國立政治大學應用數學系

碩士學位論文

譜系網路的計算：Galled Trees 與少量網點的 Tree-child Networks

Counting Phylogenetic Networks: Galled Trees and Tree-Child Networks with Few Reticulation Nodes

指導教授：符麥克　博士

研究生：黃恩宇　撰

中 華 民 國 111 年 5 月

DOI:10.6814/NCCU202200447

前言

譜系網路是演化生物學當中的一個重要工具，它們提供一個操作分類單元（operational taxonomic units）間關係的圖像化表示法，特別是演化歷史。在近年的研究當中，許多組合相關的問題諸如：實際數量的計算與漸近行為的估計已經慢慢被理解，在這篇論文當中，我們將探討在應用上常見的兩大主要譜系網路：galled trees 與 tree-child networks.

首先是 galled trees 的部分，在 [BGM20] 中，Bouvel 等人對 galled trees 實際數量的計算與漸近行為的估計有詳細的討論，然而，在實務上有兩個常見子類別 ─ 有 normal 與 one-component 性質的 galled trees─ 在這篇研究當中沒有被探討；在另外一篇研究當中 ([CZ20]) ，Cardona 跟 Zhang 對 galled trees 以及上述的兩個子類別在實際數量上都做了詳細的計算，惟漸近估計的部分有所缺乏。我們將會提出三個類別 galled trees 數量的計算公式並討論他們的漸近表現，對這兩篇研究做出結合與延伸，此外，我們也會多考慮網點數量，給予漸近分布的結果。

計算具少量網點的 Tree-child networks 已經在許多研究中藉由不同的方法討論過，舉例來說，tree-child networks 的漸近表現在 [FGM19] 與 [FGM21] 二篇論文中已被解出，當葉子數 n 趨近於無限大時，具 k 個網點的 tree-child networks 的數量會逼近

$$c_k \left(\frac{2}{e}\right)^n n^{n+2k-1}.$$

另一方面，在 [CZ20] 中所提出透過 component graphs 來計算 tree-child networks 的方式也是有效的，我們延伸這個計算方式來得到更多網點時的計算公式，並比較先前以不同方式計算出來的結果，此外，透過 component graph 的方法也對上述漸近行為提供了更直觀的證明，更進一步的，透過這個方法可以取得常數 c_k 的一般式，即 $c_k = 2^{k-1}\sqrt{2}/k!$。

關鍵詞：譜系網路、元件圖、漸近估計

DOI:10.6814/NCCU202200447

Preface

Phylogenetic networks have become an important tool in evolutionary biology; they provide a graphical representation of the relationships between the operational taxonomic units and thus can be used for visualizing the evolutionary process. In recent years, many studies on combinatorial questions such as exact enumeration and asymptotic counting problems have been published for them. In this thesis, we investigate galled trees and tree-child networks, two classes of phylogenetic networks that are important in applications.

For galled trees, exact and asymptotic enumeration has been studied in [BGM20]. However, there are two important subclasses, namely normal and one-component galled trees which frequently occur in practice and which were not treated in [BGM20]. On the other hand, in [CZ20], the authors discussed galled trees as well as normal and one-component galled trees from an enumerative perspective but provided little asymptotic information. We will combine and continue the two works by giving for all three classes of galled trees exact formulas and derive the first order asymptotics of their numbers. Moreover, distributional results of the number of reticulation nodes will also be considered.

The enumeration of tree-child networks with few reticulation nodes has been studied in many papers through different approaches. For instance, the asymptotic counting problem was solved in [FGM19] and [FGM21] where it was shown that their number has the first order asymptotics:

$$c_k \left(\frac{2}{e}\right)^n n^{n+2k-1},$$

as the number of leaves n tends to ∞, where k is the number of reticulation nodes and $c_k > 0$ is a constant. Counting tree-child networks via component graphs is an effective way which was proposed in [CZ20]. We will extend this approach to

DOI:10.6814/NCCU202200447

obtain formulas for the number of tree-child networks with more reticulation nodes and compare them with the results from previous papers (where such results were derived with different methods). Moreover, the counting method via component graphs also gives a more straightforward proof of the above asymptotic result; in addition, it yields an easy expression for c_k, namely, $c_k = 2^{k-1}\sqrt{2}/k!$.

Keywords: Phylogenetic networks, component graphs, asymptotic estimate

DOI:10.6814/NCCU202200447

Contents

DOI:10.6814/NCCU202200447

Chapter 1

Introduction

1.1 Phylogenetic Trees

Phylogenetic trees are one of the widely-used tools in evolutionary biology and their combinatorial properties are well-understood; see Figure 1.1 for an example. Although this thesis will focus on phylogenetic networks, phylogenetic trees still play an important role in the following chapters. Thus, we give a rigorous definition.

Definition 1.1 (Phylogenetic trees)**.** *A phylogenetic tree is a planted, binary tree whose leaves are bijectively labelled by* $\{1, \ldots, n\}$*. The set of all phylogenetic trees with* n *leaves is denoted by* $\mathcal{PT}_n$*.*

Note that phylogenetic trees can be defined more generally, for example, they can be m-ary or unrooted. However, in this thesis, we use the above definition since we discuss only rooted, binary trees.

Theorem 1.2. *Let* T_n *be the number of phylogenetic trees with* n *leaves. Then,*

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

DOI:10.6814/NCCU202200447

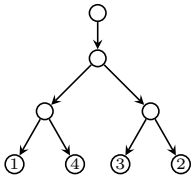


Figure 1.1: An example of a phylogenetic tree with $n = 4$.

Proof. Using the fact that each phylogenetic tree with n leaves has $2n - 1$ edges, one can insert one other leaf into any edge and therefore generate $2n - 1$ phylogenetic trees with $n + 1$ leaves. Hence, a phylogenetic tree with $n - 1$ leaves generates $2n - 3$ different phylogenetic trees with n leaves. Thus,

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

This proves the theorem. □

1.2 Rooted Phylogenetic Networks

Rooted phylogenetic networks, or RPNs for short, generalize phylogenetic trees. More specifically, they contain reticulation nodes that can represent horizontal gene transfer or hybridization in evolutionary biology.

A *rooted phylogenetic network* is a directed acyclic graph (DAG) without double edges and the direction of all edges from a unique root to leaves; moreover, the vertices of such a network can be classified into four types by the in-degree and out-degree of each node:

(1) A (unique) *root* with in-degree 0 and out-degree 1;

(2) *Tree nodes* with in-degree 1 and out-degree 2;

(3) *Reticulation nodes* with in-degree 2 and out-degree 1;

DOI:10.6814/NCCU202200447

(4) *Leaves* with in-degree 1 and out-degree 0 which are bijectively labelled with labels from the set $\{1, \ldots, n\}$.

Moreover, an edge (a, b) is a *tree edge* if b is either a tree node or a leaf and a *reticulation edge* if b is a reticulation node.

From the definition, we have a simple result about the relationship of different types of nodes and edges:

Proposition 1.3. *Let N be a rooted phylogenetic network with n leaves and k reticulation nodes. N has $k + n - 1$ tree nodes, $k + 2n - 1$ tree edges and $2k$ reticulation edges.*

Proof. Let t be the number of tree nodes. By definition, we have $t + n$ tree edges and $2k$ reticulation edges. Observe that a tree node increases the number of leaves by 1 and a reticulation node decreases the number of leaves by 1; therefore, we have that $1 + t - k = n$ from which the result follows. □

Clearly, if the number of reticulation nodes of a network equals 0, it is a phylogenetic tree. However, the counting problem of networks is much more difficult compared to the one for phylogenetic trees. To simplify such questions, topological constraints are needed which lead to the definition of many subclasses of rooted phylogenetic networks. In this thesis, we discuss the following two:

Galled trees: A rooted phylogenetic network is called a *galled tree* if each biconnected component has at most 1 reticulation node; see Figure 1.2 (a) for an example. We call the biconnected components with at least three vertices reticulation cycles. For convenience, we use $\mathcal{GT}_{n,k}$ to denote the set of all galled tree with n leaves and k reticulation nodes.

Tree-child networks: A rooted phylogenetic network is called a *tree-child network* if every non-leaf node has a child that is not a reticulation node; see Figure

DOI:10.6814/NCCU202200447

(a) (b)

Figure 1.2: (a) A one-component galled tree having $n = 6$ leaves and $k = 2$ reticulation nodes which is not normal. (b) A normal tree-child network which is not one-component where $n = 6$ and $k = 2$.

1.2 (b) for an example. We use $\mathcal{TC}_{n,k}$ to denote the set of all tree-child networks with n leaves and k reticulation nodes. Note that $\mathcal{GT}_{n,k} \subseteq \mathcal{TC}_{n,k}$.

Apart from the two network classes above, we use two more constraints that we found useful in this thesis:

One-component networks: A rooted phylogenetic network is called a *one-component network* if the child of each reticulation node is a leaf; see Figure 1.2 (a) for an example. We denote the set of all one-component networks of a class $\mathcal{S}$ of rooted phylogenetic networks by $\mathcal{OS}$.

Normal networks: A rooted phylogenetic network is called a *normal network* if it is a tree-child network and the two parents of each reticulation node are incomparable; see Figure 1.2 (b) for an example. We denote the set of all normal networks of a class $\mathcal{S}$ of tree-child networks by $\mathcal{NS}$.

1.3 Previous Results and Purpose of This Work

Phylogenetic trees were already used by Charles Darwin to visualize the ancestor-descendent relationship among species which was seen as the foundation of evolutionary biology. As our understanding of genes became more clear,

DOI:10.6814/NCCU202200447

tree-like structures are no longer enough to give an illustration of the evolutionary process. As a result, phylogenetic networks became more and more popular since they allow for a more flexible modeling. Consequently, in addition to their biological properties, algorithmic, combinatorial and probabilistic properties of phylogenetic networks have been extensively studied in recent years.

In the paper [BGM20], Bouvel et al. focused on two families of phylogenetic networks, namely, level-1 networks (galled trees) and level-2 networks, in both rooted and unrooted cases. They provided closed formulas of such networks together with asymptotic estimates of the number of the networks and they further gave distributional results of some parameters. Unlike [BGM20] who used analytic combinatorial tools, Cardona and Zhang counted galled trees in [CZ20] through an enumerative approach. Moreover, they also considered one-component and normal galled trees giving closed formula for these two subfamilies of galled trees, too.

Apart from galled trees, tree-child networks were also discussed in [CZ20]. Cardona and Zhang utilized component graphs and gave an efficient algorithmic way for computing the number of tree-child networks when the number of leaves and reticulation nodes are small. They also computed an exact formula for the number of tree-child networks with 2 reticulation nodes. As for analytic combinatorial tools, [FGM19] studied tree-child and normal networks. Fuchs et al. obtained an asymptotic expansion of the number of these two families of networks. Further in [FGM21], they gave explicit expressions for the exponential generating functions with a fixed number of reticulation nodes, k, and also computed the numbers of the two classes of networks for fixed k, where $k = 1, 2, 3$.

We are interested in both enumerative and asymptotic results for galled trees and tree-child networks. More precisely, for galled trees, we will review the results for one-component galled trees and normal galled trees from [CZ20] and then

DOI:10.6814/NCCU202200447

use similar tools as in [BGM20] to prove asymptotic results and distributional results for parameters. Next, we will use the approach from [CZ20] to give an exact formula for the number of tree-child networks with 3 reticulation nodes and compare it with the one obtained in [FGM21]. Finally, we will show that the approach from [CZ20] can be used to re-derive the main result in [FGM19]. This new method will allow us to solve an open problem from [FGM19].

We conclude by giving an outline of this thesis. First, in Chapter 2, we will explain the methods and theorems which are needed in this thesis; all topics are accompanied by examples. In Chapter 3, we present our results for galled trees and its subclasses. In Chapter 4, we focus on tree-child networks. We will show our enumerative results with few reticulation nodes and then present our proof of their first-order asymptotics. Finally, we summarize our results and give a conclusion in Chapter 5.

DOI:10.6814/NCCU202200447

Chapter 2

Methods

In this chapter, we summarize the methods used in this thesis; the topics are organized in the following order: some basics on symbolic combinatorics, exact and asymptotic enumeration, component graphs and Laplace method.

More precisely, symbolic methods for combinatorial objects are discussed in Section 2.1; they aim to find the relationship between the combinatorial structures and generating functions that encode their counting sequence. Next in Section 2.2, we will explain the exact and asymptotic Lagrange inversion formulas, powerful tools for dealing with generating functions, and combinatorial limit theorems for discrete sequences of random variables. Then, in Section 2.3 we will describe the concept of component graphs which was used to count tree-child networks by Cardona and Zhang. Finally, Section 2.4 introduces the Laplace method which is an important method in bivariate asymptotics.

Most of the contents of this chapter can be found in [FS09], Section 4.7 in [SF13] and Section 4 of [CZ20].

DOI:10.6814/NCCU202200447

2.1 Symbolic Method

Counting discrete structures is one of the pillars of combinatorics. More precisely, we want to study the number of elements of a fixed size in a combinatorial class, which is a set of discrete objects. A precise definition is as follows:

Definition 2.1. *(Combinatorial class) A combinatorial class is a finite or denumerable set of discrete objects on which a size function is defined such that:*
(1) the size of an element is a non-negative integer;
(2) the number of elements of any given size is finite.

As a convention, for a class $\mathcal{A}$, the number of elements of size n will be denoted by A_n and called the counting sequence for $\mathcal{A}$.

Generating functions are formal power series which encode these counting sequence as its coefficients. Usually, we will use ordinary generating functions, in abbreviation, OGFs, for unlabelled classes and exponential generating functions, EGFs, for labelled ones, their forms are shown below:

$$A(z) = \sum_{n=0}^{\infty} A_n z^n, \qquad \text{OGF for unlabelled class } \mathcal{A};$$

$$B(z) = \sum_{n=0}^{\infty} B_n \frac{z^n}{n!}, \qquad \text{EGF for labelled class } \mathcal{B}.$$

Since the coefficients of generating functions are usually our focus, we also introduce a notation to extract coefficients from generating functions, $[\cdot]$, where one puts the variable to any power in the bracket depending on which term one wants to extract, i.e.,

$$[z^n]A(z) = A_n, \quad [z^n]B(z) = \frac{B_n}{n!}.$$

For example, in this thesis, the combinatorial classes we will consider are

DOI:10.6814/NCCU202200447

galled trees with k reticulation nodes, tree-child networks with k reticulation nodes and some related network classes where the size function of each class will be the number of leaves of a network.

Another example is the class of phylogenetic trees which we denote by $\mathcal{PT}$. The size is again the number of leaves and for the exponential generating function of phylogenetic trees, we have

$$T(z) = \sum_{n\geq 0} \frac{T_n}{n!} z^n = \sum_{n\geq 1} \frac{(2n-3)!!}{n!} z^n = 1 - \sqrt{1-2z}; \tag{2.1}$$

see Theorem 1.2.

Finally, two classes, the *empty* and *atomic class*, are worth mentioning. An *empty class* $\mathcal{E}$ is a class having only one object, ϵ, that has size 0. An *atomic class* $\mathcal{Z}$ also has only one object this time of size 1. Obviously, the corresponding generating functions are

$$E(z) = 1, \qquad Z(z) = z.$$

Specification

In order to count the number of objects of a combinatorial class, we usually need to decompose them and use simpler structures to make the counting feasible. Then, the construction of the class from these simpler structures will be translated into different operations on generating functions resulting in an equation satisfied by the corresponding generating functions. Such a decomposition of a class is called specification.

Note that we are only concerned with labelled objects; therefore, we consider labelled constructions. As for unlabelled objects, see [FS09, Part A.I] for more information.

DOI:10.6814/NCCU202200447

Sum. The *Sum* of two classes is analogous with the concept of disjoint union; thus, in case that the classes have objects in common, we define the sum of two classes as the union of their duplicates which are given different colors. More precisely, we have

Definition 2.2. *Let $\mathcal{B}$ and $\mathcal{C}$ be two classes. The sum of $\mathcal{B}$ and $\mathcal{C}$, denoted by $\mathcal{B}+\mathcal{C}$, is*

$$\mathcal{B}+\mathcal{C} = (\mathcal{B})_{blue} \cup (\mathcal{C})_{red}.$$

The size of an object in the sum is inherited from their class. Clearly, if $\mathcal{A} = \mathcal{B} + \mathcal{C}$, then the counting sequences has the relation

$$A_n = B_n + C_n, \quad \text{and therefore} \quad A(z) = B(z) + C(z).$$

That is, the generating function of a sum of two combinatorial classes is the sum of the corresponding generating functions.

Product. *Product* is another basic constructions for labelled classes. It consists of tuples of labelled objects, however, we need to relabel them to avoid that labels are repeated. For this purpose, we define a relabelling function of an object which is a function for a labelled object that satisfies that for two labels $i < j$, we have $r(i) < r(j)$ after relabelling. Now, we can give the definition:

Definition 2.3. *Given two labelled objects, $\beta \in \mathcal{B}$ and $\gamma \in \mathcal{C}$, the product of β and γ, denoted by $\beta \star \gamma$, is the set of tuples $(r_B(\beta), r_C(\gamma))$ where r_B and r_C are any pair of relabelling functions so that the intersection of their images is $\emptyset$ and the union of their images is $\{1, \ldots, |\beta| + |\gamma|\}$.*
The product of $\mathcal{B}$ and $\mathcal{C}$ is then defined by

$$\mathcal{B} \star \mathcal{C} = \bigcup_{\beta\in\mathcal{B},\gamma\in\mathcal{C}} \beta \star \gamma.$$

DOI:10.6814/NCCU202200447

We now consider the counting sequence of $\mathcal{B} \star \mathcal{C}$. Let $\mathcal{A} = \mathcal{B} \star \mathcal{C}$. Observe that the counting sequences have the relation

$$A_n = \sum_{\substack{\beta \in \mathcal{B}, \gamma \in \mathcal{C} \\ |\beta|+|\gamma|=n}} \binom{|\beta| + |\gamma|}{|\beta|, |\gamma|} B_{|\beta|} C_{|\gamma|}.$$

Therefore,

$$\begin{aligned} A(z) &= \sum_{n \geq 0} \frac{A_n}{n!} z^n = \sum_{n \geq 0} \sum_{k=0}^{n} \frac{1}{n!} \binom{n}{k} B_k C_{n-k} z^n \\ &= \sum_{n \geq 0} \sum_{k=0}^{n} \frac{B_k}{k!} z^k \cdot \frac{C_{n-k}}{(n-k)!} z^{n-k} \\ &= B(z) \cdot C(z). \end{aligned}$$

In words, the EGF of the labelled product is the product of EGFs.

Sequence. The k-th power of $\mathcal{B}$ is defined by $\mathcal{B} \star \mathcal{B} \cdots \star \mathcal{B}$ with k factors of $\mathcal{B}$ and will be denoted as $\textsc{Seq}_k(\mathcal{B})$. The sequence class of $\mathcal{B}$ is defined by:

$$\textsc{Seq}(\mathcal{B}) = \mathcal{E} + \mathcal{B} + \textsc{Seq}_2(\mathcal{B}) + \textsc{Seq}_3(\mathcal{B}) + \cdots + \textsc{Seq}_k(\mathcal{B}) + \cdots$$

From the previous discussions, we have that the corresponding generating function of $\textsc{Seq}_k(\mathcal{B})$ is $B(z)^k$. Recall that the empty class, $\mathcal{E}$, has generating function 1. Hence, if $\mathcal{A} = \textsc{Seq}(\mathcal{B})$, then

$$A(z) = 1 + B(z)^2 + \cdots + B(z)^k + \cdots = \frac{1}{1 - B(z)}.$$

Set. The class $\textsc{Set}$ and $\textsc{Seq}$ are almost the same with the only difference that the objects of $\textsc{Set}$ have no order. The k-set of $\mathcal{B}$ is defined as $\frac{\textsc{Seq}_k(\mathcal{B})}{k!}$ and will be denoted as $\textsc{Set}_k(\mathcal{B})$. The set class of $\mathcal{B}$ is defined by:

DOI:10.6814/NCCU202200447

$$\textsc{Set}(\mathcal{B}) = \mathcal{E} + \mathcal{B} + \textsc{Set}_2(\mathcal{B}) + \textsc{Set}_3(\mathcal{B}) + \cdots + \textsc{Set}_k(\mathcal{B}) + \cdots$$

The generating function of $\textsc{Set}_k(\mathcal{B})$ is $\frac{B(z)^k}{k!}$ and thus the generating function of $\mathcal{A} = \textsc{Set}(\mathcal{B})$ is

$$A(z) = 1 + B(z) + \frac{B(z)^2}{2!} + \cdots + \frac{B(z)^k}{k!} = e^{B(z)}.$$

We summarize the above constructions and their generating functions into a theorem, see Theorem II.1 in [FS09].

Theorem 2.4. *Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be combinatorial classes and $A(z)$, $B(z)$, $C(z)$ be their corresponding generating functions. The constructions of sum, labelled product, sequence and set and the associated operations on EGFs are listed below:*

$$\begin{aligned}
&\textit{Sum:} && \mathcal{A} = \mathcal{B} + \mathcal{C} && \Longrightarrow && A(z) = B(z) + C(z),\\
&\textit{Product:} && \mathcal{A} = \mathcal{B} \star \mathcal{C} && \Longrightarrow && A(z) = B(z) \cdot C(z),\\
&\textit{Sequence:} && \mathcal{A} = \textsc{Seq}(\mathcal{B}) && \Longrightarrow && A(z) = \frac{1}{1 - B(z)},\\
&\textit{Set:} && \mathcal{A} = \textsc{Set}(\mathcal{B}) && \Longrightarrow && A(z) = e^{B(z)}.
\end{aligned}$$

Before we go on to the next topic, let us give some examples.

Example 2.5 (Phylogenetic trees)**.** *We can describe the structure of phylogenetic trees as:*

$$\mathcal{PT} = \bullet + \underset{\mathcal{PT}\ \ \mathcal{PT}}{\wedge}$$

where the black node represents a labelled leaf and the white nodes represent internal nodes. Since the order of the trees in the second term is irrelevant, we obtain

DOI:10.6814/NCCU202200447

the specification:

$$\mathcal{PT} = \mathcal{Z} + \textsc{Set}_2(\mathcal{PT}). \tag{2.2}$$

By Theorem 2.4, we can translate (2.2) into

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

Solving for $T(z)$ we have

$$T(z) = 1 - \sqrt{1-2z};$$

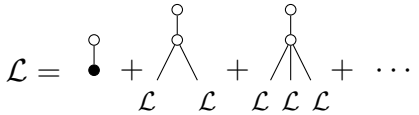
see (2.1).Consequently, by the binomial theorem, the n-th coefficient is

$$\begin{aligned}
\frac{T_n}{n!} &= [z^n]1 - \sqrt{1-2z} \\
&= [z^n] - (1-2z)^{\frac{1}{2}} \\
&= -\binom{\frac{1}{2}}{n}(-2)^n \\
&= -(-2)^n \cdot \frac{\frac{1}{2} \times \frac{-1}{2} \times \frac{-3}{2} \times \cdots \times \frac{-(2n-3)}{2}}{n!} \\
&= \frac{2^n}{2^n} \cdot \frac{1 \times 3 \times \cdots \times (2n-3)}{n!} \\
&= \frac{(2n-3)!!}{n!} \\
&= \frac{1}{n!} \cdot \frac{(2n-2)!}{2^{n-1}(n-1)!}.
\end{aligned}$$

which is the same result as in Theorem 1.2.

Example 2.6 (General phylogenetic trees)**.** *We can also generalize the concept of phylogenetic trees by allowing internal nodes to have out-degree at least 2. Denote by $\mathcal{L}$ the class of general phylogenetic trees. The size of an object is again the number of leaves. We give the construction of general phylogenetic trees below:*

DOI:10.6814/NCCU202200447



Note that the order of subtrees is again irrelevant. Thus, we have the specification

$$\mathcal{L} \;=\; \mathcal{Z} + \textsc{Set}_{\geq 2}(\mathcal{L}). \tag{2.3}$$

By Theorem 2.4, we can translate (2.3) into an equation for its generating function $L(z)$*:*

$$L(z) = z + e^{L(z)} - 1 - L(z).$$

We reorganize the terms in order to solve this equation,

$$\begin{aligned} L(z) - \frac{z-1}{2} &= \frac{1}{2} e^{L(z)} \\ &= \frac{1}{2} e^{\frac{z-1}{2}}\, e^{L(z) - \frac{z-1}{2}}. \end{aligned}$$

Now define $V(z)$ *as the function satisfying* $V(z) = ze^{V(z)}$*. Then,*

$$V\left(\frac{1}{2}\, e^{\frac{z-1}{2}}\right) = L(z) - \frac{z-1}{2}$$

and consequently,

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

The function, V*, is related to the Lambert W function by* $V(u) = -W(-z)$*. Therefore, we can obtain* L_n *with the help of the Lambert W function.*

DOI:10.6814/NCCU202200447

2.2 Asymptotic Enumeration

In Example 2.5 the EGF has a neat and simple expression which allowed us to find an exact formula for the number of phylogenetic trees. However, in practice, such a nice solution is seldomly available, e.g. see Example 2.6. Moreover, even if we obtain a formula for the counting sequence of a combinatorial class, the growth behaviour of the counting sequence as n tends to infinity might still not be clear. Therefore, it is important to have asymptotic tools in order to understand the growth of combinatorial quantities.

In the following parts, we will introduce the exact and asymptotic Lagrange inversion formulas which are powerful methods dealing with finding exact formulas and asymptotic approximations of counting sequences.

Lagrange inversion formulas

The Lagrange inversion formula relates the coefficients of the compositional inverse of a function to the coefficients of the powers of the function. In particular, since the combinatorial structures are often recursively defined, it turns out that we can often rearrange terms in a specification so that it fits into the setting of the Lagrange inversion formula.

Theorem 2.7 (Exact Lagrange inversion formula)**.** *Let $\phi(z) = \sum_{n=0}^{\infty} \phi_n z^n$ be a power series such that $\phi_0 \neq 0$. Suppose that a generating function $C(z)$ satisfies an equation of the form $C(z) = z\phi(C(z))$. Let $f(z)$ be any power series. Then, for all integer n, we have*

$$[z^n]f(C(z)) = \frac{1}{n}[z^{n-1}]f'(z)\phi(z)^n$$

DOI:10.6814/NCCU202200447

Moreover, if $f(z) = z$, then

$$[z^n]C(z) = \frac{1}{n}[z^{n-1}]\phi(z)^n.$$

We revisit Example 2.5 but this time we will obtain the exact counting formula with the help of the Lagrange inversion formula; this will turn out to be less computation-intensive compared with the derivation before.

Example 2.8 (Continuation of Example 2.5)**.** *The counting sequence of phylogenetic trees can also be derived without solving for its exponential generating function, $T(z)$. We have obtained an equation in terms of $T(z)$ from the specification of phylogenetic trees:*

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

To bring this equation into the form $T(z) = z\phi(T(z))$, we rearrange terms which gives

$$\phi(z) = \left(1 - \frac{z}{2}\right)^{-1}$$

Since $\phi(z)$ satisfies $\phi(0) \neq 0$, by the Lagrange inversion formula and the binomial theorem, we obtain that

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

We then have $T_n = \frac{(2n-2)!}{2^{n-1}(n-1)!}$ after multiplying by $n!$.

To understand the growth of combinatorial quantities, another form of the Lagrange inversion formula is helpful.

DOI:10.6814/NCCU202200447

Theorem 2.9 (Asymptotic Lagrange inversion formula)**.** *Let $\phi(z) = \sum_{n=0}^{\infty} \phi_n z^n$ be a power series such that $\phi_0 \neq 0$, ϕ_n is positive for all n, and $\phi(z) \neq \phi_0 + \phi_1 z$. Moreover, let $R > 0$ be the radius of convergence of ϕ at 0 and assume that $\phi(z) - z\phi'(z) = 0$ has a solution $\tau \in (0, R)$. Suppose that a generating function $C(z)$ satisfies an equation of the form $C(z) = z\phi(C(z))$. Then, we have the following:*

(i) $\rho = \frac{\tau}{\phi(\tau)}$ is the radius of convergence of C at 0;

(ii) We have $C(z) \sim \tau - \sqrt{\frac{2\phi(\tau)}{\phi''(\tau)}}\sqrt{1 - \frac{z}{\rho}}$ as $z \to \rho$;

(iii) We have $[z^n]C(z) \sim \sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}}\frac{\rho^{-n}}{\sqrt{\pi n^3}}$ as $n \to \infty$.

Remark 2.10. *The relationship between (ii) and (iii) is a result of the so-called transfer theorems which are used to transfer the asymptotic information around dominant singularities into asymptotic expansions of the coefficients; see [FS09, VI.3] for more detail.*

Example 2.11 (Continuation of Example 2.8)**.** *Recall from Example 2.5 that $T_n = \frac{(2n-2)!}{2^{n-1}(n-1)!}$. Thus, we can estimate the asymptotic growth of T_n directly by Stirling's formula. Since*

$$\begin{aligned}(2n-2)! &= \frac{(2n)!}{(2n)(2n-1)} \sim \frac{\sqrt{\pi}}{2} n^{-\frac{3}{2}} \left(\frac{4}{e^2}\right)^n n^{2n};\\ 2^{n-1}(n-1)! &= \quad 2^{n-1}\frac{n!}{n} \quad \sim \frac{\sqrt{2\pi}}{2} n^{-\frac{1}{2}} \left(\frac{2}{e}\right)^n n^n,\end{aligned}$$

dividing this gives

$$T_n = \frac{(2n-2)!}{2^{n-1}(n-1)!} \sim \frac{\sqrt{2}}{2}\left(\frac{2}{e}\right)^n n^{n-1}.$$

We can also apply the asymptotic form of Lagrange inversion formula. Recall that $\phi(z) = \left(1 - \frac{z}{2}\right)^{-1}$. Clearly, ϕ satisfies the assumptions of Theorem 2.9 with

DOI:10.6814/NCCU202200447

$R = 2$ and the equation $\phi(z) - z\phi'(z) = 0$ has 1 as a solution in $(0, 2)$. Thus, by Theorem 2.9 and Stirling's formula:

$$T_n = n! \, [z^n] \, T(z) \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \cdot \frac{1}{2} \frac{2^n}{\sqrt{\pi n^3}}$$
$$\sim \frac{\sqrt{2}}{2} \left(\frac{2}{e}\right)^n n^{n-1}.$$

Example 2.12 (Continuation of Example 2.6)**.** *Recall that $L(z)$ satisfies $L(z) = z + e^{L(z)} - 1 - L(z)$ which is not of the form $C(z) = z\phi(C(z))$. However, $L(z) = V\left(\frac{1}{2}e^{\frac{z-1}{2}}\right) + \frac{z}{2} - \frac{1}{2}$ and $V(z) = ze^{V(z)}$. Since the latter is of the form $C(z) = z\phi(C(z))$, we can find the asymptotics of $L(z)$ via $V(z)$.*

First for $V(z)$, we have $\phi(z) = e^z$ which satisfies the assumptions of Theorem 2.9 with $\tau = 1$. Thus, by Theorem 2.9, we have that $\rho = 1/e$ and $V(z) \sim 1 - \sqrt{2}\sqrt{1 - ez}$ as $z \to \frac{1}{e}$. Next, note that $\frac{1}{2}e^{\frac{z-1}{2}} \to 1/e$ when $z \to 2\log 2 - 1$. Thus, as $z \to 2\log 2 - 1$,

$$L(z) = V\left(\frac{1}{2}e^{\frac{z-1}{2}}\right) + \frac{z}{2} - \frac{1}{2} \sim \log 2 - \sqrt{2\log 2 - 1}\sqrt{1 - \frac{z}{2\log 2 - 1}}.$$

Finally, by the transfer theorems (see Remark 2.10) and Stirling's formula:

$$L_n = n! \, [z^n] \, L(z) \sim \frac{\sqrt{4\log 2 - 2}}{2} \left(\frac{1}{e(2\log 2 - 1)}\right)^n n^{n-1}.$$

Asymptotic distribution of parameters

So far, our generating functions have been univariate with the variable recording the size of the combinatorial objects. Including additional variables in generating functions, that is, recording other parameters, may give us a more clear picture of how most of the objects of size n look like.

Example 2.13 (The number of cherries in phylogenetic trees)**.** *A cherry in a phy-*

DOI:10.6814/NCCU202200447

logenetic tree is an internal node having two leaves as its children. We will take the number of cherries as a guiding example in this part. Let z as usual be the variable that records the number of leaves and x be the additional variable that records the number of cherries. Since the number of cherries of a tree is the sum of left and right subtree of the root except when the tree has size 2, we have the following equation for the bivariate generating function:

$$T(z,x) = z + \frac{xz^2}{2} - \frac{z^2}{2} + \frac{T(z,x)^2}{2}. \tag{2.4}$$

Under suitable assumptions, if we uniformly and randomly pick an object of size n, the following theorem shows that the additional parameter is asymptotically normal distributed. Moreover, the asymptotic mean and variance can be derived from the theorem as well.

Theorem 2.14. *Suppose that $C(z,x)$ is a power series that is the solution of the equation $C = F(C,z,x)$, where $F(C,z,x)$ satisfies:*

(i) *$F(C,z,x)$ is analytic in C, z and x around 0,*

(ii) *$F(C,0,x) = 0$,*

(iii) *$F(0,z,x) \neq 0$ and*

(iv) *all coefficients $[z^n C^m]F(C,z,1)$ are non-negative.*

Assume in addition that the region of analyticity of $F(C,z,x)$ contains non-negative solutions $z = z_0$ and $C = C_0$ of the system of equations:

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

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

DOI:10.6814/NCCU202200447

If X_n is a sequence of random variables such that $\mathbb{E}[x^{X_n}] = \frac{[z^n]C(z,x)}{[z^n]C(z,1)}$, then X_n is asymptotically normally distributed.

To be more precise, setting

$$\begin{aligned}\mu &= \frac{F_x}{z_0 F_z},\\ \sigma^2 &= \mu + \mu^2 + \frac{1}{z_0 F_z^3 F_{CC}}\left(F_z^2(F_{CC}F_{xx} - F_{Cx}^2)\right.\\ &\qquad \left. -2F_zF_x(F_{CC}F_{zx} - F_{Cz}F_{Cx}) + F_x^2(F_{CC}F_{zz} - F_{Cz}^2)\right),\end{aligned}$$

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

$$\mathbb{E}X_n = \mu n + O(1) \quad \textit{and} \quad \operatorname{Var} X_n = \sigma^2 n + O(1)$$

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

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

Example 2.15 (Continuation of Example 2.13)**.** *Rearranging terms in (2.4), we have*

$$T(z,x) = \frac{z + \frac{xz^2}{2} - \frac{z^2}{2}}{1 - \frac{T(z,x)}{2}};$$

that is, $F(T,z,x) = \frac{z+xz^2-z^2}{1-T/2}$. The conditions (i), (ii), (iii) and (iv) in the previous theorem can be easily checked. Moreover, the system of equations

$$\begin{cases} T = F(T,z,1) \\ 1 = F_T(T,z,1) \end{cases}$$

has a solution $T_0 = 1$ and $z_0 = 1/2$. Then after calculations of derivatives and applying the formulas given in the previous theorem, we have that the number of cherries in a phylogenetic trees is asymptotically normal with mean $\sim n/4$ and variance $\sim n/16$; see [MS00] where this was derived via urn models.

DOI:10.6814/NCCU202200447

2.3 Component Graphs and One-component Tree-child Networks

We saw above that the counting of phylogenetic trees (both exact and asymptotically) is relatively easy. On the other hand, due to reticulation nodes, the counting problems become much more difficult for phylogenetic networks. In order to simplify and overcome these problems, in Section 4 of [CZ20], Cardona and Zhang used component graphs to decompose a tree-child network into parts. In brief, we can view a tree-child network as a one-component tree-child network on top with networks attached below. In this section we will first introduce component graphs and then recall the counting result for one-component tree-child networks.

Component Graphs

The process of constructing the component graph from a network is as follows:

Given a network with k reticulation nodes, $\mathcal{N}$, we remove all reticulation edges turning the network into tree components. Take each tree component as a vertex of the component graph, C, induced from $\mathcal{N}$. For each removed reticulation edge in $\mathcal{N}$, add one edge between the nodes of C which contained the ends of that reticulation edge; see Figure 2.1 for examples.

From the construction of component graphs, we can easily see that each node of the component graph has in-degree 2 except the node containing the network root (having in-degree 0). This property is called **indegree constraint** in [CZ20].

DOI:10.6814/NCCU202200447

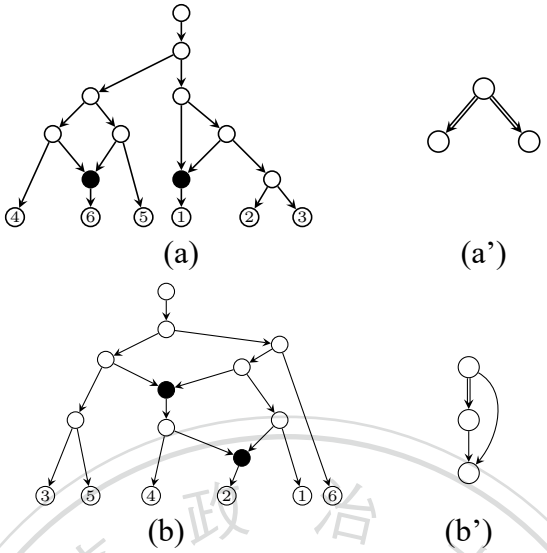


Figure 2.1: TCNs and their corresponding component graphs.

One-component tree-child networks

One-component tree-child networks are major building blocks when counting via component graphs. Recall that each reticulation node is followed by a leaf in a one-component networks. Thus, one-component networks arise from component graphs which just contain a single node. We will now count one-component tree-child networks.

Theorem 2.16. *Denote by* $\mathrm{OTC}_{n,k}$ *the number of one-component tree-child networks with* n *leaves and* k *reticulation nodes. Then, we have*

$$\mathrm{OTC}_{n,k} = \binom{n}{k} \frac{(2n-2)!}{2^{n-1}(n-k-1)!}.$$

A reticulation node is inserted into a network by selecting 2 tree edges (choosing identical edges is allowed) and then adding a node to each edge which is the starting point of a reticulation edge.

Note that for N, a one-component tree-child network, we can remove all

DOI:10.6814/NCCU202200447

reticulation nodes and edges attached to them so that we obtain a corresponding tree, T, of N where T has some degree 2 nodes. Hence, those degree 2 nodes form a map for us to insert reticulation nodes. In other words, any one-component tree-child network can be obtained from a phylogenetic tree after some reticulation insertions.

However, there is the question of whether we can have the same networks after inserting reticulation nodes into different trees. To answer this we give a lemma below. For convenience, we use $A \subseteq \{1, \ldots, n\}$, for a subset of labels of leaves below reticulation nodes, and $N \oplus A$ to represent all of the possible networks after inserting the nodes in A.

Lemma 2.17. *Let T_1, T_2 be two phylogenetic trees on $A \subseteq \{1, \cdots, n\}$ and $B \subseteq \{1, \cdots, n\} \setminus A$.* $(T_1 \oplus B) \cap (T_2 \oplus B) \neq \emptyset$ *if and only if* $T_1 = T_2$.

Proof. The "if" part is trivial.

Suppose that $T_1 \neq T_2$. Then there exists a node, say u, in T_1 such that the set of leaves below u is different from the corresponding set of leaves in T_2. Given $N_1 \in T_1 \oplus B$, let C_u be the set of leaves below u in N_1. If C_u appears in $N_2 \in T_2 \oplus B$, then $C_u{}'$ appears in both T_1 and T_2 where $C_u{}'$ is the set C_u with the leaves below reticulation nodes removed and we have a contradiction. Therefore, $(T_1 \oplus B) \cap (T_2 \oplus B) = \emptyset$. □

Now, we can give the proof of the theorem:

Proof of Theorem 2.16. Let N be a one-component tree-child network with n leaves and k reticulation nodes. We know that N can be generated from one of $\frac{(2(n-k)-1)!}{2^{n-k-1}(n-k-1)!}$ phylogenetic trees. From Proposition 1.3, we have that such a tree has $2(n-k)-1$ tree edges.

To perform our first reticulation node insertion, we can choose one of $\binom{2(n-k)-1}{2}$ pairs of distinct edges or one of $2(n-k)-1$ identical edges. In total,

DOI:10.6814/NCCU202200447

we have

$$\binom{2(n-k)-1}{2} + 2(n-k) - 1 = \frac{(2n-2k-1)(2n-2k)}{2}$$

ways and the number of tree edges is increased by 3. Here, note that the tree edges below reticulation nodes cannot be used to perform reticulation insertion (otherwise we would violate the one-component property). Continuing with the reticulation node insertion step by step, we have that there are $\frac{(2n-2k+2i-3)(2n-2k+2i-2)}{2}$ ways to perform our i-th reticulation insertion.

Moreover, there are k labels to be assigned to each reticulation which can be done in $\binom{n}{k}$ ways.

Therefore, the number of one-component tree-child networks with n leaves and k reticulation nodes equals to:

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

This concludes the proof. □

2.4 Laplace Method

The Laplace method is a powerful tool when estimating sums. The method is centered on the following threes steps:

- Restricting the range so that it contains the largest terms.
- Approximating the terms and bounding the tails.

DOI:10.6814/NCCU202200447

- Approximating by an integral, extending the range and bounding the new tails.

We will explain this via an example, namely, we will apply the Laplace method to estimate the number of one-component tree-child networks with n leaves. Recall that the number of one-component tree-child network with n leaves and k reticulation nodes is

$$\mathrm{OTC}_{n,k} = \binom{n}{k} \frac{(2n-2)!}{2^{n-1}(n-k-1)!}$$

and hence

$$\begin{aligned}\mathrm{OTC}_n &= \sum_{k=0}^{n-1} \binom{n}{k} \frac{(2n-2)!}{2^{n-1}(n-k-1)!} \\ &= \frac{n!(2n-2)!}{2^{n-1}} \sum_{k=0}^{n-1} \frac{1}{k!(n-k)!(n-k-1)!}.\end{aligned}$$

Separating the above expression into two parts, the estimate for the coefficient in front of the summation can be derived from Stirling's formula; for the second part, we will use the Laplace method.

First, for $\frac{n!(2n-2)!}{2^{n-1}}$, Stirling's formula gives

$$\frac{n!(2n-2)!}{2^{n-1}} \sim \sqrt{2}\pi \left(\frac{2}{e^3}\right)^n n^{3n-1}.$$

Second, set

$$S := \sum_{k=0}^{n-1} \frac{1}{k!(n-k)!(n-k-1)!} = \sum_{k=0}^{n-1} a_k(n)$$

whose asymptotics will be derived by the steps above.

First, to restrict to the range that contains the largest terms, we observe from

DOI:10.6814/NCCU202200447

the ratio of consecutive terms that $\frac{1}{k!(n-k)!(n-k-1)!}$ is increasing for $k \leq n-\sqrt{n+1}$ and decreasing for $k \geq n-\sqrt{n+1}$. Thus, we need to restrict k near $n-\sqrt{n+1}$. Next, setting $k = n-\sqrt{n}+x$ and expanding, we have

$$a_k(n) = \frac{1}{2\pi\sqrt{2e\pi}} n^{-\frac{1}{2}} e^{2\sqrt{n}} \left(\frac{e}{n}\right)^n \cdot e^{\left(-\frac{13}{6}-\frac{3}{2}x-x^2\right)n^{-1/2}+\left(-1-\frac{7}{4}x^2-\frac{1}{3}x^3\right)n^{-1}+\mathcal{O}\left(\frac{1+x^4}{n^{3/2}}\right)}.$$

where $x = o(n)$.

Observe from the equation above that for $x = \mathcal{O}\left(n^{\alpha}\right)$ with $n^{\frac{1}{4}} \leq n^{\alpha} \leq n^{\frac{1}{3}}$, $a_k(n)$ affects the asymptotics and is exponentially small outside that range. Thus, picking $\alpha = 3/10$, we have

$$\begin{aligned} a_k(n) &= \frac{1}{2\pi\sqrt{2e\pi}} n^{-\frac{1}{2}} e^{2\sqrt{n}} \left(\frac{e}{n}\right)^n e^{-\frac{x^2}{\sqrt{n}}} \left(1+\mathcal{O}\left(\frac{1+|x|^3}{n}+\frac{|x|}{\sqrt{n}}\right)\right) \\ &= \frac{1}{2\pi\sqrt{2e\pi}} n^{-\frac{1}{2}} e^{2\sqrt{n}} \left(\frac{e}{n}\right)^n e^{-\frac{x^2}{\sqrt{n}}} \left(1+\mathcal{O}\left(\frac{1}{n^{\frac{1}{10}}}\right)\right). \end{aligned} \tag{2.5}$$

uniformly in $x = \mathcal{O}\left(n^{3/10}\right)$.

Now, we consider

$$S = \sum_{|x|\leq n^{3/10}} \frac{1}{k!(n-k)!(n-k-1)!} + \sum_{|x|>n^{3/10}} \frac{1}{k!(n-k)!(n-k-1)!}$$

The tail $\sum_{|x|\geq n^{3/10}} \frac{1}{k!(n-k)!(n-k-1)!}$ is exponentially small with respect to the former term. Hence, we obtain that

$$S \sim \frac{1}{2\pi\sqrt{2e\pi}} n^{-\frac{1}{2}} e^{2\sqrt{n}} \left(\frac{e}{n}\right)^n \sum_{-n^{3/10}\leq x\leq n^{3/10}} e^{-x^2/\sqrt{n}} \left(1+\mathcal{O}\left(\frac{1}{n^{\frac{1}{10}}}\right)\right)$$

Finally, since $e^{-x^2/\sqrt{n}}$ is also exponentially small, we add back the tail for $e^{-x^2/\sqrt{n}}$ and approximate the sum by an integral, that is,

DOI:10.6814/NCCU202200447

$$
\begin{aligned}
S &\sim \frac{1}{2\pi\sqrt{2e\pi}} n^{-\frac{1}{2}} e^{2\sqrt{n}} \left(\frac{e}{n}\right)^n \int_{-\infty}^{\infty} e^{-x^2/\sqrt{n}} \left(1 + \mathcal{O}\left(\frac{1}{n^{\frac{1}{10}}}\right)\right) \mathrm{d}x \\
&\sim \frac{1}{2\pi\sqrt{2e}} n^{-\frac{1}{4}} e^{2\sqrt{n}} \left(\frac{e}{n}\right)^n \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right).
\end{aligned}
$$

Overall, we have that

$$
\mathrm{OTC}_n \sim \frac{1}{2\sqrt{e}} n^{-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n n^{2n};
$$

see Theorem 2 in [FYZ21].

DOI:10.6814/NCCU202200447

Chapter 3

Galled Trees

In this chapter, we will study the class of galled trees. This family was discussed both in [BGM20, Section 4] and [CZ20, Section 3]. In the former paper, the authors used the symbolic method as well as analytic tools to derive not only an exact formula for the number of galled trees but also the asymptotic growth term and distributional results (these results will be reviewed in Section 3.1 below). In the later one, the authors counted the number by relating galled trees to ordered trees and using leaf insertions. Moreover, they considered two additional families namely, normal and one-component galled trees. We list their counting formulas for each class: denote by g_n and m_n the number of galled trees and normal galled trees, respectively, and by $c_{n,k}$ denote the number of one-component galled trees with k reticulation nodes. We have

$$g_n = \sum_{(k_2,k_3,\cdots,k_n)\in A} \frac{(n+k_2+\cdots+k_n-1)!3^{k_2+k_3}4^{k_4}\cdots n^{k_n}}{k_2!k_3!\cdots k_n!2^{k_2+k_3+\cdots+k_n}};$$

$$m_n = \sum_{(k_2,k_3,\cdots,k_n)\in A} \frac{(n+k_2+\cdots+k_n-1)!1^{k_3}2^{k_4}\cdots(n-2)^{k_n}}{k_2!k_3!\cdots k_n!2^{k_2+k_3+\cdots+k_n}};$$

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

DOI:10.6814/NCCU202200447

where $A = \{(k_2, k_3, \cdots, k_n) \mid n = 1 + k_2 + 2k_3 + \cdots + (n-1)k_n; k_i \geq 0\}$.

The formulas given by [CZ20] depend on solving an integer partition problem. On the other hand, this is not needed in [BGM20]. Thus, we think that the symbolic method is more flexible and in addition also gives asymptotic results. Hence, we will review the study done in [BGM20] as a warm-up and then apply the method to the two subclasses from [CZ20] which were not considered in [BGM20]. We conclude with some numerical results in Table 3.1.

n	g_n	m_n	c_n
1	1	1	1
2	3	1	3
3	36	6	27
4	723	69	399
5	20280	1050	8205
6	730755	20025	216315
as $n \to \infty$	$h_1 \approx 0.1339$	$h_1 \approx 0.1802$	$h_1 \approx 0.2706$
$\sim h_1 h_2^n n^{n-1}$	$h_2 \approx 2.9430$	$h_2 \approx 1.5440$	$h_2 \approx 2.1442$

Table 3.1: g_n, m_n and c_n are the number of galled, normal and one-component galled trees with n leaves, respectively.

3.1 Review: Galled Trees

Exact and Asymptotic Enumeration

We start with the specification of galled trees. Since each biconnected component of a galled tree has at most one reticulation node, we observe that each biconnected component consists either of a single edge or of tree node on top and a reticulation node at the bottom with two node-disjoint paths from top to bottom (a reticulation cycle). Also, recall that there are no parallel edges; thus, the two paths contain at least one other node. Hence, galled trees can be classified as follows:

DOI:10.6814/NCCU202200447

(1) A single leaf;

(2) The tree node below the root is not part of a reticulation cycle;

(3) The tree node below the root is at the top of a reticulation cycle and one of the paths to the bottom node has no other node;

(4) The tree node below the root is at the top of a reticulation cycle and both paths to the bottom node have at least one other node.

The description is shown by the figure:

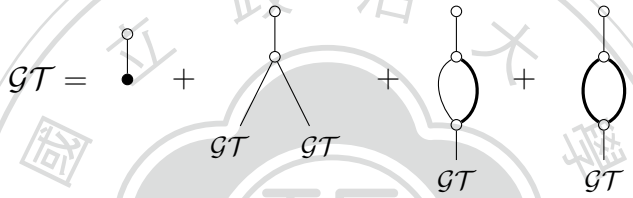


Figure 3.1: The specification of galled trees. A thick path indicates that there is at least one node on it.

Denote by $G(z) = \sum_{n\geq 0} g_n \frac{z^n}{n!}$ the generating function of galled trees. In (1), the single leaf is an atomic class having z as its generating function. In (2), there are two galled trees below the tree node which is translated into $\frac{1}{2}G(z)^2$. For (3) and (4), the thick path is a sequence of galled trees having at least one node and there is a galled tree at the bottom and hence their translations are $\frac{G(z)}{1-G(z)} \cdot G(z)$ and $\frac{1}{2}(\frac{G(z)}{1-G(z)})^2 \cdot G(z)$ respectively. Therefore, we have

$$G(z) = z + \frac{G(z)^2}{2} + \frac{G(z)^2}{1-G(z)} + \frac{G(z)}{2}\left(\frac{G(z)}{1-G(z)}\right)^2. \tag{3.1}$$

In order to apply the exact Lagrange inversion formula, we rearrange terms in (3.1) to obtain the form $G(z) = z\phi(G(z))$ where $\phi(z) = \frac{1}{1-\frac{z}{2}-\frac{z}{1-z}-\frac{z^2}{2(1-z)^2}}$. From this, we obtain the following result:

DOI:10.6814/NCCU202200447

Proposition 3.1. *For any $n \geq 1$, the number g_n of galled trees with n leaves is given by*

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

where $(m-i)^\star = max\{m-i,\ 1\}$.

Proof. By the Lagrange inversion formula, we have that

$$g_n = n![z^n]G(z) = n!\frac{1}{n}[z^{n-1}]\phi(z)^n = (n-1)![z^{n-1}]\phi(z)^n.$$

We first state the binomial theorem which plays an important role:

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

Repeatly using the binomial theorem when expanding $\phi(z)^n = (1-\frac{z}{2}-\frac{z}{1-z}-\frac{z^2}{2(1-z)^2})^{-n}$, we obtain that

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

DOI:10.6814/NCCU202200447

$$= \sum_{i\geq 0} \binom{n+i-1}{i} \frac{z^i}{2^i}$$
$$+ \sum_{i\geq 0} \sum_{k=1}^{i} \sum_{p=0}^{k} \sum_{j\geq 0} \binom{n+i-1}{i} \binom{i}{k} \binom{k}{p} \binom{2k-p+j-1}{j} \frac{z^{i+k-p+j}}{2^{i-p}}.$$

Therefore,

$$\begin{aligned} g_n &= (n-1)![z^{n-1}]\phi(z)^n \\ &= \frac{(2n-2)!}{2^{n-1}(n-1)!} \\ &\quad + \sum_{m=0}^{n-1} \sum_{\substack{i+k-p=m \\ 0\leq p\leq k\leq i,\ k\neq 0}} \frac{(n+i-1)!(n+k-i-2)!2^{p-i}}{(i-k)!(k-p)!p!(n-m-1)!(2k-p-1)!} \\ &= \frac{(2n-2)!}{2^{n-1}(n-1)!} \\ &\quad + \sum_{m=0}^{n-1} \sum_{i=1}^{m} \sum_{k=(m-i)^\star}^{i} \frac{(n+i-1)!(n+k-i-2)!2^{k-m}}{(i-k)!(m-i)!(k-m+i)!(n-m-1)!(m+k-i-1)!}. \end{aligned}$$

This proves the claim. □

Apart from exact enumeration, we can also derive the asymptotic behavior of g_n with the help of the asymptotic form of Lagrange inversion formula and Stirling's formula.

Theorem 3.2. *The number g_n of galled trees with n leaves is asymptotically equivalent to $h_1 h_2^n n^{n-1}$ where $h_1 = \frac{\sqrt{34}(\sqrt{17}-1)}{136} \approx 0.1339$ and $h_2 = \frac{8}{e} \approx 2.9430$.*

Proof. Again, we rewrite (3.1) into the form $G(z) = z\phi(G(z))$ where $\phi(z) = (1 - \frac{z}{2} - \frac{z}{1-z} - \frac{z^2}{2(1-z)^2})^{-1}$. Equivalently, $\phi(z) = \frac{2(1-z)^2}{(2-z)(z^2-3z+1)}$ which gives that the pole having smallest absolute value is $\frac{3-\sqrt{5}}{2}$. Therefore, $\phi(z)$ is analytic at 0 and the radius of convergence is $R = \frac{3-\sqrt{5}}{2} > 0$.

Considering the equation $\phi(z) - z\phi'(z) = 0$ and solving it by Maple, it has roots 1, $1 \pm i$ and $\frac{5\pm\sqrt{17}}{4}$. Note that $\frac{5-\sqrt{17}}{4} \in (0, R)$. Thus, set $\tau = \frac{5-\sqrt{17}}{4}$. We

DOI:10.6814/NCCU202200447

obtain $\rho = \frac{\tau}{\phi(\tau)} = \frac{1}{8}$ and $\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}} = \frac{\sqrt{17}(\sqrt{17}-1)}{136}$. Therefore, from the asymptotic form of Lagrange inversion formula,

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

Since $g_n = n![z^n]G(z)$ and the Stirling's formula gives that $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$, we have

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

which completes the proof. □

Asymptotic distribution of the number of reticulation nodes

We can also incorporate parameters into the above approach. The parameter we discuss is the number of biconnected components, or equivalently, the number of reticulation cycles. Multivariate generating functions are used to analyze additional parameters. Let $G(z,x) = \sum \frac{g_{n,k}}{n!} z^n x^k$ where $g_{n,k}$ is the number of galled trees with n leaves and k reticulation nodes. We can refine the specification of Figure. 3.1 to obtain for $G = G(z,x)$:

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

The equation above can be written as follows

$$G = z\Phi(G,x) \text{ where } \Phi(z,x) = \frac{1}{1 - \frac{z}{2} - x\frac{z}{1-z} - x\frac{z^2}{2(1-z)^2}}.$$

If all the assumptions of Theorem 2.14 are satisfied, then we obtain that the number of reticulation nodes is asymptotically normal distributed. This is summarized in the following theorem.

DOI:10.6814/NCCU202200447

Theorem 3.3. *Let X_n be the random variable counting the number of reticulation nodes in a galled tree with n leaves which is picked uniformly at random. X_n is asymptotically normal distributed. More precisely, we have*

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

where $\mu_X \approx 0.56$ and $\sigma_X^2 \approx 0.18$.

Proof. From the arguments above, $\Phi(z,x) = (1 - \frac{z}{2} - x\frac{z}{1-z} - x\frac{z^2}{2(1-z)^2})^{-1}$ is a function satisfying $G = z\Phi(G,x)$. From the definition of the expectation, we have

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

Set $F(G,z,x) = z\Phi(G,x)$, we can observe that $G = F(G,z,x)$ holds. The four hypotheses from of Theorem 3.2 are satisfied.

The system of equations

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

has a solution (G_0, z_0) with $G_0 \approx 0.2192$ and $z_0 = \frac{1}{8}$ such that $F_z\,(G_0,z_0,1) \neq 0$ and $F_{GG}\,(G_0,z_0,1) \neq 0$.

The result follows now from Theorem 2.14 and the numerical estimates of μ_X and σ_X^2 are computed from the formulas in the theorem. □

DOI:10.6814/NCCU202200447

3.2 Normal Galled Trees

Exact and Asymptotic Enumeration

The normal property requires that the two parents of each reticulation node are incomparable. With this additional restriction, we only need to make a small adjustment to the specification of galled trees. To be more clear, only case (3) in the specification of galled trees does not satisfy the condition. We still list each case for normal galled trees and again give a figure for the sake of clarity; the cases are:

(1) A single leaf;

(2) The tree node below the root is not part of a reticulation cycle;

(3) The tree node below the root is at the top of a reticulation cycle and this cycle contains two paths to the bottom node where both paths have at least one other node.

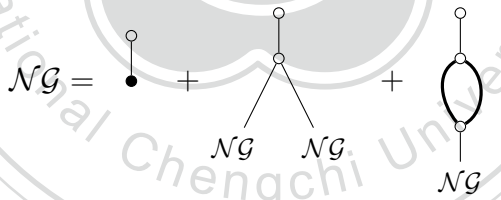


Figure 3.2: The specification of normal galled trees. A thick path indicates that there is at least one node on it.

Denote by $M(z) = \sum_{n\geq 0} m_n \frac{z^n}{n!}$ the generating function of normal galled trees. Similarly as for galled trees, the three cases can be translated into the generating functions z, $\frac{1}{2}M(z)^2$ and $\frac{1}{2}(\frac{M(z)}{1-M(z)})^2 \cdot M(z)$, respectively. Therefore, we have

$$M(z) = z + \frac{M(z)^2}{2} + \frac{M(z)}{2}\left(\frac{M(z)}{1-M(z)}\right)^2. \tag{3.2}$$

DOI:10.6814/NCCU202200447

Rearranging the equation above into the form $M(z) = z\phi(M(z))$, we have $\phi(z) = \frac{1}{1-\frac{z}{2}-\frac{z^2}{2(1-z)^2}}$. Then by Lagrange inversion formula we obtain the following result.

Proposition 3.4. *For any $n \geq 1$, the number m_n of normal galled trees with n leaves is given by*

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

Proof. Recall that $m_n = n![z^n]M(z) = (n-1)![z^{n-1}]\phi(z)^n$. Using the binomial theorem multiple times, we obtain that

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

Hence,

$$\begin{aligned}
m_n &= (n-1)![z^{n-1}]\phi(z)^n \\
&= \frac{(2n-2)!}{2^{n-1}(n-1)!} + \sum_{\ell=0}^{n-1} \sum_{\substack{i+j=\ell \\ 0\leq j\leq i}} \frac{(n+i-1)!(n+2j-\ell-2)!}{j!(i-j)!(2j-1)!(n-\ell-1)!2^i}
\end{aligned}$$

DOI:10.6814/NCCU202200447

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

This proves the claim. □

We also derive an asymptotic estimate of m_n.

Theorem 3.5. *The number m_n of normal galled trees with n leaves is asymptotically equivalent to $h_1 h_2^n n^{n-1}$ where $h_1 \approx 0.1802$ and $h_2 \approx 1.5440$.*

Proof. Again, we rewrite (3.2) into the form $M(z) = z\phi(M(z))$ where $\phi(z) = (1 - \frac{z}{2} - \frac{z^2}{2(1-z)^2})^{-1}$. Equivalently, $\phi(z) = \frac{2(1-z)^2}{-z^3+3z^2-5z+2}$. The pole of $\phi(z)$ having the smallest absolute value is $-\frac{(108+12\sqrt{177})^{1/3}}{6} + \frac{4}{(108+12\sqrt{177})^{1/3}} + 1 \approx 0.5466$.

Consider the equation $\phi(z) - z\phi'(z) = 0$, we solve it with Maple and find there is a root $\tau \approx 0.3583 \in (0, R)$ where $R = -\frac{(108+12\sqrt{177})^{1/3}}{6} + \frac{4}{(108+12\sqrt{177})^{1/3}} + 1$ is the radius of convergence. We obtain that $\rho = \frac{\tau}{\phi(\tau)} \approx 0.2383$ and $\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}} \approx 0.1274$. By Theorem 2.9, $[z^n]M(z) \sim 0.1274\ldots \frac{(0.2383\ldots)^{-n}}{\sqrt{\pi n^3}}$. Furthermore, using Stirling's formula, we have that $h_1 = \sqrt{\frac{\phi(\tau)}{\phi''(\tau)}} \approx 0.1802$ and $h_2 = \frac{\rho^{-1}}{e} \approx 1.5440$. □

From this theorem, we see that almost all galled trees are not normal.

Asymptotic distribution of the number of reticulation nodes

Similarly, we can consider the number of reticulation nodes as an additional parameter. Let $M = M(z, x) = \sum \frac{m_{n,k}}{n!} z^n x^k$. From Figure 3.2 we can derive the refinement:

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

To apply Theorem 2.14, rewrite the equation as

$$M = z\Phi(M, x), \quad \text{where } \Phi(z, x) = \frac{1}{1 - \frac{z}{2} - x\frac{z^2}{2(1-z)^2}}.$$

DOI:10.6814/NCCU202200447

Theorem 3.6. *Let X_n be the random variable counting the number of reticulation nodes in normal galled trees with n leaves. X_n is asymptotically normal distributed. More precisely, we have*

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

where $\mu_X \approx 0.23$ and $\sigma_X^2 \approx 0.05$.

Proof. The function $\Phi(M, x)$ is the one mentioned above. With the same argument as in the proof of Theorem 3.3, we have

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

Define $F(M, z, x) = z\Phi(M, x)$ such that $M = F(M, z, x)$ holds. The hypotheses of Theorem 2.14 are easily checked to hold. Also, the system

$$\begin{cases} M = F(M, z, 1) \\ 1 = F_M(M, z, 1) \end{cases}$$

admits a solution $M_0 \approx 0.3583$ and $z_0 \approx 0.2383$.

The result follows now from Theorem 2.14 with μ_X and σ_X^2 computed with Maple. □

3.3 One-Component Galled Trees

Exact and Asymptotic Enumeration

Each reticulation node of a network having the one-component property is followed by a leaf. The description of one-component is the same as the one of galled trees listed above:

DOI:10.6814/NCCU202200447

(1) A single leaf;

(2) The tree node below the root is not part of a reticulation cycle;

(3) The tree node below the root is at the top of a reticulation cycle and one of the paths to the bottom node has no other node;

(4) The tree node below the root is at the top of a reticulation cycle and both paths to the bottom node have at least one other node.

The difference is that there is a leaf rather than a network below the reticulation cycle, see the figure below for the specification.

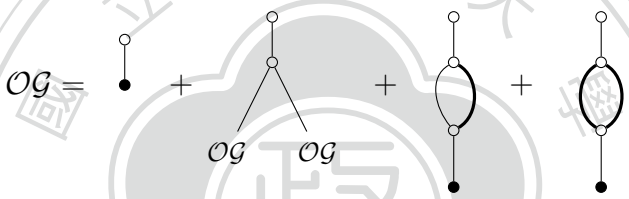


Figure 3.3: The specification of one-component galled trees. Thick paths indicate that there is at least one node on it and black nodes represent labelled leaves.

Denote by $C(z) = \sum_{n\geq 0} c_n \frac{z^n}{n!}$ the generating function of one-component galled trees. Similar as for galled trees, each case translates into the generating functions z, $\frac{1}{2}C(z)^2$, $\frac{C(z)}{1-C(z)} \cdot z$ and $\frac{1}{2}(\frac{C(z)}{1-C(z)})^2 \cdot z$, respectively.

Thus, we obtain that

$$C(z) = z + \frac{C(z)^2}{2} + z\frac{C(z)}{1-C(z)} + \frac{z}{2}\left(\frac{C(z)}{1-C(z)}\right)^2. \tag{3.3}$$

Proposition 3.7. *For any $n \geq 1$, the number c_n of one-component galled trees with n leaves is given by*

$$c_n = \sum_{m=0}^{n-1}\sum_{i=0}^{m}\sum_{\ell}^{(m-i)/2} \frac{(-1)^{m-i-2\ell}(2n+i-1)!(2n-m-2)!n!2^{m-\ell-n+1}}{(2n-1)!(n-m-1)!(n-m+i+\ell)!i!\ell!(m-i-2\ell)!}.$$

DOI:10.6814/NCCU202200447

Proof. Note that $C(z) = z\phi(C(z))$ with

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

which is obtained after rearranging (3.3). Applying the binomial theorem several times, we have

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

Hence,

$$\begin{aligned}
c_n &= (n-1)![z^{n-1}]\phi(z)^n \\
&= \sum_{m=0}^{n-1}\sum_{\substack{i+k+\ell=m \\ 0\leq\ell\leq k}}(-1)^{k-\ell}\frac{(2n+i-1)!(2n-m-2)!n!2^{i+k+1-n}}{(2n-1)!(n-m-1)!(n-k)!i!\ell!(k-\ell)!} \\
&= \sum_{m=0}^{n-1}\sum_{i=0}^{m}\sum_{\ell}^{(m-i)/2}\frac{(-1)^{m-i-2\ell}(2n+i-1)!(2n-m-2)!n!2^{m-\ell-n+1}}{(2n-1)!(n-m-1)!(n-m+i+\ell)!i!\ell!(m-i-2\ell)!}.
\end{aligned}$$

This proves the claim. □

Our formula is complicated. The result will be more simpler if we utilize $c_{n,k}$ in [CZ20] and solve it by Maple. This gives

DOI:10.6814/NCCU202200447

$$
\begin{aligned}
c_n &= \sum_{k=0}^{n-1} c_{n,k} \\
&= \sum_{k=0}^{n-1} \binom{n}{k} \frac{1}{2^{n+k-1}} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \frac{j!(2n+2j-2)!}{(2j)!(n+j-1)!} \\
&= \frac{n!}{2n-1} \sum_{i=0}^{n} \binom{n-1/2}{n-i} \binom{1/2-n}{i} (-1)^i.
\end{aligned}
$$

The asymptotic behaviour of the number of one-component galled trees is derived in the next result.

Theorem 3.8. *The number c_n of one-component galled trees with n leaves is asymptotically equivalent to $h_1 h_2^n n^{n-1}$ where $h_1 \approx 0.2706$ and $h_2 \approx 2.1442$.*

Proof. Again we rewrite equation (3.3) into the form $C(z) = z\phi(C(z))$ with $\phi(z) = \frac{1+\frac{z}{1-z}+\frac{z^2}{2(1-z)^2}}{1-\frac{z}{2}}$. Equivalently, $\phi(z) = \frac{-z^2+2z-2}{(z-1)^2(z-2)}$ from which we see that the radius of convergence $R = 1$.

Solving the equation $\phi(z) - z\phi'(z) = 0$ with Maple, it has roots $1 - i\sqrt{1+\sqrt{2}}$, $1 + i\sqrt{1+\sqrt{2}}$, $1 - \sqrt{\sqrt{2}-1}$ and $1 + \sqrt{\sqrt{2}-1}$. Note that $1 - \sqrt{\sqrt{2}-1} \in (0, R)$ thus $\tau = 1 - \sqrt{\sqrt{2}-1}$. We obtain $\rho = \frac{\tau}{\phi(\tau)} = -2\sqrt{2} + 3 \approx 0.1716$ and $\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)}} \approx 0.1913$. Using Theorem 2.9 and applying Stirling's formula, we have

$$
c_n = (n-1)![z^n]C(z) \sim 0.2706\ldots(2.1442\ldots)^n n^{n-1}
$$

which is the claimed result. □

Asymptotic distribution of the number of reticulation nodes

Taking the number of reticulation nodes into account and refining the generating function with a new variable x, we have $C = C(z, x) = \sum \frac{c_{n,k}}{n!} z^n x^k$ which

DOI:10.6814/NCCU202200447

satisfies

$$C = z + \frac{C^2}{2} + x\frac{zC}{1-C} + x\frac{z}{2}\left(\frac{C}{1-C}\right)^2$$

Rearranging terms gives that $C = z\Phi(C, x)$ with

$$\Phi(z, x) = \frac{1 + x\frac{z}{1-z} + x\frac{z^2}{2(1-z)^2}}{1 - \frac{z}{2}}.$$

Theorem 3.9. *Let X_n be the random variable counting the number of reticulation nodes in one-component galled trees with n leaves. X_n is asymptotically normal distributed. More precisely, we have*

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

where $\mu_X \approx 0.41$ and $\sigma_X^2 \approx 0.12$.

Proof. With the same argument as in the proof of Theorem 3.3, we have

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

Define $F(C, z, x) = z\Phi(C, x)$ such that $C = F = (C, z, x)$ holds where $\Phi(C, x)$ is the function mentioned above. The hypotheses of Theorem of 2.14 are easily checked to hold. Also, the system

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

has a solution $C_0 \approx 0.3564$ and $z_0 \approx 0.1716$. The result follows from Theorem 2.14 with μ_X and σ_X^2 computed with Maple. □

DOI:10.6814/NCCU202200447

Chapter 4

Tree-child Networks with Few Reticulation Nodes

In this chapter, we will study tree-child networks. Denote by $\mathrm{TC}_{n,k}$ the number of tree-child networks with n leaves and k reticulation nodes.

The first section in this chapter is concerned with exact enumeration of tree-child networks with few reticulation nodes. Exact formulas for specific k's have been given in many papers.

Cardona and Zhang used component graphs to construct tree-child networks from one-component tree-child networks and gave the following formulas for $k = 1$ and 2 in [CZ20]:

$$\mathrm{TC}_{n,1} = \frac{n(2n)!}{2^n n!} - 2^{n-1} n!; \tag{4.1}$$

$$\begin{aligned}\mathrm{TC}_{n,2} =& \frac{n!}{2^n} \sum_{j=1}^{n-2} \binom{2j}{j} \binom{2n-2j}{n-j} \frac{j(2j+1)(2n-j-1)}{2n-2j-1} \\ &+ n(n-1)n! 2^{n-3} - \frac{(2n-1)!n}{3 \cdot 2^{n-1}(n-2)!},\end{aligned} \tag{4.2}$$

where the formula for $k = 1$ was derived from a recurrence relation (see [Zha19]

DOI:10.6814/NCCU202200447

for more details).

Fuchs et al. described the shape of the exponential generating function of the number of tree-child networks with fixed k in [FGM19] and used it to obtain the asymptotics for this counting sequence. Even though the authors noted that exact formulas for generating functions for higher k are complicated to obtain, they gave exact formulas for small k. In a subsequent paper, they used these exact formulas to derive the following results:

$$\mathrm{TC}_{n,2} = n!\left(\frac{(n+1)n(n-1)(3n+2)}{6(2n+1)2^n}\binom{2n+2}{n+1} - 2^n n(n-1)\right); \tag{4.3}$$

$$\mathrm{TC}_{n,3} = n!\Big(\frac{(n+2)^2(n+1)n(n-1)(n-2)}{12(2n+3)2^n}\binom{2n+4}{n+2} - \frac{2^n}{48}n(n-1)(n-2)(48n+31)\Big). \tag{4.4}$$

Pons and Batle conjectured a bijection between tree-child networks and a certain class of words and found (based on their conjecture) a simple recurrence formula that allowed them to determine the number of tree-child networks for small n and k. Moreover, their result also gives formulas for small values of k, e.g.,

$$\mathrm{TC}_{n,2} = \binom{n}{2}\left((2n+1)!! - 2(2n)!! + \frac{1}{3}(2n-1)!!\right); \tag{4.5}$$

$$\mathrm{TC}_{n,3} = \binom{n}{3}\left((2n+3)!! - 3(2n+2)!! + (2n+1)!! + \frac{17}{8}(2n)!!\right). \tag{4.6}$$

We focus on constructing tree-child networks from component graphs. We will review the result in [CZ20], extend them to $k = 3$ and then compare the result with the ones above.

In the second section we will give a proof via component graphs for the asymptotic growth term of the counting sequence.

DOI:10.6814/NCCU202200447

In [FGM19], Fuchs et al. obtained the following first-order asymptotics for $\mathrm{TC}_{n,k}$ (see also [FGM21] for some corrections of the proof):

$$\mathrm{TC}_{n,k} \sim c_k \left(\frac{2}{e}\right)^n n^{n+2k-1}, \quad \text{as } n \to \infty. \tag{4.7}$$

Moreover, they computed c_k for small values of k, namely, $c_1 = \sqrt{2}$, $c_2 = \sqrt{2}$ and $c_3 = \frac{2\sqrt{2}}{3}$.

Surprisingly, with our approach based on component graphs, we do not only get the above first-order asymptotics for $\mathrm{TC}_{n,k}$, but also a simple, closed expression for c_k.

Theorem 4.1. *For the number* $\mathrm{TC}_{n,k}$ *of tree-child networks with* n *leaves and* k *reticulation nodes, we have*

$$\mathrm{TC}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!} \left(\frac{2}{e}\right)^n n^{n+2k-1}, \quad \textit{as } n \to \infty.$$

We submitted this result and it was recently accepted by Discrete Applied Mathematics; see [FHY21].

4.1 Enumeration with 1, 2 and 3 Reticulation Nodes

Component graphs are a suitable tool for decomposing tree-child network and they are also useful for counting networks. Given any component graph, the node containing the root of the corresponding networks will be viewed as a one-component tree-child networks whose children below the reticulation nodes are exactly $\{1, \ldots, k\}$. At these k nodes, we will attach other one-component networks and networks. The labels are used to distinguish their difference and to keep track of how we connect other networks.

We mentioned in Theorem 2.16 that there are $\binom{n}{k}\frac{(2n-2)!}{2^{n-1}(n-k-1)!}$ one-component

DOI:10.6814/NCCU202200447

TCNs with n leaves and k reticulation nodes. However, for our subsequent arguments, we need to make a slight adjustment to Theorem 2.16.

Theorem 4.2. *Denote by* $\mathrm{O}_{n,k}$ *the number of one-component networks with* n *leaves and* k *reticulation nodes where the labels of the leaves below the reticulation nodes are from* $\{1,\ldots,k\}$. *Then,*

$$\mathrm{O}_{n,k} = \frac{(2n-2)!}{2^{n-1}(n-k-1)!}.$$

We consider tree-child networks with 1 and 2 reticulation nodes as a warm-up to demonstrate how the enumeration process works. Then, we will deal with the more complicated case of 3 reticulation nodes.

In this section, most of the summations are simplified by Maple.

1 and 2 Reticulation Nodes

[$\mathrm{TC}_{n,1}$] There is only one component graph which generates all tree-child networks with 1 reticulation nodes; see Figure 4.1 (i). Suppose that the node on the top has j leaves of the generated network and the one below $n-j$ leaves where j ranges from 1 to $n-1$. Observe that the only reticulation node lies in the top node which implies that the network which we have to insert (the bottom one) is actually a phylogenetic tree. Finally, after inserting a network, we relabel the leaves in an order-consistent way. Hence, we obtain for $\mathrm{TC}_{n,1}$

$$\begin{aligned}
\mathrm{TC}_{n,1} =& \sum_{j=1}^{n-1} \binom{n}{j} \frac{(2j)!}{2^j(j-1)!} \cdot \frac{(2n-2j-2)!}{2^{n-j-1}(n-j-1)!} \\
=& \sum_{j=1}^{n-1} \binom{n}{j} \frac{(2j)!(2n-2j-2)!}{2^{n-1}(j-1)!(n-j-1)!} \\
=& \frac{n(2n)!}{2^n(n)!} - 2^{n-1}n!.
\end{aligned}$$

DOI:10.6814/NCCU202200447

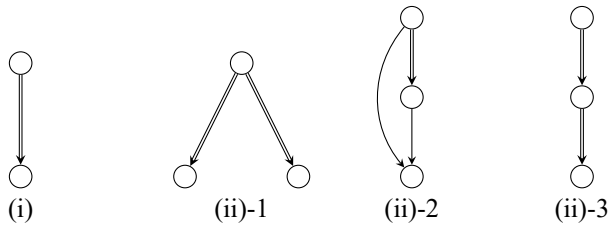


Figure 4.1: The component graphs of $\mathcal{TC}_{n,1}$ and $\mathcal{TC}_{n,2}$.

We summarize this as a theorem.

Theorem 4.3. *For the number* $\mathrm{TC}_{n,1}$ *of tree-child networks with* n *leaves and* 1 *reticulation node, we have*

$$\mathrm{TC}_{n,1} = \frac{n(2n)!}{2^n(n)!} - 2^{n-1}n!.$$

Note that it is exactly the same result as in [Zha19, Theorem 4] where it was derived using a recurrence relation.

[$\mathrm{TC}_{n,2}$] We now turn our focus on the three component graphs of tree-child networks with 2 reticulation nodes, see Figure 4.1 (ii)-1, (ii)-2 and (ii)-3. From the structure of the component graph in Figure 4.1 (ii)-1, we see that the node on top represents a one-component tree-child network with 2 reticulation nodes where the leaves below the reticulation nodes are 1 and 2. Since there is no reticulation node left, the two nodes at the bottom combined are a phylogenetic tree. As a result, we pick a phylogenetic tree and break it into two branches at the root. Then, we attach the branch having the smallest label and the other branch to reticulation node 1 and 2, respectively. Finally, we relabel in an order-consistent way. Therefore, if the one-component tree-child network on top has j leaves where j ranges from 1 to $n-2$, then we have

$$\mathrm{TC}_{n,2}^{\text{(ii)-1}} = \sum_{j=1}^{n-2} \binom{n}{j} \frac{(2j+2)!}{2^{j+1}(j-1)!} \cdot \frac{(2n-2j-2)!}{2^{n-j-1}(n-j-1)!}$$

DOI:10.6814/NCCU202200447

$$= \sum_{j=1}^{n-2} \binom{n}{j} \frac{(2j+2)!}{2^n(j-1)!} \cdot \frac{(2n-2j-2)!}{(n-j-1)!}$$
$$= \frac{(2n)!}{2^n n!}(n^3+4n^2+n) - 2^{n-3}n!(15n^2+9n). \tag{4.8}$$

The tree-child networks having the component graph from Figure 4.1 (ii)-2 can be formed by a one-component tree-child network having 1 reticulation node whose leaf is replaced by a phylogenetic tree. For the single edge that goes from the top node to the bottom node, we pick an ordered pair of edges, the first one is an edge from the top one-component network and the other from the phylogenetic tree below. Then, we create a new edge from the middle of the first edge to the middle of the second. Due to the tree-child property, it is clear that we can only pick tree edges as candidates. Moreover, some situations still violate the tree-child property even when we pick just tree edges:

(1) The starting edge is the edge below the reticulation node.

(2) The terminal edge is the root edge.

Suppose that the top one-component network has j leaves (not counting the leaf below the reticulation nodes) and the one below $n-j$ where j ranges from 1 to $n-2$. Recall that by Proposition 1.3, a network has $2n+k-1$ tree edges. Therefore, the one-component network on the top has $2j+1$ tree edges which can be chosen and $2n-2j-2$ edges can be chosen in the phylogenetic tree. Overall,

$$\begin{aligned}
\mathrm{TC}_{n,2}^{\text{(ii)-2}} &= \sum_{j=1}^{n-2} \binom{n}{j} \frac{(2j)!}{2^j(j-1)!} \cdot \frac{(2n-2j-2)!}{2^{n-j-1}(n-j-1)!} \cdot (2j+1)(2n-2j-2) \\
&= \sum_{j=1}^{n-2} \binom{n}{j} \frac{(2j+1)!}{2^{n-2}(j-1)!} \cdot \frac{(2n-2j-2)!}{(n-j-2)!} \\
&= 2^{n-2}n!(3n^2+9n) - \frac{(2n)!}{2^n n!}\left(4n^2+2n\right).
\end{aligned} \tag{4.9}$$

DOI:10.6814/NCCU202200447

The component graph in Figure 4.1 (ii)-3 is handled similarly. A tree-child network having such a component graph consists of a one-component tree-child network on top and the network that is going to be inserted is a tree-child network with one reticulation node. Suppose that the one-component network on top has j leaves and the network below is from $\mathrm{TC}_{n-j,1}$ where j is ranging from 1 to $n-2$, using Theorem 4.3, we have

$$\begin{aligned}
\mathrm{TC}_{n,2}^{\text{(ii)-3}} &= \sum_{j=1}^{n-2} \binom{n}{j} \frac{(2j)!}{2^j(j-1)!} \cdot \left(\frac{n(2n)!}{2^n n!} - 2^{n-1} n! \right) \\
&= \frac{n!}{2^n} \sum_{j=1}^{n-2} \binom{2j}{j} \binom{2n-2j}{n-j} j(n-j) - n! 2^{n-1} \sum_{j=1}^{n-2} \frac{(2j)!}{4^j j!(j-1)!} \\
&= n(n-1)n! 2^{n-3} - \frac{(2n-1)!n}{3 \cdot 2^{n-1}(n-2)!}, \qquad (4.10)
\end{aligned}$$

where we used standard identities for binomial coefficients in the last line.

Overall, $\mathrm{TC}_{n,2}$ is just the sum of $\mathrm{TC}_{n,2}^{\text{(ii)-t}}$ where $t = 1,\ 2,\ 3$. Thus,

$$\begin{aligned}
\mathrm{TC}_{n,2} &= \mathrm{TC}_{n,2}^{\text{(ii)-1}} + \mathrm{TC}_{n,2}^{\text{(ii)-2}} + \mathrm{TC}_{n,2}^{\text{(ii)-3}} \\
&= \frac{(2n)!}{3 \cdot 2^n n!} (3n+2)\, n\, (n-1) - n! 2^n n\, (n-1)\,.
\end{aligned}$$

We summarize this in the following theorem.

Theorem 4.4. *For the number* $\mathrm{TC}_{n,2}$ *of tree-child network with* n *leaves and* 2 *reticulation nodes, we have*

$$\mathrm{TC}_{n,2} = \frac{(2n)!}{3 \cdot 2^n n!} \cdot (3n+2)\, n\, (n-1) - n! 2^n n\, (n-1)\,.$$

It is not hard to see that this coincides with (4.2), (4.3) and (4.5).

DOI:10.6814/NCCU202200447

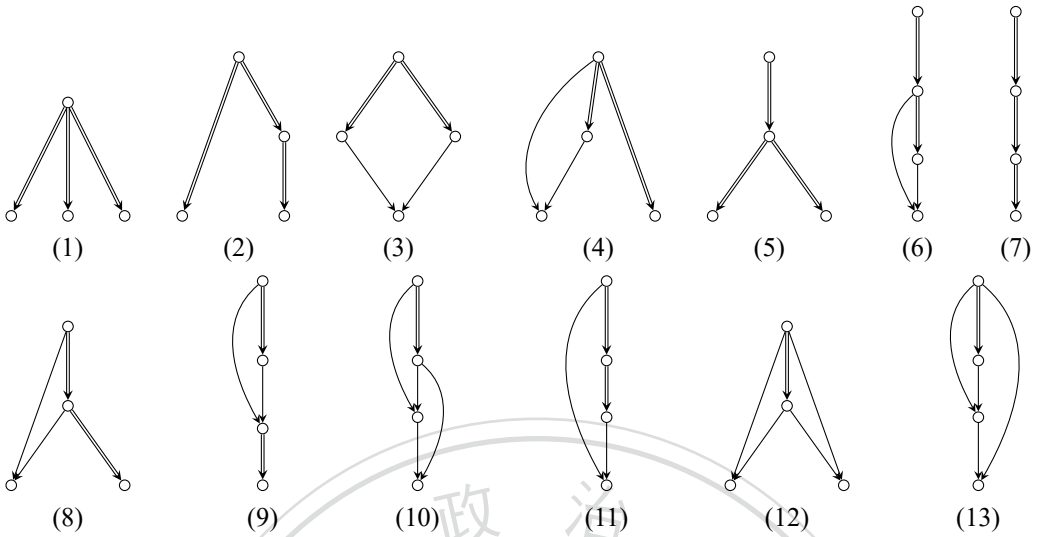


Figure 4.2: All possible component graphs of TCN with 3 reticulation nodes.

3 Reticulation Nodes

There are 13 component graphs for tree-child networks with n leaves and 3 reticulation nodes. Luckily, we do not need to count the number of tree-child networks for each single one of them; instead, some of them can be grouped according to the type of the outgoing edges of the root and handled together. The way we count is mostly the same as in the two cases ($k = 1$ and $k = 2$) before. That is, we choose a network from the $O_{j+t,t} = \frac{(2(j+t)-2)!}{2^{j+t-1}(j-1)!}$ one-component tree-child networks where t is the number of nodes connected to the root node by double edges in the component graph and j is the number of leaves that lie not below reticulation nodes. As for the single edges starting from the root node, we choose suitable pair of edges and create a new edge starting from the middle of one of these edges to the middle of the other. However, we need to consider more carefully how other networks are attached to the one-component tree-child network and how the additional edges connect different parts in the component graph.

[3 double edges, 0 single edge] The component graph is shown in Figure

DOI:10.6814/NCCU202200447

4.2 (1). Here, we will use unrooted phylogenetic trees with ℓ leaves. Note that an unrooted phylogenetic tree can be converted into a phylogenetic tree by inserting a root into one of its $2\ell-3$ edges. Therefore, the number of unrooted phylogenetic trees with ℓ leaves is

$$\frac{(2\ell-2)!}{2^{\ell-1}(\ell-1)!}\frac{1}{2\ell-3}=\frac{(2\ell-4)!}{2^{\ell-2}(\ell-2)!}.$$

Also note that we can break an unrooted phylogenetic tree into 3 parts by choosing one of its $\ell-2$ non-root nodes.

Let $G_{3,0}$ be the number of tree-child networks having the graph in Figure 4.2 (1) as their component graph. Note that the one-component tree-child network on top can have at most $n-3$ leaves. Thus, we have

$$\begin{aligned}G_{3,0}&=\sum_{j=1}^{n-3}\binom{n}{j}\frac{(2j+4)!}{2^{j+2}(j-1)!}\cdot\frac{(2n-2j-4)!}{2^{n-j-2}(n-j-2)!}(n-j-2)\\&=\frac{(2n)!}{3\cdot 2^n n!}\left(2n^5+37n^4+76n^3+53n^2+12n\right)\\&\quad-2^{n-7}n!\left(315n^4+1470n^3+1545n^2+510n\right).\end{aligned}\tag{4.11}$$

[2 double edges, 1 single edge] See Figure 4.2 (4) for the component graph. We break a phylogenetic tree into two parts using the same method as for the component graph in Figure 4.1 (ii)-1. The single edge is dealt with as for the component graph in Figure 4.1 (ii)-2. Recall that there are $2\ell+k-1$ tree edges in a network with ℓ leaves and k reticulation nodes. We need to prevent violating the tree-child property during the construction. Consequently, except for the 2 tree edges below the reticulation nodes, there are $2j+3$ tree edges which we are allowed to choose in the one-component tree-child network on the top. For the bottom part, $2(n-j)-4$ tree edges can be picked since three of them have been either deleted or are adjacent to reticulation nodes. Therefore, the number of

DOI:10.6814/NCCU202200447

tree-child networks with n leaves and 3 reticulation nodes having the component graph from Figure 4.2 (4) as their component graph is

$$\begin{aligned} G_{2,1} &= \sum_{j=1}^{n-2} \binom{n}{j} \frac{(2j+2)!}{2^{j+1}(j-1)!} \cdot \frac{(2n-2j-2)!}{2^{n-j-1}(n-j-1)!}(2j+3)(2n-2j-4) \\ &= 2^{n-5} n! \left(35n^4 + 570n^3 + 925n^2 + 390n\right) \\ &\quad - \frac{(2n)!}{2^n n!} \left(12n^4 + 50n^3 + 46n^2 + 12n\right). \end{aligned} \tag{4.12}$$

[2 double edges, 0 single edge] There are 2 component graphs of this kind, namely the ones from Figure 4.2 (2) and Figure 4.2 (3). For the one in Figure 4.2 (2), which is less complicated, the two networks that have to be attached to the 2 leaves below the reticulation nodes are a phylogenetic tree and a tree-child network with 1 reticulation node. For the one in Figure 4.2 (3), we pick two phylogenetic trees with ℓ leaves for the first tree and $n-j-\ell$ leaves for the second tree. Then we join them with an edge which starts from the first to the second. For the starting points, there are $2\ell-1$ edges that can be used; for the end points, there are $2(n-j-\ell)-2$ edges since the root edge has to be avoided. Also, since the bottom part is symmetric, we need to multiply by $1/2$ in order to avoid double counting. Finally, note that the two trees need to have at least 3 leaves so that we are able to connect them via an edge.

Overall, we have for the number $G^{(2)}$ of tree-child networks arising from the somponent graphs in Figure 4.2 (2):

$$\begin{aligned} G^{(2)} &= \sum_{j=1}^{n-3} \binom{n}{j} \frac{(2j+2)!}{2^{j+1}(j-1)!} \sum_{\ell=1}^{n-j-2} \binom{n-j}{\ell} \frac{(2\ell-2)!}{2^{\ell-1}(\ell-1)!} \cdot \mathrm{TC}_{n-j-\ell,1} \\ &= \sum_{j=1}^{n-3} \binom{n}{j} \frac{(2j+2)!}{2^{j+1}(j-1)!} \sum_{\ell=1}^{n-j-2} \binom{n-j}{\ell} \frac{(2\ell-2)!}{2^{\ell-1}(\ell-1)!} \left(\frac{\tilde{\ell}(2\tilde{\ell})!}{2^{\tilde{\ell}}(\tilde{\ell})!} - 2^{\tilde{\ell}-1}(\tilde{\ell})! \right) \end{aligned}$$

DOI:10.6814/NCCU202200447

$$
\begin{aligned}
=&\frac{n!}{2^n}\sum_{j=1}^{n-3}\frac{(2j+2)!}{j!(j-1)!}\sum_{\ell=1}^{n-j-2}\frac{(2\ell-2)!}{\ell!(\ell-1)!}\left(\frac{(2\tilde{\ell}-2)!}{\tilde{\ell}!(\tilde{\ell}-1)!}-2^{2\tilde{\ell}-1}\right)\\
=&2^n n!\cdot n(n-1)\left(\frac{5}{64}n^2+\frac{31}{64}n+\frac{14}{64}\right)\\
&-\frac{(2n)!}{2^n n!}\frac{1}{128}n(n-1)(2n+1)\left(\frac{2}{7}n+\frac{8}{35}\right),
\end{aligned}
$$

where $\tilde{\ell}=n-j-\ell$.

Moreover, for the number of tree-child networks arising from Figure 4.2 (3), we have

$$
\begin{aligned}
G^{(3)}=&\frac{1}{2}\sum_{j=1}^{n-3}\binom{n}{j}\frac{(2j+2)!}{2^{j+1}(j-1)!}\sum_{\ell=1}^{n-j-2}\binom{n-j}{\ell}\frac{(2\ell-1)(2\ell-2)!}{2^{\ell-1}(\ell-1)!}\frac{(2\tilde{\ell}-2)(2\tilde{\ell}-2)!}{2^{\tilde{\ell}-1}(\tilde{\ell}-1)!}\\
=&\frac{n!}{2^{n-1}}\sum_{j=1}^{n-3}\frac{(2j+2)!}{j!\,(j-1)!}\sum_{\ell=1}^{n-j-1}(2\ell-1)\frac{(2\ell-2)!}{\ell!(\ell-1)!}(\tilde{\ell}-1)\frac{(2\tilde{\ell}-2)!}{\tilde{\ell}!(\tilde{\ell}-1)!}\\
=&\frac{(2n+1)!}{2^{n-1}(n-1)!}\frac{(5n^2+69n+66)}{70}-n!2^n\frac{15}{16}\left(n^3+\frac{17}{5}n^2+2n\right),
\end{aligned}
$$

where $\tilde{\ell}=n-j-\ell$.

Therefore, if we let $G_{2,0}$ denote the number of tree-child networks whose correspondent component graphs have 2 double edges and 0 single edges, then we have

$$
\begin{aligned}
G_{2,0}=&G^{(2)}+G^{(3)}\\
=&2^n n!\frac{1}{64}\left(5n^4-34n^3-221n^2-134n\right)\\
&+\frac{(2n+1)!}{2^{n+1}n!}\left(-\frac{2}{7}n^3+\frac{142}{35}n^2+\frac{148}{35}n\right).
\end{aligned}
\tag{4.13}
$$

[1 double edge, 2 single edges] Here, we need to consider the component graphs in Figure 4.2 (12),(13). In both cases, the bottom part is a phylogenetic tree. Moreover, there are 2 edges connecting the upper one-component tree-child

DOI:10.6814/NCCU202200447

network with the bottom part. For this construction, if there is no path containing the end points of the two chosen edges, then the component graph from Figure 4.2 (12) will be the correspondent component graph for the resulting network; otherwise, it will be the component graph from Figure 4.2 (13).

Next, we discuss the number of possible edges which can be chosen. Observe that after an edge of the upper part is attached to the bottom part, it creates a new edge in the upper part which can be chosen as well. Therefore, the number of possible edges from the top is $(2j+1)(2j+2)$. For the bottom part, which is a phylogenetic tree with $n-j$ leaves, the two endpoints of the edges must be distinct and further, they cannot have a same source. That is, we have $\binom{2n-2j-2}{2}-(n-j-1)$ edges which are feasible. Note that there must be at least 3 edges not adjacent to the reticulation node in the phylogenetic trees. Therefore, j ranges from 1 to $n-3$.

Overall, the number of tree-child networks with 3 reticulation nodes whose component graphs have 1 double edge and 2 single edges is

$$\begin{aligned} G_{1,2} =& \sum_{j=1}^{n-3} \binom{n}{j} \frac{(2j)!}{2^j (j-1)!} \frac{(2\hat{j}-2)!}{2^{\hat{j}-1}(\hat{j}-1)!} (2j+1)(2j+2) \left(\binom{2\hat{j}-2}{2} - (\hat{j}-1) \right) \\ =& \frac{n!}{2^{n-1}} \sum_{j=1}^{n-3} \frac{(2j)!}{j!(j-1)!} \frac{(2\hat{j})!}{\hat{j}!(\hat{j}-1)!} \cdot (2j+1)(2j+2)2(\hat{j}-1)(\hat{j}-2) \\ =& 2^{n-5} n! (5n^4 - 94n^3 - 425n^2 - 254n) + \frac{(2n)!}{2^{n-3} n!} n(n+1)(2n+1), \end{aligned} \tag{4.14}$$

where $\hat{j} = n - j$.

[1 double edge, 1 single edge] The component graphs belonging to this case are in Figure 4.2 (8), (9), (10) and (11). The basic structure in these cases is a one-component tree-child network on top of a tree-child network with 1 reticulation node. Note that a tree-child network with 1 reticulation node is actually a galled tree. The terminal node of the single edge connecting the two parts determines

DOI:10.6814/NCCU202200447

which kind of graph it is. The edges of a tree-child network with 1 reticulation node are partition into four parts: (a) the path from the root to the tree node on the top of the reticulation cycle, (b) edges on the reticulation cycles, (c) edges that lie below the reticulation node and (d) all other edges. Depending on the terminal node of the edge from the top one-component network to the galled tree, we obtain the following pairs: (a)-(9), (b)-(10), (c)-(11) and (d)-(8).

For the possible edges to be chosen, there are $2j+1$ feasible edges in the one-component tree-child network and $2n-2j-4$ edges in the tree-child network with $n-j$ leaves and 1 reticulation node since the tree edge below the reticulation node and the two tree edges adjacent to the parents of the reticulation node are forbidden and the root edge gets removed.

Consequently, we have

$$\begin{aligned} G_{1,1} =& \sum_{j=1}^{n-1} \binom{n}{j} \frac{(2j)!}{2^j(j-1)!} \cdot \mathrm{TC}_{n-j,1}(2j+1)(2n-2j-4) \\ =& \sum_{j=1}^{n-1} \binom{n}{j} \frac{(2j)!}{2^j(j-1)!} \left(\frac{\hat{j}(2\hat{j})!}{2^{\hat{j}}(\hat{j})!} - 2^{\hat{j}-1}\hat{j}! \right) (2j+1)(2\hat{j}-4) \\ =& 2^n n! \frac{1}{32}(3n^4 - 14n^3 - 3n^2 + 14n) - \frac{(2n+1)!}{2^{n+1}n!}\frac{1}{35}(8n^3 - 52n^2 + 44n), \end{aligned} \tag{4.15}$$

where $\hat{j} = n - j$.

[1 double edge, 0 single edge] See the component graphs in Figure 4.2 (5), (6) and (7). This case is simply a one-component tree-child network connected with a tree-child network with 2 reticulation nodes. We have dealt with $\mathrm{TC}_{n,2}$ in Theorem 4.4. Thus, the number of networks having these component graphs is

$$G_{1,0} = \sum_{j=1}^{n-1} \binom{n}{j} \frac{(2j)!}{2^j(j-1)!} \cdot \mathrm{TC}_{\hat{j},2}$$

DOI:10.6814/NCCU202200447

$$=2^n n!\frac{5}{128}\left(n^4-\frac{26}{15}n^3-\frac{9}{5}n^2+\frac{38}{15}n\right)-\frac{(2n+1)!}{2^{n+1}n!}\frac{16}{105}(n^3-3n^2+2n), \tag{4.16}$$

where $\hat{j}=n-j$.

The number of tree-child networks with n leaves and 3 reticulation nodes is consequently the sum of all the above cases; see Table 4.1 for a summary. We summarize the result in the next theorem.

Theorem 4.5. *For the number* $\mathrm{TC}_{n,3}$ *of tree-child networks with* n *leaves and* 3 *reticulation nodes, we have*

$$\mathrm{TC}_{n,3}=\frac{(2n+1)!}{2^{n+1}n!}n\,(n-1)\,(n-2)\left(\frac{2n}{3}+\frac{4}{3}\right)-2^n n!n\,(n-1)\,(n-2)\left(n+\frac{31}{48}\right).$$

Our formula shows that the counting sequence of $\mathrm{TC}_{n,3}$ for $n\geq 4$ starts with

$$2544,\ 154500,\ 6494400,\ 241204950,\ 8609378400,\ldots$$

which is the same results as obtained from (4.4) and (4.6).

$\mathrm{TC}_{n,3}$	$\frac{(2n+1)!}{2^{n+1}n!}n\,(n-1)\,(n-2)\left(\frac{2n}{3}+\frac{4}{3}\right)-2^n n!n\,(n-1)\,(n-2)\left(n+\frac{31}{48}\right)$
$G_{3,0}$	$\frac{(2n)!}{3\cdot 2^n n!}\cdot\left(2n^5+37n^4+76n^3+53n^2+12n\right)-2^{n-7}n!\left(315n^4+1470n^3+1545n^2+510n\right)$
$G_{2,1}$	$2^{n-5}n!\cdot\left(35n^4+570n^3+925n^2+390n\right)-\frac{(2n)!}{2^n n!}\left(12n^4+50n^3+46n^2+12n\right)$
$G_{2,0}$	$2^n n!\frac{1}{64}\left(5n^4-34n^3-221n^2-134n\right)+\frac{(2n+1)!}{2^{n+1}n!}\cdot\left(-\frac{2}{7}n^3+\frac{142}{35}n^2+\frac{148}{35}n\right)$
$G_{1,2}$	$2^{n-5}n!\cdot(5n^4-94n^3-425n^2-254n)+\frac{(2n)!}{2^{n-3}n!}\cdot n(n+1)(2n+1)$
$G_{1,1}$	$2^n n!\frac{1}{32}(3n^4-14n^3-3n^2+14n)-\frac{(2n+1)!}{2^{n+1}n!}\frac{1}{35}(8n^3-52n^2+44n)$
$G_{1,0}$	$2^n n!\frac{5}{128}(n^4-\frac{26}{15}n^3-\frac{9}{5}n^2+\frac{38}{15}n)-\frac{(2n+1)!}{2^{n+1}n!}\frac{16}{105}(n^3-3n^2+2n)$

Table 4.1: The counting formulas for the numbers of tree-child networks with 3 reticulation nodes corresponding to the different groups of component graphs considered in the proof of Theorem 4.5.

DOI:10.6814/NCCU202200447

4.2 Asymptotic Enumeration

In this section we will prove Theorem 4.1, i.e., we will show the expansion

$$\mathrm{TC}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1} \quad \text{as } n \to \infty.$$

We will use the groups of component graphs from the last section for $k = 3$. When computing an asymptotic expansion for each group, we found that the so-called *star component graph* namely, the case [3 double edges, 0 single edges], contributes the most (see Table 4.2 for the numerical results). This will turn out to be true also for general k. Consequently, we will divide the component graphs into two groups, namely the star component graph and the remaining graphs, and treat their asymptotics separately.

n	$\mathrm{TC}_{n,3}$	$\mathrm{S}_{n,3}$	$\mathrm{R}_{n,3}$	R/S
5	1.5450×10^5	2.7900×10^4	1.2660×10^5	4.54
10	1.1116×10^{13}	3.4142×10^{12}	7.7015×10^{12}	2.26
15	1.1384×10^{21}	4.3490×10^{20}	7.0352×10^{20}	1.62
25	1.8594×10^{38}	8.8048×10^{37}	9.7896×10^{37}	1.11
40	3.2014×10^{66}	1.7680×10^{66}	4.4334×10^{66}	0.81

Table 4.2: The number of tree-child networks with n leaves and 3 reticulation nodes arising from different groups of component graphs where $\mathrm{S}_{n,3}$ is the number for the star component graph and $\mathrm{R}_{n,3}$ is the number for the remaining graphs.

Star Component Graph

For convenience, we use $\mathrm{S}_{n,k}$ to denote the number of all tree-child networks with n leaves and k reticulation nodes having the star component graph with k leaves as component graph, i.e., the component graph consisting of a root to which k children are attached via double edges. We have given formulas for the cases

DOI:10.6814/NCCU202200447

$k \leq 3$ in the last section. Now we give the formula for general k.

Proposition 4.6. *For $k \geq 1$, we have,*

$$\mathrm{S}_{n,k} = \frac{n!}{2^{n-1}(k-1)!} \sum_{j=1}^{n-k} \frac{(2j+2k-2)!}{j!(j-1)!} \cdot \frac{(2n-2j-k-1)!}{(n-j-k)!(n-j)!}.$$

Proof. Recall that we have

$$\mathrm{O}_{\ell,k} = \frac{(2\ell-2)!}{2^{\ell-1}(\ell-k-1)!}.$$

Further, recall that $T(z)$ denotes the exponential generating function of the number of phylogenetic trees. Now, to construct tree-child networks with component graph the star component graph, we pick a one-component tree-child network with k reticulation nodes having $j+k$ leaves and replace all the leaves below the reticulation nodes by phylogenetic trees. Then, we relabel all leaves in an order-consistent way such that there are only labels from $\{1, \ldots, n\}$.

From this, we have the following

$$\mathrm{S}_{n,k} = \sum_{j=1}^{n-k} \binom{n}{j} \frac{(2j+2k-2)!}{2^{j+k-1}(j-1)!} \cdot \frac{1}{k!}(n-j)! \left[z^{n-j}\right] T(z)^k. \tag{4.17}$$

Recall that from Example 2.5, we know that

$$T(z) = z\left(\frac{1}{1-\frac{T(z)}{2}}\right), \quad \text{and} \quad T(z) = 1 - \sqrt{1-2z}.$$

Therefore, by the Lagrange inversion formula

$$[z^n] f(C(z)) = \frac{1}{n}[z^{n-1}] f'(z)\phi(z)^n.$$

DOI:10.6814/NCCU202200447

and the binomial theorem

$$(1-z)^{-n} = \sum_{i\geq 0} \binom{n+i-1}{i} z^i,$$

we have

$$\begin{aligned}\left[z^{n-j}\right] T(z)^k &= \frac{k}{(n-j)} \left[z^{n-j-1}\right] k z^{k-1} \left(1-\frac{z}{2}\right)^{-(n-j)} \\ &= \frac{k}{(n-j)} 2^{j+k-n} \binom{2n-2j-k-1}{n-j-k}.\end{aligned}$$

Plugging this into (4.17), the claimed result is obtained with an easy computation. □

Next, the asymptotics of $\mathrm{S}_{n,k}$ can be obtained by applying the Laplace method to the sum from the last proposition.

Theorem 4.7. *For $k \geq 1$, we have*

$$\mathrm{S}_{n,k} \sim \frac{\sqrt{2} d_k}{2(k-1)!} \left(\frac{2}{e}\right)^n n^{n+2k-1}, \quad \textit{as } n \to \infty,$$

where

$$d_k := \sum_{j\geq 0} \frac{(2j+k-1)!}{j!(j+k)!} 4^{-j}.$$

Proof. In the previous proposition, the asymptotics of the term outside the summation can be deduced from Stirling's formula:

$$\frac{n!}{2^{n-1}(k-1)!} \sim \frac{2\sqrt{2\pi}}{(k-1)!} \left(\frac{1}{2e}\right)^n n^{n+1/2}, \quad \text{as } n \to \infty. \tag{4.18}$$

DOI:10.6814/NCCU202200447

Let $\Sigma_{n,k}$ be the remaining terms, that is

$$\begin{aligned}\Sigma_{n,k} &:= \sum_{j=1}^{n-k} \frac{(2j+2k-2)!}{j!(j-1)!} \cdot \frac{(2n-2j-k-1)!}{(n-j-k)!(n-j)!} \\ &= \sum_{j=0}^{n-k-1} \frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!} \cdot \frac{(2j+k-1)!}{j!(j+k)!}. \end{aligned} \tag{4.19}$$

The ratio of consecutive terms is

$$\frac{(2j+k+1)(2j+k)(n-j)(n-j-1)}{(2n+2k-2j-1)(2n+2k-2j)(j+1)(j+k+1)}$$

which is smaller than 1. Thus, we can observe that the main contribution to the sum (4.19) comes from the terms with small j. Now, expanding the first term inside (4.19) gives

$$\frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!} = \frac{1}{\sqrt{\pi}} 4^{n-j-1} n^{2k-3/2} \cdot e^{\frac{1}{n}\left(-k^2+\frac{1}{2}-\frac{4k-3}{2}j+\mathcal{O}\left(\frac{1+j^2}{n^2}\right)\right)}.$$

uniformly for j with $j = o(n)$. Thus,

$$\frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!} = \frac{1}{\sqrt{\pi}} 4^{n-j-1} n^{2k-3/2} \left(1+\mathcal{O}\left(\frac{1+j}{n}\right)\right). \tag{4.20}$$

We now split $\Sigma_{n,k}$ into two parts:

$$\begin{aligned}\sum_{j=0}^{n-k-1} \frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!} &= \sum_{j\le\sqrt{n}} + \sum_{j\ge\sqrt{n}} \\ &= \sum_{j\le\sqrt{n}} \frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!} + \Delta\end{aligned}$$

DOI:10.6814/NCCU202200447

$$\sim \sum_{j\le\sqrt{n}} \frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!}$$

where Δ is an exponentially small term.

Plugging (4.20 into this sum gives, as $n \to \infty$,

$$\begin{aligned}\Sigma_{n,k} &\sim \sum_{j\le j_0} \frac{(2n-2j-2)!}{(n-j-k)!(n-j-k-1)!} \cdot \frac{(2j+k-1)!}{j!(j+k)!} \\ &\sim \frac{d_k}{\sqrt{\pi}} 4^{n-1} n^{2k-3/2}\end{aligned}$$

and multiplying this with (4.18) gives the claimed result. □

What is left is to simplify d_k.

Lemma 4.8. *For $k \ge 1$, we have*

$$\sum_{j\ge 0} \frac{(2j+k-1)!}{j!(j+k)!} 4^{-j} = \frac{2^k}{k}.$$

Proof. Set $A(z) = \sum_{j\ge 0} \frac{(2j+k-1)!}{(j+k)!j!} z^j$ and again recall the binomial theorem.

$$(1-z)^{-n} = \sum_{i\ge 0} \binom{n+i-1}{i} z^i$$

and the Lagrange inversion formula

$$[z^n] f(C(z)) = \frac{1}{n} [z^{n-1}] f'(z) \phi(z)^n.$$

Now, note that

$$\begin{aligned}\left[z^j\right] A(z) =& \frac{(2j+k-1)!}{(j+k)!j!} \\ =& \frac{1}{(j+k)} \binom{2j+k-1}{j}\end{aligned}$$

DOI:10.6814/NCCU202200447

$$
\begin{aligned}
&= \frac{1}{(j+k)} \left[z^j\right] (1-z)^{-(j+k)} \\
&= \frac{1}{k} \frac{1}{(j+k)} \left[z^{j+k-1}\right] k z^{k-1} (1-z)^{-(j+k)} \\
&= \frac{1}{k} \left[z^{j+k}\right] C(z)^k \\
&= \frac{1}{k} \left[z^j\right] \left(\frac{C(z)}{z}\right)^k
\end{aligned}
$$

where the second last step involves an application of the Lagrange inversion formula by setting $f(z) = z^k$ and where $C(z)$ is a function such that $\frac{1}{1-C(z)} = \frac{C(z)}{z}$, or equivalently, $C(z)^2 - C(z) + z = 0$.

Thus,

$$C(z) = \frac{1 - \sqrt{1-4z}}{2}.$$

Therefore,

$$A(z) = \frac{1}{k} \cdot \left(\frac{1 - \sqrt{1-4z}}{2z}\right)^k$$

which gives $A(1/4) = \frac{2^k}{k}$. □

The first order asymptotics of $S_{n,k}$ is now straightforwardly obtained by combining the last three results.

Theorem 4.9. *For the number* $\mathrm{S}_{n,k}$ *of tree-child networks with* n *leaves and* k *reticulation nodes arising from the star component graph, we have, as* $n \to \infty$*,*

$$\mathrm{S}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!} \left(\frac{2}{e}\right)^n n^{n+2k-1}.$$

Remaining Component Graphs

Denote by $\mathrm{R}_{n,k}$ the number of networks arising from the non-star component graphs. We will show that $\mathrm{R}_{n,k}$ contributes asymptotically less than $\mathrm{S}_{n,k}$ which then completes the proof of Theorem 4.1.

DOI:10.6814/NCCU202200447

Before estimating $\mathrm{R}_{n,k}$, we estimate T_n and $\mathrm{O}_{n,k}$.

Proposition 4.10. *We have,*

$$T_n = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n-1}\right)$$

Proof. See Example 2.11. □

Proposition 4.11. *We have,*

$$\mathrm{O}_{n,k} = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+k-1}\right)$$

Proof. Recall that

$$\mathrm{O}_{n,k} = \frac{(2n-2)!}{2^{n-1}(n-k-1)!}.$$

Thus, by Stirling's formula, $n! \sim \sqrt{2\pi n}\,(\frac{n}{e})^n$, we have

$$\begin{aligned}\mathrm{O}_{n,k} &= \mathcal{O}\left(2^{-(n-1)} \cdot \frac{\sqrt{2\pi(2n-2)}}{\sqrt{2\pi(n-k-1)}} \cdot \frac{(2n-2)^{2n-2}}{(n-k-1)^{n-k-1}} \cdot e^{-(2n-2)+(n-k+1)}\right) \\ &= \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+k-1}\right)\end{aligned}$$

which is the claimed result. □

The following lemma gives bounds for certain sums.

Lemma 4.12. *Let $\alpha_0, \ldots, \alpha_t$ be real numbers none of which equal to -1. Set*

$$s := \#\{0 \le j \le t \ : \ \alpha_j > -1\}$$

Then,

$$\sum_{\ell_0+\cdots+\ell_t=\ell} \ell_0^{\alpha_0} \cdots \ell_t^{\alpha_t} = \mathcal{O}\left(\ell^{s-1+\sum_{\alpha_j>-1}\alpha_j}\right),$$

DOI:10.6814/NCCU202200447

where the sum runs over all positive integers $\ell_0, \ldots, \ell_t$.

Proof. We first assume that $0 < s < t + 1$. Then, we can write the sum as

$$\sum_{\ell_0+\cdots+\ell_t=\ell} \ell_0^{\alpha_0} \cdots \ell_t^{\alpha_t} = \sum_{i=1}^{\ell-1} \left(\sum_{\sum_{\alpha_j<-1} \ell_j=\ell-i} \left(\prod_{\alpha_j<-1} \ell_j^{\alpha_j} \right) \right) \left(\sum_{\sum_{\alpha_j>-1} \ell_j=i} \left(\prod_{\alpha_j>-1} \ell_j^{\alpha_j} \right) \right).$$

We will start by estimating the two terms inside this sum.

For the first term, we have

$$\sum_{\sum_{\alpha_j<-1} \ell_j=\ell-i} \left(\prod_{\alpha_j<-1} \ell_j^{\alpha_j} \right) = \mathcal{O}\left((\ell-i)^{\alpha}\right),$$

where $\alpha = \max\{\alpha_j \ : \ \alpha_j < -1\}$. This follows because at least one of the ℓ_j's with $\sum_{\alpha_j<-1} \ell_j = \ell - i$ is at least $(\ell - i)/(t + 1 - s)$ (giving the claimed upper bound) and the series $\sum_{\ell=1}^{\infty} \ell^{\beta}$ converges for all $\beta < -1$ (giving a constant upper bound for the remaining ℓ_j's).

For the second term, by approximating by an integral,

$$\begin{aligned} \sum_{\sum_{\alpha_j>-1} \ell_j=i} \left(\prod_{\alpha_j>-1} \ell_j^{\alpha_j} \right) &= \mathcal{O}\left(i^{s-1+\sum_{\alpha_j>-1} \alpha_j} \int_{\sum_{\alpha_j>-1} x_j=1} \left(\prod_{\alpha_j>-1} x_j^{\alpha_j} \right) \mathrm{d}\mathbf{x} \right) \\ &= \mathcal{O}\left(i^{s-1+\sum_{\alpha_j>-1} \alpha_j} \right), \end{aligned} \tag{4.21}$$

where the integral is $\mathcal{O}(1)$ since it converges.

DOI:10.6814/NCCU202200447

Finally, by combining the estimates of the two terms:

$$\begin{aligned}\sum_{\ell_0+\cdots+\ell_t=\ell} \ell_0^{\alpha_0}\cdots\ell_t^{\alpha_t} &= \mathcal{O}\left(\sum_{i=1}^{\ell-1} i^{s-1+\sum_{\alpha_j>-1}\alpha_j}(\ell-i)^{\alpha}\right)\\ &= \mathcal{O}\left(\sum_{1\le i\le \ell/2} i^{s-1+\sum_{\alpha_j>-1}\alpha_j}(\ell-i)^{\alpha}\right.\\ &\qquad \left.+\sum_{\ell/2\le i\le \ell-1} i^{s-1+\sum_{\alpha_j>-1}\alpha_j}(\ell-i)^{\alpha}\right).\end{aligned}$$

For the first sum, we have

$$\begin{aligned}\sum_{1\le i\le \ell/2} i^{s-1+\sum_{\alpha_j>-1}\alpha_j}(\ell-i)^{\alpha} &= \mathcal{O}\left(\ell^{\alpha}\sum_{1\le i\le \ell/2} i^{s-1+\sum_{\alpha_j>-1}\alpha_j}\right)\\ &= \mathcal{O}\left(\ell^{\alpha+s+\sum_{\alpha_j>-1}\alpha_j}\int_0^{1/2} x^{s-1+\sum_{\alpha_j>-1}\alpha_j}\mathrm{d}x\right)\\ &= \mathcal{O}\left(\ell^{\alpha+s+\sum_{\alpha_j>-1}\alpha_j}\right).\end{aligned}$$

For the second sum, we have

$$\begin{aligned}\sum_{\ell/2\le i\le \ell-1} i^{s-1+\sum_{\alpha_j>-1}\alpha_j}(\ell-i)^{\alpha} &= \mathcal{O}\left(\ell^{s-1+\sum_{\alpha_j>-1}\alpha_j}\sum_{\ell\ge 1}\ell^{\alpha}\right)\\ &= \mathcal{O}\left(\ell^{s-1+\sum_{\alpha_j>-1}\alpha_j}\right).\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{\ell_0+\cdots+\ell_t=\ell} \ell_0^{\alpha_0}\cdots\ell_t^{\alpha_t} &= \mathcal{O}\left(\ell^{\alpha+s+\sum_{\alpha_j>-1}\alpha_j}\right)+\mathcal{O}\left(\ell^{s-1+\sum_{\alpha_j>-1}\alpha_j}\right)\\ &= \mathcal{O}\left(\ell^{s-1+\sum_{\alpha_j>-1}\alpha_j}\right)\end{aligned}$$

which is the claimed result for $0<s<t$.

DOI:10.6814/NCCU202200447

For the missing cases $s = 0$ and $s = t + 1$, the result is already implied by (4.2) and (4.21), respectively. This concludes the proof. □

We are now ready for the last step.

Theorem 4.13. *For the number* $\mathrm{R}_{n,k}$ *of tree-child networks with* n *leaves and* k *reticulation nodes arising from the non-star component graphs, we have, as* $n \to \infty$*,*

$$\mathrm{R}_{n,k} = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-3/2}\right).$$

Proof. For a non-star component graph with $k + 1$ vertices, we consider its subgraph which contains only the root node and the nodes attached to it by double edges; all other nodes and the edges are removed.

Let t be the number of all non-root nodes in this subgraph. Note that $1 \leq t < k$ since $t = k$ is true only if the original component graph is the star component graph. Now, we bound the number using a construction similar as the one from the previous section. First, we pick a one-component tree-child network with t reticulation nodes whose leaves are replaced by phylogenetic trees with $n_1, \ldots, n_t$ leaves, respectively. Then, we relabel the leaves in an order-consistent way. Finally, we reattach the edges. Note that the number of ways this can be done is bounded by a constant times $n_0^{\delta_0} \cdots n_t^{\delta_t}$, where

$$\delta_0 + \ldots + \delta_t = 2(k - t). \tag{4.22}$$

since $k - t$ edges need to be reattached. Note that this is only an upper bound since some networks may be constructed multiple times.

Overall, we obtain (up to a constant) the following upper bound for the number of tree-child networks arising from the given component graph

$$\tilde{\mathrm{R}}_{n,k} := \sum_{n_0+\cdots+n_t=n} \binom{n}{n_0, \ldots, n_t} \mathrm{O}_{n_0+t,t} T_n \cdots T_{n_t} n_0^{\delta_0} \cdots n_t^{\delta_t}.$$

DOI:10.6814/NCCU202200447

The above sum, by Propositions 4.10 and 4.11 and Stirling's formula, can be bounded as

$$\begin{aligned}\tilde{\mathrm{R}}_{n,k} &= n! \sum_{n_0+\cdots+n_t=n} \frac{\mathrm{O}_{n_0+t,t}}{n_0!} n_0^{\delta_0} \cdot \frac{T_{n_1}}{n_1!} n_1^{\delta_1} \cdots \frac{T_{n_t}}{n_t!} n_t^{\delta_t} \\ &= \mathcal{O}\left(\left(\frac{1}{e}\right)^n n^{n+1/2} \sum_{n_0+\cdots+n_t=n} 2^{n_0} n_0^{2t+\delta_0-3/2} \cdot 2^{n_1} n_1^{\delta_1-3/2} \cdots 2^{n_t} n_t^{\delta_t-3/2} \right) \\ &= \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+1/2} \sum_{n_0+\cdots+n_t=n} n_0^{2t+\delta_0-3/2} \cdot n_1^{\delta_1-3/2} \cdots n_t^{\delta_t-3/2} \right).\end{aligned}$$

Note that $2t+\delta_0-3/2 > -1$ and $n_j^{\delta_j-3/2} > -1$ if and only if $\delta_j > 0$. Set

$$s := \#\{1 \le j \le t \,:\, \delta_j > 0\}$$

which satisfies $s \ge 1$ since $t < k$. Then, by Lemma 4.12,

$$\begin{aligned}\tilde{\mathrm{R}}_{n,k} &= \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+1/2} n^{s+2t+\sum_{j=0}^{t}(\delta_j-3/2)+(t-s)3/2} \right) \\ &= \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-1} n^{-s/2} \right)\end{aligned}$$

which gives the required bound for $\tilde{\mathrm{R}}_{n,k}$ since $s \ge 1$.

Finally, summing over all non-star component graphs with $k+1$ vertices gives the claimed result. □

DOI:10.6814/NCCU202200447

Chapter 5

Conclusion

This thesis provided both enumerative and asymptotic information about galled trees and tree-child networks.

First, we applied the symbolic method to galled trees as well as normal and one-component galled trees. This method turned out to be not only more flexible for finding exact formulas but was also capable of providing asymptotic results. The results we obtained for the different classes of galled trees with n leaves were easier then previous results since the formulas given in [CZ20] require solving an integer partition problem. Moreover, our first order asymptotics results show that most galled trees are not normal.

Second, for tree-child networks, we counted them via component graphs and enhanced the range of applicability of this method to cases with more reticulation nodes than in [CZ20]. Moreover, we showed that this approach can also be used to obtain the first order asymptotics result from [FGM19]. In addition, this new method allowed us to solve an open problem from [FGM19] and [FGM21].

As for other phylogenetic networks, component graphs may not be useful since, for example, the normal condition may be violated during the construction. We wonder whether component graphs can be generalized so that they can be used to deal with other networks as well. We leave this as an open problem.

DOI:10.6814/NCCU202200447

Bibliography

[BGM20] Mathilde Bouvel, Philippe Gambette, and Marefatollah Mansouri. Counting phylogenetic networks of level 1 and 2. *Journal of Mathematical Biology*, 81(6):1357–1395, 2020.

[CZ20] Gabriel Cardona and Louxin Zhang. Counting and enumerating tree-child networks and their subclasses. *Journal of Computer and System Sciences*, 114:84–104, 2020.

[FGM19] Michael Fuchs, Bernhard Gittenberger, and Marefatollah Mansouri. Counting phylogenetic networks with few reticulation vertices: tree-child and normal networks. *Australas. J. Combin.*, 73:385–423, 2019.

[FGM21] Michael Fuchs, Bernhard Gittenberger, and Marefatollah Mansouri. Counting phylogenetic networks with few reticulation vertices: exact enumeration and corrections. *Australas. J. Combin.*, 81:257–282, 2021.

[FHY21] Michael Fuchs, En-Yu Huang, and Guan-Ru Yu. Counting phylogenetic networks with few reticulation vertices: A second approach. *arXiv:2104.07842*, 2021.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

DOI:10.6814/NCCU202200447

[FYZ21] Michael Fuchs, Guan-Ru Yu, and Louxin Zhang. On the asymptotic growth of the number of tree-child networks. *European J. Combin.*, 93:Paper No. 103278, 20, 2021.

[MS00] Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. *Mathematical biosciences*, 164(1):81–92, 2000.

[PB21] Miquel Pons and Josep Batle. On the exact counting of tree-child networks. 2021.

[SF13] Robert Sedgewick and Philippe Flajolet. *An Introduction to The Analysis of Algorithms*. Pearson Education India, 2013.

[Zha19] Louxin Zhang. Generating normal networks via leaf insertion and nearest neighbor interchange. *BMC bioinformatics*, 20(20):1–9, 2019.

DOI:10.6814/NCCU202200447