TU WIEN
TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

DISSERTATION

COMBINATORIAL PROPERTIES OF PHYLOGENETIC NETWORKS

Ausgeführt zum Zwecke der Erlangung des akademischen
Grades eines Doktors der technischen Wissenschaften unter
der Anleitung von

AO.UNIV.PROF. DIPL.-ING. DR.TECHN. BERNHARD GITTENBERGER
E104, Institut für
Diskrete Mathematik und Geometrie
eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

MAREFATOLLAH MANSOURI
Matrikelnummer 11771258
Fruethstrasse 5 /10
1030 Wien

Wien, am 12 Mai 2020

M. Mansouri
Marefatollah Mansouri

Bernhard Gittenberger

Hsien-Kuei Hwang

DOCTORAL THESIS

Combinatorial Properties of Phylogenetic Networks

SUBMITTED TO

VIENNA UNIVERSITY OF TECHNOLOGY

Institute of Discrete Mathematics and Geometry

SUPERVISED BY

Ao.Univ.Prof. Dipl.-Ing. Dr.techn.

Bernhard Gittenberger

PRESENTED BY

Marefatollah Mansouri

ABSTRACT

This thesis is concerned with the combinatorial properties of some subclasses of rooted and unrooted phylogenetic networks. We consider several phylogenetic networks and provide enumeration formulas (exact and asymptotic) for them with a given number of leaves, also analyze some of their characterizing parameters, such as the number of biconnected components and the number of edges across all these components. The vital tool in this context is the concept of generating functions such that the proof method is based on their algebraic and analytic properties and helps us to solve the enumeration problems. Due to these methods, this thesis belongs to the field of analytic combinatorics. We begin by introducing some of the concepts from graph theory necessary to formally define a phylogenetic tree and a phylogenetic network with many other helpful theorems in the first and second chapters.

The third chapter continues the work of Semple and Steel and extends their work to the case rooted and unrooted level-2 networks. Moreover, the scheme for generating functions leading to show that mentioned above parameters are asymptotically normally distributed.

The next chapter treats the analysis of tree-child and normal networks. This part deals with the delicate problem of deriving the enumerative and asymptotic results. It also sheds light on solutions of open problems in [8] regards to presenting explicit formulas for the count of such networks with up to three reticulation vertices.

The final chapter treats applications of analytic combinatorics to general phylogenetic networks. This is done by extending results from Chapter 4. Some of the results presented in this thesis have already been published in scientific articles by the present author.

PUBLICATIONS

This thesis is based on papers written by the author and co-authored with Bernhard Gittenberger, Michael Fuchs, Mathilde Bouvel and Philippe Gambette. All these papers have been published, submitted or in the way to be submitted:

- [26] Counting Phylogenetic Networks with Few Reticulation Vertices: Tree-Child and Normal Networks. Australasian Journal of Combinatorics (2018), 385–423. With Michael Fuchs and Bernhard Gittenberger.

- [6] Counting Phylogenetic Networks of level 1 and 2. Submitted (2019). With Mathilde Bouvel and Philippe Gambette.

- [43] Counting general Phylogenetic Networks. Submitted (2020).

- [44] The structure and enumeration of galled networks. In preparation (2020).

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without assistance and encouragement of several people. It is a pleasure to express my sincere thanks to all those who helped me for the success of this study and made it an unforgettable experience. Thank God Almighty for showering His most abundant grace during all the challenging moments in completing my doctoral thesis.

First and foremost, I am deeply indebted to my doctoral thesis supervisor, Bernhard Gittenberger, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I owe Bernhard lots of gratitude for having me shown this way of research. I am really glad to be associated with a person like him in my life. On the academic level, I appreciate his contributions of time and ideas to make my work productive and stimulating. His deep insights helped me at various stages of my research and valuable suggestions, comments and constant guidance encourage me to explore more of the unknown research area. He absolutely has an amazing inspiring personality. Bernhard inspired me by his dedicated and passionate working attitudes, as well as his humble and friendly attitudes towards others.

Another person I cannot thank enough is Michael Fuchs. He continuously encouraged me to explore new fields of mathematics and he never got tired of explaining when I was lacking the knowledge. Despite the distance we always kept close contact and he was always willing to answer any of my questions. I thank him very much for inviting me several times to Taiwan. I gratefully acknowledge the funding received towards my PhD from the bilateral Austrian-Taiwanese project FWF-MOST.

I greatly appreciate the support received from Philippe Gambette and Mathilde Bouvel through the collaborative work who were always so helpful and provided me with their assistance throughout my PhD study.

Also, my sincere gratitude is reserved for my colleagues Andrei Asinowski, Lukas Spiegelhofer and Clément Requilé in our combinatorial and algorithmic group at TU Vienna. I am very grateful that they patiently listen to my ideas on my research problem and always willing to help me with my thesis. In addition, I would also like to take this opportunity to thank the secretaries Marietta Meszlenyi and Sonja Rees for helping me with many German paperwork. They made me, despite living far away from home, feel very warm and well-protected in Austria.

Last but not least, a heartfelt thanks goes to my supportive wonderful family, especially to my fiancée and love of my life, Zahra Babaei, for always being there in the moments of difficulties during the last three years. She kept me going on and this work would not have been possible without her input. This journey would not have been possible if not for her, and I dedicate this milestone to her.

Contents

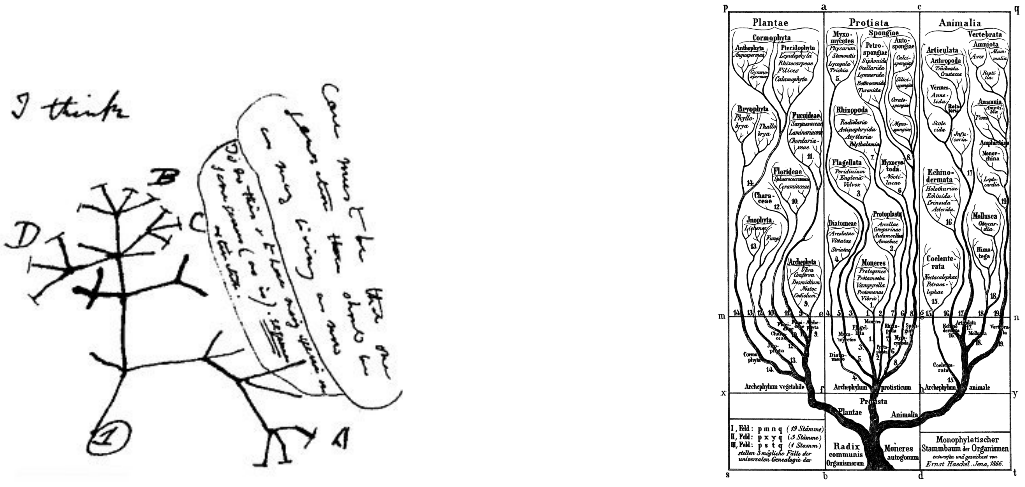
Chapter 1

Introduction

> *"I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense."*
>
> - Charles Darwin, 1828

Phylogeny is the study of relationships among different groups of organisms, as a way of classifying them. All residing creatures on this planet carry a signature of their evolutionary background inside their DNA. By reading styles and variations between the genetic makeup of different species, molecular biologists are in a position to piece collectively elements of the story of how lifestyles these days traces back to a common origin (as illustrated on Figure1.1). Phylogenetic evaluation pursuits at finding out the evolutionary relationships between unique species or taxa so that you can reap an understanding of the evolution of existence on earth. To address this task, “Phylogenetic trees” are extensively used and are normally computed from molecular sequences. Generally, phylogenetic trees are appropriate to represent evolutionary histories where the principal events are speciations (at the internal nodes) and ancestor with adjustment (along the edge of the tree). Nevertheless, these trees are inappropriate to version mechanisms of “*reticulate evolution*” [58] inclusive of hybridization, homologous evolution, or lateral gene transfer. Furthermore, events inclusive of incomplete lineage sorting or complex patterns of gene duplication and loss which can cause incompatibilities, cannot be represented on a tree. So instead of phylogenetic trees, we can use "Phylogenetic networks" when analyzing data sets whose evolution involves enormous number of reticulate events (for example see [48, 35, 61]).

Phylogenetic networks are used to model reticulation events that have an explicit biological interpretation in evolutionary biology. Even though the presence of such phenomena has been acknowledged with the aid of biologists since the


I think
1
A
B
C
D
Plantae
Protista
Animalia
Radix communis Organismorum
Moneres autogonum
Monophyletischer Stammbaum der Organismen
Ernst Haeckel, Jena, 1866

(i) (ii)

Figure 1.1: (i) Darwin's Tree of Life, July 1837. (ii) Early depiction of a "tree of life" by Ernst Haeckel, 1866, in which plants and animals dominated two of the three main branches.

advent of the development of evolution as a scientific order, for the most part, phylogenetic trees in preference to phylogenetic networks were used to model the connection between species. This is probably due to the fact that trees are a significantly easier structure than networks and thus allow a rich theory. For instance, their combinatorics is well-understood: the corresponding counting problem was already solved by Schröder in 1870 [51]. Several further studies were published to analyze parameters or variations, e.g. [2, 3, 23]. Moreover, phylogenetic trees are also important for constructing phylogenetic networks (see [4, 10]) and thus the comparison of phylogenetic trees and networks is an active area of research, see [25] and [5, 46, 52] concerning tree-embeddings in networks.

The combinatorics of phylogenetic networks, on the other hand, remains an assignment and only few papers have addressed it. Mainly, the intention of this thesis is to make some more development and specifically to resolve the counting problem for phylogenetic networks, a fundamental question which is of interest in mathematical biology; see [7]. Before stating our results in more detail, we recall some definitions and former work. Phylogenetic networks are usually labeled. we can consider two kinds of labeling wherein all labels are assumed to be different: (i) all vertices are labeled; such networks we will call *vertex-labeled networks* throughout this work, and (ii) only leaves are labeled; these are called *leaf-labeled networks*. Note that in the later case, we use $X = \{x_1, \cdots, x_n\}$ to denote a set of *taxa* whose evolutionary history is of interest to us. Note that each taxon x_i represents some species, group or individual organism. For example, $X = \{x_1, x_2, \cdots\}$ might denote a set of reptiles, with x_1 representing crocodiles, x_2 representing turtles , etc. The set X is corresponding to leaves of networks. It is vital here to clarify precisely what is meant by phylogenetic networks. Before that, we need to define the following concepts; see [64, chapter 4] .

Definition 1.0.1 (cut-edge)**.** *In graph theory a cut-edge or bridge of an undirected graph or multigraph G is an edge whose removal disconnects the graph.*

Definition 1.0.2 (bridgeless component)**.** *A bridgeless component of a graph or multigraph G is a maximal induced subgraph of G without cut-edges.*

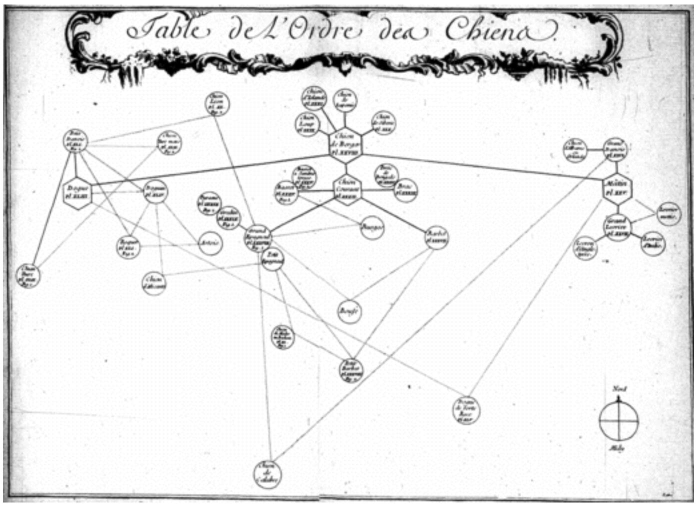


Figure 1.2: The first phylogenetic network (Buffon, 1755).

The mathematical and computational perspective of networks is perhaps the most active subject of current phylogenetics. In the following [60, chapter 10] we give a short introduction to the *binary phylogenetic networks* and describe a selection of the main concepts and results that we need for future discussion.

Definition 1.0.3 (rooted phylogenetic networks)**.** *We define a binary rooted phylogenetic network N on a set X of leaf labels as a directed acyclic multigraph having:*

1. *exactly one root, that is an indegree-0 outdegree-2 vertex (or an indegree-0 outdegree-0 vertex if N only has one leaf);*

2. *leaves, that is indegree-1 outdegree-0 vertices (or an indegree-0 outdegree-0 vertex if N only has one leaf) which are bijectively labeled by elements of X;*

3. *tree vertices, that is indegree-1 outdegree-2 vertices;*

4. *reticulation vertices, that is indegree-2 outdegree-1 vertices;*
   and such that

5. *for each bridgeless component B of N, there exist at least two cut arcs of N whose tail*[1] *belongs to B.*

   Remark. Note that in literature the last property does not explicitly hold for *vertex-labeled rooted general phylogenetic networks*, so we can ignore it for this case (For more details see chapter 5).

Typical parameters of rooted networks.

Given a rooted network $\mathcal{G}$, we use $n(\mathcal{G})$, $\ell(\mathcal{G})$, $r(\mathcal{G})$ and $t(\mathcal{G})$ to denote the set of all vertices, the set of leaves, the set of reticulation vertices and the set of tree vertices, respectively. we shall see that always

$$\ell + r = t + 2 = (n + 1)/2. \tag{1.1}$$

Thus, It means besides of t and n, any pair of ℓ, r, t and n determine the rest of parameters. Also, for large n, both $\ell + r$ and t are about $n/2$.

Lemma 1.0.4 ([7])**.** *Let $\mathcal{G}$ be a rooted network on n vertices with ℓ leaves, r reticulation vertices, and t tree vertices. Then $t = \ell + r - 2$ and $n = 2t + 3$. Also, $\mathcal{G}$ has $3r + 2\ell - 2$ edges.*

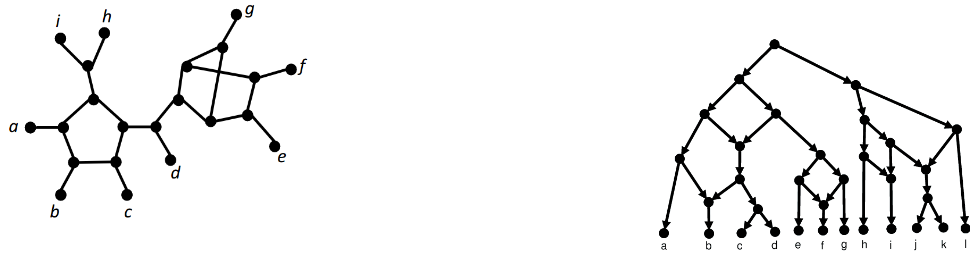
Proof. Note first that $n = r + \ell + t + 1$. Since the sum of the out-degrees equals the number e of edges which, in turn, equals the sum of the in-degrees, we have $r + 2t + 2 = e = 2r + t + \ell$. Hence $t = r + \ell - 2$, and now the lemma follows easily. □

Now, we extend the latter definition to unrooted phylogenetic networks.

Definition 1.0.5 (unrooted phylogenetic networks)**.** *An unrooted binary phylogenetic network N on a set X of leaf labels is a loopless (undirected) graph whose vertices have either degree 3 (internal vertices) or degree 1 (leaves), such that its set $L(N)$ of leaves is bijectively labeled by X and such that for each bridgeless component B of N having strictly more than one vertex, the set of cut-edges incident with some vertex of B has size at least 3.*

An unrooted binary phylogenetic tree is an unrooted binary phylogenetic network with no bridgeless component containing strictly more that one vertex.

[1] The *tail* of an arc is by definition its starting point. Its arrival point is called *head*.



(i) (ii)

Figure 1.3: (i) An unrooted phylogenetic network. (ii) A rooted phylogenetic network.

Phylogenetic networks are used to model reticulate evolution. We can consider many kinds and subclasses of phylogenetic networks based on the biological phenomenon that they represent or which data they are constructed from, or restrictions to get computationally tractable problems; see [35] for more details. Thus, biologists have defined many subclasses of the class of phylogenetic networks. In

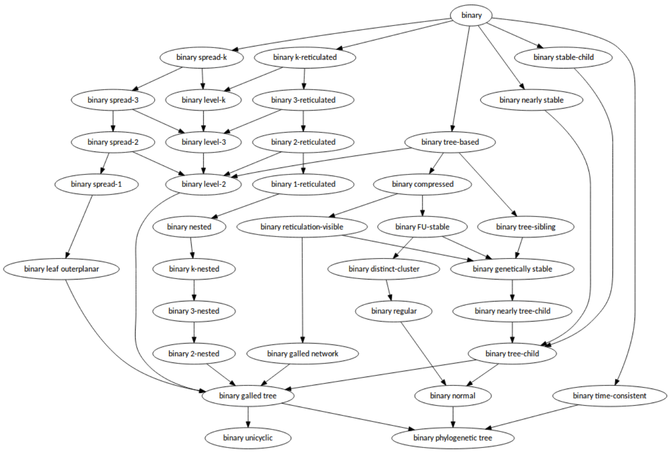


Figure 1.4: Classes of binary networks. An arrow from class A to class B means that A contains B.

this thesis we mainly study enumerative properties of *level-1*, *level-2*, *tree-child* and *normal* networks. After all in the last chapter, we will show how the results of two later phylogenetic networks can be extended to general networks as well. The language we use is the one introduced by *P. Flajolet* and *R. Sedgewick* in their reference book Analytic Combinatorics[20].

Chapter 2

Preliminaries

2.1 Phylogenetic trees

From the time of Charles Darwin [12], the reconstructing the evolutionary history of all orgasm has been the purpose of many biologists to explicit it in the form of a *phylogenetic tree*. Phylogenetic tree, also known as *Dendrogram*, is a diagram to show the evolutionary connections of a family of organisms derived from a regular ancestral configuration. The ancestor is inside the the tree; organisms which have arisen from it are placed at the ends of tree leaves. Note that the degree of relationship can be indicated by the distance of one group from the others. It means, closely related groups are positioned on branches close to one another. A depicted tree is a common way to summarize the results of phylogenetic analysis. This also presents the patterns of organisms and the essence of the evolutionary processes. However, phylogenetic trees represent a graphical model of evolutionary connection, but here we consider slightly more general class of objects which is important of mathematical view and called *binary phylogenetic X-tree* [56].

Definition 2.1.1. *A binary phylogenetic X-tree is a tree T in which every interior vertex has degree three and whose leaf set is X. The set X is often referred to as the label set of T and its elements as labels.*

Historically, enumeration of phylogenetic trees have played a significant role in combinatorial biology. Counting techniques to systemize the problem are developed to obtain information on the quantities A_n of objects of size n in a family $\mathcal{A}$. From this starting point, phylogenetic trees can be studied in more detail. These structures are widely used to express and explore evolutionary relationships. and have been well studied (see, for example, [54, 47]). Note that any binary phylogenetic X-tree on $|X| = n$, has $2n - 3$ edges and $n - 3$ interior edges. Let $B(n)$ denote the number of all binary phylogenetic trees with label set X and let

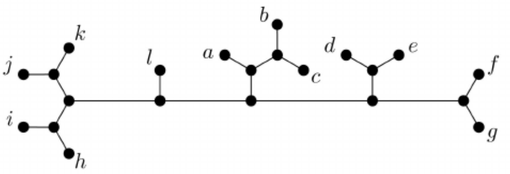


Figure 2.1: A binary phylogenetic tree where $X = \{a, b, \cdots, l\}$.

$b(n) = |B(n)|$. We have following proposition provides a well-known formula for $b(n)$.

Proposition 2.1.2 (Schröder, [51]). *For all* $n \geqslant 3$*,*

$$b(n) = \frac{(2n-4)!}{(n-2)!2^{n-2}} = 1 \times 3 \times 5 \times \cdots \times (2n-5).$$

Proof. If $n = 2$, then $b(n) = 1$. We apply induction for all $n \geqslant 3$. Since $b(3) = 1$, the result holds for $n = 3$. Suppose that the result holds for $n = k - 1$, where $k \geqslant 4$. A binary phylogenetic tree on k labeled leaves can be formed by connecting the kth leaf to a new node in the middle of any of the edges of an unrooted binary tree on $k-1$ labeled leaves. There are $2k - 5$ edges at which the kth node can be attached; therefore, the number of trees on k leaves is larger than the number of trees on $k - 1$ leaves by a factor of $2k - 5$, and so

$$b(n) = 1 \times 3 \times 5 \times \cdots \times (2n-5),$$

as required. Its not hard to see that the last expression is also equal to $\frac{(2n-4)!}{(n-2)!2^{n-2}}$. □

Recall that if k is an odd integer, $k!!$ denotes the product $k \times (k-2) \times (k-4) \times \cdots \times 1$. Using this notation, $b(n) = (2n-5)!!$. Furthermore, applying Stirling's formula $n! \sim \sqrt{2\pi n} \cdot (\frac{n}{e})^n$ to the second formula for $b(n)$ in gives the asymptotic equivalences

$$b(n) \sim \sqrt{2}\ (\frac{2}{e})^{n-2}\ n^{n-2}.$$

This shows that phylogenetic binary X-tree topologies grows exponentially with number n. Note that, the rate of growth in the number of phylogenetic trees is an important factor for the reconstruction of them from varied types of data. Usually, we want to choose a best tree under some criteria. Obviously, when n is large, it is computationally impracticable by querying all possible trees. Indeed, this is motivation for expanding novel mathematical techniques which are going to find optimal trees (or near optimal) among of all trees.

2.1.1 Rooted binary phylogenetic X-trees

In this part, we extend the notion of unrooted binary phylogenetic X-trees to rooted ones. Furthermore, we point out some of the essential interplays between unrooted and rooted binary phylogenetic X-trees. In biology binary phylogenetic trees and their rooted counterparts can be used to represent evolutionary relationships. In particular, for a rooted phylogenetic tree $\mathcal{T}$ on X, the branching indicates evolutionary relationships. Also, the edges are directed away from the root ρ. This root depicts the common ancestor of the life forms. Leaves of tree $\mathcal{T}$ represent the biological connection of extant species of the set X. Also we can view the internal vertices of $\mathcal{T}$ as corresponding to past speciation events. We can obtain a correspondent rooted phylogenetic X-tree with root vertex ρ, for given a (unrooted) binary phylogenetic X-tree $\mathcal{T}$, as follows. consider $B(n)$ as like before and let $B_R(n)$ denote the set of rooted binary phylogenetic tree with label set $\{1, 2, \cdots, n\}$. We describe a natural bijection between $B(n+1)$ and $B_R(n)$.

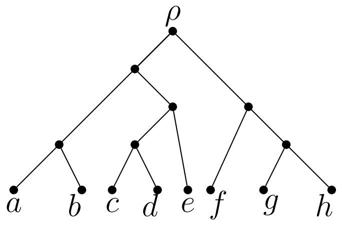


Figure 2.2: A rooted binary phylogenetic tree. Edges are directed down the page.

Let $\phi : B(n+1) \to B_R(n)$ be the map that deletes from a tree in $B(n+1)$ the leaf labeled $n+1$ and its incident edge, and then roots the resulting tree at the remaining end-vertex of this edge. We now get the following result.

Corollary 2.1.3. *For all $n \geq 2$,*

$$|B_R(n)| = |B(n+1)| = (2n-3)!! = \frac{(2n-2)!}{(n-1)!\ 2^{n-1}}.$$

During the last few years, there are many studies in phylogenetic bioinformatics in developing algorithms to reconstruct and model reticulation events (for example, see [57, 11, 34, 36]). Not surprisingly, phylogenetic networks bring many new complications. For example, for phylogenetic algorithms, the typical parameter of interest is the size of a family of phylogenetic networks. This implies that it is not always possible to establish a sufficient algorithm to find the best network without considering this parameter. Indeed, without a predetermined class of phylogenetic networks in mind, we can use the methods of analytic combinatorics to obtain precise estimates of various quantities for phylogenetic networks.

2.2 Analytic combinatorics

> *"We may loosely describe combinatorics as the branch of mathematics concerned with selecting, arranging, constructing, classifying, and counting or listing things."*
>
> - Robin J. Wilson, 2016

The focus of this thesis with regards to the preceding definition lies on the enumeration of objects, which are mostly described by recursions and boundary conditions, namely phylogenetic networks. A standard tool in this context are generating functions which were introduced as formal power series whose coefficients give the sizes of a sought family of objects with respect to a parameter encoded in the exponent. The main reference of this section is the book [20].

2.2.1 Formal power series

Informally, we can consider a formal power series as the following expression

$$A(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots ,$$

like a polynomial but continuing for ever. However, we need to proceed in a more precise mathematical manner to answer questions like what z is, or what the infinite sum means.

Definition 2.2.1 ([37])**.** *A formal power series is an infinite sequence of* $(a_0, a_1, a_2, \cdots)$ *elements taken from a commutative ring with identity* R*. A polynomial is a formal power series*(a_0, a_1, a_2) *for which there is some natural number* n *such that* $a_i = 0$ *for* $i > n$*; the smallest such* n *is the degree of the polynomial. We always think of a formal power series as represented in the form*

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots = \sum_{n \geq 0} a_n z^n.$$

For now, it is just an appropriate way of writing it, however we will see it can be connected with the concept of power series in analysis. We use language based on this; we often call a_n the "coefficient of z_n", and refer to a_0 as the "constant term" of the power series. We can consider many manipulations on formal power series which can give them their flexibility and various applicability.
Addition We define the sum of two formal power series term by term:

$$(a_0, a_1, a_2, \cdots) + (b_0, b_1, b_2, \cdots) = (a_0 + b_0, a_1 + b_1, a_2 + b_2, \cdots)$$

or said otherwise,

$$\sum_{n\geq 0} a_n z^n + \sum_{n\geq 0} b_n z^n = \sum_{n\geq 0} (a_n + b_n) z^n.$$

From now on we will just use the second form; but we can always go back to the first form if required.

Multiplication We define the product of two formal power series by the convolution formula

$$\big(\sum_{n\geq 0} a_n z^n\big)\big(\sum_{n\geq 0} b_n z^n\big) = \sum_{n\geq 0} c_n z^n,$$

where

$$c_n = \sum_{k\geq 0}^{n} a_k b_{n-k}.$$

Differentiation We can differentiate formal power series; no calculus involved, except that we steal from calculus the idea that the derivative of z^n is nz^{n-1}. So

$$\frac{d}{dz}\sum_{n\geq 0} a_n z^n = \sum_{n\geq 0} n a_n z^{n-1},$$

More details on formal power series can be found in [31, 59]. In the end, we want to recall some important power series expansions:

$$\frac{1}{1-z} = \sum_{n\geq 0} z^n, \qquad (1+z)^\alpha = \sum_{n\geq 0} \binom{\alpha}{n} z^n, \qquad e^z = \sum_{n\geq 0} \frac{1}{n!} z^n,$$

where $\binom{\alpha}{n} = \alpha(\alpha-1)\cdots(\alpha-n+1)/n!$. It is convenient to introduce another notation, namely,

$$[z^n]A(z) := a_n,$$

which extracts the coefficient of $A(z)$ at z^n.

Connection with analysis

We have seen that without any attention to whether or not formal power series converge we can manipulate them. Also, it is possible to look at formal power series over rings where convergence may not make a sense. But the good news is that, over the real or complex numbers, if our series are convergent for some non-zero values of x, then we can use all the tools of analysis on them. Note that the most vital case of that is the following point: any identity between real or complex power series, involving addition, multiplication (possibly infinite sums and products) and substitution, is an identity in the ring of formal power series. It comes from the uniqueness of the Taylor series for an analytic function. We will see some examples later. Note that the definitions of analytic, holomorphic and meromorphic functions as well as the basics of the analysis of singularities are left to more focused texts. Also, there is another less formal but often useful method which is known as *Cauchy's formula* to extract the coefficients of a power series. If $f(z) = \sum_{n\geq 0} a_n z^n$ is analytic in a disc Ω containing 0 and let λ be a simple loop around 0 in Ω that is positively oriented. Then

$$a_n = \frac{1}{2\pi i} \int_{\lambda} \frac{f(z)}{z^{n+1}} dz.$$

2.2.2 Generating functions: some basics tools and techniques

"Generating functions are the central objects of the theory, rather than a mere artifact to solve recurrences, as it is still often believed."

- Philippe Flajolet, 2007

Generating functions are well-known analytic tools in combinatorics and analysis of algorithms. We can use them for various purposes like in stating exact and recurrence formulas, finding asymptotic expansions, proving combinatorial identities, and other statistical properties, deriving averages, and variances. In particular, we construct generating functions for some popular combinatorial structures. Generally speaking, a *combinatorial class* is a collection $\mathcal{C}$ of objects of a similar kind (e.g. words, trees, graphs), endowed with a suitable notion of size or weight (which is a function $f : \mathcal{C} \longrightarrow \mathbb{N}$) in a way that there are only finitely many objects of each size. We denote by $\mathcal{C}_n$ the set of objects of size n in $\mathcal{C}$, and by c_n the cardinality of $\mathcal{C}_n$. Specifically, each combinatorial class we consider is a family of general phylogenetic networks, and the size of such a network is its number of vertices or leaves.

Objects of size n in $\mathcal{C}$ can be seen as an arrangement (following some rules to be precised) of n atoms, which are objects of size 1. In our context, these atoms are the vertices (or leaves) of the networks. In general, combinatorial objects may or may not be labeled, depending on whether the atoms constituting an object are distinguishable from one another (*labeled* case) or not (*unlabeled* case). Here, our networks will be labeled combinatorial objects.

To deal with a *labeled*[1] combinatorial class $\mathcal{C}$, we introduce the *exponential generating function* $C(z) = \sum_{n\geq 0} c_n \frac{z^n}{n!}$, which is a formal power series in z

[1] Although it is also very classical, the case of *unlabeled* objects (with their corresponding *ordinary* generating functions) will not be useful in our work, and is therefore omitted from our presentation.

displays the entire counting sequence of $\mathcal{C}$. The neutral class $\mathcal{E}$ is made of a single object of size 0, and its associated generating function is $E(z) = 1$. The atomic class $\mathcal{Z}$ is made of a single object of size 1, and its associated generating function is $Z(z) = z$.

Class	Nr. of elements	Weights	EGF
Neutral class	1	0	$E(z) = 1$
Atomic class	1	1	$Z(z) = z$

Table 2.1: Neutral and atomic classes.

A *specification* for a combinatorial class is an unambiguous description of the objects in the class using simpler classes and possibly the class itself. For instance, consider labeled rooted ordered binary trees, and define their size to be the number of their leaves. Such a tree is unambiguously described as being either a leaf or composed of a root to which a left and a right subtree are attached, which are themselves labeled rooted ordered binary trees, with a *consistent relabeling* of their atoms. By this, we mean the following: considering two trees whose atoms are labeled by $\{1, \ldots, k\}$ and $\{1, \ldots, k'\}$, we can build a tree using the first (resp. second) as left (resp. right) subtree; the atoms of this tree are labeled by $\{1, \ldots, k + k'\}$, and need to be such that the relative order between the labels in the left (resp. right) subtree is preserved (and they may be in any such way). This specification for labeled rooted ordered binary trees can be formally written as follows: $\mathcal{B} = \bullet \; \uplus \; \begin{array}{c} \circ \\ / \backslash \\ \mathcal{B} \;\; \mathcal{B} \end{array}$, where $\bullet$ represents a leaf (contributing 1 to the size of the object) and $\circ$ represents an internal node (which contributes 0 to the size).

Specifications describing (labeled) combinatorial classes can be translated into equations satisfied by the corresponding (exponential) generating functions. The precise statement that we refer to is [20, Theorem II.1]. The following proposition summarizes the simplest cases of this translation, which we will often use later in this thesis. Note that if $\mathcal{A}$ is a class, the size of an element $\alpha \in \mathcal{A}$ is denoted by $\mid \alpha \mid$, or $\mid \alpha \mid_{\mathcal{A}}$ in the few cases where the underlying class needs to be made explicit.

Proposition 2.2.2 (Dictionary). *Let $\mathcal{A}$ and $\mathcal{B}$ be two labeled combinatorial classes. Denote by $A(z)$ and $B(z)$ their respective exponential generating functions. Then the generating function of the class $\mathcal{C}$ which is the disjoint union of $\mathcal{A}$*

and $\mathcal{B}$ is $C(z) = A(z) + B(z)$ with size defined in a consistent manner : $\alpha \in \mathcal{C}$

$$\mid \alpha \mid_{\mathcal{C}}= \begin{cases} \mid \alpha \mid_{\mathcal{A}} & if \alpha \in \mathcal{A} \\ \mid \alpha \mid_{\mathcal{B}} & if \alpha \in \mathcal{B}. \end{cases}$$

Secondly, their Cartesian product $\mathcal{C} = \mathcal{A} \times \mathcal{B} = \{\gamma = (\alpha, \beta) | \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$ represents a new class with size defined consistently as $|\gamma|_{\mathcal{C}} = |\alpha|_{\mathcal{A}} + |\beta|_{\mathcal{B}}$. In this case we have to consider all possibilities in the manner of a Cauchy product, hence $\sum_{k=0}^{n} a_k b_{n-k}$, and we conclude as anticipated

$$C(z) = A(z) \cdot B(z) = \sum_{n \geq 0} \Big(\sum_{k=0}^{n} a_k b_{n-k} \Big) z^n.$$

These two constructions are enough to derive many fundamental constructions. For instance, if $\mathcal{A}$ contains no object of size 0, we can use sum and product in order to define the sequence class which consists of sequences of objects of $\mathcal{A}$ as the infinite sum

$$SEQ(\mathcal{A}) = \varepsilon + \mathcal{A} + (\mathcal{A} \times \mathcal{A}) + (\mathcal{A} \times \mathcal{A} \times \mathcal{A}) + \dots$$

(*i.e.*, m-tuples of objects of $\mathcal{A}$, for any $m \geq 0$), which gives us generating function $\dfrac{1}{1 - A(z)}$.

We now turn our attention to recursive specifications of a combinatorial class. As saw before, trees are best described recursively. Note that in the next sections we are going to describe decomposition of phylogenetic network that is based on tree structure which will then be translated into a functional equation involving their associated exponential generating functions.

Example 2.2.3. *We give here another proof of $|B_R(n)| = (2n - 3)!!$ based on generating functions and properties of Cauchy product. Clearly every tree $T \in B_R(X)$ is a single leaf or gives rise to two subtrees $T(\ell)$ and $T(r)$, where ℓ and r are the leaves of the $T(\ell)$ and $T(r)$ respectively. For the latter case, the union of those leaf-label sets is X. Conversely, given subsets X_ℓ and X_r with $X_\ell \cup X_r = X$, one can join a rooted binary tree on X_ℓ with a rooted binary tree on X_r to obtain a rooted binary tree on X. It follows that*

$$|B_R(X)| = \frac{1}{2} \sum_{X_\ell, X_r} |B_R(X_\ell)| \, |B_R(X_r)|,$$

where the sum is over all partitions $\{X_\ell, X_r\}$ of X; the factor $\frac{1}{2}$ accounts for the fact that there are two ways to designate ℓ and r as children of the root of T

. Since $|B_R(z)|$ depends only on $|X|$, and not on particular elements of X. Let $s(n) = |B_R(z)|$. We then have

$$s(n) = \frac{1}{2}\sum_{i=0}^{n}\binom{n}{i}s(i)\,s(n-i).$$

Consider the exponetial generating function $s(z) = \sum_{n\geq 0}\frac{s(n)}{n!}z^n$ for $s(n)$. We have

$$\begin{aligned}
s(z) &= \sum_{n\geq 0}\frac{1}{n!}\left(\frac{1}{2}\sum_{i=0}^{n}\binom{n}{i}s(i)\ s(n-i)\right)z^n \\
&= \frac{1}{2}\sum_{n\geq 0}\frac{1}{n!}\left(\sum_{i=0}^{n}\frac{n!\ s(i)s(n-i)}{i!\ (n-i)!}\right)z^n \\
&= \frac{1}{2}\sum_{n\geq 0}\left(\sum_{i=0}^{n}\frac{s(i)\ s(n-i)}{i!\ (n-i)!}\right)z^n \qquad (\textit{this is just a Cauchy product.}) \\
&= \frac{1}{2}\left(\sum_{n\geq 0}\frac{s(n)}{n!}z^n\right)\cdot\left(\sum_{n\geq 0}\frac{s(n)}{n!}z^n\right).
\end{aligned}$$

This expression for $s(z)$ translates into the more succinct equation

$$s(z) = \frac{s(z)^2}{2} + z.$$

The term "$+z$" accounts for the case where we have just a single isolated root vertex. Solving this equation, we have $s(z) = 1 \pm \sqrt{1-2z}$; however note that $\lim_{z\to 0}s(z) = 0$, and hence $s(z) = 1 - \sqrt{1-2z}$. We now expand $\sqrt{1-2z}$ using binomial theorem.

$$\sqrt{1-2z} = (1-2z)^{\frac{1}{2}} = \sum_{n\geq 0}\binom{\frac{1}{2}}{n}(-2z)^n.$$

Therefore,

$$\begin{aligned}
\frac{s(n)}{n!} &= [z^n]s(z) = [z^n](1-\sqrt{1-2z}) \\
&= [z^n]\left(1 - \sum_{n\geq 0}\binom{\frac{1}{2}}{n}(-2z)^n\right) \\
&= -\binom{\frac{1}{2}}{n}(-2)^n = \frac{(2n-3)!!}{n!}.
\end{aligned}$$

And thus we have the same result as in Corlllary 2.1.3.

The other possible way, especially in the case of tree-like objects, is to appeal to the *transfer theorem* (see [20], VI.1). Before going ahead, first we illustrate some concepts which help us to clarify the details. A *singularity* of an analytic function $f(z)$ is a point z_0 on the boundary of its region of analycity for which $f(z)$ is not analytically continuable. Singularities of a function analytic at 0, which lie on the boundary of the disc of convergence, are called dominant singularities. In this case, a dominant singularity is a singularity with smallest modulus. From *Pringsheim*'s theorem ([20], Theorem IV.6) we know that if $f(z)$ is representable at the origin by a series expansion that has non-negative coefficients and radius of convergence ρ, then the point $z = \rho$ is a singularity of $f(z)$. The idea behind the *transfer theorem* is that if $A(z)$ and $B(z)$ are two generating functions with the same positive real number ρ as dominant singularity; So when $z \rightarrow \rho$, we can write $A(z) \rightarrow B(z)$. We obtain the asymptotic expansion of $[z^n]A(z)$ by transferring the behaviour of $A(z)$ around its dominant singularity from a simpler function $B(z)$, from which we know the analytic behaviour.
A natural extension of the approach is to assume the error terms to be valid in the complex plane slit along the real half line $R_{\geq 1}$. In fact, weaker conditions suffice: any domain whose boundary makes an acute angle with the half line $R_{\geq 1}$ appears to be suitable.

Definition 2.2.4 (Δ-analytic)**.** *Given two numbers ϕ, R with $R > \rho$ and $0 < \phi < \frac{\pi}{2}$, the open domain $\Delta(\phi, R)$ is defined as*

$$\Delta(\phi, R) = \{z \big| |z| < R, z \neq \rho, |arg(z - \rho)| > \phi\}.$$

A domain is a Δ-domain at ρ if it is a $\Delta(\phi, R)$ for some R and ϕ. For a complex number τ, a Δ-domain at τ is the image by the mapping $z \rightarrow \tau z$ of a Δ-domain at ρ. A function is Δ-analytic if it is analytic in some Δ-domain.

Theorem 2.2.5 (Transfer Theorem)**.** *If the generating function $A(z)$ admits an expansion of the form $A(z) \sim c \cdot (1 - \frac{z}{\rho})^{-\alpha}$ as $n \rightarrow \infty$, around its (unique) dominant singularity ρ, then we have*

$$[z^n]A(z) \sim c \cdot \frac{n^{\alpha-1}}{\Gamma(\alpha)} \cdot \rho^{-n},$$

as $n \rightarrow \infty$.

Remark. Here $A(z)$ is analytic in the disk of radius ρ centered at the origin.

Example 2.2.6 (Unary-binary trees)**.** *We consider the species $\mathcal{T}$ of (unlabeled non-empty) planar unary-binary trees (i.e., each internal node has either one or two "de-scendants"). The following figure illustrates schematically the decomposition*

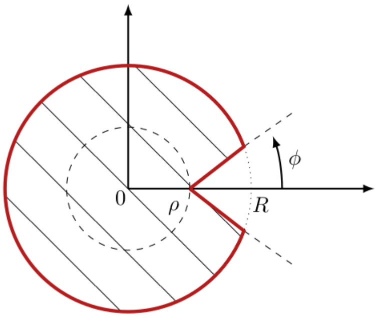
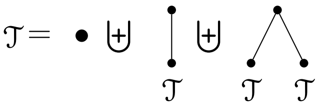


Figure 2.3: A typical Δ-domain at ρ.



of this species: This decomposition can be directly translated into the following functional equation for the (ordinary) generating function $T(z)$:

$$T(z) = z\big(1 + T(z) + T^2(z)\big).$$

Here we obtained a quadratic functional equation, which has the two possible solutions

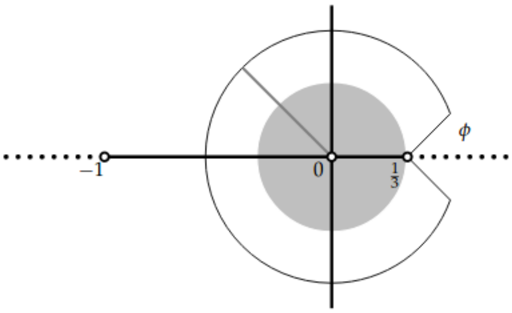
$$T_{\pm}(z) = \frac{1 - z \pm \sqrt{(1+z)(1-3z)}}{2z}.$$

Taking a closer look at $T_+(z)$, we see, that it possesses a singularity at 0, which corresponds to the constant term of the formal power series, and ought to be 1. Hence, we can dismiss this branch and arrive at the final solution

$$T(z) = \frac{1 - z - \sqrt{(1+z)(1-3z)}}{2z}.$$

The dominant singularity is visibly $z = 1/3$, and the function is analytic in a $\Delta-$domain. Around the point $1/3$, a singular expansion is obtained by multiplying $(1-3z)^{\frac{1}{2}}$ and the analytic expansion of the factor $(1+z)^{\frac{1}{2}}/(2z)$. The singularity analysis process and applying theorem 2.2.5 yields automatically

$$T(z) = 1 - 3^{\frac{1}{2}}\sqrt{1-3z} + O(1-3z) \quad \to \quad [z^n]T(z) = t_n = 3^n\sqrt{\frac{3}{4\pi n^3}} + O(3^n n^{-2}).$$



Recall that $[z^n]T(z)$ is the coefficient of z^n in $T(z)$, and so it is $\frac{c_n}{n!}$ (resp. c_n), when $T(z)$ is a exponential (resp.ordinary) generating function. Note that the location of a dominant singularity will give the exponential growth of the sequence, and the nature of this singularity the subexponential term. If $T(z)$ has several dominant singularities coming from *pure periodicities* (for more details see [20], IV.6.1), then the contributions from each of them must be combined.

These methods are fundamental results from complex analysis that allow to set up generating function in its disk of convergence, but not always. In particular, the *transfer theorem* (Theorem VI.1 of [20]) is one of the suitable tools, which allows us to derive asymptotic estimates of the coefficients of generating functions.

Theorem 2.2.7. *(Singular Inversion Theorem, [20, Theorem A.2]). Let $C(z)$ be a generating function such that $C(0) = 0$, satisfying the equation $C(z) = z\phi(C(z))$ for $\phi(z) = \sum_{n\geq 0} \phi_n z^n$ a power series such that $\phi_0 \neq 0$, all ϕ_n are non-negative real numbers, and $\phi(z) \neq \phi_0 + \phi_1 z$. Denote by R the radius of convergence of ϕ at 0. Assume that ϕ is analytic at 0 (so that $R > 0$), that the* characteristic equation *$\phi(z) - z\phi'(z) = 0$ has a solution $\tau \in (0, R)$ (that is necessarily unique), and that ϕ is aperiodic[2]. Then the following assertions hold:*

- *$\rho = \frac{\tau}{\phi(\tau)}$ is the radius of convergence of C at 0;*
- *near ρ, $C(z) \sim \tau - \sqrt{\frac{2\phi(\tau)}{\phi''(\tau)}}\sqrt{1 - \frac{z}{\rho}}$;*
- *when n grows, $[z^n]C(z) \sim \sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}} \frac{\rho^{-n}}{\sqrt{\pi n^3}}$.*

[2] Aperiodicity is needed only for the third item below. The definition of aperiodicity is omitted here, and can be found in [20, Definition IV.5]. A sufficient condition for a power series to be aperiodic (which applies to all examples considered in this thesis), is to have $\phi_n > 0$ for all n.

Example 2.2.8 (Cayley trees)**.** *Consider the class of labeled rooted unordered trees. Let $\mathcal{C}$ denote the set of these trees. Then $\mathcal{C}$ can be recursively described as a root followed by an unordered $k-$tuple of labeled rooted trees for some $k \geq 0$. This recursive description is then translated to specification*

$$C(z) = z + zC(z) + z\frac{C(z)^2}{2!} + z\frac{C(z)^3}{3!} + \cdots = ze^{C(z)}.$$

We thus get, $\phi(u) = \phi'(u) = \phi''(u) = e^u$. Also the equation $\phi(z) - z\phi'(z) = e^z - ze^z = 0$ has a solution $z = 1$. With help of Theorem 2.2.7, when n grows we get

$$[z^n]C(z) \sim \frac{1}{\sqrt{2\pi}} e^n n^{-\frac{3}{2}}.$$

Theorem 2.2.9 (Lagrange-Bürmann inversion)**.** *Let $\Phi(z)$ be a power series with $\Phi(0) \neq 0$ and $y(z)$ the (unique) power series solution of the equation*

$$y(z) = z\Phi(y(z)).$$

Then $y(z)$ is invertable and the n-th coefficient of $g(y(z))$ (where $g(z)$ is an arbitrary power series) is given by

$$[z^n]g(y(z)) = \frac{1}{n}[u^{n-1}]g'(u)\Phi(u)^n \qquad (n \geq 1).$$

We give an immediate application of Theorem 2.2.9. We have already observed that the generating function $C(z)$ of Cayley trees satisfies functional equation $C(z) = ze^{C(z)}$. With $\Phi(z) = e^z$ we obtain (for $n \geq 1$)

$$c_n = \frac{n!}{n}[z^{n-1}]e^{nz} = n^{n-1}.$$

2.2.3 Additive parameters and multivariate generating function

It is sometimes interesting to analyze the behaviour of other parameters than size. For example, interesting parameters for plane trees can be: height, number of leaves, path length, etc. These parameters are important for algorithm analysis as they correspond to the performance of algorithms that compute with or are modeled by plane trees. We now consider *multivariate* generating functions, where additional variables (x, y, ...) record the value of other parameters of our objects. One variable is used to track the size of the structure (e.g. number of nodes in a plane tree) and the other is used to track the parameter of interest (e.g. height, number of leaves, path length).

In our cases, we will consider one more such parameter, which are numbers of certain "unary nodes" occuring in our objects. Namely, denoting $c_{n,\ell}$ the number of objects of size n in the combinatorial class $\mathcal{C}$ such that the parameter has value ℓ, the multivariate exponential generating function we consider is $C(z,y) = \sum_{n,\ell} c_{n,\ell} y^{\ell} \frac{z^n}{n!}$.

For instance on the previous example of rooted plane trees consider one additional parameter, which is the number of leaves nodes. The coefficient of $z^n y^{\ell}$ in the generating function $T(z,y)$ is then the number of rooted plane trees with n nodes and exactly ℓ leaves, divided by $n!$.

The “dictionnary” translating combinatorial specifications to equations satisfied by the generating function extends to multivariate series, and our specification that shows any such tree is leaf or sequences (≥ 1) of trees that attached to the root nod. This gives $T(z,y) = zy + \dfrac{zT(z,y)}{1-T(z,y)}$. Let z be considered as a parameter. Using the Lagrange inversion theorem (2.2.9) yields

$$\begin{aligned}
t_{n,k} &= [y^k]\Big([z^n]T(z,y)\Big) \\
&= [y^k]\Big(\frac{1}{n}[v^{n-1}]\Big(y + \frac{v}{1-v}\Big)^n\Big) \\
&= \frac{1}{n}\binom{n}{k}[v^{n-1}]\frac{v^{n-k}}{(1-v)^{n-k}} \\
&= \frac{1}{n}\binom{n}{k}[v^{k-1}]\frac{1}{(1-v)^{n-k}} \\
&= \frac{1}{n}\binom{n}{k}\binom{n-2}{k-1}.
\end{aligned}$$

Moreover, under some hypotheses, the following theorem (see [14, Theorem 2.23]) allows to prove that the considered paramaters are asymptotically normally distributed. The notation used in the statement of this theorem is as follows: if F is a function of several variables, including v, F_v denotes the partial derivative of F with respect to v; as usual, $\mathbb{E}$ and $\mathbb{V}ar$ denote expectation and variance, respectively; $\mathcal{N}(0,1)$ is the standard normal distribution; and $\xrightarrow{d}$ denotes convergence in distribution.

Theorem 2.2.10. *Assume that $C(z,u)$ is a power series that is the (necessarily unique and analytic) solution of the functional equation $C = F(C,z,u)$, where $F(C,z,u)$ satisfies the following assumptions: $F(C,z,u)$ is analytic in C, z and u around 0, $F(C,0,u) = 0$, $F(0,z,u) \neq 0$, and all coefficients $[z^n C^m]F(C,z,1)$ are real and non-negative. Assume in addition that the region of convergence of $F(C,z,u)$ is large enough for having non-negative solutions $z = z_0$ and $C = C_0$*

of the system of equations

$$\begin{aligned} C &= F(C, z, 1) \\ 1 &= F_C(C, z, 1) \end{aligned}$$

with $F_z(C_0, z_0, 1) \neq 0$ *and* $F_{CC}(C_0, z_0, 1) \neq 0$.

Then, if X_n *is a sequence of random variables such that*

$$\mathbb{E}u^{X_n} = \frac{[z^n]C(z,u)}{[z^n]C(z,1)},$$

then X_n *is asymptotically normally distributed.*

More precisely, setting

$$\begin{aligned} \mu &= \frac{F_u}{z_0 F_z}. \\ \sigma^2 &= \mu + \mu^2 + \frac{1}{z_0 F_z^3 F_{CC}} \Big(F_z^2 (F_{CC} F_{uu} - F_{Cu}^2) - 2 F_z F_u (F_{CC} F_{zu} - F_{Cz} F_{Cu}) \\ &\quad + F_u^2 (F_{CC} F_{zz} - F_{Cz}^2) \Big) \end{aligned}$$

where all partial derivatives are evaluated at the point $(C_0, z_0, 1)$*, we have*

$$\mathbb{E}X_n = \mu n + O(1) \qquad and \qquad \mathbb{V}ar X_n = \sigma^2 n + O(1)$$

and if $\sigma^2 > 0$ *then*

$$\frac{X_n - \mathbb{E}X_n}{\sqrt{\mathbb{V}ar X_n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Chapter 3

Leaf-labeled Phylogenetic Networks Level 1 and Level 2

Phylogenetic networks generalize phylogenetic trees introducing reticulation vertices, which have two parents, and represent ancestral species resulting from the transfer of genetic material between coexisting species, through biological processes such as lateral gene transfer, hybridization or recombination. Recall that, binary phylogenetic networks are usually defined as rooted directed acyclic graphs whose vertices have either,the root (indegree 0 and outdegree 2), tree vertices (indegree 1 and outdegree 2), reticulation vertices (indegree 2 and outdegree 1) and leaves (indegree 1 and outdegree 0) such vertices being bijectively labeled by a set of *taxa*, which correspond to currently living species.

An important parameter that allows to measure the complexity of a phylogenetic networks is its "level". Phylogenetic trees are actually phylogenetic networks of level 0, and the level of a network N measures "how far from a tree" N is.

As trees, phylogenetic network can be rooted or unrooted. Ideally, phylogenetic networks should be rooted, the root representing the common ancestor of all taxa labeling the leaves. But several methods which reconstruct phylogenetic networks, such as distance-based or parsimony-based methods (for example see [30, 45]), do not produce inherently rooted networks.

The problem of enumerating (rooted or unrooted) trees is a very classical one in enumerative combinatorics. Solving this problem actually led to general methods for enumerating other tree-like structures, where generating functions play a key role. These methods have successfully been used by Semple and Steel [56] to enumerate two families of phylogenetic networks, namely unicyclic networks and unrooted level-1 networks (also called *galled trees*). Their results include an equation defining implicitly the generating function for unrooted level-1 networks (refined according to two parameters), which yields a closed formula for the number of unrooted level-1 networks with n (labeled) leaves, k cycles, and a total of

m edges (also called arcs) across all the cycles. An upper bound on the number of unlabeled galled trees is also provided in [9]. Other counting results have been more recently obtained on “galled networks” [32].

In this chapter, we extend the results of Semple and Steel in several ways. First, about unrooted level-1 networks, we provide an asymptotic estimate of the number of such networks with n (labeled) leaves. We also prove that the two parameters considered by Semple and Steel are asymptotically normally distributed. Second, we consider rooted level-1 networks, whose enumeration does not seem to have been considered so far in the literature. For these networks, we provide a closed formula counting them by number of leaves, together with an asymptotic estimate, and a closed formula for their enumeration refined by two parameters (the number of cycles and number of edges across all the cycles). Moreover, we show that these two parameters are asymptotically normally distributed. Finally, we consider both unrooted and rooted level-2 networks. Similarly, we provide in each case exact and asymptotic formulas for their enumeration, and prove asymptotic normality for some parameters of interest, namely: the number of bridgeless components of strictly positive level, and the number of edges across them. These parameters are a generalization for level-k networks ($k > 2$) of those considered by Semple and Steel for level-1 networks, in the sense that they quantify how different from a tree these phylogenetic networks are. Table 3.1 provides an overview of our results, and of where they can be found in this Chapter.

Type of network	Unrooted, level-1	Rooted, level-1	Unrooted, level-2	Rooted, level-2
Letter $\mathcal{X}$ denoting the class	$\mathcal{G}$ (**g**alled)	$\mathcal{R}$ (**r**ooted)	$\mathcal{U}$ (**u**nrooted)	$\mathcal{L}$ (**l**ast)
Eq. for the EGF $X(z)$	Thm. 3.2.1 $(*)$	Thm. 3.3.1	Thm. 3.4.1	Thm. 3.5.1
Exact formula for x_n	Thm. 3.2.1 $(*)$	Prop. 3.3.2	Prop 3.4.2	Prop. 3.5.2
Asymptotic estimate of x_n	Prop. 3.2.3	Prop. 3.3.3	Prop. 3.4.3	Prop. 3.5.3
Eq. for the multivariate EGF	Eq. (3.2.3) $(*)$	Eq. (3.3.4)	Eq. $(\star_U)$	Eq. $(\star_L)$
Asymptotic normality	Prop. 3.2.4	Prop. 3.3.5	Prop. 3.4.4	Prop. 3.5.4

Table 3.1: Overview of our main results. EGF means exponential generating function. The results marked with $(*)$ also appear in the work of Semple and Steel [56]. In addition, refined enumeration formulas for unrooted and rooted level-1 networks are provided in [56, Thm. 4] and Prop. 3.3.4 respectively. (Although the proof method applies to obtain such formulas for level-2 as well, the computations would however be rather intricate, and the interest *a priori* of the formulas so obtained questionable, hence our choice not to do it.)

3.1 Some definitions and notation

As illustrated in Figure 3.1, a binary rooted phylogenetic network N is said to be *level-k* (or called a *level-k network* for short) if the number of reticulation vertices contained in any bridgeless component of N is less than or equal to k. In a level-1 network N, each bridgeless component B having at least two vertices consists of the union of two directed paths, which start and end at the same vertices, called *source* and *sink* respectively. The source is actually either the root of N, or the head of a cut arc of N, and the sink is the unique reticulation vertex of B. Such brigdeless components are called *cycles*. (Note that a multiple edge is a particular case of such a cycle, where both directed paths consist of just one edge.)

Note that variations on the definition of rooted binary phylogenetic networks are around in the literature, and a few comments on our choice of definition are in order. Our definition of binary rooted phylogenetic networks allows multiple arcs, as in [28], but contrary to several other articles about phylogenetic networks. Our goal is indeed to study the most general model of leaf-labeled binary phylogenetic networks that could be counted if their number of leaves and their level are fixed. Note that for each bridgeless component B of binary rooted phylogenetic networks there exist at least two cut arcs of N whose tail[1] belongs to B and whose head does not belong to B. This condition is necessary to ensure that there are finitely many phylogenetic networks with a given number of leaves and level. Indeed, without it, such networks have unbounded number of vertices: this can be seen by replacing any cut arc of the network by a sequence of multiple arcs separated by cut arcs. Also it is a common technical condition for "recoverable"phylogenetic networks in which degenerated components known as "strongly redundant components"are excluded; see [63].

Similarly in some algorithmic-oriented papers about phylogenetic networks, bridgeless components with three vertices and two outgoing arcs are forbidden because the information needed to distinguish those components from simple tree vertices also connected with two outgoing arcs is not available in the input data. In the perspective of counting those objects we do not impose this restriction. But it could easily be added to our combinatorial descriptions and formulas below, to be taken into account if needed.

Definition 3.1.1. *An unrooted binary phylogenetic network is said to be level-k (or called an unrooted level-k network for short) if an unrooted binary phylogenetic tree can be obtained by first removing at most k edges per bridgeless component, then contracting each degree-2 vertex with one of its neighbours. We denote by* cycles *the bridgeless components of unrooted level-1 networks having*

[1] The *tail* of an arc is by definition its starting point. Its arrival point is called *head*.

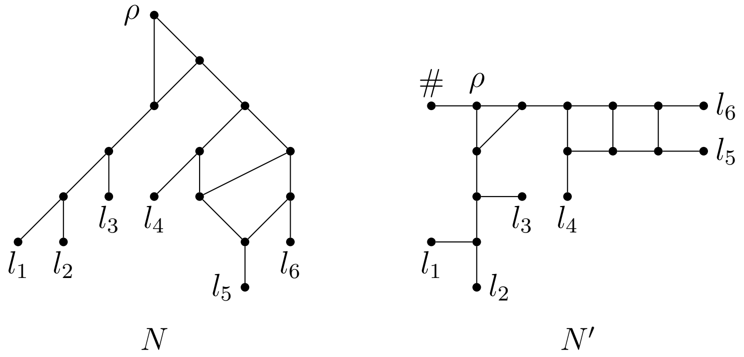


Figure 3.1: A rooted level-2 network N (where all arcs are directed downwards) and the unrooted level-2 network N' obtained by applying the unrooting procedure on N.

strictly more than one vertex. (Indeed, they are just cycles – of size at least 3 *– in the graph-theoretical sense.)*

Note that given a rooted level-k network N on n leaves, we can obtain an unrooted binary phylogenetic network N' on $n + 1$ leaves with the following *unrooting procedure*: add a vertex adjacent to the root of N, labeled with an extra leaf label (usually denoted $\#$), and ignore all arc directions. Theorem 1 of [29] implies in addition that the network N' so obtained is an unrooted level-k network. This unrooting procedure which consists of building an unrooted level-k network from a rooted level-k network, illustrated in Figure 3.1, can be reversed (see Lemma 4.13 of [38]), although not in a unique fashion. Indeed, given an unrooted level-k network N' on $n + 1$ leaves, it is possible to choose any leaf and delete it, making its neighbour the root ρ of a rooted level-k network N obtained by one of the following actions

1. placing the bridgeless component B containing ρ at the top;
2. orienting downwards all the cut edges incident with vertices of B;
3. choosing the tail t of one of these cut arcs as the sink of B;
4. computing an ρ-t numbering [41] on the vertices of B if there are more than one, that is labeling vertices of B with integers from 1 to the number n_B of vertices of B, such that the labels of ρ and t are respectively 1 and n_B and such that any vertex of B except ρ and t is adjacent both to a vertex with a lower label and a vertex with a higher label;

5. orienting each edge of B by choosing its vertex with the lower label as the tail;
   and

6. moving downwards into the network, recursively applying this procedure on all other bridgeless components.

This correspondence is not one-to-one because of the choices of the leaf which is deleted, and most importantly because of the choices of sinks in step 3 above.

3.1.1 Decomposition of rooted and unrooted level-k networks

For any bridgeless component B with $k_B \leq k$ reticulation vertices of a rooted level-k network N, the directed multigraph obtained by removing all outgoing arcs and contracting indegree-1 outdegree-1 vertices with their parent is called a *level-k_B generator* [62, 28]. For each $k > 0$, there exists a finite list of *level-k generators* which can be built from level-$(k-1)$ generators [28]. Therefore, depending on the level k_ρ of the bridgeless component B_ρ of N containing its root ρ, N can be decomposed in the following way. It is either

- a single leaf if $k_\rho = 0$ and ρ has outdegree 0;
- a root ρ being the parent of the root ρ_1 of a rooted level-k network N_1 and of the root ρ_2 of a rooted level-k network N_2 with disjoint sets of leaf labels, if $k_\rho = 0$ and ρ has outdegree 2;
- a level-k_ρ generator G_ρ containing the root, with $0 < k_\rho \leq k$, whose arcs are subdivided to create new indegree-1 outdegree-1 vertices, to which we add a set of cut arcs, whose tails are the outdegree-0 vertices of G_ρ and the newly created indegree-1 outdegree-1 vertices, and whose heads are roots of rooted level-k networks with disjoint sets of leaf labels.

Similarly, for any bridgeless component B of an unrooted level-k network N, the multigraph obtained by first removing all outgoing arcs, then contracting with one of its neighbours each vertex having exactly two distinct neighbours, is called an *unrooted level-k_B generator* [29, 33]. An unrooted level-k_B generator can also be defined as a single vertex for $k_B = 0$, as two vertices linked by a multiple edge for $k_B = 1$, and as a 3-regular bridgeless multigraph with $2k_B - 2$ vertices for $k_B > 1$ (Lemma 6 of [33]). Therefore, by considering a leaf $l_\#$ of any unrooted level-k network N and the bridgeless component B containing the vertex adjacent to this leaf, depending on the level k_B of B, N can be decomposed in the following way.

- If $k_B = 0$ and B consists of a single vertex of degree 1 in N, then N is just the leaf $l_\#$ adjacent to another leaf.

- If $k_B = 0$ and B is not a single vertex of degree 1 in N, then the leaf $l_\#$ is adjacent to a vertex v of degree 3 in N, such that the other two edges incident to v are cut edges. N is described by the edge between $l_\#$ and v, plus the two other edges incident with v, which are in turn identified with edges of two unrooted level-k networks N_1 and N_2 with disjoint sets of leaf labels (not containing $\#$) in such a way that v is identified with a leaf $l_{\#1}$ (resp. $l_{\#2}$) of N_1 (resp. N_2), removing the leaf labels of $l_{\#1}$ and $l_{\#2}$ during this identification.

- Otherwise $0 < k_B \leq k$. In this case, N is described by taking a level-k_B generator whose edges are subdivided to insert vertices, and then performing identification of these inserted vertices (in a same flavour as in the previous case). Specifically, one of these inserted vertices is identified with the neighbour of $l_\#$ in N, and all others are identified with leaves of unrooted level-k networks with disjoint sets of leaf labels (not containing $\#$). Again, each leaf that is identified with another vertex looses its label during this identification.

These decompositions of rooted and unrooted level-k networks will be the key to our counting results below.

n	g_{n-1}	r_n	u_{n-1}	ℓ_n
1	0	1	0	1
2	1	3	1	18
3	2	36	6	1 143
4	15	723	135	120 078
5	192	20 280	5 052	17 643 570
6	3 450	730 755	264 270	3 332 111 850
as $n \to \infty$	$c_1 \approx 0.20748$	$c_1 \approx 0.1339$	$c_1 \approx 0.07695$	$c_1 \approx 0.02931$
$x_n \sim c_1 c_2^n n^{n-1}$ with	$c_2 \approx 1.89004$	$c_2 \approx 2.943$	$c_2 \approx 5.4925$	$c_2 \approx 15.4333$
OEIS reference	A328121	A328122	A333005	A333006

Table 3.2: The numbers of rooted and unrooted level-1 or level-2 networks on n leaves.

3.2 Counting unrooted level-1 networks

3.2.1 Generating function and exact enumeration formula

Unrooted level-1 networks (also called unrooted galled trees) have been enumerated in [56]. The enumeration does not only consider the number of leaves of the galled trees, but is refined according to two parameters: the number of cycles (*i.e.*, level-1 generators) and the total number of edges which are part of a cycle (that we will call *inner edges*). We only reproduce in Theorem 3.2.1 and Proposition 3.2.2 a simplified version of the results of [56], taking into account the number of leaves only.

Theorem 3.2.1. *For any $n \geq 0$, let g_n denote the number of unrooted level-1 networks with $(n{+}1)$ leaves, and denote by $G(z) = \sum_{n\geq 0} g_n \frac{z^n}{n!}$ the corresponding generating function. Then G satisfies the following equation:*

$$G(z) = z + \frac{1}{2}G(z)^2 + \frac{1}{2}\frac{G(z)^2}{1-G(z)},$$

or equivalently

$$G(z) = z\phi(G(z)) \text{ with } \phi(z) = \frac{1}{1-\frac{1}{2}z(1+\frac{1}{1-z})}.$$

Proposition 3.2.2. *For any $n \geq 0$, let g_n denote the number of unrooted level-1 networks with $(n+1)$ leaves. We have*

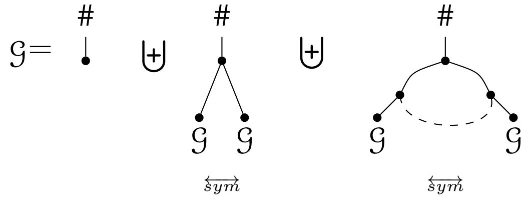
$$g_n = \frac{(2n-2)!}{2^{n-1}(n-1)!} + \sum_{1\leq i\leq k\leq n-1} \frac{(n+i-1)!(n+k-i-2)!}{k!(k-1)!(i-k)!(n-i-1)!}2^{-i}.$$

Notice that even if the formulas seem different, Proposition 3.2.2 can be recovered from Theorem 4 of [56] by summing over k and m and performing the change of variable $m = n - i + 3k - 1$. The first values of g_n have been included in Table 5.2.

Proof. We recall the main steps of the proofs of Theorem 3.2.1 and Proposition 3.2.2 given in [56].

Since counting rooted objects is far easier than counting unrooted objects, we establish a bijective correspondence between unrooted level-1 networks, and a rooted version of these networks, that we call *pointed* level-1 networks. Pointed level-1 networks on a set of taxa X are simply unrooted level-1 networks on the set of taxa $X \uplus \{\#\}$, where we declare that the leaf labeled by $\{\#\}$ is the "root" of the network. This provides a bijection between unrooted level-1 networks on the set of taxa $X \uplus \{\#\}$ and pointed level-1 networks on X, that have a root labeled by $\#$. Therefore, there are as many unrooted level-1 networks on the set of taxa $X \uplus \{\#\}$ as pointed level-1 networks on X rooted in a leaf labeled by $\# \notin X$. Hence g_n is the number of pointed level-1 networks with n leaves in addition to the root.

In a pointed level-1 network N (with at least two leaves), we consider the other extremity of the edge to which the root belongs. This vertex may belong to a cycle or not. In the latter case, N is simply described as an unordered pair of two pointed level-1 networks. In the former case, it is described as a non-oriented sequence of at least two pointed level-1 networks. Taking into account the trivial pointed level-1 network with one leaf, a specification for the pointed level-1 networks is therefore as follows:



$\mathcal{G}=$
#
⊎
#
$\mathcal{G}$ $\mathcal{G}$
$\overleftrightarrow{sym}$
⊎
#
$\mathcal{G}$
$\mathcal{G}$
$\overleftrightarrow{sym}$

where an arrow labeled by *sym* indicates that there is a symmetry w.r.t. the vertical axis to take into account, and the dashed edge corresponds to an edge or a path with internal vertices that are incident with cut-edges, themselves identified with edges of other pointed level-1 networks, the vertex lying on the cycle being identified with a leaf of corresponding network. Thanks to the “dictionary”, the generating function therefore satisfies $G(z) = z + \frac{1}{2}G(z)^2 + \frac{1}{2}\frac{G(z)^2}{1-G(z)}$ as claimed by Theorem 3.2.1. The end of Theorem 3.2.1 is obtained by simple algebraic manipulations.

From $G(z) = z\phi(G(z))$, where $\phi(z) = \frac{1}{1-\frac{1}{2}z(1+\frac{1}{1-z})}$, we can apply Lagrange inversion to find g_n. Indeed, $g_n = n![z^n]G(z) = (n-1)![z^{n-1}]\phi(z)^n$.

Recall the following expansion of $(1-z)^{-n}$, for any $n \geq 1$, which will be used here and several times later on:

$$\left(\frac{1}{1-z}\right)^n = \sum_{i\geq 0} \binom{n+i-1}{i} z^i.$$

Applying this identity twice and the binomial theorem, we get

$$\begin{aligned}\phi(z)^n =&\sum_{i\geq 0}\binom{n+i-1}{i}\left(\frac{1}{2}z(1+\frac{1}{1-z})\right)^i\\ =&\sum_{i\geq 0}\binom{n+i-1}{i}\left(1+\sum_{k=1}^{i}\sum_{p\geq 0}\binom{i}{k}\binom{k+p-1}{p}z^p\right)\frac{1}{2^i}z^i\\ =&\sum_{i\geq 0}\binom{n+i-1}{i}\frac{z^i}{2^i}+\sum_{i\geq 0}\sum_{k=1}^{i}\sum_{p\geq 0}\binom{n+i-1}{i}\binom{i}{k}\binom{k+p-1}{p}\frac{z^{i+p}}{2^i}.\end{aligned}$$

It follows that

$$\begin{aligned}[z^{n-1}]\phi(z)^n =&\binom{2n-2}{n-1}\frac{1}{2^{n-1}}+\sum_{i=0}^{n-1}\sum_{k=1}^{i}\frac{1}{2^i}\binom{n+i-1}{i}\binom{i}{k}\binom{n+k-i-2}{n-i-1}\\ \text{and } g_n =&\frac{(2n-2)!}{2^{n-1}(n-1)!}+\sum_{1\leq k\leq i\leq n-1}\frac{(n+i-1)!(n+k-i-2)!}{k!(k-1)!(i-k)!(n-i-1)!}2^{-i}. \qquad \square\end{aligned}$$

3.2.2 Asymptotic evaluation

From Theorem 3.2.1, we can furthermore derive an asymptotic evaluation of the number g_n of unrooted level-1 networks on $(n+1)$ leaves, using Theorem 2.2.7.

Proposition 3.2.3. *The number g_n of unrooted level-1 networks on $n+1$ leaves is asymptotically equivalent to $c_1 \cdot c_2^n \cdot n^{n-1}$ for constants $c_1$and c_2 such that $c_1 \approx 0.20748$ and $c_2 \approx 1.89004$.*

Proof. Recall that $G(z)$ satisfies $G(z) = z\phi(G(z))$, where $\phi(z) = \frac{1}{1-\frac{1}{2}z(1+\frac{1}{1-z})}$. Equivalently, this can be rewritten as $\phi(z) = \frac{2-2z}{z^2-4z+2}$. So, $\phi(z)$ is a rational fraction, whose pole with smallest absolute value is $2-\sqrt{2} \approx 0.5858$. As such, $\phi(z)$ is analytic at 0, with radius of convergence $R = 2-\sqrt{2}$. Moreover, owing to footnote 2, $\phi(z)$ is aperiodic. Finally, the characteristic equation $\phi(z) - z\phi'(z) = 0$ can be numerically solved (see companion Maple worksheet), showing that it admits a unique solution in the disk of convergence of ϕ, namely $\tau \approx 0.34270$. Therefore, the hypotheses of Theorem 2.2.7 are all satisfied, and denoting $\rho = \frac{\tau}{\phi(\tau)} \approx 0.19464$, Theorem 2.2.7 gives

$$[z^n]G(z) \sim \sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}}\frac{\rho^{-n}}{\sqrt{\pi n^3}}.$$

Using the Stirling estimate of the factorial $n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$, we get

$$g_n \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}}\frac{\rho^{-n}}{\sqrt{\pi n^3}} \sim \frac{n^{n-1}}{(e\rho)^n}\sqrt{\frac{\phi(\tau)}{\phi''(\tau)}}.$$

Replacing τ and ρ by their numerical approximations, we obtain the announced result. □

3.2.3 Refined enumeration and asymptotic distribution of parameters

From the specification of pointed level-1 networks seen in the proof of Theorem 3.2.1, it follows easily, as done in [56], that the multivariate generating function $G(z,x,y) = \sum_{n,k,m} \frac{g_{n,k,m}}{n!} z^n x^k y^m$, where $g_{n,k,m}$ is the number of unrooted level-1 networks with $n+1$ leaves, k cycles, and m inner edges, satisfies

$$G(z,x,y) = z + \frac{1}{2}G(z,x,y)^2 + \frac{1}{2}xy^3\frac{G(z,x,y)^2}{1-yG(z,x,y)}.$$

This equation can be rewritten as $G(z,x,y) = z\phi(G(z,x,y),x,y)$ where ϕ is defined by $\phi(z,x,y) = \frac{1}{1-\frac{1}{2}z\left(1+\frac{xy^3}{1-yz}\right)}$. As done in [56], we can apply the Lagrange inversion formula to obtain an explicit expression for $g_{n,k,m}$ – see [56, Thm. 4].

Using Theorem 2.2.10, the above equation may also be used to prove that the parameters “number of cycles” and “number of inner edges” are both asymptotically normally distributed.

Proposition 3.2.4. *Let X_n (resp. Y_n) be the random variable counting the number of cycles (resp. inner edges) in unrooted level-1 networks with $n+1$ leaves. Both X_n and Y_n are asymptotically normally distributed, and more precisely, we have*

$$\mathbb{E}X_n = \mu_X n + O(1), \qquad \mathbb{V}arX_n = \sigma_X^2 n + O(1) \quad \textit{and} \quad \frac{X_n - \mathbb{E}X_n}{\sqrt{\mathbb{V}arX_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

$$\mathbb{E}Y_n = \mu_Y n + O(1), \qquad \mathbb{V}arY_n = \sigma_Y^2 n + O(1) \quad \textit{and} \quad \frac{Y_n - \mathbb{E}Y_n}{\sqrt{\mathbb{V}arY_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mu_X \approx 0.46$, $\sigma_X^2 \approx 0.18$, $\mu_Y \approx 1.61$ and $\sigma_Y^2 \approx 1.44$.

Proof. Consider first X_n. Defining $G(z,x) := G(z,x,1)$, it holds that

$$\mathbb{E}x^{X_n} = \frac{[z^n]G(z,x)}{[z^n]G(z,1)}.$$

It follows from the equation for $G(z,x,y)$ that $G(z,x) = F(G(z,x),z,x)$, where F is defined by $F(G,z,x) = z\frac{1}{1-\frac{1}{2}G\left(1+\frac{x}{1-G}\right)}$. Being rational, we see immediately that $F(G,z,x)$ is analytic in G, z and x around 0. Moreover, performing the substitution $z = 0$ (resp. $G = 0$) gives $F(G,0,x) = 0$ (resp. $F(0,z,x) = z$, which is not identically 0). Finally, it is readily checked that F satisfies $[z^nG^m]F(G,z,1) \geq 0$ for all n,m In addition, we can determine numerically that the system

$$\begin{aligned} G =& F(G,z,1) \\ 1 =& F_G(G,z,1) \end{aligned}$$

admits a solution (G_0, z_0) such that $G_0 \approx 0.3427$ and $z_0 \approx 0.1946$, which satisfies the hypothesis of Theorem 2.2.10 (see the companion Maple worksheet to determine the solution and to check it satisfies the required hypotheses). The result then follows from Theorem 2.2.10, and the numerical estimates of μ_X and σ_X^2 are obtained plugging the numerical estimates for G_0 and z_0 into the explicit formulas given by Theorem 2.2.10 (see again companion Maple worksheet for details). The proof for Y_n follows the exact same steps, considering this time $G(z,y) := G(z,1,y)$ instead, and adjusting the definition of F accordingly. As expected, the solution (G_0, z_0) of the associated system is the same as above. □

Remark 1. In the above proof of Proposition 3.2.3 (resp. Proposition 3.2.4), we have provided some details on how Theorem 2.2.7 (resp. Theorem 2.2.10) was used and on how its hypotheses were checked. This is omitted in later proofs using Theorem 2.2.7 (see Propositions 3.3.3, 3.4.3 and 3.5.3) or Theorem 2.2.10 (see Propositions 3.2.4, 3.3.5 and 3.4.4), since they work following the exact same steps. Note also that all numerical resolutions of equations are done in the companion Maple worksheet[2].

3.3 Counting rooted level-1 networks

3.3.1 Combinatorial specification and generating function

As for unrooted level-1 networks, we start by a combinatorial specification that describes rooted level-1 networks (also called rooted galled trees). Because every cycle in a rooted level-1 network not only has a tree vertex above all other vertices of the cycle, but also a reticulation vertex which is below all other vertices of the cycle, notice that these objects are different from the pointed level-1 networks that we considered in the proof of Theorem 3.2.1 and Proposition 3.2.2.

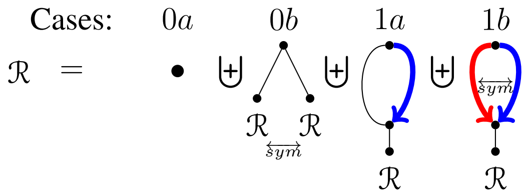
[2] at http://user.math.uzh.ch/bouvel/publications/BouvelGambetteMansouri_Version2_WithoutMultipleEdges.mw

Recall that each cycle of a level-1 network has stricly more than one outgoing arc (otherwise there would be an infinite number of level-1 networks on n taxa).

Let us denote by $\mathcal{R}$ the set of rooted level-1 networks. The size of a network of $\mathcal{R}$ is the number of its leaves. Distinguishing on the level (0 or 1) of the bridgeless component containing its root, a network of $\mathcal{R}$ is described in exactly one of the following ways. It may be:

- a single leaf (case $0a$);
- a binary root vertex with two children that are roots of networks of $\mathcal{R}$, whose left-to-right order is irrelevant (case $0b$);
- a cycle containing the root with at least two outgoing cut arcs leading to networks of $\mathcal{R}$. This last possibility splits into two subcases, since the reticulation vertex of the cycle may be a child of the root:
  - a cycle whose reticulation vertex is attached to a network of $\mathcal{R}$, is a child of the root and is the lowest vertex of a path coming from the root, where a sequence of at least one network of $\mathcal{R}$ is attached (case $1a$);
  - a cycle whose reticulation vertex is attached to a network of $\mathcal{R}$, and such that a sequence of at least one network of $\mathcal{R}$ is attached to each path of this cycle, the left-to-right order of these two paths being irrelevant (case $1b$).

The specification for $\mathcal{R}$ is therefore given by (all arcs are directed downwards, the thick arcs each represent a directed path which contains at least one internal vertex incident with a cut arc):



Denoting r_n the number of rooted level-1 networks on n leaves, and $R(z) = \sum_{n\geq 0} r_n \frac{z^n}{n!}$ the associated exponential generating function, we deduce from the specification that

$$R = z + \frac{1}{2}R^2 + \frac{R^2}{1-R} + \frac{R}{2}\left(\frac{R}{1-R}\right)^2 .$$

Unlike for the other generating functions considered in this section, the above equation for R allows to find a closed formula for R. Indeed, the above equation has four solutions that can be made explicit with the help of a solver. We can further notice that evaluating the generating function $R(z)$ at $z = 0$, we must obtain $R(0) = r_0 = 0$. Among the four candidate solutions for R, we therefore select the only one which has value 0 for $z = 0$ and obtain an explicit form for $R(z)$, given in Theorem 3.3.1.

Theorem 3.3.1. *The exponential generating function $R(z)$ of rooted level-1 networks is expressed as*

$$R(z) = \frac{5 - \sqrt{1-8z} - \sqrt{18 - 8z - 2\sqrt{1-8z}}}{4}$$

, within its disk of convergence of radius $\frac{1}{8}$.

3.3.2 Exact enumeration formula

The first terms of the sequence $(r_0, r_1, r_2, \ldots)$ can be read on the Taylor expansion of $R(z)$, and have been collected in Table 5.2:

$$R(z) = z + 3\frac{z^2}{2!} + 36\frac{z^3}{3!} + 723\frac{z^4}{4!} + 20280\frac{z^5}{5!} + o(z^5).$$

More generally, we have:

Proposition 3.3.2. *For any $n \geq 1$, the number r_n of rooted level-1 networks with n leaves is given by*

$$\frac{(2n-2)!}{2^{n-1}(n-1)!} + \sum_{\substack{1 \leq k \leq i \leq n-1 \\ 0 \leq p \leq k}} \frac{(n+i-1)!(n+k-i-2)! \quad 2^{p-i}}{(i-k)!(k-p)!p!(n-1-i-k+p)!(2k-p-1)!}.$$

Proof. To obtain a generic formula for r_n, we apply the Lagrange inversion formula, rewriting $R(z)$ as $R(z) = z\phi(R(z))$ where $\phi(z) = \frac{1}{1-\frac{1}{2}z-\frac{z}{1-z}-\frac{1}{2}\left(\frac{z}{1-z}\right)^2}$.

Using twice the usual development of $(1-z)^{-n}$ (for $n \geq 1$) and twice the binomial theorem, we obtain

$$\begin{aligned}\phi(z)^n =& \sum_{i \geq 0} \binom{n+i-1}{i} \frac{z^i}{2^i} \\ &+ \sum_{i \geq 0} \sum_{k=1}^{i} \sum_{p=0}^{k} \sum_{j \geq 0} \binom{n+i-1}{i} \binom{i}{k} \binom{k}{p} \binom{2k-p+j-1}{j} \frac{z^{i+k-p+j}}{2^{i-p}},\end{aligned}$$

and we deduce that

$$
\begin{aligned}
r_n &= n![z^n]R(z) = n!\frac{1}{n}[z^{n-1}]\phi(z)^n = (n-1)![z^{n-1}]\phi(z)^n \\
&= \frac{(2n-2)!}{2^{n-1}(n-1)!} \\
&+ \sum_{\substack{1 \le k \le i \le n-1 \\ 0 \le p \le k}} \frac{(n+i-1)!(n+k-i-2)!}{(i-k)!(k-p)!p!(n-1-i-k+p)!(2k-p-1)!} 2^{p-i}
\end{aligned}
$$

as announced. □

3.3.3 Asymptotic evaluation

The equation for $R(z)$ also enables us to derive an asymptotic estimate of r_n.

Proposition 3.3.3. *The number r_n of rooted level-1 networks on n leaves is asymptotically equivalent to $c_1 \cdot c_2^n \cdot n^{n-1}$ for $c_1 = \frac{\sqrt{34}(\sqrt{17}-1)}{136} \approx 0.1339$ and $c_2 = \frac{8}{e} \approx 2.943$.*

Proof. Recall that $R(z) = z\phi(R(z))$ where $\phi(z) = \frac{1}{1-\frac{1}{2}z-\frac{z}{1-z}-\frac{1}{2}\left(\frac{z}{1-z}\right)^2}$ so that we can apply the Singular Inversion Theorem. Unlike in the case of unrooted level-1 networks, the solution τ of the characteristic equation $\phi(z) - z\phi'(z) = 0$ to be considered has a nice explicit expression here, and we have $\tau = \frac{5-\sqrt{17}}{4}$. We obtain $\rho = \frac{\tau}{\phi(\tau)} = \frac{1}{8}$ and $\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}} = \frac{\sqrt{17}(\sqrt{17}-1)}{136}$. Consequently, from Theorem 2.2.7 we have

$$[z^n]R(z) \sim \frac{\sqrt{17}(\sqrt{17}-1)}{136} \frac{8^n}{\sqrt{\pi n^3}}.$$

Since $r_n = n![z^n]R(z)$, using the Stirling estimate of the factorial, we finally get

$$r_n \sim \frac{\sqrt{34}(\sqrt{17}-1)}{136} \left(\frac{8}{e}\right)^n n^{n-1}.$$

□

Notice that with the explicit expression of the generating function $R(z)$ in Theorem 3.3.1, another way of proving Proposition 3.3.3 would have been to use the Transfer Theorem (Corollary VI.1 of [20]). We do not enter the details of this other method here, but we can check that it gives the same result.

3.3.4 Refined enumeration formula

As in the work of Semple and Steel [56], we can refine the enumeration of rooted level-1 networks according to two additional parameters, which are typical of the "level-1" nature of our networks: their number of cycles and their total number of arcs among cycles. To do so, let us introduce the multivariate generating function $R(z, x, y) = \sum \frac{r(n,k,m)}{n!} z^n x^k y^m$, where $r(n, k, m)$ is the number of rooted level-1 networks with n leaves, k cycles and m inner arcs (i.e. the total number of arcs inside those k cycles is m). The specification for $\mathcal{R}$ translates into the following equation for $R = R(z, x, y)$:

$$R = z + \frac{1}{2}R^2 + x\frac{R^2y^3}{1 - yR} + \frac{1}{2}xR\left(\frac{Ry^2}{1 - yR}\right)^2.$$

The equation can be rewritten as follows:

$$R = z\phi(R, x, y) \text{ where } \phi(z, x, y) = \frac{1}{1 - \frac{1}{2}z - x\frac{zy^3}{1-yz} - \frac{1}{2}xy^4\left(\frac{z}{1-yz}\right)^2}.$$

Applying the Lagrange inversion formula again, we have

$$\frac{r(n, k, m)}{n!} = [z^n x^k y^m]R(z, x, y) = \frac{1}{n}[z^{n-1}x^k y^m]\phi(z, x, y)^n,$$

and by the exact same steps of computation as in the proof of Proposition 3.3.2, we get:

Proposition 3.3.4. *The number $r(n, k, m)$ of level-1 networks with n leaves, k cycles and m inner arcs (with $k \geq 1$ and $m \geq 1$) is*

$$r(n, k, m) = \sum_{p=0}^{k} \frac{(2n + 3k - m - 2)!(m - 2k - 1)!2^{m-n-2k-p+1}}{(n + 2k - m - 1)!p!(k - p)!(m - 3k - p)!(k + p - 1)!}.$$

Notice that from $r_n = r(n, 0, 0) + \sum_{k=1}^{n-1} \sum_{m=3k}^{n+2k-1} r(n, k, m)$ and the above theorem, we can recover Proposition 3.3.2 by the change of variable $m = n + 3k - i - 1$.

3.3.5 Asymptotic distribution of parameters

As we have seen with Proposition 3.2.4, the equation for the refined generating function does not only give access to the explicit formula of Proposition 3.3.4 above, but also allows to prove that the two parameters of interest are each asymptotically normally distributed.

Proposition 3.3.5. *Let X_n (resp. Y_n) be the random variable counting the number of cycles (resp. inner arcs) in rooted level-1 networks with n leaves. Both X_n and Y_n are asymptotically normally distributed, and more precisely, we have*

$$\mathbb{E}X_n = \mu_X n + O(1), \qquad \mathbb{V}ar X_n = \sigma_X^2 n + O(1) \quad \textit{and} \quad \frac{X_n - \mathbb{E}X_n}{\sqrt{\mathbb{V}ar X_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

$$\mathbb{E}Y_n = \mu_Y n + O(1), \qquad \mathbb{V}ar Y_n = \sigma_Y^2 n + O(1) \quad \textit{and} \quad \frac{Y_n - \mathbb{E}Y_n}{\sqrt{\mathbb{V}ar Y_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mu_X \approx 0.56$, $\sigma_X^2 \approx 0.18$, $\mu_Y \approx 1.93$ and $\sigma_Y^2 \approx 1.24$.

Proof. Recall that, defining $\phi(z,x,y) = \frac{1}{1-\frac{1}{2}z - x\frac{zy^3}{1-yz} - \frac{1}{2}xy^4\left(\frac{z}{1-yz}\right)^2}$, $R(z,x,y)$ satisfies $R = z\phi(R,x,y)$. We focus first on X_n, setting $y = 1$, and we consider $R(z,x) := R(z,x,1)$. We have

$$\mathbb{E}x^{X_n} = \frac{[z^n]R(z,x)}{[z^n]R(z,1)}.$$

Defining the function F by $F(R,z,x) = z\phi(R,x,1)$, we infer $R(z,x) = F(R(z,x),z,x)$. It is readily checked that F satisfies all hypotheses of Theorem 2.2.10. Moreover, the system

$$\begin{aligned} R =& F(R,z,1) \\ 1 =& F_R(R,z,1) \end{aligned}$$

admits a solution (R_0, z_0) with $z_0 = 1/8$ and $R_0 \approx 0.2192$, which satisfies the hypothesis of Theorem 2.2.10. The result and numerical estimates of μ_X and σ_X^2 then follow from Theorem 2.2.10.

For Y_n instead of X_n, the proof works in the exact same way, considering this time $R(z,y) := R(z,1,y)$ instead, and adjusting the definition of F accordingly. As in the proof of Proposition 3.2.4, we find the same solution (R_0, z_0) of the associated system, as it should be. □

3.4 Counting unrooted level-2 networks

3.4.1 Combinatorial specification

First of all, let us recall that any bridgeless component in an unrooted level-2 network has at least three outgoing cut-edges (since otherwise there would be an infinite number of such networks with a given number of leaves).

As in the case of level-1 unrooted networks, we consider *pointed* level-2 networks, that are unrooted level-2 networks equipped with a fictitious root, which is a new leaf labeled by the special taxa $\#$. This provides a bijection between unrooted level-2 networks on the set of taxa $X \uplus \{\#\}$ and pointed level-2 networks on X. Therefore, there are as many unrooted level-2 networks of the set of taxa $X \uplus \{\#\}$ as pointed level-2 networks on X rooted in a leaf labeled by $\# \notin X$. Notice that pointed level-2 networks do not correspond to classical rooted level-2 networks. Indeed, every bridgeless component in a pointed level-2 network has a distinguished vertex which could be considered as the equivalent of a root, but no reticulation vertices, whereas it has both in the usual definition of rooted level-2 networks.

Let us denote by $\mathcal{U}$ the set of such pointed level-2 networks, the size of a network of $\mathcal{U}$ being the number of its leaves different from the root. Let u_n be the number of networks of size n in $\mathcal{U}$, the above argument shows that u_n counts the number of unrooted level-2 networks on $(n+1)$ leaves. We introduce $U(z) = \sum_{n\geq 0} u_n \frac{z^n}{n!}$ the associated exponential generating function.

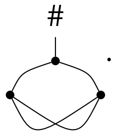
To obtain a combinatorial specification for $\mathcal{U}$, and hence an equation satisfied by $U(z)$, we describe the possible shapes of a network N of $\mathcal{U}$, depending on the level (0, 1 or 2) of the bridgeless component that contains the neighbouring vertex of the fictitious root.

- Of course, we must start with the trivial case in which the fictitious root is attached directly to a leaf.
  For the remaining cases, denote by v the vertex at the other extremity of the edge incident to the fictitious root.

- If v does not belong to a cycle nor to a bridgeless component of level 2, then N is described as an unordered pair of two pointed level-2 networks.

- If v belongs to a cycle but not to a bridgeless component of level 2, then N is described as an unoriented sequence of at least two pointed level-2 networks.
  (These first three cases are the same as in Section 3.3.)

- The last possibility is that v belongs to a bridgeless component of level 2.

  The underlying level-2 generator, G, is necessarily of the shape  .

In the following, we distinguish many cases depending on whether each edge of the level-2 generator contains exactly one vertex incident with a cut-edge, several or none.

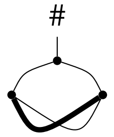
3.4.2 Case analysis for unrooted level-2 generators

In the pictures below, we use thick lines to represent paths containing at least 2 internal nodes incident with a cut-edge which is incident with another pointed unrooted level-2 network. We use # to represent the fictitious root in the pointed network, v to denote its neighbour, and $\mathcal{U}$ to represent any pointed network.

Case 1: One edge with attached networks

One edge of the generator carries a sequence of at least two outgoing arcs. Because multiple edges are not allowed, it cannot be one of the two edges incident to v. So, it can be only one of the two edges not incident to v (which are not distinguished). The sequence is unoriented, because of symmetry, explaining the factor $\frac{1}{2}$ below.

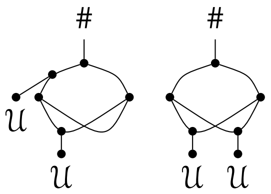
$$\frac{U^2}{2(1-U)}$$



Case 2: Two edges with attached networks

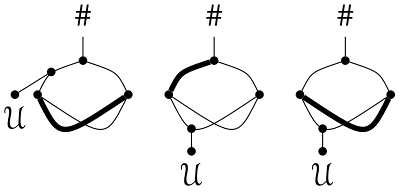
Case 2A - Two edges of the generator carry exactly one outgoing arc. Since multiple edges are not allowed, it can either be one edge incident to v and one not, or both edges not incident to v. In the latter case, the two edges should not be distinguished, hence the factor $\frac{1}{2}$.

$$U^2 + \frac{U^2}{2} = \frac{3}{2}U^2$$



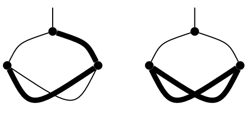
Case 2B - One edge of the generator carries a single outgoing arc and another edge carries a sequence of at least two outgoing arcs. Again, these cannot be the two edges incident to v. The only case where symmetries need to be taken care of is when the two edges are those not incident to v: in this case, the sequence is not oriented, hence the factor $\frac{1}{2}$. In all other cases, the orientation of the sequence is determined by the presence of the fictitious root or the outgoing arc from the other edge with and attached network.

$$\frac{U^3}{1-U} + \frac{U^3}{1-U} + \frac{U^3}{2(1-U)} = \frac{5U^3}{2(1-U)}$$



Case 2C - Two edges of the generator (but not the two incident to v, as before) carry a sequence of at least two outgoing arcs. If one arc is incident to v and the other not, then both sequences are oriented and there is no symmetry factor. If the two arcs are those not incident to v, then the two sequences they carry can be seen as an unordered pair of oriented sequences, seen up to symmetry w.r.t. the vertical axis. This yields a factor $\frac{1}{2}$ since the pair is unordered, and another factor $\frac{1}{2}$ to account for the symmetry w.r.t. the vertical axis.

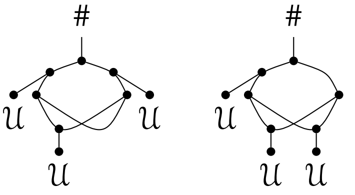
$$\frac{U^4}{(1-U)^2} + \frac{U^4}{4(1-U)^2} = \frac{5U^4}{4(1-U)^2}$$



Case 3: Three edges with no attached networks

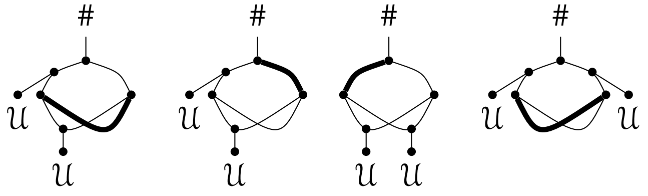
Case 3A - Three edges of the generator carry exactly one outgoing arc. The unused edge can either be incident v or not. In both cases, we have a factor $\frac{1}{2}$ because of symmetry.

$$\frac{U^3}{2} + \frac{U^3}{2} = U^3$$



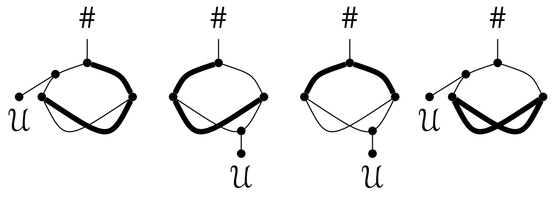
Case 3B - Two edges of the generator carry a single outgoing arc and one carries a sequence of at least two outgoing arcs. The only cases where a symmetry comes into play here are when the edges carrying single outgoing arcs are either the two edges incident to v or the two edges not incident to v. This yields the factor $\frac{1}{2}$ in these two cases. Moreover, all sequences are oriented, because of the presence of the fictitious root or the single outgoing arcs.

$$\frac{U^4}{1-U} + \frac{U^4}{1-U} + \frac{U^4}{2(1-U)} + \frac{U^4}{2(1-U)} = \frac{3U^4}{1-U}$$



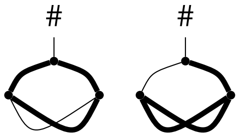
Case 3C - One edge of the generator carries a single outgoing arc and two edges carry a sequence of at least two outgoing arcs. Similarly to the previous case, we obtain a factor $\frac{1}{2}$ for symmetry reasons when the two edges carrying sequences are either the two edges incident to v or the two edges not incident to v. Moreover, all sequences are oriented, because of the presence of the fictitious root or the single outgoing arc.

$$\frac{U^5}{(1-U)^2} + \frac{U^5}{(1-U)^2} + \frac{U^5}{2(1-U)^2} + \frac{U^5}{2(1-U)^2} = \frac{3U^5}{(1-U)^2}$$



Case 3D - Three edges of the generator carry a sequence of at least two outgoing arcs. In both cases, we have a factor $\frac{1}{2}$ for symmetry reason, but all sequences are oriented by the presence of the fictitious root, or of the sequence on the edge(s) incident to v.

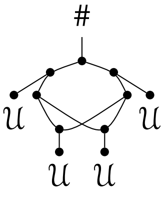
$$\frac{U^6}{2(1-U)^3} + \frac{U^6}{2(1-U)^3} = \frac{U^6}{(1-U)^3}$$



Case 4: Four edges with no attached networks

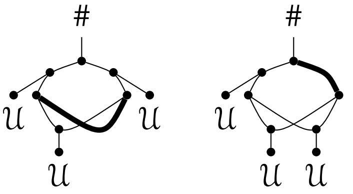
Case 4A - The four edges of the generator each carry exactly one outgoing arc. In this case, the two edges incident to v can be exchanged without modifying the network, and the same holds for the two edges not incident to v. This yields a factor $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ due to symmetries.

$$\frac{U^4}{4}$$



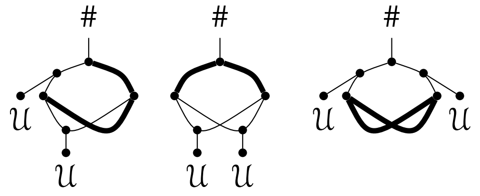
Case 4B - Three edges of the generator carry a single outgoing arc and the fourth one carries a sequence of at least two outgoing arcs. If this fourth edge is one incident to v, then the sequence it carries is oriented by the presence of the fictitious root, but the two arcs pending on the edges not incident to v are symmetric, hence a factor $\frac{1}{2}$. If on the contrary the edge carrying the sequence is not incident to v, then the sequence is also oriented, this time because of the arcs attached to the edges incident to v. Moreover, the picture has a symmetry w.r.t. the vertical axis, hence a factor $\frac{1}{2}$.

$$\frac{U^5}{2(1-U)} + \frac{U^5}{2(1-U)} = \frac{U^5}{1-U}$$



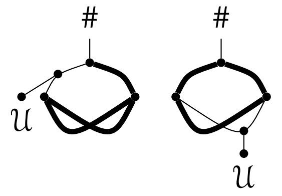
Case 4C - Two edges carry a single outgoing arc and the two others carry a sequence of at least two outgoings arcs. In all cases, the sequences are oriented, by the presence of either the fictitious root or of the single arcs attached to edges. If the edges carrying sequences are one incident to v and the other not incident to v, all edges are in addition distinguished from each other. In the other two cases, both edges incident to v form an unordered pair, as well as the two edges not incident to v. In each case, we therefore have a factor $\frac{1}{4}$.

$$\frac{U^6}{(1-U)^2} + \frac{U^6}{4(1-U)^2} + \frac{U^6}{4(1-U)^2} = \frac{3U^6}{2(1-U)^2}$$



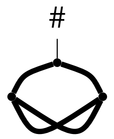
Case 4D - One edge of the generator carries a single outgoing arc and three edges carry a sequence of at least two outgoing arcs. As in the previous case, all sequences are oriented. However, if the two edges incident to v carry a sequence, the picture has a symmetry w.r.t. the vertical axis, hence a factor $\frac{1}{2}$ arises. If on the contrary the two edges not incident to v carry a sequence, these two edges are indistinguishable, hence we get a factor $\frac{1}{2}$ also in this case.

$$\frac{U^7}{2(1-U)^3}+\frac{U^7}{2(1-U)^3}=\frac{U^7}{(1-U)^3}$$



Case 4E - All four edges of the generator carry a sequence of at least two outgoing arcs. Then all sequences are oriented, but the two edges not incident to v are indistinguishable. The picture has in addition a symmetry w.r.t. the vertical axis. This yields a factor $\frac{1}{4}$.

$$\frac{U^8}{4(1-U)^4}$$



3.4.3 Generating function

The specification is directly translated into the following equation for the generating function U :

$$U=z+\frac{U^2}{2}+\frac{U^2}{2(1-U)}+\frac{U^2}{2(1-U)}+\frac{3}{2}U^2+\frac{5U^3}{2(1-U)}+\frac{5U^4}{4(1-U)^2}+U^3+\frac{3U^4}{1-U}$$
$$+\frac{3U^5}{(1-U)^2}+\frac{U^6}{(1-U)^3}+\frac{U^4}{4}+\frac{U^5}{1-U}+\frac{3U^6}{2(1-U)^2}+\frac{U^7}{(1-U)^3}+\frac{U^8}{4(1-U)^4}. \tag{3.1}$$

This equation for the generating function allows to derive the first coefficients of the series expansion of $U(z)$, namely

$$U(z)=z+3z^2+\frac{45}{2}z^3+\frac{421}{2}z^4+\frac{8809}{4}z^5+\cdots .$$

Expansion of a function near a singularity $\rho \sim 0.067$ is of the form

$$U(z) = U_0 + U_1 X + \cdots,$$

with $X = \sqrt{1 - \frac{z}{\rho}}$, $U_0 \sim 0.121$, and $U_1 \sim -0/0109$. The corresponding first values of u_n have been included in Table 5.2. (Recall indeed that $U(z) = \sum_{n \geq 0} u_n \frac{z^n}{n!}$ and that u_n is the number of unrooted level-2 networks on $(n+1)$ leaves).

The above equation for $U(z)$ can also be rewritten as follows:

Theorem 3.4.1. *The generating function $U(z)$ satisfies*

$$U(z) = z\phi(U(z)) \text{ where } \phi(z) = \frac{1}{1 - \frac{3z^5 - 16z^4 + 32z^3 - 30z^2 + 12z}{4(1-z)^4}}.$$

Proof. This is simply obtained from the above equation for U by algebraic manipulations. □

3.4.4 Exact enumeration formula

To obtain a closed form for u_n, we start from the equation for U given in Theorem 3.4.1. By the Lagrange inversion formula we obtain

$$u_n = n![z^n]U(z) = \frac{n!}{n}[z^{n-1}]\phi^n(z) = (n-1)![z^{n-1}]\phi^n(z),$$

so, to compute the first values of u_n, we can compute the Taylor expansions of $\phi^n(z)$.

As for the case of level-1 networks, we may also deduce with routine algebra an explicit formula for u_n.

Proposition 3.4.2. *For any $n \geq 1$, the number u_n of unrooted level-2 phylogenetic networks with $(n+1)$ leaves is given by*

$$u_n = (n-1)! \sum_{\substack{0 \leq s \leq q \leq p \leq k \leq i \leq n-1 \\ j = n-1-i-k-p-q-s \geq 0 \\ i \neq 0}} \begin{array}{c} \binom{n+i-1}{i}\binom{4i+j-1}{j}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s} \\ \times (3)^i \left(\frac{-15}{6}\right)^k \left(-\frac{17}{15}\right)^p \left(-\frac{1}{2}\right)^q \left(-\frac{3}{16}\right)^s. \end{array}$$

Proof sketch. Recall that $U(z) = z\phi(U(z))$ with $\phi(z) = \frac{1}{1 - \frac{3z^5 - 16z^4 + 32z^3 - 30z^2 + 12z}{4(1-z)^4}}$. Using first the classical expansion of $(1-z)^{-n}$ into a series, and then the binomial theorem, we obtain

$$\begin{aligned}\phi(z)^n =&\sum_{i\geq 0}\binom{n+i-1}{i}\left(\frac{12z}{4(1-z)^4}+\frac{-30z^2+32z^3-16z^4+3z^5}{4(1-z)^4}\right)^i\\ =&\sum_{i\geq 0}\sum_{k=0}^{i}\binom{n+i-1}{i}\binom{i}{k}\left(\frac{12z}{4(1-z)^4}\right)^{i-k}\left(\frac{-30z^2+32z^3-16z^4+3z^5}{4(1-z)^4}\right)^k.\end{aligned}$$

We continue applying the binomial theorem inside the above formula, isolating each time the term with the lowest degree in the numerator (that is, first $\frac{-30z^2}{4(1-z)^4}$, second $\frac{32z^3}{4(1-z)^4}$, …). This yields

$$\begin{aligned}\phi(z)^n =&\sum_{i\geq 0}\sum_{k=0}^{i}\sum_{p=0}^{k}\sum_{q=0}^{p}\sum_{s=0}^{q}\binom{n+i-1}{i}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s}\\ &\left(\frac{12z}{4(1-z)^4}\right)^{i-k}\left(\frac{-30z^2}{4(1-z)^4}\right)^{k-p}\left(\frac{32z^3}{4(1-z)^4}\right)^{p-q}\left(\frac{-16z^4}{4(1-z)^4}\right)^{q-s}\left(\frac{3z^5}{4(1-z)^4}\right)^{s}\\ =&\sum_{i\geq 0}\sum_{k=0}^{i}\sum_{p=0}^{k}\sum_{q=0}^{p}\sum_{s=0}^{q}\binom{n+i-1}{i}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s}\frac{(3)^i(\frac{-15}{6})^k(\frac{-17}{15})^p(\frac{-1}{2})^q(\frac{-3}{16})^s}{(1-z)^{4i}}z^{i+k+p+q+s}.\end{aligned}$$

The result then follows from expanding of $(1-z)^{-4i}$ into a series as $(1-z)^{-4i} = \sum_{j\geq 0}\binom{4i+j-1}{j}z^j$ and using the Lagrange inversion formula. □

3.4.5 Asymptotic evaluation

From Theorem 3.4.1, we can furthermore derive an asymptotic evaluation of the number u_n of unrooted level-2 networks on $(n+1)$ leaves, using Theorem 2.2.7.

Proposition 3.4.3. *The number u_n of unrooted level-2 networks on $(n+1)$ leaves is asymptotically equivalent to $c_1 \cdot c_2^n \cdot n^{n-1}$ for constants c_1 and c_2 such that $c_1 \approx 0.0669$ and $c_2 \approx 5.492$.*

Proof. Denoting by $\tau \approx 0.12116$ the unique solution of the characteristic equation $\phi(z) - z\phi'(z) = 0$ in the disk of convergence of ϕ, and $\rho = \frac{\tau}{\phi(\tau)} \approx 0.06697$, we have

$$[z^n]U(z) \sim \sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}}\frac{\rho^{-n}}{\sqrt{\pi n^3}}.$$

Using the Stirling estimate of the factorial, we get

$$u_n \sim \left(\frac{n}{e}\right)^n\sqrt{2\pi n}\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}}\frac{\rho^{-n}}{\sqrt{\pi n^3}} \sim \frac{n^{n-1}}{(e\rho)^n}\sqrt{\frac{\phi(\tau)}{\phi''(\tau)}}.$$

Replacing τ and ρ by their numerical approximations, we get the announced result. □

3.4.6 Refined enumeration formula and asymptotic distribution of parameters

Consider the refined generating function $U(z,x,y)$ for unrooted level-2 networks, where the variable z counts the size as before, the variable x counts the number of bridgeless components of level 1 or 2 (or equivalently, the number of level-1 or level-2 generators in the decomposition of these networks), and the variable y counts the number of inner edges, defined as the total number of edges across all level-1 and level-2 bridgeless components. The specification provided in the Equation 3.1 can be refined for these statistics, yielding the following equation for $U := U(z,x,y)$:

$$\begin{aligned} U = z &+ \frac{U^2}{2} + \frac{xy^3U^2}{2(1-yU)} + \frac{xy^6U^2}{2(1-yU)} + \frac{3}{2}xy^6U^2 + \frac{5xy^7U^3}{2(1-yU)} + \frac{5xy^8U^4}{4(1-yU)^2} \\ &+ xy^7U^3 + \frac{3xy^8U^4}{1-yU} + \frac{3xy^9U^5}{(1-yU)^2} + \frac{xy^{10}U^6}{(1-yU)^3} + \frac{xy^8U^4}{4} + \frac{xy^9U^5}{1-yU} \\ &+ \frac{3xy^{10}U^6}{2(1-yU)^2} + \frac{xy^{11}U^7}{(1-yU)^3} + \frac{xy^{12}U^8}{4(1-yU)^4}. \end{aligned} \tag{$\star_U$}$$

From the above equation, and similarly to Proposition 3.3.4, it would be possible (although computations and result are not reported in this part) to derive an explicit formula for the number of unrooted level-2 networks with n leaves, k bridgeless components of level 1 or 2, and m edges across them. Furthermore, some information on the asymptotic behavior of these parameters can be obtained from Equation ($\star_U$).

Proposition 3.4.4. *Let X_n (resp. Y_n) be the random variable counting the number of level-1 or level-2 bridgeless components (resp. the number of edges across them) in unrooted level-2 networks with $n+1$ leaves. Both X_n and Y_n are asymptotically normally distributed, and more precisely, we have*

$$\mathbb{E}X_n = \mu_X n + O(1), \qquad \mathbb{V}ar X_n = \sigma_X^2 n + O(1) \quad \text{and} \quad \frac{X_n - \mathbb{E}X_n}{\sqrt{\mathbb{V}ar X_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

$$\mathbb{E}Y_n = \mu_Y n + O(1), \qquad \mathbb{V}ar Y_n = \sigma_Y^2 n + O(1) \quad \text{and} \quad \frac{Y_n - \mathbb{E}Y_n}{\sqrt{\mathbb{V}ar Y_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mu_X \approx 0.7039$, $\sigma_X^2 \approx 0.1672$, $\mu_Y \approx 4.0295$ and $\sigma_Y^2 \approx 4.6961$.

Proof. Consider first X_n. Defining $U(z,x) := U(z,x,1)$, we obtain

$$\mathbb{E}x^{X_n} = \frac{[z^n]U(z,x)}{[z^n]U(z,1)}.$$

It follows from the equation for $U(z,x,y)$ that $U(z,x) = F(U(z,x),z,x) = z\frac{1}{1-A(U(z,x),z,x)}$ with

$$\begin{aligned}A(U,z,x) =&\frac{U}{2} + \frac{xU}{2(1-U)} + \frac{xU}{2(1-U)} + 2xU + \frac{5xU^2}{2(1-U)} + \frac{5xU^3}{4(1-U)^2} + xU^2\\ &+ \frac{3xU^3}{1-U} + \frac{3xU^4}{(1-U)^2} + \frac{xU^5}{(1-U)^3} + \frac{xU^3}{4} + \frac{xU^4}{1-U} + \frac{3xU^5}{2(1-U)^2}\\ &+ \frac{xU^6}{(1-U)^3} + \frac{xU^7}{4(1-U)^4}.\end{aligned}$$

It is readily checked that F satisfies all hypotheses of Theorem 2.2.10. The system

$$\begin{aligned}U =&F(U,z,1)\\ 1 =&F_U(U,z,1)\end{aligned}$$

admits a solution (U_0, z_0) such that $U_0 \approx 0.0897$ and $z_0 \approx 0.04801$, which satisfies the hypothesis of Theorem 2.2.10. The result then follows from Theorem 2.2.10, and the numerical estimates of μ_X and σ_X^2 are obtained plugging the numerical estimates for U_0 and z_0 into the explicit formulas given by Theorem 2.2.10. The proof for Y_n follows the exact same steps, considering this time $U(z,y) := U(z,1,y)$ instead, and adjusting the definition of F accordingly. Again, as expected, the solution (U_0, z_0) of the associated system is the same as above. □

3.5 Counting rooted level-2 networks

3.5.1 Combinatorial specification and generating function

To derive a specification for rooted level-2 networks, we distinguish cases depending on the level (0, 1 or 2) of the generator to which the root belongs. The cases corresponding to levels 0 and 1 will be the same as in Section 3.3. When the root of a rooted level-2 network belongs to a level-2 generator, we have to remember that these generators have one vertex which is above all their other vertices (which is the root of the network) and not just one but two reticulation vertices. As for rooted level-1 networks, it is important to keep in mind that any bridgeless component of level 2 in a rooted level-2 network has at least two outgoing cut-arcs (since otherwise there would be an infinite number of such networks with a given number of leaves).

We denote by $\mathcal{L}$ the set of rooted level-2 networks, where the size corresponds to the number of leaves. And we denote by $L(z)$ the corresponding exponential generating function. Distinguishing on the level (0, 1 or 2) of the bridgeless component containing the root, we can see that any network N of $\mathcal{L}$ satisfies exactly one of the following (see Figure 3.2).

- N is just a leaf. This contributes z to the generating function (case $0a$).
- The root of N belongs to a bridgeless component of level 0, that is to say it is a binary root vertex. Its children are themselves networks of $\mathcal{L}$ whose left-to-right order is irrelevant. This contributes $\frac{L^2}{2}$ to the generating function (case $0b$).
- The root of N belongs to a bridgeless component of level 1. This case splits into two subcases, just as in Section 3.3.
  - Either N consists of a cycle whose reticulation vertex is attached to a network of $\mathcal{L}$, is a child of the root and is the lowest vertex of a path from the root where a sequence of at least one network of $\mathcal{L}$ is attached. This contributes $\frac{L^2}{1-L}$ to the generating function (case $1a$).
  - Or N consists of a cycle whose reticulation vertex is attached to a network of $\mathcal{L}$, and a sequence of at least one network of $\mathcal{L}$ is attached to each path of this cycle (case $1b$). This contributes $\frac{L}{2}\left(\frac{L}{1-L}\right)^2$.
- The root of N belongs to a bridgeless component of level 2. The level-2 generators are displayed in Figure 3.2, cases $2a$ to $2d$. From these generators, the networks whose root belong to a bridgeless component of level 2 are obtained attaching networks of $\mathcal{L}$ to their reticulation vertex or vertices with out-degree 0, and possibly replacing their arcs with sequences of at least one network of $\mathcal{L}$. Note that in cases $2b$ and $2d$, depending on our choices for such arcs, we may have to cope with horizontal and vertical symmetry. We study these cases in order, and find their contribution to the generating function $L(z)$.

  - We first deal with the case where the level-2 generator to which the root belongs is of type $2a$. This generator has 5 internal arcs, all distinguished from each other by the structure of the generator. A network of $\mathcal{L}$ is attached to its reticulation vertex of outdegree 0. Moreover, recalling that each bridgeless component must have at least two outgoing cut-arcs, at least one of the five internal arcs of the generator must carry a non-empty sequence of networks of $\mathcal{L}$. Therefore, the contribution of case $2a$ to the generating function of $\mathcal{L}$ is

$$L \cdot \sum_{i=1}^{5} \binom{5}{i} \left(\frac{L}{1-L}\right)^i .$$

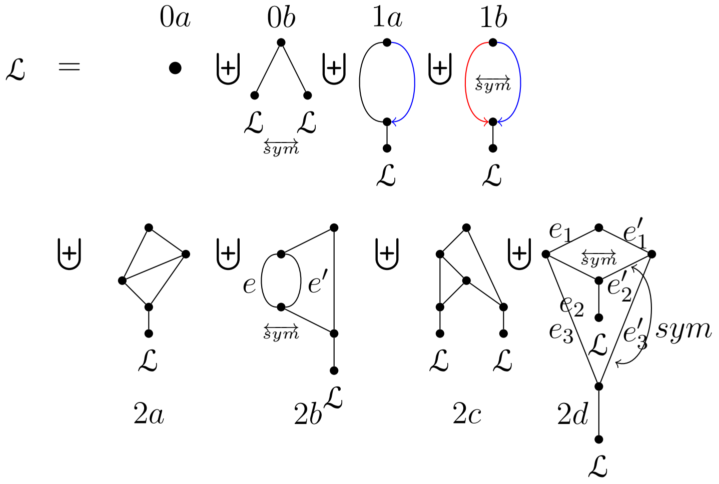


Figure 3.2: The specification of the class $\mathcal{L}$.

- In the case where the level-2 generator to which the root belongs is of type $2b$, we similarly have 5 internal arcs in the generator, at least one of which must be replaced by a non-empty sequence of networks of $\mathcal{L}$. However, the two arcs e and e' are not distinguishable. The contribution to the generating function is therefore more subtle to analyse, and we perform this detailed analysis in the following.

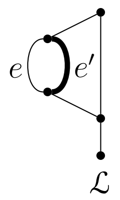
3.5.2 Case analysis for the rooted level-2 generator 2b

In the pictures below, we use thick lines to represent paths containing at least one internal node incident with a cut-arc which is incident with the root of another rooted level-2 network. All arcs are directed downwards. We use $\mathcal{L}$ to represented any rooted level-2 network.

Case 1:

Only one arc of the generator carries a sequence of at least one outgoing arc. This arc can only be e or e' (and these cases are indistinguishable), since otherwise the network would contain multiple edges, and this is not allowed.

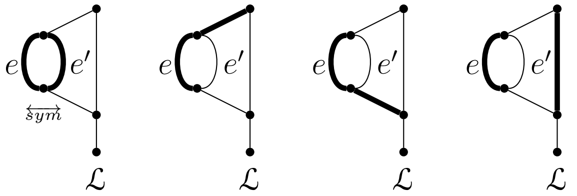
$$L\frac{L}{1-L}$$



Case 2:

Exactly two arcs of the generator carry a sequence of at least one outgoing arc. To avoid multiple edges, either these two arcs are e and e' (and those two arcs are symmetric, hence the factor $\frac{1}{2}$), or one of them is e or e' (which are not distinguished) and the other arc is chosen among the three arcs different from e and e'.

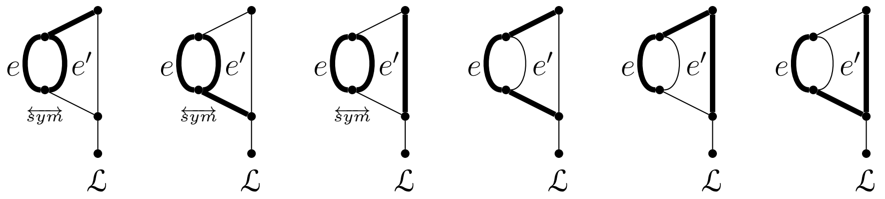
$$\frac{1}{2}L\left(\frac{L}{1-L}\right)^2 + 3L\left(\frac{L}{1-L}\right)^2 = \frac{7}{2}L\left(\frac{L}{1-L}\right)^2$$



Case 3:

Exactly three arcs of the generator carry a sequence of at least one outgoing arc. Here, there are two possibilities. Either both e and e' are among those three arcs (and those two arcs are symmetric, hence the factor $\frac{1}{2}$). Or, to avoid multiple edges, we must choose one of e and e' (which are not distinguished from each other), and two additional arcs among the three remaining arcs.

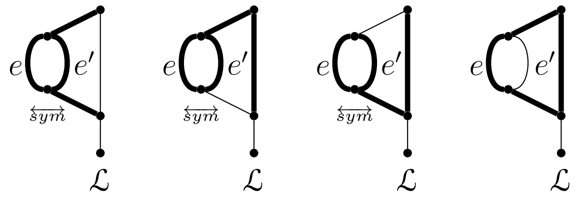
$$\frac{3}{2}L\left(\frac{L}{1-L}\right)^3 + 3L\left(\frac{L}{1-L}\right)^3 = \frac{9}{2}L\left(\frac{L}{1-L}\right)^3$$



Case 4:

Exactly four arcs of the generator carry a sequence of at least one outgoing arc. Either both e and e' are among those four arcs (and those two arcs are symmetric, hence the factor $\frac{1}{2}$), so the last two are chosen among the three other arcs of the generator. Or we choose the three arcs of the generator other than e and e', and e (which is undistinguishable from e').

$$\frac{\binom{3}{2}}{2}L\left(\frac{L}{1-L}\right)^4 + L\left(\frac{L}{1-L}\right)^4 = \frac{5}{2}L\left(\frac{L}{1-L}\right)^4$$



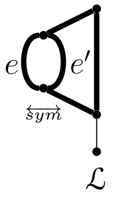
Case 5:

All five arcs of the generator carry a sequence of at least one outgoing arc. The fact that e and e' are symmetric explains the factor $\frac{1}{2}$.

$$\frac{1}{2}L\left(\frac{L}{1-L}\right)^5.$$

The overall contribution of case $2b$ to the generating function of $\mathcal{L}$ is then shown to be

$$L\frac{L}{1-L} + \frac{7}{2}L\left(\frac{L}{1-L}\right)^2 + \frac{9}{2}L\left(\frac{L}{1-L}\right)^3 + \frac{5}{2}L\left(\frac{L}{1-L}\right)^4 + \frac{1}{2}L\left(\frac{L}{1-L}\right)^5.$$



- We now consider the case where the level-2 generator to which the root belongs is of type $2c$. This generator has 6 internal arcs, all distinguished from each other by the structure of the generator. Moreover, two networks of $\mathcal{L}$ are attached to its reticulation vertices, so that the condition that each bridgeless component must be have at least two outgoing cut-arcs is already satisfied. Therefore, all 6 internal arcs of the generator carry possibly empty sequences of networks of $\mathcal{L}$. As a consequence, the contribution of case $2c$ to the generating function of $\mathcal{L}$ is
$$L^2\left(\frac{1}{1-L}\right)^6.$$
- Similarly, when the root belongs to a level-2 generator of type $2d$, the 6 arcs of the generator carry possibly empty sequences of networks of $\mathcal{L}$. However, this generator enjoys both a horizontal symmetry (mapping e_i to e'_i for $i = 1, 2, 3$) and a vertical symmetry (exchanging the indices 2 and 3 and the corresponding pending networks of $\mathcal{L}$). In case all arcs carry empty sequences, the horizontal symmetry is actually the identity, so that only the vertical symmetry applies, yielding a factor $\frac{1}{2}$. Otherwise, both the horizontal and the vertical symmetry need to be taken into account, yielding a factor $\frac{1}{4}$. The total contribution of case $2d$ to the generating function of $\mathcal{L}$ is therefore
$$\frac{1}{2}L^2 + \frac{1}{4}L^2 \cdot \sum_{i=1}^{6}\binom{6}{i}\left(\frac{L}{1-L}\right)^i.$$

Following this case analysis we obtain an equation characterizing the generating function of $\mathcal{L}$.

Theorem 3.5.1. *The exponential generating function $L(z)$ of rooted level-2 networks counted by number of leaves satisfies*

$$L = z + L^2 + \frac{7L^2}{1-L} + \frac{3L^3}{2(1-L)} + \frac{14L^3}{(1-L)^2} + \frac{15L^4}{4(1-L)^2} + \frac{29L^4}{2(1-L)^3} + \frac{5L^5}{(1-L)^3} + \frac{15L^5}{2(1-L)^4} + \frac{15L^6}{4(1-L)^4} + \frac{3L^6}{2(1-L)^5} + \frac{3L^7}{2(1-L)^5} + \frac{L^2}{(1-L)^6} + \frac{L^8}{4(1-L)^6},$$

or equivalently

$$L(z) = z\phi(L(z)) \quad \textit{where} \quad \phi(z) = \frac{1}{1 - \frac{36z-102z^2+159z^3-148z^4+81z^5-24z^6+3z^7}{4(1-z)^6}}.$$

We therefore obtain the first terms of the series expansion of $L(z)$,

$$L(z) = z + 9z^2 + \frac{381}{2}z^3 + \frac{20013}{4}z^4 + \frac{588119}{4}z^5 + \frac{129927717}{8}z^6 + \cdots,$$

as reported in Table 5.2.

3.5.3 Exact enumeration formula

As in the previous sections, Theorem 3.5.1 allows to derive an explicit formula for the number ℓ_n of rooted level-2 phylogenetic networks with n leaves.

Proposition 3.5.2. *For any $n \geq 1$, the number ℓ_n of rooted level-2 phylogenetic networks with n leaves is given by*

$$\ell_n = (n-1)! \sum_{\substack{0 \leq t \leq m \leq s \leq q \leq p \leq k \leq i \leq n-1 \\ j = n-1-i-k-p-q-s-m-t \geq 0 \\ i \neq 0}} \begin{array}{l} \binom{n+i-1}{i}\binom{6i+j-1}{j}\binom{i}{k}\binom{k}{p}\binom{p}{q}\binom{q}{s}\binom{s}{m}\binom{m}{t} \\ \times (9)^i \left(\frac{-17}{6}\right)^k \left(\frac{-53}{34}\right)^p \left(\frac{-148}{159}\right)^q \left(\frac{-81}{148}\right)^s \left(\frac{-8}{27}\right)^m \left(\frac{-1}{8}\right)^t. \end{array}$$

Proof sketch. This follows again from the Lagrange inversion formula, using the equation $L(z) = z\phi(L(z))$ for the function ϕ given in Theorem 3.5.1. The computations involve the usual development of $(1-z)^{-n}$ and the binomial formula, applied following exactly the same steps as in the proof of Proposition 3.4.2. Details of the computations are left to the reader. □

3.5.4 Asymptotic evaluation

Similarly, from Theorem 3.5.1, we can also derive the asymptotic behavior of ℓ_n.

Proposition 3.5.3. *The number ℓ_n of rooted level-2 phylogenetic networks with n leaves behaves asymptotically as*

$$\ell_n \sim c_1 c_2^n n^{n-1},$$

where $c_1 \approx 0.02931$ and $c_2 \approx 15.433$.

Proof. Recall that

$$L(z) = z\phi(L(z)) \quad \textit{where} \quad \phi(z) = \frac{1}{1 - \frac{36z-102z^2+159z^3-148z^4+81z^5-24z^6+3z^7}{4(1-z)^6}}.$$

Denoting by $\tau \approx 0.0444$ the unique solution of the characteristic equation $\phi(z) - z\phi'(z) = 0$ in the disk of convergence of ϕ, and $\rho = \frac{\tau}{\phi(\tau)} \approx 0.0238$, the Singular Inversion Theorem gives

$$[z^n]L(z) \sim \sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}} \frac{\rho^{-n}}{\sqrt{\pi n^3}}.$$

Like before, we get the claimed result from $\ell_n = n![z^n]L(z)$ and the Stirling estimate of the factorial. □

3.5.5 Refined enumeration formula and asymptotic distribution of parameters

Let $L(z,x,y) = \sum_{n,k,m} \ell_{n,k,m} \frac{z^n}{n!} x^k y^m$ be the multivariate generating function counting rooted level-2 networks w.r.t. their number of leaves (variable z), their number of bridgeless components of level 1 or 2 (variable x) and their number of arcs across all these (variable y). From the specification of $\mathcal{L}$ discussed earlier, $L(z,x,y) = L$ is easily seen to satisfy the following equation:

$$\begin{aligned} L = z + \frac{L^2}{2} + x\Big(&\frac{y^6L^2}{2} + (y^3+6y^6)\frac{L^2}{1-yL} + \frac{3y^7L^3}{2(1-yL)} + (\tfrac{y^4}{2} + \tfrac{27y^7}{2})\frac{L^3}{(1-yL)^2} \\ &+ \frac{15y^8L^4}{4(1-yL)^2} + \frac{29y^8L^4}{2(1-yL)^3} + \frac{5y^9L^5}{(1-yL)^3} + \frac{15y^9L^5}{2(1-yL)^4} + \frac{15y^{10}L^6}{4(1-yL)^4} \\ &+ \frac{3y^{10}L^6}{2(1-yL)^5} + \frac{3y^{11}L^7}{2(1-yL)^5} + \frac{y^6L^2}{(1-yL)^6} + \frac{y^{12}L^8}{4(1-yL)^6}\Big). \end{aligned} \tag{$\star_L$}$$

From the above equation, an explicit formula for $\ell_{n,k,m}$ could routinely be derived, as in Proposition 3.3.4, although the computations are more involved. We decided not to report this formula here. Equation ($\star_L$) also allows to study the asymptotic behavior of the considered parameters.

Proposition 3.5.4. *Let X_n (resp. Y_n) be the random variable counting the number of bridgeless components of level 1 or 2 (resp. the number of edges across them) in rooted level-2 networks with n leaves. Both X_n and Y_n are asymptotically normally distributed, and more precisely, we have*

$$\mathbb{E}X_n = \mu_X n + O(1), \qquad \mathbb{V}ar X_n = \sigma_X^2 n + O(1)$$

$$\mathbb{E}Y_n = \mu_Y n + O(1), \qquad \mathbb{V}ar Y_n = \sigma_Y^2 n + O(1)$$

where $\mu_X \approx 0.8242$, $\sigma_X^2 \approx 0.1232$, $\mu_Y \approx 4.8132$ and $\sigma_Y^2 \approx 3.5523$.

Proof. To prove the result for X_n (resp. Y_n), we specialize Equation ($\star_L$) for $y = 1$ (resp. $x = 1$) and rewrite it as $L(z, x, 1) = F(L(z, x, 1), z, x)$ for some explicit function F (resp. $L(z, 1, y) = F(L(z, 1, y), z, y)$, for an explicit different F). It is readily checked that F satisfies all hypotheses of Theorem 2.2.10, as well as the solutions (L_0, z_0) of the system

$$\begin{aligned} L =& F(L, z, 1) \\ 1 =& F_L(L, z, 1) \end{aligned}$$

whose approximate values are $L_0 \approx$ and $z_0 \approx$. The result then follows from Theorem 2.2.10, and the numerical estimates of μ_X and σ_X^2 (resp. μ_Y and σ_Y^2) are obtained plugging the numerical estimates for L_0 and z_0 into the explicit formulas given by Theorem 2.2.10. □

Chapter 4

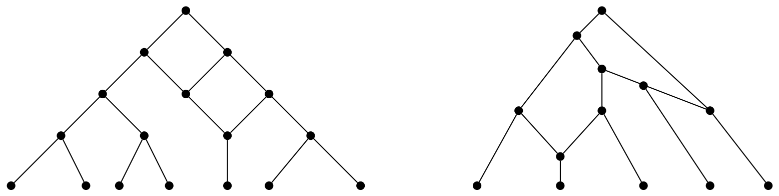
Tree-child and Normal Networks

4.1 Introduction and Results

Phylogenetic networks are used to model reticulate evolution. However, the process of evolution is driven by specific principles which add further restrictions on phylogenetic networks. Thus, biologists have defined many subclasses of the class of phylogenetic networks. Bounding the level of a network is one way to establish some control over the class of phylogenetic networks. Another way is to impose a structural condition on the network, for example, by considering how the vertices connect via paths to the leaves. This is the motivation behind a related class of networks that is particularly amenable to analysis. We start with a class that was defined fairly recently (2009), in [27], and which has turned out to be one of the most natural and important classes of networks. This chapter will be concerned with the counting of two subclasses of the class of phylogenetic networks that are called *tree-child networks* and *normal networks*; see e.g. [65].

In tree-child networks, one has the additional requirement that reticulation events cannot happen in close proximity, or more formally, every tree vertex must have at least one child which is not a reticulation vertex and no reticulation vertex is directly followed by another reticulation vertex. Normal networks, on the other hand, form a subclass of the class of tree-child networks with the additional requirement that evolution does not take shortcuts, or again more formally, if there is a path of at least length 2 from a vertex u to a vertex v, then there is no direct edge from u to v. For examples of such networks see Figure 4.1.

Remark. Note that in general phylogenetic networks, multiple edges are not explicitly forbidden (except when dealing with enumeration of leaf-labeled networks, since otherwise the counting problem is not meaningful). In fact, only double edges may occur because of the degree constraints. The tree-child condition, however, makes double edges impossible. Thus tree-child and normal networks do not contain double edges.



(i) (ii)

Figure 4.1: Two phylogenetic networks, where (i) is a general network that is not a tree-child network and (ii) is a tree-child network that is not a normal network. Edges are directed downwards.

Next, let us recall what is known about the number of tree-child and normal networks. Denote by T_n and N_n the number of vertex-labeled tree-child networks and vertex-labeled normal networks, respectively, where n is the total number of vertices. Similarly, denote by $\tilde{T}_\ell$ and $\tilde{N}_\ell$ the number of leaf-labeled tree-child networks and leaf-labeled normal networks, where ℓ denotes the number of leaves. Then, it was proved in [7] that for all odd n,

$$(e_1 n)^{5n/4} \leq N_n \leq T_n \leq (e_2 n)^{5n/4},$$

where $e_1, e_2 > 0$ are suitable constants. (It is easy to see that $N_n = T_n = 0$ when n is even.) Similarly, there are $f_1, f_2 > 0$ such that for all ℓ,

$$(f_1 \ell)^{2\ell} \leq \tilde{N}_\ell \leq \tilde{T}_\ell \leq (f_2 \ell)^{2\ell}.$$

Note that the first result can be equivalently stated as

$$N_n = n^{5n/4 + \mathcal{O}(n/\log n)} \qquad \text{and} \qquad T_n = n^{5n/4 + \mathcal{O}(n/\log n)}$$

and the second as

$$\tilde{N}_\ell = \ell^{2\ell + \mathcal{O}(\ell/\log \ell)} \qquad \text{and} \qquad \tilde{T}_\ell = \ell^{2\ell + \mathcal{O}(\ell/\log \ell)}.$$

Thus, one is still quite far away from getting precise asymptotics for these counting sequences and this was left as an open problem in [7].

In this section, we will consider tree-child and normal networks with a fixed number k of reticulation vertices. It should be mentioned that they form (very) small subclasses of the class of all tree-child and normal networks since it was also proved in [7] that almost all vertex-labeled tree-child resp. normal networks have $k \sim n/4$ and almost all leaf-labeled tree-child resp. normal networks have

$k \sim \ell$. Nevertheless, these subclasses are interesting from a combinatorial point of view since we can get precise asymptotics of their numbers. Moreover, they are more suitable for modelling phylogenesis in environments where reticulation is a very rare event (although even then it may be sometimes desirable to admit $k \to \infty$ as $n \to \infty$). Models with a fixed number of reticulation vertices were for instance considered in [42, 56]. Likewise, in the construction of phylogenetic networks from trees, models with bounded reticulation number do play a role, see [39, 40].

Recently, people studying phylogenetic networks or related structures have become more and more interested in enumerative aspects. We mentioned already the shape analysis of phylogenetic trees [2, 3, 23, 24] and the bounds for the counting sequences of some classes of phylogenetic networks [7]. But other counting problems were studied in [1, 16, 13, 50, 49, 53, 56]. Though combinatorial counting problems are often amenable to the rich tool box of analytic combinatorics [20], generating functions have been rarely used in phylogenetic networks enumeration problems.

Here we focus on the already mentioned class of phylogenetic networks with a low number of reticulation events, more specifically on the above two subclasses of this class, and demonstrate how analytic combinatorics can be used to obtain general (asymptotic) enumeration results for those classes. We believe that our result is of interest to experts working on the mathematics of phylogenetics and that many more enumeration problems in phylogenetics can be approached in a similar way.

Now, denote by $N_{k,n}$ resp. $T_{k,n}$ the number of normal resp. tree-child networks with k reticulation vertices in the vertex-labeled case and $\tilde{N}_{k,\ell}$ resp. $\tilde{T}_{k,\ell}$ in the leaf-labeled case. Then, our results are as follows.

Theorem 4.1.1. *For the number $N_{k,n}$ of vertex-labeled normal networks with $k \geq 1$ reticulation vertices, there is a positive constant c_k such that*

$$N_{k,n} \sim c_k \left(1-(-1)^n\right) \left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1}, \qquad (n \to \infty).$$

In particular,

$$c_1 = \frac{\sqrt{2}}{4}; \qquad c_2 = \frac{\sqrt{2}}{32}; \qquad c_3 = \frac{\sqrt{2}}{384}.$$

Remark. Note that this result also holds for $k = 0$ where it becomes the result of Schröder; see above and [51].

Surprisingly, the same result also holds for vertex-labeled tree-child networks. (It was proved in [7] that $N_n = o(T_n)$.) This shows in particular that if one considers only first-order asymptotics, then the additional requirement for normal networks does not matter. Note, however, that we are considering networks with an *a priori* fixed number k of reticulation vertices. Thus, we do not claim that the asymptotic equivalence given in Theorem 4.1.1 holds uniformly in k (and neither do we claim this in Theorem 4.1.2 below). Indeed, such a claim would surely be wrong since otherwise one could sum up both sides over k and would get a contradiction to the above mentioned result from [7].

Theorem 4.1.2. *For the number $T_{k,n}$ of vertex-labeled tree-child networks with $k \geq 1$ reticulation vertices,*

$$T_{k,n} \sim c_k \left(1 - (-1)^n\right) \left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1}, \qquad (n \to \infty)$$

with c_k as in the previous theorem.

Remark. Again the result also holds for $k = 0$ where it is Schröder's result. Moreover, note that for $k = 0$ and $k = 1$, $T_{k,n}$ is identical with the number of all vertex-labeled phylogenetic networks. We can show for $k \geq 2$, the latter number becomes strictly larger than $T_{k,n}$, however, the leading term of the asymptotic expansion is likely to be again the same.

Corollary 4.1.3. *Let $k \geq 1$. Then, asymptotically almost all tree-child networks with k reticulation vertices are normal networks.*

Here we present the numerical values for the number of normal and tree-child networks with three reticulation vertices. They are compared to the first and second order asymptotics. The error is $\mathcal{O}\left(1/\sqrt{n}\right)$ in the first order asymptotics and $\mathcal{O}\left(1/n\right)$ in the second order asymptotics. Thus convergence is slow. So, we chose a quadratic scale to better visualize the convergence.

The data indicate that the more reticulation vertices the networks have, the bigger is the constant factor in the third order term. In particular, for normal networks even the second order asymptotics is still fairly inaccurate when the size is around 1000 vertices.

n	$N_{3,n}$	first order asymptotics	second order asymptotics
21^2	$0.742450513 \times 10^{1052}$	$1.380613859 \times 10^{1052}$	$0.639038343 \times 10^{1052}$
23^2	$6.765254066 \times 10^{1301}$	$11.83574504 \times 10^{1301}$	$6.031172891 \times 10^{1301}$
25^2	$4.878745045 \times 10^{1581}$	$8.117965422 \times 10^{1581}$	$4.455195521 \times 10^{1581}$
27^2	$1.074095703 \times 10^{1892}$	$1.713861489 \times 10^{1892}$	$0.997859210 \times 10^{1892}$
29^2	$2.503773287 \times 10^{2233}$	$3.855495246 \times 10^{2233}$	$2.355863510 \times 10^{2233}$
31^2	$1.957523560 \times 10^{2606}$	$2.923628151 \times 10^{2606}$	$1.859821038 \times 10^{2606}$

n	$T_{3,n}$	first order asymptotics	second order asymptotics
21^2	$1.076588119 \times 10^{1052}$	$1.380613859 \times 10^{1052}$	$1.133422020 \times 10^{1052}$
23^2	$9.485462012 \times 10^{1301}$	$11.83574504 \times 10^{1301}$	$9.900887645 \times 10^{1301}$
25^2	$6.651391940 \times 10^{1581}$	$8.117965422 \times 10^{1581}$	$6.897042119 \times 10^{1581}$
27^2	$1.430044284 \times 10^{1892}$	$1.713861489 \times 10^{1892}$	$1.475194062 \times 10^{1892}$
29^2	$3.266427497 \times 10^{2233}$	$3.855495246 \times 10^{2233}$	$3.355617999 \times 10^{2233}$
31^2	$2.509177651 \times 10^{2606}$	$2.923628151 \times 10^{2606}$	$2.569025778 \times 10^{2606}$

Remark. When going beyond first-order asymptotics, one sees that the additional requirement for normal networks does indeed matter; see below for longer asymptotic expansions for $k = 1, 2, 3$ which show a difference in the second order term for vertex-labeled normal and tree-child networks.

Similar results to the results above will be shown for leaf-labeled tree-child and normal networks, too; see Section 4.5.

The remainder of the chapter is as follows. In the next section, we will explain how to use generating functions to count tree-child and normal networks. In other words, we will use methodology of Analytic Combinatorics [20] which is relies on the symbolic method [20, Sec. I.1–I.2] and the treatment of labeled structures [20, Sec. II.1–II.2] as well as the pointing operation [20, Sec. II.6]. This counting procedure will then be applied in Section 4.3 to vertex-labeled normal networks. (This section will contain the proof of Theorem 4.1.1.) In Section 4.4, we apply the same approach to vertex-labeled tree-child networks and prove Theorem 4.1.2. In Section 4.5, we will briefly discuss results for leaf-labeled networks which are obtained from those for vertex-labeled networks in Section 4.3 and Section 4.4.

4.2 Decomposing Phylogenetic Networks

In order to count the above classes of phylogenetic networks, we will decompose them and use this decomposition to obtain a reduction which can be easily analyzed by means of generating functions. Then the reduction is extended to get back the original network in such a way that the extension procedure has a counterpart in generating function algebra, hence allowing an asymptotic analysis of the number of phylogenetic networks. We start with normal networks, since tree-child networks differ from normal ones just by dropping a condition which allows a similar analysis.

Consider a normal network having exactly k reticulation vertices. Then each such vertex has two incoming edges. Choose one of them and remove it. The remaining graph is a (labeled and nonplane) Motzkin tree[1], *i.e.*, a tree consisting

[1] We mention that we slightly abuse the word here: A Motzkin tree (also known as unary-binary

of leaves (zero children), unary vertices (one child) and binary vertices (two children). All edges in this Motzkin tree are directed away from the root. In particular, it is a Motzkin tree with exactly $2k$ unary vertices, where k of them are the starting points of the removed edges, the other k their end points (note that here the tree-child property was used).

Now consider the following procedure (see Figure 4.2 for an illustration): Start with a Motzkin tree M with exactly $2k$ unary vertices and n vertices in total. Then add edges such that *(i)* each edge connects two unary vertices, *(ii)* no two of the added edges have a vertex in common, and *(iii)* the resulting graph is a normal network N. Finally, color the start vertices of the added edges green and their end vertices red. We say then that M (keeping the colors from the above generation of N, but not the edges) is a *colored Motzkin skeleton* (or simply *Motzkin skeleton*) of N. In this way all normal networks with n vertices are generated and each of them exactly 2^k times, since every network N with k reticulation vertex has exactly 2^k different Motzkin skeletons.

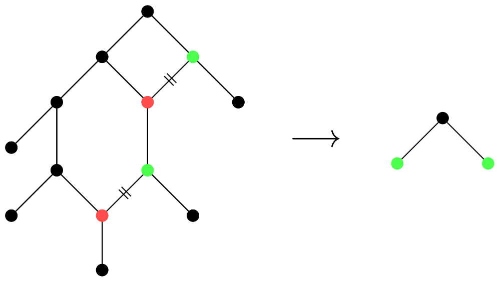


Figure 4.2: A normal network with colored Motzkin skeleton and coresponding sparsened skeleton. Note that there are three more possible colored Motzkin skeletons which one can obtain from the same network and that all but one yield the same sparsened skeleton.

In order to set up generating functions for phylogenetic networks, we will construct them as follows: for a given network N first pick one of its 2^k possible Motzkin skeletons. Then, look for the minimal subtree T which contains all green vertices. This tree contains all the green vertices as well as all last common ancestors[2] of any two green vertices. These particular vertices form a tree whose edges

tree) is usually an unlabeled and plane. The concept stems from computer science, see [17, 19, 21]. In contrast, the trees we are considering here are labeled and nonplane, but nevertheless still unary-binary trees. Thus, they are the labeled and nonplane counterpart of classical Motzkin trees. For a comprehensive introduction into recursive structures like Motzkin trees and also labeled and nonplane combinatorial structures see [20].

[2] Note that we use the name which is common in the combinatorial literature. In the phylogenetics literature this is usually called *most recent common ancestor*.

are paths in T. Contract each of these paths to one single edge. The resulting tree, which is again a Motzkin tree, is called the *sparsened skeleton* of N. The structure of this tree tells us how the green vertices are distributed within N (again See Figure 4.2).

In order to construct networks with k reticulation vertices, we start with a sparsened skeleton having k green vertices. Then we replace all edges by paths that are made of red vertices or binary vertices with a Motzkin tree (whose unary vertices are all colored red) as second child and add a path of the same type on top of the root of the sparsened skeleton. Moreover, we attach a Motzkin tree (again with all unary vertices colored red) to each leaf of the sparsened skeleton such that this new subtree is linked to the (now former) leaf by an edge (for normal networks, this tree can be a binary tree). Do all of the above in such a way that the new structure has k red vertices altogether. What we obtain so far is a Motzkin skeleton of a phylogenetic network. Finally, add edges connecting the green vertices to the red ones in such a way that the corresponding mapping is bijective and that the normality condition for phylogenetic networks is respected.

Let us set up the exponential generating function for Motzkin trees which appear in the above construction. This means that the tree-child condition for networks has to be respected, but the number of unary vertices need not be even. After all, the unary vertices in those trees will be the red vertices of our network.

Denote by $M_{\ell,n}$ the number of all vertex-labeled Motzkin trees with n vertices and ℓ unary vertices (all colored red) that respect the tree-child condition for networks, which means that the child of a unary vertex cannot be a unary vertex and each binary vertex has at least one child which is not a unary vertex. Let $\mathcal{M}$ denote the set of all these Motzkin trees. The exponential generating function associated to $\mathcal{M}$ is

$$M(z,y) = \sum_{n\geq 1}\sum_{\ell\geq 0} M_{\ell,n} y^{\ell}\frac{z^n}{n!}. \tag{4.1}$$

Furthermore, let $M_u(z,y)$ and $M_b(z,y)$ denote the generating function associated to all Motzkin trees in $\mathcal{M}$ whose root is a unary vertex and a binary vertex, respectively. Then

$$M_u(z,y) = zy(z + M_b(z,y))$$

since a unary vertex cannot have a unary child. In a Motzkin tree with a binary root, the root may have two children being either a leaf or a binary vertex, or one of the children is a unary vertex and the other either a leaf or a binary vertex. This yields

$$M_b(z,y) = \frac{z}{2}((z + M_b(z,y))^2 + 2zy(z + M_b(z,y))^2).$$

Solving gives

$$M_b(z,y) = \frac{1-\sqrt{1-2z^2-4yz^3}}{z(1+2yz)} - z$$

and

$$M_u(z,y) = y\frac{1-\sqrt{1-2z^2-4yz^3}}{1+2yz} \tag{4.2}$$

and thus

$$M(z,y) = z + M_u(z,y) + M_b(z,y) = \frac{(1+yz)\left(1-\sqrt{1-2z^2-4yz^3}\right)}{z(1+2yz)}. \tag{4.3}$$

The first few coefficients can be seen from

$$z+yz^2+\frac{1}{2}z^3+\frac{3}{2}yz^4+\left(y^2+\frac{1}{2}\right)z^5+\frac{5}{2}yz^6+\left(4y^2+\frac{5}{8}\right)z^7+\left(2y^3+\frac{35}{8}y\right)z^8+\cdots.$$

4.3 Counting Vertex-Labeled Normal Networks

In this section, we will count (vertex-labeled) normal networks with a fixed number k of reticulation vertices. We will start with the cases $k = 1, 2, 3$ which will be discussed in the next three subsections and for which we will derive asymptotic expansions up to the second order term (in fact, our method allows one to obtain full asymptotic expansions as well). From these three cases, we will observe a general pattern which will be proved in the fourth subsection.

4.3.1 Normal networks with one reticulation vertex

In this subsection we will determine the asymptotic number of normal networks with one reticulation vertex and then discuss their relationship to unicyclic networks that were studied in [56].

Counting

In order to count normal networks with only one reticulation vertex we use Motzkin trees from the class $\mathcal{M}$, which have generating function (4.3), and (sparsened) skeletons, as described in the previous section: We delete one of the two incoming edges of the reticulation vertex which then gives a unary-binary tree satisfying the tree-child property with exactly two unary vertices. Conversely, we can start with the general tree or even the sparsened skeleton (which only consists of one vertex) and then construct the network from this.

Proposition 4.3.1. *The exponential generating function for vertex-labeled normal networks with one reticulation vertex is*

$$N_1(z) = \frac{z\left(1-\sqrt{1-2z^2}\right)^3}{2(1-2z^2)^{3/2}} = z\frac{a_1(z^2)-b_1(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{3/2}}, \tag{4.4}$$

where

$$a_1(z) = 2-3z \qquad \textit{and} \qquad b_1(z) = 2-z.$$

Proof. As already mentioned, we start with the general tree as depicted in Figure 4.3, which arises from the sparsened skeleton, *i.e.*, the tree consisting of a single green vertex g as follows: we add a sequence of trees on top of g which consist of a root to which a tree in $\mathcal{M}$ is attached. Moreover, we attach also a tree from $\mathcal{M}$ to g as a subtree.

Next, in order to obtain all normal networks arising from these Motzkin skeletons, we have to add an edge starting from g and ending at the red vertex. Note that for a normal network, this edge is neither allowed to point to a vertex on the path from g to the root (since the network must be a DAG), nor to the root of one of the trees which are connected to the vertices on the path from g to the root (since this violates the normality condition) nor to any vertex in the subtree of g (since this again violates the normality condition). Overall, the red vertex must be contained in the forest attached to the path from g to the root, but not in the tree attached to g. Moreover, note that since there is only one red vertex, the requirement that trees in this forest satisfy the tree-child property could actually be dropped.

The networks arising from these skeletons can therefore be specified as a tree without red vertices (the one attached to g) and a sequence of structures of the form "vertex plus Motzkin tree with non-unary root" (*cf.* Figure 4.3). In terms of generating functions this gives

$$N_1(z) = \frac{1}{2}\frac{\partial}{\partial y}\frac{zM(z,0)}{1-z\tilde{M}(z,y)}\Big|_{y=0}$$

where

$$\tilde{M}(z,y) = z + M_b(z,y) = M(z,y) - zy(z+M_b(z,y)). \tag{4.5}$$

The factor $1/2$ makes up for the fact that each network is counted exactly twice by the above procedure. Evaluating this and writing M_y for the partial derivative of M (w.r.t. y) yields

$$\begin{aligned} N_1(z) &= \frac{z}{2}M(z,0)(M_y(z,0)-z^2-zM_b(z,0))\sum_{\ell\geq 1}\ell z^\ell M(z,0)^{\ell-1} \\ &= \frac{z^2M(z,0)(M_y(z,0)-z^2-zM_b(z,0))}{2(1-zM(z,0))^2}. \end{aligned}$$

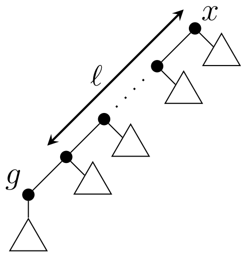


Figure 4.3: The structure of Motzkin skeletons of networks with one reticulation vertex. It originates from a sparsened skeleton which consists of only one green vertex. It has one green vertex, denoted by g, and one red vertex which is hidden within the forest made of the triangles in the picture, which are attached to g and all the vertices on the path of length ℓ. Note that the position of the red vertex in this forest is restricted by the normality condition.

Now, by using

$$M(z,0) = \frac{1-\sqrt{1-2z^2}}{z}, \quad M_y(z,0) = \frac{1}{\sqrt{1-2z^2}} - 1, \quad M_b(z,0) = \frac{1-\sqrt{1-2z^2}}{z} - z, \tag{4.6}$$

we obtain (4.4). □

From this result we can now easily obtain the asymptotic number of normal networks.

Corollary 4.3.2. *Let $N_{1,n}$ denote the number of vertex-labeled normal networks with n vertices and one reticulation vertex. If n is even then $N_{1,n}$ is zero, otherwise*

$$N_{1,n} = n![z^n]N_1(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+1}\left(\frac{\sqrt{2}}{2} - \frac{3\sqrt{\pi}}{2}\cdot\frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Proof. The function (4.4) has two dominant singularities, namely at $\pm 1/\sqrt{2}$, with singular expansions

$$N_1(z) \overset{z\to\pm 1/\sqrt{2}}{\sim} \pm\frac{1}{8(1\mp\sqrt{2}z)^{3/2}} \mp \frac{3\sqrt{2}}{8(1\mp\sqrt{2}z)} + \mathcal{O}\left(\frac{1}{\sqrt{1\mp\sqrt{2}z}}\right).$$

Applying a transfer lemma (see [18, 20]) for these two singularities and using Stirling's formula completes the proof. □

Remark. Note that the periodicity is not surprising since, as mentioned in the introduction, phylogenetic networks always have an odd number of vertices.

Relationship to unicyclic networks

In [56], the authors counted unicyclic networks which are (vertex-labeled or leaf-labeled) pointed[3] graphs with only one cycle to which complete binary trees are attached. The enumeration was done only for leaf-labeled networks there.

On the other hand, phylogenetic networks with exactly one reticulation vertex are the same as unicyclic networks, if one disregards the direction of the edges.[4]

So, another way of counting normal networks with exactly one reticulation vertex is by using a modification of the approach of [56]: either the root is in a cycle in which case one of the vertices on this cycle except the root and its two neighbours are the reticulation vertex and to each vertex may be attached a complete binary rooted tree or the root is not in the cycle in which case exactly one subtree contains the cycle. This translates into

$$N_1(z) = zM(z,0)N_1(z) + \frac{1}{2}\sum_{\ell\geq 3}(\ell-2)z^{\ell+1}M(z,0)^{\ell}.$$

Solving this equation gives

$$N_1(z) = \frac{\sum_{\ell\geq 3}(\ell-2)z^{\ell+1}M(z,0)^{\ell}}{2(1-zM(z,0))} = \frac{z^4M(z,0)^3}{2(1-zM(z,0))^3}.$$

Plugging (4.6) into this reveals

$$N_1(z) = \frac{z(1-\sqrt{1-2z^2})^3}{2(1-2z^2)^{3/2}}$$

as it must be.

4.3.2 Normal networks with two reticulation vertices

For this case, we use two variables y_1, y_2 to express the possible pointings of the two green vertices of the Motzkin skeletons. Furthermore, we have now more complicated paths (and attached trees) which replace the edges of the sparsened skeleton and thus we first set up the generating function corresponding to theses paths. To govern the situation where an edge from one of the two green vertices

[3] Pointed means that an edge is chosen to which a vertex v is attached (with an edge, of course). The chosen edge itself is thus split into two edges and the point where the new edge is attached becomes a further new vertex. The vertex v is then the root vertex of the network

[4] Combinatorially, there is no big difference between rooted and pointed, since we can always drop the attached vertex and edge in the latter case and direct all edges. Thus, if one can solve the counting problem for a subclass of rooted networks also the corresponding counting problem for pointed graphs can be solved.

must not point into a certain subtree or to a particular vertex, we distinguish several types of unary vertices, which are the red vertices of our construction.

To simplify the explanation, let us use the following conventions: If the root of a Motzkin tree is a unary vertex (so, a red vertex) we call the tree a *red tree*, otherwise a *white tree*. Note that the class of white trees has generating function $\tilde{M}(z, y)$ given in (5.4), whereas the class of red trees has generating function $M_u(z, y)$, see (4.2).

The structure we will need is a class $\mathcal{P}$ of paths which serve as the essential building blocks for Motzkin skeletons. In this class the rules for pointing to particular red vertices differ, depending on whether (i) the red vertex lies on the path itself, but is not the very first vertex there,(ii) it is the root of one of the (red) subtrees attached to the vertices of the path, (iii) it is one of the non-root vertices of one of the attached subtrees or (iv) the red vertex is the first vertex of the path. To distinguish these three classes of red vertices, we will mark the red vertices of type (i) with the variable y, those of type (ii) with $\bar{y}$ and the vertex of type (iii) with $\tilde{y}$ and finally, the red vertices of type (iv) with the variable $\hat{y}$.

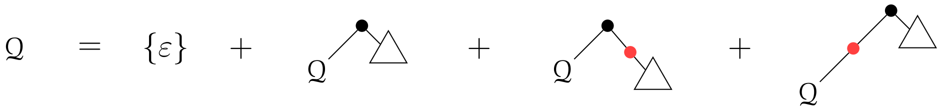


Figure 4.4: The specification of the class $\mathcal{Q}$. In this picture, the paths are drawn such that they are going from upper right to lower left. The triangles represent the trees attached to the path. These are white trees, *i.e.*, trees which do not have a unary root. The variables $\bar{y}$ and y mark the red vertices that are shown in the middle and right one of figure respectively. Others may be hidden in the white trees and are marked by $\tilde{y}$. The last part of the specification guarantees that there are no consecutive red vertices on the path.

Moreover, we have to respect the tree-child condition. Normality does not play a role on this level, it actually only causes the need for the second class of red vertices. The tree-child condition implies that the successor of a red vertex on the path itself must not be red. Moreover, if the tree attached to some vertex x is a red tree, then the successor of x on the path must not be a red vertex. This gives rise to the a combinatorial specification. Take a set of three possible atomic items: a vertex with a white tree, a vertex with a red tree (which is itself a red vertex having a white tree as subtree), and a vertex having a red vertex and a white tree as subtrees. Then a path in $\mathcal{P}$ is either (a) a sequence made of these atomic items or (b) a red vertex followed by a sequence of type (a). More formally, let $\tilde{\mathcal{M}}$ denote the class of white trees, $\circ$ denote a binary vertex and $\bullet$ denote a red (unary) vertex.

We write $\{x\} \times S \times T$ if x is a vertex having subtrees S and T, where T is omitted if x is a red vertex and the edge x — S is an edge of the path. Then we consider a class $\mathcal{Q}$ which contains all path in $\mathcal{P}$ of type (a) above. The specification of this class is

$$\mathcal{Q} = \{\varepsilon\} \cup \{\circ\} \times \mathcal{Q} \times \tilde{\mathcal{M}} \cup \{\circ\} \times \mathcal{Q} \times (\{\bullet\} \times \tilde{\mathcal{M}}) \cup \{\circ\} \times (\{\bullet\} \times \mathcal{Q}) \times \tilde{\mathcal{M}}, \tag{4.7}$$

where ε denotes the empty tree; see Figure 4.4. Since a path in $\mathcal{P}$ may also start with a red vertex, which then belongs to the third class of red vertices, we specify $\mathcal{P}$ as

$$\mathcal{P} = \mathcal{Q} \cup \{\bullet\} \times \mathcal{Q}. \tag{4.8}$$

This leads to the generating function

$$P(z, y, \tilde{y}, \bar{y}, \hat{y}) = \frac{1 + z\hat{y}}{1 - (z + z^2 y + z^2 \bar{y})\tilde{M}(z, \tilde{y})} \tag{4.9}$$

after all.

Let us summarize what we just defined. In our analysis the variables y, $\tilde{y}$, $\hat{y}$ and $\bar{y}$ will be replaced by a sum of variables y_i where the presence of a particular y_i indicates that the corresponding g_i is allowed to point, its absence that pointing is forbidden. In particular, y represents the permission to point to vertices of the path (except its first vertex) and $\bar{y}$ is corresponding permission to point to the roots of the trees attached to the path. The variable $\tilde{y}$ describes the permission to point to non-root vertices of these trees and $\hat{y}$ allows pointing to the first vertex of the path.

Now we are ready to state the following result.

Proposition 4.3.3. *The exponential generating function for vertex-labeled normal networks with two reticulation vertices is*

$$N_2(z) = z\frac{a_2(z^2) - b_2(z^2)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{7/2}}, \tag{4.10}$$

where

$$a_2(z) = 11z^4 - 66z^3 + 50z^2 - 8z \quad \textit{and} \quad b_2(z) = -28z^3 + 42z^2 - 8z.$$

Proof. Note that, in the current situation, there are only two possible sparsened skeletons: either a path of length one (with both vertices green) or a cherry (with both leaves being green vertices). From this, one builds two possible types of Motzkin skeletons that are depicted in Figure 4.5.

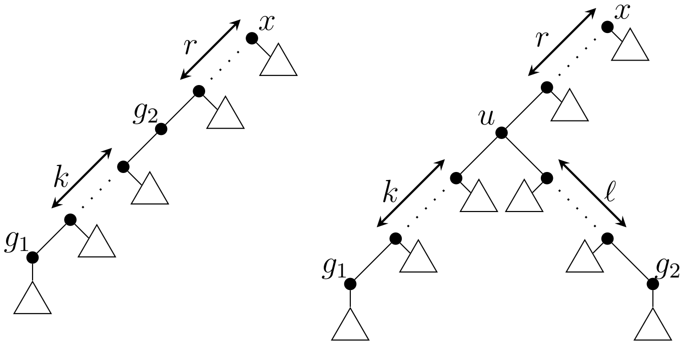


Figure 4.5: The possible structures of Motzkin skeletons of networks with two reticulation vertices. These originate from the two possible sparsened skeletons made of two green vertices: The path of length one, which gives rise to the left Motzkin skeleton, and the cherry leading to the right Motzkin skeleton.
Note: In this figure (as well as in all the subsequent figures of this part) the triangles are placeholders for trees which may but need not necessarily be white trees (see beginning of Section 4.3.2). The class they belong to depends on their position with in the normal network.

For the first type (see Figure 4.5, left), we have to complete the Motzkin skeletons by adding two egdes having start vertex g_1 and g_2, respectively. The one starting from g_1 may point to any non-root vertex within the subtrees that are attached to the skeleton's spine (*i.e.*, the paths k and ℓ and g_2). By normality, it can neither point to the root of one of those subtrees nor to a vertex in the subtree attached to g_1 itself, but g_1 does not belong to what we called the spine anyway. Similarly, the edge starting at g_2 may point to any non-root vertex in the subtrees attached to the path ℓ, the path from the root to the parent of g_2. Thus the generating function of the subtrees attached to the vertices of ℓ is $\tilde{M}(z, y_1 + y_2)$, that of the subtrees attached to the vertices of k is $\tilde{M}(z, y_1)$. The tree attached to g_1 corresponds to $M(z, 0)$ since it must not contain any red vertices. Finally, note that we have to point at two red vertices, one targeted by g_1 and one targeted by g_2. Pointing (and not counting it any more as red vertex) corresponds to differentiation in the world of generating functions. Since we do not want any other red vertices to be present, we set $y_1 = y_2 = 0$ after the differentiations. After all, we get the generating function

$$\begin{aligned} N_{2,1}(z) =& \partial_{y_1}\partial_{y_2} \frac{z^2 M(z,0)}{(1 - z\tilde{M}(z,y_1))(1 - z\tilde{M}(z, y_1 + y_2))}\Big|_{y_1=0,y_2=0} \\ &- \frac{z^7 M(z,0)^4}{(1 - zM(z,0))^5}. \end{aligned}$$

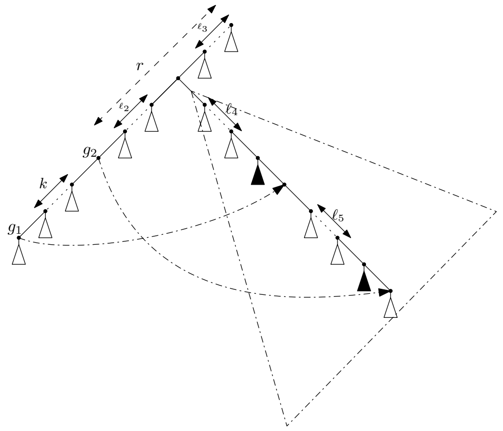


Figure 4.6: The case of Motzking skeleton that pointing of green vertices violate normality condition by making a shortcut.

The subtraction factor comes from the fact that normal networks need the additional requirement that evolution does not take shortcuts. So we have to take care of this property and subtract cases that violate the normality condition by making shortcuts. To see this better, consider an attached subtree (dashed triangle) on the path r with two red vertices which that one of them is an ancestor of another one and are targeted with green vertices g_1 and g_2 which are on the same path, and the pointers cross each other. This subtrees can be replaced with fixed paths ℓ_4 and ℓ_5 and two more subtree $\tilde{M}(z,0)$. To see this, note that from the terminology of the path definition, g_1 is not allowed to point to the first vertex of path l_4. Also the path ℓ_5 between two red vertices cannot be empty. To govern these situations and establish generating functions, add two attached subtrees $\tilde{M}(z,0)$ just before each red vertex (filled subtrees). These two subtrees with the subtree of g_1 and a tree which is attached to the last red vertex, contribute $\tilde{M}(z,0)^4$ altogether in second term.

For the second type (see Figure 4.5, right), none of the two green vertices g_1 and g_2 is the ancestor of the other, they have a common ancestor u. Moreover, there is a path on top of u connecting u with the root of the network, called r. Also, there are paths from u to g_1 and g_2, namely k and ℓ, respectively. To each of the vertices of k, ℓ and r as well as to the two green vertices a tree from $\mathcal{M}$ is attached.

In order to meet the constraints imposed by the tree-child and the normality property there are certain restrictions for the target vertices of the edges we add to the green vertices. We will analyse the parts of the structure. First, since the targets of the added edges are certainly reticulation vertices, the trees attached to a green vertex cannot be red trees (*cf.* the terminology at the beginning of Section 4.3.2) and have generating function $\tilde{M}(z,y)$. We only have to replace y by y_1 or y_2 or their sum, depending on whether g_1 or g_2 or both green vertices, respectively, are allowed to point at the red vertices in this tree (the last situation cannot happen here). The vertices u, g_1 and g_2 cause a factor z^3.

First, we analyse the contribution of the paths with this assumption, that there are not any red vertices on the paths.

- Path r: Both green vertices may point into the attached subtrees, except to their root. The trees are therefore white trees and the generating function of the path is $1/(1 - z\tilde{M}(z, y_1 + y_2)$.
- Path k: The vertex g_1 may point to any non-root vertex of the attached trees, g_2 may point to any vertex of the attached trees on k. Thus the generating function of this path is $P(z, 0, y_1 + y_2, y_2, 0)$.
- Path ℓ: The situation for this path is symmetric to k.

This leads to the first term of the following expression. Now consider the case where there is just one red vertex on the path ℓ or k , not being the first one, (here we consider ℓ) which is targeted by the green vertex that lies on the other side path (g_1); see Figure 4.7. Also, to avoid shortcut, g_2 is not allowed to point at any root vertex of the attached subtrees. This case contributes the second term. Overall, this yields the generating function

$$N_{2,2}(z) = \frac{1}{2}\partial_{y_1}\partial_{y_2}\frac{z^3\tilde{M}(z,y_1)\tilde{M}(z,y_2)P(z,0,y_1+y_2,y_1,0)P(z,0,y_1+y_2,y_2,0)}{1-z\tilde{M}(z,y_1+y_2)}\Bigg|_{y_1=0,y_2=0}$$
$$+\,\partial_{y_1}\partial_{y_2}\frac{z^3\tilde{M}(z,y_2)\tilde{M}(z,0)P(z,y_1,y_2,0,0)P(z,0,y_2,0,0)}{1-z\tilde{M}(z,y_2)}\Bigg|_{y_1=0,y_2=0}.$$

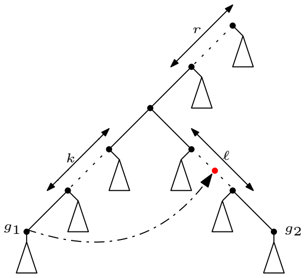


Figure 4.7: The structure of Motzkin skeletons of networks which one of the two possible reticulation vertices lies on a path. Should be noted that in this situation, phylogenetic properties not allowed to have two reticulation vertices on the paths.

The exponential generating function for normal networks with two reticulation vertices is then $N_2(z) = (N_{2,1}(z) + N_{2,2}(z))/4$, where the factor 4 appears, because each normal network is generated four times. Simplifying the resulting expression gives (5.3.1). □

As an easy consequence, we obtain the asymptotic number of networks.

Corollary 4.3.4. *Let $N_{2,n}$ denote the number of vertex-labeled normal networks with n vertices and two reticulation vertices. If n is even then $N_{2,n}$ is zero, otherwise*

$$N_{2,n} = n![z^n]N_2(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+3} \left(\frac{\sqrt{2}}{16} - \frac{3\sqrt{\pi}}{8} \cdot \frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Proof. This follows by singularity analysis as in the proof of Corollary 4.3.2. □

Remark. It turns out that the asymptotic main term is determined by $N_{2,2}(z)$. In hindsight, this is no surprise, because the corresponding sparsened skeleton has two edges, which leads to three paths made of sequences of trees after all. This leads to three expressions contributing a singularity in the denominator and together with the number of differentiations this eventually determines the growth rate of the coefficients of the generating function.

4.3.3 Normal networks with three reticulation vertices

In the case of three reticulation vertices we follow the same procedure: We decompose the network according to how the reticulation vertices are distributed in the network. There are four cases.

Case 1: The three green vertices lie on one path, *i.e.*, one green vertex is ancestor of another, which itself is ancestor of the third one.

Case 2: One green vertex is a common ancestor of the other two, but none of those two is ancestor of the other one.

Case 3: One green vertex is ancestor of another one, but not of both of them, and the third one is not ancestor of any other green vertex.

Case 4: No green vertex is ancestor of any other green vertex.

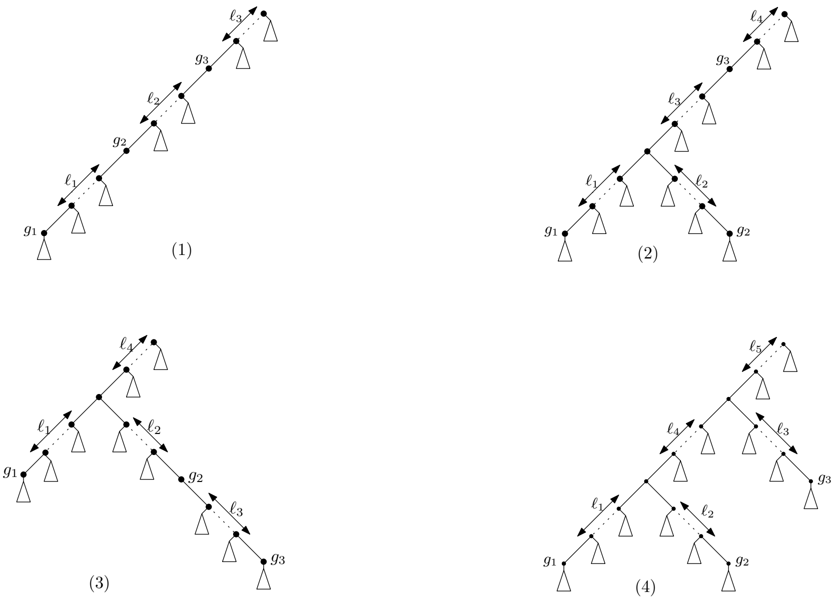


Figure 4.8: The four possible structures of Motzkin skeletons of networks with three reticulation vertices. The item (1) arises from the sparsened skeleton which is a path of length two and the case (2) arises from a unary vertex to which a cherry is attached. The third one arises from the sparsened skeleton which consists of a root with a left child and path of length 2 as right subtree. The fourth possible structure of Motzkin skeletons of network arises from a sparsened skeleton which is a rooted binary caterpillar with three leaves.

For all Motzkin skeletons, the lengths of the paths connecting two green vertices or connecting a green vertex with the last common ancestor of two green vertices are the free parameters. To each vertex of such a path we may attach a Motzkin tree which must be shaped in such a way that the condition for normality of the network is respected. So, we will set up generating functions $N_{3,1}(z)$, $N_{3,2}(z)$, $N_{3,3}(z)$, $N_{3,4}(z)$ associated to Motzkin skeletons of the four respective cases and since the procedure will generate each normal network eight times, the generating function of normal networks with three reticulation vertices is

$$N_3(z) = \frac{N_{3,1}(z) + N_{3,2}(z) + N_{3,3}(z) + N_{3,4}(z)}{8}.$$

We start with Case 1, see Figure 4.8, top left tree. As in the previous section we call the path from the root to the parent of g_1 the spine of the Motzkin skeleton.

•

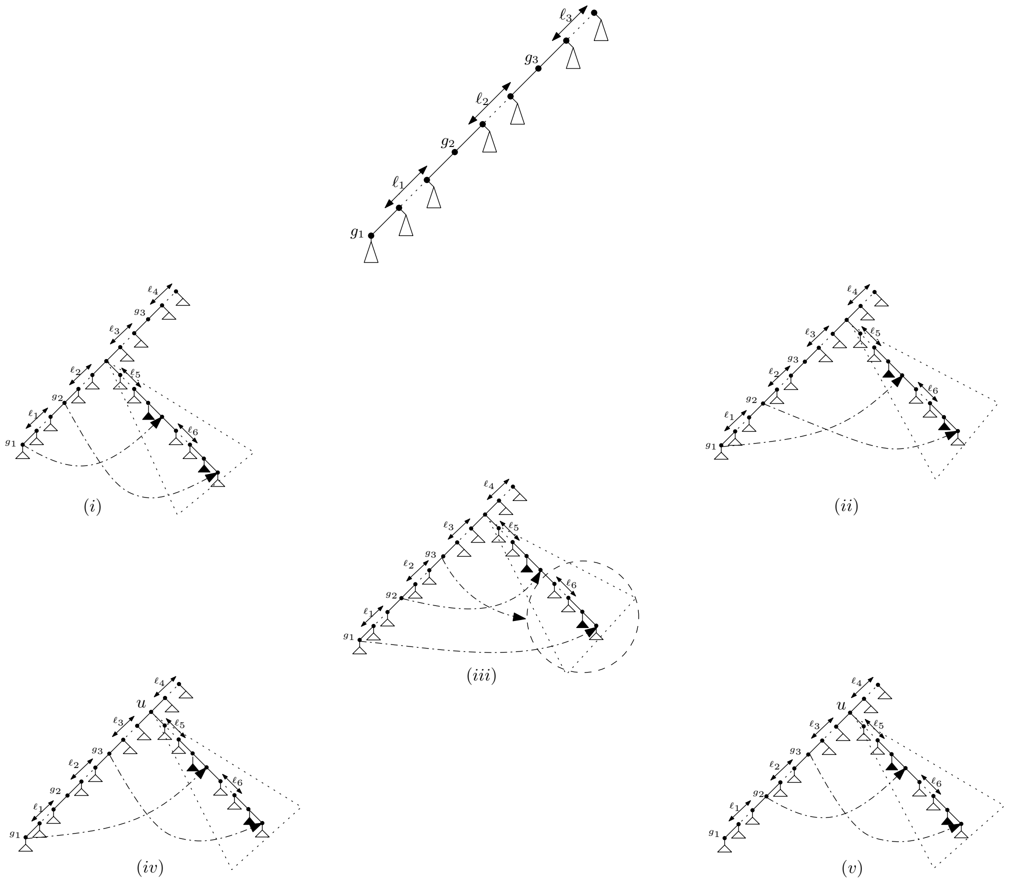


Figure 4.9: The colored Motzkin skeletons arising from the sparsened skeleton Case 1 with the five subtraction cases due to the creation of shortcuts. (i): the pointers of g_1 and g_2 cross each other and g_3 can point to any vertex such that the normal condition is preserved; (ii) and (iii): these are the cases where g_1 and g_2 point to the same path. in the former case, shortcut made up by pointing of g_1 and g_2, in the latter case, it comes up with pointing g_3 to the vertex inside the circle area. (iv) and (v): the remaining two cases with g_1 and g_2 not pointing on the same path. Thus, in (iv) resp. (v), g_2 resp. g_1 is not allowed to point to the path between u and the endpoint of the pointer from g_1 resp. g_2 and no vertex after that endpoint.

$$
\begin{aligned}
N_{3,1}(z) = \mathbf{Y}_{1,2,3} & \left(\frac{z^3 M(z,0)}{(1 - z\tilde{M}(z,y_1))(1 - z\tilde{M}(z,y_1+y_2))(1 - z\tilde{M}(z,y_1+y_2+y_3))} \right) \\
& - \partial_{y_3} \frac{z^8 \tilde{M}(z,0)^4}{(1 - z\tilde{M}(z,0))^5 (1 - z\tilde{M}(z,y_3))} \Big|_{y_3=0} && (i) \\
& - \partial_{y_3} \frac{z^8 \tilde{M}(z,0)\tilde{M}(z,y_3)^3 P(z,y_3,y_3,y_3,0)^2}{(1 - z\tilde{M}(z,0))^2 (1 - z\tilde{M}(z,y_3))^2} \Big|_{y_3=0} && (ii) \\
& - \partial_{y_3} \frac{z^8 \tilde{M}(z,0)^2\tilde{M}(z,y_3)^2 P(z,y_3,y_3,y_3,0)}{(1 - z\tilde{M}(z,0))^5} \Big|_{y_3=0} && (iii) \\
& - \partial_{y_2} \frac{z^8 \tilde{M}(z,0)^3\tilde{M}(z,y_2) P(z,0,y_2,y_2,0)}{(1 - z\tilde{M}(z,0))^2 (1 - z\tilde{M}(z,y_2))^3} \Big|_{y_2=0} && (iv) \\
& - \partial_{y_1} \frac{z^8 \tilde{M}(z,0)^3\tilde{M}(z,y_1) P(z,0,y_1,y_1,0)}{(1 - z\tilde{M}(z,0)) (1 - z\tilde{M}(z,y_1))^4} \Big|_{y_1=0}, && (v)
\end{aligned}
$$

where $\tilde{M}(z,y)$ is as in the last subsection (*cf.* (4.5)) and $\mathbf{Y}_{1,2,3}$ denotes the operator differentiating with respect to $y_1,\ y_2,\ y_3$ and setting $y_1 = y_2 = y_3 = 0$ afterwards, *i.e.*, $\mathbf{Y}_{1,2,3} f(z,y_1,y_2,y_3) = (\partial_{y_1}\partial_{y_2}\partial_{y_3} f)(z,0,0,0)$. The first expression comes from the fact that g_1 can point to each non-root vertex of the subtrees attached to any of the vertices of the spine. Likewise, the pointing options for g_2 are the non-root vertices of the subtrees attached to the vertices of the sub-paths $\ell_2 \cup \ell_3$ of the spine. The situation for g_3 is analogous. Now in the following we subtract cases that violate normality condition by making shortcuts and do not considered so far.

(i) Fix the shortcut which is made up of pointing g_1 and g_2 to an attached subtree between g_2 and g_3. So in this case, green vertex g_3 may point to any non-root vertex of the attached trees of path ℓ_4;

(ii) g_3 may point to any vertex on paths ℓ_5 and ℓ_6, except the first one, and any vertex of their attached trees (thus the generating function of these paths are $P(z,y_3,y_3,y_3,0)^2$) or point to any non-root vertex of the attached trees of path ℓ_3 and ℓ_4, filled attached white trees, and a white subtree which is attached to the red node in the consequence of the path ℓ_6;

(iii) g_3 may point to any vertex on path ℓ_6, except the first one, any vertex of its attached trees and non-root vertex of the two attached subtrees afterwards;

(iv) a green vertex g_2 may point to any non-root vertex of the attached trees of the paths $\ell_2 \cup \ell_3 \cup \ell_4 \cup \ell_5$ with a white subtree after that (marked as a filled tree).

(v) a green vertex g_1 may point to any non-root vertex of the attached trees of paths $\ell_1 \cup \ell_2 \cup \ell_3 \cup \ell_4 \cup \ell_5$ and a white subtree thereafter ℓ_5.

Next we will determine the generating function of all normal networks belonging to Case 2, which have Motzkin skeletons as shown on the top right of Figure 4.8. As in the previous section we analyse the substructures. There are four vertices in the sparsened skeleton, yielding a factor z^4. The red vertices in the (white) subtree attached to g_1 may only be targets of the edge coming from g_2, for the subtree attached to g_2 *vice versa*.

- Paths ℓ_3 and ℓ_4: These paths are sequences of vertices, each with a white subtree attached to it. For ℓ_4 each green vertex is allowed to point at the red vertices in these white subtrees. Pointing to the vertices of the path is not allowed. Likewise, the corresponding vertices in the subtrees of ℓ_3 are forbidden for g_3 by the normality condition.

- Paths ℓ_1 and ℓ_2: There are two possible cases. first, none of green vertices g_1 and g_2 point to the paths vertices. Then in this way they are symmetric, so we discuss ℓ_1. The non-root vertices of the subtrees are the only allowed targets for the edge from g_1. The edge from g_2 may end at each vertex of the subtrees. There are no options for g_3.

$$\begin{aligned}
N_{3,2}(z) = \frac{1}{2}\mathbf{Y}_{1,2,3}&\left(\frac{z^4\tilde{M}(z,y_1)\tilde{M}(z,y_2)P(z,0,y_1+y_2,y_2,0)P(z,0,y_1+y_2,y_1,0)}{(1-z\tilde{M}(z,y_1+y_2))(1-z\tilde{M}(z,y_1+y_2+y_3))}\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,0)\tilde{M}(z,y_2)P(z,y_1,y_2,0,0)P(z,0,y_2,0,0)}{(1-z\tilde{M}(z,y_2))(1-z\tilde{M}(z,y_2+y_3))}\right)\\
&-\frac{1}{2}\partial y_1\partial y_2\frac{z^7\tilde{M}(z,0)^4P(z,y_1+y_2,0,0,0)}{(1-z\tilde{M}(z,0))^5}\Big|_{y_1=0,y_2=0} \qquad (i)\\
&-\partial y_1\partial y_2\frac{z^7\tilde{M}(z,0)\tilde{M}(z,y_1)^3P(z,y_2,y_1,y_1,0)P(z,y_1,y_1,y_1,0)}{(1-z\tilde{M}(z,y_1))^4}\Big|_{y_1=0,y_2=0}. \qquad (ii)
\end{aligned}$$

Then,we obtain the first term. The second expression comes from restricting the green vertex g_1 to point on at a vertex on the path ℓ_2 except the first one. To avoid shortcut, The edge from g_2 cannot point to the root-vertex of the subtrees ℓ_1

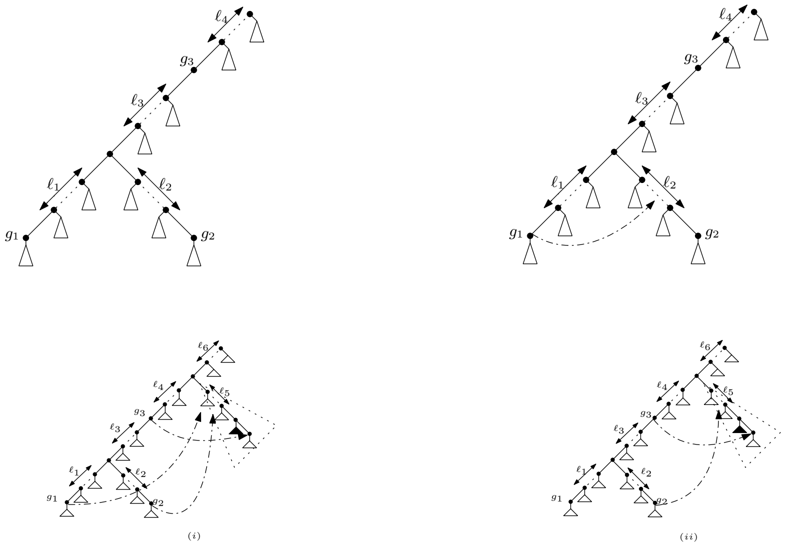


Figure 4.10: Case 2 with two possible Motzkin skeletons which are not respecting the normality condition. In (i) The three reticulation vertices lie on one path, and a shortcut is created by pointing of g_3. In the case (ii) the green vertex g_1 has some freedom in pointing but not allowed to point any vertex on path ℓ_5.

anymore. Now we take care of two possible situations which violate the normality condition and we did not consider yet. Fix the target (red colored) vertex for pointing g_3, as depicted in items (i) and (ii). So we may have,

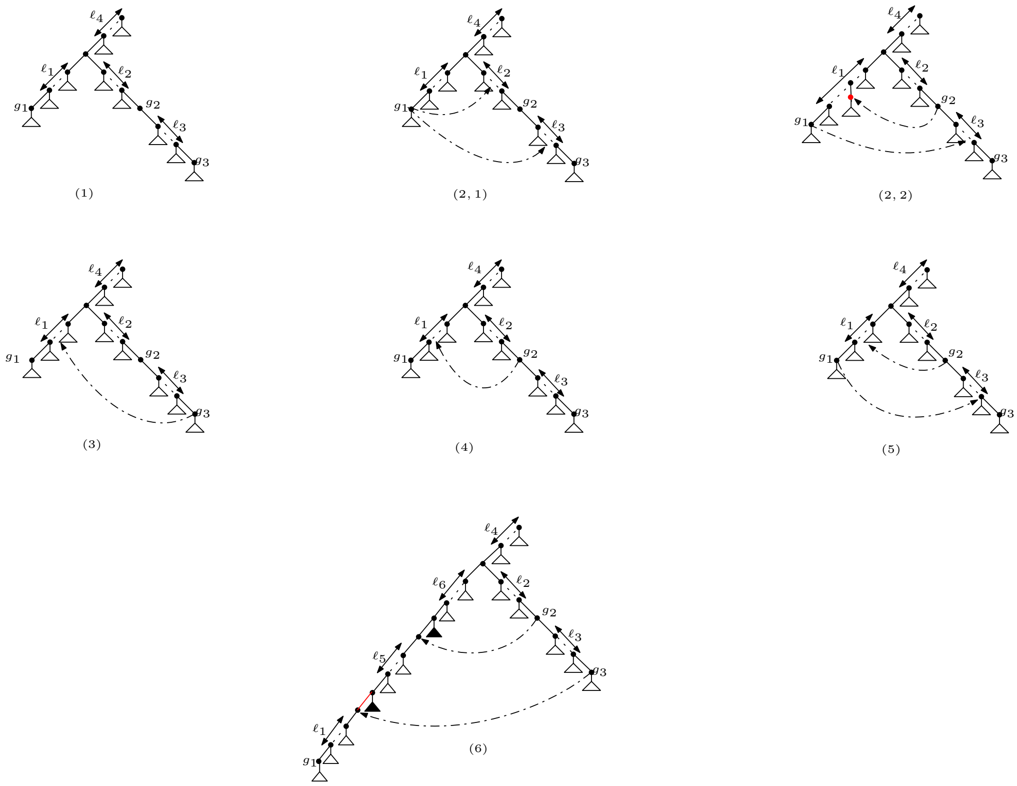
(i) both green vertices g_1 and g_2 point at vertices (not the very first vertex) on path ℓ_5;

(ii) Only g_2 points to a vertex on path ℓ_5 but not first one.

Case 3 is the one shown in down left of Figure 4.8. The sparsened skeleton has 4 vertices and the subtrees attached to g_1 and g_3 are white trees. The red vertices of the subtree of g_1 may be targeted by the edges starting either in g_2 or in g_3, the red vertices of the other tree by edges from g_1. In order to better understanding, we obtain the generating function of case 3 by analyzing the contribution of each case regarding the possible distribution of red vertices on the paths.

Item (1). Start with a Motzkin tree skeleton with no red vertex (for all paths $y = \hat{y} = 0$) on the paths (Figure 4.11 (1)).

- Path ℓ_4: All green vertices may point to the non-root vertices of the (white) subtrees. Pointing to the path itself is not allowed.



(1) (2, 1) (2, 2)

(3) (4) (5)

(6)

Figure 4.11: Motzkin skeletons of case 3 which may have up to 2 reticulation vertices on paths. Missing pointers of green vertices are not allowed to point to a path.

- Path ℓ_3: The edge starting at g_3 may point to non-root vertices of the subtrees. There is no option for g_2. All tree vertices can be the end point of the edge starting at g_1.
- Path ℓ_1: Similar to ℓ_3. The edges from g_2 and g_3 may point anywhere tree vertices. The non-root vertices of the subtrees may be targeted by g_1 as well.
- Path ℓ_2: g_2 and g_3 may point to the non-root vertices of the subtrees. To point at the root vertices of the subtrees is only allowed for g_1.

Item $(2, 1)$. Consider a red vertex, but not the first one, that lies on the paths ℓ_2 or ℓ_3, and is targeted by g_1. Except the path ℓ_1 the pointing roles for g_2 and g_3 are like before. For this path green nodes not allowed point to the root vertices of the subtrees anymore.
Item $(2, 2)$. Consider a red vertex, again not the first one, on the path ℓ_3 that is targeted by g_1. Then consider g_2 which point to a root vertex of subtrees on path ℓ_1. So g_3 does not allow to point any root vertices of subtrees.
Item (3). Consider a non-initial red vertex on the paths ℓ_1 which is targeted by g_3. The pointing roles for g_2 is similar as first item but g_1 not allowed to point root vertices of the subtrees on paths ℓ_2 and ℓ_3 anymore.
Item (4). Here we consider that g_2 pointed a red vertex on the path ℓ_1 (except first one). Consequently, because of avoiding a shortcut, g_1 is not allowed to point at root vertices of subtrees on path ℓ_2. The pointing option for g_3 is similar as first part.
Item (5). Suppose there are two (not the first one) red vertices on paths ℓ_1 and ℓ_3, which are targeted by g_2 and g_1, respectively. The green node g_3 may point to a non-root vertex of the subtrees on all paths.
Item (6).In the last case, consider two red vertices on the path that reaches g_1 which are targeted by g_2 and g_3, respectively, as like show in Figure 4.11 (7). So g_1 cannot point to any root vertices of the attached subtrees anymore.
Note that the shortcut may only occur in the first three cases. Then, we subtract from this the exponential-generating functions of all cases where networks contain shortcut. We can see all possible cases in Figure 4.12. The first column contains the cases from Figure 4.11 which contain the shortcuts in the second column. Altogether, we obtain for the generating function $N_{3,3}(z)$ of Case 3 the following expression.

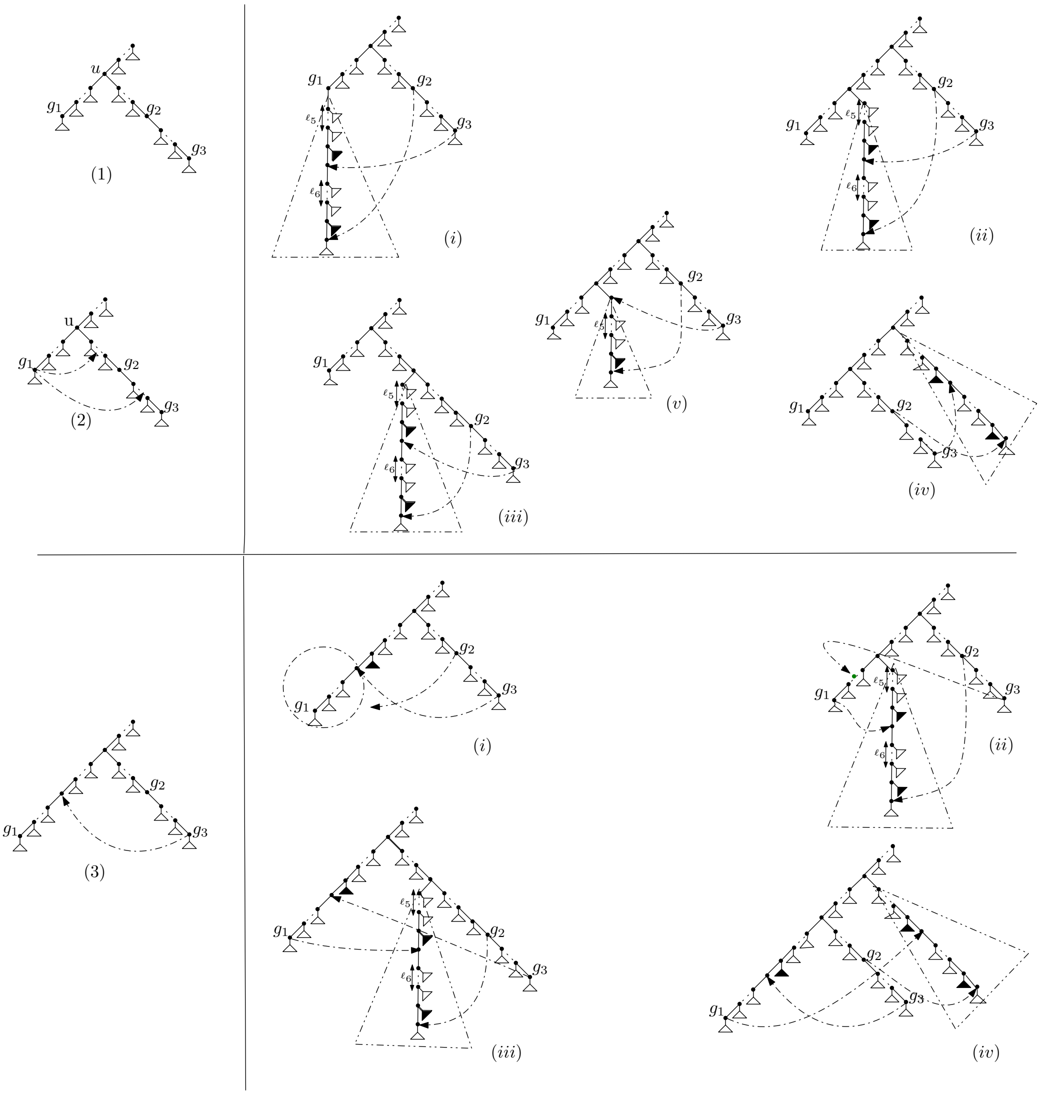


Figure 4.12: All possible Motzkin skeletons that violate normality condition by creating shortcuts. In the first row, only g_2 and g_3 are on the same path, so a shortcut is only created if their pointers cross and point on the same path. That path can be after u (cases (i), (ii), and (v) in the second column); between u and g_2 (case (iii) in the second column); or before u (case (iv) in the second column). In the second row g_2, g_3, and g_1 are all on the same path (due to the pointer of g_3). So, one needs to subtract the cases where the pointers from g_2 and g_3 cross and point to the same path (case (i) in the second column) and where the pointers from g_1 and g_2 cross and point to the same path (cases (ii), (iii), and (iv) in the second column).

$$
\begin{aligned}
N_{3,3}(z) = {} & \mathbf{Y}_{1,2,3}\Bigg(\frac{z^4\tilde{M}(z,y_1)\tilde{M}(z,y_2+y_3)}{(1-z\tilde{M}(z,y_1+y_2+y_3))}P(z,0,y_1+y_3,y_1,0)P(z,0,y_1+y_2+y_3,y_1,0)\\
& \times P(z,0,y_1+y_2+y_3,y_2+y_3,0)\Bigg)\\
& + \mathbf{Y}_{1,2,3}\Bigg(\frac{z^4\tilde{M}(z,0)\tilde{M}(z,y_2+y_3)}{(1-z\tilde{M}(z,y_2+y_3))}P(z,y_1,y_3,0,0)P(z,y_1,y_2+y_3,0,0)P(z,0,y_2+y_3,0,0)\Bigg)\\
& + \mathbf{Y}_{1,2,3}\Bigg(\frac{z^4\tilde{M}(z,0)\tilde{M}(z,y_3)}{(1-z\tilde{M}(z,y_3))}P(z,y_1,y_3,0,0)P(z,0,y_3,0,0)P(z,0,y_3,y_2,0)\Bigg)\\
& + \mathbf{Y}_{1,2,3}\Bigg(\frac{z^4\tilde{M}(z,y_1)\tilde{M}(z,y_2)}{(1-z\tilde{M}(z,y_1+y_2))}P(z,y_3,y_1+y_2,y_2,0)P(z,0,y_1+y_2,0,0)P(z,0,y_1,0,0)\Bigg)\\
& + \mathbf{Y}_{1,2,3}\Bigg(\frac{z^4\tilde{M}(z,y_1)\tilde{M}(z,y_3)}{(1-z\tilde{M}(z,y_1+y_3))}P(z,y_2,y_1+y_3,y_3,0)P(z,0,y_1+y_3,0,0)P(z,0,y_1+y_3,y_1,0)\Bigg)\\
& + \mathbf{Y}_{1,2,3}\Bigg(\frac{z^4\tilde{M}(z,y_1)\tilde{M}(z,y_3)}{(1-z\tilde{M}(z,y_3))^2}P(z,y_2,y_3,0,0)P(z,y_1,y_3,0,0)\Bigg)\\
& + \partial y_1\frac{z^8\tilde{M}(z,0)\tilde{M}(z,y_1)^3}{(1-z\tilde{M}(z,y_1))^6}\Big|_{y_1=0}\\
& - \partial y_1\Bigg(\frac{z^8\tilde{M}(z,0)^3\tilde{M}(z,y_1)P(z,y_1,y_1,y_1,0)^2}{(1-z\tilde{M}(z,y_1))^2(1-z\tilde{M}(z,0))^2}\\
& + \frac{z^9\tilde{M}(z,0)\tilde{M}(z,y_1)^4P(z,y_1,y_1,y_1,0)^4}{(1-z\tilde{M}(z,y_1))^3}\\
& + \frac{z^9\tilde{M}(z,0)\tilde{M}(z,y_1)^4P(z,y_1,y_1,y_1,0)^3P(z,y_1,y_1,y_1,y_1)^2}{(1-z\tilde{M}(z,y_1))^2}\\
& + \frac{z^8\tilde{M}(z,0)\tilde{M}(z,y_1)^3P(z,0,y_1,y_1,0)^2P(z,y_1,y_1,y_1,0)}{(1-z\tilde{M}(z,y_1))^3}\\
& + \frac{z^9\tilde{M}(z,0)\tilde{M}(z,y_1)^4P(z,y_1,y_1,y_1,0)^4}{(1-z\tilde{M}(z,y_1))^3}\Bigg)\Big|_{y_1=0}\\
& - \partial y_1\partial y_2\frac{z^6\tilde{M}(z,y_2)\tilde{M}(z,y_1)^2P(z,0,y_1+y_2,y_2,0)}{(1-z\tilde{M}(z,y_1))^4}\Big|_{y_1=0,y_2=0}\\
& - \frac{z^{10}\tilde{M}(z,0)^5}{(1-z\tilde{M}(z,0))^8} - \frac{z^{11}\tilde{M}(z,0)^6}{(1-z\tilde{M}(z,0))^8} - \frac{z^{11}\tilde{M}(z,0)^6}{(1-z\tilde{M}(z,0))^8}.
\end{aligned}
$$

The last case of normal networks has Motzkin skeletons as shown in down right of Figure 4.8. The restriction for the target vertex of the edges to be added at g_1, g_2 and g_3 follow the analogous rules in order to meet the normality constraint. Setting up the generating function follows the same pattern as before. We omit now the details and get from the path analysis after all,

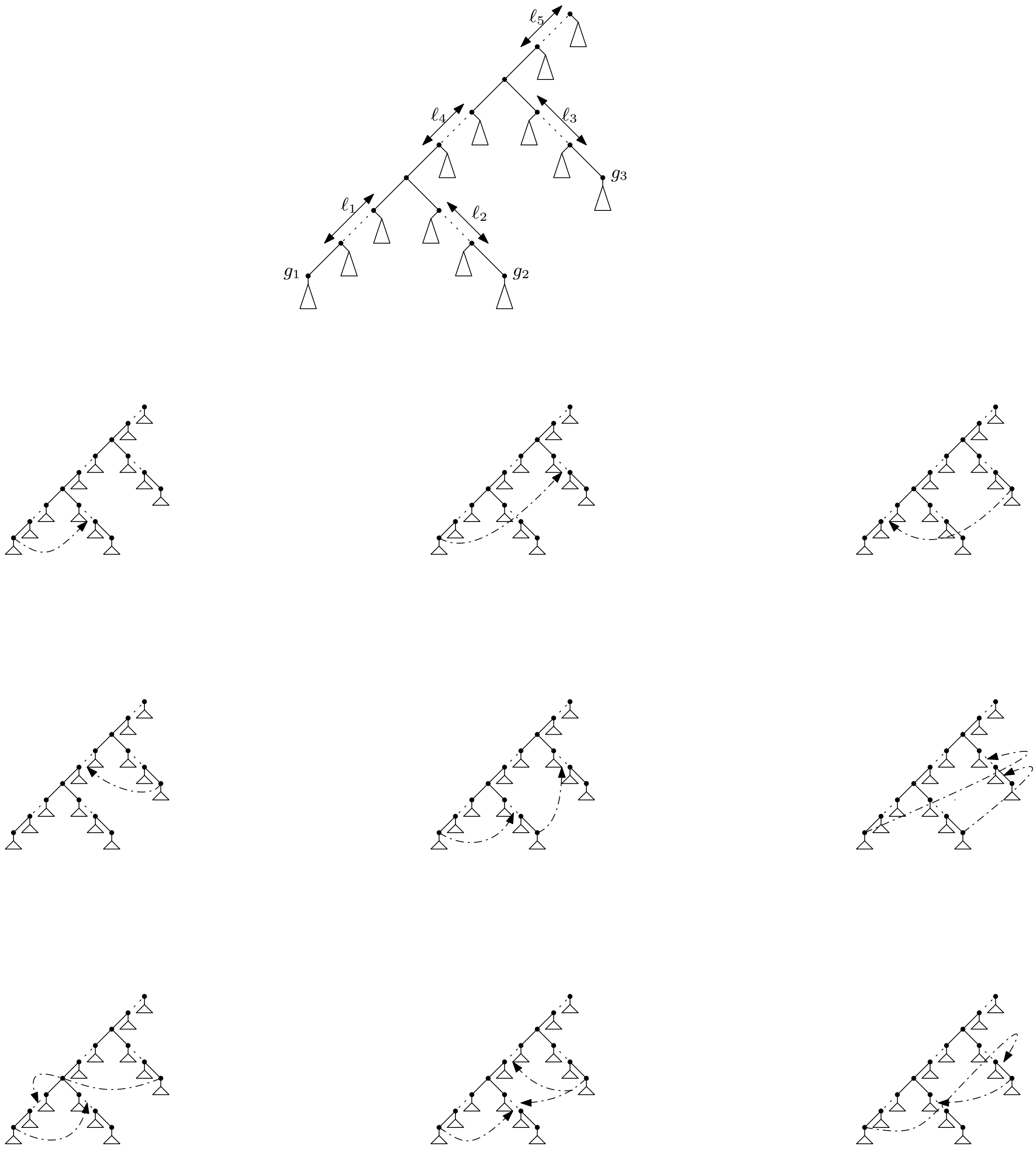


Figure 4.13: The case 4 with Motzkin skeletons of networks with three reticulation vertices such that maximum 2 of them can be lie on the paths.

$$
\begin{aligned}
N_{3,4}(z) = \frac{1}{2}\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_1+y_2)\tilde{M}(z,y_1+y_3)\tilde{M}(z,y_2+y_3)}{(1-z\tilde{M}(z,y_1+y_2+y_3))}P(z,0,y_1+y_2+y_3,y_1+y_2,0)\\
&\times P(z,0,y_1+y_2+y_3,y_1+y_3,0)P(z,0,y_1+y_2+y_3,y_2+y_3,0)P(z,0,y_1+y_2+y_3,y_3,y_3,0)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_2)\tilde{M}(z,y_3)\tilde{M}(z,y_2+y_3)}{(1-z\tilde{M}(z,y_2+y_3))}P(z,y_1,y_2+y_3,y_3,0)P(z,0,y_2+y_3,y_2,0)\\
&\times P(z,0,y_2+y_3,y_3,0)^2\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_2)\tilde{M}(z,y_3)\tilde{M}(z,y_2+y_3)}{(1-z\tilde{M}(z,y_2+y_3))}P(z,y_1,y_2+y_3,y_2,0)\\
&\times P(z,0,y_2+y_3,y_3,0)P(z,0,y_2+y_3,y_2,0)P(z,0,y_2+y_3,0,0)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_1)\tilde{M}(z,y_2)\tilde{M}(z,y_1+y_2)}{(1-z\tilde{M}(z,y_1+y_2))}P(z,y_3,y_1+y_2,y_2,y_3)\\
&\times P(z,0,y_1+y_2,y_1,0)P(z,0,y_1+y_2,y_2,0)P(z,0,y_1+y_2,0,0)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_1)\tilde{M}(z,y_2)\tilde{M}(z,y_1+y_2)}{(1-z\tilde{M}(z,y_1+y_2))}P(z,y_3,y_1+y_2,0,0)\\
&\times P(z,0,y_1+y_2,y_1,0)P(z,0,y_1+y_2,y_2,0)P(z,0,y_1+y_2,0,0)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_3)^2\tilde{M}(z,0)}{(1-z\tilde{M}(z,y_3))^3}P(z,y_1,y_3,0,0)P(z,y_2,y_3,0,0)\Bigg)\\
+\frac{1}{2}\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_3)^2\tilde{M}(z,0)}{(1-z\tilde{M}(z,y_3))^4}P(z,y_1+y_2,y_3,0,0)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_2)^2\tilde{M}(z,0)}{(1-z\tilde{M}(z,y_2))^3}P(z,y_1,y_2,0,0)P(z,y_3,y_2,0,y_3)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_2)^2\tilde{M}(z,0)}{(1-z\tilde{M}(z,y_2))^3}P(z,y_1+y_3,y_2,0,y_3)P(z,y_3,y_2,0,0)\Bigg)\\
+\mathbf{Y}_{1,2,3}\Bigg(&\frac{z^5\tilde{M}(z,y_2)^2\tilde{M}(z,0)}{(1-z\tilde{M}(z,y_2))^3}P(z,y_1,y_2,0,0)P(z,y_3,y_2,0,0)\Bigg).
\end{aligned}
$$

Overall, by collecting everything, we obtain the following result.

Proposition 4.3.5. *The exponential generating function for vertex-labeled normal* *networks with three reticulation vertices is*

$$N_3(z) = z\frac{a_3(z^2) - b_3(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{11/2}}, \tag{4.11}$$

where

$$\tilde{a}_3(z) = 877z^6 - 3065z^5 + 2392z^4 - 628z^3 + 64z^2$$

and

$$\tilde{b}_3(z) = 110z^6 - 1455z^5 + 1860z^4 - 564z^3 + 64z^2$$

As a consequence we obtain the following result.

Corollary 4.3.6. *Let $N_{3,n}$ denote the number of vertex-labeled normal networks with n vertices and three reticulation vertices. If n is even then $N_{3,n}$ is zero, otherwise*

$$N_{3,n} = n![z^n]N_3(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+5}\left(\frac{\sqrt{2}}{192} - \frac{3\sqrt{\pi}}{64}\cdot\frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Proof. This follows by singularity analysis as for $k = 1$ and $k = 2$ above. □

4.3.4 Normal networks with a fixed number of reticulation vertices

By looking at Proposition 4.3.1, Proposition 4.3.3 and Proposition 4.3.5, one clearly sees a pattern for the exponential generating function of normal networks. In this section, we will prove that this pattern continues to hold for the exponential generating function of normal networks with k reticulation vertices. This will then be used to prove the remaining claims of Theorem 4.1.1.

We start with a technical lemma. Therefore, consider the following function

$$G(z,y) = \frac{a(z,y) - b(z,y)\sqrt{1-2z^2-4yz^3}}{1+2yz}, \tag{4.12}$$

where $a(z,y), b(z,y)$ are polynomials in z and y with $a(z,0) = b(z,0) = 1$. This function will be used as a building block for construction the exponential generating function of normal networks. We need the following simple properties of this function.

Lemma 4.3.7. *(a) For all $\ell \geq 1$,*

$$\frac{\partial^\ell}{\partial y^\ell} G(z,y)\Big|_{y=0} = \frac{c_\ell(z) - d_\ell(z)\sqrt{1-2z^2}}{(1-2z^2)^{\ell-1/2}},$$

where $c_\ell(z)$ and $d_\ell(z)$ are suitable polynomials.

(b) For all $\ell \geq 0$,

$$\frac{\partial^\ell}{\partial y^\ell} \frac{1}{1-G(z,y)}\Big|_{y=0} = \frac{e_\ell(z) - f_\ell(z)\sqrt{1-2z^2}}{(1-2z^2)^{\ell+1/2}},$$

where $e_\ell(z)$ and $f_\ell(z)$ are suitable polynomials.

Proof. For the proof of part (a), by differentiation

$$\frac{\partial^\ell}{\partial y^\ell} G(z,y) = \frac{a_\ell(z,y) - b_\ell(z,y)\sqrt{1-2z^2-4yz^3}}{(1+2yz)^{\ell+1}(1-2z^2-4yz^3)^{\ell-1/2}}$$

with suitable polynomials $a_\ell(z,y)$ an $b_\ell(z,y)$. (Note that this becomes incorrect for $\ell = 0$). The claim follows now by setting $y = 0$.

For the proof of part (b), we use induction. Note that $\ell = 0$ is trivial. Now, assume that the claim holds for all $\tilde{\ell} < \ell$. Then, by Leibnitz rule

$$\frac{\partial^\ell}{\partial y^\ell} \frac{1}{1-G(z,y)}\Big|_{y=0} = \frac{\partial^{\ell-1}}{\partial y^{\ell-1}} \left(\frac{1}{1-G(z,y)} \cdot \frac{1}{1-G(z,y)} \cdot G'(z,y) \right)\Big|_{y=0}$$

$$= \sum_{k_1+k_2+k_3=\ell-1} \binom{\ell-1}{k_1,k_2,k_3} \frac{\partial^{k_1}}{\partial y^{k_1}} \frac{1}{1-G(z,y)}\Big|_{y=0} \cdot \frac{\partial^{k_2}}{\partial y^{k_2}} \frac{1}{1-G(z,y)}\Big|_{y=0} \cdot G^{(k_3+1)}(z,y)\Big|_{y=0}.$$

Plugging into this the induction hypothesis and part (a) gives the claimed form with power of the denominator equal to

$$k_1 + 1/2 + k_2 + 1/2 + k_3 + 1/2 = \ell + 1/2.$$

This proves the result. □

Now, we can prove the following result which generalizes Proposition 4.3.1, Proposition 4.3.3 and Proposition 4.3.5.

Proposition 4.3.8. *The exponential generating function for vertex-labeled normal networks with k reticulation vertices is*

$$N_k(z) = \frac{a_k(z) - b_k(z)\sqrt{1-2z^2}}{(1-2z^2)^{2k-1/2}},$$

where $a_k(z)$ and $b_k(z)$ are suitable polynomials.

Proof. Fix a type of Motzkin skeletons (arising from a sparsened skeleton) for generating normal networks with k reticulation vertices. As explained in the cases $k = 1, 2, 3$, the exponential generating function of the normal networks arising from these skeletons is a product of generating functions for the paths which are either counted by $1/(1 - z\tilde{M})$ or P multiplied with a z for each vertex of the sparsened skeleton and the generating functions of the Motzkin trees attached to the leaves. In particular note that $z\tilde{M}$ is of the form (4.12) and the denominators of P is one minus a function of the form (4.12). Also, note that all these functions G have polynomials satisfying $a(z, 0) = b(z, 0) = 1$.

In summary, we have that the exponential generating function $N_k(z)$ for normal networks is a sum of terms of the form

$$\partial_{y_1} \cdots \partial_{y_k} \frac{G_1(z, y) \cdots G_s(z, y)}{(1 - G_{s+1}(z, y)) \cdots (1 - G_{s+t}(z, y))}\Big|_{y_1=0,\dots,y_k=0}, \tag{4.13}$$

where the number of functions $G_{s+i}(z, y)$ is bounded by the number of edges of the sparsened skeleton increased by one (for the sequence of trees added above the root when constructing the Motzkin skeletons). Moreover, y is the sum of the y_i's where not all of them must be present and the missing ones can differ from one occurrence to the next in the above formula. Note that because of this special form of y, we can apply the above lemma after expanding (4.13) and obtain that

$$N_k(z) = \frac{a_k(z) - b_k(z)\sqrt{1 - 2z^2}}{(1 - 2z^2)^p}. \tag{4.14}$$

What remains is to show that $p = 2k - 1/2$. For this observe that (4.13) without the derivatives is of the general form given in (4.14) with the exponent of the denominator equals $t/2$ which reaches its maximum for the sparsened skeleton with the maximal number of edges, indeed sparsened skeletons which all green vertices lie on leaves, and is thus at most $k - 1/2$. Also, from the above lemma, we see that each differentiation increases the exponent by 1. Thus, the exponent of (4.13) when written as (4.14) is at most $2k - 1/2$. Adding up this terms gives the claim. □

Corollary 4.3.9. *We have*

$$N_k(z) = z\frac{\tilde{a}_k(z^2) - \tilde{b}_k(z^2)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{2k-1/2}},$$

where $\tilde{a}_k(z)$ and $\tilde{b}_k(z)$ are suitable polynomials.

Proof. Observe that $N_k(-z) = -N_k(z)$ since phylogenetic networks necessarily have an odd number of vertices. Thus,

$$\frac{a_k(-z) - b_k(-z)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{2k-1/2}} = -\frac{a_k(z) - b_k(z)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{2k-1/2}}.$$

This implies

$$a_k(-z) + a_k(z) = (b_k(-z) + b_k(z))\sqrt{1-2z^2}$$

which is obviously only possible if

$$a_k(-z) = -a_k(z) \qquad \text{and} \qquad b_k(-z) = -b_k(z),$$

i.e., both are odd functions. From this the result follows. □

Now, we can finish the proof of Theorem 4.1.1.

Corollary 4.3.10. *Let $N_{k,n}$ denote the number of vertex-labeled normal networks with n vertices and k reticulation vertices. If n is even then $N_{k,n}$ is zero, otherwise there is a positive constant $\tilde{c}_k$ such that*

$$N_{k,n} = n![z^n]N_k(z) \sim \tilde{c}_k \left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1},$$

$n \to \infty$.

Proof. From the above corollary,

$$N_{n,k} = n![z^n]z\frac{\tilde{a}_k(z^2) - \tilde{b}_k(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{2k-1/2}}.$$

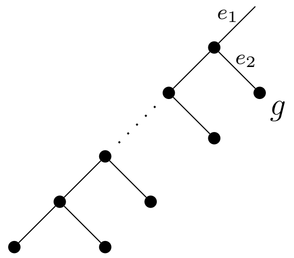
From this, by singularity analysis and Stirling's formula, the claimed expansion follows with

$$\tilde{c}_k = \frac{2\sqrt{2\pi}\tilde{a}_k(1/2)}{4^k\Gamma(2k-1/2)}.$$

What is left is to prove that $\tilde{c}_k > 0$ (note that we already showed this for $k = 1, 2, 3$ directly). This will follow from Proposition 4.3.11 below which shows that already a subset of the set of normal networks with k reticulation vertices satisfies the above claimed asymptotics with a positive constant. □

The proof of Corollary 4.3.10 relies on the fact that a certain constant (called $\tilde{c}_k$ there) is positive. This constant is related to the number of normal phylogenetic networks; it is the multiplicative constant of the asymptotic main term. We will construct a subclass of the class of normal networks and show that the number of networks in that subclass is the same as for normal networks up to a positive multiplicative constant. The result is presented in Proposition 4.3.11 below and closes the small gap left in the proof of Corollary 4.3.10.

For this purpose, we consider all the normal networks which are generated (possibly with duplicity) from a sparsened skeleton which is a rooted binary caterpillar, *i.e.*, a sparsened skeleton of the form



(For the discussion below, we have added an edge from the root.) Note that by the same arguments as above, these networks are also counted by an exponential generating function of the form

$$C_k(z) = z\frac{\tilde{e}_k(z^2) - \tilde{f}_k(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{2k-1/2}},$$

where $\tilde{e}_k(z)$ and $\tilde{f}_k(z)$ are suitable polynomials.

Now, we are in position to prove the following proposition.

Proposition 4.3.11. *Let $C_{k,n}$ denote the number of vertex-labeled normal networks with n vertices and k reticulation vertices which arise from the above caterpillar-skeleton. If n is even then $C_{k,n}$ is zero, otherwise there is a positive constant $\tilde{d}_k$ such that*

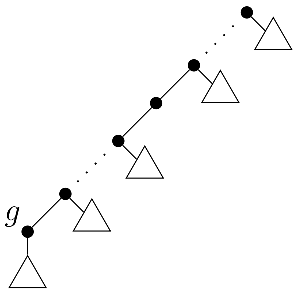
$$C_{k,n} = n![z^n]C_k(z) \sim \tilde{d}_k\left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1},$$

as $n \to \infty$.

Proof. As in the proof of Corollary 4.3.10, the asymptotic formula follows from (4.3.4), where

$$\tilde{d}_k = \frac{2\sqrt{2\pi}\tilde{e}_k(1/2)}{4^k\Gamma(2k-1/2)}.$$

For the positivity claim, we will show that $\tilde{e}_k(1/2)$ is non-decreasing in k from which the claim follows by our result for $k = 1$. In order to prove this, consider the caterpillar-skeleton above with k leaves. Denote the path consisting of the edges e_1 and e_2 by P. Then, a subset of all normal networks generated by this caterpillar-skeleton of k leaves is formed by normal networks which are generated by a caterpillar-skeleton with $k-1$ leaves to which a normal network with one reticulation vertex generated by P is added. More precisely, for the latter networks g is connected to one of the subtrees attached to e_1 or e_2 (such that the normal condition is satisfied), i.e., these networks arise from



and are counted by

$$P(z) = \partial_y \frac{z^2 M(z,0)}{(1 - z\tilde{M}(z,y))^2}\Big|_{y=0} = \frac{8z^2 - 12z^4 - (8z^2 - 4z^4)\sqrt{1-2z^2}}{(1-2z^2)^2},$$

where $\tilde{M}(z,y)$ is as above. Consequently, the normal networks from the above mentioned subset are counted by

$$C_{k-1}(z)P(z) = z\frac{\tilde{p}_k(z^2) - \tilde{q}_k(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{2k-1/2}},$$

where

$$\begin{aligned}\tilde{p}_k(z^2) &= (8z^2 - 12z^4)\tilde{e}_{k-1}(z^2) + (8z^2 - 4z^4)(1-2z^2)\tilde{f}_{k-1}(z^2);\\ \tilde{q}_k(z^2) &= (8z^2 - 4z^4)\tilde{e}_{k-1}(z^2) + (8z^2 - 12z^4)\tilde{f}_{k-1}(z^2).\end{aligned}$$

This gives, for odd n,

$$n![z^n]C_{k-1}(z)P(z) \sim \tilde{g}_k \left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1}$$

with

$$\tilde{g}_k = \frac{2\sqrt{2\pi}\tilde{e}_{k-1}(1/2)}{4^k\Gamma(2k-1/2)} > 0.$$

Moreover, since this counts a subclass of normal networks generated by a caterpillar-skeleton with k leaves, we have $\tilde{d}_k \geq \tilde{g}_k$ which gives $\tilde{e}_k(1/2) \geq \tilde{e}_{k-1}(1/2)$. This proves our claim and thus the proposition is also proved. □

Finally, we would like to remark that in order to compute the multiplicative constant in the asymptotic expression given in Corollary 4.3.10 one has to understand the precise structure of the generating functions for each Motzkin skeleton. Our investigations show that the main contribution comes from the Motzkin skeletons for which the sparsened skeleton is a (rooted, nonplane) tree with k vertices. Since there is no explicit formula for the number of such trees (but in fact there is an asymptotic solution; see [20]), we cannot expect to get some explicit form for the constant, but only some expression in terms of the number of rooted trees of size k. This observation may also be exploited to derive upper bounds for the constant. With the help of Proposition 4.3.11 lower bounds may be derived as well. However, this needs some further investigations to understand the shape of the polynomials $\tilde{e}_k(z)$ appearing in (4.3.4).

4.4 Counting Vertex-Labeled Tree-Child Networks

In this section, we will count (vertex-labeled) tree-child networks. As in the last section, we will first work out in detail the cases $k = 1, 2, 3$, where, as for normal networks, we will show more precise results than stated in Theorem 4.1.2. The general case (and thus the proof of Theorem 4.1.2) is then done in the last subsection below.

4.4.1 Tree-child networks with one reticulation vertex

We start with tree-child networks with one reticulation vertex which are again counted by using the Motzkin skeletons in Figure 4.3.

Proposition 4.4.1. *The exponential generating function for vertex-labeled tree-child networks with one reticulation vertex is*

$$T_1(z) = \frac{z^3\left(1-\sqrt{1-2z^2}\right)}{(1-2z^2)^{3/2}} = z\frac{\tilde{a}_1(z^2) - \tilde{b}_1(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{3/2}}, \tag{4.15}$$

where

$$\tilde{a}_1(z) = \tilde{b}_1(z) = z.$$

Proof. We have to add an edge from g in the Motzkin skeletons in Figure 4.3 which points to a unary (or red) vertex. Note that this edge is not allowed to point on a vertex on the path from g to the root (since the resulting network must be a DAG), but is allowed to point to any vertex on the subtrees attached to these vertices. Moreover, the edge can also point to any non-root vertex in the subtree attached to g (pointing on the root of this subtree is not allowed because we do not allow double edges).

This gives

$$T_1(z) = \frac{z}{2}\partial_y \frac{\tilde{M}(z,y)}{1-zM(z,y)}\bigg|_{y=0} = \frac{z}{2}\left(\frac{M_y(z,0) - z^2 - zM_b(z,0)}{1-zM(z,0)} + \frac{zM_y(z,0)M(z,0)}{(1-zM(z,0))^2}\right)$$

where $\tilde{M}(z,y)$ is given in (4.5). Similar to the normal network case, the factor $1/2$ compensates for the fact that each network is counted exactly twice by the above procedure. Now, by using (4.6) we obtain (4.15). □

From this, we obtain the following consequence.

Corollary 4.4.2. *Let $T_{1,n}$ denote the number of vertex-labeled tree-child networks with n vertices and one reticulation vertex. If n is even then $T_{1,n}$ is zero, otherwise*

$$T_{1,n} = n![z^n]T_1(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+1}\left(\frac{\sqrt{2}}{2} - \frac{\sqrt{\pi}}{2}\cdot\frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Remark. Note that the constant of the second order term in the asymptotic expansion above is $-\sqrt{\pi}/2$ whereas that of the asymptotic expansion of $N_{1,n}$ is $-3\sqrt{\pi}/2$. Thus, the difference between normal networks and tree-child networks becomes visible only in the second order term (and the number of normal networks is of course smaller than the number of tree-child networks). The behavior for $k = 2$ and $k = 3$ is similar; see below.

Relationship to unicyclic networks revisited.

Again there is a close relationship to unicyclic networks and the alternative approach from Section 4.3.1 can be used: either the root is in a cycle, but in which case now each vertex except the root can be the reticulation vertex, or the root is not in a cycle. This gives

$$T_1(z) = zM(z,0)T_1(z) + \frac{1}{2}\sum_{\ell\geq 2}\ell z^{\ell+1}M(z,0)^\ell.$$

Solving gives

$$T_1(z) = \frac{\sum_{\ell\geq 2}\ell z^{\ell+1}M(z,0)^\ell}{2(1-zM(z,0))} = \frac{z^3M(z,0)^2(2-zM(z,0))}{2(1-zM(z,0))^3}.$$

which by using the expression (4.5) for $M(z,0)$ simplifies to (4.15).

4.4.2 Tree-child networks with two reticulation vertices

As for normal networks, the counting is done by using two variables y_1 and y_2 and the two types of Motzkin skeletons depicted in Figure 4.5.

For trees attached to paths the situation is different from normal networks. We never encounter different pointing rules between roots and internal vertices, but very well between vertices on the path and vertices within the trees. Thus the red vertices in the third and fourth term on the right-hand side of the specification for $\mathcal{Q}$, see (4.7), fall into different classes of red vertices. In the third term, $\{\circ\} \times \mathcal{Q} \times$

$(\{\bullet\} \times \tilde{\mathcal{M}})$, the red vertex is the root of the attached (red) tree. It can be treated like the red vertices within the tree and therfore we mark it with $\tilde{y}$. A consequence of this is that we do not need to distinguish between red and white trees any more. Indeed, the second term of the specification corresponds to having a white tree attached, the third one to having a red tree attached (to the path, in both cases). Since the red vertices fall into the same class and are both marked by $\tilde{y}$, we may replace these two terms by one term corresponding to attaching simply a Motzkin tree. The red vertex in the last term of (4.7) is on the path itself, thus marked by y. The other subtree cannot be a red tree by the tree-child condition.

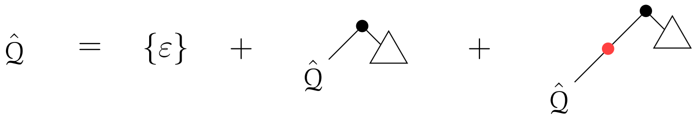


Figure 4.14: The specification of the class $\hat{\mathcal{Q}}$ which is similar to that of $\mathcal{Q}$ (*cf.* Figure 4.4) but with the second and third term merged. Also, now the subtree of the second term can be either red or white and that of the third term must be white. All the red vertices in these subtrees are counted by $\tilde{y}$; the other red vertices arising from the third term are counted by y.

Altogether, this modification leads to a new class $\hat{\mathcal{Q}}$, specified by

$$\hat{\mathcal{Q}} = \{\varepsilon\} \cup \{\circ\} \times \hat{\mathcal{Q}} \times \mathcal{M} \cup \{\circ\} \times (\{\bullet\} \times \hat{\mathcal{Q}}) \times \tilde{\mathcal{M}},$$

see Figure 4.14. We use this new class in (4.8) instead of $\mathcal{Q}$ to specify the paths forming the basic building block for the Motzkin skeletons of tree-child networks. Call this new structure $\hat{\mathcal{P}}$. Then, we obtain the generating function

$$\hat{P}(z, y, \tilde{y}, \hat{y}) = \frac{1 + z\hat{y}}{1 - zM(z, \tilde{y}) - z^2 y\tilde{M}(z, \tilde{y})}.$$

To summarize: The variable y tells us which green vertex is allowed to point to vertices of the path (with the first vertex as possible exception), $\tilde{y}$ which may point to vertices in the trees attached to the path, and $\hat{y}$ which may point to the first vertex of the path. We also make explicit a frequently appearing function:

$$\hat{P}(z, 0, \tilde{y}, 0) = \frac{1}{1 - zM(z, \tilde{y})}.$$

Now, the result for tree-child networks with two reticulation vertices is as follows.

Proposition 4.4.3. *The exponential generating function for vertex-labeled tree-child networks with two reticulation vertices is*

$$T_2(z) = z\frac{\tilde{a}_2(z^2) - \tilde{b}_2(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{7/2}}, \tag{4.16}$$

where

$$\tilde{a}_2(z) = -z^4 + 8z^3 \qquad \textit{and} \qquad \tilde{b}_2(z) = 8z^3$$

Proof. We start with the tree-child networks arising from the Motzkin skeletons on the left in Figure 4.5. Here, g_1 and g_2 can point to all vertices in the attached subtrees except the root of the subtree attached to g_1. In addition, g_2 can also point to all vertices on the path between g_1 and g_2 except the vertex directly followed by g_2.

Overall, we obtain

$$\begin{aligned} T_{2,1}(z) &= \partial_{y_1}\partial_{y_2} z^2 \tilde{M}(z, y_1+y_2)\hat{P}(z, y_2, y_1+y_2, 0)\hat{P}(z, 0, y_1+y_2, 0)\Big|_{y_1=0, y_2=0} \\ &= \partial_{y_1}\partial_{y_2} \frac{z^2\tilde{M}(z, y_1+y_2)}{(1-zM(z,y_1+y_2))(1-(z+z^2y_2)M(z,y_1+y_2)}\Big|_{y_1=0,y_2=0}. \end{aligned}$$

Now, consider the Motzkin skeletons on the right of Figure 4.5. For the trees attached to the green vertices only pointing to the root is forbidden, for all the other trees there is no pointing restriction. Note that in this case no green vertex is allowed to point to the vertices on the paths. This leads to the first line of following generating function. In this way, the Motzkin skeleton which is depicted in Figure 4.7 gives second the term.

$$\begin{aligned} T_{2,2}(z) =& \frac{1}{2}\partial_{y_1}\partial_{y_2}\frac{z^3\tilde{M}(z,y_1+y_2)^2}{1-zM(z,y_1+y_2)}\hat{P}(z,0,y_1+y_2,0)\hat{P}(z,0,y_1+y_2,0)\Big|_{y_1=0,y_2=0} \\ &+ \partial_{y_1}\partial_{y_2}\frac{z^3\tilde{M}(z,y_2)^2}{1-zM(z,y_2)}\hat{P}(z,y_1,y_2,y_1)\hat{P}(z,0,y_2,0)\Big|_{y_1=0,y_2=0} \end{aligned}$$

The exponential generating function for vertex-labeled tree-child networks is now obtained as $T_2(z) = (T_{2,1}(z) + T_{2,2}(z))/4$. Plugging in the above expressions and simplifying gives the result. □

As a consequence, we have the following result.

Corollary 4.4.4. *Let $T_{2,n}$ denote the number of vertex-labeled tree-child networks with n vertices and two reticulation vertices. If n is even then $T_{2,n}$ is zero, otherwise*

$$T_{2,n} = n![z^n]T_2(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+3}\left(\frac{\sqrt{2}}{16} - \frac{\sqrt{\pi}}{8}\cdot\frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

4.4.3 Tree-child networks with three reticulation vertices

In this case we use the four different types of Motzkin skeletons depicted in Figure 4.8. Moreover, we use the Y operator from Section 4.3.3.

We start with Case 1 the tree-child networks arising from the Motzkin skeletons depicted on the top left of Figure 4.8. The possibilities for the pointings of the edges starting at g_1, g_2 and g_3 are similar as in the first case for $k = 2$ (see above). All these edges may target any non-root vertex in the tree attached to g_1 and any vertex in all the other trees. Concerning the vertices on the spine, we have some restrictions. The edge from g_1 may not end at any vertex from ℓ_1, for the first vertex this applies even to g_2. Similarly, the edges from g_1 and g_2 may not point to any vertex of ℓ_2, and no green vertex may point to the first vertex of ℓ_2 as well as to any vertex of ℓ_1. Note that for tree-child netwotks we do not care about shortcuts cases any more.

Overall, we obtain for this Motzkin skeleton

$$T_{3,1}(z) = \mathbf{Y}_{1,2,3}\left(\frac{z^3\tilde{M}(z, y_1+y_2+y_3)}{1-zM(z, y_1+y_2+y_3)}\hat{P}(z, y_3, y_1+y_2+y_3, 0)\right.$$
$$\left.\times\hat{P}(z, y_2+y_3, y_1+y_2+y_3, 0)\right).$$

For the other cases, a similar reasoning for the possible pointings of the edges starting from g_1, g_2 and g_3 can be used. Furthermore, we have to pay attention to the Motzkin skeletons we generate which are not tree-child. These cases are those where two green vertices point to the children of a latest common ancestor of two green vertices, but the third green vertex has some freedom in pointing (see Figure 4.15 (i)). We refrain from giving details and just list the obtained expressions. The reader is invited to derive them herself.

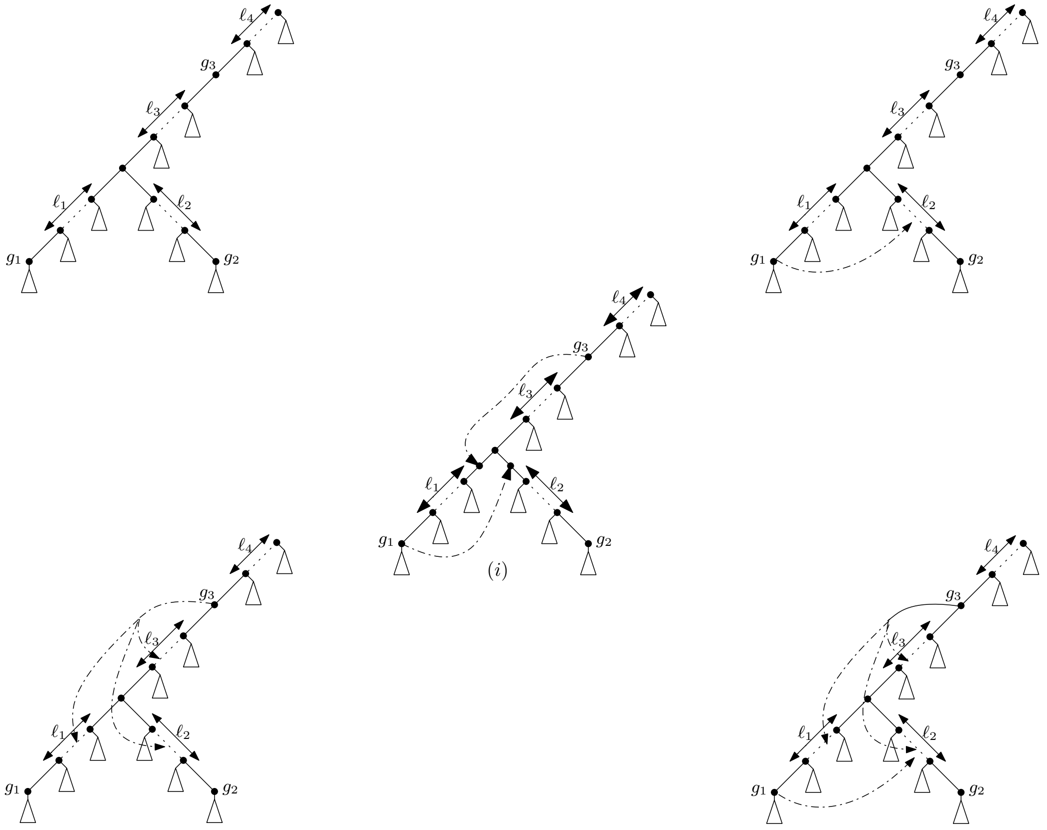


Figure 4.15: The subtraction terms comes from (i), where two green vertices point to the children of a latest common ancestor of two green vertices.

$$\begin{aligned}
T_{3,2}(z) =& \frac{1}{2}\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_1+y_2+y_3)^2}{1-zM(z,y_1+y_2+y_3)}\hat{P}(z,0,y_1+y_2+y_3,0)^3\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_2+y_3)^2}{1-zM(z,y_2+y_3)}\hat{P}(z,y_1,y_2+y_3,y_1)\hat{P}(z,0,y_2+y_3,0)^2\right)\\
&+\frac{1}{2}\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_1+y_2)^2}{1-zM(z,y_1+y_2)}\hat{P}(z,y_3,y_1+y_2,y_3)^2\hat{P}(z,y_3,y_1+y_2,0)\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_2)^2}{1-zM(z,y_2)}\hat{P}(z,y_1+y_3,y_2,y_1+y_3)\hat{P}(z,y_3,y_2,y_3)\hat{P}(z,y_3,y_2,0)\right)\\
&-\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_2)^2}{(1-zM(z,y_2))^2}\hat{P}(z,0,y_2,y_1)\hat{P}(z,0,y_2,y_3)\right).
\end{aligned}$$

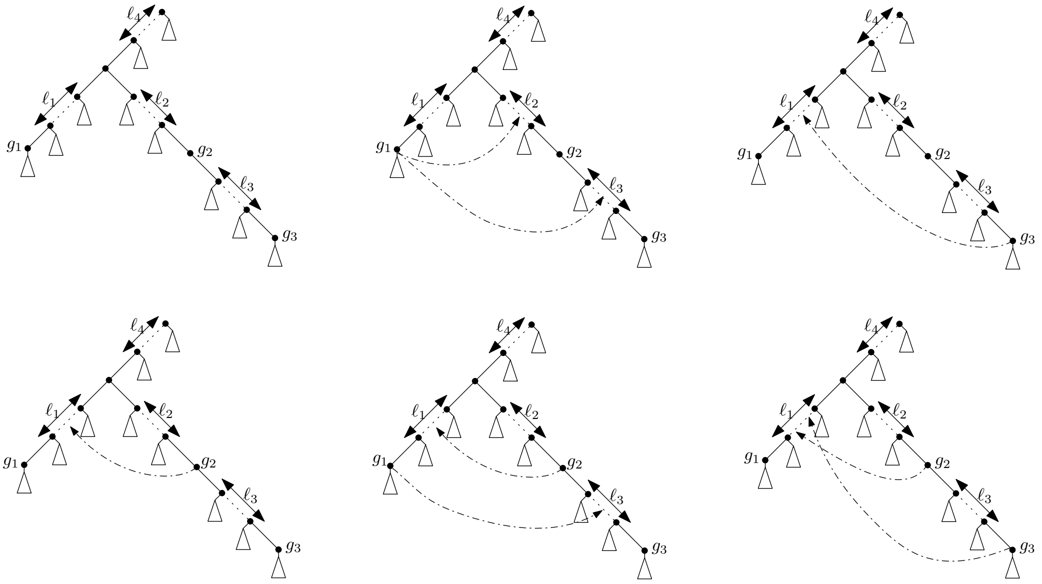


Figure 4.16: All Motzkin tree skeletons of case 3 with possible distribution of reticulation vertices on paths. Recall that, Missing pointers of green vertices are not allowed to point to a path.

For the Motzkin skeletons Case 3 which are depicted in Figure 4.16, we obtain

$$
\begin{aligned}
T_{3,3}(z) =&\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_1+y_2+y_3)^2}{1-zM(z,y_1+y_2+y_3)}\hat{P}(z,0,y_1+y_2+y_3,0)^3\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_2+y_3)^2}{1-zM(z,y_2+y_3)}\hat{P}(z,y_1,y_2+y_3,y_1)\hat{P}(z,y_1,y_2+y_3,0)\hat{P}(z,0,y_2+y_3,0)\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_1+y_2)^2}{1-zM(z,y_1+y_2)}\hat{P}(z,y_3,y_1+y_2,y_3)\hat{P}(z,0,y_1+y_2,0)^2\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_1+y_3)^2}{1-zM(z,y_1+y_3)}\hat{P}(z,y_2,y_1+y_3,y_2)\hat{P}(z,0,y_1+y_3,0)^2\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_3)^2}{1-zM(z,y_3)}\hat{P}(z,y_2,y_3,y_2)\hat{P}(z,y_1,y_3,0)\hat{P}(z,0,y_3,0)\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}(z,y_1)^2}{1-zM(z,y_1)}\hat{P}(z,y_2+y_3,y_1,y_2+y_3)\hat{P}(z,0,y_1,0)^2\right).
\end{aligned}
$$

For the final case, consider the Motzkin skeletons depicted in Figure 4.17. Note that ignoring shortcut cases unable us to merge some of the cases in Motzkin tree skeleton that we have already considered for Normal networks. Here, the generating function is given by

$$\begin{aligned}
T_{3,4}(z) = \frac{1}{2}\mathbf{Y}_{1,2,3}&\left(\frac{z^5\tilde{M}(z,y_1+y_2+y_3)^3}{1-zM(z,y_1+y_2+y_3)}P(z,0,y_1+y_2+y_3,0)^4\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_2+y_3)^3}{(1-zM(z,y_2+y_3))^3}P(z,y_1,y_2+y_3,y_1)^2\right)\\
&+\frac{1}{2}\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_1+y_2)^3}{(1-zM(z,y_1+y_2))^2}P(z,y_3,y_1+y_2,y_3)^3\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_3)^3}{(1-zM(z,y_3))^3}P(z,y_1,y_3,y_1)P(z,y_2,y_3,y_2)\right)\\
&+\frac{1}{2}\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_3)^3}{(1-zM(z,y_3))^4}P(z,y_1+y_2,y_3,y_1+y_2)\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_2)^3}{(1-zM(z,y_2))^2}P(z,y_1+y_3,y_2,y_1+y_3)P(z,y_3,y_2,y_3)^2\right)\\
&+\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_2)^3}{(1-zM(z,y_2))^3}P(z,y_1,y_2,y_1)P(z,y_3,y_2,y_3)\right)\\
&-\mathbf{Y}_{1,2,3}\left(\frac{z^5\tilde{M}(z,y_2)^3}{(1-zM(z,y_2))^3}P(z,0,y_2,y_1)P(z,0,y_2,y_3)\right).
\end{aligned}$$

The exponential generating function for vertex-labeled tree-child networks is obtained as $T_3(z) = (T_{3,1}(z) + T_{3,2}(z) + T_{3,3}(z) + T_{3,4}(z))/8$ after all. This gives the following result.

Proposition 4.4.5. *The exponential generating function for vertex-labeled tree-child networks with three reticulation vertices is*

$$T_3(z) = z\frac{\tilde{a}_3(z^2) - \tilde{b}_3(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{11/2}}, \tag{4.17}$$

where

$$\tilde{a}_3(z) = -35z^6 + 175z^5 \qquad \textit{and} \qquad \tilde{b}_3(z) = 34z^6 + 175z^5.$$

As a consequence, we have the following result.

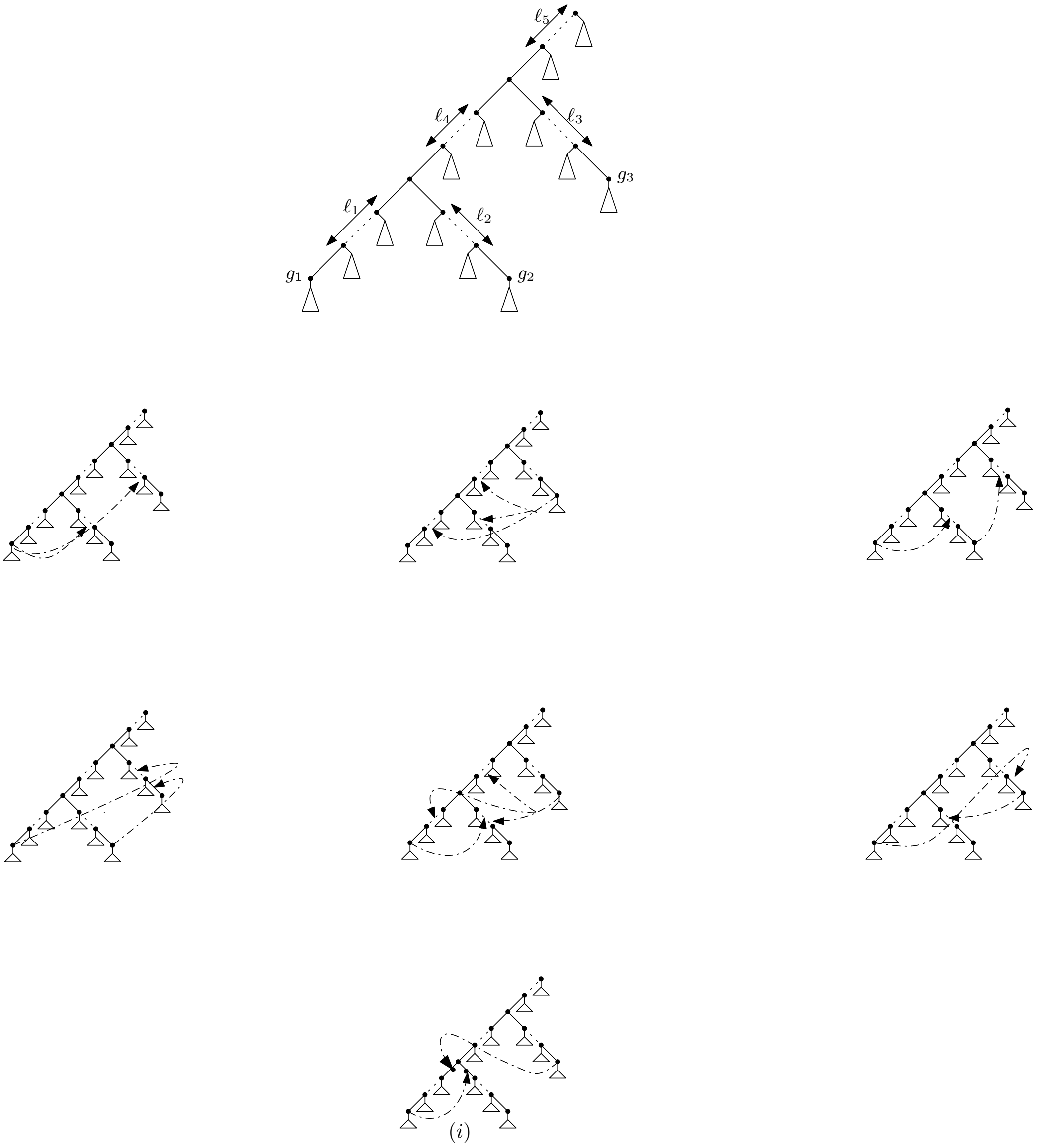


Figure 4.17: Motzkin skeletons of case 4 with all possible pointing of green vertices to red vertices which some of them may lie on the paths as well. Note that, we don't care shortcut structures in tree-child networks anymore. Thus ignoring this restriction causes the less numbers of Motzkin skeletons compare to normal networks with same reticulation vertices.

Corollary 4.4.6. *Let $T_{3,n}$ denote the number of vertex-labeled tree-child networks with n vertices and three reticulation vertices. If n is even then $T_{3,n}$ is zero, otherwise*

$$T_{3,n} = n![z^n]T_3(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+5}\left(\frac{\sqrt{2}}{192} - \frac{\sqrt{\pi}}{64}\cdot\frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

4.4.4 Tree-child networks with a fixed number of reticulation vertices

In this subsection, we will prove Theorem 4.1.2 which is deduced from the following proposition.

Proposition 4.4.7. *For the numbers of vertex-labeled normal networks $N_{k,n}$ and vertex-labeled tree-child networks $T_{k,n}$,*

$$T_{k,n} = N_{k,n}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right), \text{ as } n \to \infty.$$

Proof. First, observe that $T_{k,n} - N_{k,n}$ is bounded by the number of networks which arise from all types of Motzkin skeletons where for each green vertex we consider all possibilities of adding an edge such that the normal condition is violated (note that this is an over-estimate of the difference). Thus, we only have to count the number of such networks which arise from a fixed type of Motzkin skeletons and a fixed green vertex. Similar to the proof of Proposition 4.3.8, the largest number will come from the Motzkin skeletons where the green vertices are the leaves (this will become clear by applying the same arguments as below to all other Motzkin skeletons).

Now, fix such a type of Motzkin skeletons and one of its green vertices. Then, for this vertex, we will have the following options.

- The green vertex points to one of the subtrees attached to the leaves of the skeletons. For the exponential generating function this gives

$$\partial_{y_2}\cdots\partial_{y_k}\frac{G_1'(z,y)\cdots G_s(z,y)}{(1-G_{s+1}(z,y))\cdots(1-G_{s+2k-1}(z,y))}\Big|_{y_2=0,\ldots,y_k=0},$$

  where the derivative comes from choosing a vertex in the subtree as end point of the green vertex. (Here, and below y is the sum of y_i's with $2 \le i \le k$ and not all of the y_i's must be present; also which are present can differ from one occurrence to the next.)

- The green vertex points to the root of a Motzkin tree from $\mathcal{M}$ attached to the path from the green vertex to the root or attached to some of the edges of the sparsened skeleton on a path from the green vertex to a leaf. Then, we have

  $$\partial_{y_2}\cdots\partial_{y_k}\frac{G_1(z,y)\cdots G_s(z,y)}{(1-G_{s+1}(z,y))\cdots(1-G_{s+2k-1}(z,y))(1-G_{s+2k}(z,y))}\Big|_{y_2=0,\dots,y_k=0},$$

  where the additional term comes from the fact that now one edge was split into two edges by the above pointing.

- The green vertex attaches to a red vertex on the path which is the ancestor of some other green vertices and may one of them is pointed to the root of a Motzkin tree from $\mathcal{M}$ attached to the paths from the first green vertices to the root of sparsened skeleton.

  $$\partial_{y_3}\cdots\partial_{y_k}\frac{G_1(z,y)\cdots G_s(z,y)}{(1-G_{s+1}(z,y))\cdots(1-G_{s+2k}(z,y))(1-G_{s+2k+1}(z,y))}\Big|_{y_3=0,\dots,y_k=0},$$

- The green vertex points to the first vertex on one of the branches attached to the path from the green vertex to the root. Then, we have

  $$\partial_{y_2}\cdots\partial_{y_k}\frac{G_1(z,y)\cdots G_s(z,y)}{(1-G_{s+1}(z,y))\cdots(1-G_{s+2k-1}(z,y))}\Big|_{y_2=0,\dots,y_k=0}.$$

The exponential generating function of all networks arising from these Motzkin skeletons and the green vertex are a sum of generating functions of the above three types. Thus, from Lemma 4.3.7, we obtain that this generating function has the form

$$\frac{c(z)-d(z)\sqrt{1-2z^2}}{(1-2z^2)^p},$$

where $c(z)$ and $d(z)$ are suitable polynomials and the maximum of p is as follows: note that without the derivatives in the above expressions, p would be at most k (this bound is taken on in the first two cases, but not in the last case where p is at most $k-1/2$); also, because of Lemma 4.3.7, each derivative increases this bound by one. Thus, p is at most $2k-1$.

Now, with the same arguments as in the proof of Corollary 4.3.9, we obtain that the exponential generating function of the above number has the form

$$z\frac{\tilde{c}(z^2)-\tilde{d}(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{2k-1}},$$

where $\tilde{c}(z)$ and $\tilde{d}(z)$ are suitable polynomials. Singularity analysis gives then the bound

$$\mathcal{O}\left(\left(\frac{\sqrt{2}}{e}\right)^{n} n^{n+2k-3/2}\right).$$

Summing over all possible type of Motzkin skeletons and all green vertices, we obtain the same bound for $T_{k,n} - N_{k,n}$ which proves the claimed result. □

4.5 Explicit Formulas for the Number of Tree-Child and Normal Networks with $k = 1, 2, 3$.

In this section, we will count leaf-labeled normal and tree-child networks with ℓ leaves and k reticulation vertices (recall that we denoted their numbers by $\tilde{N}_{k,\ell}$ and $\tilde{T}_{k,\ell}$, respectively). The counting results will follow from those for vertex-labeled networks since there is a close relationship between leaf-labeled normal and tree-child networks and vertex-labeled ones. To see this, first recall the equation 1.1 that shows for any phylogenetic network with ℓ leaves, k reticulation vertices and n vertices, we have (n is always odd)

$$\ell + k = \frac{n+1}{2}.$$

Also we need the following lemma.

Lemma 4.5.1 (see [7]). *The descendant sets for any two non-leaf vertices in a tree-child network (and thus also normal network) are different.*

Proof. For each vertex v, let $D(v)$ denote the set of strict descendants of v, that is, the set of vertices other than v that can be reached by a directed path from v. For two distinct vertices v and v', if $v \in D(v')$, then $v' \notin D(v)$ since the network is acyclic, so $D(v) \neq D(v')$. Suppose $v \notin D(v')$, and consider a leaf ℓ which can be reached from v by a path of tree edges. Then we must have $\ell \in D(v)\ D(v')$, and so again $D(v) \neq D(v')$. □

These two lemmas immediately imply that

$$N_{k,2\ell+2k-1} = \binom{2\ell+2k-1}{\ell}(\ell+2k-1)!\tilde{N}_{k,\ell}.$$

To see this, note that all vertex-labeled normal networks with $2\ell + 2k - 1$ vertices and k reticulation vertices can be constructed as follows: start with a (fixed) leaf-labeled normal network with ℓ leaves and k reticulation vertices. Then, choose ℓ labels from the set $2\ell+2k-1$ labels and re-label the leaves of the fixed network such that the order is preserved. Finally, label the remaining $\ell+2k-1$ vertices by any permutation of the set of remaining $\ell+2k-1$ labels. By the above two lemmas, in this way every vertex-labeled normal network is obtained exactly once.

The above now implies that

$$\tilde{N}_{k,\ell} = \frac{\ell!}{(2\ell+2k-1)!} N_{k,2\ell+2k-1}$$

from which an asymptotic result follows by Theorem 4.1.1 and Stirling's formula. Similarly, an asymptotic result for leaf-labeled tree-child networks is obtained from

$$\tilde{T}_{k,\ell} = \frac{\ell!}{(2\ell+2k-1)!} T_{k,2\ell+2k-1}.$$

Overall, we obtain the following theorem.

Theorem 4.5.2. *For the numbers $\tilde{N}_{k,\ell}$ and $\tilde{T}_{k,\ell}$ of leaf-labeled normal networks resp. leaf-labeled tree-child networks with $k \geq 1$ reticulation vertices, we have*

$$\tilde{N}_{k,\ell} \sim \tilde{T}_{k,\ell} \sim 2^{3k-1} c_k \left(\frac{2}{e}\right)^{\ell} \ell^{\ell+2k-1}, \qquad (\ell \to \infty)$$

where c_k is as in Theorem 4.1.1.

Examples

Now we are ready to present explicit formulas for vertex and leaf-labeled of normal and tree child networks up to 3 reticulation vertices. Start with normal networks with one reticulation vertex. First, set $n = 2m+1$. Then, from (4.4) we obtain

$$[z^n]N_1(z) = [z^m]\bar{N}_1(z)$$

with

$$\bar{N}_1(z) = \frac{a_1(z) - b_1(z)\sqrt{1-2z}}{(1-2z)^{3/2}},$$

where $a_1(z)$ and $b_1(z)$ are as in (4.4). So this implies

$$[z^m]\bar{N}_1(z) = [z^m]\frac{a_1(z)}{(1-2z)^{\frac{3}{2}}} - [z^m]\frac{b_1(z)}{(1-2z)}.$$

After some simple computation we have

$$[z^m]\bar{N}_1(z) = 2^m\big((m+2)\frac{\binom{2m}{m}}{4^m} - \frac{3}{2}\big).$$

By replacing $m = (n-1)/2$ we get

$$N_{1,n} = n![z^n]N_1(z) = n!2^{(n-1)/2}\big((n+3)\frac{\binom{n-1}{(n-1)/2}}{2^n} - \frac{3}{2}\big). \tag{4.18}$$

By considering the $\tilde{N}_{1,\ell} = \frac{\ell!}{n!}N_{1,n}$, and lemma 1.0.4 we have following explicit formula for leaf-labeled networks with one reticulation vertex.

$$\tilde{N}_{1,\ell} = \ell!2^\ell\big((\ell+2)\frac{\binom{2\ell}{\ell}}{4^\ell} - \frac{3}{2}\big). \tag{4.19}$$

In the same way, from 5.3.1 we get

$$[z^m]\bar{N}_2(z) = [z^m]\frac{a_2(z)}{(1-2z)^{\frac{7}{2}}} - [z^m]\frac{b_2(z)}{(1-2z)^3}.$$

Suppose the coefficient extractor function , $\mathcal{F}_i(m) := [z^m]\bar{N}_i(z)$ Computing the mth coefficient of the power series representing a given generating function

$$\begin{aligned}\mathcal{F}_2(m) := [z^m]\bar{N}_2(z) = [z^m]\frac{11z^4 - 66z^3 + 50z^2 - 8z}{(1-2z)^{\frac{7}{2}}} - [z^m]\frac{-28z^3 + 42z^2 - 8z}{(1-2z)^3} \\ = \frac{1}{3}\ 2^{m-1}(3m-7)\Big(2m(m^2+9m-4)\frac{\binom{2m}{m}}{(2m-1)4^m} - 3(m+1)\Big).\end{aligned}$$

Then, $N_{2,n} = n!\mathcal{F}_2((n-1)/2)$, that leads to an expansion of the following form

$$N_{2,n} = n!\ \frac{2^{(n-3)/2}}{3}(3n-17)\Big((n-1)(n^2+16n-33)\frac{\binom{n-1}{(n-1)/2}}{(n-2)2^n} - 3(n+1)\Big). \tag{4.20}$$

By replacing $m = l+1$, we have following expression for leaf-labeled normal networks with two reticulation vertices.

$$\tilde{N}_{2,\ell} = \ell!\ \frac{2^\ell}{3}\ (3\ell-4)\Big(2(\ell+1)(\ell^2+11\ell+6)\frac{\binom{2\ell+2}{\ell+1}}{(2\ell+1)\ 4^{\ell+1}} - 3(\ell+2)\Big). \tag{4.21}$$

As similar as before for normal networks with 3 reticulation vertices (see 4.11) we obtain the following results.

$$\mathcal{F}_3(m) := [z^m]\bar{N}_3(z) = \frac{1}{3}\, 2^{m-6}\Big(m(m-1)\frac{\binom{2m}{m}}{(2m-1)4^{m-3}}A(m) - B(m)\Big). \tag{4.22}$$

where

$$\begin{aligned} A(m) &= m^4 + 15m^3 - 158m^2 + 324m + 40, \\ B(m) &= 144m^4 - 751m^3 - 1089m^2 - 9106m - 7080. \end{aligned}$$

Immediately we get $N_{3,n} = n!\mathcal{F}_3((n-1)/2)$ and $\tilde{N}_{3,n} = \ell!\mathcal{F}_3(\ell+2)$. This leads following formulas for normal network with 3 reticulation vertices in vertex and leaf-labeled cases respectively.

$$\tilde{N}_{3,n} = n!\, \frac{2^{(n-13)/2}}{3}\Big((n-1)(n-3)\frac{\binom{n-1}{(n-1)/2}}{(n-2)2^{n-5}}A((n-1)/2) - B((n-1)/2)\Big) \tag{4.23}$$

$$\tilde{N}_{3,\ell} = \ell!\, \frac{2^{\ell-4}}{3}\Big((\ell+1)(\ell+2)\frac{\binom{2\ell+4}{\ell+2}}{(2\ell+3)4^{\ell-1}}A(\ell+2) - B(\ell+2)\Big). \tag{4.24}$$

Let us recall the equation 4.15, the exponential generating function for vertex-labeled tree-child networks with one reticulation vertex. Therefore, we have

$$[z^n]T_1(z) = [z^m]\bar{T}_1(z)$$

with

$$\bar{T}_1(z) = \frac{\tilde{a}_1(z) - \tilde{b}_1(z)\sqrt{1-2z}}{(1-2z)^{3/2}},$$

where $\tilde{a}(z)$ and $\tilde{b}(z)$ are as in (4.15). Hence, we obtain

$$[z^m]\bar{T}_1(z) = [z^m]\frac{z}{(1-2z)^{\frac{3}{2}}} - [z^m]\frac{z}{(1-2z)}.$$

which leads to

$$[z^m]\bar{T}_1(z) = 2^m\big(m\frac{\binom{2m}{m}}{4^m} - \frac{1}{2}\big).$$

Then for $i = 1, 2, 3$ the function $\hat{\mathfrak{F}}_i(m) = [z^m]\bar{T}_i(z)$ satisfies

$$T_{1,n} = n!\hat{\mathfrak{F}}_1((n-1)/2) \qquad \text{and} \qquad \tilde{T}_{1,\ell} = \ell!\hat{\mathfrak{F}}_1(\ell).$$

which gives us explicit formula

$$T_{1,n} = n!\ 2^{(n-1)/2}\ \Big((n-1)\frac{\binom{n-1}{(n-1)/2}}{2^n} - \frac{1}{2}\Big), \tag{4.25}$$

for vertex-labeled tree-child network with one reticulation vertex and for leaf-labeled we have

$$\tilde{T}_{1,\ell} = \ell!\ 2^\ell\ \big(\ell\frac{\binom{2\ell}{\ell}}{4^\ell} - \frac{1}{2}\big). \tag{4.26}$$

By the same argument, the exponential generating function for vertex-labeled tree-child networks with two reticulation vertex leads to

$$\begin{aligned}\hat{\mathfrak{F}}_2(m) := [z^m]\bar{T}_2(z) &= [z^m]\frac{-z^4+8z^3}{(1-2z)^{\frac{7}{2}}} - [z^m]\frac{8z^3}{(1-2z)^3}\\ &= 2^{m-1}(m-1)(m-2)\Big(2m(3m-1)\frac{\binom{2m}{m}}{3(2m-1)4^m} - 1\Big).\end{aligned}$$

Then

$$T_{2,n} = n!\hat{\mathfrak{F}}_2((n-1)/2) = n!2^{(n-7)/2}(n-3)(n-5)\Big((n-1)(3n-5)\frac{\binom{2(n-1)}{(n-1)/2}}{3(n-2)2^n} - 1\Big). \tag{4.27}$$

By replacing $m = l+1$, we have following expression for leaf-labeled normal networks with two reticulation vertices.

$$\tilde{T}_{2,\ell} = \ell!\ \ell(\ell-1)2^\ell\Big(2(\ell+1)(3\ell+2)\frac{\binom{2\ell+2}{\ell+1}}{3(2\ell+1)\cdot 4^{\ell+1}} - 1\Big). \tag{4.28}$$

Finally, we derive explicit formulas for leaf-labeled and vertex-labeled tree-child network with 3 reticulation vertices. To do that first consider that

$$\begin{aligned}\hat{\mathfrak{F}}_3(m) := [z^m]\bar{T}_3(z) &= [z^m]\frac{-35z^6+175z^5}{(1-2z)^{\frac{11}{2}}} - [z^m]\frac{34z^6+175z^5}{(1-2z)^5}\\ &= \frac{1}{3}2^{m-6}\tilde{P}(m)\Big(m^2(m-1)\frac{\binom{2m}{m}}{(2m-1)4^{m-3}} - (48m-65)\Big).\end{aligned}$$

where

$$\tilde{P}(m) = (m-4)(m-3)(m-2).$$

So for vertex-labeled case we have

$$T_{3,n} = n!\,\frac{2^{(n-13)/2}}{3}\tilde{P}((n-1)/2)\Big((n-1)^2(n-3)\frac{\binom{n-1}{(n-1)/2}}{(n-2)2^{n-4}} - (24n-89)\Big). \tag{4.29}$$

Finally, we have an equation of the form

$$\tilde{T}_{3,\ell} = \ell!\,\frac{2^{\ell-4}}{3}\tilde{P}(\ell+2)\Big((\ell+2)^2(\ell+1)\frac{\binom{2\ell+4}{\ell+2}}{(2\ell+3)4^{\ell-1}} - (48\ell+31)\Big). \tag{4.30}$$

for leaf-labeled of tree-child networks.

Chapter 5

Counting General Phylogenetic Networks

A family of general phylogenetic networks is a set of all subclasses of binary rooted networks. From the enumeration point of view, it would be interesting under some assumption (fixed number of reticulation vertices) we can find out the connection between studied subsets, tree-child and normal networks, with the whole set that is included them as well. So, this chapter aims to extend the approach of the previous chapter to general networks. For this some further work has to be done. First, for vertex-labeled networks our method for normal and tree child networks relied on the use of Motzkin skeletons, which have green and red vertices, and all of them are unary vertices. Recall that these vertices arise from deleting an edge for each reticulation vertex which was colored red (the green vertices are then the other endpoints of the deleted edges). However, if one considers general phylogenetic networks, then the colored vertices in the Motzkin skeleton can be leaves as well. See Figure 5.1 cases (2) and (3). This Figure shows that in the network (1), if the indicated edges are deleted, we have red and green colored vertices as like tree-child classes networks that we saw in Chapter 4. For middle one, after deleting indicated edges then so-called vertex r_g becomes a leaf (which is colored both red and green). On the other hand, for the network depicted on the right side, after removing the marked edges, the vertex which was connected to r_2 and r_3 becomes a leaf (which is called *double-green*). So, in order to consider the counting problem for general networks, more possibilities for the Motzkin skeletons must be considered which are the combination of mentioned above situations. This chapter will aim to take some steps towards this by exploring some details. The next section details its design and compilation.

As before we want to show that, by using analytic methods, we can obtain precise asymptotic estimates for the number of general phylogenetic networks. Note that variations on the definition of general phylogenetic networks are around in the literature. In general phylogenetic networks, as defined before, multiple edges are not explicitly forbidden. Our goal is indeed to study the most general model of general phylogenetic networks that could be counted if their number of reticulation vertices is fixed and provide a more detailed investigation regarding enumeration properties of general networks with or without multiple edges on their structures. Now, denote by $G_{k,n}$ resp. $\tilde{G}_{k,\ell}$ the number of general networks with k reticulation vertices in the vertex-labeled (leaf-labeled) case. We focus mainly on proves the following results.

Theorem 5.0.3. *For the number $G_{k,n}$ of vertex-labeled phylogenetic networks with $k \geq 1$ reticulation vertices, there is a positive constant d_k such that*

$$G_{k,n} \sim d_k \left(1-(-1)^n\right)\left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1}, \qquad (n \to \infty).$$

In particular,

$$d_1 = \frac{\sqrt{2}}{4}; \qquad d_2 = \frac{\sqrt{2}}{32}; \qquad d_3 = \frac{\sqrt{2}}{384}.$$

The result reveals that the first and second order asymptotics are the same as vertex-labeled tree-child networks. In other words, we can show that for the general networks with fixed number of reticulation vertices, the additional networks that not satisfying the tree-child conditions are asymptotically negligible as $n \to \infty$. Another goal will be to show how this approach help us to have following result as well.

Theorem 5.0.4. *For the numbers $\tilde{G}_{k,\ell}$ of leaf-labeled general networks with $k \geq 1$ reticulation vertices, we have*

$$\tilde{G}_{k,\ell} \sim 2^{3k-1} d_k \left(\frac{2}{e}\right)^\ell \ell^{\ell+2k-1}, \qquad (\ell \to \infty)$$

where d_k is as in Theorem 5.0.3.

Remark. Note that this result only holds for fixed k as n goes to infinity. The case when k approaches to infinity cannot be done in this way.

5.1 Decomposing general phylogenetic networks

In order to count general phylogenetic networks, we will adjust the procedure of sparsened skeleton decomposition for general networks. This method is well studied for tree-child networks, with k of reticulation vertices in Chapter 4. Similar as before, we use the decomposition to obtain a reduction which can be easily analyzed by means of generating functions. Consider a general phylogenetic network having k reticulation vertices. Then each such vertex has two incoming edges. If one edge is removed for each of the k reticulation vertices, then the remaining graph is again Motzkin tree (labeled and nonplane). Depending on our choice for removed edges, this Motzkin tree has at most $2k$ unary nodes. Recall that for tree-child networks this method gives exactly $2k$ unary nodes.

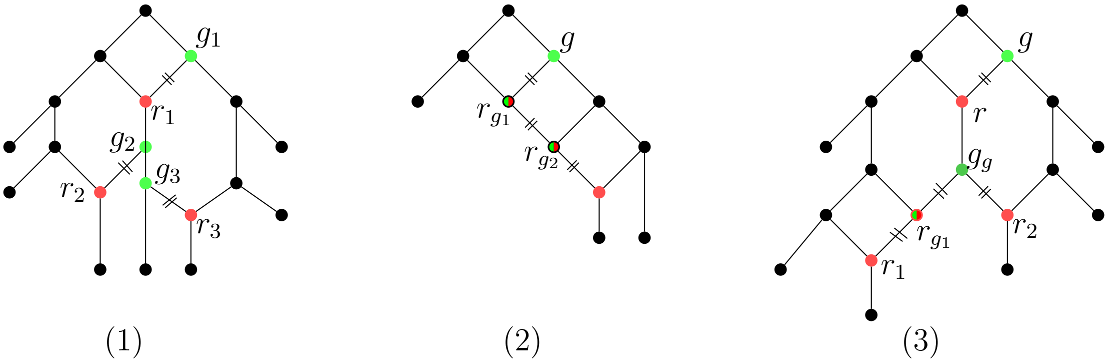


Figure 5.1: Three general phylogenetic networks with colored Motzkin skeletons where, (1) after adding marked edges there is bijection between green and red (reticulation) vertices in the Motzkin skeleton. General networks depicted in (2) and (3) where the red-green and double-green vertices appear after adding the marked edges. Edges are directed downwards.

Now consider the following procedure: start with a Motzkin tree T with not more that $2k$ unary nodes and n vertices in total.

- Add directed edges such that each edge connects two unary nodes and any two edges do not have a vertex in common. Color the started vertices of the added directed edges green and their end vertices red. Note that if motzkin tree has exactly $2k$ unary nodes, the coloring procedure imposes that there will be equal k colored green and red nodes (see Figure 5.1, (1)).

- Consider two unary vertices and joint them by using a sequence ($\geqslant 2$) of fixed number of directed edges in the following way. One of edge in the sequence connects the first unary vertex of the Motzin tree to a leaf which we call r_g. Then connect r_g to another leaf and continue this process for disjoint leaves until use all directed edges but one. Now connect the last leaf to a second unary vertex by the remaining edge. As similar before color first unary vertex green and consider red color for second ones, then mark (color) all leaves on the path (leaf) *red-green* (Figure 5.1, (2)). Note that for a general network with k reticulation vertices, the number of directed edges in the considered sequence, cannot be exceeded of k, because each marked red-green vertex is reticulation vertex.

- Consider a leaf g_g of the Motzkin tree. As similar before connect g_g to the two distinct unary vertices by using two sequences of outgoing directed edges. Mark g_g as *double-green* vertex and then color targeted unary vertices red. Also, consider red-green color for all the leaves on the paths of g_g to the unary vertices. (see Figure 5.1 (3)).

Note that in the above procedure the resulting graph must be a general phylogenetic network $\mathcal{G}$. We say then that T (keeping the colors from the above generation of $\mathcal{G}$, but not the edges) is a *colored Motzkin skeleton* (or simply Motzkin skeleton) of $\mathcal{G}$. Now, consider two sets, but not necessarily disjoint, of colored vertices obtained of above procedure. The member of first set is all colored vertices with outgoing edges and then assume all colored vertices with ingoing edges in the second set. Call them *pointer* and *target* sets respectively. In this way, red-green vertices are considered in both pointer and target sets. It is not hard to see that the size of target set is correspondent with number of reticulation nodes on a general phylogenetic network. Note that in this procedure any general network with no multiple edges and n vertices is generated and each of them exactly 2^k times, so in this case every network $\mathcal{G}$ with k reticulation vertices has exactly 2^k different Motzkin skeletons. However, note that as opposed to defined subclasses of phylogenetic networks like tree-child networks, here we assume multiple edges (reticulation vertex with one parent) are allowed to be in general networks. So for a reticulation vertex with just one parent, any arbitrary choice and removing of multiple edges, causes the same Motzkin skeleton. It means the described procedure generates a network with k reticulation vertices and r multiple edges exactly 2^{k-r} times. In the first step our aim is to set up an exponential generating functions for general networks with no multiple edges and then get the correspondent exponential generating function for other networks with at least one multiple edge.

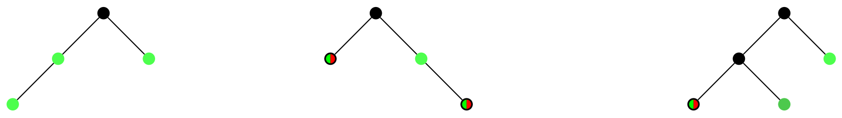


Figure 5.2: Corresponding sparsened skeletons of Figure 5.1 networks.

In order to set up generating functions for the number of general phylogenetic networks, we will construct them as follows: For a given network $\mathcal{G}$ fix one of its possible Motzkin tree skeletons, that shows us how the pointer set vertices are distributed within $\mathcal{G}$ (for instance consider networks in Figure 5.1 without marked

edges). Now look for sparsened skeleton of $\mathcal{G}$ which contains all pointer set vertices and contract all paths between any two vertices which are either pointer vertices or an ancestor of them to one edge. Note that this ancestor may be pointer vertices itself (also see, 4.2). In order to construct general networks with k target vertices (reticulations), we consider a sparsened skeleton with less than k pointer vertices. Then we replace all edges by paths that are made of red vertices or binary vertices with a Motzkin tree (whose unary vertices are all colored red) as second child and add a path of the same type on top of the root of the sparsened skeleton. Moreover, we attach a Motzkin tree (again with all unary vertices colored red) only to those leaves of the sparsened skeleton such that are just colored green (not red-green or double green). Note that red-green and double-green colored always are as leaves of sparsened skeleton. Do all of the above in such a way that the new structure has k target vertices (red and red-green) altogether. What we obtain so far is a Motzkin skeleton of a phylogenetic network. Finally, add edges connecting the pointer vertices to the target ones in such a way that the general phylogenetic networks condition is respected. As an advantage, a similar procedure can be used to set up generating functions for other kinds of phylogenetic networks, with fixed number of reticulation vertices, such as "stack-free" and "galled" networks that are defined in [55, 35].
Let us set up the exponential generating function for the Motzkin trees which appear in the above construction. After all, the unary vertices in those trees will be the red nodes of our network.

Denote by $M_{\ell,n}$ the number of all vertex-labeled Motzkin trees n vertices and ℓ unary vertices. Furthermore, let $\mathcal{M}$ be the set of all these Motzkin trees. The exponential generating function associated to $\mathcal{M}$ is

$$M(z,y) = \sum_{n\geq 1}\sum_{\ell\geq 0} M_{\ell,n} y^{\ell} \frac{z^n}{n!}.$$

Furthermore, let $M(z,y)$ denote the generating function associated to all Motzkin trees in $\mathcal{M}$ whose root is a unary node or a binary node, so we have

$$M(z,y) = z + zyM(z,y) + \frac{z}{2}M^2(z,y).$$

and thus

$$M(z,y) = \frac{1 - zy - \sqrt{1 + (y^2-2)z^2 - 2zy}}{z}. \tag{5.1}$$

The first few coefficients can be seen from

$$z + yz^2 + (y^2 + \frac{1}{2})z^3 + (y^3 + \frac{3}{2}y)z^4 + \cdots .$$

5.2 Counting Vertex-Labeled General Phylogenetic Networks With One Reticulation Vertex

In this section, Our main goal is to present a precise asymptotic result for the number of general phylogenetic networks with a fixed number k of reticulation vertices. To clear up the methods we start with simple cases, determine the asymptotic number of general phylogenetic networks with up to 3 reticulation nodes. After that, we will show how this approach helps us to present explicit formulas for the exact number of vetrex and leaf-labeled of them. Finally, we will focus on the general case and show how previous results lead us to prove theorems 5.0.3 and 5.0.4, for general phylogenetic networks with k reticulation vertices. As a warm-up consider general phylogenetic network with only one reticulation node we use the procedure to obtain (4.3) and the (sparsened) skeleton, as described in the previous section: Consider a general network with no multiple edges, delete one of the two incoming edges of the reticulation node which then gives a unary-binary tree with exactly two unary nodes which are colored green and red (we will consider general networks with multiple edges separately). Conversely, we can start with the general tree or even the sparsened skeleton and then construct the network from this. For more explicitly, Let $G_i^{\nparallel}(z)$ resp. $G_i^{\shortparallel}(z)$ denote exponential generating functions for networks with no multiple edges (with multiple edges) and i reticulation vertices.

Proposition 5.2.1. *The exponential generating function for vertex-labelled general phylogenetic networks with one reticulation node is*

$$G_1(z) = G_1^{\nparallel}(z) + G_1^{\shortparallel}(z) = z\frac{\tilde{a}_1(z^2) - \tilde{b}_1(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{\frac{3}{2}}}, \tag{5.2}$$

where,

$$\tilde{a}_1(z) = \tilde{b}_1(z) = 1 - z. \tag{5.3}$$

Proof. We start with the general Motzkin tree as depicted in Figure 5.3 (a) and add an edge starting from g and ending at a red vertex. Note that for all phylogenetic network, this edge is not allowed point to a node on the path from g to the root (since the network must be a DAG). Thus, when starting from the sparsened skeleton, *i.e.*, the single green vertex g, then we must add a sequence of trees on top of g which consist of a root (these vertices make the path from g to the root of the network) to which either a leaf or a binary node with two trees is attached. The red vertex must be contained in the forest made by this sequence or the tree attached to g. Note that the second expression refers to the depicted structure (b)

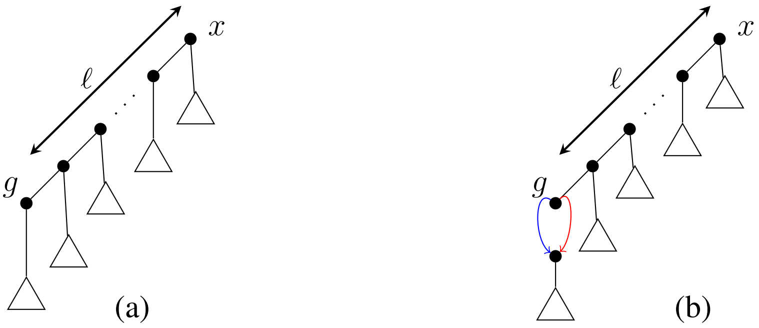


Figure 5.3: (a) The structure of Motzkin skeletons of networks with one reticulation vertex. It originates from a sparsened skeleton which consists of only one green vertex. It has one green vertex, denoted by g, and one red vertex which is hidden within the forest made of the triangles in the picture, which are attached to g and all the vertices on the path of length ℓ. Note that the position of the red vertex in this forest is restricted by the general condition. (b) There is multiple edge when the root of the subtree which is attached to g is the red node.

which is for general networks with a multiple edge. In terms of generating functions altogether gives

$$G_1(z) = \frac{1}{2}\frac{\partial}{\partial y}\frac{z\tilde{M}(z,y)}{1-zM(z,y)}\Big|_{y=0} + \frac{z^2M(z,0)}{1-zM(z,0)},$$

where,

$$\tilde{M}(z,y) = M(z,y) - zyM(z,y) = (1-zy)M(z,y). \tag{5.4}$$

The factor $1/2$ makes up for the fact that each network in case (a) is counted by the above procedure exactly twice. □

From this result we can now easily obtain the asymptotic number of networks.

Proposition 5.2.2. *Let $G_{1,n}$ denote the number of vertex-labelled general phylogenetic network with n vertices and one reticulation vertex. If n is even then $G_{1,n}$ is zero, otherwise*

$$G_{1,n} = n![z^n]G_1(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+1}\left(\frac{\sqrt{2}}{2} - \frac{\sqrt{\pi}}{2}\cdot\frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Proof. The function (5.2) has two dominant singularities, namely at $\pm 1/\sqrt{2}$, with singular expansions

$$G_1(z) \overset{z\to 1/\sqrt{2}}{\sim} \frac{1}{8(1-\sqrt{2}z)^{3/2}}, \qquad G_1(z) \overset{z\to -1/\sqrt{2}}{\sim} -\frac{1}{8(1+\sqrt{2}z)^{3/2}}.$$

Applying a transfer lemma for these two singularities completes the proof. □

Exact value of vertex-labeled general phylogenetic networks with one reticulation vertex

First, set $n = 2m+1$. Then, from (5.2) we obtain

$$[z^n]G_1(z) = [z^m]\bar{G}_1(z)$$

with

$$\bar{G}_1(z) = \frac{\tilde{a}_1(z) - \tilde{b}_1(z)\sqrt{1-2z}}{(1-2z)^{3/2}},$$

where $\tilde{a}_1(z)$ and $\tilde{b}_1(z)$ are as in (5.3). So we have

$$[z^m]\bar{G}_1(z) = [z^m]\frac{\tilde{a}_1(z)}{(1-2z)^{\frac{3}{2}}} - [z^m]\frac{\tilde{b}_1(z)}{(1-2z)}.$$

after some computation we have

$$[z^m]\bar{G}_1(z) = 2^m\big((m+1)\frac{\binom{2m}{m}}{4^m} - \frac{1}{2}\big).$$

By replacing $m = (n-1)/2$ this implies

$$G_{1,n} = n!2^{(n-3)/2}\big((n+1)\frac{\binom{n-1}{(n-1)/2}}{2^{n-1}} - 1\big). \tag{5.5}$$

5.2.1 Counting Leaf-Labeled General Phylogenetic Network

Let $G_{n,k}$ (resp.$\tilde{G}_{\ell,k}$) denote the number of vertex-labeled (leaf-labeled) general phylogenetic networks with n vertices (ℓ leaves) and k reticulation nodes. It is well studied in part 4.5, that for all subclasses of general networks containing only networks in which any two vertices have different sets of descendant, we have the following equation

$$G_{k,n} = \binom{n}{\ell}(n-\ell)!\,\tilde{G}_{k,\ell}. \tag{5.6}$$

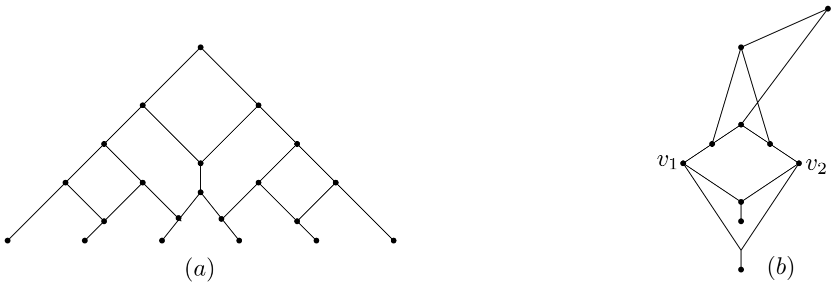


Figure 5.4: Two general phylogenetic networks, where in (a) the descendent set for any two vertices are different, and (b) is a general network which vertices v_1 and v_2 have a set of same descendent.

Here we briefly recall the argument that leads to equation 5.6. First consider $\ell + k = \dfrac{n+1}{2}$ for any phylogenetic network with ℓ leaves, k reticulation vertices and n vertices (Recall that n is always odd). Now all vertex-labeled general networks with n vertices and k reticulation vertices can be constructed as follows: start with a (fixed) leaf-labeled general network with ℓ leaves and k reticulation vertices. Then, choose ℓ labels from the set n labels and re-label the leaves of the fixed network such that the order is preserved. Finally, label the remaining $n - \ell$ vertices by any permutation of the set of remaining $n - \ell$ labels. By the above structure, in this way every vertex-labeled general network is obtained exactly once.

But for classes of networks where not all networks have the above mentioned property it is difficult to obtain a simple connection between the vertex-labeled and leaf-labeled phylogenetic networks. For that we have to cope with symmetry in some of generated general networks. Here, we will present complete details to show how to deal with symmetry for general networks with up to 3 fixed reticulation vertices. However, it will later be shown that as n goes infinity (resp.ℓ), the family of general networks that need to deal with symmetry are asymptotically negligible and thus one again expects $\tilde{G}_{k,\ell} \sim \dfrac{\ell!}{n!}G_{k,n}$, be a good approximation for all leaf-labeled general networks with fixed number of reticulation vertices when n goes to infinity.

As a warm up, we are going to take exact formula for leaf-labeled general phylogenetic networks with one reticulation vertex. By the above points we get,

$$\tilde{G}_{1,\ell} = \frac{\ell!}{n!}G_{1,n}.$$

After seting $n = 2\ell + 1$ in (5.5) we have

$$\tilde{G}_{1,l} = \ell!\, 2^{\ell}\Big((\ell+1)\frac{\binom{2\ell}{\ell}}{4^{\ell}} - \frac{1}{2}\Big). \tag{5.7}$$

Remark. Relationship to Tree-child networks. Note that general phylogenetic networks with exactly one reticulation vertex and no multiple edges are tree-child networks. It means that $G_1^{\natural}(z)$ is exactly correspondent to generating function for vertex-labeled tree-child networks with one reticulation vertex. This translates into

$$T_1(z) = G_1^{\natural}(z) = \frac{1}{2}\frac{\partial}{\partial y}\frac{z\tilde{M}(z,y)}{1 - zM(z,y)}\Big|_{y=0}.$$

as it must be. In the same way as before the mentioned approaches immediately implies the formulas 4.25 and 4.26 for the number of vertex-labeled and leaf-labeled general phylogenetic networks with one reticulation vertex.

This approach for leaf labeled case can be saw in [66] with different methods.

5.3 General Phylogenetic Network With Two Reticulation Vertices

Now we expand the methods for general phylogenetic networks with 2 reticulation nodes. For this case,we use some variables $y_1, y_2, y_{r_g}, y_{g_g}$ to express the possible pointing of the pointer set vertices of the Motzkin skeletons. Furthermore, we have now more complicated paths (and attached trees) which replace the edges of the sparsened skeleton and thus we first set up the generating function corresponding to theses paths. To govern the situation where an edge from one of the two pointer set vertices must not point to a certain vertex on the paths in order to avoid multiple edges in the first step, we distinguish three types of unary vertices, which are the red vertices of our construction. We will define is a class $\mathcal{P}$ of paths which serve as the essential building blocks for Motzkin skeletons. In this class the rules for pointing to particular red vertices differ, depending on whether (i) the red vertex lies on the path itself, (ii) it is one of the vertices of one of the attached subtrees (iii) the red vertex is the first vertex of the path. We will mark the red vertices of type (i) with the variable y, those of type (ii) with $\tilde{y}$ and the vertex of type (iii) with $\hat{y}$ which is introduced in order to manage structures analysis multiple edges in phylogenetic networks.

To simplify the explanation, let us use the following conventions where the ε denotes the empty tree. Each path is a sequence of vertices with trees attached. Note that each red vertex may belong to different categories (if it is first vertex of

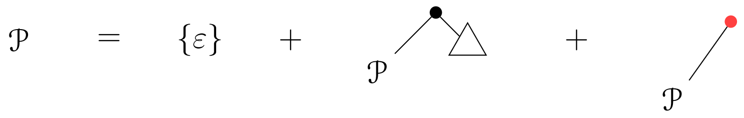


Figure 5.5: The specification of the class $\mathcal{P}$.

path marked with $\hat{y}$, otherwise with y). In our analysis the variables y, $\tilde{y}$ and $\hat{y}$ will be replaced by a sum of variables y_i for $i \in \{1, 2, r_g, g_g\}$, where the presence of a particular y_i indicates that the corresponding g_i is allowed to point, its absence that pointing is forbidden. In particular, y represent the permission to point to vertices of the path (except its first vertex) and $\tilde{y}$ describes the permission to point to vertices of attached trees and $\hat{y}$ allows pointing to the first vertex of the path. We specify $\mathcal{P}$ as

$$\mathcal{P} = \frac{1 - zy}{1 - z(y + M(z, \tilde{y}))} + \frac{z\hat{y}}{1 - z(y + M(z, \tilde{y}))}.$$

This leads to the generating function

$$P(z, y, \tilde{y}, \hat{y}) = \frac{1 - zy + z\hat{y}}{1 - z(y + M(z, \tilde{y}))},$$

after all. Start with this assumption that in the Motzkin skeletons added directed edges not allowed to make multiple edges., see Figure 5.6, and then add the contribution of other all possible Motzkin tree skeletons with multiple edges as shown in Figure 5.7. Now we are ready to state the following result.

Proposition 5.3.1. *The exponential generating function for vertex-labeled general phylogenetic networks with two reticulation nodes is*

$$G_2(z) = G_2^{\nparallel}(z) + G_2^{\parallel}(z) = z \cdot \frac{\tilde{a}_2(z^2) - \tilde{b}_2(z^2)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{7/2}},$$

where

$$\tilde{a}_2(z) = z^4 - 2z^3 - \frac{1}{2}z^2 + \frac{5}{2}z \quad \textit{and} \quad \tilde{b}_2(z) = -z^2 + \frac{5}{2}z.$$

Proof. Consider the general phylogenetic networks arising from the Motzkin skeleton on the Figure 5.6 (a) and complete the Motzkin skeleton by adding two egdes having start vertex g_1 and g_2, respectively. Due to this, note that pointings of the green vertices do not violate the general phylogenetic network properties

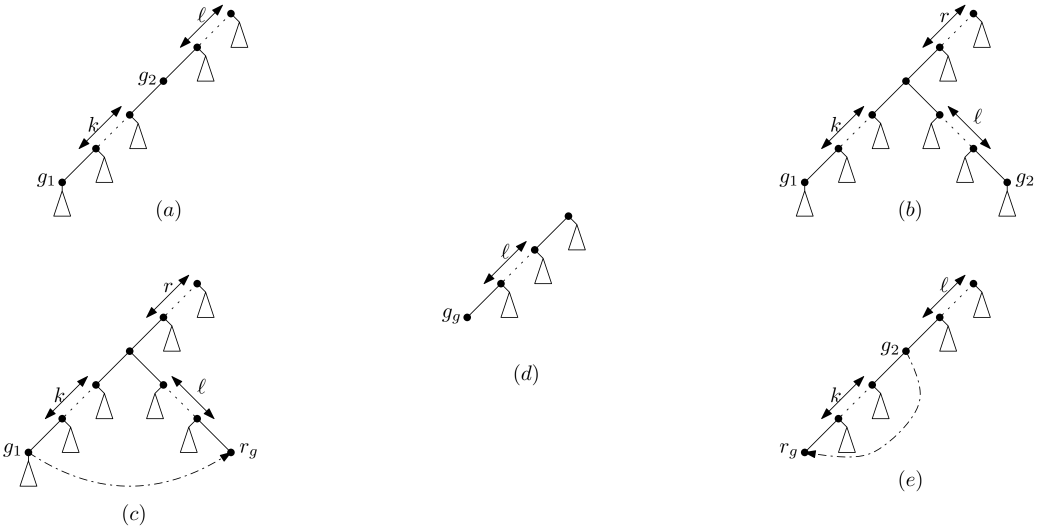


Figure 5.6: The possible structures of the Motzkin skeletons of general phylogenetic networks with 2 reticulation nodes such that added edges not allowed to make multiple edges.

by making a directed cyclic component. Also, to avoid multiple edges, set up the generating function $\tilde{M}_1(z, y_1 + y_2)$ for the tree attached to the green vertex g_1. In general $\tilde{M}_i(z, y_1 + y_2) = (1 - zy_i)M(z, y_1 + y_2)$ is the specification of unary root Motzkin trees such that pointer vertex which already marked by variable y_i, is not allowed to point to the root vertex. So this means pointing to the root of this tree is forbidden for g_1 but not for g_2. For all the other trees there is no pointing restriction. The analysis of the vertices on the paths is done path by path.

- Path ℓ: No green vertex is allowed to point to the vertices of that path.
- Path k: Except first node, pointing to all vertices is allowed for g_2, but g_1 may not point to that path at all. So we have

$$G^{\nparallel}_a(z) = \partial_{y_1}\partial_{y_2} z^2 \tilde{M}_1(z, y_1 + y_2)P(z, y_2, y_1 + y_2, 0)P(z, 0, y_1 + y_2, 0)\Big|_{y_1=0, y_2=0}.$$

Now, consider the Motzkin skeleton (b). For the trees attached to the green vertices only pointing to the root is forbidden for parent vertices, for all the other trees there is no pointing restriction. The analysis of the vertices on the paths is done path by path.

- Path r: No green vertex is allowed to point to the vertices of that path.
- Path k: Pointing to all vertices is allowed for g_2, but g_1 may not point to that path at all. The situation for path ℓ is symmetric.

In this way, Motzkin skeletons which are not respecting the general condition are generated as well: Indeed, g_1 may point to the vertex of ℓ and g_2 to the vertex of k, thus violating the generality condition by making a cycle. The factor $\frac{1}{2}$ in the beginning of expression comes from the "horizontal symmetry" (This can be briefly shown by *H*-*S*) of the Motzkin skeleton. This yields the generating function

$$G^{\natural}_b(z) = \frac{1}{2}\partial_{y_1}\partial_{y_2}\left(\frac{z^3\tilde{M}_1(z,y_1+y_2)\tilde{M}_2(z,y_1+y_2)}{1-zM(z,y_1+y_2)}P(z,y_2,y_1+y_2,y_2)P(z,y_1,y_1+y_2,y_1)\right.$$
$$\left. - \frac{z^3M(z,0)^2}{1-zM(z,0)}P(z,y_2,0,y_2)P(z,y_1,0,y_1)\Big|_{y_1=0,y_2=0}\right).$$

The other case of general networks has the Motzkin skeleton as shown in Figure 5.6, (c). The property of red-green leaf entails, first one added directed edges connects g_1 to r_g. After that, there is no restriction for pointing of r_g except the vertices on the paths. This gives

$$G^{\natural}_c(z) = \partial_{y_r}\frac{z^3M(z,y_r)}{(1-zM(z,y_r))^3}\Big|_{y_{r_g}=0}.$$

Now, consider the Motzkin skeleton (d) of Figure 5.6. The double-green vertex g_g can point to all vertices (the pointing order does not matter, so we divide by 2) in the attached subtrees.

$$G^{\natural}_d(z) = \frac{1}{2}(\partial_{y_g})^2\frac{z}{1-zM(z,y_g)}\Big|_{y_g=0}.$$

For the final case, consider the Motzkin skeleton (e). Generality condition entails that r_g be only possible target vertex for pointing of g_2. For all the other trees there is no pointing restriction for r_g. To avoid of multiple edges, the path k cannot be a simple edge. To do that set the generating function

$$P^\star(z,y,\tilde{y},\hat{y}) = P(z,y,\tilde{y},\hat{y}) - 1 = \frac{zM(z,\tilde{y})+z\hat{y}}{1-z(y+M(z,\tilde{y}))},$$

for a nonempty path. Then we get

$$G^{\natural}_e(z) = \partial_{y_r}\frac{z}{1-zM(z,y_r)}P^\star(z,0,y_r,0)\Big|_{y_r=0} = \partial_{y_r}\frac{z^2M(z,y_r)}{(1-zM(z,y_r))^2}\Big|_{y_r=0}.$$

The exponential generating function for vertex-labeled general networks (with no multiple edges) is obtained as $G^{\natural}_2(z) = G^{\natural}_a(z)+G^{\natural}_b(z)+G^{\natural}_c(z)+G^{\natural}_d(z)+G^{\natural}_e(z)/4$ after all. This gives the following result.

$$G^{\natural}_2(z) = z\cdot\frac{a^{\natural}_2(z^2)-b^{\natural}_2(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{7/2}}, \tag{5.8}$$

where

$$a_2^{\sharp}(z) = z^4 + \frac{1}{2}z^2 + \frac{3}{2}z \quad \text{and} \quad b_2^{\sharp}(z) = z^2 + \frac{3}{2}z. \tag{5.9}$$

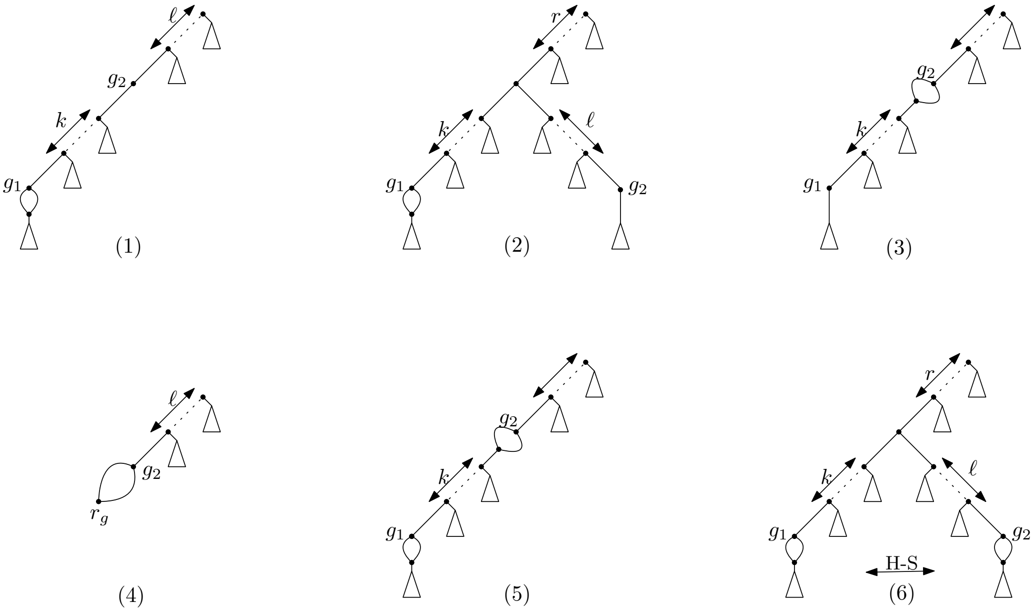


Figure 5.7: The possible structures of the Motzkin skeletons of phylogenetic networks with 2 reticulation nodes and all possible considered of fixed multiple edges contributions.

Next we will set up the exponential generating function for general networks with at least one multiple edges on their structure (see Figure 5.7). Altogether, we obtain

$$G_2^{\Vert}(z) = \frac{1}{2}\left(\partial_{y_2} z^3 M(z,y_2)\frac{P(z,y_2,y_2,0)}{1-zM(z,y_2)}\bigg|_{y_2=0} + \partial_{y_2}\frac{z^4 M(z,y_2)\tilde{M}_2(z,y_2)P(z,y_2,y_2,y_2)}{(1-zM(z,y_2))^2}\bigg|_{y_2=0}\right.$$
$$\left. +\partial_{y_1}\frac{z^3\tilde{M}_1(z,y_1)}{(1-zM(z,y_1))^2}\bigg|_{y_1=0} + \partial_{y_r}\frac{z^2}{(1-zM(z,y_r))}\bigg|_{y_r=0}\right)$$
$$+ \frac{z^4 M(z,0)}{(1-zM(z,0))^2} + \frac{1}{2}\frac{z^5 M^2(z,0)}{(1-zM(z,0))^3},$$

where the factor 2 appears for the expression of (1) to (4), because in these cases each general phylogenetic network is generated two times. Note that, there is just a unique general network which arises from the Case 5. Also, the factor 2 appears in last term, because of horizontal symmetry. □

So the exponential generating function for vertex-labeled general phylogenetic networks with two reticulation nodes is then $G_2(z) = G_2^{\nparallel}(z) + G_2^{\parallel}(z)$. As an easy consequence, we obtain the asymptotic number of networks.

Corollary 5.3.2. *Let $G_{2,n}$ denote the number of vertex-labeled general phylogenetic networks with n vertices and exactly two reticulation vertex. If n is even then $G_{2,n}$ is zero, otherwise*

$$G_{2,n} = n![z^n]G_2(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+3}\left(\frac{\sqrt{2}}{16} - \frac{\sqrt{\pi}}{8} \cdot \frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Proof. This follows by singularity analysis as before. □

Explicit formula for vertex-labeled general networks with two reticulation vertices

We can use generating functions $G_2^{\nparallel}(z)$ and $G_2^{\parallel}(z)$ to extract closed formulas for vertex-labeled general networks. To see that, consider the contribution of each of them separately. Start with the exponential generating function $G_2^{\nparallel}(z)$ for general networks that do not have double edges in own structures.

First, set $n = 2m + 1$. Then, from 5.8 we obtain

$$[z^n]G_2^{\nparallel}(z) = [z^m]\bar{G}_2^{\nparallel}(z)$$

with

$$\bar{G}_2^{\nparallel}(z) = \frac{a_2^{\nparallel}(z) - b_2^{\nparallel}(z)\sqrt{1-2z}}{(1-2z)^{7/2}},$$

where $a_2^{\nparallel}(z)$ and $b_2^{\nparallel}(z)$ are as in 5.9. So we have

$$[z^m]\bar{G}_2^{\nparallel}(z) = [z^m]\frac{a_2^{\nparallel}(z)}{(1-2z)^{\frac{7}{2}}} - [z^m]\frac{b_2^{\nparallel}(z)}{(1-2z)^3}.$$

After some computation we have

$$[z^m]\bar{G}_2^{\nparallel}(z) = 2^{m-2}\left(P_1(m)\frac{2m\binom{2m}{m}}{15(2m-1)4^m} - P_2(m)\right),$$

where

$$P_1(m) = 30m^3 + 20m^2 + 15m - 20 \quad \text{and} \quad P_2(m) = 2m^2 + m. \tag{5.10}$$

By replacing $m = (n-1)/2$ this implies

$$G^{\nparallel}_{2,n} = n!2^{(n-5)/2}\Big(P_1((n-1)/2)\frac{(n-1)\binom{n-1}{(n-1)/2}}{15(n-2)2^{n-1}} - P_2((n-1)/2)\Big). \tag{5.11}$$

Note that correspondent generating function for general networks with multiple edges is

$$G^{\parallel}_2(z) = z \cdot \frac{a^{\parallel}_2(z^2) - b^{\parallel}_2(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{1/2}},$$

such that

$$a^{\parallel}_2(z) = z^2 + z \quad \text{and} \quad b^{\parallel}_2(z) = z.$$

In the same way, it can be used to get exact formula for vertex-labeled general networks that are belong to this subclass. We refrain from giving details and just list the obtained expressions. The reader is invited to derive them herself.

$$G^{\parallel}_{2,n} = n!2^{(n-3)/2}(n-1)\Big(\frac{(n-1)\binom{n-1}{(n-1)/2}}{2^n} - \frac{1}{2})\Big). \tag{5.12}$$

After all by summing up 5.11 and 5.12 we have

$$G_{2,n} = G^{\nparallel}_{2,n} + G^{\parallel}_{2,n} = n!2^{(n-3)/2}\Big(A((n-1)/2)\frac{(n-1)\binom{n-1}{(n-1)/2}}{15(n-2)2^{n-1}} - B((n-1)/2)\Big), \tag{5.13}$$

where

$$A(m) = 30m^3 + 80m^2 - 15m - 20 \quad \text{and} \quad B(m) = m^2 + \frac{3}{2}m. \tag{5.14}$$

Explicit formula for leaf-labeled general networks with two reticulation vertices

Note that, the Equation (5.6) which comes from the described procedure in Section 5.2.1 for construction all vertex-labeled networks from fixed leaf-labeled ones does not work anymore. It is because by applying the method there are some leaf-labeled networks which generate some vertex-labeled networks more than one (here twice). Thus for normalization, and deal with symmetry the correspondent generating functions of such networks can be considered separately (see Figure 5.8).

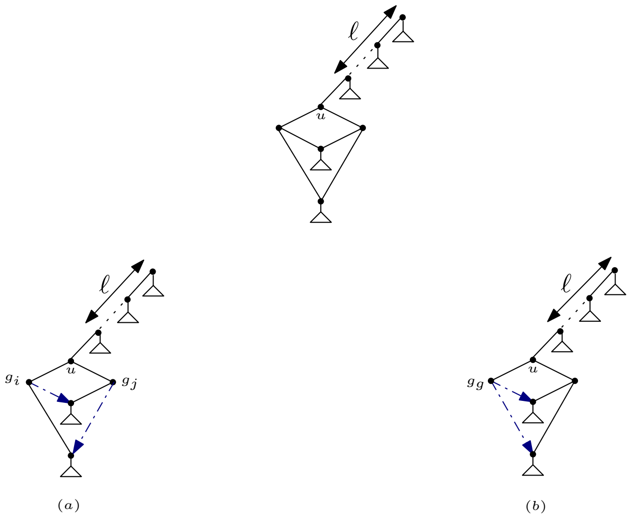


Figure 5.8: (Top) The structure of general network with two reticulation vertices such that two vertices have the same set of descendant which can be generate from (a) by pointing green vertices to the root of each other attached trees or (b) a double-green vertex points to unary vertices with same parent.

So we have

$$G_2^{\natural}(z) = \dot{G}_2^{\natural}(z) + \ddot{G}_2^{\natural}(z),$$

where $\dot{G}_2^{\natural}(z)$ correspondent generating function for general networks such that the procedure set out in Section 5.2.1 can be applied directly for them.

$$\dot{G}_2^{\natural}(z) = z \cdot \frac{\dot{a}_2^{\natural}(z^2) - \dot{b}_2^{\natural}(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{7/2}},$$

where,

$$\dot{a}_2^{\natural}(z) = -4z^5 + 11z^4 - 9z^3 + 4z^2 + z \quad \text{and} \quad \dot{b}_2^{\natural}(z) = 4z^4 - 6z^3 + 4z^2 + z.$$

For this set of general networks we can directly use Equation (5.6). Thus, same procedure as like before gives us

$$\dot{G}_{2,\ell}^{\natural} = \ell! 2^{\ell-1}\left((6\ell^4 + 19\ell^3 + 18\ell^2 - 7\ell - 3)\frac{(\ell+1)\binom{2\ell+2}{\ell+1}}{(6\ell-3)(2\ell+1)4^{\ell}} - (2\ell^2 + 5\ell + 3)\right). \tag{5.15}$$

Now we set up generating function for family of networks which are shown in the top of Figure 5.8. It is not hard to see that, by using the previous methods each related (fixed) leaf-labeled general network can construct a vertex-labeled general network exactly twice (because of symmetry). For this case Equation (5.6) can modify as $\ddot{G}^{\mathrm{H}}_{2,\ell} = 2\frac{\ell!}{n!}\ddot{G}^{\mathrm{H}}_{2,n}$. The generating function for this subfamily of general networks is

$$\ddot{G}^{\mathrm{H}}_2(z) = z \cdot \frac{\ddot{a}^{\mathrm{H}}_2(z^2) - \ddot{b}^{\mathrm{H}}_2(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{1/2}},$$

where

$$\ddot{a}^{\mathrm{H}}_2(z) = -\frac{1}{2}z^2 + \frac{1}{2}z \quad \text{and} \quad \ddot{b}^{\mathrm{H}}_2(z) = \frac{1}{2}z.$$

After some manipulation we get

$$\ddot{G}^{\mathrm{H}}_{2,\ell} = \ell!2^{\ell-1}\Big(\frac{(\ell+1)\binom{2\ell+2}{\ell+1}}{(2\ell-1)(2\ell+1)4^{\ell}}\Big).$$

The explicit formula for leaf-labeled general networks with no multiple edges and two reticulation vertices is

$$\begin{aligned} G^{\mathrm{H}}_{2,\ell} &= \dot{G}^{\mathrm{H}}_{2,\ell} + \ddot{G}^{\mathrm{H}}_{2,\ell} \\ &= \ell!2^{\ell-1}\Big((6\ell^4+19\ell^3+18\ell^2-7\ell)\frac{(\ell+1)\binom{2\ell+2}{\ell+1}}{(6\ell-3)(2\ell+1)4^{\ell}} - (2\ell^2+5\ell+3)\Big). \end{aligned} \tag{5.16}$$

To complete the details, we can get explicit formula for the number of leaf-labeled networks that are generated by sparsened skeletons which as depicted in Figure 5.7. Note that for this case, all generated networks belong to the first subclass of general networks which the Equation 5.6 can be used directly. So we have

$$G^{\shortparallel}_{2,\ell} = \ell!2^{\ell+1}(\ell+1)\Big(\frac{(\ell+1)\binom{2\ell+2}{\ell+1}}{4^{\ell+1}} - \frac{1}{2}\Big).$$

After all, we get

$$\tilde{G}_{2,\ell} = G^{\mathrm{H}}_{2,\ell} + G^{\shortparallel}_{2,\ell} = \ell!2^{\ell-1}\Big(A(\ell)\frac{(\ell+1)\binom{2\ell+2}{\ell+1}}{(6\ell-3)(2\ell+1)4^{\ell}} - B(\ell)\Big)\Big), \tag{5.17}$$

where

$$A(\ell) = 6\ell^4+31\ell^3+30\ell^2-10\ell-3 \quad \text{and} \quad B(\ell) = 2\ell^2+\frac{41}{8}\ell+\frac{25}{8}. \tag{5.18}$$

5.4 General Phylogenetic Network With Three Reticulation Nodes

In the same way, the methods can be used to study of specifications for general phylogenetic networks with $k \geq 3$ reticulation nodes. Its obvious by increasing number of rediculation nodes we have to consider more variates of Motzkin skeletons to cover all possible cases. As similar as $k = 2$, first we consider structures with no generated multiple edges and then for each Motzkin skeleton respectively, we look for possible contributions of multiple edges on the structures and add them to the results. Here we want to prove following results.

Proposition 5.4.1. *The exponential generating function for vertex-labeled general phylogenetic networks with three reticulation nodes is*

$$G_3(z) = G_3^{\nparallel}(z) + G_3^{\parallel}(z) = z \cdot \frac{a_3(z^2) - b_3(z^2)\sqrt{1-2z^2}}{(1-2z^2)^{11/2}},$$

where

$$a_3(z) = z^6 + 5z^5 - 10z^4 - \frac{23}{2}z^3 + \frac{109}{4}z^2,$$

and,

$$b_3(z) = z^5 - \frac{7}{2}z^4 - 5z^3 + \frac{109}{4}z^2.$$

Corollary 5.4.2. *Let $G_{3,n}$ denote the number of vertex-labeled general phylogenetic networks with n vertices and exactly three reticulation vertices. If n is even then $G_{3,n}$ is zero, otherwise*

$$G_{3,n} = n![z^n]G_3(z) = \left(\frac{\sqrt{2}}{e}\right)^n n^{n+5}\left(\frac{\sqrt{2}}{192} - \frac{\sqrt{\pi}}{64} \cdot \frac{1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$.

Also, as before we can take the explicit formulas for vertex and leaf-labeled general networks with 3 reticulation vertices. For vertex labeled case, as like before we set $n = 2m + 1$, so we have

$$[z^n]G_3(z) = [z^m]\bar{G}_3(z),$$

such that,

$$[z^m]\bar{G}_3(z) = [z^m]\frac{a_3(z)}{(1-2z)^{\frac{11}{2}}} - [z^m]\frac{b_3(z)}{(1-2z)^5}.$$

It gives coefficient extractor function

$$\mathcal{F}(m) := [z^m]\bar{G}_3(z) = \frac{2^{m-6}}{3}\left(A_1(m)\frac{m(m-1)\binom{2m}{m}}{35(2m-1)4^{m-2}} - B_1(m)\right).$$

where

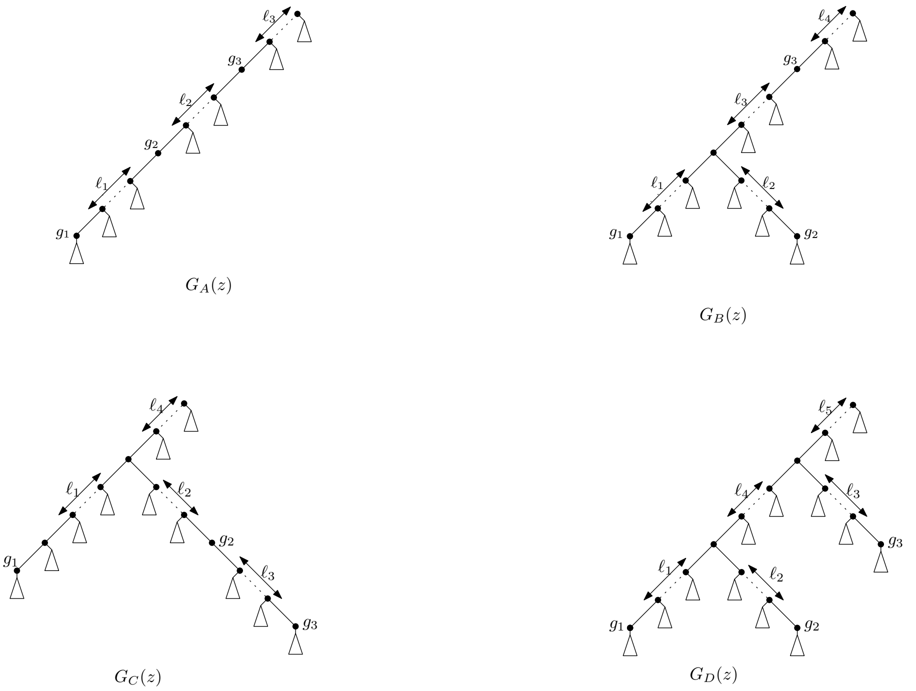
$$\begin{aligned} A_1(m) &= 104m^4 + 836m^3 + 876m^2 - 454m - 79, \\ B_1(m) &= 48m^4 + 127m^3 - 60m^2 - 121m + 6. \end{aligned}$$

By replacing $m = (n-1)/2$, we have,

$$G_{3,n} = n! \cdot \mathcal{F}((n-1)/2).$$

Now, we want to present a theoretical extension of the studied procedure for general phylogenetic network with three reticulation vertices to prove Proposition (5.4.1). After that we can show how extract an explicit formula for leaf-labeled general networks with 3 reticulation vertices. As before, we decompose the network according to how the pointer vertices are distributed in the networks. More explicitly, first consider the Motzkin skeletons with just green vertices (Figure 5.9). We can use them to figure out the rest of Motzkin skeletons with red-green and double-green vertices as well. In the end, we add the contribution of the Motzkin skeletons with multiple edges. Recall that, for $i, \ j \in \{1, 2, 3, r, g\}$, $\mathbf{Y}_{i,\ldots,j}$ denote the operator differentiating with respect to $y_i, \ldots, \ y_j$ and setting $y_i = \ldots = y_j = 0$ afterwards, *i.e.*, $\mathbf{Y}_{i,\ldots,j} f(z, y_i, \ldots, y_j) = \left(\partial_{y_i} \ldots \partial_{y_j} f\right)(z, 0, \ldots, 0)$. First we set up generating functions for figure 5.9 cases. To do that, we follow the same procedure that used for general phylogenetic networks with two reticulation vertices. Start with simple case where the three green vertices lie on one path, *i.e.*, one green vertex is ancestor of another, which itself is ancestor of the third one. All possibilities for the pointings of the edges starting at g_1, g_2 and g_3 may target any vertex in all the other trees. Concerning the vertices on the spine, we have some restrictions. The edge from g_1 may not end at any vertex from $\ell_1 \cup \ell_2$ and the root of its attached subtree. The edges from g_2 may not point to first vertex of ℓ_1 (to avoid of multiple edges) and any vertex of ℓ_2. Finally, no green vertex may point to the vertex of ℓ_3. Note that the contribution of multiple edges will be considered in later cases. Overall, this yields the generating function

$$\begin{aligned} G_A(z) = \mathbf{Y}_{1,2,3}\Big(&\frac{z^3\tilde{M}_1(z, y_1+y_2+y_3)}{1 - zM(z, y_1+y_2+y_3)} P(z, y_3, y_1+y_2+y_3, 0) \\ &\times P(z, y_2+y_3, y_1+y_2+y_3, y_3)\Big). \end{aligned}$$



$G_A(z)$

$G_B(z)$

$G_C(z)$

$G_D(z)$

Figure 5.9: The possible structures of Motzkin skeletons of networks with 3 reticulation vertices. All of them, originate from a sparsened skeleton with this assumption that pointer set consists of only green vertices.

Next we will determine the generating function of all general networks belonging the case where one green vertex is a common ancestor of the other two, but none of those two is ancestor of the other one. As in the previous section we analyse the substructures. There are four vertices in the sparsened skeleton, yielding a factor z^4 . Any non-root red vertices in the subtree attached to g_1 may be targets of the edge coming from any green nodes and for root one, pointing is allowed for g_2 and g_3 (but not g_1 to avoid multiple edges). for the subtree attached to g_2 vice versa.

- Paths ℓ_3 and ℓ_4: These paths are sequences of vertices, each with a subtree attached to it. For ℓ_4 each green vertex is allowed to point at the red vertices in these subtrees. Pointing to the vertices of the path is not allowed. Likewise, just the corresponding vertices on the path of ℓ_3 are forbidden for g_1 and g_2 by the generality condition but g_3 may points non-first vertex of that as well.

- Paths ℓ_1 and ℓ_2: They are symmetric, so we discuss ℓ_1. The vertices of the subtrees are allowed targets for the edge from all green vertices. The edge from g_2 and g_3 may end at each vertex of the path.

Note that the generality condition will be violated by making a cycle, if g_2 points at a red vetex on the path ℓ_1 and g_1 does it vice versa. We subtract this cases from the result. Overall, this gives, again using the operator $\mathbf{Y}_{i,\dots,j}$ defined above, the generating function

$$\begin{aligned}G_B(z) =&\frac{1}{2}\mathbf{Y}_{1,2,3}\left(\frac{z^4\tilde{M}_1(z,y_1+y_2+y_3)\tilde{M}_2(z,y_1+y_2+y_3)}{1-zM(z,y_1+y_2+y_3)}P(z,y_1+y_3,y_1+y_2+y_3,y_1+y_3)\right.\\ &\times P(z,y_3,y_1+y_2+y_3,0)P(z,y_3+y_2,y_1+y_2+y_3,y_3+y_2)\\ &\left.-\frac{z^4M(z,y_3)^2}{(1-zM(z,y_3))}P(z,y_1+y_3,y_3,y_1+y_3)P(z,y_2+y_3,y_3,y_2+y_3)P(z,y_3,y_3,0)\right).\end{aligned}$$

Next we pay attention to the case where one green vertex is ancestor of another one, but not of both of them, and the third one is not ancestor of any other green vertex. in Figure 5.9. The sparsened skeleton has 4 vertices and the subtrees attached to g_1 and g_3. The red vertices of the subtree of g_1 and g_3 may be targeted by any edges starting from green vertices. Note that if g_1 and g_3 have the red root attached subtrees, they are not allowed to point at own attached red root vertex respectively to avoid multiple edges. Next we inspect the paths:

- Path ℓ_4: All green vertices may point to the vertices of the subtrees. Pointing to the path itself is not allowed.
- Path ℓ_3: The edge starting at g_3 may point to vertices of the subtrees, but not to the vertices of the path itself. All but the first vertex for g_2 of the path as well as all tree vertices can be the end point of the edge starting at g_1 and g_2.
- Path ℓ_1: Similar to ℓ_3. The edges from g_2 and g_3 may point anywhere of the path. The vertices of the subtrees may be targeted by g_1 as well.
- Path ℓ_2: All green vertices may point to the vertices of the subtrees. To point at the vertices on the path is only allowed for g_1.

Altogether, we obtain the generating function $G_C(z)$ with the expression

$$\begin{aligned}G_C(z) = \mathbf{Y}_{1,2,3}&\left(\frac{z^4\tilde{M}_3(z,y_1+y_2+y_3)\tilde{M}_1(z,y_1+y_2+y_3)}{1-zM(z,y_1+y_2+y_3)}P(z,y_2+y_3,y_1+y_2+y_3,y_2+y_3)\right.\\ &P(z,y_1,y_1+y_2+y_3,y_1)P(z,y_1+y_2,y_1+y_2+y_3,y_1)\\ &-\frac{z^4M(z,0)^2}{1-zM(z,0)}P(z,y_1,0,y_1)^2P(z,y_2+y_3,0,y_2+y_3)\\ &-\frac{z^4\tilde{M}_3(z,y_3)M_3(z,y_3)}{(1-zM(z,y_3))^2}P(z,y_1,y_3,y_1)P(z,y_2,y_3,y_2)\\ &\left.-\frac{z^4M_2(z,y_2)^2}{1-zM(z,y_2)}P(z,y_1,y_2,y_1)P(z,y_1+y_2,y_2,y_1)P(z,y_3,y_2,y_3)\right).\end{aligned}$$

In this way, Motzkin skeletons which are not respecting the generality condition are generated as well: Indeed, g_1 may point to the vertex on the paths ℓ_2 or ℓ_3 when both or one of g_2 and g_3 point to vertex of ℓ_1, such that makes directed cyclic component.

The last case of general networks has Motzkin skeletons as shown in Figure 5.9. The restriction for the target vertex of the edges to be added at g_1, g_2 and g_3 follow the analogous rules in order to meet the generality constraint. Setting up the generating function follows the same pattern as before. We omit now the details and get from path analysis after all

$$\begin{aligned}
G_D(z) = {} & \frac{1}{2}\mathbf{Y}_{1,2,3} \\
& \Bigg(\frac{z^5 \tilde{M}_3(z, y_1+y_2+y_3)\tilde{M}_2(z, y_1+y_2+y_3)\tilde{M}_1(z, y_1+y_2+y_3)}{1 - zM(z, y_1+y_2+y_3)} P(z, y_1+y_2, y_1+y_2+y_3, y_1+y_2) \\
& \times P(z, y_1+y_3, y_1+y_2+y_3, y_1+y_3)P(z, y_2+y_3, y_1+y_2+y_3, y_2+y_3)P(z, y_3, y_1+y_2+y_3, y_3) \\
& - \frac{z^5 \tilde{M}_3(z, y_3)M(z, y_3)^2}{(1 - zM(z, y_3))^2} P(z, y_1+y_3, y_3, y_1+y_3)P(z, y_2+y_3, y_3, y_2+y_3)P(z, y_3, y_3, y_3) \Bigg) \\
& - \mathbf{Y}_{1,2,3} \left(\frac{z^5 \tilde{M}_1(z, y_3)M(z, y_1)^2}{(1 - zM(z, y_1))^2} P(z, y_1+y_3, y_1, y_1+y_3)P(z, y_2+y_1, y_1, y_2+y_1)P(z, y_3, y_1, y_3) \right. \\
& \left. - \frac{z^5 M(z, 0)^3}{(1 - zM(z, 0))^2} P(z, y_1, 0, y_1)P(z, y_2, 0, y_2)P(z, y_3, 0, y_3) \right).
\end{aligned}$$

So far we just have considered the Motzkin skeletons in Figure 5.9 with three reticulation vertices such that only green vertices are as pointer set vertices. Now we consider the structure of the Motzkin skeletons with red-green and double-green vertices and set up generating functions for them separately. Note that, the crucial point is that distribution of pointer nodes on the Motzkin skeleton must be such a way that, after adding directed edges, we get a general phylogenetic network with 3 reticulation vertices. Recall that, for any red-green leaf first we consider another pointer vertex such that connects to this nodes by adding a directed edge. Let's start with the Motzkin skeletons that contain at least one red-green vertex. Consider a case with three pointer vertices lie on a path (two green colored vertices with a red-green leaf), such a way that a red-green once lies on the bottom of the path (left of Figure 5.10). Note that, we have two different expressions depends on our choice that which green vertex (g_2 or g_3) is considered first to point red-green leaf.

$$G_A^{r_g}(z) = \mathbf{Y}_{r,3} \frac{z^3 P^{\star}(z, y_3, y_r+y_3, y_3)P(z, y_3, y_r+y_3, 0)}{1 - zM(z, y_r+y_3)} + \mathbf{Y}_{r,2} \frac{z^3 P(z, y_2, y_r+y_2, 0)}{(1 - zM(z, y_r+y_2))^2}.$$

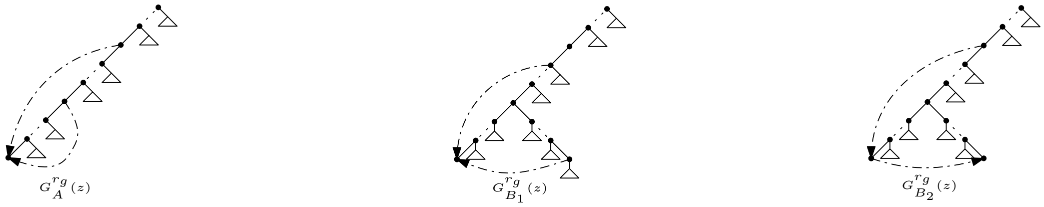


Figure 5.10: The structure of Motzkin skeletons with red-green leaves that can be generated from the Motzkin skeletons $G_A(z)$ and $G_B(z)$. In the left and right figures, g_1 is replaced with a red-green leaf, and in the right one two red-green leaves are considered as a pointer vertex at the end of cherry.

Note that to avoid multiple edges, the path between g_2 and red-green vertex cannot be empty edge, in the case of added directed edge connects g_2 to the red-green leaf. As before, there are two possible cases for the general networks arising from the Motzkin skeletons depicted on the middle of Figure 5.10. In the first case, if we fix an added directed edge from g_2 to red-green leaf, the only restriction for pointing of g_3 will be the vertices on the path that connects it to the root and its first child vertex (to avoid multiple edge). The red-green vertex may point to any non-path vertex. The second term is regards the situation that a shortcut connects g_3 to the red-green vertex. After subtracting Motzkin skeletons which are not respecting general network condition, we obtain

$$\begin{aligned}G^{rg}_{B_1}(z) =&\mathbf{Y}_{r,3}\left(\frac{z^4M(z,y_r+y_3)}{1-zM(z,y_r+y_3)}P(z,y_3,y_r+y_3,0)P(z,y_3,y_r+y_3,y_3)^2\right)\\&+\mathbf{Y}_{r,2}\left(\frac{z^4\tilde{M_2}(z,y_2+y_r)}{(1-zM(z,y_r+y_2))^2}P(z,y_2,y_r+y_2,y_2)P(z,y_r,y_r+y_2,y_r)\right)\\&-\mathbf{Y}_{r,2}\left(\frac{z^4M(z,0)}{(1-zM(z,0))^2}P(z,y_2,0,y_2)P(z,y_r,0,y_r)\right).\end{aligned}$$

Another case such that one green vertex is a common ancestor of the other two red-green vertices, is depicted in right of Figure 5.10. First, g_3 points to the one of red-green leaf then another directed edge connects this leaf to second red-green leaf in the Motzkin skeleton. The edge starting at latter red-green leaf may point to any vertex except on the paths ones. This yields the generating function

$$G^{rg}_{B_2}(z) =\mathbf{Y}_r\left(\frac{z^4}{(1-zM(z,y_r))^4}\right).$$

Consider the Motzkin skeletons depicted in Figure 5.11. For the first one, the

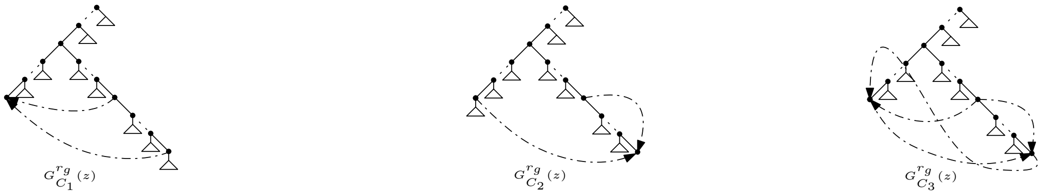


Figure 5.11: The structures of Motzkin skeletons that are correspondent with the Motzkin skeleton $G_C(z)$ by replacing respectively g_1, g_3 or both of them with red-green leaves.

generating function is given by

$$\begin{aligned}G^{rg}_{C_1}(z) = \mathbf{Y}_{r,3}\Bigg(&\frac{z^4\tilde{M}_3(z,y_r+y_3)}{(1-zM(z,y_r+y_3))^2}P(z,y_r,y_r+y_3,y_r)P(z,y_3,y_r+y_3,y_3)\\ &-\frac{z^4M(z,0)}{(1-zM(z,0))^2}P(z,y_r,0,y_r)P(z,y_3,0,y_3)\Bigg)\\ &+\mathbf{Y}_{r,2}\left(\frac{z^4M(z,y_2+y_r)}{(1-zM(z,y_r+y_2))}P(z,y_2,y_r+y_2,y_2)P(z,y_2,y_r+y_2,0)\right).\end{aligned}$$

For the Motzkin skeletons on the middle of Figure 5.11, we obtain

$$\begin{aligned}G^{rg}_{C_2}(z) = \mathbf{Y}_{r,2}&\left(\frac{z^4M(z,y_r+y_2)}{(1-zM(z,y_r+y_2))^2}P(z,y_2,y_r+y_2,y_2)P(z,y_2,y_r+y_2,0)\right)\\ &+\mathbf{Y}_{r,1}\Bigg(\frac{z^4\tilde{M}_1(z,y_1+y_r)}{1-zM(z,y_r+y_1)}P(z,y_1,y_r+y_1,y_1)P^{\star}(z,y_1,y_r+y_1,y_1)\\ &\times P(z,y_r,y_r+y_1,y_r)\\ &-\frac{z^4M(z,0)}{1-zM(z,0)}P(z,y_r,0,y_r)P(z,y_1,0,y_1)P^{\star}(z,y_1,0,y_1)\Bigg).\end{aligned}$$

For the right one, we will take two terms for exponential generating function depending on which red-green leaf is pointed by g_2 first. After all, we get from path analysis

$$G^{rg}_{C_3}(z) = \mathbf{Y}_r\left(\frac{z^4}{(1-zM(z,y_r))^3}P^{\star}(z,0,y_r,0)+\frac{z^4}{(1-zM(z,y_r))^4}\right).$$

The last case of general networks with at least one red-green vertex have Motzkin skeletons as shown in Figure 5.12. The restriction for the target vertex of the edges to be added at pointer set vertices follow the analogous rules in order to meet the

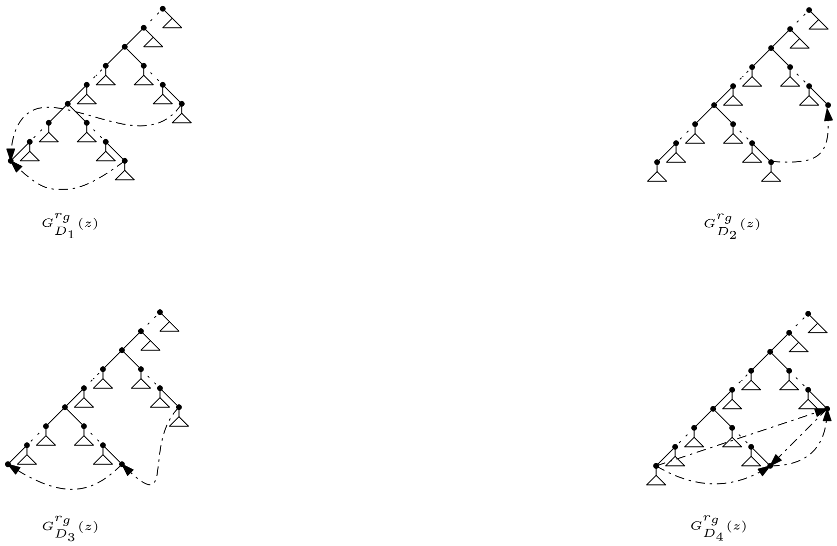


Figure 5.12: The Motzkin skeletons which arise from the $G_D(z)$ by considering contribution of all possible red-green leaves.

generality constraint. Setting up the generating function follows the same pattern as before. We omit now the details and get from path analysis after all

$$\begin{aligned}G^{rg}_{D_1}(z) = \mathbf{Y}_{r,3}&\Bigg(\frac{z^5\tilde{M}_3(z,y_r+y_3)M(z,y_r+y_3)}{1-zM(z,y_r+y_3)}P(z,y_r,y_r+y_3,y_r)P(z,y_3,y_r+y_3,y_3)^3\\&-\frac{z^5M(z,0)^2}{(1-zM(z,0))}P(z,y_r,0,y_r)P(z,y_3,0,y_3)^3\Bigg)\\&+\mathbf{Y}_{r,2}\Bigg(\frac{z^5\tilde{M}_2(z,y_2+y_r)M(z,y_2+y_r)}{(1-zM(z,y_r+y_2))^2}P(z,y_2,y_r+y_2,y_2)^2P(z,y_r,y_r+y_2,y_r)\\&-\frac{z^5M(z,0)^2}{(1-zM(z,0))^2}P(z,y_2,0,y_2)^2P(z,y_r,0,y_r)\Bigg).\end{aligned}$$

$$\begin{aligned}G^{rg}_{D_2}(z) = \mathbf{Y}_{r,1}&\Bigg(\frac{z^5\tilde{M}_1(z,y_r+y_1)M(z,y_r+y_1)}{(1-zM(z,y_r+y_1))^2}P(z,y_r,y_r+y_1,y_r)P(z,y_1,y_r+y_1,y_1)^2\\&-\frac{z^5M(z,0)^2}{(1-zM(z,0))^2}P(z,y_r,0,y_r)P(z,y_1,0,y_1)^2\Bigg).\\G^{rg}_{D_3}(z) = \mathbf{Y}_r&\left(\frac{z^5M(z,y_r)}{(1-zM(z,y_r))^5}\right).\\G^{rg}_{D_4}(z) = \mathbf{Y}_r&\left(\frac{z^5M(z,y_r)}{(1-zM(z,y_r))^5}+\frac{z^5M(z,y_r)}{(1-zM(z,y_r))^5}\right).\end{aligned}$$

In the end, we consider the Motzkin skeletons with the contribution double-green vertices as depicted in Figure 5.13. Note that, the extra factor $\frac{1}{2}$ appears in the expression of $G_E^2(z)$ and $G_E^3(z)$, because the order of pointing for double-green vertices is not matter. After normalization we obtain

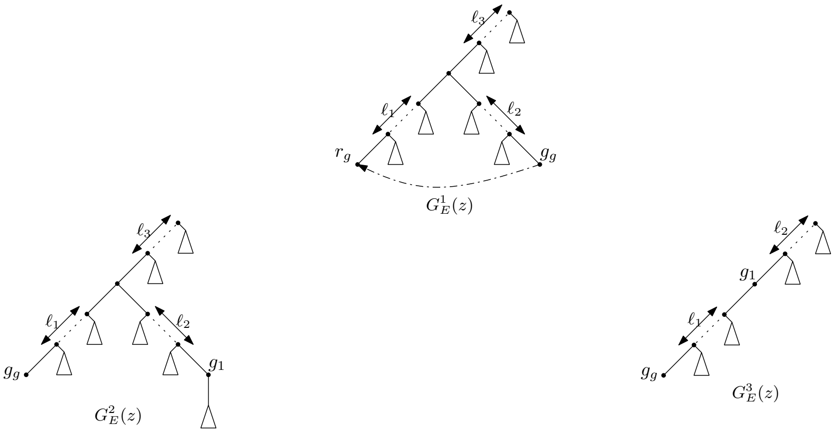


Figure 5.13: Possible structures of Motzkin skeletons with three reticulation vertices and contribution of double-green vertices.

$$
\begin{aligned}
G_E^1(z) &= \mathbf{Y}_{g,r} \frac{z^3 P(z, y_g, y_r + y_g, y_g)}{(1 - zM(z, y_g + y_r))^2}. \\
G_E^2(z) &= \frac{1}{2} (\mathbf{Y}_g)^2 \mathbf{Y}_1 \frac{z^3 \tilde{M}_1(z, y_1 + y_g)}{1 - zM(z, y_1 + y_g)} P(z, y_g, y_1 + y_g, y_g) P(z, y_1, y_g + y_r, y_1) \\
&\quad - \frac{1}{2} (\mathbf{Y}_g)^2 \mathbf{Y}_1 \frac{z^3 M(z, 0)}{1 - zM(z, 0)} P(z, y_1, 0, y_1) P(z, y_g, 0, y_g) \\
&\quad - \mathbf{Y}_g \frac{z^5 M(z, y_g)}{(1 - zM(z, y_g))^5}. \\
G_E^3(z) &= \frac{1}{2} (\mathbf{Y}_g)^2 \mathbf{Y}_1 \frac{z^2 P(z, y_1, y_1 + y_g, 0)}{1 - zM(z, y_1 + y_g)}.
\end{aligned}
$$

Now, we sum up all obtained generating functions so far. For normalization, the result must be divided by 8, Since the procedure will generate each general network eight times. Overall, by collecting everything, the exponential generating function for vertex-labelled general phylogenetic networks with three reticulation nodes is

$$
G_3^{\natural}(z) = z \cdot \frac{a_3^{\natural}(z^2) - b_3^{\natural}(z^2)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{11/2}},
$$

where

$$a_3^{\natural}(z) = 3z^6 + 2z^5 + 4z^4 + 2z^3 + \frac{69}{4}z^2,$$

and,

$$b_3^{\natural}(z) = z^5 + \frac{9}{2}z^4 + 11z^3 + \frac{69}{4}z^2.$$

Also, as before we can take the explicit formulas for vertex and leaf-labeled general networks with 3 reticulation vertices. To see them set $n = 2m+1$, so we have

$$[z^n]G_3^{\natural}(z) = [z^m]\bar{G}_3^{\natural}(z),$$

such that,

$$[z^m]\bar{G}_3^{\natural}(z) = [z^m]\frac{a_3^{\natural}(z)}{(1-2z)^{\frac{11}{2}}} - [z^m]\frac{b_3^{\natural}(z)}{(1-2z)^5}.$$

It gives

$$\mathfrak{F}^{\natural}(m) := [z^m]\bar{G}^{\natural}3(z) = \frac{2^{m-6}}{3}\Big(A_3^{\natural}(m)\frac{m(m-1)\binom{2m}{m}}{35(2m-1)4^{m-2}} - B_3^{\natural}(m)\Big).$$

where

$$\begin{aligned}
A_3^{\natural}(m) &= 104m^4 + 416m^3 + 596m^2 - 384m + 61,\\
B_3^{\natural}(m) &= 48m^4 + 31m^3 - 12m^2 - 73m + 6.
\end{aligned}$$

By replacing $m = (n-1)/2$ we have, $G_{3,n}^{\natural} = n! \cdot \mathfrak{F}^{\natural}((n-1)/2)$.
With some more steps but similar as before we can present explicit formula for the number of leaf-labeled general network with three reticulation vertices. Let $\dot{G}_3^{\natural}(z)$ denote corresponding generating function for general networks where the Equation (5.6) holds true for them and $\ddot{G}_3^{\natural}(z)$ be generating function for general networks which arise from the Motzkin skeletons Figure 5.14. We have $G_3^{\natural}(z) = \dot{G}_3^{\natural}(z) + \ddot{G}_3^{\natural}(z)$. So for the first subfamily (for $m > 3$) we get

$$\dot{\mathfrak{F}}^{\natural}(m) = \frac{2^{m-5}}{3}\Big(\dot{A}_3^{\natural}(m)\frac{m(m-1)\binom{2m}{m}}{35(2n-5)(2n-3)(2m-1)4^{m-2}} - \dot{B}_3^{\natural}(m)\Big), \tag{5.19}$$

where,

$$\dot{A}_3^{\natural}(m) = 280m^6 - 288m^5 - 1086m^4 - 2626m^3 + 9239m^2 - 7463m + 4290,$$

and

$$\dot{B}_3^{\natural}(m) = 24m^4 - \frac{31}{2}m^3 + 6m^2 + \frac{85}{2}m - 21.$$

Also for $m = 3$ ($\ell = 1$), we have $\dot{\mathfrak{F}}^{\natural}(3) = 8$.

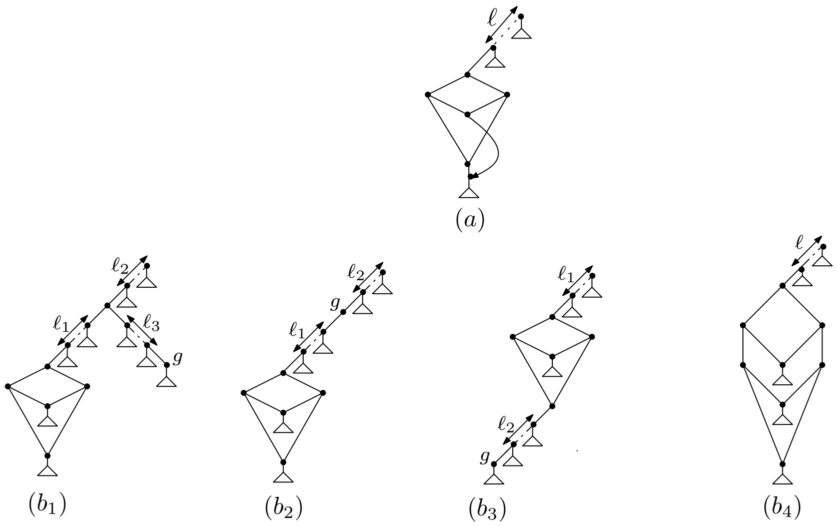


Figure 5.14: The family of general networks with correspondent generating function $\ddot{G}^{\natural}_{3}(z)$. Note that, applying procedure (5.2.1) for each fixed leaf-labeled general network of (a) can construct correspondent vertex-labeled general network four times. For leaf-labeled general network arose from second row figures, it would be exactly twice. Also, in the first step we complete the structure of figures b_1, b_2 and b_3 by adding one more directed edge from g to an unary red vertex.

Now we consider the family of general networks with 3 reticulation vertices such that there is a pair of vertices that have the set of same descendent and applying the procedure (5.2.1) needs to cope with symmetry for them; see Figure 5.14. First, we set up generating function, let's show it $G^{\natural}_{s_1}(z)$, for case (a) as shown at the top of figure 5.14. Each fixed leaf-labeled general network which is arisen from this structure can generate corresponding vertex-labeled networks four times. So for this case we normalize equation 5.6 by considering $G^{\natural}{}_{s_1,\ell} = 4\frac{\ell!}{n!}G^{\natural}_{s_1,n}$. Let $G^{\natural}_{s_2}(z)$ denote the corresponding generating function for second row structures of Figure 5.14. Note that each fixed leaf-labeled network belongs to these family can construct vertex-labeled network two times, so we get $G^{\natural}{}_{s_2,\ell} = 2\frac{\ell!}{n!}G^{\natural}_{s_2,n}$. Overall, we obtain the $\ddot{G}^{\natural}_{3}(z) = G^{\natural}_{s_1}(z) + G^{\natural}_{s_2}(z)$, where

$$G^{\natural}_{s_1}(z) = \frac{1}{4}\frac{z^6 M(z,0)}{1 - zM(z,0)},$$

and then we get

$$\ddot{\mathcal{F}}^{\natural}_{s_1}(m) := [z^m]\bar{G}^{\natural}_{s_1}(z) = 2^{m-2}\Bigg(\frac{m(m-1)\binom{2m}{m}}{(2n-3)(2m-1)4^m}\Bigg). \tag{5.20}$$

Also we have

$$\begin{aligned}G^{\text{ʯ}}_{s_2}(z) = &\frac{1}{2}\partial y \frac{z^7\tilde{M}(z,y)M(z,y)^2}{4(1-zM(z,y))^2}P(z,y,y,y)\\ &+\frac{1}{2}\partial y \frac{z^6 M(z,y)^2}{4(1-zM(z,y))}P(z,y,y,0)\\ &+\frac{1}{2}\partial y \frac{z^6\tilde{M}(z,y)M(z,y)}{2(1-zM(z,y))^2}\\ &+\frac{1}{4}\frac{z^8 M(z,0)^3}{1-zM(z,0)}.\end{aligned}$$

such that for or $m > 3$ ($\ell > 1$) we have

$$\ddot{\mathfrak{F}}^{\text{ʯ}}_{s_2}(m) := [z^m]\bar{G}^{\text{ʯ}}_{s_2}(z) = 2^{m-3}\Big((2m^3-15m^2+38m-34)\frac{m(m-1)\binom{2m}{m}}{(2m-5)(2m-3)(2m-1)4^{m-1}} - \frac{1}{2}(m-3)\Big), \tag{5.21}$$

and $\ddot{\mathfrak{F}}^{\text{ʯ}}_{s_2}(3) = 0$. Obviously, this means there is no any general network with one leaf which can be generated by second row structures of Figure 5.14. Overall, by collecting everything, we have $G^{\text{ʯ}}_{3,1} = 51$ and for $\ell > 1$ we use

$$\begin{aligned}G^{\text{ʯ}}_{3,\ell} &= \ell!\cdot\Big(\dot{\mathfrak{F}}^{\text{ʯ}}(\ell+2) + 4\ddot{F}^{\text{ʯ}}_{s_1}(\ell+2) + 2\ddot{F}^{\text{ʯ}}_{s_2}(\ell+2)\Big)\\ &= \ell!\,\Big(r_3(\ell)2^{-\ell}\binom{2\ell+4}{\ell+2} - 2^{\ell}p_3(\ell)\Big),\end{aligned} \tag{5.22}$$

where after manipulation we get $r_3(\ell)$ and $p_3(\ell)$ as show in Table 5.1 for the

	The explicit formula	
$G^{\text{ʯ}}_{1,\ell}$	$\ell!\,\Big(r_1(\ell)2^{-\ell}\binom{2\ell}{\ell} - 2^{\ell}p_1(\ell)\Big)$	$r_1(\ell) = \ell,$ and $p_1(\ell) = \frac{1}{2}.$
$G^{\text{ʯ}}_{2,\ell}$	$\ell!\,\Big(r_2(\ell)2^{-\ell}\binom{2\ell+2}{\ell+1} - 2^{\ell}p_2(\ell)\Big)$	$r_2(\ell) = \frac{(\ell+1)(6\ell^4+19\ell^3+18\ell^2-7\ell)}{2(6\ell-3)(2\ell+1)},$ and $p_2(\ell) = \frac{2\ell^2+5\ell+3}{2}.$
$G^{\text{ʯ}}_{3,\ell}$	$\ell!\,\Big(r_3(\ell)2^{-\ell}\binom{2\ell+4}{\ell+2} - 2^{\ell}p_3(\ell)\Big)$	$r_3(\ell) =$ $\frac{(\ell+1)(\ell+2)(280\ell^6+3072\ell^5+12834\ell^4+22386\ell^3+10949\ell^2-5211\ell-3990)}{840(2\ell+3)(2\ell+1)(2\ell-1)}.$ $p_3(\ell) = \frac{48\ell^4+415\ell^3+1326\ell^2+1799\ell+816}{768}.$

Table 5.1: The numbers of leaf-labeled general networks with ℓ leaves and no multiple edges.

number of leaf-labeled general networks with three reticulation vertices and no multiple edges. In the following, we want to set up exponential generating functions for general networks with three reticulation vertices and at least one multiple edges. It can be done by a case by case analysis of each sparsened skeletons which are depicted in Figures 5.15 to the 5.21. Note that, each factor of the expression makes up for the fact that each network is generated many times. So we use them to normalize counting values of each case separately.

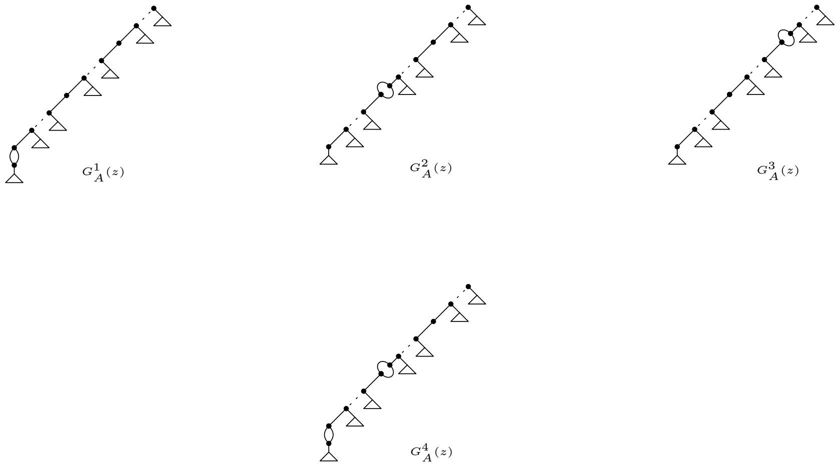


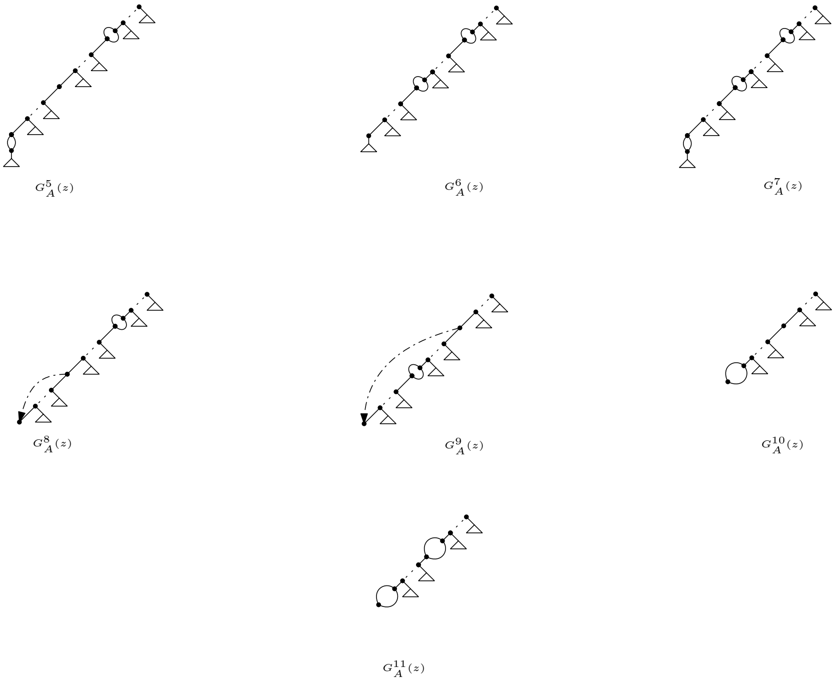
Figure 5.15: The structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G_A(z)$.

$$G^1_A(z) = \frac{1}{4}\mathbf{Y}_{2,3}\frac{z^4 M(z, y_2+y_3)}{1 - zM(z, y_2+y_3)}P(z, y_2+y_3, y_2+y_3, y_3)P(z, y_3, y_2+y_3, 0).$$

$$G^2_A(z) = \frac{1}{4}\mathbf{Y}_{1,3}\frac{z^4 \tilde{M}_1(z, y_1+y_3)}{1 - zM(z, y_1+y_3)}P(z, y_3, y_1+y_3, y_3)P(z, y_3, y_1+y_3, 0).$$

$$G^3_A(z) = \frac{1}{4}\mathbf{Y}_{1,2}\frac{z^4 \tilde{M}_1(z, y_1+y_2)}{(1 - zM(z, y_1+y_2))^2}P(z, y_2, y_1+y_2, 0).$$

$$G^4_A(z) = \frac{1}{2}\mathbf{Y}_{3}\frac{z^5 M(z, y_3)}{1 - zM(z, y_3)}P(z, y_3, y_3, y_3)P(z, y_3, y_3, 0).$$



$G_A^5(z)$

$G_A^6(z)$

$G_A^7(z)$

$G_A^8(z)$

$G_A^9(z)$

$G_A^{10}(z)$

$G_A^{11}(z)$

Figure 5.16: The rest of structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G_A(z)$.

$$G_A^5(z) = \frac{1}{2}\mathbf{Y}_2 \frac{z^5 M(z, y_2)}{(1 - zM(z, y_2))^2} P(z, y_2, y_2, 0).$$

$$G_A^6(z) = \frac{1}{2}\mathbf{Y}_1 \frac{z^5 \tilde{M}(z, y_1)}{(1 - zM(z, y_1))^3}.$$

$$G_A^7(z) = \frac{z^6 M(z, 0)}{(1 - zM(z, 0))^3}.$$

$$G_A^8(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^4 P^\star(z, 0, y_r, 0)}{(1 - zM(z, y_r))^2}.$$

$$G_A^9(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^4}{(1 - zM(z, y_r))^3}.$$

$$G_A^{10}(z) = \frac{1}{4}\mathbf{Y}_{r,3} \frac{z^3 P(z, y_3, y_r + y_3, 0)}{1 - zM(z, y_3 + y_r)}.$$

$$G_A^{11}(z) = \frac{1}{2}\mathbf{Y}_r \frac{z^4}{(1 - zM(z, y_r))^2}.$$

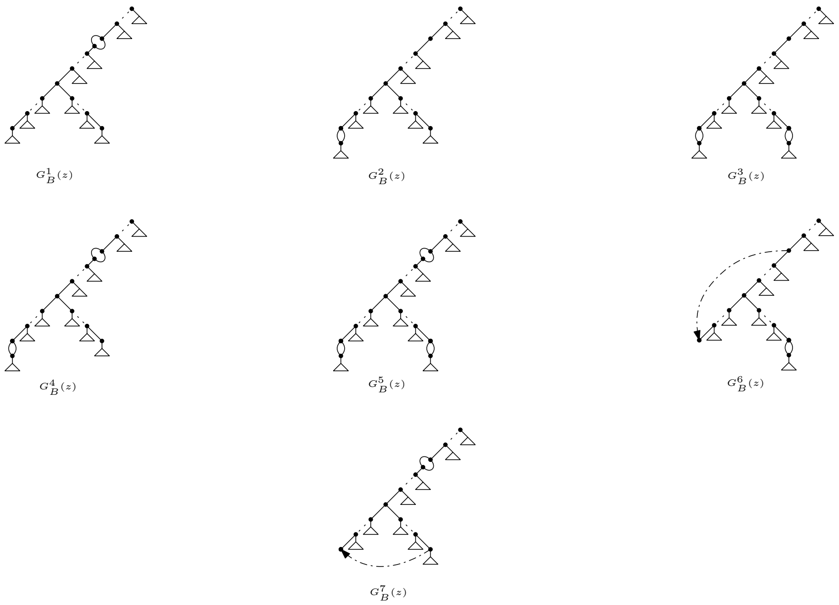


Figure 5.17: The structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G_B(z)$.

$$\begin{aligned}
G_B^1(z) =& \frac{1}{8}\mathbf{Y}_{1,2}\Bigg(\frac{z^5\tilde{M}_1(z,y_1+y_2)\tilde{M}_2(z,y_1+y_2)}{(1-zM(z,y_1+y_2))^2}P(z,y_1,y_1+y_2,y_1)P(z,y_2,y_1+y_2,y_2)\\
&-\frac{z^5M(z,0)^2}{(1-zM(z,0))^2}P(z,y_1,0,y_1)P(z,y_2,0,y_2)\Bigg).\\
G_B^2(z) =& \frac{1}{4}\mathbf{Y}_{2,3}\Bigg(\frac{z^5M(z,y_2+y_3)\tilde{M}_2(z,y_2+y_3)}{1-zM(z,y_2+y_3)}P(z,y_2+y_3,y_2+y_3,y_2+y_3)\\
&\times P(z,y_3,y_2+y_3,0)P(z,y_3,y_2+y_3,y_3)\Bigg).\\
G_B^3(z) =& \frac{1}{4}\mathbf{Y}_3\frac{z^6M(z,y_3)^2}{1-zM(z,y_3)}P(z,y_3,y_3,y_3)^2P(z,y_3,y_3,0).\\
G_B^4(z) =& \frac{1}{2}\mathbf{Y}_2\frac{z^6\tilde{M}_2(z,y_2)M(z,y_2)}{(1-zM(z,y_2))^3}P(z,y_2,y_2,y_2).\\
G_B^5(z) =& \frac{1}{2}\frac{z^7M(z,0)^2}{(1-zM(z,0))^4}.\\
G_B^6(z) =& \frac{1}{4}\mathbf{Y}_r\frac{z^5M(z,y_r)}{(1-zM(z,y_r))^3}P(z,y_r,y_r,y_r).\\
G_B^7(z) =& \frac{1}{4}\mathbf{Y}_r\frac{z^5M(z,y_r)}{(1-zM(z,y_r))^4}.
\end{aligned}$$

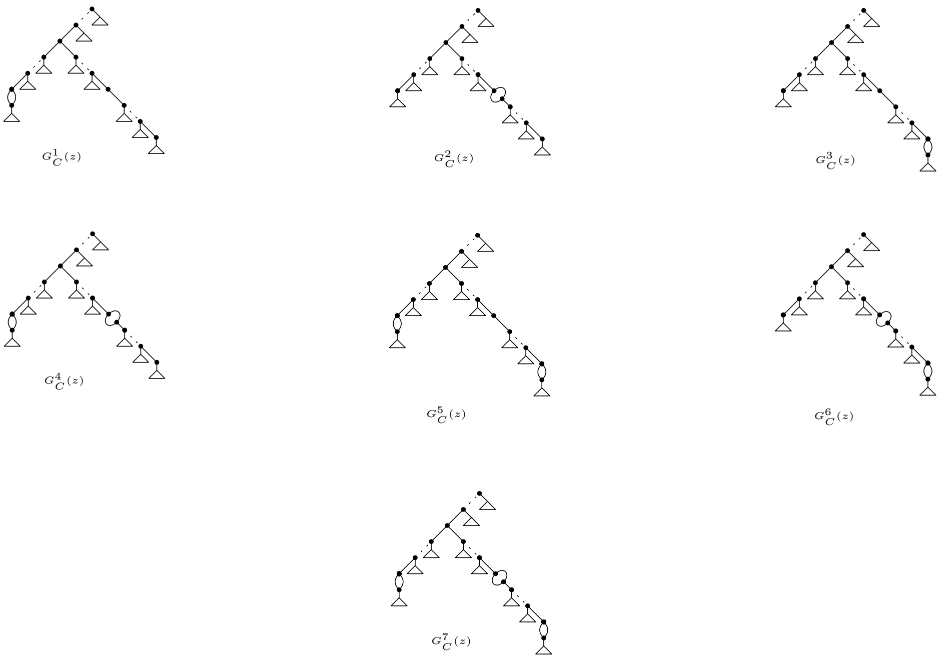


Figure 5.18: The structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G_C(z)$.

$$G_C^1(z)=\frac{1}{4}\mathbf{Y}_{2,3}\frac{z^5M(z,y_2+y_3)\tilde{M}_3(z,y_2+y_3)}{(1-zM(z,y_2+y_3))^2}P(z,y_2+y_3,y_2+y_3,y_2+y_3)P(z,y_2,y_2+y_3,0).$$

$$\begin{aligned}G_C^2(z)=\frac{1}{4}\mathbf{Y}_{1,3}&\left(\frac{z^5\tilde{M}_1(z,y_3+y_1)\tilde{M}_3(z,y_3+y_1)}{1-zM(z,y_3+y_1)}P(z,y_1,y_3+y_1,y_1)^2P(z,y_3,y_3+y_1,y_3)\right.\\&\left.-\frac{z^5M(z,0)^2}{1-zM(z,0)}P(z,y_1,0,y_1)^2P(z,y_3,0,y_3)\right).\end{aligned}$$

$$\begin{aligned}G_C^3(z)=\frac{1}{4}\mathbf{Y}_{1,2}&\left(\frac{z^5\tilde{M}_1(z,y_2+y_1)M(z,y_2+y_1)}{1-zM(z,y_2+y_1)}P(z,y_1,y_2+y_1,y_1)\right.\\&\times P(z,y_1+y_2,y_1+y_2,y_1)P(z,y_2,y_2+y_1,y_2)\\&\left.-\frac{z^5M(z,0)^2}{(1-zM(z,0))^2}P(z,y_1,0,y_1)P(z,y_2,0,y_2)\right).\end{aligned}$$

$$G_C^4(z)=\frac{1}{2}\mathbf{Y}_3\frac{z^6\tilde{M}_3(z,y_3)M(z,y_3)}{(1-zM(z,y_3))^3}P(z,y_3,y_3,y_3).$$

$$G_C^5(z)=\frac{1}{2}\mathbf{Y}_2\frac{z^6M(z,y_2)^2}{(1-zM(z,y_2))^2}P(z,y_2,y_2,y_2)P(z,y_2,y_2,0).$$

$$G_C^6(z)=\frac{1}{2}\mathbf{Y}_1\frac{z^6\tilde{M}_1(z,y_1)M(z,y_1)}{(1-zM(z,y_1))^2}P(z,y_1,y_1,y_1)^2.$$

$$G_C^7(z)=\frac{z^7M(z,0)^2}{(1-zM(z,0))^4}.$$

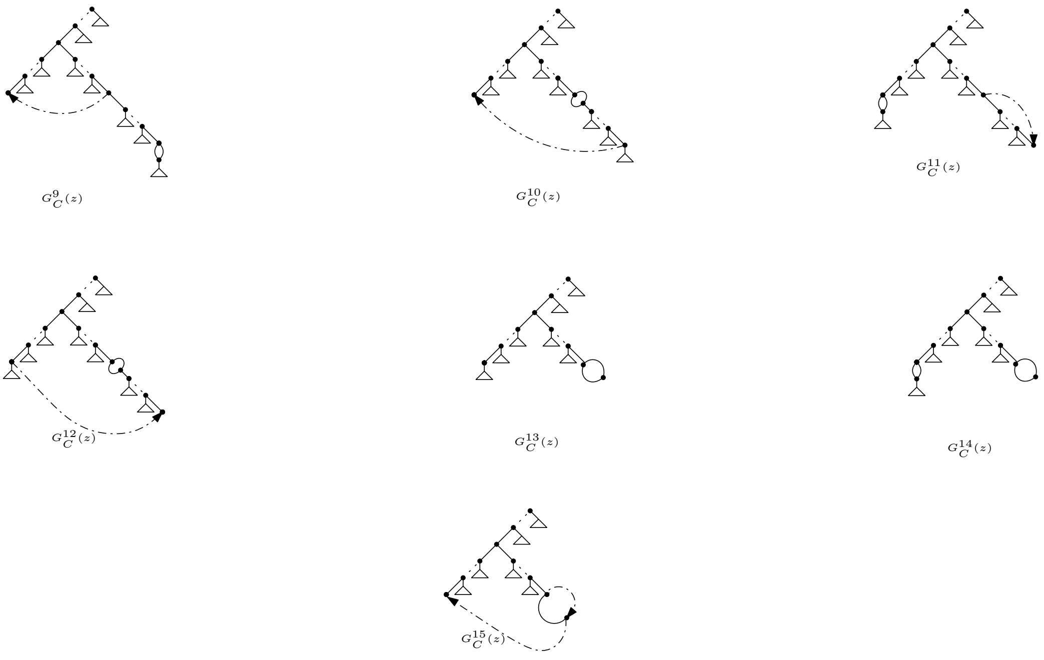


Figure 5.19: The structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G_C(z)$.

$$G_C^9(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^5 M(z,y_r)}{(1-zM(z,y_r))^3} P(z,y_r,y_r,y_r).$$

$$G_C^{10}(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^5 M(z,y_r)}{(1-zM(z,y_r))^4}.$$

$$G_C^{11}(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^5 M(z,y_r)}{(1-zM(z,y_r))^2} P(z,y_r,y_r,y_r) P^\star(z,0,y_r,0).$$

$$G_C^{12}(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^5 M(z,y_r)}{(1-zM(z,y_r))^4}.$$

$$G_C^{13}(z) = \frac{1}{4}\mathbf{Y}_{r,1} \left(\frac{z^4 \tilde{M}_1(z,y_1+y_r)}{1-zM(z,y_1+y_r)} P(z,y_1,y_1+y_r,y_1) P(z,y_r,y_1+y_r,y_r) \right.$$
$$\left. - \frac{M(z,0)}{1-zM(z,0)} P(z,y_1,0,y_1) P(z,y_r,0,y_r) \right).$$

$$G_C^{14}(z) = \frac{1}{2}\mathbf{Y}_r \left(\frac{z^5 M(z,y_r)}{(1-zM(z,y_r))^2} P(z,y_r,y_r,y_r) \right).$$

$$G_C^{15}(z) = \frac{1}{4}\mathbf{Y}_r \frac{z^4}{(1-zM(z,y_r))^3}.$$

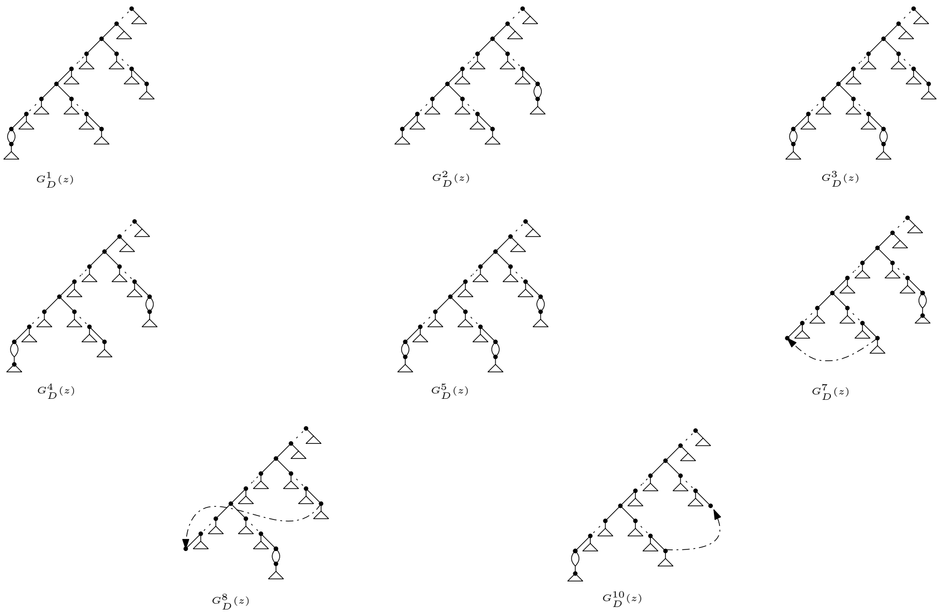


Figure 5.20: The structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G_D(z)$.

$$G_D^1(z) = \frac{1}{4}\mathbf{Y}_{2,3}\left(\frac{z^6 M(z,y_2+y_3)\tilde{M}_2(z,y_2+y_3)\tilde{M}_3(z,y_2+y_3)}{1-zM(z,y_2+y_3)}P(z,y_2+y_3,y_2+y_3,y_2+y_3)\right.$$
$$\left.\times P(z,y_2,y_2+y_3,y_2)P(z,y_3,y_2+y_3,y_3)^2 - \frac{z^6 M(z,0)^3}{(1-zM(z,0))^2}P(z,y_2,0,y_2)P(z,y_3,0,y_3)^2\right).$$
$$G_D^2(z) = \frac{1}{8}\mathbf{Y}_{1,2}\left(\frac{z^6 M(z,y_1+y_2)\tilde{M}_1(z,y_1+y_2)\tilde{M}_2(z,y_1+y_2)}{(1-zM(z,y_1+y_2))^2}P(z,y_1+y_2,y_1+y_2,y_1+y_2)\right.$$
$$\left.\times P(z,y_1,y_1+y_2,y_1)P(z,y_2,y_1+y_2,y_2) - \frac{z^6 M(z,0)^3}{(1-zM(z,0))^3}P(z,y_1,0,y_1)P(z,y_2,0,y_2)\right).$$
$$G_D^3(z) = \frac{1}{4}\mathbf{Y}_3\frac{z^7 M(z,y_3)^2\tilde{M}_3(z,y_3)}{(1-zM(z,y_3))^2}P(z,y_3,y_3,y_3)^3.$$
$$G_D^4(z) = \frac{1}{2}\mathbf{Y}_2\frac{z^7 M(z,y_2)^2\tilde{M}_2(z,y_2)}{(1-zM(z,y_2))^3}P(z,y_2,y_2,y_2)^2.$$
$$G_D^5(z) = \frac{1}{2}\frac{z^8 M(z,0)^3}{(1-zM(z,0))^5}.$$
$$G_D^6(z) = \frac{1}{4}\mathbf{Y}_r\left(\frac{z^6 M(z,y_r)^2}{(1-zM(z,y_r))^4}P(z,y_r,y_r,y_r)\right).$$
$$G_D^7(z) = \frac{1}{4}\mathbf{Y}_r\left(\frac{z^6 M(z,y_r)^2}{(1-zM(z,y_r))^4}P(z,y_r,y_r,y_r)\right).$$
$$G_D^8(z) = \frac{1}{4}\mathbf{Y}_r\left(\frac{z^6 M(z,y_r)^2}{(1-zM(z,y_r))^4}P(z,y_r,y_r,y_r)\right).$$

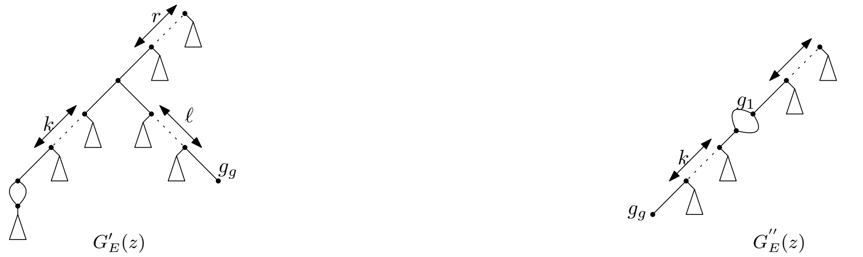


Figure 5.21: The structures of the Motzkin skeletons of general phylogenetic networks with at least one multiple edges which are arised from $G^2_E(z)$ and $G^3_E(z)$.

$$G'_E(z) = \frac{1}{8}(\mathbf{Y}_g)^2 \frac{z^4 M(z, y_g)}{(1 - zM(z, y_g))^2} P(z, y_g, y_g, y_g).$$
$$G''_E(z) = \frac{1}{8}(\mathbf{Y}_g)^2 \frac{z^3}{(1 - zM(z, y_g))^2}.$$

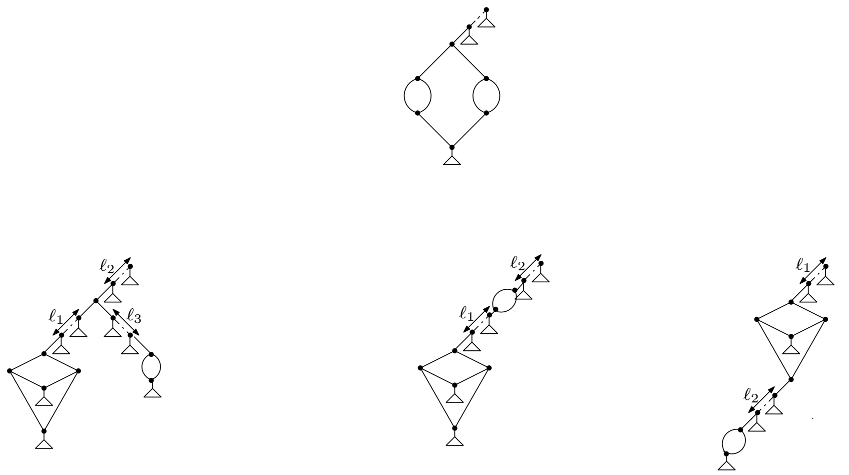


Figure 5.22: General networks with multiple edges and corresponding generating function $\ddot{G}^{\shortparallel}_3(z)$ such that any fixed leaf-labeled of them can generate all-vertex labeled exactly twice.

Overall, by collecting everything, we obtain the following result.

$$G^{\shortparallel}_3(z) = z \cdot \frac{a^{\shortparallel}_3(z^2) - b^{\shortparallel}_3(z^2)\sqrt{1 - 2z^2}}{(1 - 2z^2)^{11/2}},$$

where

$$a_3^{\shortparallel}(z) = z^5 - z^4 + \frac{13}{2}z^3 + 10z^2 \quad \text{and,} \quad b_3^{\shortparallel}(z) = 4z^3 + 10z^2.$$

After some computation, it gives

$$\mathcal{F}^{\shortparallel}(m) := [z^m]\bar{G}^{\shortparallel}{}_3(z) = 2^{m-1}\Big(A_3^{\shortparallel}(m)\frac{m(m-1)\binom{2m}{m}}{3(2m-1)4^m} - B_3^{\shortparallel}(m)\Big). \tag{5.23}$$

where

$$A_3^{\shortparallel}(m) = 6m^3 + 4m^2 - m - 2 \quad \text{and,} \quad B_3^{\shortparallel}(m) = m^3 - \frac{1}{2}m^2 - \frac{1}{2}.$$

By replacing $m = (n-1)/2$ we have, $G_{3,n}^{\shortparallel} = n! \cdot \mathcal{F}^{\shortparallel}((n-1)/2)$ for the number of vertex-labeled general phylogenetics with 3 reticulation vertices and at least one multiple edge in their structures.

Now we set up generating function for leaf-labeled. we Consider $G^{\shortparallel}{}_3(z) = \dot{G}_3^{\shortparallel}(z) + \ddot{G}_3^{\shortparallel}(z)$ which respectively right side of equation denote generating functions for two subfamilies of this class (general networks with multiple edges) that we can use the equation directly or not (needs to cope with symmetry); see Figure 5.22. For the first subfamily we get

$$\dot{\mathcal{F}}^{\shortparallel}(m) := [z^m]\dot{\bar{G}}_3^{\nparallel}(z) = 2^{m-2}\Big(\dot{A}_3^{\nparallel}(m)\frac{m(m-1)\binom{2m}{m}}{3(2n-3)(2m-1)4^{m-1}} - \dot{B}_3^{\nparallel}(m)\Big).$$

where,

$$\dot{A}_3^{\shortparallel}(m) = 6m^4 - 5m^3 - 7m^2 - 2m + 6 \quad \text{and,} \quad \dot{B}_3^{\shortparallel}(m) = 2m^3 - m^2 - m. \tag{5.24}$$

Also the generating function corresponding to the general networks in Figure 5.22 is

$$\begin{aligned}\ddot{G}_3^{\shortparallel}(z) &= \frac{1}{2}\frac{z^6M(z,0)}{1-zM(z,0)} + \frac{1}{4}\frac{z^8M(z,0)^3}{(1-zM(z,0))^3} + \frac{1}{4}\frac{z^7M(z,0)^2}{(1-zM(z,0))^2} + \frac{1}{2}\frac{z^7M(z,0)^2}{(1-zM(z,0))^2}\\ &= \frac{1}{2}\frac{z^3}{(1-2z^2)^{\frac{3}{2}}},\end{aligned}$$

such that

$$\ddot{\mathfrak{F}}^{\shortparallel}(m) := [z^m]\ddot{\tilde{G}}_3^{\nparallel}(z) = 2^{m-1}m(m-1)(m-2)\Big(\frac{\binom{2m}{m}}{(2n-3)(2m-1)4^m}\Big). \tag{5.25}$$

Note that, every member of leaf-labeled general networks arising from Figure 5.22 construct corresponding vertex-labeled networks twice. Overall, by replacing $m = \ell + 2$ we have

$$\begin{aligned} G_{3,\ell}^{\shortparallel} &= \ell! \cdot \Big(\dot{\mathfrak{F}}^{\shortparallel}(\ell+2) + 2\ddot{\mathfrak{F}}^{\shortparallel}(\ell+2)\Big) \\ &= \ell! \cdot 2^{\ell} \cdot \Big(\frac{(\ell+1)(\ell+2)^2(6\ell^3+31\ell^2+45\ell+15)\binom{2\ell+4}{\ell+2}}{3(2\ell+1)(2\ell+3)4^{\ell+1}} - (2\ell^3+11\ell^2+19\ell+10)\Big), \end{aligned}$$

for the number of leaf-labeled general networks with three reticulation vertices and at least one multiple edge. Finally, we have $\tilde{G}_{3,\ell} = G_{3,\ell}^{\nparallel} + G_{3,\ell}^{\shortparallel}$, for the number of general phylogenetic networks with three reticulation vertices.

Now, the defined structure for paths of sparsened skeletons with a considered specification for attached trees on them, capable us to prove the theorem 5.0.3.

Proof of Theorem 5.0.3. In particular note that function $G(z, y)$ is the form zM (4.3), which z refers to vertices lie on the pathes of sparsened skeleton.

$$G(z, y) = a(z, y) - b(z, y)\sqrt{1 + (y^2-2)z^2 - 2zy}, \tag{5.26}$$

where $a(z, y), b(z, y)$ are polynomials in z and y with $a(z, 0) = b(z, 0) = 1$. In summary, we have exponential generating function G_k for phylogenetic network in sum of terms of the form

$$\partial_{y_1} \cdots \partial_{y_k} \frac{G_1(z, y) \cdots G_s(z, y)}{(1 - G_{s+1}(z, y)) \cdots (1 - G_{s+t}(z, y))}\Big|_{y_1=0,\ldots,y_k=0}, \tag{5.27}$$

Note that in this expression, numerator refers to generating function of subtrees which rooted at green vertices. The denominator refers to sequences of subtrees which rooted the vertices on the paths of sparsened skeleton. Also where the number of functions $G_{s+i}(z, y)$ is bounded by the number of edges of the sparsened skeleton increased by one (for the sequence of trees added above the root when constructing the Motzkin skeletons). Now, recall lemma 4.3.7 from previous chapter which can be used for any similar structures as $G(z, y)$. We can apply this lemma after expanding (5.27) and obtain that

$$G_k(z) = \frac{a_k(z) - b_k(z)\sqrt{1-2z^2}}{(1-2z^2)^p}. \tag{5.28}$$

We proceed to show that $p = 2k - 1/2$. For that, observe (5.27) without the derivatives is of the general form given in (5.28) with the exponent of the denominator equals $t/2$ which reaches its maximum for the sparsened skeleton with the maximal number of edges and is thus at most $k - 1/2$. Also, from the above lemma, we see that each differentiation increases the exponent by 1. Thus, the exponent of (5.27) when written as (5.28) is at most $2k - 1/2$. Adding up this terms gives

$$G_k(z) = \frac{a_k(z) - b_k(z)\sqrt{1-2z^2}}{(1-2z^2)^{2k-1/2}},$$

where $a_k(z)$ and $b_k(z)$ are suitable polynomials. Let $G_{k,n}$ denote the number of vertex-labeled general phylogenetic networks with n vertices and k reticulation vertices. If n is even then $G_{k,n}$ is zero, otherwise there is a positive constant d_k such that as $n \to \infty$,

$$G_{k,n} = n![z^n]G_k(z) \sim d_k \left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-1},$$

Where by singularity analysis and Stirling's formula we get

$$d_k = \frac{2\sqrt{2\pi}a_k(1/\sqrt{2})}{4^k\Gamma(2k-1/2)}.$$

□

Remark. For the positivity claim, we already see in chapter 4 that corresponding constant $\tilde{d}_k$ for normal and tree-child networks is positive which is lower bound of d_k for general networks.

Proposition 5.4.3. *For the numbers of vertex-labeled general phylogenetic networks $G_{k,n}$ and vertex-labeled tree-child networks $T_{k,n}$,*

$$G_{k,n} = T_{k,n}\left(1 + \mathcal{O}(\frac{1}{n})\right), \tag{5.29}$$

as $n \to \infty$.

Proof. First, observe that $G_{k,n} - T_{k,n}$ is bounded by the number of networks which arise from all types of Motzkin skeletons where for each green vertex, the considered all possibilities of adding an edge violates the tree-child condition. As similar before (see,4.4.4), the largest number will come from the sparsened skeletons where all pointer vertices are the leaves. Now, fix such a type of Motzkin skeletons and one of its green vertices. Then, for this vertex, we will have the following options.

- The green vertex points to the root of the subtree which is attached to the one of green vertices in the Motzkin skeletons. Note that if it points to the root of its subtree, tree-child condition violates by making multiple edge. For the exponential generating function this gives

  $$\partial_{y_2}\cdots\partial_{y_k}\frac{G_1(z,y)\cdots G_s(z,y)}{(1-G_{s+1}(z,y))\cdots(1-G_{s+2k-1}(z,y))}\Big|_{y_2=0,\ldots,y_k=0},$$

  Here, and below y is the sum of y_i's with $2 \le i \le k$ and not all of the y_i's must be present; also which are present can differ from one occurrence to the next.

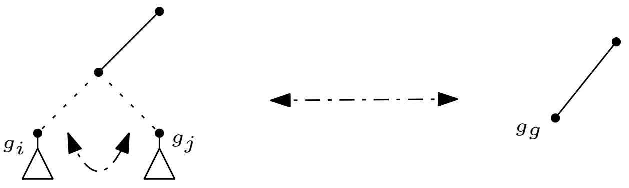
- There is a red-green vertex on the Motzkin skeleton. Note that the red-green property entails that one another pointer vertex joints to this leaf by adding directed edge which reduces the number of the derivative by one. Then we get

  $$\partial_{y_2}\cdots\partial_{y_k}\frac{G_1(z,y)\cdots G_{s-1}(z,y)}{(1-G_{s+1}(z,y))\cdots(1-G_{s+2k-1}(z,y))}\Big|_{y_2=0,\ldots,y_k=0}.$$

- There is double-green vertex in the Motzkin skeleton that points to the branches of sparsened skeleton. Then, we have

  $$\partial_{y_3}\cdots\partial_{y_k}\frac{G_1(z,y)\cdots G_{s-2}(z,y)}{2\cdot(1-G_{s+1}(z,y))\cdots(1-G_{s+2k-1}(z,y))}\Big|_{y_2=0,\ldots,y_k=0}.$$

  The existence of double green node in considered skeleton is like that two green vertices are merged to each others. Consequently, the number of edges reduce by two, which also leads to a contribution of smaller order.



The exponential generating function of all networks arising from these Motzkin skeletons and the pointer vertices are a sum of generating functions of the above three types. Thus, we obtain that this generating function has the form

$$\frac{c(z)-d(z)\sqrt{1-2z^2}}{(1-2z^2)^p},$$

where $c(z)$ and $d(z)$ are suitable polynomials and the maximum of p is as follows: note that without the derivatives in the above expressions, p would be at most $k - 1/2$. Also, because of above lemma, each derivative increases this bound by one. Thus, p is at most $2k - \frac{3}{2}$.

Now, we obtain that the exponential generating function of the above number has the form

$$\frac{c(z) - d(z)\sqrt{1-2z^2}}{(1-2z^2)^{2k-\frac{3}{2}}},$$

where $c(z)$ and $d(z)$ are suitable polynomials. Singularity analysis gives then the bound

$$\mathcal{O}\left(\left(\frac{\sqrt{2}}{e}\right)^n n^{n+2k-2}\right).$$

Summing over all possible type of Motzkin skeletons and all green vertices, we obtain the claimed result. □

Vertex Labeled Phylogenetic Networks	k=1		k=2		k=3	
	c_1	c'_1	c_2	c'_2	c_3	c'_3
$N_{k,n}$	$\frac{\sqrt{2}}{2}$	$-\frac{3\sqrt{\pi}}{2}$	$\frac{\sqrt{2}}{16}$	$-\frac{3\sqrt{\pi}}{8}$	$\frac{\sqrt{2}}{192}$	$-\frac{3\sqrt{\pi}}{64}$
$T_{k,n}$	$\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{\pi}}{2}$	$\frac{\sqrt{2}}{16}$	$-\frac{\sqrt{\pi}}{8}$	$\frac{\sqrt{2}}{192}$	$-\frac{\sqrt{\pi}}{64}$
$G_{k,n}$	$\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{\pi}}{2}$	$\frac{\sqrt{2}}{16}$	$-\frac{\sqrt{\pi}}{8}$	$\frac{\sqrt{2}}{192}$	$-\frac{\sqrt{\pi}}{64}$

Table 5.2: The first two asymptotic orders of normal, tree-child and general phylogenetic networks with at most 3 reticulation vertices. For all of them the first coefficient is same.

5.4.1 Asymptotic counting of leaf-labeled general phylogenetic Networks

In this part we want to prove Theorem 5.0.4 and argue that for the number of leaf-labeled general phylogenetic networks with $k \geq 1$ reticulation vertices (as like leaf-labeled tree-child ($\tilde{T}_{k,\ell}$) and normal networks) we can use,

$$\tilde{G}_{k,\ell} \sim 2^{3k-1} d_k \left(\frac{2}{e}\right)^{\ell} \ell^{\ell+2k-1}, \qquad (\ell \to \infty) \tag{5.30}$$

as a relative precise estimate of leaf-labeled general phylogenetic network, where d_k is as in Theorem 5.0.3.

It is enough to show that the number of a subfamily $\mathcal{G}$ of general networks such that some of their vertices have the same set of descendant are rare. Indeed, $\mathcal{G}$ consists of general networks that equation 5.6 can not be used directly for them. It's because of that in the described method (see 5.2.1) some of the fix leaf-labeled networks generate vertex-labeled networks more than one. In other words, this condition (a pair of vertices with a set of the same descendant) is necessary condition but not sufficient for the repeated generation of vertex-labeled networks. For instance, consider a leaf which is attached edge (u, g_i) in Figure 5.23 (a). Though, g_1 and g_2 have a set of the same descendant but applying the procedure (5.2.1), generates each vertex-labeled uniquely.

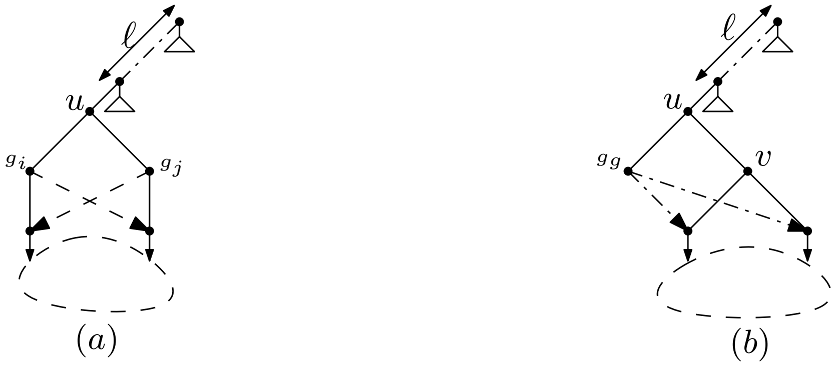


Figure 5.23: The structures of general phylogenetic networks where pair of vertices have a same descendant set after adding the directed edges in Motzkin skeletons.

Proof of Theorem 5.0.4. Consider a subfamily $\mathcal{G}$ as similar before. It is sufficient for our purposes to show that when $\ell \to \infty$, the number of these networks are asymptotically negligible. Assume, without loss of generality, these networks are without multiple edges because each of them reduces the number of differentiations in the expression of the exponential generating function by one, that causes the contribution of lower-order.

Note that, $\mathcal{G}$ is bounded by the number of networks which arise from two types of Motzkin skeletons that are depicted in Figure 5.23. First, when two green vertices point to the child vertices of each others (Figure 5.23, (a)) and second, a double-green vertex points unary vertices with the same parent (b). In the former case, two green vertices and in the later case double-green vertex with vertex v have set of the same descendant. Note that in each of described cases, the number of derivatives and consequently, the power of denominator in exponential generating function will be reduce by two. So The first two asymptotic orders are as in theorem 5.0.3. That implies

$$G_{k,2\ell+2k-1} \sim \binom{2\ell+2k-1}{\ell}(\ell+2k-1)!\tilde{G}_{k,\ell}. \tag{5.31}$$

Now we have $\tilde{G}_{k,\ell} \sim \frac{\ell!}{(2\ell+2k-1)!}G_{k,2\ell+2k-1}$, which an asymptotic result (5.30) follows by Theorem 5.0.3 and Stirling's formula.

□

Chapter 6

Future work

In the course of this thesis, we considered the counting problem of phylogenetic networks which is largely unsolved. The results rely heavily on analytic combinatorics [20]. We devised an approach, based on generating functions and analytic combinatorics, to solve this problem for some important subclasses of phylogenetic networks. In the following, we discuss some possible directions for future researches.

A precise look at the framework of chapter 3 for level-1 and level-2 networks reveals that it can also be used to derive *uniform random generators* (for example with the recursive method [22] or with a Boltzmann sampler [15]) directly from the specifications of the classes of phylogenetic networks given there. This could be useful for applications in bioinformatics, especially to generate simulated data in order to evaluate the speed or the quality of the output of algorithms dealing with phylogenetic networks. Also, we are confident that one could adapt the methods from Chapter 3 to level-k networks for $k > 2$ too, but for this some further work has to be done. A successful analysis of the case of level-1 and level-2 would constitute an important step, as it would open the way to the study of this kind of families of phylogenetic networks.

The results presented in Chapter 4 and 5 may facilitate improvements in the future studies for phylogenetic networks with fixed number of reticulation vertices as the size of the network tends to infinity. The latter restriction is necessary for our method to work. Indeed, the combinatorial setup we developed in this thesis is the construction of a sequence of combinatorial classes (for each given number of reticulation vertices, we contruct a separate class). The actual distribution of the reticulation vertices is then – on the level of generating functions – realized by differentiations. Letting k tend to infinity, when n tends to infinity, means that we have to cope with a growing number of differentiations and it is not clear how this changes the qualitative nature of the generating function. We certainly cannot expect that $N_k(z)$ keeps of the shape (5.28) when k depends on n and gets large with growing n. Thus, we have to leave the question of counting phylogenetic networks when k is allowed to grow with n open.

Apart from this, the next step would be to carry the study of the distribution of the number of reticulation vertices of a given phylogenetic networks over to the random case.

Bibliography

[1] N. Alexeev and M. A. Alekseyev. Combinatorial scoring of phylogenetic networks. *Computing and Combinatorics, Lecture Notes in Comput. Sci*, 9797:560–572, 2016.

[2] M. Bóna. On the number of vertices of each rank in k-phylogenetic trees. *Discrete Math. Theor. Comput. Sci.*, 18(3):7, 2016.

[3] M. Bóna and P. Flajolet. Isomorphism and symmetries in random phylogenetic trees. *J. Appl. Probab.*, 46(4):1005–1019, 2009.

[4] M. Bordewich and C. Semple. Determining phylogenetic networks from inter-taxa distances. *J. Math. Biol.*, 73(2):283–303, 2016.

[5] M. Bordewich and C. Semple. Reticulation-visible networks. *Adv. in Appl. Math.*, 78:114–141, 2016.

[6] Mathilde Bouvel, Philippe Gambette, and Marefatollah Mansouri. Counting phylogenetic networks of level 1 and 2. *arXiv:1909.10460*, 2020.

[7] C. Semple C. McDiarmid and D. Welsh. Counting phylogenetic networks. *Ann. Comb.*, 19:205–224, 2015.

[8] Gabriel Cardona and Louxin Zhang. Counting tree-child networks and their subclasses. *arXiv:1908.01917*, 2020.

[9] Kuang-Yu Chang, Wing-Kai Hon, and Sharma V. Thankachan. Compact encoding for galled-trees and its applications. *2018 Data Compression Conference*, pages 297–306, 2018.

[10] Z.-Z. Chen and L. Wang. Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 9(2):372–384, 2012.

[11] C. Langley D. Gusfield, S. Eddhu. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinf. Comput. Biol.*, 2:173–213, 2004.

[12] Francis Darwin. The life and letters of charles darwin, including an autobiographical chapter. *F. Darwin, ed., John Murray, London*, 1887.

[13] F. Disanto and N. A. Rosenberg. Enumeration of ancestral configurations for matching gene trees and species trees. *J. Comput. Biol.*, 24(9):831–850, 2017.

[14] M. Drmota. Random trees: An interplay between combinatorics and probability. *Springer, Wien-New York*, 2009.

[15] Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. A boltzmann samplers for the random generation of combinatorial structures. *Combinatorics, Probability and Computing*, 13:577–625, 2004.

[16] V. Johnson É. Czabarka, P. L. Erdős and V. Moulton. Generating functions for multi-labeled trees. *Discrete Appl. Math.*, 161:107–117, 2013.

[17] P. Flajolet and A. Odlyzko. The average height of binary trees and other simple trees. *J. Comput. System Sci.*, 25:171–213, 1982.

[18] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3(2):216–240, 1990.

[19] P. Flajolet and H. Prodinger. Register allocation for unary-binary trees. *SIAM J. Comput.*, 15(3):629–640, 1986.

[20] P. Flajolet and R. Sedgewick. Analytic combinatorics. *Cambridge University Press*, (1), 2009.

[21] P. Flajolet and J.-M. Steyaert. On the analysis of tree-matching algorithms. *Automata, Languages and Programming (Proc. Seventh Internat. Colloq., Noordwijkerhout), Springer, Berlin-New York*, 85:208–219, 1980.

[22] Philippe Flajolet, Paul Zimmermann, and Bernard Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theoretical Computer Science*, 132((1-2):1–35, 1994.

[23] L. R. Foulds and R. W. Robinson. Determining the asymptotic number of phylogenetic trees,. *Combinatorial Mathematics, VII (Proc. Seventh Australian Conf., Univ. Newcastle, Newcastle, Springer, Berlin*, 829:110–126, 1980.

[24] L. R. Foulds and R. W. Robinson. Enumerating phylogenetic trees with multiple labels. *Proceedings of the First Japan Conference on Graph Theory and Applications*, 72:129–139, 1988.

[25] A. R. Francis and M. Steel. Tree-like reticulation networks—when do tree-like distances also support reticulate evolution? *Math. Biosci.*, 259:12–19, 2015.

[26] Michael Fuchs, Bernhard Gittenberger, and Marefatollah Mansouri. Counting phylogenetic networks with few reticulation vertices: Tree-child and normal networks. *Australasian Journal of Combinatorics*, 73(2):385–423, 2018.

[27] F. Rossello G. Cardona and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Trans. Comput. Biol. Bioinform*, 6:552–569, 2009.

[28] Philippe Gambette, Vincent Berry, and Christophe Paul. The structure of level-k phylogenetic networks. *In CPM09. Springer*, 5577(7):289–300, 2009.

[29] Philippe Gambette, Vincent Berry, and Christophe Paul. Quartets and unrooted phylogenetic networks. *Journal of Bioinformatics and Computational Biology*, 10(4):1250004.1–1250004.23, 2012.

[30] Olivier Gascuel, Fabio Pardi, and Jakub Truszkowski. Distance-based phylogeny reconstruction: Safety and edge radius. *Encyclopedia of Algorithms, Springer New York*, pages 567–571, 2016.

[31] Bernhard Gittenberger. Diskrete Methoden. course notes, Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien. Version summer term. `http://www.dmg.tuwien.ac.at/gittenberger/dmeth.pdf`, (2), 2013.

[32] Andreas DM Gunawan, Jeyaram Rathin, and Louxin Zhang. Counting and enumerating galled networks. *https://arxiv.org/abs/1812.08569*, 2018.

[33] Katharina Huber, Vincent Moulton, and Taoyang Wu. Transforming phylogenetic networks: Moving beyond tree space. *JTB.*, 404:30–39, 2016.

[34] Daniel H. Huson, Tobias Klöpper, Pete J. Lockhart, and Mike A. Steel. Reconstruction of reticulate networks from gene trees. In Satoru Miyano, Jill Mesirov, Simon Kasif, Sorin Istrail, Pavel A. Pevzner, and Michael Waterman, editors, *Research in Computational Molecular Biology*, pages 233–249, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[35] D.H. Huson, R. Rupp, and C. Scornavacca. Phylogenetic networks: Concepts, algorithms and applications. *Cambridge University Press*, 2010.

[36] Trinh N. D. Huynh, Jesper Jansson, Nguyen Bao Nguyen, and Wing-Kin Sung. Constructing a smallest refining galled phylogenetic network. In Satoru Miyano, Jill Mesirov, Simon Kasif, Sorin Istrail, Pavel A. Pevzner, and Michael Waterman, editors, *Research in Computational Molecular Biology*, pages 265–280, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[37] Peter J. Cameron. Advanced combinatorics. course notes, university of st andrews. version summer term. `http://www-groups.mcs.st-andrews.ac.uk/~pjc/Teaching/MT5821/1/`, 2014.

[38] Remie Janssen, Mark Jones, Péter L. Erdös, Leo van Iersel, and Celine Scornavacca. Exploring the tiers of rooted phylogenetic network space using tail moves. *Bulletin of Mathematical Biology*, 80(8):2177–2208, 2009.

[39] S. Kelk and C. Scornavacca. Constructing minimal phylogenetic networks from softwired clusters is fixed parameter tractable. *Algorithmica*, 68(4):886–915, 2014.

[40] S. Kelk L. van Iersel and C. Scornavacca. Kernelizations for the hybridization number problem on multiple nonbinary trees. *J. Comput. System Sci.*, 82(6):1075–1089, 2016.

[41] Abraham Lempel, Shimon Even, and Israel Cederbaum. An algorithm for planarity testing of graphs. *Theory of Graphs: International Symposium*, pages 215–232, 1967.

[42] S. Linz M. Bordewich and C. Semple. Lost in space? Generalising subtree prune and regraft to spaces of phylogenetic networks. *J. Theoret. Biol.*, 423:1–12, 2017.

[43] Marefatollah Mansouri. Counting general phylogenetic networks. *in preparation*, 2020.

[44] Marefatollah Mansouri. The structure and enumeration of galled networks. *in preparation*, 2020.

[45] Zhao F Mellor-Crummey J. Markstein V. Nakhleh L, Jin G. Reconstructing phylogenetic networks using maximum parsimony. *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005)*, pages 93–102, 2005.

[46] S. Linz P. Cordue and C. Semple. Phylogenetic networks that display a tree twice. *Bull. Math. Biol.*, 76(10):2664–2679, 2014.

[47] David Penny. Inferring Phylogenies.Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts. *Systematic Biology*, 53(4):669–670, 08 2004.

[48] Andrew R. Francis and Mike Steel. Which phylogenetic networks are merely trees with additional arcs? *Systematic biology*, 64, 2015.

[49] N. A. Rosenberg. Counting coalescent histories. *J. Comput. Biol.*, 14(3):360–377, 2007.

[50] K. St. John S. Linz and C. Semple. Counting trees in a phylogenetic network is #P-complete. *SIAM J. Comput.*, 42(4):1768–1776, 2013.

[51] E. Schröder. Vier kombinatorische Probleme,. *Z. Math. Phys.*, 15:361–376, 1870.

[52] C. Semple. Phylogenetic networks with every embedded phylogenetic tree a base tree. *Bull. Math. Biol.*, 78(1):132–137, 2016.

[53] C. Semple. Size of a phylogenetic network. *Discrete Appl. Math.*, 217(2):362–367, 2017.

[54] C. Semple and M. Steel. Phylogenetics. *Oxford University Press*, 2003.

[55] Charles Semple and Jack Simpson. When is a phylogenetic network simply an amalgamation of two trees? *Bulletin of Mathematical Biology*, 80(9):2338–2348, 2018.

[56] Charles Semple and Mike Steel. Unicyclic networks: compatibility and enumeration. *Transactions on Computational Biology and Bioinformatics*, 3(1):84–91, 2006.

[57] J. R. Simpson. Tree structure in phylogenetic networks, , PhD thesis, University of Canterbury, New Zealand . 2019.

[58] P. H. A. Sneath. Cladistic representation of reticulate evolution. *Systematic Zoology. Oxford University Press, Society of Systematic Biologists, Taylor and Francis, Ltd.*, 24(3):360–368, 1975.

[59] Richard P. Stanley. Enumerative combinatorics. *Cambridge Studies in Advanced Mathematics*, 1, 2011.

[60] Mike Steel. Phylogeny discrete and random processes in evolution. *Society for Industrial and Applied Mathematics*, 2016.

[61] A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):517–535, 2011.

[62] Leo van Iersel, Judith Keijsper, Steven Kelk, Leen Stougie, Ferry Hagen, and Teun Boekhout. Constructing level-2 phylogenetic networks from triplets. *TCBB. http://www.win.tue.nl/ liersel/level2full.pdf*, 6(4):667–681, 2009.

[63] Leo van Iersel and Vincent Moulton. Trinets encode tree-child and level-2 phylogenetic networks. *Journal of Mathematical Biology*, 68:1432–1416, 2014.

[64] Douglas B. West. Introduction to graph theory. *Prentice Hall*, 2000.

[65] S. J. Willson. Properties of normal phylogenetic networks. *Bull. Math. Biol.*, 72(2):340–358, 2010.

[66] Louxin Zhang. Generating normal networks via leaf insertion and nearest neighbor interchange. *BMC Bioinformatics*, 20(20):642, 2019.

Wiedner Hauptstr. 8
1040 Vienna

TEL: +43 1 58801 - 104 584
marefatollah.mansouri@tuwien.ac.at

Marefatollah Mansouri

“ PhD student under Supervision of Dr. Prof. **Bernhard Gittenberger** started on April 2017. Programs in the field of analytic combinatorial methods, and its applications on studying phylogenetic networks. Special interests in Algorithm design and probabilistic method in Graph, optimization, and Graph theory application in Computer sciences.”
Doctoral Dissertation: Combinatorial Properties of Phylogenetic Networks.

Education

2012–2014 M.sc. Student. **Institute for Advanced Studies in Basic Sciences (IASBS)**, Zanjan, IRAN.
Masters Thesis: Evaluation of Algorithms for Learning a Hidden Subgraph. (With probabilistic method)
Total GPA: (17.56/20)

2007–2011 BS. Student. **Shahid Beheshti College, University of Farhangiyan**, Zanjan, IRAN.
Total GPA: (17.43/20.0)

Publications

[1] Michael Fuchs, Bernhard Gittenberger, and Marefatollah Mansouri. Counting Phylogenetic Networks with Few Reticulation Vertices: Tree-Child and Normal Networks, *Australasian Journal of Combinatorics.* **73** (2) (2018), 385–423.

[2] Mathilde Bouvel and Philippe Gambette and Marefatollah Mansouri, Counting Phylogenetic Networks of level 1 and 2. *arXiv:1909.10460* (2019).

[3] Marefatollah Mansouri, Counting general Phylogenetic Networks. *in preparation* (2020).

[4] Marefatollah Mansouri, The structure and enumeration of galled networks. *in preparation* (2020).

Presentation

- Counting Phylogenetic Networks with k Reticulation Nodes. Tutorial and Workshop on Analytic and Enumerative Aspects of Combinatorics. **The Institute of Statistical Sciences at Academia Sinica, Taiwan.** October 30th to October 31st, 2017.
- Combinatorics of Phylogenetic Networks. **ANR-FWF-MOST Meeting Caen, France.** October 29 and 30, 2018.
- Counting general phylogenetic networks. **ANR-FWF-MOST meeting, Wien.** August 30th, 2019.
- Combinatorial Optimization and Its Applications in Computer Science. **Institute for Advanced Studies in Basic Sciences.** 21 January 2013.
- Finding a Large Hidden Clique in a Random Graph. **Institute for Advanced Studies in Basic Sciences.** 10 June 2014.
- Solving group testing problem, with recognizing hidden spanning tree. **Institute for Advanced Studies in Basic Sciences.** 10 October, 2013.

Visiting experiences and Seminars

(1)−**Dagstuhl Seminar 19443.** Algorithms and Complexity in Phylogenetics. Schloss Dagstuhl Leibniz Center for Informatics. Visiting ans Seminar. October 27 - 31, 2019.

(2)−**AofA. International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms.** CIRM, Luminy, France, June 24-28, 2019.

- June 29–5 July, 2019, Visiting, **Est Marne-la-Valle University, Paris**.
- June 25–29, 2018, AofA, **Uppsala University, Sweden**.
- June 19-23, 2017, AofA 2017, **Princeton University, USA**.

(3)−**5th Algorithmic and Enumerative Combinatorics Summer School.** Hagenberg, Austria, July 29-August 2, 2019.

- 4th Algorithmic and Enumerative Combinatorics Summer School 2018.

(4)−**ALEA in Europe Workshop.** Vienna, Austria, October 9-13, 2017.

(5)−**European Conference on Combinatorics, Graph Theory and Applications.** Vienna, Austria, August 28 - September 1, 2017.

3rd Workshop on Analytic and Enumerative Aspects of Combinatorics. university of Caen, France, August 30-31, 2019.

- October 28, 2017, **National Chiao Tung University**, TAIWAN, Tutorial and Workshop on Analytic and Enumerative Aspects of Combinatorics.
- October 29-30, 2018, **MOST (Taiwan) − ANR (France) − FWF (Austria)** workshop, TU Wien, Vienna, Austria.
- June 14–27, 2019, Visiting, **National Chiao Tung University, TAIWAN**.

Honors and Awards

2017 **FWF grant,Vienna University of Technology**, Wien, Austria.

2007–2014 **Financial scholarship of the Ministry of education.** The criteria of this selection are: GPA, Research Activities and Extracurricular and Social Activities.

2014 **Ranked 1th**, among all students in in Dept. of Mathematics In the field of graph theory, at the end of M.S period. Zanjan, Iran.

2011 **Ranked 1th**, 1th among all students of the Shahid Beheshti University, at the end of B.S period, Zanjan, Iran.

2009 **Honored for the best B.S. paper,** presented in Iranian Student's Conference on Mathematics Education. Zanjan, Iran.

Teaching Experiences

2013 **Teaching Assistant of Discrete Mathematics II and advanced graph Theory for IT students.** Spring semester, IASBS, Zanjan, Iran.

2014 **Teaching Assistant of Differential Equations for IT students.** Spring semester, IASBS, Zanjan, Iran.

2017 **Teaching advanced mathematics.** Zanjan, Iran.

2009–2013 **Teaching mathematics for High school students**, Zanjan, Iran.

Program Committees

2013–2015 **Management Of The Mathematics Teacher team**, Zanjan, Iran.

2014 **Member of Executive manager of the International Mathematical Olympiad.** IASBS, Zanjan, Iran.

2018–Present
Member of The Vienna School of Mathematics. (https://www.vsmath.at/) Wien, Austria.

Extracurricular Activity and Projects

Nov. 2012 **Study on Group Testing Problems**, Zanjan, Iran.

Jun. 2013 **Studying Probabilistic methods on Random Graphs.** IASBS, Zanjan, Iran.

2015–2016 **Survey on Algorithm design especially Randomize Algorithms for Combinatorial Optimization problems.** IASBS, Zanjan, Iran.

2017 **Survey on M-ary search tree and Polya urn models.** Wien, Austria.

Skills

Programming. Maple, LaTeX, IPE, Matlab, Python (Elementary), Also Theoretical background in Algorithm design.

Languages. English, Persian (Native), Azerbaijani (Native), German (Elementary).