind that enumerate partitions of a fixed set into a given number of classes of size at least two, and obtain their asymptotic normality as the number of classes varies. The Erdős-Székely bijection translates this result into the asymptotic normality of the number of phylogenetic trees with given number of vertices, when the number of leaves varies. We also show the asymptotic normality of the number of phylogenetic trees with given number of leaves and varying number of internal vertices, which is more interesting to students of phylogeny. This is accomplished by showing the asymptotic normality of the number of partitions of $n + m$ elements into m classes of size at least two, when n is fixed and m varies, which with the Erdős-Székely bijection gives the result we want. The proofs are adaptations of the techniques of L.H. Harper [Ibid.].

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Summary

The enumeration of trees has a rich history with many applications. Kirchoff's Laws led to a natural interest in trees and in counting them [29]. Various formulae have been developed for counting leaf-labeled trees, many of them included in the monograph by Moon [34]. Cayley [7] formulated that the number of labeled trees on n vertices is n^{n-2}. Similar formulae have also been derived for the number of rooted binary leaf-labeled trees [24] (a *rooted tree* is a tree with one distinguished vertex called the *root*).

Harding [24] described ordinary generating functions for rooted, binary *tree-shapes* (i.e. isomorphism classes of unlabeled trees) with or without a specified number of internal vertices. Counting rooted unlabeled trees with the Pólya–Redfield method can be found, e.g., in [33]. Otter contributed a method for relating the counts of unlabeled trees to the counts of rooted unlabeled trees [36]. The functional equation for the ordinary generating function of the number of rooted unlabeled trees was already known (see Cayley [36]). Using methods due to Otter and Pólya (described in e.g. [23]), Dobson [11] also gave the generating function for unrooted, binary tree-shapes in terms of Harding's function. In addition, in [40, p.22], a formula involving the exponential generating function for rooted binary trees is given.

Studies in evolutionary biology have led to the enumeration of another type of trees. It is common practice to use *leaf-labeled* (or *phylogenetic*) trees to represent

the evolution of species, populations, organisms, and the like [40]. A leaf-labeled tree is a simple, connected graph with no cycles, and each of its leaves (i.e. vertices of degree 1) is labeled by precisely one element from a given label set. The set of labels corresponds to the set of species, populations or organisms under consideration. For *phylogenetic trees* the non-root, non-leaf vertices must have degree at least three. A simple example of such a tree is presented in Figure 1.1 (a).

Recently it has become apparent that it is useful to employ a more general type of tree when trying to understand, for example, gene evolution. In particular, due to processes such as gene (or genome) duplication or lateral gene transfer, trees can often arise in which more than one leaf is labeled by the same element of the label set. We will call such trees *leaf-multi-labeled trees*. Leaf-multi-labeled trees in which the root has degree at least two and internal vertices with degree at least three are known as MUL-trees [27]. An example of such a tree, and how it may arise, is presented in Figure 1.1 (b) and (c). Note that leaf-labeled trees form a subclass of leaf-multi-labeled trees. In addition their usefulness in the study of gene versus species evolution (e.g. [14, 39]), leaf-multi-labeled trees have been used to construct phylogenetic networks (e.g. [28, 27, 32]), and they naturally arise in biogeography (e.g.[19]).

As with leaf-labeled trees, for the purposes of applications it is important to develop a mathematical understanding of leaf-multi-labeled trees. Although at first sight leaf-multi-labeled trees do not seem very different from leaf-labeled trees, the theory of leaf-multi-labeled trees is quite rich in its own right, and several results on theoretical and algorithmic properties of such trees have recently appeared (cf. e.g. [14, 19, 20, 26]).

In this thesis, we shall derive formulae for ordinary generating functions for leaf-multi-labeled trees, and describe how they can be used to develop recursions for counting such trees. As we only consider ordinary generating functions we drop

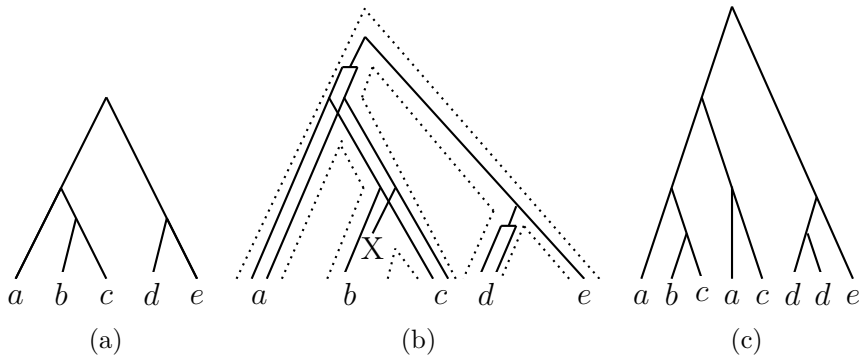


Figure 1.1: [a] A leaf-labeled "species tree" labeled by the set of species $\{a, b, c, d, e\}$. [b] A "gene tree" (in bold) representing the evolution of a gene, depicted within the species tree (in dotted) from [a] — we see two gene duplication events, and a gene loss (indicated with a cross). [c] The leaf-multi-labeled tree corresponding to the gene tree in [b], for which the label set is $\{a, b, c, d, e\}$.

the term "ordinary" from now on; the basics on generating functions that we shall use may be found in *Introductory Combinatorics* by R. Brualdi [2]. We then show the asymptotic normality of the number of phylogenic trees with a given number of vertices where the number of internal vertices varies using adaptations of the the method developed by Harper [25]. The same approach leads to the asymptotic normality of phylogenetic trees with a fixed number of leaves where the number of internal vertices is allowed to vary.

We begin in Chapter 2 with a formula (Theorem 2.1) involving the generating function for the number of rooted binary leaf-multi-labeled trees, and use this to develop a recursion for counting such trees (see equation (2.2)). This formula is a straightforward extension of Harding's [24] formula for generating functions of tree-shapes (see also equation (2.1)), since the class of leaf-multi-labeled trees includes the class of tree shapes. (A tree-shape can be considered as a leaf-multi-labeled tree in which only one label is used to label all leaves.) In this chapter we also develop generating functions for rooted gene trees and for rooted leaf-multi-labeled trees. In Chapter 3, we will present a theorem (Theorem 3.3), which will allow us to relate generating functions of rooted binary leaf-multi-trees to unrooted versions of these trees.

Otter [36] gave a formula for unrooted trees that provided a relationship between counts for rooted trees and counts for unrooted trees. F. Harary [22] generalized Otter's theorem to include unlabeled graphs. Unfortunately the proof he gave seems to contain a flaw. However, Harary's theorem can easily be proved for semi-labeled graphs (Theorem 3.3), as the introduction of labels allows us to use Harary's original approach to prove this extension. This, in turn, gives us an extension of Otter's theorem for semi-multi-labeled trees, which allows us to use our generating functions for rooted trees to find generating functions of unrooted trees. In Chapter 4 we consider unrooted trees, giving formulae for generating functions in the unrooted binary trees, unrooted gene trees and unrooted leaf-multi-labeled trees.

Turning our attention to the asymptotic normality and phylogenetic trees, we lay the ground work in Chapter 5. We use a bijection developed by P.L. Erdős and L.A. Székely [13] to relate semi-labeled trees with a fixed number of vertices and a varying number of leaves to the Stirling numbers of the second kind. We also provide an overview of the method used by Harper [25] to show the asymptotic normality of the Stirling numbers of the second kind. In Chapter 6 we show the asymptotic normality of a variant of the Stirling numbers and hence the asymptotic normality of the phylogenetic trees mentioned. These results are extended to phylogenetic trees in which the number of leaves is fixed and the number of internal vertices is allowed to vary.

We also present three programs in Sage (open-source programming language) designed to use the recursive functions for the leaf-multi-labeled trees to calculate the numbers of the various categories of these trees. This code can be found in Appendix 1. In Appendix 2 and 3 we provide the Maple programs used in our calculations.

1.2 Basic definitions, statements, and notation

For the general terminology describing graphs the reader is referred to *Graphical Enumeration*, by Harary [22].

By *graph*, we will mean simple finite graphs, i.e. the vertex set is finite and there are no loops or multiple edges. Formally:

Definition 1.1. A graph $G = (V_G, E_G)$ has a finite vertex set V_G and and edge set E_G is a set of 2-subsets of $V(G)$.

We will use the notation xy for an edge $\{x, y\} \in E_G$; thus, $xy = yx$ when we talk about edges of a graph.

Definition 1.2. A *trivial graph* consists of one vertex and no edges.

Definition 1.3. A *labeled graph* is a graph in which every vertex is labeled from a set X and each element of the label set X is used at most once. If G is a labeled graph there exists an injective function $\alpha_G : V_G \to X$.

Definition 1.4. A *multi-labeled graph* is a graph in which every vertex is labeled, but elements of the label set may be used for more than one vertex. So we have a function $\alpha_G : V_G \to X$.

The family of multi-labeled graphs includes the family of labeled graphs.

Definition 1.5. A *semi-labeled graph* is a graph in which a subset of the vertices are labeled and each element of the label set is used at most once. Given a graph G, a fixed subset L_G of the vertex set V_G, and an injective function $\alpha_G : L_G \to X$, G is a semi-labeled graph. The set L_G is the *set of labeled vertices.*

Again, the family of semi-labeled graphs contains the family of labeled graphs.

Definition 1.6. A *semi-multi-labeled graph* is a graph in which a subset of the vertices are labeled. Labels may be used more than once. Given such a graph G, if L_G is the labeled (fixed) subset of the vertex set V_G, there exists a function $\alpha_G : L_G \to X$.

Unless otherwise specified, label set of all graphs in this dissertation will be $[k] = \{1, 2, \ldots, k\}$.

Definition 1.7. If G is a semi-multi-labeled graph with labeling $\alpha_G : L_G \to [k]$ then we define $\alpha^\star : V_C \to [k] \cup \{0\}$ as

$$\alpha^\star_G(v) = \begin{cases} \alpha(v) & \text{if } v \in L_G \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\alpha^\star_G\big|_{L_G} = \alpha_G$ and $\alpha^\star_G\big|_{V_G \setminus L_G} \equiv 0$. Notice that semi-multi-labeled graphs and multi-labeled graphs are not fundamentally different. If G is a semi-multi-labeled graph, with labeling given by the function $\alpha_G : L_G \to [k]$, then we may view it as a multi-labeled graph using $\alpha^\star : V_G \to [k] \cup \{0\}$. Thus we can now consider unlabeled graphs, semi-labeled graphs, labeled graphs and multi-labeled graphs as subfamilies of the family of semi-multi-labeled graphs. The label 0 is a special label that can be reused even if we require the other labels to be used only once, and the original labeling α can be reconstructed from $\alpha^\star$ with $L_G = V_G \setminus (\alpha^\star)^{-1}(0)$. Consequently, any definition referring to semi-multi-labeled graphs using the labeling function $\alpha^\star_G$ will refer to these subclasses as well.

Definition 1.8. A special vertex in the graph G is a single vertex $\rho_G \in V_G$. Depending on our goals, we will call this special vertex a *root* or a *marked vertex*, and the graph a *rooted graph* or *marked graph*. Note that from now on we will use the notation ρ_G exclusively to indicate the special vertex.

Using Definition 1.8, the rooted and marked graphs are the same. We will however still use these separate terms. The reason for the distinction is that certain families

of trees consist of rooted trees where the root has stated properties. When we wish to use a special vertex that may not have these stated properties, we will refer to a marked graph instead of a rooted graph to emphasize the distinction.

Definition 1.9. A *graph isomorphism* ϕ between two semi-multi-labeled graphs G and H is a bijection between vertex sets that has the following properties

1. Both ϕ and ϕ^{-1} are adjacency preserving, hence $v_iv_j \in E_G \Leftrightarrow \phi(v_i)\phi(v_j) \in E_H$.
2. ϕ is label preserving: for every $v \in V_G$ we have $\alpha^\star_H(\phi(v)) = \alpha^\star_G(v)$.
3. ϕ preserves the special vertex; either both G and H have a special vertex and $\phi(\rho_G) = \rho_H$, or neither of them has a special vertex .

Definition 1.10. Two graphs G and H are considered to be identical (the same) if there exists graph isomorphism ϕ between them.

Definition 1.11. A *graph automorphism* is a graph isomorphism between a graph and itself.

The set of graph automorphisms is a group with the composition being the group operation, the identity function is the identity, and inverse being the usual inverse of a function.

Definition 1.12. Given a graph G, two vertices, $v_1, v_2 \in V_G$ are *equivalent* if there is an automorphism, ϕ of G such that $\phi(v_1) = v_2$.

It is a routine exercise to prove that the relationship in Definition 1.12 is an equivalence relation. This motivates the following definition

Notation 1.13. *The number of equivalence classes under the relation in Definition 1.12 is denoted by* p_G.

Definition 1.14. A *cut-vertex* of a non-trivial graph is a vertex of the graph whose removal increases the number of components of the graph.

Definition 1.15. A *non-separable graph* is a connected non-trivial graph which does not have a cut-vertex.

Definition 1.16. A *block* of a graph is a maximal non-separable subgraph of the graph.

Definition 1.17. Given a non-trivial graph G with blocks $B_1, \ldots, B_k$ and cut-vertices $v_1, v_2, \ldots, v_m$, the *block-cutpoint graph*, $\mathfrak{b}(G)$ is a bipartite graph in which one partite set consists of the cut-vertices of G and the other set contains a vertex b_i for each block B_i of G. We include $v_j b_i$ as an edge of $\mathfrak{b}(G)$ if and only if $v_j \in B_i$.

The proof of the following can be found in standard graph theory books, i.e. [9] We will use this fact later.

Claim 1.18. *If G is a connected nontrivial graph, then $\mathfrak{b}(G)$ is a tree whose leaves are precisely the vertices corresponding to the blocks of G with exactly one cut-vertex. Consequently, G is either non-separable (is a single block) or it has at least one block with precisely one cut-vertex, and the removal of any blocks that have one cut-vertex does not disconnect G.*

Definition 1.19. Two blocks B_1 and B_2 of G are *equivalent* if there exists an automorphism ϕ of G such that $\phi(V(B_1)) = V(B_2)$.

Definition 1.20. A *tree* is an acyclic connected graph. If the tree has only one vertex, it will be referred to as a *trivial tree*.

Note that many authors refer to (unlabeled) trees as *tree shapes*, emphasizing the fact that they consider two such trees different only if they are not isomorphic.

Definition 1.21. A leaf of a non-trivial tree is a vertex of degree 1. Unless stated otherwise, in this dissertation, the vertex of the trivial tree will also be considered a leaf.

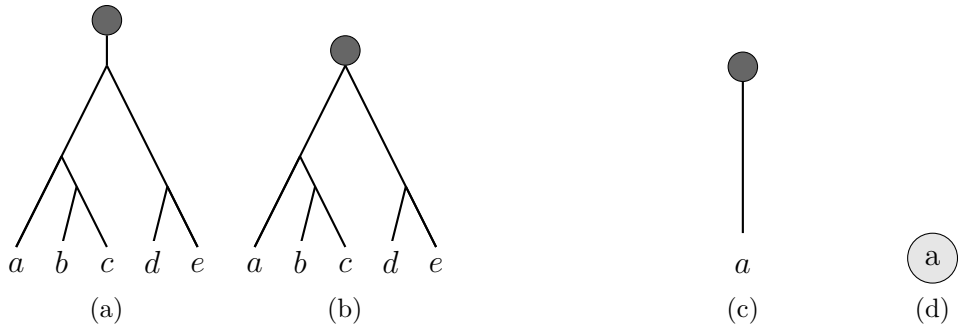


Figure 1.2: (a) A leaf-labeled "species tree" labeled by the set of species $\{a, b, c, d, e\}$ where the root has degree one. (b) The same information depicted using a tree where the root has degree two. (c) A tree with one leaf and root with degree one. (d) The same information depicted by a singleton vertex which is considered both a leaf and a root and is labeled.

Definition 1.22. A *leaf-labeled tree* is a semi-labeled tree in which the set of labeled vertices is the set of non-root vertices of degree one.

Definition 1.23. *Leaf-multi-labeled trees* are trees in which the set of labeled vertices is the set of non-root vertices of degree one. The labels are not necessarily unique and may be used for more than one leaf.

The following definition is the motivation for introducing the terminology of marked graphs earlier, as this definition is standard for rooted binary trees. We will make use of binary trees whose special vertex is not a root in the sense of the standard definition and we will refer to these trees as marked binary trees.

Definition 1.24. A *rooted binary tree* is either a trivial tree (where the root is the single vertex) or a tree in which the root has degree two and all non-root, non-leaf vertices have degree three.

Since phylogenetic trees represent the evolutionary relationships between species with internal non-root vertices corresponding to speciation events, such internal non-root vertices must have an edge that leads towards the root, and and at least two edges corresponding to the new species that were created by the speciation event.

Therefore such vertices should have degree at least three, and the root would correspond to the common ancestor of all the species represented in the phylogenetic tree. Non-root leaves corresponding to existing species are labeled with the name of the species. What are the properties of the root of such a tree? As the edges represent the time-period when the corresponding species existed, having a root of degree one would mean that we draw the edge corresponding to this time period of the common ancestor, and having a root of degree greater than one would mean that we do not draw this edge. Clearly, there is a one-to-one correspondence between these representations (removing the degree one root and rooting the resulting tree at the neighbor of the original root). Therefore we can use species trees where the root has degree one, or species trees where the degree of the root is at least two (see Figure 1.2). As these two depictions are equivalent, the choice one of these conventions is made according to convenience. For the techniques used in this dissertation it will be more convenient to require that the root does not have degree one. This implies that for trees which have only one leaf, that vertex will be considered both a leaf and a root, and will be labeled.

Gene trees or MUL-trees, as they are also referred to in the literature, represent the evolutionary relationships of copies of the same gene across several species, and due to processes such as duplication or deletion of genetic material, the topology of a gene tree may look very different from its corresponding species tree. See Figure 1.1. As the leaves still are labeled with the name of the species the corresponding gene sample came from, any label that appeared in the species tree may appear several times or not at all in the gene tree. The same reasoning regarding the root applies as on phylogenetic trees.

Since it is not reasonable to assume that during a speciation event more than two new species is created, ideally a phylogenetic tree is a rooted leaf-labeled binary tree. However, these trees are created from data, which may not be sufficient to completely

resolve the tree, and the placement of the root is difficult. Thus, these trees may or may not be binary or rooted. These facts motivate the following definitions.

Definition 1.25. *MUL-trees* or *gene trees* are leaf-multi-labeled trees that may be rooted or unrooted. Every leaf is labeled whether it is a root or not. Non-root, non-leaf vertices have degree at least three. The root, if exists, does not have degree one.

Definition 1.26. *Phylogenetic trees* are MUL-trees where labels are not reused. They are leaf-labeled trees that may be rooted or unrooted. Every leaf is labeled whether it is a root or not. Non-root, non-leaf vertices have degree at least three. The root, if exists, does not have degree one.

We reiterate one of our earlier remarks as these definitions are the main reason to introduce the terminology for marked trees. In rooted binary trees the root must have degree two, and in non-trivial rooted phylogenetic or MUL-trees, the root must have degree at least two and is unlabeled. While a marked tree, just as rooted tree, is a tree with a special vertex identified, the terminology "marked gene tree", "marked phylogenetic tree" and "marked binary tree" will refer to the cases where the underlying tree is an unrooted version of the tree class (i.e. unrooted gene tree, phylogenetic tree or binary tree) and the marked vertex is any vertex of this tree (either a labeled leaf or an unlabeled vertex of degree at least three).

Finally we will return to the graph automorphisms and the idea of equivalence, and define two more concepts for trees.

Definition 1.27. For a semi-multi-labeled tree T, two edges, $e_1, e_2 \in E_T$ are *equivalent* if there exists an automorphism ϕ of T that maps the end vertices of e_1 to the end vertices of e_2.

Notation 1.28. *The number of equivalence classes on the set of edges of a tree T defined by the equivalence relation in Definition 1.27 is denoted by q_T*

Definition 1.29. An edge e of a (semil-multi-labeled) tree T is said to be *symmetric* if there exists a graph automorphism ϕ that exchanges the endpoints of the edge.

As the removal of a symmetry edge must result in two trees that have the same number of vertices, it is clear that there can be at most one symmetry edge for any tree.

Notation 1.30. *The number of symmetry edges of a tree T is denoted by s_T. By the preceding remark, $s_T \in \{0, 1\}$.*

1.3 Generating functions

In this section we will define ordinary and exponential generating functions and state without proof some basic results about them. The interested reader should refer to one of the standard books, such as *Generatingfunctionology* [44] for more details.

As usual, for k-dimensional vectors $\vec{x} = (x_1, \ldots, x_k)$ and $\vec{y} = (y_1, \ldots, y_k)$ over an additive semigroup $\vec{x} + \vec{y}$ will denote the vector $(x_1 + y_1, \ldots, x_k + y_k)$.

Definition 1.31. Let $F(x_1, \ldots, x_k)$ be a function on k variables and $n \in \mathbb{N}$ where $\mathbb{N}$ is the set of nonnegative integers. . For shortness, we denote $F(x_1^n, \ldots, x_k^n)$ by $F(\cdot^n)$, and $F(\cdot^1)$ by $F(\cdot)$.

Definition 1.32. Let $\mathcal{A}$ be a set and and $k \in \mathbb{Z}^+$. The function β is a *k-type* on $\mathcal{A}$, if β is a function from $\mathcal{A}$ to $\mathbb{N}^k$. A *type* is a k-type for some k.

Definition 1.33. Let $\mathcal{A}$ be a set equipped with a k-type β and $a \in \mathcal{A}$. The *term* of a with respect to β (or the term of x, for short, if the choice of β is clear) on variables $x_1, \ldots, x_n$ defined as

$$\mathrm{term}_\beta(a) = \prod_{j=1}^{n} x_j^{n_j},$$

where $\beta(a) = (n_1, \ldots, n_k)$. When β is clear from the text, we will use the notation $\mathrm{term}(a)$.

At this point we are ready to define ordinary generating functions.

Definition 1.34. Let $\mathcal{B}$ be a set equipped with a k-type β. The ordinary generating function of $\mathcal{B}$ with respect to the type β on variables $x_1, \ldots, x_k$ is

$$B(x_1, x_2, \ldots, x_k) = \sum_{b \in \mathcal{B}} \mathrm{term}_\beta(b) = \sum_{(n_1, n_2, \ldots, n_k) \in \mathbb{N}^k} a_{n_1, n_2, \ldots, n_k} \prod_{j=1}^{k} x_j^{n_j},$$

where $a_{n_1,\ldots,n_k} = \Big|\{b \in B : \mathrm{type}(b) = (n_1, \ldots, n_k)\}\Big|$. We will also refer to $B(x_1, \ldots, x_k)$ as the ordinary generating function for the counts $a_{n_1,\ldots,n_k}$.

The following claims are well known, and also easily follow from the definitions. Their proofs will be omitted.

Claim 1.35. *Let $\mathcal{A}_1, \mathcal{A}_1$ be disjoint sets and let β_i be a k-type on $\mathcal{A}_i$ for $i \in \{1, 2\}$. For $\mathcal{B} = \mathcal{A}_1 \cup \mathcal{A}_2$ define the k-type β by $\beta_1 \cup \beta_2$, i.e. $\beta(a) = \beta_1(a)$ if $a \in \mathcal{A}_1$ and $\beta(a) = \beta_2(a)$ otherwise. Denote the ordinary generating function of $\mathcal{A}_i$ by $A_i(\cdot)$ and the ordinary generating function of $\mathcal{B}$ by $B(\cdot)$. Then $B(\cdot) = A_1(\cdot) + A_2(\cdot)$.*

Claim 1.36. *Let $\mathcal{A}_1, \mathcal{A}_2$ be sets and let β_i be a k-type on $\mathcal{A}_i$ for $i \in \{1, 2\}$. For $\mathcal{B} = \mathcal{A}_1 \times \mathcal{A}_2$ define the k-type β by $\beta(a_1, a_2) = \beta_1(a_1) + \beta_2(a_2)$. Denote the ordinary generating function of $\mathcal{A}_i$ by $A_i(\cdot)$ and the ordinary generating function of $\mathcal{B}$ by $B(\cdot)$. Then $B(\cdot) = A_1(\cdot) \cdot A_2(\cdot)$.*

Tğhe first part of this last claim easily follows by induction from the previous claim.

Claim 1.37. *Let $\mathcal{A}$ be a set equipped with a k-type γ, and $n \in \mathbb{Z}^+$. Let $\mathcal{B}_1 = \prod_{i=1}^{n} \mathcal{A}$, $\mathcal{B}_2 \subseteq \mathcal{B}_1$ by $(a_1, \ldots, a_n) \in \mathcal{B}_2$ iff $a_1 = \cdot = a_n$. Define the k-type β on $\mathcal{B}_1$ (and consequently on $\mathcal{B}_2$ by $\beta(a_1, \ldots, a_n) = \prod_{j=1}^{n} \gamma(a_j)$. Denote the ordinary generating function of $\mathcal{A}$ by $A(\cdot)$ and the ordinary generating function of $\mathcal{B}_i$ by $B_i(\cdot)$. Then $B_1(\cdot) = A^n(\cdot)$ and $B_2(\cdot) = A(\cdot^n)$.*

In the rest of the thesis, we will refer to ordinary generating functions simply as generating functions. We also use exponential generating functions with one variable, so we define those here.

Definition 1.38. Let $\mathcal{B} = \cup\mathcal{B}_n$, where $\mathcal{B}_n$ is a set of structures defined on $[n]$, and $b_n = |\mathcal{B}_n|$. The exponential generating function (EGF) $B(t)$ of $\mathcal{B}$ (or alternatively, of the counts b_n) is

$$B(t) = \sum_{n\in\mathbb{N}} b_n \frac{t^n}{n!}.$$

The following claim is immediate from the definition

Claim 1.39. *Let $B(t)$ be the exponential generating function of the counts b_n. Then $\frac{d}{dt}(B(t))$ is the exponential generating function of the counts $c_n = b_{n+1}$.*

The following is the Product Rule of Exponential Generating Functions:

Claim 1.40. *Let $\mathcal{A}$ and $\mathcal{B}$ be two classes of objects with exponential generating functions $A(t)$ and $B(t)$. Let $\mathcal{C} = \cup\mathcal{C}_n$ be the set of objects, where $\mathcal{C}_n$ is the set of objects on $[n]$ that consist of all pairs of objects that can be obtained by taking an ordered pair $(A, [n] \setminus A)$ of possibly empty subsets of $[n]$, and inserting an object from $\mathcal{A}_{|A|}$ on A and an object from $\mathcal{B}_{|[n]\setminus A|}$ on $[n] \setminus A$. The exponential generating function $C(t)$ of $\mathcal{C}$ is $A(t) \cdot B(t)$.*

Chapter 2

Rooted leaf-multi-labeled trees

2.1 Rooted binary trees

We begin by considering the generating function for rooted, binary leaf-multi-labeled trees (see Definition 1.23). Let t_n denote the number of rooted unlabeled binary tree shapes with n leaves. (This is equivalent to the set of rooted leaf-multi-labeled binary tree shapes in which all the leaves are labeled with one label.) Harding [24] observed (see also Wedderburn [43]) that the ordinary generating function for $\{t_n\}_{n=0}^{\infty}$,

$$T(z) = \sum_{n=0}^{\infty} t_n z^n$$

satisfies the equation

$$T(z) = z + \frac{1}{2}T^2(z) + \frac{1}{2}T(z^2). \tag{2.1}$$

This can be argued as follows: It is clear that $t_0 = 0$ and $t_1 = 1$. For $n \geq 2$, since the root has degree 2, the tree is composed of two subtrees, the roots of which are neighbors of the original root. Since the new roots have degree two, they are rooted binary trees. $T^2(z)$ counts the subtree pairs (T_1, T_2). When $T_1 \neq T_2$ the pair is counted twice. When $T_1 = T_2$ the pair is counted once. The trees with two isomorphic subtrees are counted by $T(z^2)$. Putting this information together yields the formula.

The same argument can be used to find a formula for the ordinary generating function for rooted, binary leaf-multi-labeled trees using the label set $[k]$:

$$R(x_1, \ldots, x_k) = \sum_{n=0}^{\infty} r_{n_1,\ldots,n_k} x_1^{n_1} \cdots x_k^{n_k},$$

where $r_{n_1,\ldots,n_k}$ is the number of rooted, binary leaf-multi-labeled trees with $\sum_{i=1}^{k} n_i$ leaves in which each label $j \in [k]$ is used on n_j leaves. Note that n_j may be 0. We have:

Theorem 2.1.

$$R(x_1,\ldots,x_k) = (x_1 + \cdots + x_k) + \frac{1}{2}R^2(x_1,\ldots,x_k) + \frac{1}{2}R(x_1^2,\ldots,x_k^2).$$

This theorem can be used in a straight-forward fashion to find obtain a recursion for calculating the numbers $r_{n_1,\ldots,n_k}$ as follows. Let

$$h_{n_1,\ldots,n_k} = \sum_{m_1=0}^{n_1}\sum_{m_2=0}^{n_2}\cdots\sum_{m_i=0}^{n_i}\cdots\sum_{m_k=0}^{n_k} r_{m_1,\ldots,m_k} r_{n_1-m_1,\ldots,n_k-m_k}.$$

Thus,

$$R^2(x_1\ldots,x_k) = \sum_{m_1\ldots,m_k} h_{m_1,\ldots,m_k}\prod_{j=1}^{k} x_j^{m_j}.$$

Then

$$r_{n_1,\ldots,n_k} = \begin{cases} 0 & \text{if } \sum_{i=1}^{k} n_i = 0,; \\ 1 & \text{if } \sum_{i=1}^{k} n_i = 1,; \\ \frac{1}{2}\left(r_{n_1/2,\ldots,n_k/2} + h_{n_1,\ldots,n_k}\right) & \text{if all } n_i \text{ are even;} \\ & \text{and } \sum_{j=1}^{k} n_i \geq 2, ; \\ \frac{1}{2}h_{n_1,\ldots,n_k} & \text{else.} \end{cases} \tag{2.2}$$

Two observations are of interest. Suppose we let $r_{n;k}$ denote the number of rooted binary leaf-multi-labeled trees with n leaves on the set $[k]$, and let $R_k(z) = \sum_n r_{n;k} z^k$ be the associated generating function. If we let $x_1 = x_2 = \cdots = x_k = z$, then we obtain $R(z,z,\ldots,z) = \sum_n \sum_{\substack{(n_1,\ldots,n_k) \\ n_1+\cdots+n_k=n}} r_{n_1,\ldots,n_k} z^n = \sum_n r_{n;k} z^n = R_k(z)$. By Theorem 2.1 we now have

$$R_k(z) = kz + \frac{1}{2}R_k^2(z) + \frac{1}{2}R_k(z^2),$$

The case $k = 1$ yields (2.1), as expected. Note that this formula also yields the recursion:

$$r_{n;k} = \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ \frac{1}{2}\sum_{j=1}^{n-1} r_{j;k} r_{n-j;k} & \text{if } n > 1 \text{ odd}, \\ \frac{1}{2}\left(r_{n/2;k} + \sum_{j=1}^{n-1} r_{j;k} r_{n-j;k}\right) & \text{else.} \end{cases} \tag{2.3}$$

Secondly, we consider the case where we only count those trees which use every label in $[k]$ (i.e. the numbers $r_{n_1,\ldots,n_k}$ where each n_i is positive). Let $v_{n,k}$ denote the number of rooted binary leaf-multi-labeled trees with label set $[k]$ that use each label at least once and let $V_k(z)$ be the corresponding generating function. Then the inclusion-exclusion principle yields

$$v_{n;k} = \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} r_{n;k-j}. \tag{2.4}$$

Consequently we have

$$V_k(z) = \sum_{n=0}^{\infty} v_{n;k} z^n = \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} R_{k-j}(z).$$

We include some values of $r_{n;k}$ in Table 2.1 and some values for $v_{n,k}$ in Table 2.2. The program used to calculate these numbers is in Appendix A.1

Table 2.1: The first few values of $r_{n;k}$, the number of rooted binary MUL-trees with n leaves on the label set $[k]$, obtained using recursion equation (2.3).

$n\backslash k$	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	1	3	6	10	15	21	28
3	1	6	18	40	75	126	196
4	2	18	75	215	495	987	1778
5	3	54	333	1260	3600	8568	17934
6	6	183	1620	8010	28275	80136	194628
7	11	636	8202	53240	232500	785106	2213036
8	23	2316	43188	366680	1979385	7960638	26037431
9	46	8610	232947	2590420	17287050	82804806	314260765
10	98	32763	1282824	18674660	154041450	878729418	3869500208

Table 2.2: The first few values of $v_{n;k}$, the number of rooted binary leaf-multi-labeled trees with n leaves on the label set $[k]$, obtained using equation (2.4).

$n\backslash k$	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	1	1	0	0	0	0	0
3	1	4	3	0	0	0	0
4	2	14	27	15	0	0	0
5	3	48	180	240	105	0	0
6	6	171	1089	2604	2625	945	0
7	11	614	6333	24180	42075	34020	10395
8	23	2270	36309	207732	554820	755370	509355
9	46	8518	207255	1710108	6578550	13408740	14963130
10	98	32567	1184829	13739550	73169250	209434995	343863135

2.2 Rooted gene trees

In this next section, we will consider rooted leaf-multi-labeled trees.

Let $\mathcal{R}_k$ denote the set of isomorphism classes of rooted leaf-multi-labeled trees on label set $[k]$. $\mathcal{R}_k$ includes the single vertex trees and the trees where the degree of every non-root, non-leaf vertex is at least three, and the degree of the root is least two. Note that for a binary tree with $n \geq 2$ leaves, the number of internal vertices can be given as a function of n ($(n-1)$ if rooted and $(n-2)$ if unrooted);

however, for non-binary trees this is not the case. In particular, an element of $\mathcal{R}_k$ with $n \geq 2$ leaves can have any number of internal vertices between 1 and $n-1$. It is therefore useful to keep track of the number of internal, unlabeled vertices. For this reason, we define the $(k+1)$-type β on $\mathcal{R}_k$ by $\beta(T) = (u, n_1, \ldots, n_k)$ if the tree T has u unlabeled vertices and n_i leaves labeled with i. Let $a_{u,n_1,\ldots,n_k}$ to be the number of trees in $\mathcal{R}_k$ with u unlabeled nodes and n_j nodes with label j, and $A(z; x_1, \ldots, x_k) = \sum a_{u,n_1,\ldots,n_k} z^u x_1^{n_1} \ldots x_k^{n_k}$ be the corresponding generating function.

We can now give a Cayley-type equality for $A(\cdot)$. Consistent with our earlier notation, for any $T \in \mathcal{R}_k$ let $\ell_j(T)$ be the number of vertices that have label j, by $un(T)$ the number of unlabeled vertices, and let

$$\text{term}(T) \quad = \quad z^{un(T)} \prod_{j=1}^{k} x_j^{\ell_j(T)}.$$

Theorem 2.2.

$$A(z; x_1, \ldots, x_k) = \frac{(x_1 + \cdots + x_k - z) + z \cdot \text{Exp}\Big(\sum_{n=1}^{\infty} \frac{1}{n} A(z^n; x_1^n, \ldots, x_k^n) \Big)}{z+1}$$

Proof. There is precisely one tree in $\mathcal{R}_k$ that is a single vertex and is labeled by j. Thus, $A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)$ counts non-trivial trees in $\mathcal{R}_k$. If we take a non-trivial tree in $\mathcal{R}_k$ the root has degree at least two. Remove the unlabeled root of this tree and root each tree of the resulting forest at the neighbors of the old root. Since the neighbors at the old root are either leaves or vertices of degree at least three, the roots of this forest are either labeled vertices of a singleton or unlabeled vertices of degree at least two. Therefore all of the trees in the resulting forest are trees in $\mathcal{R}_k$. Also, any forest of trees from $\mathcal{R}_k$ with at least two components cant be obtained this way from a tree of $\mathcal{R}_k$. Let $H_1(\cdot)$ count the rooted finite forests that have at least two components. Note that $H_1(\cdot)$ counts the rooted finite forests that are not just a single tree (i.e. disjoint unions of at least two elements in $\mathcal{R}_k$). Thus the trees in $\mathcal{R}_k$

having at least two vertices are in one-to-one correspondence with the rooted forests that have at least two components. Subtracting the number of singleton trees from $A(z; x_1, \ldots, x_k)$ and dividing by z to reduce the number of unlabeled vertices by one (removal of the root), we have

$$H_1(\cdot) = \frac{A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)}{z}.$$

Let $H_2(\cdot)$ be the number of all rooted finite nonempty forests. Since $A(x_1, \ldots, x_k)$ counts the rooted forests with precisely one component,

$$H_2(\cdot) = A(z; x_1, \ldots, x_k) + H_1(\cdot) = \frac{(1+z)A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)}{z}.$$

If $H_3(\cdot)$ is the number of all rooted finite forests of trees, including the empty forest, then

$$H_3(\cdot) = H_2(\cdot) + 1 = \frac{(1+z)A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k - z)}{z}.$$

Any rooted forest (including the empty one) is determined by the number of copies of any tree in $\mathcal{R}_k$ that appears within it. Therefore $H_2(\cdot)$ is an infinite sum where each term is of the following form: Let D be a (possibly empty) finite subset of $\mathcal{R}_k$, for each $T \in D$ let m_T be a positive integer. Then the product $\prod_{T \in D} (\mathrm{term}(T))^{m_T}$ is the term corresponding to the forest where each $T \in D$ appears precisely m_T times. Moreover, $H_3(\cdot)$ is the sum of all terms of this type. Therefore

$$\begin{aligned}
H_3(\cdot) &= \left(\prod_{T \in \mathcal{R}_k} \left(\sum_{j=0}^{\infty} \mathrm{term}(T)^j \right) \right) = \left(\prod_{T \in \mathcal{R}_k} \Big(1 - \mathrm{term}(T) \Big)^{-1} \right) \\
&= \prod_{(u; n_1, \ldots, n_k)} \left(\prod_{\substack{T \in \mathcal{R}_k \\ \beta(T) = (u; n_1, \ldots, n_k)}} \Big(1 - \mathrm{term}(T) \Big)^{-1} \right) \\
&= \prod_{(u; n_1, \ldots, n_k)} \left(\left(\Big(1 - \mathrm{term}(T) \Big)^{-1} \right)^{\left| \{T \in \mathcal{R}_k : \beta(T) = (u; n_1, \ldots, n_k)\} \right|} \right) \\
&= \prod_{(u; n_1, \ldots, n_k)} \left(\Big(1 - z^u x_1^{n_1} \cdots x_k^{n_k} \Big)^{-a_{u; n_1, \ldots, n_k}} \right).
\end{aligned}$$

This follows from collecting the terms corresponding to the trees that have the same form for term(T) and the definition of the numbers $a_{u;n_1,\ldots,n_k}$. This implies that

$$\begin{aligned}
\log(H_3(\cdot)) &= -\sum_{(u;n_1,\ldots,n_k)} a_{n_1,\ldots,n_k} \log(1 - z^u x_1^{n_1} \cdots x_k^{n_k}) \\
&= \sum_{(u;n_1,\ldots,n_k)} a_{u;n_1,\ldots,n_k} \sum_{n=1}^{\infty} \frac{\left(z^u x_1^{n_1} \cdot x_k^{n_k}\right)^n}{n} \\
&= \sum_{n=1}^{\infty} \frac{1}{n} \sum_{(u;n_1,\ldots,n_k)} a_{n_1,\ldots,n_k} \Big((z^n)^u (x_1^n)^{n_1} \cdot (x_k^n)^{n_k} \Big) \\
&= \sum_{n=1}^{\infty} \frac{1}{n} A(z^n; x_1^n, \ldots, x_k^n),
\end{aligned}$$

from which the statement of the theorem follows. □

As an immediate corollary, we can now give a formula involving the generating function for the number of trees in $\mathcal{R}_k$ where the label j is used precisely n_j times: Let $g_{n_1,\ldots,n_k}$ be the number of such trees in $\mathcal{R}_k$, with corresponding generating function

$$G(x_1, \ldots, x_k) = \sum_{(n_1,\ldots,n_k)} g_{n_1,\ldots,n_k} \prod_{j=1}^{k} x_j^{n_j},$$

Then $g_{n_1,\ldots,n_k} = \sum_u a_{u;n_1,\ldots,u_k}$ and we have

$$A(1; x_1, \ldots, n_k) = \sum_{(n_1,\ldots,n_k)} \left(\left(\sum_u a_{u;n_1,\ldots,n_k} \cdot 1^u \right) \prod_{j=1}^{k} x_j^{n_j} \right) = G(x_1, \ldots, x_k),$$

from which we obtain the following.

Corollary 2.1.

$$G(x_1, \ldots, x_k) = \frac{1}{2} \left((x_1 + \cdots + x_k - 1) + \mathrm{Exp}\Big(\sum_{n=1}^{\infty} \frac{1}{n} G(x_1^n, \ldots, x_k^n) \Big) \right).$$

We use this formula to derive a recursion for the number $g_{n;k}$ of trees in $\mathcal{R}_k$ on n leaves using $[k]$ as label set. Clearly $G_k(x) = \sum_n g_{n;k} x^k = G(x, \ldots, x)$. Let

$$G_k^\star(x) = \sum_{n \geq 1} \frac{1}{n} G_k(x^n) = \sum_{n \geq 0} g_{n;k}^\star x^n.$$

Then $g^\star_{0;k} = g_{0;k} = 0$. We have

$$\begin{aligned}\sum_{m\geq 1} g^\star_{m;k} x^m &= \sum_{n\geq 1} \frac{1}{n} G_k(x^n) = \sum_{n\geq 1} \frac{1}{n} \sum_{j\geq 1} g_{j;k} x^{nj} \\ &= \sum_{n\geq 1}\sum_{j\geq 1} \frac{g_{j;k}}{n} x^{nj} = \sum_{m\geq 1} x^m \sum_{n\geq 1} \sum_{\substack{j\geq 1: \\ jn=m}} \frac{g_{j;k}}{n} \\ &= \sum_{m\geq 1} x^m \sum_{j:j|m} \frac{j g_{j;k}}{m}\end{aligned}$$

Then it follows that

$$g^\star_{n;k} = \frac{1}{n}\sum_{d:d|n} d g_{d;k} = g_{n;k} + \frac{1}{n}\sum_{\substack{d:d|n \\ d<n}} d g_{d;k}.$$

Therefore $g^\star_{1;k} = g_{1;k}$. From Corollary 2.1 it follows that

$$\begin{aligned}G_k(x) &= \frac{1}{2}\left(kx - 1 + e^{G^\star_k(x)}\right) = \frac{1}{2}\left(kx - 1 + \sum_{m\geq 0} \frac{(G^\star_k(x))^m}{m!}\right) \\ &= \frac{1}{2}\left(kx + \sum_{m\geq 1} \frac{(G^\star_k(x))^m}{m!}\right).\end{aligned}$$

So:

$$2G_k(x) = \left(kx + \sum_{m\geq 1} \frac{(G^\star_k(x))^m}{m!}\right).$$

In particular, we get $g_{1;k} = \frac{1}{2}(k + g_{1;k})$ (i.e. $g_{1;k} = k$, as expected, since $g_{1;k}$ counts the labeled single vertex trees). Moreover, for $n \geq 2$ we get

$$\begin{aligned}2g_{n;k} &= \sum_{m=1}^{n} \left(\frac{1}{m!} \sum_{\substack{(n_1,\ldots,n_m):n_i\geq 1 \\ n_1+\cdots+n_m=n}} \prod_{j=1}^{m} g^\star_{n_j;k}\right) \\ &= g^\star_{n;k} + \sum_{m=2}^{n} \left(\frac{1}{m!} \sum_{\substack{(n_1,\ldots,n_m):n_i\geq 1 \\ n_1+\cdots+n_m=n}} \prod_{j=1}^{m} g^\star_{n_j;k}\right),\end{aligned}$$

from which, using, we can obtain (for $n \geq 2$) that

$$g_{n;k} = \frac{1}{n}\sum_{\substack{d:d|n \\ d<n}} d g_{d;k} + \sum_{m=2}^{n} \left(\frac{1}{m!} \sum_{\substack{(n_1,\ldots,n_m):n_i\geq 1 \\ n_1+\cdots+n_m=n}} \prod_{j=1}^{m} \left(\frac{1}{n_j}\sum_{d:d|n_j} d g_{d;k}\right)\right). \tag{2.5}$$

We include some values of $g_{n;k}$ in Table 2.3.

Table 2.3: The first few values of $g_{n;k}$, the number of rooted gene trees with n leaves on the label set $[k]$. These counts were obtained using recursion (2.5).

$n\backslash k$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	1	3	6	10	15	21
3	2	10	28	60	110	182
4	5	40	156	430	965	890
5	12	170	948	3396	9376	21798
6	33	785	6206	28818	97775	269675
7	90	3770	42504	256172	1068450	3496326
8	261	18805	301548	2357138	12081605	46897359
9	766	96180	2195100	22253672	140160650	645338444
10	2312	502381	16307598	214370398	1658936806	9059465175

2.3 Alternative recursive function for rooted gene trees.

In the interest of developing a time efficient program to calculate counts of rooted MUL-trees, an alternative recursive function for $g_{n;k}$ was found. For the singleton tree we have $n = 1$ and $g_{1;k} = k$, so we now consider rooted gene trees that are non-trivial. As before, we establish a bijection between non-trivial rooted gene trees and forests of rooted gene trees with at least two components. When the number of leaves $n \geq 2$, this bijection can be described as follows: We remove the root of the tree and designate the neighbors of the original root as roots of the trees in the resulting forest. The total number of leaves in the forest is still n. The forest can be described as a partition of n into at least two classes, where the elements in each class represent the number of leaves for the corresponding tree in the forest. Thus, our goal is to have a suitable description of such partitions of n and the counts for the forests that result in this partition. Let $\mathcal{P}_n$ be the set of all partitions of n into at least 2 classes. Each such partition can be written as a unique sequence $\alpha = (a_1^{\beta_1}, a_2^{\beta_2}, \ldots, a_j^{\beta_j})$ with $n > a_1 \geq a_2 \geq \cdots \geq a_j \geq 1$, β_i are positive integers with $\beta_1 + \cdots + \beta_j \geq 2$ and $n = \beta_1 a_1 + \cdots + \beta_j a_j$ [42]. Each such partition describes a forest of trees. For each

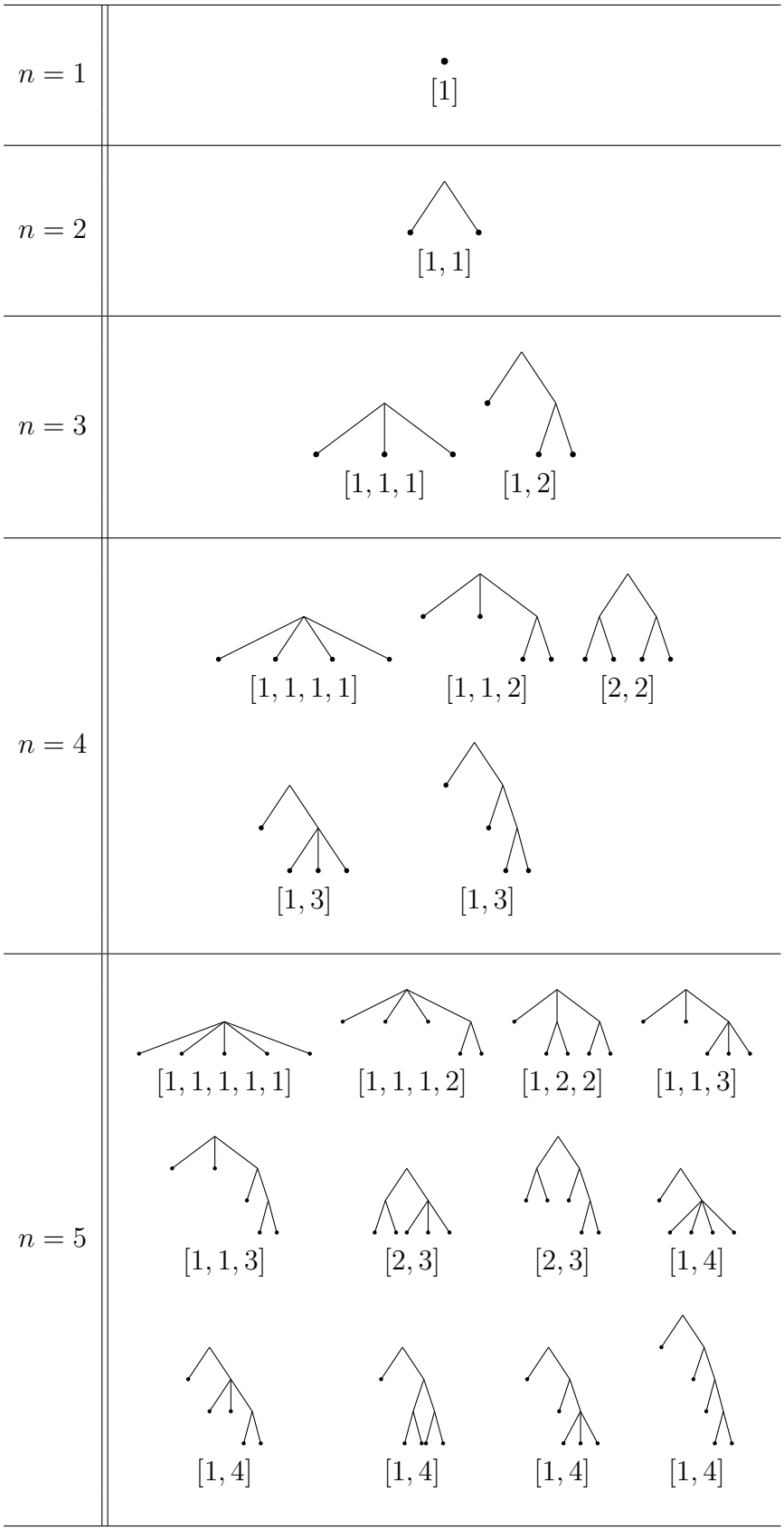


Figure 2.1: MUL-trees with one to five leaves on label set $[1]$ generated using the recursion 2.6.

a_i, the forest will contain a multiset of size β_i of MUL-trees which have a_i leaves. The number of rooted MUL-trees with a_i leaves is $g_{a_i;k}$. The number of ways to take a multiset of cardinality β_i from a set of cardinality $g_{a_i;k}$ is $\binom{g_{a_i;k}+\beta_i-1}{\beta_i}$ (choosing β_i objects from a set of $g_{a_i;k}$ items with replacement). Note that if $\beta_i = 0$, then $\binom{g_{a_i;k}-1}{0} = 1$. It follows that:

$$g_{n;k} = \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ \displaystyle\sum_{\alpha \in \mathcal{P}_n} \prod_{i=1}^{j} \binom{g_{a_i;k} + (\beta_i - 1)}{\beta_i} & \text{if } n > 1. \\ \alpha = \left(a_1^{\beta_1}, \ldots, a_j^{\beta_j}\right) \end{cases} \tag{2.6}$$

Figure 2.1 depicts the MUL-trees with one to five leaves on label set $[1]$. The partition under each tree is the one used in the construction of the tree.

2.4 Rooted leaf-multi-labeled trees in general

This section considers a different set of isomorphism classes of rooted leaf-multi-labeled trees on label set $[k]$, $\mathcal{F}_k$. This set includes the single vertex trees, trees in which unlabeled degree two vertices are allowed and trees in which the root may have degree one. The singleton tree in $\mathcal{F}_k$ is a root and a labeled leaf, but for all other trees in $\mathcal{F}_k$, the root is not labeled and is not considered a leaf, even if it is of degree one (see Definition 1.23). As before, we define the $(k+1)$-type β on $\mathcal{F}_k$ by $\beta(T) = (u, n_1, \ldots, n_k)$ if the tree T has u unlabeled vertices and n_i leaves labeled with i. Let $f_{u,n_1,\ldots,n_k}$ to be the number of trees in $\mathcal{F}_k$ with u unlabeled nodes and n_j nodes with label j, and $F(z; x_1, \ldots, x_k) = \sum f_{u,n_1,\ldots,n_k} z^u x_1^{n_1} \ldots x_k^{n_k}$ be the corresponding ordinary generating function. As in the previous chapter, for a leaf-multi-labeled $T \in \mathcal{F}_k$, let $\ell_j(T)$ be the number of vertices that have label j, by $un(T)$ the number

of unlabeled vertices, and let

$$\text{term}(T) \quad = \quad z^{un(T)} \prod_{j=1}^{k} x_j^{\ell_j(T)}.$$

Theorem 2.3.

$$F(z; x_1, \ldots, x_k) = (x_1 + \cdots + x_k - z) + z \cdot \text{Exp}\left(\sum_{n=1}^{\infty} \frac{1}{n} F(z^n; x_1^n, \ldots, x_k^n) \right)$$

Proof. There is exactly one tree on a single vertex with label j and this tree has no unlabeled vertices. Thus, $F(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)$ counts the trees in $\mathcal{F}_k$ with more than one vertex and is therefore divisible by z. The trees from $\mathcal{F}_k$ with at least one unlabeled vertex are in one to one correspondence with the nonempty forests, composed of trees from $\mathcal{F}_k$. This correspondence is obtained by removing the root and designating the neighbors of the removed root as the roots of the appropriate trees in the forest. The forest has at least one component, since the degree of the root was at least one. If a root in the forest has a label, the corresponding vertex in the original tree was a leaf. If the degree of the new root was $m \geq 2$ in the original tree, it is an unlabeled root of degree $m - 1$ in the forest. Let $\mathbf{H}_2(\cdot)$ count the non empty rooted finite forests of trees from $\mathcal{F}_k$. Then

$$\mathbf{H}_2(\cdot) = \frac{F(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)}{z}$$

Let $\mathbf{H}_3(\cdot) = \mathbf{H}_2(\cdot) + 1$, that is all finite rooted forests of trees in $\mathcal{F}_k$, including the empty forest. Using the same argument an in Theorem 2.2 we have

$$\begin{aligned} \mathbf{H}_3(\cdot) \quad &= \quad \prod_{T \in \mathcal{F}_k} \left(\sum_{j=0}^{\infty} \text{term}(T)^j \right) = \prod_{T \in \mathcal{F}_k} \left(1 - \text{term}(T) \right)^{-1} \\ &= \quad \prod_{(u; n_1, \ldots, n_k)} \left(\left(1 - z^u x_1^{n_1} \cdots x_k^{n_k} \right)^{-f_{u; n_1, \ldots, n_k}} \right). \end{aligned}$$

Thus $\log(\mathbf{H}_3(\cdot)) = \sum_{n=1}^{\infty} \frac{1}{n} F(z^n; x_1^n, \ldots, x_k^n)$, from which the theorem follows.

□

Chapter 3

Otter's Theorem

3.1 Background and statement

R. Otter presented a theorem in [36], which can be used to relate counts of rooted unlabeled trees to counts of unrooted unlabeled trees, using the idea of equivalent vertices (Definition 1.12), equivalent edges (Definition 1.27), and the symmetry edge (Definition 1.29) of a given tree.

More specifically, he showed the following:

Theorem 3.1. *In any tree the number of nonequivalent vertices minus the number of nonequivalent lines (symmetry line excepted) is one.*

Using our notation (see Notations 1.13, 1.28, 1.30), the above can be expressed as

$$p_T - (q_T - s_T) = 1.$$

F. Harary has stated a generalization of this theorem for unlabeled graphs [22]. Recall that for any semi-multi-labeled graph G, p_G denotes the number of non-equivalent vertices (Definition 1.12). We will let $q^\star_G$ be the number of non-equivalent blocks (Definition 1.19), and $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{q^\star_G}\}$ be the set of classes of isomorphic blocks. Also, we will use $b_{G,i}$ be the number of nonequivalent vertices in $\mathcal{B}_i$. Then the theorem as stated by Harary is:

Theorem 3.2. *For any unlabeled connected nontrivial graph G,*

$$p_G - 1 = \sum_{i=1}^{q^\star_G} (b_{G,i} - 1).$$

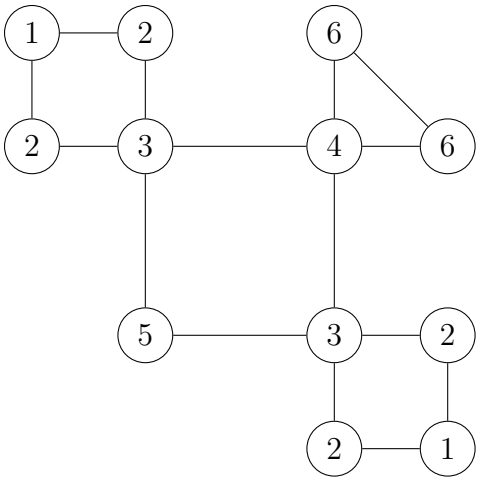


Figure 3.1: The numbers on the vertices are not labels, but are used to indicate which vertices are equivalent. There are three classes of blocks; one contains the two small 4-cycles ($\mathcal{B}_1$), the one large 4-cycle ($\mathcal{B}_2$) and the 3-cycle ($\mathcal{B}_3$). In this example, $q^\star_G = 3$, $p_G = 6$, $b_{G,1} = 3$, $b_{G,2} = 3$, and $b_{G,3} = 2$.

The example in figure 3.1 will help illustrate the theorem.

The proof of his theorem in *Graphical Enumeration* [22] is not entirely correct (for explanation, see Section 3.3). However, by introducing labels, the theorem can easily be proved for semi-multi-labeled graphs using the line of thought suggested by Harary.

3.2 Harary's Theorem and its consequences

This section will be devoted to the proof of Harary's Theorem for semi-multi-labeled graphs:

Theorem 3.3. *For any semi-multi-labeled connected nontrivial graph G,*

$$p_G - 1 = \sum_{i=1}^{q^\star_G} (b_{G;i} - 1). \tag{3.1}$$

Proof. Given any graph G with the corresponding labeling function $\alpha(v_i)$, we use induction on k, the number of blocks $q^\star_G$. If $q^\star_G = 1$, either G has only one block or G has several isomorphic blocks and a single cut-vertex. In either case, equation

(3.1) trivially holds. Let $k \geq 1$ and assume the statement holds for any graph G' with $q^\star_{G'} = k$. Consider a semi-labeled graph G with $q^\star_G = k+1 \geq 2$ and assume that α_G uses the label set $[n]$. Choose any block of G that has exactly one cut-vertex (such a block exists by Claim 1.18). This block belongs to one of the classes in $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{k+1}$. Without loss of generality we may assume that it belongs to block class $\mathcal{B}_{k+1}$. Delete all the vertices of the blocks in class $\mathcal{B}_{k+1}$ except the cut vertices of G to obtain G', which is a connected nontrivial subgraph of G by Claim 1.18 and the fact that $q^\star_G \geq 2$. Define the function $\alpha^\star_{G'} : V_{G'} \to \{0, 1, \ldots, n+1\}$ as follows. If $v_i \notin B$ for some $B \in \mathcal{B}_{k+1}$, then $\alpha^\star_{G'}(v_i) = \alpha^\star_G(v_i)$. If $v_i \in B \cap V(G')$ for some $B \in \mathcal{B}_{k+1}$ (v_i is a cut-vertex of G in a block of $\mathcal{B}_{k+1}$) then $\alpha^\star_{G'}(v_i) = n+1$. Note the label $n+1$ has not been used by $\alpha^\star_G$, so we have not inadvertently created any new equivalencies—a cut-vertex in a block of $\mathcal{B}_k$ can only be equivalent to another such cut-vertex in G', and therefore no new equivalencies between blocks or vertices have been created.

At this point we will argue that

$$\left\{\phi\Big|_{V(G)'} : \phi \text{ is an automorphism of } G\right\} = \left\{\phi : \phi \text{ is an automorphism of } G'\right\}$$

First we will show that the left-hand side of this equation is a subset of the right-hand side. Given any automorphism of ϕ of G, it is clear that $\phi\Big|_{V(G')}$ is an automorphism of the graph G' which preserves labels for those vertices v of G' which are not vertices in any block in $\mathcal{B}_{k+1}$, since in this case we must have $\alpha^\star_{G'}(v) = \alpha^\star_G(v) = \alpha^\star_G(\phi(v)) = \alpha^\star_{G'}(\phi(v))$ by definition of $\alpha^\star_{G'}$. If v is a cut-vertex in a block belonging to the class $\mathcal{B}_{k+1}$, then, because the labeling $\alpha^\star_{G'}$ uses a new label for these vertices, v is equivalent precisely with the cut vertices in blocks within $\mathcal{B}_{k+1}$ both in G and in G'. In particular, v is equivalent in G' with $\phi(v)$, and $\alpha^\star_{G'}(v) = n+1 = \alpha^\star_{G'}(\phi(v))$. Therefore we have that $\phi\Big|_{V(G')}$ is an automorphism of G' with the labeling $\alpha^\star_{G'}$.

What remains to be seen that the right hand side of the above equation is a subset of the left hand size. Given an automorphism ϕ' of (the semi-labeled graph) G', then

ϕ' must map the vertices that were cut-vertices of a block in $\mathcal{B}_{k+1}$ to a cut-vertex in a block in $\mathcal{B}_{k+1}$ since ϕ' must preserve the label $n+1$. Since any two blocks in $\mathcal{B}_{k+1}$ were isomorphic with the corresponding cut vertices mapped to each other, ϕ' can be extended to G by using these isomorphisms to some automorphism ϕ of G, thus, $\phi' = \phi\big|_{V(G')}$.

Therefore G' has the nonequivalent block classes $\mathcal{B}_1, \ldots, \mathcal{B}_k$ from the nonequivalent block classes of G and for $i \in \{1, \ldots, k\}$, we have $b_{G';i} = b_{G;i}$. Consequently, $p_{G'} = p_G - (b_{G;k+1} - 1)$. By the induction hypothesis equation (3.1) holds for G', thus,

$$\begin{aligned} p_G - 1 &= (b_{G;k+1} - 1) + (p_{G'} - 1) = (b_{G;k+1} - 1) + \sum_{i=1}^{k} (b_{G';i} - 1) \\ &= \sum_{i=1}^{k+1} (b_{G;i} - 1) = \sum_{i=1}^{q_G^\star} (b_{G;i} - 1) \end{aligned}$$

□

We can now obtain Otter's Theorem as a corollary, but it will be helpful to use notation referring specifically to trees. Given a nontrivial unrooted semi-labeled tree T, p_T is the number of non-equivalent vertices and $q_T^\star$ is the number of non-equivalent block classes in T. In a nontrivial tree the blocks are the edges with their end-vertices. Two edges are equivalent in the sense of Definition 1.27 when their blocks are equivalent in the sense of Definition 1.19, thus we have $q_T^\star = q_T$, motivating the strong similarity in the notations. As before let $b_{T;i}$ be the number of non-equivalent vertices in $\mathcal{B}_i$. If $\mathcal{B}_i$ consists of a symmetry edge (Definition 1.29) then $b_{T;i} = 1$, otherwise $b_{T;i} = 2$. We know that s_T, the number of symmetry edges is 0 or 1.

The generalization of Otter's Theorem to semi-multi-labeled trees is stated in the following corollary.

Corollary 3.4. *For any semi-labeled tree T, we have*

$$p_T - (q_T - s_T) = 1 \tag{3.2}$$

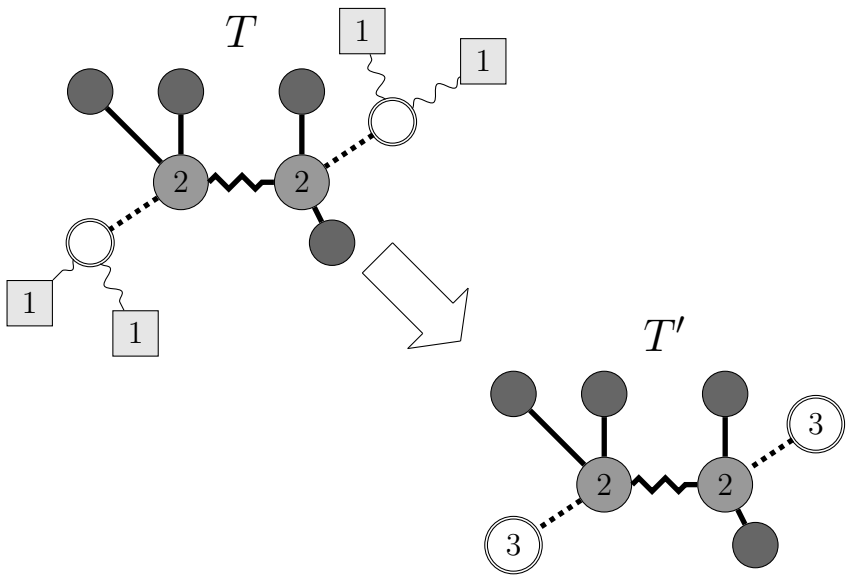


Figure 3.2: A semi-labeled tree T on label set $\{1, 2\}$ and a semi-labeled tree T' on label set $\{1, 2, 3\}$. The shapes, coloring and line types illustrate equivalence: vertices and edges that are depicted by the same kind of shape or line are equivalent. The jagged edge connecting the two vertices labeled by 2 is a symmetry edge. Note that $p_T = q_T = 4$, $s_T = s_{T'} = 1$ and $p_{T'} = q_{T'} = 3$. The equivalent blocks in T are the white circular nodes connected to the labeled leaves where the white circular nodes are the cut-vertices. Removing the leaves attached to these vertices and relabeling them as in the proof results in the tree T'.

Proof. If T is a singleton vertex, then $p_T = 1$, $q_T = s_T = 0$, and the statement holds.

Assume that T is nontrivial, so Theorem 3.3 applies, and we only need to show that $\sum_{i=1}^{q_T}(b_{T;i} - 1) = q_T - s_T$.

For each class of blocks other than one containing the symmetry edge the number of non-equivalent vertices is two. If an edge is a symmetry edge, the two vertices in this block are equivalent. Therefore, if there is no symmetry edge, $s_T = 0$, and $\sum_{i=1}^{q_T}(b_{T;i} - 1) = q_T = q_T - s_T$. If there is a symmetry edge, $s_T = 1$, and $\sum_{i=1}^{q_T}(b_{T;i} - 1) = q_T - 1 = q_T - s_T$. □

We are now ready to use Corollary 3.2 to relate counts of rooted leaf-multi-labeled trees to counts of unrooted leaf-multi-labeled trees, as Otter did for unlabeled trees. For this, the concept of marking will be used extensively.

Let T be an unrooted leaf-multi-labeled tree and mark one of its vertices. Clearly,

the number of non-isomorphic markings is p_T, since marking at two vertices gives rise to different marked trees if and only if the marked vertices are not equivalent. We use the term marking instead of rooting here, since, for example, if T is a nontrivial binary tree, the degree of the marked vertex is one (in the case of a labeled leaf) or three (in the case of an unlabeled vertex), unlike the root of a nontrivial rooted binary tree which must have degree two.

We can also obtain a marked tree by subdividing an edge of T into two edges and marking the resulting vertex of degree 2. If T was a nontrivial binary tree, the resulting marked tree can be considered a rooted binary tree with the marked vertex as root. Thus, q_T corresponds to the number of ways to root the tree T at one of its edges, and s_T corresponds to the number of ways to root the tree T at one of its edges so that the subtrees resulting from the removal of this root are isomorphic.

3.3 Counterexamples

The proof stated in of Harary's theorem for unlabeled graphs uses the same idea as our proof, claiming that removing a class of equivalent blocks in which the blocks each have exactly one cut-vertex results in a new graph in which the number of nonequivalent blocks is one less than in the original graph. Unfortunately, this statement is not true for unlabeled graphs in general, and is false even for trees, as shown by the counterexamples shown in figures 3.3 and 3.4

Generalizing the proof to include multi-labeled graphs removes this difficulty, since relabeling of the cut vertices insures that any set of blocks in G have the same equivalency relationships in the resulting subgraph G'.

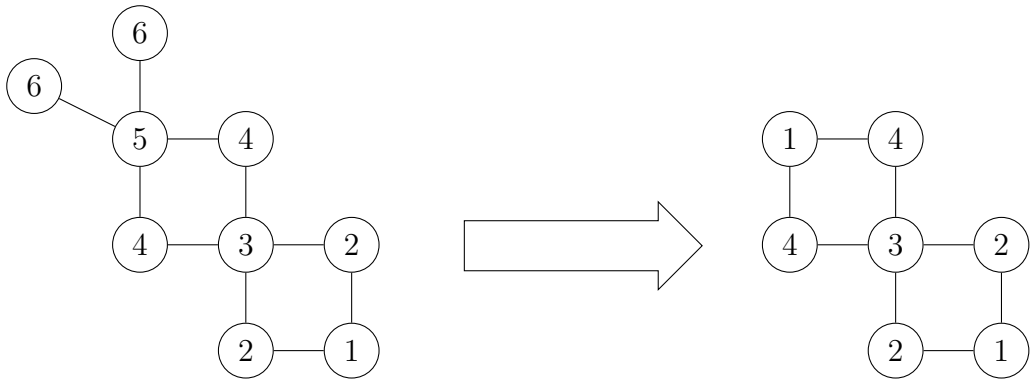


Figure 3.3: First counterexample: The numbers shown here are not labels, but indicate the equivalence classes of the vertices. The unlabeled graph G has two equivalent bridges and two nonequivalent 4-cycles. Thus, $q^\star_G = 3$ and $p_G = 6$. If the class of equivalent bridges is removed, for the resulting G', $q^\star_{G'} = 1$, not 2 as claimed, and $p_{G'} = 3$. Thus, $p_G - 1 \neq 1 + p_{G'}$ as claimed.

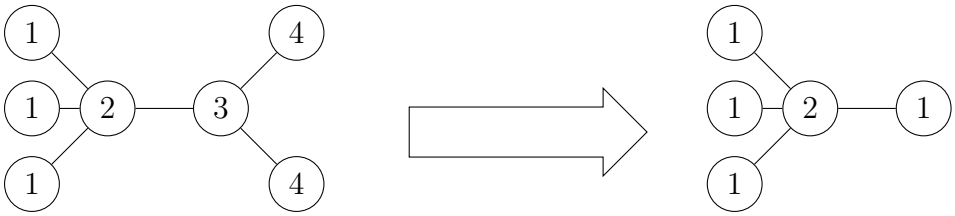


Figure 3.4: Second counterexample: as above, the numbers on the vertices are not labels, but indicate equivalence classes. The unlabeled tree T has three sets of nonequivalent bridges and four sets of nonequivalent vertices. Thus, $q_T = 3$ and $p_T = 4$. If the class with two equivalent bridges is removed, for the resulting T' is a star, so, $q_{T'} = 1$, not 2 as claimed, and $p_{T'} = 2$.Thus, $p_T - 1 \neq 1 + p_{T'}$as claimed.

Chapter 4

Unrooted leaf multi-labeled trees

4.1 Unrooted binary trees

In this section, we will present an equation for the generating function for unrooted binary leaf-multi-labeled trees.

As indicated in the previous section, in order to count unrooted binary trees it will be helpful to first count marked binary trees, where the marked vertices are either labeled leaves or internal vertices of degree three. We will denote the set of such marked binary trees with label set $[k]$ by $\mathcal{M}_k$, the corresponding k-type, as usual, is $(n_1, \ldots, n_k)$ where n_i is the number of leaves with label i, $m_{n_1,\ldots,n_k}$ is the number of trees in $\mathcal{M}_k$ with type $(n_1, \ldots, n_k)$, and the corresponding generating function is $M(x_1, \ldots, x_k) = \sum m_{n_1,\ldots,n_k} x_1^{n_1} \cdots x_k^{n_k}$.

We have the following:

Theorem 4.1.

$$\begin{aligned} M(x_1, \ldots, x_k) &= (x_1 + \cdots + x_k)\Big(1 + R(x_1, \ldots, x_k)\Big) + \frac{1}{6}R^3(x_1, \ldots, x_k) \\ &\quad + \frac{1}{2}R(x_1, \ldots, x_k)R(x_1^2, \ldots, x_k^2) + \frac{1}{3}R(x_1^3, \ldots, x_k^3). \end{aligned}$$

Proof. Let $T \in \mathcal{M}_k$ with marked vertex ρ_T. If ρ_T is a leaf of T marked with label j, then either T is a single vertex or the degree of ρ_T is one. In the latter case we can obtain a rooted binary tree $T' \in \mathcal{R}_k$ from T by setting $T' = T \setminus \{\rho_T\}$ and $\rho_{T'}$ be the unique neighbor of ρ_T in T. As $\rho_{T'}$ is either a (labeled) leaf of T or it has degree three in T, T' is either a (labeled) singleton tree or it has degree two in T', therefore $T' \in \mathcal{R}_k$ as claimed.

It follows that the counts for the trees in $\mathcal{M}_k$ with the marked vertex being a leaf have generating function $(x_1 + \cdots x_k)(1 + R(x_1, \ldots, x_k))$. It only remains to describe the generating function for marked trees where an internal vertex (i.e. vertex of degree three) is marked.

This is determined by the collection of *forests* consisting of three not necessarily different rooted binary leaf-multi-labeled trees. From any tree $T \in \mathcal{M}_k$ where the marked vertex ρ_T has degree three we can obtain such a forest by removing ρ_T and rooting each of the resulting trees at the corresponding neighbor of ρ_T. Since any neighbor of ρ_T was either a leaf, or it had degree three in T, the new root is either a vertex or it has degree two, as required.

Now, consider the the three terms $\frac{1}{6}R^3(x_1, \ldots, x_k)$, $\frac{1}{2}R(x_1, \ldots, x_k)R(x_1^2, \ldots, x_k^2)$, and $\frac{1}{3}R(x_1^3, \ldots, x_k^3)$. We will use Claims 1.35 and 1.37. A forest with three non-isomorphic trees in $\mathcal{R}_k$ is counted by $\frac{1}{6} \cdot 6 = 1$ times by the first term, and is not counted by the other two terms. A forest with two isomorphic trees and the third non-isomorphic to the first two is counted by the first term $\frac{1}{6} \cdot 3 = \frac{1}{2}$ times, by the second term $\frac{1}{2}$ times and the third term does not count it. A forest with three isomorphic trees forest is counted $\frac{1}{6} + \frac{1}{2} + \frac{1}{3} = 1$ times by the sum of these three terms. Thus, the forests with three trees from $\mathcal{R}_k$ are counted by $\frac{1}{6}R^3(\cdot) + \frac{1}{2}R(\cdot)R(\cdot^2) + \frac{1}{3}R(\cdot^3)$. This completes the proof of the theorem. □

Now, let $u_{n_1,\ldots,n_k}$ denote the number of unrooted leaf-multi-labeled binary trees where the label j is used n_j times, and let $U(x_1, \ldots, x_k) = \sum u_{n_1,\ldots,n_j} x_1^{n_1} \cdots x_k^{n_k}$. Using Corollary 3.4 we obtain the following:

Theorem 4.2.

$$\begin{aligned} U(x_1,\ldots,x_k) &= M(x_1,\ldots,x_k)+(x_1+\cdots+x_k)-R(x_1,\ldots,x_k) \\ &\quad +R(x_1^2,\ldots,x_k^2) \\ &= \Big(R(x_1,\ldots,x_k)+2\Big)\Big(x_1+\cdots+x_k-1+\frac{1}{2}R(x_1^2,\ldots,x_k^2)\Big) \\ &\quad +2+\frac{1}{3}R(x_1^3,\ldots,x_k^3)+\frac{1}{6}R^3(x_1,\ldots,x_k). \end{aligned}$$

Proof. Fix $n_1\ldots,n_k$ and sum equation (3.2) over all leaf-multi-labeled binary trees T where for all $j\in[k]$ the label j is used precisely n_j times. If we start from a non-singleton tree, p_T is the number of marked trees that are isomorphic to T, q_T is the number of rooted binary trees that are isomorphic to T after suppressing the root, and s_T is the number of rooted binary trees isomorphic to T, where the two rooted subtrees obtained by removing the root and rooting the remaining trees at the neighbor of the root are isomorphic to one another. So we obtain

$$u_{n_1,\ldots,n_k}=\begin{cases} 1 & \text{if } \sum n_j=1, \\ m_{n_1,\ldots,n_k}-r_{n_1,\ldots,n_k}+r_{n_1/2,\ldots,n_k/2} & \text{if } 2|n_j \text{ for all } j\in[k], \\ m_{n_1,\ldots,n_k}-r_{n_1,\ldots,n_k} & \text{otherwise.} \end{cases}$$

We obtain the theorem by multiplying both sides with $x_1^{n_1}\cdots x_k^{n_k}$ and summing over all values of $n_1,\ldots,n_k$. □

We note that if we let $u_{n;k}$ denote the number of unrooted leaf-multi-labeled binary trees using label set $[k]$ that have n leaves , and let

$$h^\star_{n;k}=kr_{n-1;k}-r_{n;k}+\frac{1}{6}\sum_{i=1}^{n-2}\sum_{j=1}^{n-i-1}\sum_{\ell=1}^{n-i-j}r_{i;k}r_{j;k}r_{\ell;k}+\frac{1}{2}\sum_{\substack{(i,j)\\2i+j=n}}r_{i;k}r_{j;k},$$

with $r_{n;k}$ as defined in Chapter 2.1, we can use the last theorem to obtain the following

recursion for computing $u_{n;k}$.

$$u_{n;k} := \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ h^{\star}_{n;k} + \frac{1}{3} r_{n/3;k} + r_{n/2;k} & \text{if } n = 6\ell, \ell \in \mathbb{N}, \\ h^{\star}_{n;k} & \text{if } n = 6\ell \pm 1, \ell \in \mathbb{N}, \\ h^{\star}_{n;k} + r_{n/2;k} & \text{if } n = 6\ell \pm 2 \geq 2, \ell \in \mathbb{Z}, \\ h^{\star}_{n;k} + \frac{1}{3} r_{n/3;k} & \text{if } n = 6\ell + 3 \geq 2, \ell \in \mathbb{Z}. \end{cases} \tag{4.1}$$

We include some values of $u_{n;k}$ in Table 4.1. We can also count only those trees which use every label in $[k]$ using the inclusion-exclusion principle and equation (4.1). Table 4.2 shows counts of these trees for trees with between 1 and 10 leaves. Notice that the first column in both tables gives the number of unlabeled unrooted binary trees with the indicated number of leaves.

Table 4.1: The first few values of $u_{n;k}$, the number of unrooted binary leaf-multi-labeled trees with n leaves on the label set $[k]$, obtained using recursion (4.1)

$n\backslash k$	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	1	3	6	10	15	21	28
3	1	4	10	20	35	56	84
4	1	6	21	55	120	231	406
5	1	12	63	220	600	1386	2842
6	2	31	227	1040	3530	9772	23366
7	2	78	891	5480	23250	77112	214718
8	4	234	3876	31420	165510	655599	2122099
9	6	722	17790	190360	1243825	5878446	22102577
10	11	2376	85536	1202930	9733950	54845721	239432081

Table 4.2: The first few values of $u_{n;k}$, the number of unrooted binary leaf-multi-labeled trees with n leaves on the label set $[k]$, with each label used at least once. These counts were obtained using the inclusion-exclusion principle with recursion (4.1).

$n\backslash k$	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	1	1	0	0	0	0	0
3	1	2	1	0	0	0	0
4	1	4	6	3	0	0	0
5	1	10	30	36	15	0	0
6	2	27	140	310	300	105	0
7	2	74	663	2376	3990	3150	945
8	4	226	3186	17304	44850	59805	39690
9	6	710	15642	123508	462735	925890	1018710
10	11	2354	78441	874998	4550955	12810825	20766375

4.2 Unrooted gene trees

Using Corollary 3.4, we now obtain analogous results for counting unrooted non-binary leaf-multi-labeled trees. Let $\mathcal{W}_k$ denote the class of unrooted leaf-multi-labeled trees where every internal vertex has degree at least 3. We define the $(k+1)$-type β on $\mathcal{W}_k$ by $\beta(T) = (u, n_1, \ldots, n_k)$ if the tree T has u unlabeled vertices and n_i leaves labeled with i. Let $w_{u,n_1,\ldots,n_k}$ to be the number of trees in $\mathcal{W}_k$ with u unlabeled nodes and n_j nodes with label j, and $W(z; x_1, \ldots, x_k) = \sum w_{u,n_1,\ldots,n_k} z^u x_1^{n_1} \cdots x_k^{n_k}$ be the corresponding generating function.

To give a formula for the function W in terms of A, it is helpful to slightly extend the definition of p_T given in Section 3.2. We denote by $p_{T;un}$ the number of nonequivalent, unlabeled points of a leaf-multi-labeled unrooted tree, and by $p_{T;j}$ the number of nonequivalent points of T that are labeled with j. Clearly, $p_T = p_{T;un} + \sum_{j=1}^{k} p_{T;j}$, and

$$p_T - q_T + s_T \;=\; p_{T;un} + \sum_{j=1}^{k} p_{T;j} - q_T + s_T = 1. \tag{4.2}$$

Using this we obtain

Theorem 4.3.

$$\begin{aligned} W(z;x_1,\dots,x_k) &= (1+x_1+\cdots+x_k)A(z;x_1,\dots,x_k) \\ &\quad -\frac{1}{2}\Big((z+1)A^2(z;x_1,\dots,x_k)+(z-1)A(z^2;x_1^2,\dots,x_k^2)\Big). \end{aligned}$$

Proof. By (4.2),

$$W(\cdot)=\sum_{T\in\mathcal{W}_k}\text{term}(T)=\sum_{T\in\mathcal{W}_k}\text{term}(T)(p_{T;un}+\sum_{j=1}^{k}p_{T;j}-q_T+s_T).$$

For any unrooted leaf-multi-labeled tree T, $p_{T;un}$ is the number of trees in $\mathcal{R}_k$ that are isomorphic to T and whose root is an unlabeled vertex of T (note that the root has degree at least 3). In addition, $p_{T;j}$ is the number of leaf-multi-labeled trees that are isomorphic to T and have a leaf-vertex with label j marked; q_T is the number of trees in $\mathcal{R}_k$ where the root has degree 2 and, after suppressing the root vertex, we obtain a tree that is isomorphic to T; and s_T is the number of trees that are counted by q_T for which the two subtrees at the root are isomorphic.

Now, to obtain the terms of $W(\cdot)$ corresponding to $\sum_T \text{term}(T)\sum_j p_{T;j}$, first note that the contribution of the single vertex trees marked at a (leaf-)vertex is counted by $\sum_j x_j$. Also, the contribution of the trees with at least two vertices that are marked at a leaf-vertex is counted by $A(\cdot)\sum_j x_j$, since removing the marked vertex and rooting the remaining tree at the neighbor of this marked vertex gives a tree in $\mathcal{R}_k$. Thus $\sum_T \text{term}(T)\sum_j p_{T;j} = (A(\cdot)+1)\sum x_j$.

We now consider the terms corresponding to $\sum_T \text{term}(T)p_{T;un}$. If we consider the unlabeled marked vertex root, we get a tree in $\mathcal{R}_k$ whose root must have degree at least 3. Also, using similar arguments to those used in the proof of Theorem 2.1, The trees in $\mathcal{R}_k$ with root having degree less than 3 (so 2 or 0) are counted by $\frac{z}{2}(A^2(\cdot)+A(\cdot^2))+\sum_j x_j$, therefore

$$\sum_T \text{term}(T)p_{T;un} = A(\cdot)-\frac{z}{2}(A^2(\cdot)+A(\cdot^2))-\sum_j x_j$$

.

Therefore, $\sum_{T\in\mathcal{B}} \mathrm{term}(T)(p_{T;un} + \sum_j p_{T;j}) = (1 + \sum_j x_j)A(\cdot) - \frac{z}{2}(A^2(\cdot) + A(\cdot^2))$.

To complete the proof, note that $\sum_{T\in\mathcal{W}_k} \mathrm{term}(T)(q_T - s_T)$ counts those rooted gene trees (without counting their roots) where the root has degree 2 and the two rooted subtrees obtained when removing the original root are non-isomorphic. Again, using arguments similar to the ones used in Theorem 2.1 we obtain

$$\sum_{T\in\mathcal{W}_k} \mathrm{term}(T)(q_T - s_T) = \frac{1}{2}(A^2(\cdot) - A(\cdot^2)).$$

□

We now use this result to give a formula for the generating function for the unrooted leaf-multi-labeled trees without having to keep track of the number of unlabeled vertices: Let $s_{n_1,\ldots,n_k}$ denote the unrooted leaf-multi-labeled trees where no vertex has degree 2, and where exactly n_j copies of the label j used. Let the generating function be $S(x_1,\ldots,x_k) = \sum s_{n_1,\ldots,n_k} x_1^{n_1}\cdots x_k^{n_k}$. Then setting $z = 1$ in the statement of Theorem 4.3 we obtain the following corollary.

Corollary 4.1.

$$S(x_1,\ldots,x_k) \;=\; G(x_1,\ldots,x_k)(x_1+\cdots+x_k+1) - G^2(x_1,\ldots,x_k).$$

Using this in a similar way to that described above for $g_{n;k}$, we obtain a recursion for counting the number $s_{n;k}$ of unrooted leaf-multi-labeled trees on n leaves using $[k]$ as label set:

$$s_{n;k} = \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ k g_{n-1;k} + g_{n;k} + \displaystyle\sum_{j=1}^{n-1} g_{j;k} g_{n-j;k} & \text{if } n \geq 2. \end{cases} \tag{4.3}$$

We include some values of $s_{n;k}$ in Table 4.3.

Table 4.3: The first few values of $s_{n;k}$, the number of unrooted non-binary leaf-multi-labeled trees with n leaves on the label set $[k]$. These counts were obtained using the recursion 4.3.

$n\backslash k$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	3	11	24	42	65	93
3	5	28	82	180	335	560
4	12	109	444	1250	2840	5607
5	31	470	2688	9756	27151	63462
6	83	2145	17394	81770	279465	774543
7	233	10300	118470	721508	3028655	9953952
8	670	51135	835980	6599982	34035550	132664149
9	1981	260930	6062392	62041488	393044405	1816894738
10	5966	1359391	44897274	595614158	4635468832	25412433213

4.3 Unrooted leaf-multi-labeled trees in general

Using Corollary 3.4, we now obtain analogous results for counting unrooted trees without any degree restrictions. These trees may have internal non-root vertices of degree two. Since we can always obtain a new tree from an old one by replacing an edge with a path of any length, there are infinitely many different trees with the same number of labeled leaves. Note that we are absolutely forced to keep track of the number of internal vertices in this case. For example infinitely many different paths exist with the two leaves labeled by 1, and those paths are distinguished by the number of their internal vertices. Let $\mathcal{D}$ denote the class of unrooted leaf-multi-labeled trees, where these trees do not have any restrictions on the degree of tinternal vertices (see Definition 1.23). Let $d_{u;n_1,\ldots,n_k}$ denote the number of trees in $\mathcal{D}$ that have u unlabeled vertices and in which precisely n_j copies of the label j are used, and let $D(z;x_1,\ldots,x_k) = \sum d_{u;n_1,\ldots,n_k} z^u x_1^{n_1}\cdots x_k^{n_k}$.

To give a formula for the function D in terms of F, we will again denote by $p_{T;un}$ the number of nonequivalent, unlabeled points of a leaf-multi-labeled unrooted tree T, and by $p_{T;j}$ the number of nonequivalent points of T that are labeled with j. Using

equation (4.2) we obtain

Theorem 4.4.

$$\begin{aligned} D(z;x_1,\ldots,x_k) &= (1-z+x_1+\cdots+x_k)F(z;x_1,\ldots,x_k) \\ &\quad -\frac{1}{2}\Big(F^2(z;x_1,\ldots,x_k)+F(z^2;x_1^2,\ldots,x_k^2)\Big). \end{aligned}$$

Proof. By (4.2),

$$D(z;x_1,\ldots,x_k)=\sum_{T\in\mathcal{D}}\operatorname{term}(T)=\sum_{T\in\mathcal{D}}\operatorname{term}(T)(p_{T;un}+\sum_{j=1}^{k}p_{T;j}-q_T+s_T).$$

For any unrooted leaf-multi-labeled tree T, $p_{T;un}$ is the number of trees in $\mathcal{D}$ that are isomorphic to T and whose root is an unlabeled vertex of T (in particular, the root has degree at least 2). In addition, $p_{T;j}$ is the number of leaf-multi-labeled trees that are isomorphic to T and have a leaf-vertex with label j marked; q_T is the number of trees in $\mathcal{D}$ where the root has degree 2 and, after suppressing the root vertex, we obtain a tree that is isomorphic to T; and s_T is the number of trees that are counted by q_T for which the two subtrees at the root are isomorphic.

Now, to obtain the terms of $D(\cdot)$ corresponding to $\sum_T \operatorname{term}(T)\sum_j p_{T;j}$, first note that the contribution of the single vertex trees marked at a (leaf-)vertex is counted by $\sum_j x_j$. Also, the contribution of the trees with at least two vertices that are marked at a leaf-vertex is counted by $F(\cdot)\sum_j x_j$, since removing the marked vertex and rooting the remaining tree at the neighbor of this marked vertex gives a tree in $\mathcal{D}$. Thus $\sum_T \operatorname{term}(T)\sum_j p_{T;j}=(F(\cdot)+1)\sum x_j$.

We now consider the terms corresponding to $\sum_T \operatorname{term}(T)p_{T;un}$. If we consider the unlabeled marked vertex a root, we have a tree in $\mathcal{F}$ whose root must have degree at least 2. Also, trees in $\mathcal{F}$ with root of degree less than 2 (so 1 or 0) are counted by the singleton trees, $\sum_j x_j$, and $z(F(\cdot))$, where an unlabled root has been added to the root of any tree in $\mathcal{F}$. Therefore $\sum_T \operatorname{term}(T)p_{T;un}=F(\cdot)-z(F(\cdot))-\sum_j x_j$.

Therefore, $\sum_{T\in\mathcal{D}}\operatorname{term}(T)(p_{T;un}+\sum_j p_{T;j})=(1-z+\sum_j x_j)F(\cdot)$.

To complete the proof, note that $\sum_{T\in\mathcal{B}} \text{term}(T)(q_T - s_T)$ counts those rooted gene trees (without counting their roots) where the root has degree 2 and the two rooted subtrees obtained when removing the original root are non-isomorphic. Again, using arguments similar to the ones used in Theorem 2.1 we obtain

$$\sum_{T\in\mathcal{D}} \text{term}(T)(q_T - s_T) = \frac{1}{2}(F^2(\cdot) - F(\cdot^2)),$$

from which the theorem follows.

□

Chapter 5

Asymptotics for leaf-labeled trees

5.1 Leaf-labeled trees and set partitions

We now turn our attention to rooted phylogenetic trees. Our aim is to develop asymptotic formulae for such trees.

To this end, we first describe a bijection between the set of rooted leaf-labeled trees with n non-root vertices and k leaves, and partitions of an n element set into $n-k+1$ classes, developed by Erdős and Székely [13]. As is customary, the Stirling number $S(n,k)$ denotes the number of partitions of $[n]$ into k partition classes. We will use $F(n,k)$ to denote the number of rooted leaf-labeled (not necessarily phylogenetic) trees with k uniquely labeled non-root leaves and n non-root vertices, where the root, if it is of degree one, is unlabeled and is not counted as a leaf. The vertex of the trivial tree, as usual, will be both a root and a leaf, and will be labeled. Note that when $k \geq 2$ then our tree can not be trivial, and therefore the non-root vertices include all the k labeled leaves. Thus we must have $F(n,k) = 0$ for all $k > 1$ and $0 \leq n < k$. Also, $F(n,1) = 1$ for all $n \geq 0$ (there is precisely one such tree, a path of length n). For all $n \geq 0$, we have $F(n,0) = 0$.

The label set for such trees with k leaves is assumed to be $[k]$, the root may have degree one and internal vertices may have degree two, so these are not yet the phylogenetic trees of interest. Péter Erdős and László Székely [13] gave a bijection between the trees counted by $F(n,k)$ and partitions of an n-element set into $n-k+1$ classes. We give a brief sketch of this bijection after a few definitions. The first are

terms that help us refer to the structure of the rooted tree:

Definition 5.1. Let T be a rooted tree with root ρ_T. If the path from the root ρ_T to a vertex a contains the vertex b, the vertex a is said to be *below* vertex b. This relationship is a well-known partial order on the vertices of T.
A *child* of a vertex a is any vertex c adjacent to and below a. The vertex a is referred to as the *parent* of c.

The Erdős-Székely bijection uses the antilexicographic order on subsets of an ordered set

Definition 5.2. Let $\mathcal{X}$ be an ordered set. The *antilexicographic order* $<_{AL}$ on the power set of $\mathcal{X}$ is defined as follows:

$$A <_{AL} B \Leftrightarrow \max(A \Delta B) = \max\{(A \backslash B) \cup (B \backslash A)\} \in B.$$

The bijection can be described as follows.

If T is a trivial tree, i.e. a single vertex labeled with 1, then $n = 0$, $k = 1$. This is the only tree that has these parameters, so $F(0, 1) = 1$. In this case $n - k + 1 = 0$, so we need to assign a partition of the empty set to no partition classes (the empty partition) to this. This agrees with the usual definition $S(0, 0) = 1$.

Given a non-trivial leaf-labeled tree T with n non-root vertices and k labeled leaves we have $n \geq k \geq 1$, and $n - k + 1 \geq 1$. Since the root is not a leaf, T has $n + 1$ vertices, and $n - k + 1$ is the number of non-leaf vertices in T. We will give a partition of $[n]$ into $n - k + 1$ classes by first establishing a bijection ϕ between $[n]$ and the set of non-root vertices of T, and then assigning to each non-leaf vertex x the set $\{\phi(c) : c \text{ is a child of } x\}$. Since each non-root vertex of T is a child of precisely one non-leaf node, and non-leaf nodes have at least one child, the sets assigned to the non-root vertices will form a partition of $[n]$, as required. The number of partition classes is the number of non-leaf vertices, $n - k + 1$. By construction, the size of

each partition class is the number of children of the corresponding non-leaf vertex, a property we will want to exploit later. The special properties of ϕ ensure that for any appropriate partition we can determine the tree that gave rise to it.

The set of labels $[k]$ is clearly an ordered set, where the ordering is the usual ordering on the numbers. We need to construct the bijection ϕ between the n non-root vertices and $[n]$. Given a leaf-labeled tree, each non-root vertex is assigned a subset of $[k]$ as follows. Every leaf is assigned the set consisting of its label. Each non-leaf, non-root vertex is assigned the set containing the labels of the leaves below this vertex. Once every non-leaf vertex has been assigned a subset of $[k]$, these subsets are ordered using the antilexicographic ordering. If some of the internal vertices have degree two, it may happen that some sets occur more than once. In this instance the set of the vertex closer to the root is considered the"larger". Each non-root vertex is then given a new label corresponding to the position of its assigned set in the ordering. The tree is then assigned the partition in which there is a partition class corresponding to each non-leaf node containing the numbers assigned to its children.

The properties of the antilexicographic ordering together with the way we define the partition for the tree ensure the following:

1. The size of each partition class is equal to the number of children of the corresponding vertex.

2. The partition class which contains n is the set containing the children of the root.

3. The partition class corresponding to a non-root, non-leaf vertex a with $\phi(a) = m$ contains the number $m-1$, and all other numbers in this class are smaller than $m-1$.

Note that in the context of this terminology, a phylogenetic tree is simply a leaf-labeled tree where all non-leaf vertices have at least two children.

Given a partition $\mathcal{P}$ of $[n]$, with $n > 0$ we can find the corresponding tree T as follows. We must have a rooted tree with $n+1$ vertices and $k = n+1-|\mathcal{P}|$ leaves. Since $1 \leq |\mathcal{P}| \leq n$, we have that $1 \leq k \leq n$, so this, at first glance, is possible.

Begin with $n+1$ vertices; one is designated the root and the others are labeled by $1, 2 \ldots, n$, which correspond the values of ϕ taken on the tree.

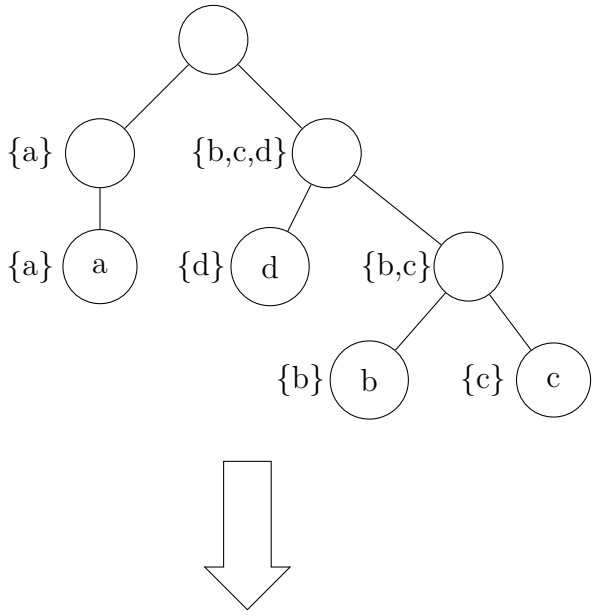
Let $A \in \mathcal{P}$ be a partition class that contains n; connect the vertices labeled by elements of A to the root. For any $B \in \mathcal{P}$, if $B \neq A$ then $b := \max(B) < n$. Connect the vertices labeled by elements of B to the vertex labeled $b+1$.

It is easy to show (and is omitted) that the resulting graph is cycle-free. Since the graph has $n+1$ vertices and n edges, it is a tree. Since elements of each partition class have the same parent, and elements of different partition classes have different parents, we have $|\mathcal{P}|$ vertices that are parents of some vertex, and so we have $n+1-|\mathcal{P}| = k$ leaves, as required. We omit the proof that the resulting tree indeed gives rise to the partition $\mathcal{P}$, as claimed above. For further details, the reader should consult [13].

See Figures 5.1 and 5.2 for an example of the bijection. (A similar result was established independently by Haiman and Schmitt [21].)

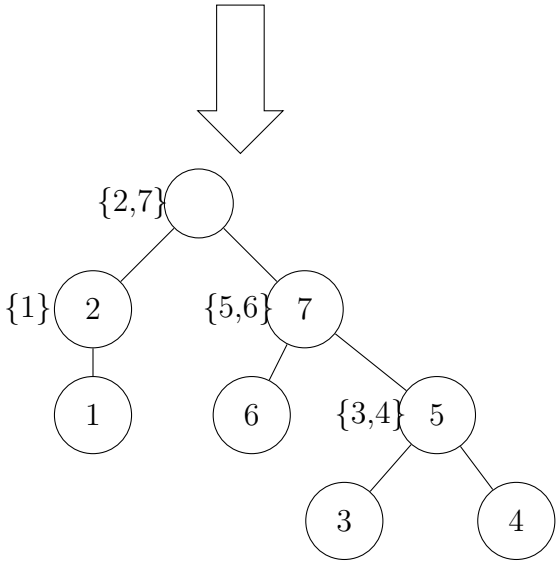
For all other (n,k) pairs, i.e. when $(n,k) \notin \{(0,1)\} \cup \{(a,b) \in \mathbb{Z}^+ : a \geq b\}$, we have $F(n,k) = 0$, since there are no trees with those parameters. Also, it is easy to see that $S(n, n-k+1) = 0$ for these (n,k) pairs. Thus, the Erdős-Székely bijection means that $F(n,k) = S(n, n-k+1)$ for all integers n, k.

It immediately follows that $\sum_k F(n,k) = \sum_i S(n,i) = B(n)$, the n-th *Bell number*, the number of ways to partition $[n]$, A000110 in *The On-Line Encyclopedia of Integer Sequences* [41]. Inverting the relationship $S(n,i) = F(n, n-i+1)$, and the abundant information available on the Stirling numbers of the second kind translates to information on the counts of rooted leaf-labeled trees. In this section we discuss some results on the Stirling numbers of the second kind for two reasons: they immediately apply to the counts of these trees and will provide guidelines for Harper's



$$\{a\} < \{a\} < \{b\} < \{c\} < \{b,c\} < \{d\} < \{b,c,d\}$$

1 2 3 4 5 6 7



$$\{2,7\}, \{5,6\}, \{3,4\}, \{1\}$$

Figure 5.1: Demonstrating the steps of the Erdős-Székely bijection from a rooted leaf-labeled tree to a partition of [7].

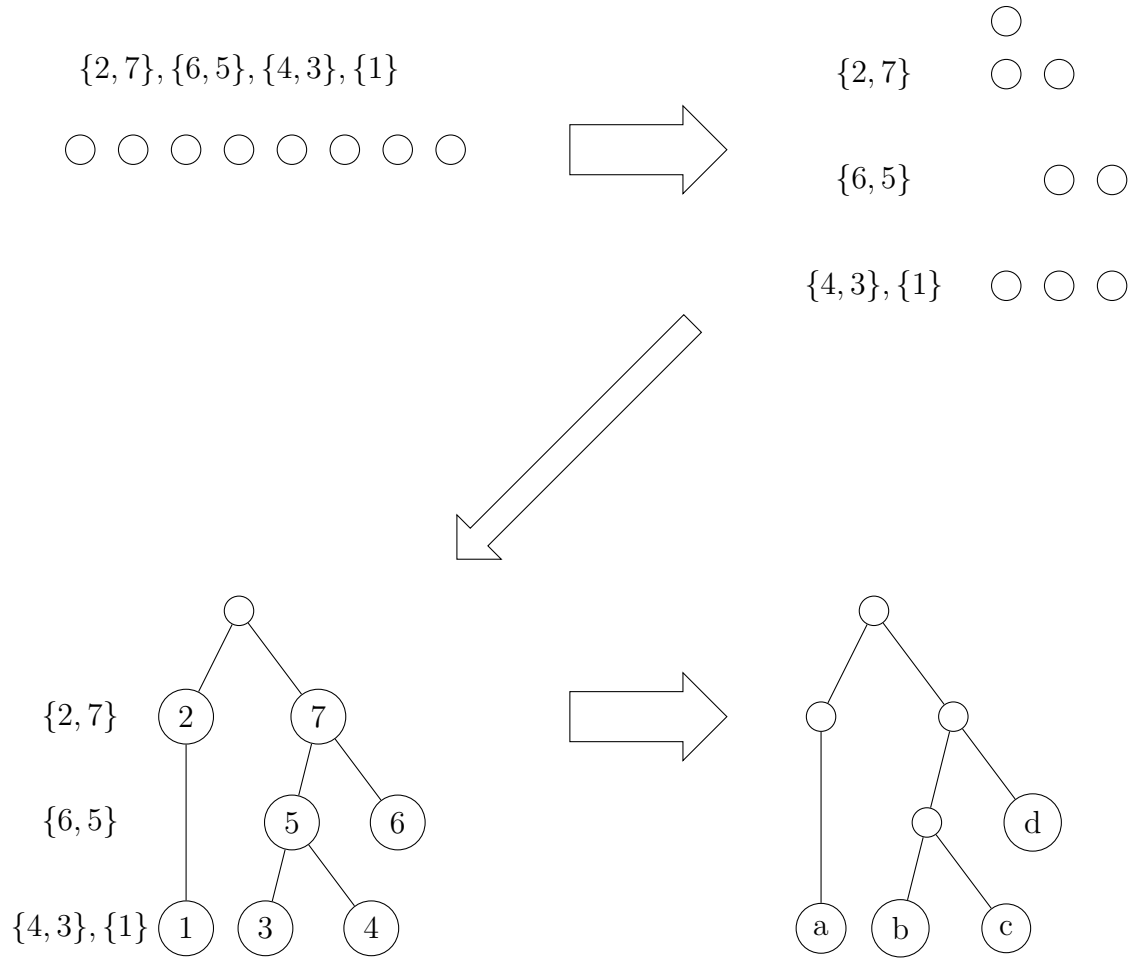


Figure 5.2: Demonstrating the steps of the Erdős-Székely bijection from a partition of $[7]$ to a rooted leaf-labeled tree.

method to obtain results in sections 6.1 and 6.3.

The bivariate generating function (page 88 [44])

$$\sum_{n\geq 0}\sum_{k\geq 0} S(n,k)x^k\frac{t^n}{n!} = e^{x(e^t-1)} \tag{5.1}$$

becomes

$$\sum_{n}\sum_{k} F(n,k)x^k\frac{t^n}{n!} = xe^{(e^{tx}-1)/x}$$

after substituting $1/x$ into x, tx into t, and multiplication by x as shown below. Since $F(n,k) = 0$ when $\min(n,k) = 0$ and $k \neq 1$, we have $\sum_{(n,k):\min(n,k)=0} F(n,k)x^k\frac{t^n}{n!} = x$.

Also, $\sum_{(n,k):\min(n,k)=0} S(n,k)x^k\frac{t^n}{n!} = 1$. Thus,

$$\begin{aligned}
x\left(e^{\frac{e^{tx}-1}{x}}-1\right) &= x\left(\sum_{n=0}^{\infty}\sum_{k=0}^{n} S(n,k)x^{-k}\frac{(tx)^n}{n!} - \sum_{(n,k):\min(n,k)=0} S(n,k)x^k\frac{t^n}{n!}\right)\\
&= x\sum_{n=1}^{\infty}\sum_{k=1}^{n} S(n,k)x^{-k}\frac{(tx)^n}{n!} = \sum_{n=1}^{\infty}\sum_{k=1}^{n} S(n,k)x^{n-k+1}\frac{t^n}{n!}\\
&= \sum_{n=1}^{\infty}\sum_{j=1}^{n} S(n,n-j+1)x^j\frac{t^n}{n!} = \sum_{n=1}^{\infty}\sum_{j=1}^{n} F(n,j)x^j\frac{t^n}{n!}\\
&= \left(\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} F(n,k)x^k\frac{t^n}{n!}\right) - \left(\sum_{(n,k):\min(n,k)=0} F(n,k)x^k\frac{t^n}{n!}\right)\\
&= \left(\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} F(n,k)x^k\frac{t^n}{n!}\right) - x
\end{aligned}$$

For $1 \le k \le n$ we have the recurrence relation

$$S(n,k) = S(n-1,k-1) + kS(n-1,k), \tag{5.2}$$

since $S(n-1,k-1)$ counts the partitions of $[n]$ where $\{n\}$ is a partition class, and $kS(n-1,k)$ counts those partitions of $[n]$ where the partition class containing n contains some other element of $[n-1]$ as well. Since $1 \le k \le n$ is equivalent with $1 \le n-k+1 \le n$, this translates to $F(n,k) = F(n-1,k)+(n+1-k)F(n-1,k-1)$, as follows

$$\begin{aligned}
F(n,k) &= S(n,n-k+1)\\
&= S(n-1,n-k) + (n-k+1)S(n-1,n-k+1)\\
&= F\Big(n-1,(n-1)-(n-k)+1\Big)\\
&\qquad +(n-k+1)F\Big(n-1,(n-1)-(n-k+1)+1\Big)\\
&= F(n-1,k) + (n-k+1)F(n-1,k-1)
\end{aligned}$$

Applying formula (5.2) for the polynomials $R_n(x) = \sum_k S(n,k)x^k$ one obtains the recurrence relation

$$R_n(x) = x\Big(R'_{n-1}(x) + R_{n-1}(x)\Big) \tag{5.3}$$

with initial condition $R_1(x) = x$.

5.2 Harper's Method

Harper [25] gave a very elegant proof for the asymptotic normality of the array $S(n,k)$. We follow the interpretation of Canfield [4] and Clark [8], who clarified and explained the details of Harper's method. Let $A(n,j)$ be an array of non-negative real numbers for $j = 0, 1, \ldots, d_n$, and define $A_n(x) = \sum_j A(n,j)x^j$.

Observe that $\sum_j A(n,j) = A_n(1)$. Let Z_n denote the random variable, for which the probability $\mathcal{P}(Z_n = j) = \frac{A(n,j)}{A_n(1)}$. In terms of $A_n(x)$, there is a well-known [8] expression for the expectation and variance of Z_n:

$$\mathcal{E}(Z_n) = \frac{A_n'(1)}{A_n(1)} \text{ and } \mathcal{D}^2(Z_n) = \frac{A_n'(1)}{A_n(1)} + \left(\frac{A_n'(x)}{A_n(x)}\right)'\Bigg|_{x=1}. \tag{5.4}$$

As $\mathcal{E}(Z_n)$ and $\mathcal{D}(Z_n)$ are determined by the array $A(n,j)$, we will also write them as $\mathcal{E}(A(n,.))$ and $\mathcal{D}(A(n,.))$

The array $A(n,j)$ is called *asymptotically normal* in the sense of a *central limit theorem*, if

$$\frac{1}{A_n(1)} \sum_{j=1}^{\lfloor x_n \rfloor} A(n,j) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \tag{5.5}$$

as $n \to \infty$ uniformly in x, where

$$x_n = \mathcal{E}(Z_n) + x\mathcal{D}(Z_n).$$

Note that the left side of (5.5) is $\mathcal{P}(Z_n \leq x_n)$, so asymptotic normality of the array $A(n,k)$ means that the cumulative density function of $\frac{Z_n - \mathcal{E}(Z_n)}{\mathcal{D}(Z_n)}$ approaches the standard normal cumulative density function uniformly everywhere.

Let $\{-y_{nk} : k = 1, 2, \ldots, d_n\}$ be the set of roots of the polynomial $A_n(x)$ and assume that all $-y_{n,k}$ are non-positive. Define the independent random variables Y_{nk} by $\mathcal{P}(Y_{nk} = 0) = y_{nk}/(1 + y_{nk})$ and $\mathcal{P}(Y_{nk} = 1) = 1/(1 + y_{nk})$.

Then the probability generating function of the random variable Z_n is $A_n(x)/A_n(1)$; and the probability generating function of the random variable Y_{nk} is $\frac{x+y_{nk}}{1+y_{nk}}$. Since the

probability generating function of a sum of independent random variables is the product of their probability generating functions, we have that the probability generating function of $\sum_k Y_{nk}$ is $\prod_{k=1}^{d_n} \frac{x+y_{nk}}{1+y_{nk}}$. However, as

$$\prod_{k=1}^{d_n} \frac{x+y_{nk}}{1+y_{nk}} = \frac{A_n(x)}{A_n(1)},$$

we conclude that Z_n and $\sum_k Y_{nk}$ have identical distribution.

Let $G_{nj}(x) = \mathcal{P}\Big(\frac{Y_{nj}-\mathcal{E}(Y_{nj})}{\mathcal{D}(Z_n)} \le x\Big)$ denote the cumulative distribution function of $\frac{Y_{nj}-\mathcal{E}(Y_{nj})}{\mathcal{D}(Z_n)}$ for $j = 1, \ldots, d_n$. The Lindeberg–Feller Theorem applies ([12] pp. 98–101) to the sequence $\frac{Z_n-\mathcal{E}(Z_n)}{\mathcal{D}_n(Z_n)} = \sum_j \frac{Y_{nj}-\mathcal{E}(Y_{nj})}{\mathcal{D}_n(Z_n)}$. The condition of the cited theorem, for all $\epsilon > 0$

$$\lim_{n\to\infty} \sum_{j=1}^{d_n} \int_{|y|>\epsilon} y^2 dG_{nj}(y) = 0$$

follows from

$$\lim_{n\to\infty} \mathcal{D}(Z_n) = \infty. \tag{5.6}$$

Therefore, the cited theorem proves the normal convergence (5.5), provided (5.6) holds and all the roots of the polynomials $A_n(x)$ have non-positive real numbers.

A sequence a_k is called *unimodal*, if first it increases, and then decreases. An array $A(n,k)$ is called *unimodal*, if for every n, the sequence $a_k = A(n,k)$ is unimodal. A sequence a_k, which is 0 for $k < t$ and $\ell < k$, with $a_t \neq 0$ and $a_\ell \neq 0$, is called *strictly log-concave* (SLC) if $a_k^2 - a_{k-1}a_{k+1} > 0$ for $t+1 \le k \le \ell-1$. An array $A(n,k)$ is called *strictly log-concave* (SLC), if for every fixed n, the sequence $a_k = A(n,k)$ is strictly log-concave. It is clear that any SLC sequence is unimodal in the variable k. Some LC sequences may not be unimodal, like 0,1,1,0,0,1,1,0. However, LC sequences, which do not have 0 terms both preceded and followed by non-zero terms (have *no internal zeroes* property) are also unimodal. Dobson [10] showed the unimodality of $S(n,k)$, Klarner [31] was the first to show the SLC property of $S(n,k)$.

Using Newton's Inequality, Lieb [31] showed that if a polynomial $\sum_{k=1}^{N} C_k x^k$ has

only real roots, then for $k = 2, \ldots, N-1$

$$C_k^2 \geq C_{k+1} C_{k-1} \Big(\frac{k}{k-1} \Big) \Big(\frac{N-k+1}{N-k} \Big). \tag{5.7}$$

Therefore, the C_k sequence is SLC. E.R. Canfield [4] noted that for asymptotically normal sequences (5.5), the SLC property and $\mathcal{D}(Z_n) \to \infty$ implies the following *local limit theorem*:

$$\lim_{n \to \infty} \frac{\mathcal{D}(Z_n)}{A_n(1)} A(n, \lfloor x_n \rfloor) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{5.8}$$

uniformly in x.

Again, the left side of (5.8) is

$$\mathcal{D}(Z_n) \mathcal{P}(Z_n = \lfloor x_n \rfloor) = \frac{\mathcal{P}\left(\frac{x_n - 1}{\mathcal{D}(Z_n)} < \frac{Z_n - \mathcal{E}(Z_n)}{\mathcal{D}(Z_n)} \leq \frac{x_n}{\mathcal{D}(Z_n)} \right)}{\frac{1}{\mathcal{D}(Z_n)}},$$

which gives a justification why we want this local condition.

Furthermore, from the fact that the convergence of the array $A(n, j)$ to the Gaussian function is actually uniform, Canfield concluded that the number $k = J_n$ maximizing $A(n, k)$ satisfies

$$J_n - \mathcal{E}(Z_n) = o(\mathcal{D}(Z_n)); \tag{5.9}$$

and

$$A(n, J_n) \sim \frac{1}{\sqrt{2\pi}} \frac{A_n(1)}{\mathcal{D}_n(Z_n)}. \tag{5.10}$$

For the Stirling numbers of the second kind, $A(n, j) = S(n, j)$, $A_n(1) = B_n$, and one has

$$\begin{aligned} \mathcal{E}(S(n, .)) &= \frac{B_{n+1}}{B_n} - 1, \\ \mathcal{D}^2(S(n, .)) &= \frac{B_{n+2}}{B_n} - \left(\frac{B_{n+1}}{B_n} \right)^2 - 1. \end{aligned} \tag{5.11}$$

Harper [25] showed that $\sum_k S(n, k) x^k$ has distinct nonpositive roots, and that (5.11) goes to infinity, which is sufficient for the asymptotic normality of the Stirling numbers of the second kind. In showing the former, Harper observed that the function $H_n(x) = e^x R_n(x)$ has the same roots as $R_n(x)$ and by (5.3), $H_n(x) = x H'_{n-1}(x)$ as follows.

$$x H'_{n-1}(x) = x \frac{d}{dx} \left(e^x R_{n-1}(x) \right) = x e^x \Big(R_{n-1}(x) + R'_{n-1}(x) \Big) = e^x R_n(x) = H_n(x).$$

$R_n(x)$ is a polynomial of degree n with a leading coefficient of one, so $R_n(x)$ and $H_n(x)$ have at most n different real roots. By induction on n we can see that $H_n(x)$ has precisely n different nonpositive real roots, one of which is $x = 0$. For $n = 1$, we have $H_1(x) = e^x R_1(x) = (e^x)(x)$ has one root at $x = 0$. Let $n \geq 2$. Then since $xH'_{n-1}(x) = H_n(x)$, the real roots of $H_n(x)$ are $x = 0$ and the roots of $H'_{n-1}(x)$. Assume by the induction hypothesis that the real roots of H_{n-1} are $0 = \alpha_0 > \alpha_2 > \cdots > \alpha_{n-2}$. By Rolle's Theorem, H'_{n-1} has at least one root between any two consecutive roots of H_{n-1}. Since $H_{n-1}(\alpha_{n-2}) = 0 = \lim_{x\to-\infty} H_{n-1}(x)$ and $H_{n-1}(x)$ is continuous and nonzero on $(-\infty, \alpha_{n-2})$, $H'_{n-1}(x)$ has a root $\beta_{n-1} \in (-\infty, \alpha_{n-2})$. Therefore H'_{n-1} has $n-1$ different negative roots, so H_n has n different nonpositive real roots, one of which is $x = 0$.

The SLC property of $S(n,k)$ implies the SLC property and unimodality of $F(n,k)$. Consequently, the $F(n,k)$ array is also asymptotically normal, in the sense of both the central and local limit theorems, with

$$\mathcal{E}(F(n,.)) = n + 1 - \mathcal{E}(S(n,.))$$

and

$$\mathcal{D}(F(n,.)) = \mathcal{D}(S(n,.)).$$

5.3 Asymptotics for Bell numbers

An asymptotic formula for the Bell numbers, in terms of the solution of the unique real solution of the equation $re^r = n$, was obtained by Moser and Wyman [35]:

$$B_n \sim (r+1)^{-\frac{1}{2}} e^{n(r+r^{-1}-1)-1} \left(1 - \frac{r^2(2r^2+7r+10)}{24n(r+1)^3}\right).$$

Iteration gives

$$r = r(n) = \ln n - \ln\ln n + O(1).$$

The function $r(n)$ is also known as the Lambert function and is also denoted by $LambertW(n)$. The explicit form of their result is not convenient to obtain asymp-

totics for the expectation and the variance, as r will vary with n. Canfield and Harper [6], and Canfield [5] made minor modifications on the proof of Moser and Wyman [35] to develop an estimate for B_{n+h}, which holds uniformly for $h = O(\ln n)$, using a *single* $r = r(n)$ value, as $n \to \infty$:.

$$\begin{aligned} B_{n+h} &= \frac{(n+h)!}{r^{n+h}} \frac{e^{e^r-1}}{(2\pi B)^{1/2}} \\ &\times \left(1 + \frac{P_0 + hP_1 + h^2P_2}{e^r} + \frac{Q_0 + hQ_1 + h^2Q_2 + h^3Q_3 + h^4Q_4}{e^{2r}} \right. \\ &+ \left. O\Big(e^{-3r}\Big)\right), \end{aligned} \tag{5.12}$$

where $B = (r^2 + r)e^r$, P_i and Q_i are explicitly known rational functions of r. We list and use in the Maple worksheet B their exact values from Canfield [3]. Using those, formula (5.12) provides asymptotics for $\mathcal{E}(S(n,.))$ and $\mathcal{D}(S(n,.))$, as in [3] (note that [3] only claimed $O(r/n)$ error term in (5.14)):

$$\mathcal{E}(S(n,.)) = \frac{n}{r} - 1 + \frac{r}{2(r+1)^2} + O\Big(\frac{1}{n}\Big). \tag{5.13}$$

$$\mathcal{D}^2(S(n,.)) = \frac{n}{r(r+1)} + \frac{r(r-1)}{2(r+1)^4} - 1 + O\Big(\frac{1}{n}\Big). \tag{5.14}$$

With symbolic calculations Salvy and Shackell [37] obtained the following asymptotics *just* in terms of n, with a compromise at the error term:

$$\mathcal{E}(S(n,.)) = \frac{n}{\ln n} + \frac{n(\ln\ln n + O(1/\ln n))}{\ln^2 n}, \tag{5.15}$$

$$\mathcal{D}^2(S(n,.)) = \frac{n}{\ln^2 n} + \frac{n(2\ln\ln n - 1 + O(1/\ln n))}{\ln^3 n}. \tag{5.16}$$

Chapter 6

Asymptotics for rooted phylogenetic trees

6.1 Set partitions corresponding to phylogenetic trees

We now turn our attention to rooted phylogenetic trees.

In Chapter 5.1 we discussed the Erdős and Székely [13] bijection between the trees counted by $F(n,k)$ and partitions of an n-element set into $n-k+1$ classes, under which the number of children of each of the non-leaf vertices corresponds to class sizes in the partition. As mentioned in the previous chapter, *phylogenetic trees* are precisely the leaf-labeled trees where every non-leaf vertex has at least two children. Let $F^\star(n,k)$ denote the number of phylogenetic trees with k leaves and n non-root vertices and $S^\star(n,k)$ denote the number partitions of an n element set into k classes such that each class contains at least two elements. The bijection still provides $F^\star(n,k) = S^\star(n, n-k+1)$ and $S^\star(n,i) = F^\star(n, n-i+1)$. Any information available on the array $S^\star(n,k)$ translates to information on the array $F^\star(n,k)$. In this section we will prove central and local limit theorems for $S^\star(n,k)$ (Theorem 6.7) which translate into such theorems for $F^\star(n,k)$, with $\mathcal{E}(F^\star(n,.)) = n+1-\mathcal{E}(S^\star(n,.))$ and $\mathcal{D}(F^\star(n,.)) = \mathcal{D}(S^\star(n,.))$.

First we derive a bivariable generating function (which is neither completely exponential nor completely ordinary). To this end, weight the partitions as follows: Assign to a partition class of size k the weight x^k, and to the entire partition the product of the weight of its partition classes. In particular, the counts of the number of partitions that contain only singleton classes are $S(n,n) = 1$. The weight of such

a partition on $[n]$ is x^n, since the partition must have n singleton classes. The exponential generating function of the weighted partitions that contain singleton classes only is

$$\sum_{n=0}^{\infty} S(n,n)x^n \frac{t^n}{n!} = \sum_{n=0}^{\infty} \frac{(xt)^n}{n!} = e^{tx}. \tag{6.1}$$

Now consider all weighted partitions, regardless of class sizes. Every weighted partition can be identified with a pair of (possibly empty) partitions on a pair of disjoint underlying sets: the first partition has only singleton classes and covers some (possibly empty) subset A of $[n]$, the second partition covers the remaining set $[n] \setminus A$ and has no singleton classes. Using equations (5.1), (6.1) and the multiplication rule of EGF's (see claim 1.40), we obtain that the EGF of weighted partitions is

$$e^{tx} \sum_n \sum_k S^\star(n,k)x^k \frac{t^n}{n!} = \sum_n \sum_k S(n,k)x^k \frac{t^n}{n!} = e^{x(e^t-1)},$$

or

$$\sum_n \sum_k S^\star(n,k)x^k \frac{t^n}{n!} = e^{-tx} \sum_n \sum_k S(n,k)x^k \frac{t^n}{n!} = e^{-tx} \cdot e^{x(e^t-1)}$$

At this point we have the mixed bivariate generating function

$$\sum_n \sum_k S^\star(n,k)x^k \frac{t^n}{n!} = e^{x(e^t-t-1)}. \tag{6.2}$$

Inclusion-exclusion or (6.2) implies that

$$S^\star(n,k) = \sum_{\ell=0}^{n} (-1)^\ell \binom{n}{\ell} S(n-\ell, k-\ell).$$

After substituting $1/x$ into x, tx into t, and multiplication by x into equation (6.2), we obtain

$$\sum_n \sum_i F^\star(n,i)x^i \frac{t^n}{n!} = xe^{(e^{tx}-tx-1)/x}$$

as shown below. Since $F^\star(n,k) = 0$ when $\min(n,k) = 0$ and $k \neq 1$, we have

$\sum_{(n,k):\min(n,k)=0} F^\star(n,k)x^k\frac{t^n}{n!} = x$. Also, $\sum_{(n,k):\min(n,k)=0} S^\star(n,k)x^k\frac{t^n}{n!} = 1$. Thus

$$\begin{aligned}
x\left(e^{\frac{e^{tx}-tx-1}{x}}-1\right) &= x\left(\sum_{n=0}^{\infty}\sum_{k=0}^{n} S^\star(n,k)x^{-k}\frac{(tx)^n}{n!} - \sum_{(n,k):\min(n,k)=0} S^\star(n,k)x^k\frac{t^n}{n!}\right)\\
&= x\sum_{n=1}^{\infty}\sum_{k=1}^{n} S^\star(n,k)x^{-k}\frac{(tx)^n}{n!} = \sum_{n=1}^{\infty}\sum_{k=1}^{n} S^\star(n,k)x^{n-k+1}\frac{(t)^n}{n!}\\
&= \sum_{n=1}^{\infty}\sum_{j=1}^{n} S^\star(n,n-j+1)x^j\frac{(t)^n}{n!} = \sum_{n=1}^{\infty}\sum_{j=1}^{n} F^\star(n,j)x^j\frac{(t)^n}{n!}\\
&= \left(\sum_{n=0}^{\infty}\sum_{j=0}^{n} F^\star(n,j)x^j\frac{(t)^n}{n!}\right) - \left(\sum_{(n,k):\min(n,k)=0} F^\star(n,k)x^k\frac{t^n}{n!}\right)\\
&= \left(\sum_{n=0}^{\infty}\sum_{j=0}^{n} F^\star(n,j)x^j\frac{(t)^n}{n!}\right) - x
\end{aligned}$$

Define $B_n^\star = \sum_k S^\star(n,k)$; this is the number of all partitions of an n-element set which do not contain singleton classes [41] A000296 in *The On-Line Encyclopedia of Integer Sequences* [41]. Then the exponential generating function of the counts $B_n^\star$ is

$$\sum_n B_n^\star\frac{t^n}{n!} = e^{e^t-t-1} = 1+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{4t^4}{4!}+\frac{11t^5}{5!}+....$$

Becker [1] observed that

$$B_n = B_{n+1}^\star + B_n^\star. \tag{6.3}$$

This identity can be shown as follows. Given a partition of $[n]$, either the partition has no singleton sets in which case it is counted in $B_n^\star$, or it contains at least one singleton class. In the later case, there is a bijection between these partitions and partitions without singleton classes of an $(n+1)$-element set where a new class has been built with all the element of all singletons with the addition of $n+1$. These sets are counted by $B_{n+1}^\star$.

Using Claim 1.39, the generating function proof of identity (6.3) is simply

$$e^{e^t-1} = \frac{d}{dt}(e^{e^t-t-1}) + e^{e^t-t-1}.$$

From $B_i = B_i^\star + B_{i+1}^\star$ for $i = 1,2,\ldots,n$, and $B_0^\star = 1$, we obtain $\sum_{i=1}^n B_i(-1)^{n-i} = B_{n+1}^\star + (-1)^{n-1}B_0^\star$. As the B_n sequence is strictly increasing, we immediately obtain

the following: $B_t - B_{t-1} < \sum_{i=1}^{t} B_i(-1)^{t-i} < B_t$ for $t > 4$, and with $t = n - h$ the asymptotical formula

$$B^{\star}_{n+1} = B_n - B_{n-1} + \ldots + (-1)^h B_{n-h} + O(B_{n-h-1}). \tag{6.4}$$

In the special case $h = 0$, using (5.12), we obtain:

$$B^{\star}_{n+1} = B_n - O(B_{n-1}) = B_n\Big(1 - O\big(\frac{r}{n}\big)\Big). \tag{6.5}$$

The following recurrence relation

$$S^{\star}(n,k) = (n-1)S^{\star}(n-2,k-1) + kS^{\star}(n-1,k) \tag{6.6}$$

can be easily seen by considering the placement of the n^{th} element in any partition counted by $S^{\star}(n,k)$. If the n^{th} element is not in a partition class of size two, then it can be removed and the resulting partition is counted in $S^{\star}(n-1,k)$. There are k classes in this count that could contain the n^{th} element. If the n^{th} element is in a partition class of size two, the removal of that class results in a partition of $n-2$ elements into $k-1$ partition classes. There are $n-1$ elements that could have been paired with n. Notice that the recursion drops back two steps.

We define the polynomial sequence $S_n(x) = \sum_k S^{\star}(n,k)x^k$. It is easy to see that $S_1(x) = 0$, $S_2(x) = x$, and for $n \geq 3$ equation (6.6) gives

$$S_n(x) = (n-1)xS_{n-2}(x) + xS'_{n-1}(x). \tag{6.7}$$

It is useful to note that the polynomial $S_i(x)$ has zero constant term, and for all $1 \leq k \leq \deg(S_i(x))$ the coefficient $S(i,k)$ is positive.

Induction immediately gives the following lemma.

Lemma 6.1. *For $n \geq 2$, $S'_n(0) > 0$, the degree of $S_n(x)$ is $\deg(S_n(x)) = \left\lfloor \frac{n}{2} \right\rfloor$, and the root 0 has multiplicity one.*

Proof. Since $S'_n(0) = S^{\star}(n,1) > 0$ for $n \geq 0$, the first part of the claim is true.

For $n = 2, 3$, $S_2(x) = S_3(x) = x$ has degree $1 = \lfloor \frac{2}{2} \rfloor = \lfloor \frac{3}{2} \rfloor$ and the polynomial has 0 as a root of multiplicity one. Assume the statement is true for $n \leq k$ and consider $S_{k+1}(x)$. By the induction hypothesis, $xkS_{k-1}(x)$ has degree $\lfloor \frac{k-1}{2} \rfloor + 1 = \lfloor \frac{k+1}{2} \rfloor$, and $xS'_k(x)$ has degree $\lfloor \frac{k}{2} \rfloor - 1 + 1 \leq \lfloor \frac{k+1}{2} \rfloor$. Since the leading coefficients of both of these polynomials are positive, regardless of the parity of k the polynomial $S_{k+1}(x) = xkS_{k-2}(x) + xS'_{k-1}(x)$ has degree $\lfloor \frac{k+1}{2} \rfloor$. By the induction hypothesis, 0 is a root of $S_k(x)$ of multiplicity one. The constant term of S'_k is positive by the first, already proven part of this lemma, therefore no power of x divides $kS_{k-1}(x) + S'_k(x)$. Since $S_{k+1}(x) = x\left(kS_{k-1}(x) + S'_k(x)\right)$, we have that x^2 is a not factor of $S_{k+1}(x)$, and the root $x = 0$ has multiplicity one. □

To be able to refer to the roots of $S_n(x)$ in order, we will introduce the following notation

Notation 6.2. *The $\lfloor \frac{n}{2} \rfloor$ roots of $S_n(x)$ are denoted by*

$$\gamma_1^{(n)} \leq \gamma_2^{(n)} \leq \cdots \leq \gamma_{\lfloor \frac{n}{2} \rfloor}^{(n)}$$

We will also use

Notation 6.3. *For a real number r*

$$\operatorname{sgn}(r) = \begin{cases} 1, & \textit{if } r{>}0 \\ 0 & \textit{if } r{=}0 \\ -1 & \textit{otherwise.} \end{cases}$$

It is easy to see that for real numbers a, b we have $\operatorname{sgn}(ab) = \operatorname{sgn}(a)\sin(b)$.

6.2 The roots of the polynomial $S_n(x)$.

In order to use Harper's method, we need to show that the roots of $S_n(x)$ are non-positive real numbers and that every root occurs with multiplicity one. This section is devoted to the task.

The following lemma must be divided into two cases, as depending on the parity of n, the number of roots of $S_n(x)$ and $S_{n+1}(x)$ may or may not be the same.

Lemma 6.4. *Let $k \geq 2$ be an integer. Then the following are true:*
First, if the roots of $S_{2k-2}(x)$ and $S_{2k-1}(x)$ occur with multiplicity one and satisfy

$$\gamma_1^{(2k-2)} < \gamma_1^{(2k-1)} < \gamma_2^{(2k-2)} < \gamma_2^{(2k-1)} < \cdots < \gamma_{k-2}^{(2k-1)} < \gamma_{k-1}^{(2k-2)} = 0 = \gamma_{k-1}^{(2k-1)},$$

then the roots, $\{\gamma_i^{(2k)}\}$ of $S_{2k}(x)$ satisfy

$$\gamma_1^{(2k)} < \gamma_1^{(2k-1)} < \gamma_2^{(2k)} < \gamma_2^{(2k-1)} < \cdots < \gamma_{k-1}^{(2k)} < \gamma_{k-1}^{(2k-1)} = 0 = \gamma_k^{(2k)}.$$

Second, if the roots of $S_{2k-1}(x)$ and $S_{2k}(x)$ occur with multiplicity one and satisfy

$$\gamma_1^{(2k)} < \gamma_1^{(2k-1)} < \gamma_2^{(2k)} < \gamma_2^{(2k-1)} < \cdots < \gamma_{k-1}^{(2k)} < \gamma_{k-1}^{(2k-1)} = 0 = \gamma_k^{(2k)}$$

then the $\{\gamma_i^{2k+1}\}$ roots of S_{2k+1} satisfy

$$\gamma_1^{(2k)} < \gamma_1^{(2k+1)} < \gamma_2^{(2k)} < \gamma_2^{(2k+1)} < \cdots < \gamma_{k-2}^{(2k+1)} < \gamma_{k-1}^{(2k)} < \gamma_{k-1}^{(2k+1)} < \gamma_k^{(2k)} = 0 = \gamma_k^{(2k+1)}.$$

Proof. In proving the first statement, our initial goal will be to show that under the assumption $S_{2k}(x)$ has a root in the interval $(\gamma_i^{(2k-1)}, \gamma_{i+1}^{(2k-1)})$ for each $i \in [k-2]$. Since $S_{2k}(x)$ has k roots, one of which is 0, all that will remain to show is that $S_{2k}(x)$ has a root that is less than $\gamma_1^{(2k-1)}$. To achieve this goal, it is enough to show that for each $i \in [k-1]$ we have

$$\operatorname{sgn}\left((2k-1)S_{2k-2}(\gamma_i^{(2k-1)}) + S'_{2k-1}(\gamma_i^{(2k-1)})\right) = (-1)^{k-1-i}, \tag{6.8}$$

since using Rolle's Theorem and equation (6.7) we get that $\frac{S_{2k}(x)}{x}$ has a root in the interval $(\gamma_i^{(2k-1)}, \gamma_{i+1}^{(2k-1)})$ for each $i \in [k-2]$. We determine the right side of equation (6.8) as follows. We know that $S'_{2k-1}(x)$ is a polynomial of degree $k-2$ with exactly one root between the $k-1$ distinct consecutive roots of $S_{2k-1}(x)$, therefore

we must have $\operatorname{sgn}\left(S'_{2k-1}(\gamma_i^{(2k-1)})\right) = -\operatorname{sgn}\left(S'_{2k-1}(\gamma_{i+1}^{(2k-1)})\right)$ for $1 \leq i \leq k-2$. Recall (Lemma 6.1) that $S'_{2k-1}(\gamma_{k-1}^{(2k-1)}) = S'_{2k-1}(0) > 0$. Therefore, $\operatorname{sgn}\left(S'_{2k-1}(\gamma_{k-1}^{(2k-1)})\right) = 1$ and

$$\operatorname{sgn}\left(S'_{2k-1}(\gamma_i^{(2k-1)})\right) = (-1)^{k-1-i} \text{ for each } i \in [k-1]. \tag{6.9}$$

Observe that $\operatorname{sgn}\left(S_{2k-2}(\gamma_i^{(2k-1)})\right) = -\operatorname{sgn}\left(S_{2k-2}(\gamma_{i+1}^{(2k-1)})\right)$ for $1 \leq i \leq k-3$, since by the hypothesis, for these values of i the polynomial $S_{2k-2}(x)$ has exactly one root in the interval $\left(\gamma_i^{(2k-1)}, \gamma_{i+1}^{(2k-1)}\right)$. The polynomial $S_{2k-2}(x)$ has positive coefficients and $k-1$ non positive roots, with $S_{2k-2}(\gamma_{k-1}^{(2k-1)}) = 0$. We know that $S'_{2k-2}(0) > 0$ and that $S_{2k-2}(x)$ has no roots between the roots $\gamma_{k-1}^{(2k-2)} = 0$ and $\gamma_{k-2}^{(2k-2)}$. Therefore, since $\gamma_{k-2}^{(2k-1)} \in (\gamma_{k-2}^{(2k-2)}, \gamma_{k-1}^{(2k-2)})$, we must have that $\operatorname{sgn}\left(S_{2k-2}(\gamma_{k-2}^{(2k-1)})\right) = -1$, which implies that

$$\operatorname{sgn}\left(S_{2k-2}(\gamma_i^{(2k-1)})\right) = (-1)^{k-1-i} = \operatorname{sgn}\left(S'_{2k-1}(\gamma_i^{(2k-1)})\right) \text{ for all } i \in [k-2]. \tag{6.10}$$

The required equation (6.8) now follows from the facts that $2k-1 > 0$, equations (6.9) and (6.10), and the fact that $\operatorname{sgn}\left(S_{2k-2}(\gamma_{k-1}^{(2k-1)})\right) = 0$.

It remains to be shown that $\frac{S_{2k}(x)}{x}$ (and consequently $S_{2k}(x)$) changes sign, and therefore has a root in$\left(-\infty, \gamma_1^{(2k-1)}\right)$. Since the degree of S_{2k-2} is greater than the degree of S'_{2k-1}, by equations (6.7) and (6.10), it is enough to show that S_{2k-2} changes sign in this interval. However, this follows from the fact that $\gamma_1^{(2k-2)} \in \left(-\infty, \gamma_1^{(2k-1)}\right)$.

In proving the second statement, we will show that under the assumption, $S_{2k+1}(x)$ has a root in the interval $(\gamma_i^{(2k)}, \gamma_{i+1}^{(2k)})$ for each $i \in [k-1]$. Since $S_{2k+1}(x)$ has k roots, one of which is 0, this achieves our goal. For this, it is enough to show that for each $i \in [k]$ we have

$$\operatorname{sgn}\left(2kS_{2k-1}(\gamma_i^{(2k)}) + S'_{2k}(\gamma_i^{(2k)})\right) = (-1)^{k-i}, \tag{6.11}$$

since using Rolle's Theorem and equation (6.7) we know that $\frac{S_{2k+1}(x)}{x}$ has a root in the interval $(\gamma_i^{(2k)}, \gamma_{i+1}^{(2k)})$ for each $i \in [k-1]$. We determine the right side of equation (6.11) as in the previous case. We know that $S'_{2k}(x)$ is a polynomial of degree $k-1$

with exactly one root between the k distinct consecutive roots of $S_{2k}(x)$. Therefore we must have $\operatorname{sgn}\left(S'_{2k}(\gamma_i^{(2k)})\right) = -\operatorname{sgn}\left(S'_{2k}(\gamma_{i+1}^{(2k)})\right)$ for $1 \leq i \leq k-1$. Recall (Lemma 6.1) that $S'_{2k}(\gamma_k^{(2k)}) = S'_{2k}(0) > 0$. Thus, $\operatorname{sgn}\left(S'_{2k}(\gamma_k^{(2k)})\right) = 1$ and

$$\operatorname{sgn}\left(S'_{2k}(\gamma_i^{(2k)})\right) = (-1)^{k-i} \text{ for each } i \in [k]. \tag{6.12}$$

Observe that $\operatorname{sgn}\left(S_{2k-1}(\gamma_i^{(2k)})\right) = -\operatorname{sgn}\left(S_{2k-1}(\gamma_{i+1}^{(2k)})\right)$ for $1 \leq i \leq k-2$, since by the hypothesis, for these values of i the polynomial $S_{2k-1}(x)$ has exactly one root in the interval $\left(\gamma_i^{(2k)}, \gamma_{i+1}^{(2k)}\right)$. The polynomial $S_{2k-1}(x)$ has positive coefficients and $k-1$ non positive roots, with $S_{2k-1}(\gamma_k^{(2k)}) = 0$. By hypothesis, $S_{2k-1}(x)$ has no roots between the roots $\gamma_{k-1}^{(2k-1)} = 0$ and $\gamma_{k-2}^{(2k-1)}$. Furthermore $S'_{2k-1}(0) > 0$ and, since $\gamma_{k-1}^{(2k)} \in (\gamma_{k-2}^{(2k-1)}, \gamma_{k-1}^{(2k-1)})$, we must have that $\operatorname{sgn}\left(S_{2k-1}(\gamma_{k-1}^{(2k)})\right) = -1$. This implies that

$$\operatorname{sgn}\left(S_{2k-1}(\gamma_i^{(2k)})\right) = (-1)^{k-i} = \operatorname{sgn}\left(S'_{2k}(\gamma_i^{(2k)})\right) \text{ for all } i \in [k-1]. \tag{6.13}$$

The required equation (6.11) now follows from the facts that $2k > 0$, equations (6.12) and (6.13), and the fact that $\operatorname{sgn}\left(S_{2k-1}(\gamma_k^{(2k)})\right) = 0$. □

Lemma 6.5. *Let $n \geq 2$ be an integer. The roots of $S_n(x)$ are non positive real numbers each of which occurs with multiplicity one. Furthermore, for $k \geq 2$ the roots of $S_{2k}(x)$ and S_{2k-1} satisfy the following inqualities:*

$$\gamma_1^{(2k)} < \gamma_1^{(2k-1)} < \gamma_2^{(2k)} < \gamma_2^{(2k-1)} < \cdots < \gamma_{k-1}^{(2k)} < \gamma_{k-1}^{(2k-1)} = 0 = \gamma_k^{(2k)}.$$

while the roots of $S_{2k}(x)$ and S_{2k+1} satisfy

$$\gamma_1^{(2k)} < \gamma_1^{(2k+1)} < \gamma_2^{(2k)} < \gamma_2^{(2k+1)} < \cdots < \gamma_{k-2}^{(2k+1)} < \gamma_{k-1}^{(2k)} < \gamma_{k-1}^{(2k+1)} < \gamma_k^{(2k)} = 0 = \gamma_k^{(2k+1)}.$$

Proof. We will show this for all $S_n(x)$ by induction on n.

The lemma is vacuously true for $S_2(x) = S_3(x) = x$. The roots of $S_4(x) = 3x^2 + x$ are $\gamma_1^{(4)} = \frac{-1}{3}$ and $\gamma_2^{(4)} = 0$, are ordered as stated, satisfying the lemma. The roots of $S_5(x) = 10x^2 + x$ are $\gamma_1^{(5)} = \frac{-1}{10}$ and $\gamma_2^{(5)} = 0$ also satisfying the lemma.

Let $n \geq 4$. and assume that the statement is true for all $S_m(x)$ where $2 \leq m \leq n-1$.

If $n = 2k$ for some integer k, then the statement follows from the induction hypothesis and the first part of Lemma 6.4.

If $n = 2k+1$, for some integer k, then the statement follows from the induction hypothesis and the second part of Lemma 6.4. □

Let the roots of $S_n(x)$ be $\{-y_{nk} : k = 1, 2, \ldots, \lfloor n/2 \rfloor\}$. Define the independent random variables Y_{nk} by $\mathcal{P}(Y_{nk} = 0) = y_{nk}/(1+y_{nk})$ and $\mathcal{P}(Y_{nk} = 1) = 1/(1+y_{nk})$. Set $W_n = \sum_k Y_{nk}$. We have for the expectation and variance, from (5.4), using (6.7) repeatedly,

$$\begin{aligned}
\mathcal{E}(W_n) &= \frac{B^\star_{n+1}}{B^\star_n} - n\frac{B^\star_{n-1}}{B^\star_n}; \\
\mathcal{D}^2(W_n) &= \frac{B^\star_{n+2}}{B^\star_n} + 2n\frac{B^\star_{n+1}B^\star_{n-1}}{(B^\star_n)^2} + n(n-1)\frac{B^\star_{n-2}}{B^\star_n} \\
&- \Big(\frac{B^\star_{n+1}}{B^\star_n}\Big)^2 - n^2\Big(\frac{B^\star_{n-1}}{B^\star_n}\Big)^2 - n\frac{B^\star_{n-1}}{B^\star_n} - (2n+1).
\end{aligned}$$

Lemma 6.6. *We have the asymptotic formulae*

$$\begin{aligned}
\mathcal{E}(W_n) &= \frac{n}{r} - r - \frac{1}{2r} + \frac{1}{2r(r+1)^2} + O\Big(\frac{1}{n}\Big), \\
\mathcal{D}^2(W_n) &= \frac{n}{r(r+1)} - r + 1 - \frac{2}{r+1} - \frac{1}{2(r+1)^2} - \frac{1}{2(r+1)^3} + \frac{1}{(r+1)^4} + O\Big(\frac{1}{n}\Big).
\end{aligned}$$

Proof. We started with the closed forms above, used (6.4) to substitute the $B^\star$ numbers, and then substituted the B numbers with (5.12), changed e^{-r} to r/n, using Maple. For details, see the Maple worksheet. □

Note that $\mathcal{E}(W_n) - \mathcal{E}(Z_n) = O(r)$ and $\mathcal{D}^2(W_n) - \mathcal{D}^2(Z_n) = O(r)$, where Z_n still denotes the random variable associated with the Bell numbers in Section 5.2. It follows from these remarkably small differences that (5.15) and (5.16) still hold when Z_n is changed to W_n.

Theorem 6.7. *For the sequence $A(n,j) = S^\star(n,j)$ the central limit theorem 5.5) and the local limit theorem (5.8) holds with $E_n = B_n^\star$. Furthermore, the number $k = J_n$ that maximizes $S^\star(n,k)$ satisfies*

$$J_n = \frac{n}{r} + o(\frac{\sqrt{n}}{r})$$

and

$$S^\star(n, J_n) = \frac{rB_{n-1}}{\sqrt{2n\pi}}(1 + o(1)).$$

Proof. The central and local limit theorems hinge on $\mathcal{D}(W_n) \to \infty$ that we have from Lemma 6.6. The arguments leading to (5.9) and (5.10) hold for $S^\star(n,k)$ instead of $S(n,k)$. B_n^* is approximated with B_{n-1} by (6.5). □

We obtain for free the asymptotically normal distribution of $F^\star(n,k)$. Defining a random variable Y_n with $\mathcal{P}(Y_n = j) = F^\star(n,j)/B_n^\star = \mathcal{P}(W_n = n - j + 1)$, we have $\mathcal{E}(Y_n) = n + 1 - \mathcal{E}(W_n) = n - n/r + r + 1 + o(1)$ and $\mathcal{D}^2(Y_n) = \mathcal{D}^2(W_n)$, and we have the asymptotic normality results on the $F^\star(n,k)$ numbers instead of $F(n,k)$, with $B_n^\star$ instead of B_n.

Felsenstein [15, 16], and also Foulds and Robinson [18] investigated the numbers $T_{n,m}$. $T_{n,m}$ is the number of rooted phylogenetic trees with n labeled leaves, m unlabeled internal vertices (the root, if it is not a leaf, is one of them). Clearly, for $m \geq 2$ we have

$$T_{n,m} = F^{\star}(n+m-1, n) = S^{\star}(n+m-1, m). \tag{6.14}$$

If we are interested only in evaluating certain $T_{n,m}$ numbers, the results in Section 6.7 would suffice. However, as the $T_{n,m}$ notation suggests, the distributions of $F(n,k)$ and $F^{\star}(n,k)$ studied in Sections 5.1, and 6.7 for large but fixed number of vertices n and varying number of leaves k, albeit is mathematically interesting, is not really relevant for phylogenetics. The relevant distribution for phylogenetics is large but fixed number of leaves and varying number of internal vertices, with which total number of vertices must vary as well. Let $t_n = \sum_k T_{n,k}$ denote the number of all phylogenetic trees with n labeled leaves. This sequence is A000311 in *The On-Line Encyclopedia of Integer Sequences* [41], which is the solution to Schroeder's fourth problem [38].

Felsenstein [16, 15] proved the recurrence relation

$$T_{n,k} = (n+k-2)T_{n-1,k-1} + kT_{n-1,k} \tag{6.15}$$

for $k > 1$ with the initial condition $T_{n,1} = 1$ for $n > 1$. Let T' be a [phylogenetic tree with n leaves (and label set $[n]$). The removal of the leaf labeled n will result in a phylogenetic tree with $n-1$ leaves if n is a child of a vertex of T' that has at least two more children. If n is a child of a vertex of T' that has just one other child than the removed leaf, then the removal of n results either in a tree that can be obtained by subdividing an edge of a phylogenetic tree with $n-1$ leaves (and the subdividing vertex is the parent of n in T', which is not a root), or a tree that can be obtained from a rooted phylogenetic tree with $n-1$ leaves by adding a new root of degree

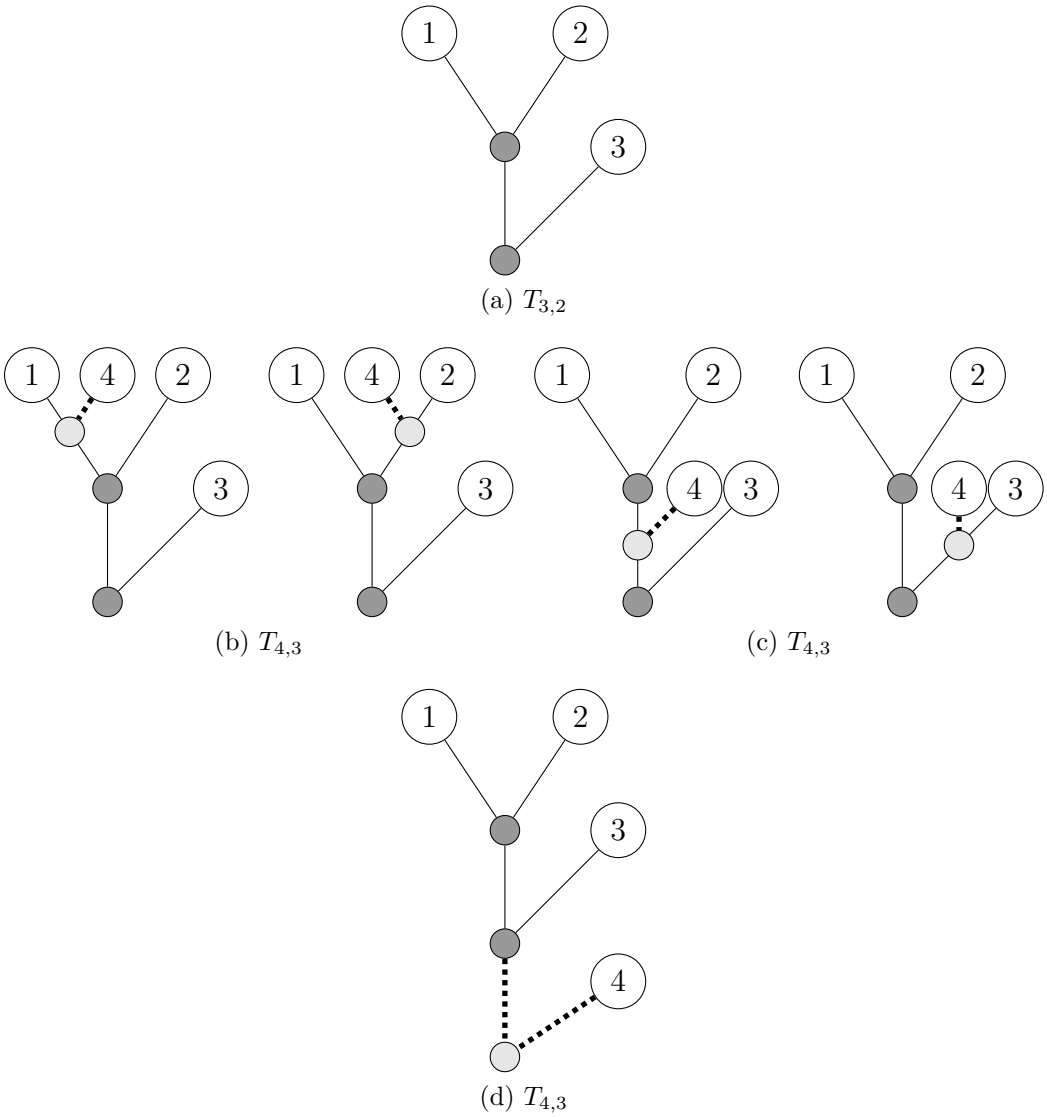
1 to the old root (and the new root is the parent of the removed leaf). Using this logic, we can obtain this recurrence relation by considering the addition of an n^{th} leaf to an already existing tree with $n-1$ leaves. There are k ways to add a new leaf labeled n as a child of an existing internal vertex of a rooted phylogenetic tree T with k internal vertices, and this takes care of the second term of the right hand side of equation (6.15). All other cases that we need to take care of change the number of internal vertices. Fix a rooted phylogenetic tree T with $n-1$ leaves (and label set $[n-1]$), and assume it has $k-1$ internal vertices. There are $n+k-3$ ways to add a leaf labeled n by subdividing an edge of T with an additional (internal) vertex an make this new leaf the child of the subdividing vertex. The n^{th} leaf can also be added to T by adding a root and two edges; one edge between the new and old root and on edge between the new root and the n^{th} leaf, which takes care of the first term of (6.15). See figure 6.1 for an example using $T_{4,2}$

Consider the polynomials $P_n(x) = \sum_k T_{n+1,k}x^k$. Then $P_n(1) = t_{n+1}$ and the degree of $P_n(x)$ is n. Felsenstein's recurrence relation (6.15) implies the identity

$$P_n(x) = nxP_{n-1}(x) + (x+x^2)P'_{n-1}(x) \tag{6.16}$$

with initial terms $P_0(x) = 1$, $P_1(x) = T_{2,1}x = x$, $P_2(x) = 3x^2 + x$, and $P_3(x) = 15x^3 + 10x^2 + x$. We show this identity as follows. For $n \geq 2$,

$$\begin{aligned} P_{n-1} &= \sum_{k=1}^{n-1} T_{n,k}x^k \text{ so:} \\ nxP_{n-1} &= \sum_{k=1}^{n-1} nT_{n,k}x^{k+1} = \sum_{k=2}^{n} nT_{n,k-1}x^k \end{aligned}$$



(a) $T_{3,2}$

(b) $T_{4,3}$

(c) $T_{4,3}$

(d) $T_{4,3}$

Figure 6.1: (a) The original $T_{3,2}$ tree. (b Adding an internal vertex an leaf by subdividing the edges adjacent to existing leaves. (c) Adding an internal vertex and leaf by subdividing the edges between non-leaf vertices. (d) Adding one non-leaf and one leaf vertex by re-rooting the tree at the new non-leaf vertex.

Also,

$$\begin{aligned} P'_{n-1}(x) &= \sum_{k=1}^{n-1} kT_{n,k}x^{k-1} \text{ so:} \\ (x+x^2)P'_{n-1}(x) &= \sum_{k=1}^{n-1} kT_{n,k}\left(x^k + x^{k+1}\right) \\ &= x + x^2 + 2T_{n,2}(x^2+x^3) + 3T_{n,3}(x^3+x^4) + \dots \\ &= x + \sum_{k=2}^{n} \left(kT_{n,k} + (k-1)T_{n,k-1}\right) x^k \end{aligned}$$

Now, using these with the recursion (6.15) one easily obtains

$$\begin{aligned} P_n(x) &= \sum_{k=1}^{n} T_{n+1,k}x^k \\ &= T_{n+1,1}x + T_{n+1,2}x^2 + T_{n+1,}x^3 + \dots \\ &= x + \sum_{k=2}^{n} \left((n+k-1)T_{n,k-1} + kT_{n,k}\right) x^k \\ &= \sum_{k=2}^{n} nT_{n,k-1}x^k + x + \sum_{k=2}^{n} \left(kT_{n,k} + (k-1)T_{n,k-1}\right) x^k \\ &= nxP_{n-1} + (x+x^2)P'_{n-1}(x) \end{aligned}$$

Theorem 6.8. *For $n \geq 1$, the polynomial $P_n(x)$ has n distinct real roots, one of them is zero, and the other $n-1$ roots are in the open interval $(-1, 0)$.*

Proof. We prove the theorem with mathematical induction on n. The small cases ($n \leq 2$) above are easy to verify. It is easy to see (by a different induction) that $P_1(-1) = -1$ and from (6.16), $P_n(-1) = (-n)P_{n-1}(-1)$, thus

$$\operatorname{sgn}(P_n(-1)) = (-1)^n. \tag{6.17}$$

So assume that $n \geq 2$, and, using the induction hypothesis, let the roots of $P_n(x)$ be

$$-1 < \alpha_1 < \cdots < \alpha_{n-2} < \alpha_{n-1} < \alpha_n = 0.$$

By Rolle's theorem, $P'_n(x)$ has a root β_i in (α_i, α_{i+1}) for $i = 1, 2, ..., n-1$. From (6.16), observe that $\operatorname{sgn}(P_{n+1}(\beta_i)) = -\operatorname{sgn}(P_n(\beta_i))$. As the sign of $P_n(x)$ must alternate on the β_i, so must the sign of $P_{n+1}(x)$, and therefore $P_{n+1}(x)$ has a root in (β_i, β_{i+1}) for

$i = 1, 2, ..., n-2$. We have to find 3 more roots: one is $x = 0$, and we will show that the other two are in the intervals $(-1, \beta_1)$ and $(\beta_{n-1}, 0)$, respectively.

Indeed, $\operatorname{sgn}(P_n(x))$ differs in -1 and β_1, since $P_n(x)$ has a single root α_1 between. Also, $\operatorname{sgn}(P_{n+1}(-1)) = -\operatorname{sgn}(P_n(-1))$ by (6.17) and from our earlier observation, $\operatorname{sgn}(P_{n+1}(\beta_1)) = -\operatorname{sgn}(P_n(\beta_1))$. Hence, $\operatorname{sgn}(P_{n+1}(x))$ differs in -1 and β_1, and therefore $P_{n+1}(x)$ has a root in $(-1, \beta_1)$.

Observe (6.16) with induction implies that for $n \geq 1$ the coefficient of x^n in $P_n(x)$ is positive. On one hand, we have that for $x < 0$ but x sufficently close to zero, $\operatorname{sgn}(P_{n+1}(x)) = -1$. On the other hand, $\operatorname{sgn}(P_{n+1}(\beta_1)) = -\operatorname{sgn}(P_{n+1}(-1)) = (-1)^n$, $\operatorname{sgn}(P_{n+1}(\beta_i)) = (-1)^{n+i-1}$, and $\operatorname{sgn}(P_{n+1}(\beta_n)) = 1$. Therefore $P_{n+1}(x)$ has a root in $(\beta_{n-1}, 0)$. □

As $P_n(x)$ has distinct real roots, Lieb's result (5.7) applies and the coefficients of $P_n(x)$ have the SLC property. An alternative way to prove this is the following:

Kurtz [30] studied triangular arrays of numbers defined with a recurrence relation $A(n,k) = f(n,k)A(n-1,k-1)+g(n,k)A(n-1,k)$ with initial conditions $A(1,1) = 1$, $A(n,0) = A(n,n+1) = 0$. He showed that if

$$2f(n,k) - f(n,k-1) - f(n,k+1) \geq 0 \text{ for } 1 < k < n; n = 1, 2, \ldots$$

and

$$2g(n,k) - g(n,k-1) - g(n,k+1) \geq 0 \text{ for } 1 < k < n; n = 1, 2, \ldots,$$

then the $A(n,k)$ array has the SLC property.

Note that the array $A(n,k) = T_{n+1,k}$ satisfies the conditions of Kurtz' result with $f(n,k) = n+k-1$ and $g(n,k) = k$; therefore $A(n,k)$ and $T_{n,k}$ have the SLC property.

Consider the following bivariate generating function for $T_{n,k}$:

$$H(x,z) = \sum_{n\geq 1} \sum_{k} T_{n,k} x^k \frac{z^n}{n!} = \sum_{n\geq 1} P_{n-1}(x) \frac{z^n}{n!},$$

in particular, $H(1,z) = \frac{z}{1!} + \frac{z^2}{2!} + \frac{4z^3}{3!} + \frac{26z^4}{4!} + \ldots$. Flajolet [17] observed the functional equation

$$H(x,z) = z + x\Big(e^{H(x,z)} - 1 - H(x,z)\Big),$$

which immediately follows from the Exponential Formula, and obtained from this equation an expression for $H(1,z)$ in terms of the Lambert function, which is the compositional inverse of xe^{-x}:

$$H(1,z) = -LambertW\Big(-\frac{1}{2}e^{\frac{z-1}{2}}\Big) + \frac{z-1}{2}.$$

He also observed that $H(1,z)$, the EGF of the t_n sequence, has a singularity at $\rho = -1 + 2\log 2$, and it is the only singularity at this radius; and furthermore, for $|z| < \rho$, there is a singular expansion of $H(1,z)$ in terms of $\Delta = \sqrt{1 - z/\rho}$, of which the first few terms are

$$H(1,z) = \log 2 - \sqrt{\rho}\Delta + \Big(\frac{1}{6} - \frac{1}{3}\log 2\Big)\Delta^2 - \frac{\rho^{3/2}}{36}\Delta^3 + O(\Delta^4). \tag{6.18}$$

Flajolet [17] used (6.18) to obtain asymptotic formula for t_n as

$$t_n \sim \frac{n!}{2\sqrt{\pi}n^{3/2}\rho^{n-1/2}},$$

and noted that asymptotic expansion can be obtained by this method. Using Maple, we went further and actually obtained the following asymptotic expansion:

$$t_n \sim \frac{n!}{\sqrt{\pi}\rho^{n-\frac{1}{2}}}\left(\frac{1}{2n^{3/2}} + \frac{3}{16n^{5/2}} + \frac{25}{256n^{7/2}} + O\Big(\frac{1}{n^{9/2}}\Big)\right).$$

The details are on the Maple worksheet in Appendix C.

Let the roots of $P_n(x)$ be $\{-y_{nk} : k = 1, 2, \ldots, n\}$. Define the independent random variables Y_{nk} by $\mathcal{P}(Y_{nk} = 0) = y_{nk}/(1 + y_{nk})$ and $\mathcal{P}(Y_{nk} = 1) = 1/(1 + y_{nk})$. Set $Z_{n+1} = \sum_k Y_{nk}$. Clearly $\mathcal{P}(Z_{n+1} = j) = T_{n+1,j}/t_{n+1}$. We have for the expectation and variance, from (5.4), using (6.16) repeatedly,

$$\mathcal{E}(Z_{n+1}) = \frac{t_{n+2}}{2t_{n+1}} - \frac{n+1}{2}; \tag{6.19}$$

$$\mathcal{D}^2(Z_{n+1}) = \frac{t_{n+3}}{4t_{n+1}} - \frac{t_{n+2}^2}{4t_{n+1}^2} - \frac{t_{n+2}}{2t_{n+1}} - \frac{n+1}{4}. \tag{6.20}$$

Flajolet [17] computed asymptotics for $\mathcal{E}(Z_{n+1})$. In addition, we computed the needed variance. The details are in a Maple worksheet.

Lemma 6.9. *We have the asymptotic formulae*

$$\mathcal{E}(Z_{n+1}) = \frac{1-\rho}{2\rho}n + O(1) \qquad \textit{and} \qquad \mathcal{D}^2(Z_{n+1}) = \frac{n}{4}\Big(\frac{1}{\rho^2} - \frac{2}{\rho} - 1\Big) + O(1).$$

Theorem 6.10. *For the sequence $A(n,j) = T_{n+1,j}$ the central limit theorem (5.5) and the local limit theorem (5.8) holds with $E_n = t_{n+1}$. Furthermore, the number $k = J_n$ that maximizes $T_{n+1,k}$ satisfies*

$$J_n = \frac{1-\rho}{2\rho}n + o(\sqrt{n})$$

and

$$T_{n+1,J_n} = \frac{n!(1+o(1))}{\pi\sqrt{2}n\rho^{n+\frac{1}{2}}\sqrt{(\frac{1}{\rho^2} - \frac{2}{\rho} - 1)}}.$$

Proof. The central and local limit theorems hinge on $\mathcal{D}(Z_n) \to \infty$ that we have from Lemma 6.9. The arguments leading to (5.9) and (5.10) hold for $T_{n+1,k}$ instead of $S(n,k)$. □

From the identity (6.14) we immediately obtain the following central and local limit theorems:

$$\frac{1}{t_{n+1}} \sum_{j=1}^{\lfloor x_n \rfloor} S^\star(n+j, j) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

and

$$\lim_{n\to\infty} \frac{\mathcal{D}(Z_n)}{t_{n+1}} S^\star(n + \lfloor x_n \rfloor, \lfloor x_n \rfloor) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

as $n \to \infty$ uniformly in x, $x_n = \mathcal{E}(Z_n) + x\mathcal{D}(Z_n)$, and $\mathcal{E}(Z_n)$ and $\mathcal{D}(Z_n)$ are defined by (6.19) and (6.20).

Bibliography

[1] D. H. Browne and H. W. Becker, *Problems and Solutions: Elementary Problems: Solutions: E461*, Amer. Math. Monthly **48** (1941), no. 10, 701–703. 1525304

[2] R. A. Brualdi, *Introductory combinatorics*, third ed., Prentice Hall, New York, 1992.

[3] E. R. Canfield, *bellmoser.pdf*, 6 pages manuscript.

[4] ———, *Central and local limit theorems for the coefficients of polynomials of binomial type*, J. Combinatorial Theory Ser. A **23** (1977), no. 3, 275–290. 0450076 (56 #8375)

[5] ———, *Engel's inequality for Bell numbers*, J. Combin. Theory Ser. A **72** (1995), no. 1, 184–187. 1354972 (96m:05012)

[6] E. R. Canfield and L. H. Harper, *A simplified guide to large antichains in the partition lattice*, Proceedings of the Twenty-fifth Southeastern International Conference on Combinatorics, Graph Theory and Computing (Boca Raton, FL, 1994), vol. 100, 1994, pp. 81–88. 1382307 (96k:06005)

[7] A. Cayley, *A theorem on trees*, Quart. J. Math. **23** (1889), 376–378.

[8] L. Clark, *Central and local limit theorems for excedances by conjugacy class and by derangement*, Integers **2** (2002), Paper A3, 9. 1896148 (2003c:60043)

[9] Reinhard Diestel, *Graph theory*, third ed., Graduate Texts in Mathematics, vol. 173, Springer-Verlag, Berlin, 2005. 2159259 (2006e:05001)

[10] A. J. Dobson, *A note on Stirling numbers of the second kind*, J. Combinatorial Theory **5** (1968), 212–214. 0228352 (37 #3933)

[11] ———, *Unrooted trees for numerical taxonomy*, J. Appl. Probability **11** (1974), 32–42. 0357179 (50 #9647)

[12] R. Durrett, *Probability*, The Wadsworth & Brooks/Cole Statistics/Probability Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991, Theory and examples. 1068527 (91m:60002)

[13] P. L. Erdős and L. A. Székely, *Applications of antilexicographic order. I. An enumerative theory of trees*, Adv. in Appl. Math. **10** (1989), no. 4, 488–496. 1023945 (91e:05037)

[14] M. Fellows, M. Hallett, and U. Stege, *Analogs & duals of the MAST problem for sequences & trees*, J. Algorithms **49** (2003), no. 1, 192–216, 1998 European Symposium on Algorithms (Venice). 2027064 (2005f:68041)

[15] J. Felsenstein, *The number of evolutionary trees.*, Systematic Zoology **27** (1978), 27–33.

[16] ———, *Inferring phylogenies*, vol. 24, Sinauer Associates, Inc, Sunderland, Massachusetts, 2004.

[17] P. Flajolet, *A problem in statistical classification theory,*, `http:http://algo.inria.fr/libraries/autocomb/schroeder-html/schroeder.html`.

[18] L. R. Foulds and R. W. Robinson, *Enumeration of phylogenetic trees without points of degree two*, Ars Combin. **17** (1984), no. A, 169–183. 746182 (85f:05045)

[19] G. Ganapathy, B. Goodson, R. Jansen, V. Ramachandran, and T. Warnow, *Pattern identification in biogeography*, Algorithms in bioinformatics, Lecture Notes in Comput. Sci., vol. 3692, Springer, Berlin, 2005, pp. 116–127. 2226830 (2007d:92062)

[20] S. Guillemot, J. Jansson, and W. Sung, *Computing a smallest multi-labeled phylogenetic tree from rooted triplets*, Algorithms and computation, Lecture Notes in Comput. Sci., vol. 5878, Springer, Berlin, 2009, pp. 1205–1214. 2792817

[21] M. D. Haiman, *On mixed insertion, symmetry, and shifted Young tableaux*, J. Combin. Theory Ser. A **50** (1989), no. 2, 196–225. 989194 (90j:05014)

[22] F. Harary and E. M. Palmer, *Graphical enumeration*, Academic Press, New York, 1973. 0357214 (50 #9682)

[23] F. Harary and G. Prins, *The number of homeomorphically irreducible trees, and other species.*, Acta Math. **101** (1959), 141–162. 0101846 (21 #653)

[24] E. F. Harding, *The probabilities of rooted tree-shapes generated by random bifurcation*, Advances in Appl. Probability **3** (1971), 44–77. 0282451 (43 #8162)

[25] L. H. Harper, *Stirling behavior is asymptotically normal*, Ann. Math. Statist. **38** (1967), 410–414. 0211432 (35 #2312)

[26] K. T. Huber, M. Lott, V. Moulton, and A. Spillner, *The complexity of deriving multi-labeled trees from bipartitions*, J. Comput. Biol. **15** (2008), no. 6, 639–651. 2425447 (2009h:92045)

[27] K. T. Huber and V. Moulton, *Phylogenetic networks from multi-labelled trees*, J. Math. Biol. **52** (2006), no. 5, 613–632. 2235520 (2007c:92038)

[28] K. T. Huber, B. Oxelman, M. Lott, and V. Moulton, *The number of evolutionary trees.*, Molecular Biology and Evolution **23** (2006), 1784–1791.

[29] G. Kirchoff, *über die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird*, Ann. Phys. Chem. **72** (1847), 497–508.

[30] D. C. Kurtz, *A note on concavity properties of triangular arrays of numbers*, J. Combinatorial Theory Ser. A **13** (1972), 135–139. 0304296 (46 #3431)

[31] E. H. Lieb, *Concavity properties and a generating function for Stirling numbers*, J. Combinatorial Theory **5** (1968), 203–206. 0230635 (37 #6195)

[32] M. Lott, A. Spillner, K. T. Huber, A. Petri, B. Oxelman, and V. Moulton, *Inferring polyploid phylogenies from multiply-labeled gene trees.*, BMC Evolutionary Biology **9** (2009), 216.

[33] L. Lovász, *Combinatorial problems and exercises*, second ed., North-Holland Publishing Co., Amsterdam, 1993. 1265492 (94m:05001)

[34] J. W. Moon, *Counting labelled trees*, From lectures delivered to the Twelfth Biennial Seminar of the Canadian Mathematical Congress (Vancouver, vol. 1969, Canadian Mathematical Congress, Montreal, Que., 1970. 0274333 (43 #98)

[35] L. Moser and M. Wyman, *An asymptotic formula for the Bell numbers*, Trans. Roy. Soc. Canada. Sect. III. (3) **49** (1955), 49–54. 0078489 (17,1201c)

[36] R. Otter, *The number of trees*, Ann. of Math. (2) **49** (1948), 583–599. 0025715 (10,53c)

[37] B. Salvy and J. Shackell, *Asymptotics of the Stirling numbers of the second kind,* Studies in Automatic Combinatorics II, Published electronically., 1997.

[38] E. Schroder, *Vier combinatorische Probleme*, Z. f. Math. Phys. **15** (1870), no. 10, 361–376.

[39] C. Scornavacca, V. Berry, and V. Ranwez, *From gene trees to species trees through supertree approach*, Language and automata theory and applications, Lecture Notes in Comput. Sci., vol. 5457, Springer, Berlin, 2009, pp. 702–714. 2544458

[40] C. Semple and M. Steel, *Phylogenetics*, Oxford Lecture Series in Mathematics and its Applications, vol. 24, Oxford University Press, Oxford, 2003. 2060009 (2005g:92024)

[41] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, `http://www.research.att.com/~njas/sequences/`, 2012, [Online; accessed 23-March 2012].

[42] R. P. Stanley, *Enumerative combinatorics. Vol. 1*, Cambridge Studies in Advanced Mathematics, vol. 49, Cambridge University Press, Cambridge, 1997, With a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original. 1442260 (98a:05001)

[43] J. H. M. Wedderburn, *The functional equation* $g(x^2) = 2\alpha x + [g(x)]^2$, Ann. of Math. (2) **24** (1922), no. 2, 121–140. 1502633

[44] H. S. Wilf, *generatingfunctionology*, third ed., A K Peters Ltd., Wellesley, MA, 2006. 2172781 (2006i:05014)

Appendix A

Sage programs which count mul-trees

A.1 Rooted and unrooted binary MUL-trees

This program counts the various types of rooted and unrooted binary MUL-trees described in Chapters 2 and 4.

```
#Calculates the number of different types of
#Semi-labelled Binary Trees with n leaves and k labels.
#Answers given in this order. Rooted (R), Rooted using all labels (V),
#Marked (M),  Marked using all labels (VM),
#Unrooted (U), Unrooted using all labels (VU)
#The number of times each label is used is not specified in first set.
#Each label used at least once in second answer set
#AUTHOR: Virginia Johnson (2011-07)  version 1


def T(n,k):
    #Gets input and will return the number of trees
    #with leaves 0-n on k labels"""
    #first section calculates the rooted binary trees
    #(R_k in documentation) number of leaves varies,
    #number of labels fixed
    LL=[]        #stores r_n,0, r_n,1, ...r_n,k
    for p in range(k+1):
```

```
    L=[0]*(n+1)     #stores r_0,k,  r_1,k, ...r_n,k
    LL.append(L)
    for i in range(n+1):
        #"0 if no leaves"
        if i==0:
            L[i]=0
        #"p if one leaf"
        elif i==1:
            L[i]=p
        #"if number of leaves is even"
        elif (mod(i,2)==0) and (i!=0):
            L[i]=1/2*L[i/2]
            for j in range(1,i):
                L[i]+=1/2*L[j]*L[i-j]
        else:
            for j in range(1,i):
                 L[i]+=1/2*L[j]*L[i-j]
        #Calculates Rooted semi-labeled binary trees
        #n= number of leaves,
        #k= number of labels
        #Each label is used at least once.
V=[0]*(n+1)
for i in range(n+1):
     for j in range (0,k):
         V[i]+=(-1)^j*binomial(k,j)*LL[k-j][i]

          #this section  calculates the sums
```

```
          #needed for a_n;k in documentation"""
BA=[]    # this holds values for smaller number of leaves0-k
for h in range(k+1):

    B=[0]*(n+1)
    BA.append(B)

    for i in range(1,n+1):
        if i==0:
            B[i]=0
        else:
            B[i]=h*LL[h][i-1]            #adds in first term
            for j in [0..floor(i/3)]:
                #selects combinations of i,j,k,which sum to n
                for m in [j..floor((i-j)/2)]:
                    p = i-j-m
                    t=[j,m,p]
                       #t is created to determine how many
                       #elements in set to create
                       #c_i,j,l documentation
                    if (2*j)+p==i and len(set(t))!=1:
                        #adds in third term first
                        #testing for j=m
                        B[i]+=(1/2)*LL[h][j]*LL[h][p]
                        #and eliminating j=m=p which
                        #is included  in
                        #next if statement
```

```
                    if j+(2*m)==i:
                        #this gets j=m=p and
                        #m=p all needed
                        #in third term
                        B[i]+=(1/2)*LL[h][j]*LL[h][m]
                        # have now added in third term


                    if len(set(t))==1:
                        #sets the coefficient c and
                        #adds in second term
                        c=1
                        B[i]+=1/6*c*LL[h][j]*LL[h][m]*LL[h][p]
                    elif len(set(t))==2:
                        c=3
                        B[i]+=1/6*c*LL[h][j]*LL[h][m]*LL[h][p]
                    elif len(set(t))==3:
                        c=6


                        B[i]+=1/6*c*LL[h][j]*LL[h][m]*LL[h][p]
                        #have now completed adding in
                        #2nd term




#this section calculates the numbers of
#Marked trees...(M in documentation)
```

```
MA=[]
# this holds values for smaller number of leaves0-k
for h in range(k+1):

    M=[0]*(n+1)
    MA.append(M)
    #calculates the final sum
    for i in range(n+1):
        if i==0:
            M[i]=0
        elif i==1:
            M[i]=h
        elif (mod(i,3)==0) and (i!=0):
            M[i]=BA[h][i]+(1/3)*LL[h][i/3]
        else:
            M[i]=BA[h][i]

#This section calculates M^* trees in documentation.
#Each label is used
VM=[0]*(n+1)
for i in range(n+1):
        for j in range (0,k):
            VM[i]+=(-1)^j*binomial(k,j)*MA[k-j][i]
#This section calculated unrooted binary trees.
#(U in documentation)
AU=[]
# this holds values for smaller number of leaves0-k
```

```
    for h in range (k+1):
        U=[0]*(n+1)
        AU.append(U)

        for i in range(n+1):
            if i==0:
                U[i]=0
            elif i==1:
                U[i]=h
            elif (mod(i,2)==0) and (i!=0):
                U[i]=MA[h][i]-LL[h][i]+LL[h][i/2]
            else:U[i]=MA[h][i]-LL[h][i]
    #This section calculates U^*
    #in documentation
    #unrooted  binary MUL trees using all k labels
    VU=[0]*(n+1)
    for i in range(n+1):
        for j in range (0,k):
            VU[i]+=(-1)^j*binomial(k,j)*AU[k-j][i]
#_________________________



    #This section returns the calculated numbers"""

    print "Number of leaves= ", n,  "    number of labels= ",k
```

```
    print "Rooted MUL Binary Trees"
    print L

    print "Rooted MUL Binary Trees using all k labels"
    print V

    print "Marked  MUL Binary Trees"
    print M
    print "Marked MUL Binary Trees using all k labels"
    print VM
    print "Unrooted MUL Binary Trees"
    print U
    print "Unrooted MUL Binary Trees using all k labels"
    print VU
```

A.2 Rooted and unrooted non-binary trees; first program

This program counts rooted and unrooted non-binary MUL-trees using the recursive function 5.4

```
   #Given the number of leaves "n" and number of labels "k"
#this program returns the number of rooted multi-leafllabeled
#trees where the degree of the root is >=2, degree of
#non-root, non-leaf vertices is >=3
#AUTHOR: Virginia Johnson (2011-10)  version 1
def G(n,k):
    #Gets input and will return the number of trees
    #with leaves 0-n where  k is the size of the label set.
```

```
T=[0]*(n+1)
for i in range (n+1):
    #easy cases
    #no leaves
    if i==0:
        T[i]=0
    #1 leaf
    elif i==1:
        T[i]=k
    #for n>=2
    else:
        #find m= how many partitions there are of i
        m=Partitions(i).cardinality()
        #set up a counter that will stop the loop
        #when finished with all partitions (m-1)
        count=0
        #get the partitions 1 at a time
        #and omit the first one
        g=iter(Partitions(i))
        g.next()

        while count != m-1:
            #fix this partition for the duration
            #of the first calculation
            L=g.next()
            #print "L"
            #print L
```

```
            #set up a string which holds counts
            S=[]

            #count the number of times each integer
            #in{1,...i-1} appears in partition
            for c in range (0,i):
                S.append(list(L).count(c))

            #create string for product
            P=[0]*(i)
            P[0]=1

            for d in range (1,len(list(S))):
                P[d]=binomial(T[d]+S[d]-1,S[d])

            T[i]+=prod(P)

            count=count+1
#Uses T to calculate number of unrooted trees
#on n leaves using label set  size k.
U=[0]*(n+1)
for i in range (n+1):
    #easy cases first
    #no leaves
    if i==0:
        U[i]=0
```

```
        #1 leaf
        elif i==1:
            U[i]=k
        #for n >=2
        else:
            U[i]=k*T[i-1]+T[i]
            for j in range(1,i):
              U[i]+=T[j]*T[i-j]
    print "Number of leaves=", n, "    Number of labels=", k
    print "Rooted Non-binary Multi-leaf-labeled Trees"
    print T
    print  "Unrooted Non-binary Multi-leaf-labeled Trees"
    print U
```

A.3 Rooted and unrooted non-binary trees; second program

This program counts rooted and unrooted non-binary MUL-trees using the recursive function 5.2.

```
##Given the number of leaves "n" and number of labels "k"
#this program returns the number of rooted multi-leaf- labeled
#trees where the degree of the root is >=2, degree of non-root,
#non-leaf vertices is >=3
#Author:Virginia Johnson  11/2011
def G(n,k):
    #Gets input and will return the number of trees
    #with leaves 0-n where  k is the size of the label set.
    T=[0]*(n+1)
```

```
for i in range (n+1):
    #easy cases
    #no leaves
    if i==0:
        T[i]=0
    #1 leaf
    elif i==1:
        T[i]=k
    #for n>=2
    else:
        #find d= divisors of i
        d=divisors(i)
        #set up a counter that will stop the loop
        #when finished with all divisors
        #except last one (m-1)
        m=len(d)
        g=0
        #create the first sum
        while g != m-1:
            T[i]+=d[g]/i*T[d[g]]
            g=g+1
        outsum=0
        for mm in range(2,i+1):
            for c in Compositions(i,length=mm):
                insum=0
                inprod=1
                for nj in c:
```

```
                    divlist = divisors(nj)
                    divsum=0
                    for d in divlist:
                        divsum+=d*T[d]
                    inprod=inprod*divsum/nj
                insum+=inprod
                outsum+=insum/factorial(mm)
        T[i]+=outsum

print "Number of leaves=", n, "    Number of labels=", k
print T
```

Appendix B

Maple Code: Bell Numbers

$$P0 := -\frac{2\cdot r^4 + 9\,r^3 + 16\cdot r^2 + 6\cdot r + 2}{24\cdot r\cdot (r+1)^3}$$

$$-\frac{1}{24}\,\frac{2\,r^4 + 9\,r^3 + 16\,r^2 + 6\,r + 2}{r\,(r+1)^3} \tag{1}$$

$$P1 := -\frac{r^2 + 3\cdot r + 1}{2\cdot r\cdot (r+1)^2}$$

$$-\frac{1}{2}\,\frac{r^2 + 3\,r + 1}{r\,(r+1)^2} \tag{2}$$

$$P2 := -\frac{1}{2\cdot r\cdot (r+1)}$$

$$-\frac{1}{2\,r\,(r+1)} \tag{3}$$

$$Q0 := \frac{6 + 24\cdot r + 100\cdot r^2 - 636\cdot r^3 - 588\cdot r^4 - 384\cdot r^5 - 143\cdot r^6 - 12\cdot r^7 + 4\cdot r^8}{1152\cdot r^2\cdot (r+1)^6}$$

$$\frac{1}{1152}\,\frac{6 + 24\,r + 100\,r^2 - 636\,r^3 - 588\,r^4 - 384\,r^5 - 143\,r^6 - 12\,r^7 + 4\,r^8}{r^2\,(r+1)^6} \tag{4}$$

$$Q1 := \frac{6 + 32\cdot r + 56\cdot r^2 + 135\cdot r^3 + 101\cdot r^4 + 37\cdot r^5 + 6\cdot r^6}{48\cdot r^2\cdot (r+1)^5}$$

$$\frac{1}{48}\,\frac{6 + 32\,r + 56\,r^2 + 135\,r^3 + 101\,r^4 + 37\,r^5 + 6\,r^6}{r^2\,(r+1)^5} \tag{5}$$

$$Q2 := \frac{20 + 90\cdot r + 190\cdot r^2 + 105\cdot r^3 + 20\cdot r^4}{48\cdot r^2\cdot (r+1)^4}$$

$$\frac{1}{48}\,\frac{20 + 90\,r + 190\,r^2 + 105\,r^3 + 20\,r^4}{r^2\,(r+1)^4} \tag{6}$$

$$Q3 := \frac{5 + 15\cdot r + 5\cdot r^2}{12\cdot r^2\cdot (r+1)^3}$$

$$\frac{1}{12}\,\frac{5 + 15\,r + 5\,r^2}{r^2\,(r+1)^3} \tag{7}$$

$$Q4 := \frac{1}{8\cdot r^2\cdot (r+1)^2}$$

$$\frac{1}{8\,r^2\,(r+1)^2} \tag{8}$$

$$B := (n,h) \to \frac{(n+h)!}{r^h}\cdot\left(1+\left(P0+h\cdot P1+h^2\cdot P2\right)\cdot\frac{r}{n}+\left(Q0+h\cdot Q1+h^2\cdot Q2+h^3\cdot Q3+h^4\cdot Q4\right)\cdot\frac{r^2}{n^2}+r^3\cdot O\left(\frac{1}{n^3}\right)\right);$$

$$(n,h)\to\frac{1}{r^h}\left((n+h)!\left(1+\frac{\left(P0+h\,P1+h^2\,P2\right)r}{n}+\frac{\left(Q0+h\,Q1+h^2\,Q2+h^3\,Q3+h^4\,Q4\right)r^2}{n^2}+r^3\,O\left(\frac{1}{n^3}\right)\right)\right) \tag{9}$$

$$Bstar := (n,h) \to B(n,h-1)-B(n,h-2)+B(n,h-3)-B(n,h-4)+B(n,h-5)-B(n,h-6)+B(n,h-7)+C\cdot B(n,h-8);$$

$$(n,h)\to B(n,h-1)-B(n,h-2)+B(n,h-3)-B(n,h-4)+B(n,h-5)-B(n,h-6)+B(n,h-7)+C\,B(n,h-8) \tag{10}$$

$$sort\left(simplify\left(asympt\left(\frac{Bstar(n,2)}{Bstar(n,0)}+2\cdot n\cdot\frac{Bstar(n,1)\cdot Bstar(n,-1)}{Bstar(n,0)^2}+n\cdot(n-1)\cdot\frac{Bstar(n,-2)}{Bstar(n,0)}-\frac{Bstar(n,1)^2}{Bstar(n,0)^2}-n^2\cdot\frac{Bstar(n,-1)^2}{Bstar(n,0)^2}-n\cdot\frac{Bstar(n,-1)}{Bstar(n,0)}-(2\cdot n+1),\ n,5\right)\right),\ order=(plex(n,r))\right);$$

$$\frac{1}{2}\,\frac{1}{(r+1)^4\,r}\left(2\,n\,r^3+6\,n\,r^2+6\,n\,r+2\,n-2\,r^6+2\,O\left(\frac{1}{n}\right)r^5-6\,r^5+8\,O\left(\frac{1}{n}\right)r^4-8\,r^4+12\,O\left(\frac{1}{n}\right)r^3-9\,r^3+8\,O\left(\frac{1}{n}\right)r^2-9\,r^2+2\,O\left(\frac{1}{n}\right)r-2\,r\right) \tag{11}$$

$$sort\left(simplify\left(asympt\left(\frac{Bstar(n,1)}{Bstar(n,0)}-n\cdot\frac{Bstar(n,-1)}{Bstar(n,0)},\ n,2\right)\right),\ order=(plex(n,r))\right);$$

$$\frac{1}{2}\,\frac{1}{(r+1)^2\,r}\left(2\,n\,r^2+4\,n\,r+2\,n-2\,r^4-4\,r^3+2\,O\left(\frac{1}{n}\right)r^3-3\,r^2+4\,O\left(\frac{1}{n}\right)r^2-2\,r+2\,O\left(\frac{1}{n}\right)r\right) \tag{12}$$

Appendix C

Maple code: Phylogenetic trees

$$\frac{\log_{10}(a)}{b} H_z := -\text{LambertW}\left(-\frac{1}{2}\exp\left(\frac{1}{2}\cdot z-\frac{1}{2}\right)\right)+\frac{1}{2}\cdot z-\frac{1}{2};$$

$$-\text{LambertW}\left(-\frac{1}{2}\,e^{\frac{1}{2}z-\frac{1}{2}}\right)+\frac{1}{2}z-\frac{1}{2} \tag{1}$$

$$H_s := subs\left(z=(-1+2\cdot\ln(2))\cdot\left(1-\Delta^2\right), H_z\right);$$

$$-\text{LambertW}\left(-\frac{1}{2}\,e^{\frac{1}{2}(-1+2\ln(2))\left(1-\Delta^2\right)-\frac{1}{2}}\right)+\frac{1}{2}(-1+2\ln(2))\left(1-\Delta^2\right)-\frac{1}{2} \tag{2}$$

$$\text{rho} := -1+2\cdot\ln(2);$$

$$-1+2\ln(2) \tag{3}$$

$$H_{sing} := map\left(simplify, series\left(H_s, \text{Delta}=0, 10\right)\right); \text{Delta} = \text{sqrt}(``(1-z/\text{rho}));$$

$$\ln(2)-\frac{1}{2}\sqrt{2}\sqrt{-2+4\ln(2)}\,\Delta+\left(\frac{1}{6}-\frac{1}{3}\ln(2)\right)\Delta^2-\frac{1}{72}\sqrt{2}\sqrt{-2+4\ln(2)}\,(-1$$
$$+2\ln(2))\Delta^3-\frac{1}{270}(-1+2\ln(2))^2\Delta^4-\frac{1}{8640}\sqrt{2}\sqrt{-2+4\ln(2)}\,(1-4\ln(2)$$
$$+4\ln(2)^2)\Delta^5+\frac{1}{17010}(-1+2\ln(2))^3\Delta^6+\frac{139}{10886400}\sqrt{2}\sqrt{-2+4\ln(2)}\,(-1$$
$$+6\ln(2)-12\ln(2)^2+8\ln(2)^3)\Delta^7+\frac{1}{204120}(-1+2\ln(2))^4\Delta^8$$
$$+\frac{571}{4702924800}\sqrt{2}\sqrt{-2+4\ln(2)}\,(1-8\ln(2)+24\ln(2)^2-32\ln(2)^3$$
$$+16\ln(2)^4)\Delta^9+O\left(\Delta^{10}\right)$$

$$\Delta = \sqrt{\left(1-\frac{z}{-1+2\ln(2)}\right)} \tag{4}$$

$$H_{asympt} := n!\cdot asympt\left(coeff\left(H_{sing}, \text{Delta}, 1\right)\cdot\text{rho}\,\hat{}\,(-n)\cdot subs(\{\cos(\text{Pi}\cdot n)=1, O=0\},\right.$$
$$\left.simplify(asympt(\text{binomial}(1/2, n), n, 2))), n\right);$$

$$\frac{1}{4}\,\frac{n!\sqrt{2}\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{3/2}}{\sqrt{\pi}\,(-1+2\ln(2))^n} \tag{5}$$

$$H_{asymptexpansion} := n!\cdot asympt\left(coeff\left(H_{sing}, \text{Delta}, 1\right)\cdot\text{rho}\,\hat{}\,(-n)\cdot subs(\{\cos(\text{Pi}\cdot n)=1\},\right.$$
$$\left.simplify(asympt(\text{binomial}(1/2, n), n, 4))), n, 8\right);$$

$$\frac{1}{(-1+2\ln(2))^n}\left(n!\left(\frac{1}{4}\,\frac{\sqrt{2}\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{3/2}}{\sqrt{\pi}}\right.\right. \tag{6}$$

$$+\frac{3}{32}\,\frac{\sqrt{2}\,\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{5/2}}{\sqrt{\pi}}+\frac{25}{512}\,\frac{\sqrt{2}\,\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{7/2}}{\sqrt{\pi}}$$
$$+\mathrm{O}\left(\left(\frac{1}{n}\right)^{9/2}\right)\Bigg)\Bigg)$$

$A := unapply(\ (6), n\ );$

$$n\to \frac{1}{(-1+2\ln(2))^{n}}\left(n!\left(\frac{1}{4}\,\frac{\sqrt{2}\,\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{3/2}}{\sqrt{\pi}}\right.\right. \tag{7}$$
$$+\frac{3}{32}\,\frac{\sqrt{2}\,\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{5/2}}{\sqrt{\pi}}+\frac{25}{512}\,\frac{\sqrt{2}\,\sqrt{-2+4\ln(2)}\left(\frac{1}{n}\right)^{7/2}}{\sqrt{\pi}}$$
$$\left.\left.+\mathrm{O}\left(\left(\frac{1}{n}\right)^{9/2}\right)\right)\right)$$

$$expect := simplify\left(asympt\left(\frac{A(n+2)}{2\cdot A(n+1)}-\frac{n+1}{2}, n, 5\right)\right);$$
$$\frac{1}{4}\,\frac{4n-4n\ln(2)+3-4\ln(2)-4\,\mathrm{O}\left(\frac{1}{n}\right)+8\,\mathrm{O}\left(\frac{1}{n}\right)\ln(2)}{-1+2\ln(2)} \tag{8}$$

$$dsquare = simplify\left(asympt\left(\frac{A(n+3)}{4\cdot A(n+1)}-\frac{A(n+2)^2}{4\cdot A(n+1)^2}-\frac{A(n+2)}{2\cdot A(n+1)}-\frac{n+1}{4}, n, 7\right)\right);$$

$$dsquare = \frac{1}{8}\,\frac{1}{(-1+2\ln(2))^2}\left(4n-8n\ln(2)^2+1+4\ln(2)-8\ln(2)^2+8\,\mathrm{O}\left(\frac{1}{n}\right)\right. \tag{9}$$
$$\left.-32\,\mathrm{O}\left(\frac{1}{n}\right)\ln(2)+32\,\mathrm{O}\left(\frac{1}{n}\right)\ln(2)^2\right)$$