國立政治大學應用數學系

碩士學位論文

樹子網路及其變體的計數和分布結果

Enumerative and Distributional Results for Tree-Child Networks and Their Variants

指導教授：符麥克　博士

研究生：劉赫煊　撰

中 華 民 國 111 年 6 月

DOI:10.6814/NCCU202201049

致　　謝

回望碩士的幾年，屬實是一段難忘的旅程。

在相遇過的人中，首先特別地，要感謝我的指導老師Michael Fuchs(符麥克)教授。老師足具敏銳又待人尊重，總是肯耐心傾聽我支離破碎的表達並予以鼓勵、幫助梳理思路。他認真負責，卻從不恣意干預他人的學習和生活節奏，給了我很大的自主空間。老師還富有熱枕，有著龐大的知識儲量：不僅是課堂上與研究中，那些在各個國家的旅行見聞、或同或異的思維碰撞，都是我想要珍藏的回憶。很榮幸能夠成為老師的學生。

感謝MF組的各位：冠儒學長、俞讚城博士、裕昇學長、恩宇、鈞齊。謝謝大家在生活、學術及未來規劃上的諸多幫忙和建議。透過大家，我也能有機會更深入地瞭解身邊人們的經歷與思考；感謝應數系的陳天進老師、陳隆奇老師、符聖珍老師、班榮超老師、姜祖恕老師、許順吉老師、洪芷漪老師、張宜武老師。謝謝各位老師的悉心教導，撥冗回答我淺薄的提問；感謝黃顯貴老師及中研院研討會的各位。開拓了我的視野，讓我見識到更廣闊的數學；感謝系辦的各位。謝謝大家的關照。

感謝遠方的家人。數年未見，離別時一語成讖。謝謝你們的關心和支持，雖然甚少直接向你們表露情感，但我都牢記在心，真的辛苦了。

感謝很多很多我想要感謝的人。

.....

此外，還要感謝這座島嶼本身。雖然略有遺憾，最終仍無法前往心往神馳的札幌、珀斯等地，但無論是阿里山的櫻花和黎明、風櫃港的濤聲和黃昏、伴隨太平洋的海風響起的only time還是作為緣起之地的彰化小城，種種過往，都經久不息迴盪在我的記憶。這為期幾年的旅程，也給了我某個長久追尋的疑問以部分解答。

最後，我要感謝來自深淵這部動漫。謝謝其帶來的震撼與靈魂的共鳴。

2022.7.24

DOI:10.6814/NCCU202201049

摘要

近年來，作為演化網路的眾多分類中最著名的子類之一，樹子網路吸引了許多數學家與生物學家的注意。然而直到幾年前，樹子網路的精確和漸進計數仍然很困難，遑論其它問題。在本碩論中，我們將回顧以往樹子網路及其變體的一些重要結果，並添加幾個新的結果。

藉由組合學和概率論中的工具，我們實現的主要貢獻有：在單組分樹子網路下證明了最近的一個對於樹子網絡精確計數的猜想；此外，得到了第一個在均勻隨機選取樹子網路時的隨機結果；同時，還擴展了樹子網路的定義並將前人的和我們的結果推廣到這一新類；另外，也使先前對於有序樹子網路中圖案極限規律的研究更進一步，且提供了首個對一類演化網路中一般圖形的研究。

該碩論的簡短概述如下：首先在第1章中，我們給出了樹子網路、樹子網路的擴展以及有序樹子網路的定義和基本性質；然後在第2章中，我們介紹了所用的工具。其次在第3和第4章，我們分別對樹子網路及其擴展進行研究。接著在第5章，我們將以往對於有序樹子網路的研究推廣到所有高度為1和2的圖案，再給出對於任意高度圖案的推論。最後在第6章，我們總結全文。

關鍵詞：演化網路，樹子網路，解析計數，極限法則，雙射法，拉普拉斯方法，動差估計。

DOI:10.6814/NCCU202201049

Abstract

In recent years, as one of the most prominent subclass among the many different classes of phylogenetic networks, the class of *tree-child networks* has attracted the attention of many mathematicians and biologists. However, until a few years ago, both exact and asymptotic counting for tree-child networks was still difficult, not to mention other problems. In this thesis, we will review the most important previous results for tree-child networks and their variants and add several new results.

Our main contributions, which are mainly proved with tools from Combinatorics and Probability Theory, are as follows. For a recent conjecture on the exact counting of tree-child networks, we give a proof for the special case when the tree-child network is a one-component network. In addition, we prove the first stochastic results for tree-child networks which are picked uniformly at random. Also, we can extend the definition of tree-child networks and generalize previous and our results to the new class. Moreover, we have taken the previous research on limit laws of patterns in ranked tree-child network a step further and provided the first general patterns study for a class of phylogenetic networks.

A short outline of the thesis is as follows: in Chapter 1, we give definitions and show some basic properties for tree-child networks, their extensions and ranked tree-child networks. Then, in Chapter 2, we introduce our tools. In Chapter 3 and Chapter 4, we focus on results for tree-child networks and their extensions, respectively. Next in Chapter 5, we generalize the former study on patterns of ranked tree-child networkto all patterns of height 1 and 2 and make a conjecture for patterns of any height. Finally, we finish the thesis in Chapter 6 with a conclusion.

Keywords: Phylogenetic network, tree-child network, analytic counting, limit laws, bijective proof, Laplace method, method of moments.

DOI:10.6814/NCCU202201049

Contents

DOI:10.6814/NCCU202201049

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202201049

Chapter 1

Introduction

In this chapter, we will describe background of the research topics discussed in this thesis and give precise definitions of the objects we are going to investigate.

1.1 Phylogenetic trees and networks

The purpose of phylogenetic analysis is to reveal the evolution between different species, so as to obtain an understanding of the evolution of life. Two important tools for visualizing the evolution are phylogenetic trees and networks which will be discussed in the two subsections below.

1.1.1 Phylogenetic trees

Phylogenetic trees are widely used in evolutionary biology and are usually calculated from molecular sequences. For example, they are used to understand the age and rate of diversification of taxa, to comprehend the evolutionary history of gene families, to study the co-evolution of hosts and parasites, and to track the origin and transmission of infectious diseases (like dominant species of COVID-19) in epidemiology.

We first give a precise (mathematical) definition of phylogenetic trees.

Definition 1.1.1. *A phylogenetic tree is a rooted binary tree with leaves labeled bijectively by the elements of* $\{1, \ldots, n\}$.

Remark 1.1.2. We will often assume that the root has also an incoming edge.

DOI:10.6814/NCCU202201049

There has been a lot of work focusing on phylogenetic trees over the past few decades; tools from enumerative and analytic combinatorics, graph theory, probability theory and other methods have been used to obtain a wealth of results. Two of them, which are relevant to this thesis, are concerned with the the counting problem and the pattern problem. The former is actually easy and summarized in the next result.

Proposition 1.1.3. *The number* PT_n *of phylogenetic trees equals* $(2n-3)!!$.

Proof. We use proof by induction.

Clearly, $\mathrm{PT}_2 = 1$ since the rooted tree which consists of a root to which two leaves labeled by 1 and 2 are attached is the only phylogenetic tree of size 2.

Next, for PT_{n-1}, there are $n-2$ internal nodes in a rooted binary tree with $n-1$ leaves due to the binary property. In addition, the number of nodes is equal to the number of edges (recall that we assume that the root has also an incoming edge). Thus a rooted binary tree with $n-1$ leaves has a total of $2n-3$ edges. We can create any rooted binary tree with n leaves from one with $n-1$ leaves by picking an edge, splitting it in two with a new node between them and attaching the new leaf (labeled by n) to the newly created node. So by the induction hypothesis, we have

$$\mathrm{PT}_n = (2n-3) \times \mathrm{PT}_{n-1} = (2n-3)!!.$$

This proves the claimed result. ∎

Remark 1.1.4. By Stirling's formula

$$n! = \sqrt{2\pi n}\left(\frac{e}{n}\right)^n \left(1+\mathcal{O}\left(\frac{1}{n}\right)\right),$$

we have the asymptotic result:

$$\begin{aligned}\mathrm{PT}_n = (2n-3)!! &= \frac{(2n-2)!}{2^{n-1}(n-1)!} \\ &\sim \frac{1}{\sqrt{2}}\left(\frac{2}{e}\right)^n n^{n-1}, \quad \text{as} \quad n \to \infty.\end{aligned}$$

As for the pattern problem, there have been many studies which have investigated the number of occurrence of patterns on the fringe of phylogenetic trees under various random models; see [6, 8–10, 18, 19, 21, 23, 24] for some results. In these studies mean, variance, limit laws and even more detailed stochastic results have been obtained.

DOI:10.6814/NCCU202201049

1.1.2 Phylogenetic networks

Phylogenetic trees are not always suitable as a model since the evolutionary process of, e.g., chromosomes, species and populations is not necessarily tree-like. This is because of meiotic recombination on the chromosome level, specification and horizontal gene transfer on the species level, and sexual recombination on the population level which all lead to reticulation events. Due to this, phylogenetic networks have been proposed as an alternative to phylogenetic trees. In fact, they have become a standard tool in evolutionary biology over the last decades.

We start by defining a phylogenetic network.

Definition 1.1.5. *A phylogenetic network is a rooted simple DAG (directed acyclic graph) with a unique root of indegree 0 and outdegree 1 such that all other nodes belong to one of the following types:*

(i) leaves, which are nodes with indegree 1 and outdegree 0; these are bijectively labeled by the elements of $\{1, \ldots, n\}$*;*

(ii) tree nodes, which are nodes with indegree 1 and outdegree 2; and

(iii) reticulation nodes, which are nodes with indegree 2 and outdegree 1.

From the definition, we can observe that phylogenetic trees are a special case of phylogenetic networks; they are phylogenetic networks without reticulation nodes.

Moreover, for any phylogenetic network, by counting incoming and outgoing edges, we obtain the following result.

Proposition 1.1.6. *For a phylogenetic network with* n *leaves,* k *reticulation nodes, and* t *tree nodes, we have,*

$$n + k = t + 1.$$

Proof. Note that the sum of the outgoing edges equals $2t + k + 1$. This sum is equal to the sum of the incoming edges which is $t + 2k + n$. Equating these two sums and a simple computation gives the claimed result. ∎

Clearly, the number of phylogenetic networks with n leaves is infinite. Thus, in order to get a meaningful counting problem, we have to consider one of the many subclasses of

DOI:10.6814/NCCU202201049

phylogenetic networks which have been proposed in phylogenetics. One of them is the class of tree-child networks which will be investigated in this thesis.

1.2 Tree-child networks

In recent years, the class of tree-child networks has become one of the most prominent subclass among the many different classes of phylogenetic networks. The definition of this class uses one fact which is often observed in evolution: newly born dominant species are usually not quickly wiped out, in other words, reticulation events seldomly occur in class vicinity of each other. Because of this, tree-child network are more important in real-world applications than other classes of phylogenetic trees.

In this master thesis, we will mainly focus on counting, bijections and the distributional behavior of tree-child networks. We will start with the definition.

Definition 1.2.1. *A phylogenetic network is a tree-child network if each non-leaf node has at least one child which is not a reticulation node.*

Thus, in a tree-child network a tree node has least one child which is not a reticulation node.

In addition, a reticulation node should not be followed by another reticulation node. This is because reticulation nodes have only one outgoing edge, in other words, they only have one child. And by definition this child cannot also be a reticulation node.

Finally, the child of the root cannot be a reticulation node, too.

We will denote by $\mathcal{TC}_{n,k}$ the set of tree-child networks with n leaves and k reticulation nodes and by $\mathrm{TC}_{n,k}$ the cardinality of this set. In addition, we use t to denote the number of tree nodes in a network from $\mathcal{TC}_{n,k}$.

Remark 1.2.2. When $k = 0$, the set $\mathcal{TC}_{n,0}$ is exactly the set of *phylogenetic trees* with n leaves. Thus, $\mathrm{TC}_{n,0} = \mathrm{PT}_n$, where PT_n was defined in Proposition 1.1.3.

Next, we will introduce *free tree nodes* which will play an important role in our subsequent study of tree-child networks.

Definition 1.2.3. *We call a tree node a free tree node if both its children are not reticulation nodes. Moreover, we call an outgoing edge of a free tree node a free edge.*

DOI:10.6814/NCCU202201049

Note that every node in a phylogenetic tree is a free tree node. More generally, also in a tree-child network, the number of free tree nodes can be determined.

Lemma 1.2.4. *Every tree-child network with n leaves and k reticulation nodes has $n-k-1$ free tree nodes.*

Proof. From Proposition 1.1.6, we see that each phylogenetic network and thus each tree-child network has $n+k-1$ tree nodes. Any tree node which has one child which is a reticulation node is not free and each reticulation node has two incoming edges. Thus, a tree-child network has exactly $2k$ tree nodes which are not free. (Here, we used the tree-child property.) So the total number of free tree nodes is $n+k-1-2k = n-k-1$ as claimed. ∎

The previous lemma immediately gives the following consequence.

Corollary 1.2.5. *In every tree-child network with n leaves and k reticulation nodes, we have that $k \leq n-1$.*

Thus, we have for the number of tree-child networks with n leaves, which will be denoted by TC_n:

$$\mathrm{TC}_n = \sum_{k=0}^{n-1} \mathrm{TC}_{n,k}.$$

In addition the set of tree-child networks with n leaves will be denoted by $\mathcal{TC}_n$; thus, $\mathrm{TC}_n = |\mathcal{TC}_n|$.

We will next introduce a subclass, an extension and a variant of the class of tree-child networks.

One-component Tree-child networks. The class of tree-child networks is rather complex. Because of this, further restrictions on tree-child networks have been imposed to obtain subclasses of networks whose investigation is easier.

One important subclass of tree-child network is the class of one-component tree-child network. It is defined as follows.

Definition 1.2.6. *A tree-child network is called a one-component tree-child network if every reticulation node is directly followed by a leaf.*

DOI:10.6814/NCCU202201049

One-component tree-child networks are more "tree-like" than general tree-child networks. In fact, as we will see below, we know much more about the one-component case than the general case. We believe that studying one component tree-child networks will be helpful for the general case because one-component networks constitute an important building block in the construction of general networks; see [5].

We denote by $\mathcal{OTC}_{n,k}$ the set of one-component tree-child networks with n leaves and k reticulation nodes and by $\mathrm{OTC}_{n,k}$ the cardinality of this set. Moreover, we denote by $\mathcal{OTC}_n$ resp. OTC_n the number resp. set of one-component tree-child networks with n leaves. Note that by Corollary 1.2.5:

$$\mathrm{OTC}_n = \sum_{k=0}^{n-1} \mathrm{OTC}_{n,k}.$$

d-combining tree-child network. Note that in a tree-child network, all reticulation nodes have two incoming edges. Thus, a natural generalization is to change the "two" into "d", where $d \geq 2$ is a fixed integer; such networks are called *d-combining tree-child network*. (The definition of tree nodes remains the same.) In particular, we call the case with $d = 2$ the bi-combining case. (Subsequently, this is the case we mean if no d is indicated.)

We use $\mathcal{TC}_{n,k}^{[d]}$ to denote the set of d-combining tree-child networks with n leaves and k reticulation nodes and $\mathrm{TC}_{n,k}^{[d]}$ to denote the cardinality of this set. Also the notion of one-component networks can be extended to d-combining tree-child networks, and we use $\mathcal{OTC}_{n,k}^{[d]}$ resp. $\mathrm{OTC}_{n,k}^{[d]}$ to denote the set resp. its cardinality of d-combining one-component tree-child networks with n leaves and k reticulation nodes.

This extension will give some interesting (and unexpected) results. For example, for the case of d-combining one-component tree-child network, the first order asymptotics of $\mathrm{OTC}_n^{[d]}$ (the number of d-combining one-component tree-child networks with n leaves) are different when $d = 2$, $d = 3$ and $d \geq 4$. We will prove these (and other) results in the chapters below.

Proposition 1.1.6, Lemma 1.2.4 and Corollary 1.2.5 can all be easily generalized to d-combining tree-child networks. We summarize the results in the proposition below.

Proposition 1.2.7. *For a d-combining tree-child network with n leaves, k reticulation nodes and t tree nodes, we have*

DOI:10.6814/NCCU202201049

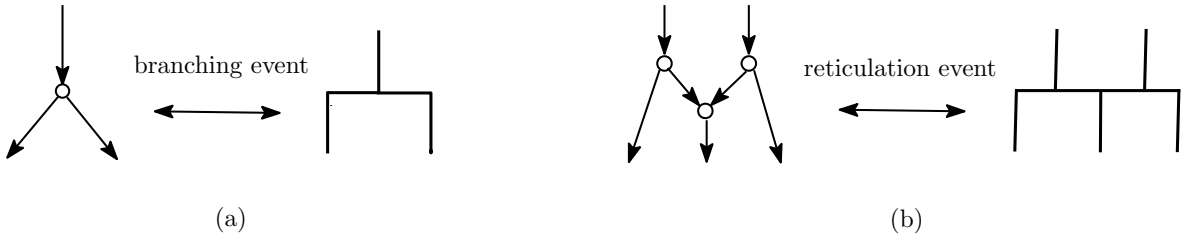


Figure 1.1: The two kinds of events used in the construction of RTCNs.

(i) $n + (d-1)k = t + 1$;

(ii) the number of free tree nodes is $n - k - 1$;

(iii) $k \leq n - 1$.

Thus, the number of d-combining tree-child networks with n leaves, denoted by $\mathrm{TC}_n^{[d]}$, and the number of d-combining one-component tree-child networks with n leaves satisfy:

$$\mathrm{TC}_n^{[d]} = \sum_{k=0}^{n-1} \mathrm{TC}_{n,k}^{[d]}, \qquad \mathrm{OTC}_n^{[d]} = \sum_{k=0}^{n-1} \mathrm{OTC}_{n,k}^{[d]}.$$

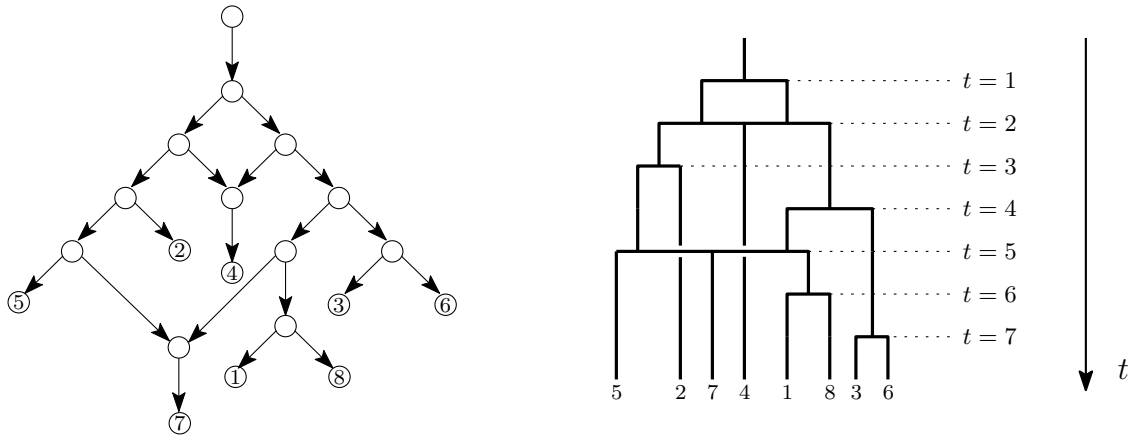
In addition, we will use the notations $\mathcal{TC}_n^{[d]}$ and $\mathcal{OTC}_n^{[d]}$ for the corresponding sets.

Ranked tree child network (RTCN). As mentioned above, in the past few decades, many studies on statistical properties of the shape of phylogenetic tree have been performed (e.g., pattern studies; see the references above). On the other hand, until recently, no corresponding studies have been done for phylogenetic networks. This is due to a fact that even counting problems for almost all subclasses of phylogenetic network (including tree-child networks) were still open until recently.

In order to overcome this, the idea of ranking phylogenetic networks was proposed in [1] which made the investigation of the above mentioned problems feasible. Ranking for tree-child networks yields *ranked tree-child networks* or RTCNs for short.

In order to define them, first in a tree-child network, we call a tree node a *branching event* and a reticulation node with its two parents a *reticulation event*; see Figure 1.1 for the graphical depiction of these two events. (Vertical edges in this depiction will subsequently be called *lineages*.) Then, a RTCN is defined as follows.

DOI:10.6814/NCCU202201049



(a) (b)

Figure 1.2: (a) A tree-child network which is rankable; (b) A RTCN corresponding to (a) with a (fixed) ranking.

Definition 1.2.8. *A ranked tree-child network is a tree-child network which is drawn starting with a branching event and consecutively adding either a branching event or a reticulation event until all events are used; see Figure 1.2 an example.*

Compared to tree-child networks, RTCNs have two advantages. First, they are easily counted and their stochastic properties can be understood. Second, they are equipped with a growth model since RTCNs start from the root and the network is constructed step by step. On the other hand, it is not clear how this has happened in other phylogenetic networks.

We will obtain some results of patterns of RTCN in Chapter 5 below.

1.3 Previous research and purpose of this work

Previous research. Tree-child networks were introduced by Cardona, Rossello and Valiente (2009); see [4]. Apart from giving the definition, the authors of this paper defined a distance that extends the well-known Robinson-Foulds distance for phylogenetic trees. Moreover, they provided algorithms for reconstructing a tree-child network and computing the distance between two tree-child networks.

The first enumerative study of tree-child networks was done a few years later by McDiarmid, Semple and Welsh (2015) who showed that the largest term in the main growth term of TC_n is n^{2n}; they obtained this result by using lower and upper bounds, see [20]. In addition,

DOI:10.6814/NCCU202201049

they also proved a similar result for vertex-labeled tree-child networks. Moreover, they found that almost all tree-child networks with n leaves have $(1+o(1))n$ reticulation nodes, a preliminary step towards understanding the distribution of $\mathrm{TC}_{n,k}$. They ended their paper with a long list of open problems.

In recent years, many follow-up studies on enumerative properties of tree-child networks (and other classes of phylogenetic networks) have appeared. We give a (brief) summary of the studies which are relevant to this thesis.

First, Fuchs, Gittenberger and Mansouri (2019 and 2021) and Fuchs, Huang and Yu (2022) solved the asymptotic counting problem for tree-child networks with a fixed number of reticulation nodes. More precisely, two different methods, namely sparsened skeletons (in [12] and [13]) and component graphs (in [14]) were used to show that the number $\mathrm{TC}_{n,k}$ has the first-order asymptotics:

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

On the other hand, Cardona and Zhang (2020) devised an algorithmic approach (based on component graphs) for obtaining values of $\mathrm{TC}_{n,k}$ for all small n and k; see [5] for a table which displays the values of $\mathrm{TC}_{n,k}$ for $1 \leq k < n$ and $3 \leq n \leq 8$. In addition, Cardona and Zhang also used their approach to derive exact formulas for $\mathrm{TC}_{n,k}$ for $k = 1, 2$; see also [13] where these formulas and a formula for $k = 3$ were obtained with a different approach.

Thus, exact and asymptotic counting for tree-child networks with few reticulation nodes is well understood.

For general tree-child networks, the most important result so far was obtained by Fuchs, Yu and Zhang (2021) who almost solved the asymptotic counting problem for TC_n; see [17]. They showed that TC_n has the same growth order as $\mathrm{TC}_{n,n-1}$ and they used a bijective proof to obtain (up to a constant) the main growth term of $\mathrm{TC}_{n,n-1}$. More precisely, they gave the following formula:

$$\mathrm{TC}_n = \Theta\left(n^{-2/3}e^{\alpha_1(3n)^{1/3}}\left(\frac{12}{e}\right)^n n^{2n}\right),$$

where $\alpha_1 = -2.338107410\ldots$ is the largest root of the Airy function of the first kind. In other words, they significantly improved the n^{2n} result of McDiarmid, Semple and Welsh (2015).

Finally, there was also some recent progress on the exact counting of tree-child networks by Pons and Batle (2021) who made a conjecture on $\mathrm{TC}_{n,k}$ in [22]. This conjecture yields a

DOI:10.6814/NCCU202201049

recurrence formula for $\mathrm{TC}_{n,k}$ which can be used to quickly re-derive the values in the above mentioned table from [5]. Moreover, also the above mentioned exact formulas for small values of k can be derived from their conjecture in an elegant way.

Apart from the above progress on tree-child networks, one-component tree-child networks were studied as well. Cardona and Zhang proposed them in [5], where they also proved a simple exact formula for $\mathrm{OTC}_{n,k}$. This formula was then used in [17] to prove a local limit theorem for the number of reticulation nodes of one-component tree-child networks picked uniformly at random and to derive the first-order asymptotics for OTC_n (which is a consequence of the local limit theorem).

Finally, as mentioned above, RTCN where proposed by Bienvenu, Lambert and Steel in [1]. They used tools from combinatorics and probability theory to prove a lot of results for them, e.g., they obtained exact counting results and distributional results for patterns and other parameters.

Purpose of this work. In this thesis, we will review (in detail) some of the above results and contribute several results on our own. The main contributions of the thesis are:

- a proof of the conjecture of Pons and Batle for the subclass of one-component tree-child networks;
- announcing the first stochastic results for random tree-child networks; the proof will be contained in the journal version of the extended abstract [7];
- extending previous and our new results to d-combining networks;
- continuing the pattern study from [1] for random RTCNs; this will constitute the first such study for a class of phylogenetic networks.

Outline of this thesis. We give a short outline of the structure of this thesis.

In Chapter 2, we introduce tools from two different areas in mathematics, namely, from Combinatorics and from Probability Theory. In the chapters after that, we will extensively use these tools.

In Chapter 3, we recall the exact counting formula for $\mathrm{OTC}_{n,k}$ from [5] (and provide a new proof) and review some details of the proof of the asymptotic counting formula for TC_n.

DOI:10.6814/NCCU202201049

We also introduce the conjecture for the exact counting of $\mathrm{TC}_{n,k}$ from [22] and prove it for one-component tree-child networks.

In Chapter 4, we generalize some of the $d = 2$ results from Chapter 3 to $d \geq 3$. Apart from this, we also give distributional results of the number of reticulation nodes obtained by picking uniformly at random one-component d-combining tree-child networks and d-combining tree-child networks with n leaves.

In Chapter 5, we extend the study of patterns of height 1 from [1] to height 2, giving for each pattern the expectation and limiting distribution. In addition, we have a conjecture for arbitrary patterns.

Finally, we finish the thesis in Chapter 6 with a conclusion.

DOI:10.6814/NCCU202201049

Chapter 2

Tools

2.1 Tools from Combinatorics

In this section, we will introduce two methods from Combinatorics: (i) bijective proofs and (ii) the Laplace method. Both these methods are used in the proof of the results in Chapter 3 and Chapter 4.

First, we will give some examples of bijective proofs, where we use Catalan numbers as guiding example. Recall that a phylogenetic tree is a rooted binary tree whose leaves are labeled bijectively by the elements of $\{1, \ldots, n\}$. A related class of trees are rooted binary ordered trees with n internal nodes. (Here, ordered means that the two children of every internal node have a left-right order.) These trees are also easily enumerated.

Proposition 2.1.1. *The number C_n of rooted binary ordered trees with n internal nodes is* $\frac{1}{n+1}\binom{2n}{n}$

Proof. Clearly, $C_0 = 1$ since the tree without internal nodes just consists of a single leaf.

Next, for $n \geq 0$, the root in a rooted binary ordered tree RT with $n+1$ internal nodes has children which root trees RT_1 and RT_2. The root is an internal node, thus the total number of internal nodes in RT_1 and RT_2 is n. In other words, if RT_1 has i internal nodes, then RT_2 has $n-i$ internal nodes. Thus, we have the following recurrence for C_n:

$$C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}, \qquad (n \geq 0) \tag{2.1}$$

with initial condition $C_0 = 1$.

DOI:10.6814/NCCU202201049

We use an ordinary generating function to solve the recurrence. Let

$$f(x) := \sum_{n\geq 0} C_n x^n.$$

Since $C_0 = 1$, $f(x)$ satisfies

$$f(x) = 1 + \sum_{n\geq 1}\sum_{i=0}^{n-1} C_i C_{n-1-i} x^n = 1 + x\sum_{n\geq 0}\sum_{i=0}^{n} C_i C_{n-i} x^n = 1 + x(f(x))^2.$$

Solving this yields

$$f(x) = \frac{1 \pm \sqrt{1-4x}}{2x}.$$

Since $C_0 = 1$, by L'Hôspital's rule as $x \to 0$, we must choose

$$f(x) = \frac{1 - \sqrt{1-4x}}{2x}.$$

Next, by the binomial theorem,

$$\frac{1-\sqrt{1-4x}}{2x} = \sum_{n=1}^{\infty} -\frac{1}{2}\binom{1/2}{n}(-4)^n x^{n-1} = \sum_{n=0}^{\infty} -\frac{1}{2}\binom{1/2}{n+1}(-4)^{n+1}x^n.$$

Thus,

$$C_n = -\frac{1}{2}\binom{1/2}{n+1}(-4)^{n+1} = \frac{1}{n+1}\binom{2n}{n},$$

where we used straightforward manipulations for binomial coefficients in the last step. ∎

Remark 2.1.2. The sequence

$$C_n = \frac{1}{n+1}\binom{2n}{n}$$

is called the sequence of Catalan numbers; it is an important sequence in Combinatorics due to it being the answer to many different counting problems. For example,

(a) The number of rooted ordered trees with n vertices and a root with degree 1.

(b) The number of ways one can decompose a convex n-gon into triangles by $n-3$ noninteresting diagonals.

(c) Consider walks in the x-y plane with steps $U : (x, y) \to (x+1, y+1)$ or $D : (x, y) \to (x+1, y-1)$. The number of walks which start at $(0, 0)$ and reach $(2n, 0)$ without crossing the x-axis.

Next we will introduce another set whose cardinality is the Catalan number.

DOI:10.6814/NCCU202201049

Definition 2.1.3. *Suppose we have a nonassociative product operation, i.e., an operation for which the order of computation is important. (E.g,* $((x_1x_2)x_3)$ *is different from* $(x_1(x_2x_3))$*.) Then, let* $\mathcal{S}_n$ *be the number of ways to compute* $x_1 \cdots x_n$*.*

Proposition 2.1.4. *We have,*

$$|\mathcal{S}_n| = C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}.$$

Proof. Each product contains within the outer brackets two expressions, where the first is a product of i factors and the second is a product of $n-i$ factors. Thus the number $u_n = |\mathcal{S}_n|$ satisfies the recurrence :

$$u_n = \sum_{i=1}^{n-1} u_i u_{n-i}, \qquad (n \geq 2).$$

Note that (2.1) can be rewritten into

$$C_{n-1} = \sum_{i=0}^{n-2} C_i C_{n-i-2} = \sum_{i=1}^{n-1} C_{i-1} C_{n-i-1}, \qquad (n \geq 2).$$

Comparing with the above recurrence (and comparing initial values), we obtain that

$$u_n = C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}$$

which is the claimed result. ∎

Because we have that the set $\mathcal{S}_n$ has the same cardinality as the set of rooted binary ordered trees with $n-1$ internal nodes, we wonder whether there is a bijection between these two sets. Finding such a bijection is called a bijective proof in Combinatorics.

Theorem 2.1.5. *There is a bijection between the set of rooted binary ordered trees with* n *internal nodes and the set* $\mathcal{S}_{n+1}$*.*

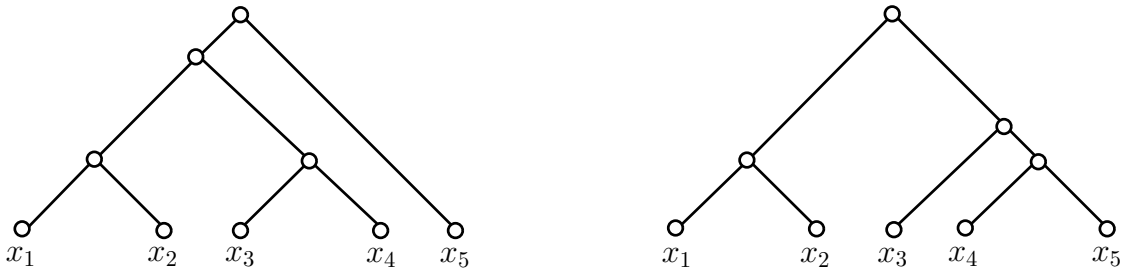
Proof. Label the leaves in a rooted binary ordered tree with n internal nodes from left to right by $x_1, x_2, \ldots, x_{n+1}$. Every bracket is represented by an internal node and the two subtrees of the node are the two components in the bracket; see Figure 2.1 for examples. ∎

In addition, we also have the following relation between Catalan numbers and the counting sequence PT_n of phylogenetic trees; see Proposition 1.1.3.

Theorem 2.1.6. *The number of phylogenetic tree with* n *leaves satisfies*

$$\mathrm{PT}_n = \frac{n! C_{n-1}}{2^{n-1}}$$

DOI:10.6814/NCCU202201049



$(((x_1x_2)(x_3x_4))x_5)$ $((x_1x_2)(x_3(x_4x_5)))$

Figure 2.1: Two rooted binary plane trees with 4 internal nodes and their corresponding images from $\mathcal{S}_5$ under the one-to-one map from Theorem 2.1.5.

Proof. Even though this can be proved by a simple computation, we give a bijective proof.

Therefore, rewrite the above equation into:

$$n!C_{n-1} = 2^{n-1}\mathrm{PT}_n.$$

Now, the left-hand side of this equation is the number of rooted binary trees together with a permutation (which gives a labeling of the n leaves).

On the other hand, the right-hand side is the number of phylogenetic trees with n leaves, where for each internal node, a left-to-right order of the two subtrees of the node is fixed. (Recall that a binary tree with n leaves has $n-1$ internal nodes.)

Clearly, the above two objects are identical and we have found a bijection. ∎

Now, we will introduce another related class of trees which are called unary-binary trees. A unary-binary tree is a rooted ordered tree with all internal nodes having 1 or 2 children; such trees are counted by Motzkin numbers.

Before giving a proof of this, we first recall the Lagrange Inversion Formula, which is an important tool in Combinatorics (and will also be used in the next proof).

Theorem 2.1.7 (Lagrange Inversion Formula)**.** *Let $f(\omega)$ be an analytic function at 0 with $f(0) \neq 0$ and $z = \frac{\omega}{f(\omega)}$. Then $z = \sum_{n\leq 0} a_n\omega^n$ is analytic at 0 with*

$$a_n = \frac{1}{n}[\omega^{n-1}](f(\omega))^n,$$

where $[\omega^n]$ is an operator which extracts the n-th coefficient in the Maclaurin series of a function in ω.

DOI:10.6814/NCCU202201049

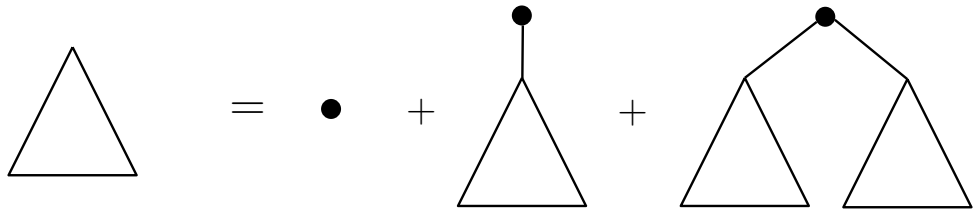


Figure 2.2: A unary-binary tree can be classified into 3 subclasses: a single root; a root with outdegree 1 followed by a unary-binary tree; a root with outdegree 2 followed by two unary-binary trees.

Proposition 2.1.8. *The number M_n of unary-binary trees with n nodes is*

$$\frac{1}{n}\sum_{k=1}^{\lfloor (n+1)/2 \rfloor}\binom{n}{k}\binom{n-k}{k-1}.$$

Proof. In contrast to Proposition 2.1.1, here we deduce the generating function of unary-binary tree directly from a composition of these trees.

We use $M(z)$ to denote the ordinary generating function of unary-binary trees, that is, part (a) in Figure 2.2. Clearly, each such tree can be decomposed into a tree just consisting of a leaf (part (b) with generating function z) or a root which is either followed by one unary-binary tree (part (c) with generating function $zM(z)$) or two unary-binary trees (part (d) with generating function $z(M(z))^2$).

Thus, $M(z)$ satisfies

$$M(z) = z + zM(z) + z(M(z))^2 \tag{2.2}$$

with $M(0) = 1$. Solving this, we get

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

and we have to use the minus sign due to $M(0) = 1$.

Note that from this formula for $M(z)$, it is not easy to obtain coefficients. However, we can use the Lagrange Inversion Formula.

So, we write (2.2) as

$$z = \frac{M(z)}{1 + M(z) + (M(z))^2}.$$

DOI:10.6814/NCCU202201049

Let $\omega = M(z)$, then $f(\omega) = 1 + \omega + \omega^2$. Then, by Theorem 2.1.7:

$$\begin{aligned}
[z]^n M(z) &= \frac{1}{n}[\omega^{n-1}](1+\omega+\omega^2)^n \\
&= \frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}[\omega^{n-1}]\omega^{n-k}(1+\omega)^{n-k} \\
&= \frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}[\omega^{k-1}](1+\omega)^{n-k} \\
&= \frac{1}{n}\sum_{k=1}^{\lfloor (n+1)/2\rfloor}\binom{n}{k}\binom{n-k}{k-1}.
\end{aligned}$$

Thus, we have,

$$M_n = \frac{1}{n}\sum_{k=1}^{\lfloor (n+1)/2\rfloor}\binom{n}{k}\binom{n-k}{k-1}$$

as claimed. ∎

Laplace method. We now have an exact counting formula for M_n. However, this formula contains a summation and we cannot see how fast the sequence M_n grows as $n \to \infty$. Note that the terms and summation indices of this sum depend on n. For such sums, we have a method to give an asymptotic results which is called Laplace method.

In general, the Laplace method is applied to sums of the forms

$$\sum_k a_k(n).$$

The asymptotics is then derived via the following steps (see Figure 2.3):

(i) Find the part of the sum which has the largest contribution to the asymptotics;

(ii) Show that the tails are neligible;

(iii) Approximate $a_k(n)$ in the major range;

(iv) Add back the tails;

(v) Approximate the sum by an integral using the Euler-Maclaurin summation formula.

Now, we apply this method to M_n in the above result. First, writing binomial coefficients as factorials, we obtain

$$M_n = (n-1)!\sum_{k=1}^{\lfloor (n+1)/2\rfloor}\frac{1}{k!(k-1)!(n-2k+1)!}.$$

DOI:10.6814/NCCU202201049

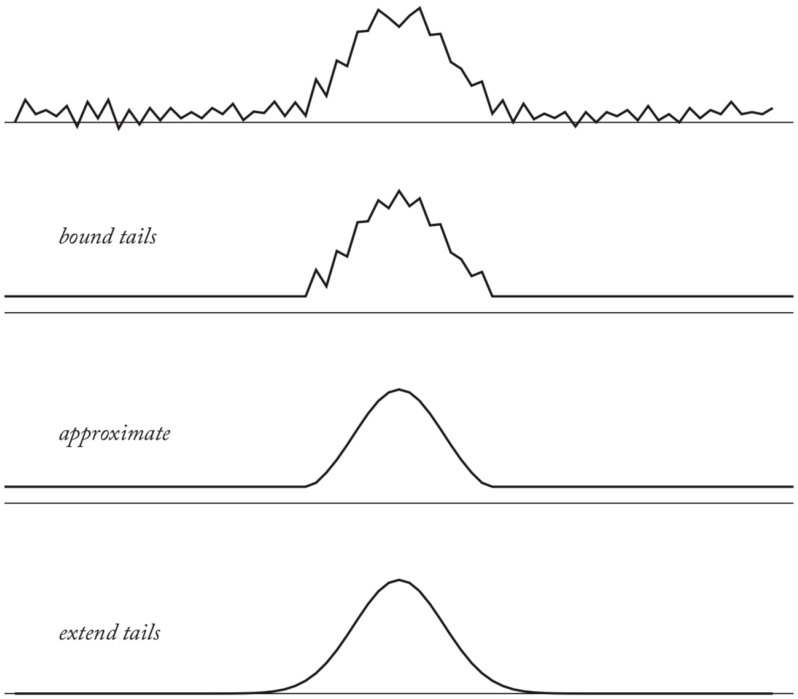


Figure 2.3: The major steps in the Laplace method.

Stirling's formula gives

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

Thus, we only need the asymptotics of

$$S(n) = \sum_{k=1}^{\lfloor (n+1)/2 \rfloor} \frac{1}{k!(k-1)!(n-2k+1)!} = \sum_{k=1}^{\lfloor (n+1)/2 \rfloor} a_k(n).$$

In order to understand where the main contribution comes from, first observe that $1/(k!(k-1)!(n-2k+1)!)$ is increasing for $1 \le k \le \frac{n}{3}$ and decreasing for $\frac{n}{3} < k \le \lfloor (n+1)/2 \rfloor$. Next, by straightforward expansion,

$$a_k(n) = \left(\frac{3}{2\pi n}\right)^{\frac{3}{2}} \left(\frac{3e}{n}\right)^n e^{-9x^2/n} \left(1 + \mathcal{O}\left(\frac{1+|x|}{n} + \frac{|x|^3}{n^2}\right)\right)$$

uniformly for $x = o(n^{2/3})$ where $k = \frac{n}{3} + x$. Thus, we have

$$a_k(n) = \begin{cases} \left(\frac{3}{2\pi n}\right)^{3/2} \left(\frac{3e}{n}\right)^n e^{-9x^2/n} \left(1 + \mathcal{O}\left(\frac{1}{n^{1/2-3\epsilon}}\right)\right), & \text{if } |x| < n^{1/2+\epsilon}; \\ \mathcal{O}\left((3e)^n n^{-n-3/2} e^{-9n^{2\epsilon}}\right), & \text{if } |x| \ge n^{1/2+\epsilon}, \end{cases} \tag{2.3}$$

where $\epsilon > 0$ is an arbitrary small constant.

DOI:10.6814/NCCU202201049

Next, we split the sum into two parts (Step (i) in the above procedure):

$$S(n) = \sum_{|x|<n^{1/2+\epsilon}} a_k(n) + \sum_{|x|\geq n^{1/2+\epsilon}} a_k(n).$$

The second part (which are the tails in Step (ii) above) is bounded by:

$$\sum_{|x|\geq n^{1/2+\epsilon}} a_k(n) = \mathcal{O}\left((3e)^n n^{-n-1/2} e^{-9n^{2\epsilon}}\right),$$

where we used the monotonicity of $a_k(n)$ and (2.3).

For the first sum, we again use (2.3) which gives

$$\begin{aligned}\sum_{|x|<n^{1/2+\epsilon}} a_k(n) &\sim \left(\frac{3}{2\pi n}\right)^{3/2} \left(\frac{3e}{n}\right)^n \sum_{|x|<n^{1/2+\epsilon}} e^{-9x^2/n} \\ &\sim \left(\frac{3}{2\pi n}\right)^{3/2} \left(\frac{3e}{n}\right)^n \sum_{x=-\infty}^{\infty} e^{-9x^2/n},\end{aligned}$$

where we have added back the tails in the second line above (these are Steps (iii) and (iv) in the outline above); note that adding back the tails introduces just another exponential small error.

Next, by the Euler-Maclaurin summation formula (Step (v) above),

$$\sum_{x=-\infty}^{\infty} e^{-9x^2/n} = \int_{-\infty}^{\infty} e^{-9x^2/n} dx + \mathcal{O}\left(\frac{1}{n}\right) = \frac{\sqrt{\pi n}}{3} + \mathcal{O}\left(\frac{1}{n}\right).$$

Plugging this into the above first sum and combining with the bound for the second sum gives:

$$S(n) \sim \frac{\sqrt{3}}{2^{3/2}\pi n} \left(\frac{3e}{n}\right)^n.$$

Now, we can give the asymptotic result for M_n:

$$M_n \sim (n-1)! \frac{\sqrt{3}}{2^{3/2}\pi n} \left(\frac{3e}{n}\right)^n \sim \sqrt{\frac{3}{4\pi n^3}} \times 3^n. \tag{2.4}$$

Thus, M_n grows (up to subexponential terms) like 3^n.

2.2 Tools from Probability Theory

In this section, we will explain some tools from probability theory which will be used in Chapter 5. For the proofs, we refer to [2].

We start with the method of moments which was introduced by Pafnuty Chebyshev in 1887 who used it to prove the central limit theorem.

First, we need a uniqueness theorem for moments.

DOI:10.6814/NCCU202201049

Theorem 2.2.1. *Let X be a random variable having finite moments $\alpha_k = \mathbb{E}(X^k)$ of all orders. If the exponential generating function of α_k (the power series $\sum_{k\geq 0} \alpha_k z^k/k!$) has a positive radius of convergence, then X is uniquely determined by its moments.*

We give two examples which will be used in Chapter 5.

- The moments of the standard normal distribution are given by:

$$\alpha_k = \mathbb{E}(N(0,1))^k = \begin{cases} \dfrac{k!}{2^{k/2}(k/2)!}, & \text{if } k \text{ is even;} \\ 0, & \text{if } k \text{ is odd.} \end{cases} \tag{2.5}$$

  The exponential generating function of this sequences is

$$\sum_{k\geq 0} \alpha_k \frac{z^k}{k!} = \sum_{k\geq 0} \frac{z^{2k}}{2^k k!} = e^{z^2/2}.$$

  Clearly, this series has radius of convergence equal to ∞. Thus, Theorem 2.2.1 implies that the standard normal distribution is uniquely determined by its moments.

- As a second example, consider a random variable X which has a Poisson distribution with parameter λ. Then, the factorial moments of X are given by

$$\mathbb{E}(X^{\underline{k}}) = \lambda^k$$

  and thus

$$\alpha_k = \mathbb{E}(X^k) = \sum_{j=0}^{k} S(k,j)\lambda^j,$$

  where $S(k,j)$ denotes the Stirling numbers of the second kind. The exponential generating function of this sequences is

$$\sum_{k\geq 0} \alpha_k \frac{z^k}{k!} = \sum_{k=0}^{\infty}\sum_{j=0}^{k} S(k,j)\lambda^j \frac{z^k}{k!} = e^{z(e^\lambda - 1)}.$$

  Thus, the exponential generating function of α_k has again radius of convergence equal to ∞ and by Theorem 2.2.1, we obtain that also the Poisson distribution is uniquely determined by its moments.

The method of moments is now encapsulated in the following theorem.

DOI:10.6814/NCCU202201049

Theorem 2.2.2. *Assume that X is uniquely determined by its moments and let X_n be a sequence of random variables whose moments of all orders exist and satisfy:*

$$\lim_{n\to\infty} \mathbb{E}(X_n^k) = \mathbb{E}(X^k) \qquad \textit{for } k = 1, 2, \ldots.$$

Then, X_n converges in distribution to X, i.e., $X_n \xrightarrow{w} X$.

Thus, if one for instance has that

$$\lim_{n\to\infty} \mathbb{E}(X_n^k) = \begin{cases} \dfrac{k!}{2^{k/2}(k/2)!}, & \text{if } k \text{ is even;} \\ 0, & \text{if } k \text{ is odd,} \end{cases}$$

then X_n satisfies a central limit theorem, i.e., $X_n \xrightarrow{w} N(0,1)$.

Also, Theorem 2.2.2 still holds with moments replaced by factorial moments. Thus, if one has that

$$\lim_{n\to\infty} \mathbb{E}(X_n^{\underline{k}}) = \lambda^k,$$

then X_n convergences in distribution to a Poisson random variable with parameter λ.

In fact, the above two theorems can be generalized to multivariate random variables.

Theorem 2.2.3. *Let $X = (X_1, \ldots, X_d)$ be a random variable for which $\alpha_{i,k} = \mathbb{E}(|X_i|^k) < \infty$ for $i = 1, \ldots, d$ and $k = 1, 2, \ldots$. Consider the mixed moments:*

$$\beta(k_1, \ldots, k_d) = \mathbb{E}(X_1^{k_1} \cdots X_d^{k_d})$$

for nonegative integers k_i. If for each i, the exponential generating function of $\alpha_{i,k}$ has a positive radius of convergence, then X is uniquely determined by its mixed moments.

We discuss as example the multivariate normal distribution. Thus, assume that $X = (X_1, \ldots, X_d)$ has a multivariate normal distribution with mean 0 and covariance matrix Σ. Recall that this implies that $X_i \sim N(0, {\sigma_i}^2)$ with some $\sigma_i > 0$. First, we compute $\alpha_{i,k} = \mathbb{E}(|X_i|^k)$ for $i = 1, \ldots, d$ and $k = 1, 2, \ldots$. We have

$$\begin{aligned} \mathbb{E}(|X_i|^k) &= 2\int_0^\infty \frac{1}{\sqrt{2\pi\sigma_i^2}} x^k e^{-x^2/(2\sigma_i^2)} dx \\ &= \frac{2^{k/2}\sigma_i^k}{\sqrt{\pi}} \int_0^\infty u^{(k-1)/2} e^{-u} du \\ &= \frac{2^{k/2}\sigma_i^k}{\sqrt{\pi}} \Gamma\left(\frac{k+1}{2}\right) \sim \sqrt{2}\sigma_i^k \left(\frac{k}{e}\right)^{k/2}, \quad \text{as} \quad k \to \infty, \end{aligned}$$

DOI:10.6814/NCCU202201049

where in the last step, we used Stirling's formula for the gamma function. Thus,

$$\sum_{k\geq 0} \alpha_{i,k} \frac{z^k}{k!}$$

has radius of convergence equal to ∞. Consequently, by Theorem 2.2.3, the multivariate normal distribution is uniquely determined by its mixed moments.

As for the computation of the mixed moments, we can apply the following theorem.

Theorem 2.2.4 (Isserlis' theorem)**.** *If $(X_1, X_2, \cdots, X_d)$ is a multivariate normal random variable with zero mean, then*

$$\mathbb{E}(X_1^{k_1} X_2^{k_2} \cdots X_d^{k_d}) = \sum_{\mathcal{P}} \prod_{\{P_1,P_2\}\in\mathcal{P}} \mathrm{Cov}(P_i, P_j),$$

where the sum runs over all partitions into pairs of the multiset

$$\{\underbrace{X_1, \ldots, X_1}_{k_1}, \underbrace{X_2, \ldots, X_2}_{k_2}, \cdots, \underbrace{X_d, \ldots, X_d}_{k_d}\}.$$

For example, if $d = 4$ and $k_1 = k_2 = k_3 = k_4 = 1$, then

$$\mathbb{E}(X_1X_2X_3X_4) = \mathbb{E}(X_1X_2)\mathbb{E}(X_3X_4) + \mathbb{E}(X_1X_3)\mathbb{E}(X_2X_4) + \mathbb{E}(X_1X_4)\mathbb{E}(X_2X_3).$$

Next, we give the multivariate version of Theorem 2.2.2.

Theorem 2.2.5. *Let $X = (X_1, \ldots, X_d)$ be a random variable satisfying the assumptions of Theorem 2.2.3. Assume that $X_n = (X_{n,1}, \ldots, X_{n,d})$ is a sequence of random variables whose mixed moments all exist and satisfy:*

$$\lim_{n\to\infty} \mathbb{E}(X_{n,1}^{k_1} \cdots X_{n,d}^{k_d}) = \beta(k_1, \ldots, k_d)$$

for all non-negative integers k_i. Then, X_n converges in distribution to X, i.e., $X_n \xrightarrow{w} X$.

Thus, if we have

$$\lim_{n\to\infty} \mathbb{E}(X_{n,1}^{k_1} \cdots X_{n,d}^{k_d}) = \mathbb{E}(X_1^{k_1} \cdots X_d^{k_d}),$$

where $(X_1, \ldots, X_d) \sim N(\mathbf{0}, \Sigma)$ with $N(\mathbf{0}, \Sigma)$ denoting the multivariate normal distribution with mean vector 0 and covariance matrix Σ, then X_n satisfies a multivariate central limit theorem, i.e.,

$$X_n \xrightarrow{w} N(\mathbf{0}, \Sigma).$$

Finally, we recall another well-known theorem from probability which will also be used in Chapter 5.

DOI:10.6814/NCCU202201049

Theorem 2.2.6 (Slutsky's theorem)**.** *Let X_n and Y_n be sequences of random variables. If X_n converges in distribution to a random element X and Y_n converges in probability to a constant c, then*

- $X_n + Y_n \xrightarrow{w} X + c$;
- $X_n Y_n \xrightarrow{w} Xc$;
- $X_n / Y_n \xrightarrow{w} X/c$, *provided that* $c \neq 0$.

DOI:10.6814/NCCU202201049

Chapter 3

Enumeration of bi-combining tree-child networks

Most of the studies on tree-child networks have focused on bi-combining tree-child networks. In this chapter, we are going to review some results for these networks and add some new results.

3.1 Results for $\mathrm{OTC}_{n,k}$

In this section, we first give a formula for $\mathrm{OTC}_{n,k}$ which is proved through a recurrence. Then, an asymptotic result can also be obtained by straightforward calculation. Finally, we use the Laplace method mentioned in Section 2.1 to derive an asymptotic result for OTC_n and prove a distributional result for the number of reticulation nodes if one picks a network of $\mathcal{OTC}_n$ uniformly at random.

Theorem 3.1.1. *The number of one-component bi-combining tree-child networks with n leaves and k reticulation nodes is given by*

$$\mathrm{OTC}_{n,k} = \binom{n}{k} \frac{(2n-2)!}{2^{n-1}\,(n-k-1)!}, \qquad (0 \le k \le n-1).$$

Consequently,

$$\mathrm{OTC}_n = \sum_{k=0}^{n-1} \binom{n}{k} \frac{(2n-2)!}{2^{n-1}\,(n-k-1)!};$$

see (1.2).

DOI:10.6814/NCCU202201049

Remark 3.1.2. This formula was first obtained in [5]. The proof we give below is, however, new.

Proof. We use a network N of $\mathcal{OTC}_{n-1,k-1}$ to construct corresponding networks N' in $\mathcal{OTC}_{n,k}$ by adding one additional reticulation node.

First, note that by Proposition 1.1.6, we have that N has $n-1+(k-1)-1$ tree nodes which is also the number of edges leading to a tree node. Also we have $(n-1)-(k-1)$ edges which lead to a leaf that is not under a reticulation node. Thus, we have

$$n-1+(k-1)-1+(n-1)-(k-1)=2n-3.$$

edges where we can add parent nodes of the new reticulation node. (Here, we used the tree-child property.) Consequently, there are

$$\binom{2n-3+2-1}{2}=\binom{2n-2}{2}.$$

choices. Moreover, label the leaf of the newly formed reticulation node with a label from $\{1,\ldots,n\}$ and increase all (old) labels in N which are at least as large as the new label by $+1$ (if there are any). Note that this construction gives each networks N' exactly k times.

Thus, from the above,

$$\mathrm{OTC}_{n,k}=\frac{n}{k}\binom{2n+k-2}{2}\mathrm{OTC}_{n-1,k-1}$$

and by iteration,

$$\mathrm{OTC}_{n,k}=\binom{n}{k}\frac{(2n-2)!}{2^k(2n-k-2)!}\mathrm{OTC}_{n-k,0}. \tag{3.1}$$

Note that $\mathrm{OTC}_{n-k,0}$ is the number of phylogenetic trees with $n-k$ leaves, so by Proposition 1.1.3,

$$\mathrm{OTC}_{n-k,0}=\mathrm{PT}_{n-k,0}=(2(n-k)-3)!!=\frac{(2n-2k-2)!}{2^{n-k-1}(n-k-1)!}.$$

Inserting this into (3.1) gives the claimed result. ∎

If we plot the numbers $\mathrm{OTC}_{n,k}$ using software such as Maple (fixing n and letting k run from 0 to $n-1$), we see that the maximum value of $\mathrm{OTC}_{n,k}$ is near n and that this plot looks like the normal distribution in the vicinity of the peak (see Figure 3.1). In the next result, we show this and use the Laplace method to derive the limit distribution of the number of reticulation nodes.

DOI:10.6814/NCCU202201049

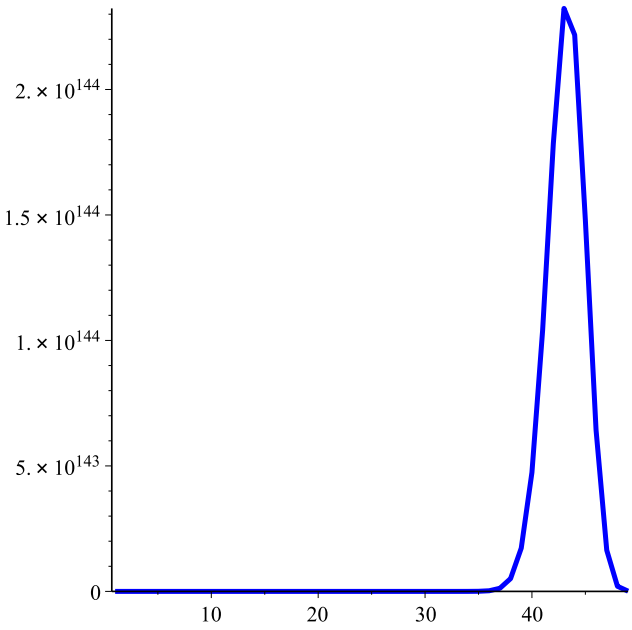


Figure 3.1: A plot of $\mathrm{OTC}_{n,k}$ with $n = 50$ and $0 \leq k \leq 49$. The peak appears at $k = 43$ and the curve looks similar to the density of the normal distribution.

Theorem 3.1.3. *Let R_n be the number of reticulation nodes of a network picked uniformly at random from the set $\mathcal{OTC}_n$. Then, we have the following limit distribution result of R_n:*

$$\frac{R_n - n + \sqrt{n}}{\sqrt[4]{n/4}} \xrightarrow{w} N(0,1).$$

Remark 3.1.4. This result was stated without proof in [15].

Proof. By Theorem 3.1.1, we have

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

The ratio of two consecutive terms in this sum is:

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

From this, we see that $1/k!(n-k)!(n-k-1)!$ is monotonic increasing for $0 \leq k \leq n - \sqrt{n+1}$ and monotonic decreasing for $n - \sqrt{n+1} < k \leq n-1$. Substituting $k = n - \sqrt{n} + x$ and using Stirling's formula gives:

$$\mathrm{OTC}_{n,k} = \frac{1}{2\sqrt{e\pi}} n^{2n-3/2} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n e^{-x^2/\sqrt{n}} \left(1 + \mathcal{O}\left(\frac{1+|x|}{\sqrt{n}} + \frac{|x|^3}{n}\right)\right) \tag{3.2}$$

DOI:10.6814/NCCU202201049

uniformly for $|x| \le n^{3/10}$ and $n \to \infty$.

Next, recall that

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

We can deal with the summation by a standard application of the Laplace method. More precisely,

$$\begin{aligned}
\mathrm{OTC}_n &\sim \frac{1}{2\sqrt{e\pi}} n^{2n-3/2} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n \sum_{x=-n^{3/10}}^{n^{3/10}} e^{-x^2/\sqrt{n}} \left(1 + \mathcal{O}\left(\frac{1+|x|}{\sqrt{n}} + \frac{x^3}{n}\right)\right) \\
&\sim \frac{1}{2\sqrt{e\pi}} n^{2n-3/2} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n \int_{-\infty}^{\infty} e^{-x^2/\sqrt{n}} \left(1 + \mathcal{O}\left(\frac{1+|x|}{\sqrt{n}} + \frac{x^3}{n}\right)\right) dx \qquad (3.3) \\
&\sim \frac{1}{2\sqrt{e}} n^{2n-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right),
\end{aligned}$$

where this actually holds as equality since all errors are exponentially small.

Combining the latter with (3.2) gives

$$\frac{\mathrm{OTC}_{n,k}}{\mathrm{OTC}_n} = \frac{1}{\sqrt{\pi} n^{1/4}} e^{-x^2/\sqrt{n}} \left(1 + \mathcal{O}\left(\frac{1+|x|}{\sqrt{n}} + \frac{|x|^3}{n}\right)\right)$$

uniformly for $k = n - \sqrt{n} + x$ with $|x| \le n^{3/10}$ and $n \to \infty$, or

$$\frac{\mathrm{OTC}_{n,k}}{\mathrm{OTC}_n} = \frac{1}{\sqrt{\pi} n^{1/4}} e^{-t^2/2} \left(1 + \mathcal{O}\left(\frac{1+|t|}{n^{1/4}}\right)\right)$$

uniformly for $k = n - \sqrt{n} + t\sqrt[4]{n/4}$ with $|t| \le n^{1/20}$ and $n \to \infty$. Thus, by similar arguments as used in the Laplace method,

$$\begin{aligned}
P\left(\frac{R_n - n + \sqrt{n}}{\sqrt[4]{n/4}} \le x\right) &= P\left(R_n \le n - \sqrt{n} + x\sqrt[4]{n/4}\right) \\
&\sim \frac{1}{\sqrt{\pi} n^{1/4}} \sum_{t=-\infty}^{x} e^{-t^2/2} \left(1 + \mathcal{O}\left(\frac{1+|t|}{n^{1/4}}\right)\right) \\
&\sim \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.
\end{aligned}$$

This proves the claimed result. ∎

The proof contains the following corollary which is Theorem 3 in [17].

Corollary 3.1.5. *We have,*

$$\mathrm{OTC}_n \sim \frac{1}{4\pi\sqrt{e}} (n!)^2 2^n e^{2\sqrt{n}} n^{-9/4}.$$

Proof. This is obtained from (3.3) and Stirling's formula. ∎

DOI:10.6814/NCCU202201049

3.2 Results for $\mathrm{TC}_{n,k}$

In this section we review the asymptotic result for $\mathrm{TC}_{n,k}$ from [17] and mention a result for the distribution of the number of reticulation nodes (which will be contained in a forthcoming paper). More precisely, we first show relations for $\mathrm{TC}_{n,k}$ with different k when n is fixed. From these relations, it can be observed that $\mathrm{TC}_{n,n-1}$ is the dominant term for the sequence $\mathrm{TC}_{n,k}$. Finally, we use a bijective proof and a method in [11] to get a Θ-result for $\mathrm{TC}_{n,n-1}$.

We start with the following bound.

Lemma 3.2.1. *For any* $1 \leq k \leq n-1$,

$$\mathrm{TC}_{n,n-1-k} \leq \frac{1}{2k}\mathrm{TC}_{n,n-k}.$$

Proof. By Lemma 1.2.4, each tree-child network in $\mathcal{TC}_{n,n-1-k}$ has k free tree nodes and $2k$ free edges. Insert a reticulation node into one of free edges; then the initial node of the free edge can be considered as one of the parents of the inserted node. The other parent is constructed by inserting a node between the root of the network and its child; then add an edge from this parent to the newly formed reticulation node. Since this generation method is injective, the claimed result follows. (Note that in the case $k = 1$, the generating method is even bijective.) ∎

Lemma 3.2.1 shows that $\mathrm{TC}_{n,k}$ increases with k (fixing n). Moreover, it implies that the order of TC_n is that of $\mathrm{TC}_{n,n-1}$.

Proposition 3.2.2. *We have* $\mathrm{TC}_{n,n-1} \leq \mathrm{TC}_n \leq \sqrt{e} \cdot \mathrm{TC}_{n,n-1}$ *and thus* $\mathrm{TC}_n = \Theta(\mathrm{TC}_{n,n-1})$

Proof. For $0 \leq k \leq n-1$, by iterating the inequality in Lemma 3.2.1,

$$\mathrm{TC}_{n,k} \leq \frac{1}{2^{n-1-k}(n-k-1)!}\mathrm{TC}_{n,n-1}. \tag{3.4}$$

Then,

$$\mathrm{TC}_n \leq \sum_{k=0}^{n-1}\left(\frac{1}{2^{n-1-k}(n-k-1)!}\right)\mathrm{TC}_{n,n-1} \leq \sqrt{e} \cdot \mathrm{TC}_{n,n-1}$$

which gives the upper bound. The lower bound is trivial. ∎

Since TC_n has the same growth order as $\mathrm{TC}_{n,n-1}$, we can now focus on finding the growth order of $\mathrm{TC}_{n,n-1}$.

Next we will introduce a bijection which will help us in solving this problem. (Since this bijection will play an important role in this thesis, we discuss it in an own paragraph.)

DOI:10.6814/NCCU202201049

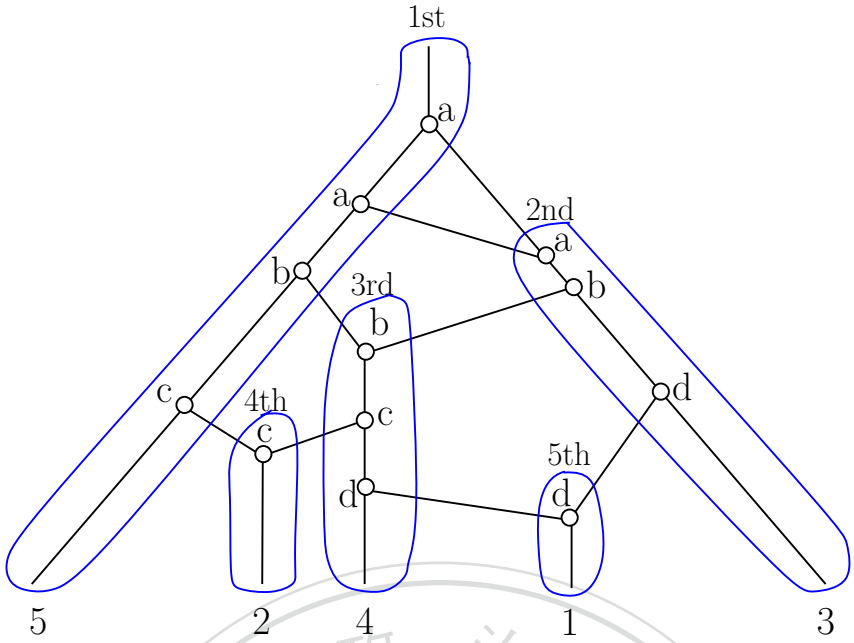


Figure 3.2: An example of a network from $\mathcal{TC}_{5,4}$ with its path-components. This tree-child network is encoded as $aabc5abd3bcd4c2d1$ (where the permutation is inserted into the word with each label at the end of its path-component).

Bijection between $\mathrm{TC}_{n,n-1}$ and words. We start with a graph-theoretical property of $\mathcal{TC}_{n,n-1}$.

Lemma 3.2.3. *A tree-child network with n leaves has $n-1$ reticulation nodes if and only if there is exactly one path from every node to a leaf whose intermediate nodes are all tree nodes.*

Proof. Notice that networks in $\mathcal{TC}_{n,n-1}$ are the only tree-child networks with no free tree nodes. Thus, they cannot have two paths as indicated since two such paths must meet in a free tree node. On the other hand, if there is a free tree node, then two paths as indicated exist because of the tree-child property. ∎

According to this, we can decompose an arbitrary network from $\mathcal{TC}_{n,n-1}$ into n path-components; see Figure 3.2 for an example. We will use this in order to bijectively map $\mathcal{TC}_{n,n-1}$ onto the following class of words.

Definition 3.2.4. *Let $\mathcal{C}_n$ denote the class of words built from n letters $\{\omega_1, \ldots, \omega_n\}$ in which each letter occurs exactly 3 times such that in every prefix the letter ω_i has either not yet occurred, or, if it has, then the number of occurrences of ω_i is at least as large as the number of occurrences of ω_j for all $j > i$.*

DOI:10.6814/NCCU202201049

For example,

$$C_2 = \{aaabbb, aababb, abaabb, baaabb, aabbab, ababab, baabab\}$$

and thus $|\mathcal{C}_2| = 7$.

Proposition 3.2.5. *There is a bijection between $\mathcal{TC}_{n,n-1}$ and $\mathcal{C}_{n-1} \times \mathcal{P}_n$, where $\mathcal{P}_n$ denotes the set of permutations of length n.*

Proof. First, note that networks in $\mathcal{TC}_{n,n-1}$ have no symmetry. Thus, if we remove labels from the networks in $\mathcal{TC}_{n,n-1}$, we obtain $\mathrm{TC}_{n,n-1}/n!$ unlabeled networks. Consequently, it suffices to show that these unlabeled networks can be bijectively mapped onto words of $\mathcal{C}_{n-1}$.

We begin by ordering the path-components of an unlabeled network from $\mathcal{TC}_{n,n-1}$. We do this recursively. First, the path-component starting from the child of the root is considered as the first path-component. Next, assume that k path-components have been indexed. We read these indexed path-components from the path-component with smallest index to the path-component with largest index and each path-component from the beginning node (which is a reticulation node or the child of the root) to its corresponding leaf. Now, consider all un-indexed path-components whose two parents are already in those indexed path-components. Order its two parents so that the first parent to be read is called *first parent* and the second parent to be read is called *second parent*. Now, order all these un-indexed path-components according to the order of their second parents. Due to the connectivity of tree-child network, we can continue doing this until all path-components are indexed which will eventually happen; see Figure 3.2 for an example.

Next, label a reticulation node and both its parents by the index of the path-component which starts from it. (This gives labels from the set $\{2, \ldots, n-1\}$ which alternatively can be replaced by the letters of the alphabet as follows: $2 \to a, 3 \to b, 4 \to c, \ldots$; Figure 3.2.) Read all path-components and nodes in these path-components in order. Then, we can get a word in $\mathcal{C}_{n-1}$.

Finally, it is not difficult to see that the above construction can be reversed. Thus, it is a bijection and the proof is finished. ∎

The previous proposition shows that we can (asymptotically and exactly) count $\mathcal{TC}_{n,n-1}$ via $\mathcal{C}_{n-1}$.

DOI:10.6814/NCCU202201049

The authors in [17] found a recurrence for the cardinality of $\mathcal{C}_{n-1}$ to which they applied the method in [11] to get an asymptotic result for the cardinality of $\mathcal{C}_{n-1}$. (We will introduce an extension of this recurrence in Section 4.2.) More precisely, they obtained

$$|\mathcal{C}_{n-1}| = \Theta\left(n! n^{-5/3} e^{\alpha_1 (3n)^{1/3}} 12^n\right), \tag{3.5}$$

where $\alpha_1 = -2.338107410\ldots$ is the largest root of the Airy function of the first kind. Thus, by Stirling's formula,

$$\mathrm{TC}_{n,n-1} = \Theta\left(n^{-2/3} e^{\alpha_1 (3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\right),$$

and finally from Proposition 3.2.2,

$$\mathrm{TC}_n = \Theta\left(n^{-2/3} e^{\alpha_1 (3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\right). \tag{3.6}$$

This was the main result in [17].

As for the number of reticulation nodes for a random network from $\mathcal{TC}_n$, we recently managed to show a corresponding lower bound in Proposition 3.2.1 which is asymptotically sharp for k close to n. This lower bound (together with the upper bound in Proposition 3.2.1) implies that for k close to n, (3.4) can be sharpened to

$$\mathrm{TC}_{n,n-1-k} \approx \frac{1}{2^k k!} \mathrm{TC}_{n,n-1}, \tag{3.7}$$

i.e., the number of reticulation nodes satisfies a Poisson limit law.

Theorem 3.2.6. *Let I_n be the number of reticulation nodes of a network picked uniformly at random from the set $\mathcal{TC}_n$. Then, we have the following limit distribution result of I_n:*

$$n - 1 - I_n \xrightarrow{w} \text{Poisson}\left(\frac{1}{2}\right).$$

The proof of this result will be presented in the journal version of [7]. (It was mentioned as a conjecture in the conclusion of [7].) We also remark that (3.7) allows us to improve the Θ-result in Proposition 3.2.2.

Corollary 3.2.7. *We have,*

$$\mathrm{TC}_n \sim \sqrt{e} \cdot \mathrm{TC}_{n,n-1}.$$

This, however, does not yield an improvement of (3.6) since the method in [11] is only capable of giving a Θ- result for $\mathrm{TC}_{n,n-1}$.

DOI:10.6814/NCCU202201049

3.3 Pons and Batle's conjecture

In this section, we will present a conjecture of Pons and Batle for $\mathrm{TC}_{n,k}$. This conjecture was found by them by generalizing the bijection between $\mathcal{TC}_{n,k}$ and $\mathcal{S}_{n-1} \times \mathcal{P}_n$ from the last section. We start with a generalization of Definition 3.2.4.

Definition 3.3.1. *Let $\mathcal{C}_{n,k}$ denote the class of words built from n letters $\{\omega_1, \ldots, \omega_n\}$ in which $\{\omega_1, \ldots, \omega_k\}$ occur exactly 3 times and $\{\omega_{k+1}, \ldots, \omega_n\}$ occur 2 times such that in every prefix the letter ω_i has either not yet occurred, or, if it has, then the number of occurrences of ω_i is at least as large as the number of occurrences of ω_j for all $j > i$.*

For example:

$$|\mathcal{C}_{2,0}| = 3: \quad \mathcal{C}_{2,0} = \{aabb, abab, baab\};$$
$$|\mathcal{C}_{2,1}| = 7: \quad \mathcal{C}_{2,1} = \{aaabb, aabab, abaab, baaab, aabba, ababa, baaba\}.$$

Note that $\mathcal{C}_n$ in Definition 3.2.4 is exactly $\mathcal{C}_{n,n}$.

The number of words in $\mathcal{C}_{n,k}$ satisfy an easy recurrence.

Proposition 3.3.2. *The number $\mathrm{c}_{n,k}$ of words in $\mathcal{C}_{n,k}$ satisfies the recurrence:*

$$\mathrm{c}_{n,k} = \mathrm{c}_{n,k-1} + (2n + k - 1)\mathrm{c}_{n-1,k}.$$

Proof. Because the last letter of a word in $\mathcal{C}_{n,k}$ must be the third occurrence of ω_k or the second occurrence of ω_n, we can divide $\mathcal{C}_{n,k}$ into two parts according two whether we assume the former (first part) or the latter (second part).

For the first part, we remove the third ω_k. This gives a one-to-one map to $\mathcal{C}_{n,k-1}$.

For the second part, we remove the two ω_n's. Note that the position of the second ω_n was fixed and the first ω_n can appear in $2n + k - 1$ positions.Thus, the number of words from the second part is $(2n + k - 1)\mathrm{c}_{n-1,k}$.

Combining the two parts gives the claimed recurrence. ∎

The conjecture of Pons and Batle is now as follows.

Conjecture 3.3.3 (Pons and Batle)**.** *$\mathrm{TC}_{n,k}$ and $\mathrm{c}_{n,k}$ are related as follows:*

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

DOI:10.6814/NCCU202201049

Two cases of the conjecture are clear.

Proposition 3.3.4. *The conjecture is true for $k = 0$ and $k = n$.*

Proof. For $k = n$, the conjecture holds by Proposition 3.2.5.

For $k = 0$, consider a permutation of the letters $\{\omega_1, \ldots, \omega_n\}$ with each letter repeated twice. This gives $(2n)!/2^n$ such permutations. Next, assume that the second letters in order of their occurrences are ω_1' to ω_n'. Then, we rename them (together with the corresponding first letters) by $\omega_1, \ldots, \omega_n$. There are $n!$ possibilities for this. Consequently,

$$\mathrm{c}_{n,0} = \frac{(2n)!}{2^n n!}$$

and thus

$$c_{n-1,0} = \frac{(2n-2)!}{2^{n-1}(n-1)!} = (2n-3)!! = \mathrm{PT}_n;$$

see Proposition 1.1.3. This proves the conjecture also for $k = n$. ∎

3.4 Bijection for $\mathcal{OTC}_{n,k}$

In this section, we give a proof of the conjecture mentioned in Section 3.3 for one-component tree-child networks. This part is currently under submission; see [15].

We start by defining a class of words.

Definition 3.4.1. *Let $\mathcal{H}_{n,k}$ denote the subset of words from $\mathcal{C}_{n,k}$ which satisfy that the letter right before the third occurrence of a letter is the second occurrence of that letter.*

An example for a word in $\mathcal{H}_{4,2}$ with the letters in the order a, b, c, d, e is $badacbacdb$.

Lemma 3.4.2. *There is a bijection between $\mathcal{H}_{n,k}$ and $\mathcal{C}_{n,0} \times \binom{\{\omega_1,\ldots,\omega_n\}}{k}$, where $\binom{A}{k}$ denotes the set consisting of subsets of A of cardinality k.*

Proof. We can remove the third occurrence of the letters $\omega_1, \ldots, \omega_k$ and the resulting word is in $\mathcal{C}_{n,0}$. Next, record the letters which were right before the removed letters. Clearly, these letters form a subsets of $\{\omega_1, \ldots, \omega_n\}$ of cardinality k. Also, this process can be reversed. ∎

Taking the word $efbadacb\mathbf{a}c\mathbf{b}de f\mathbf{c}$ as example, we see that the third **a** follows the second b, the third **b** follows the second c, and the third **c** follows the second f. Thus, the set in the

DOI:10.6814/NCCU202201049

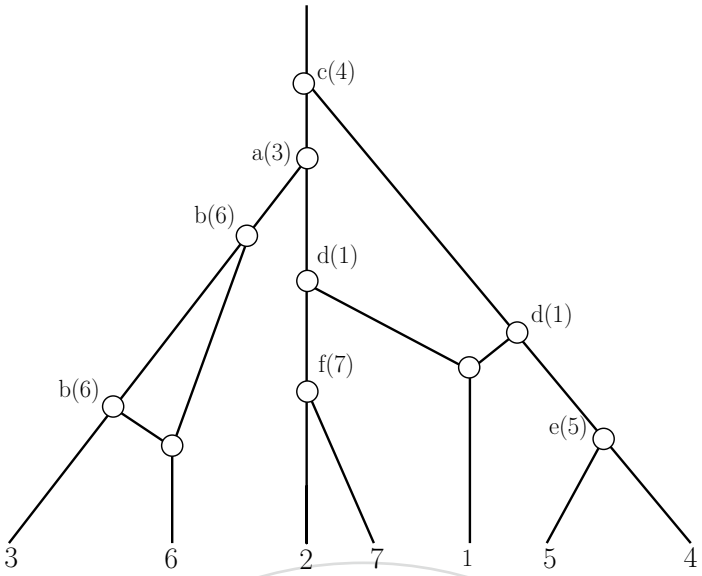


Figure 3.3: A one-component tree-child network with the internal nodes labeled as in the proof of Theorem 3.4.4.

above lemma is $\{b, c, f\}$. Moreover, we remove **a,b** and **f** and get: $efbadacbcdef$. (We will subsequently write $efbadacbcdef + \{b, c, f\}$ for the value of the bijiection from the above lemma.)

Lemma 3.4.3. *We have,*

$$|\mathcal{H}_{n,k}| = \binom{n}{k}(2n-1)!!.$$

Proof. Because $\binom{\{\omega_1,\ldots,\omega_n\}}{k}$ consists of all subsets of $\{\omega_1, \ldots, \omega_n\}$ of cardinality k, the number of these subsets is $\binom{n}{k}$. Thus, the cardinality of $\mathcal{H}_{n,k}$ equals $\binom{n}{k}$ times $|\mathcal{C}_{n,0}|$ and the result follows from Proposition 3.3.4. ∎

Let $A^{\underline{k}}$ denote the set consisting of k-tuples without repetition of elements from A. For example, $\{1, 2, 3\}^{\underline{2}} = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)\}$.

We can now prove Pons and Batle's conjecture for one-component tree-child network.

Theorem 3.4.4. *There is a bijection between $\mathcal{OTC}_{n,k}$ and $\mathcal{H}_{n-1,k} \times \{1, \ldots, n\}^{\underline{k}}$. Thus,*

$$\mathrm{OTC}_{n,k} = \frac{n!}{(n-k)!} \times |\mathcal{H}_{n-1,k}|.$$

Proof. We will directly give the map between the two sets from the theorem; see Figure 3.3 for an example.

DOI:10.6814/NCCU202201049

From $\mathcal{OTC}_{n,k}$ to $\mathcal{H}_{n-1,k} \times \{1,\dots,n\}^{\underline{k}}$. Consider a network from $\mathcal{OTC}_{n,k}$. First, we remove the reticulation nodes, their two incoming edges and their outgoing edge and leaf. This gives a tree with some unary nodes (nodes of indegree and outdegree 1), where these unary nodes are parents of the reticulation nodes which we have just removed. We label the unary nodes with the labels of the removed leaves below their removed reticulation nodes.

Next, we define path-components in the resulting tree where every path-component has an index. The first path-component is the unique path from the child of the root to the leaf with smallest label. The second path-component is from a unique node on an indexed path-component (currently there is only one path-component) to the second smallest label. Continuing like this step by step, we can find all $n-k$ path-components.

Now, the nodes without label are nodes where two path-components overlap; label them by the label of the larger leaf at the end of the two path-components. Then, remove the labels of the leaves.

Next, we read the labels of the nodes along each path-component starting from the the one with the smallest index and proceeding until the one with the largest index; call the resulting word $\tilde{\omega}$. Also, put the labels of the unary nodes into a set $\tilde{A}$. For example, for the network in Figure 3.3 this gives $\tilde{\omega} = 431736641557$ and $\tilde{A} = \{1,6\}$.

Next, we convert $\tilde{\omega}$ and $\tilde{A}$ into $\mathcal{H}_{n-1,k}$ and $\{1,\dots,n\}^{\underline{k}}$.

First, we form a k tuple by listing the elements from $\tilde{A}$ in the order they occur for the second time in $\tilde{\omega}$. For the example in Figure 3.3 this gives the tuple $(6,1)$.

Next, we rename the numbers in $\tilde{\omega}$ and $\tilde{A}$ as follows: if a number in $\tilde{\omega}$ appears the second time and there are already $j-1$ numbers appearing twice before it, then rename the number by ω_j. This converts $\tilde{\omega}$ and $\tilde{A}$ into ω and A. Moreover, via the map from Lemma 3.4.2, this gives a word from $\mathcal{H}_{n-1,k}$. For the example from Figure 3.3, the mapping is:

$$3 \to a, \quad 6 \to b, \quad 4 \to c, \quad 1 \to d, \quad 5 \to e, \quad 7 \to f.$$

Thus, $\omega = cadfabbcdeef$ and $A = \{b,d\}$ and hence $cadfabbcdeef + \{b,d\}$. Finally, this gives the word $cadfabbacdbeef$.

Therefore, the corresponding word is $cadfabbacdbeef$ and k-tuple is $(6,1)$.

From $\mathcal{H}_{n-1,k} \times \{1,\dots,n\}^{\underline{k}}$ to $\mathcal{OTC}_{n,k}$. We use again the word $cadfabbacdbeef$ and the tuple

DOI:10.6814/NCCU202201049

$(6,1)$ from above which is (considered as a tuple) an element of $\mathcal{H}_{6,2} \times \{1,\ldots,7\}^{\underline{2}}$.

We first separate the given word into a word ω and a set A by the method from Lemma 3.4.2. (For our example: $\omega = cadfabbcdeef$ and $A = \{b,d\}$.)

Then, we use the second occurrence of the letters from $B := \{\omega_1,\ldots,\omega_{n-1}\} \setminus A$ to separate ω into $n-k$ subwords: the first subword is from the beginning to the letter before the second occurrence of the first letter from B; the next is from the second occurrence of the first letter from B to the letter before the second occurrence of the second letter from B; etc.

For example, $cadf\mathbf{a}bb\mathbf{c}de\mathbf{ef} \to cadf|\mathbf{a}bb|\mathbf{c}de|\mathbf{e}|\mathbf{f}$ giving the subwords: $cadf, abb, cde, e, f$.

These subwords correspond to nodes along paths in a tree that is successively constructed as follows: the i-th subword corresponds to the i-th path which ends in a leaf which is labeled by the i-th smallest integer from $\{1,\ldots,n\}$ which is not in the k-tuple (in our example, the leaf of the first path equals 2, the leaf of the second path equals 3; etc.) Also, the i-th subword starts with a letter which is contained on one of the former $i-1$-st paths; thus, we can recursively construct the tree by starting from this letter and attaching the i-th path (assuming that the first $i-1$-st path are already attached).

Next, reticulation nodes are attached whose parents are the unary nodes on the tree with identical labels. Moreover, the leaves of the reticulation nodes are labeled with the integers from the k-tuple in the order of the second occurrence of the letters of the parents in ω.

Finally, removing all labels of the internal nodes gives the preimage of the map from above. This completes the proof. ∎

Remark 3.4.5. From the above result and Lemma 3.4.3, we obtain

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

Note that this is Theorem 3.1.1. Thus, we found a second proof of this theorem via a bijection.

One immediate question is whether our proof can be extended to yield the full conjecture of Pons and Batle? Since one-component networks are still relatively easy, this is, at present, still unclear. We leave this as an open problem.

DOI:10.6814/NCCU202201049

Chapter 4

Enumeration of d-combining tree-child networks

In this chapter, we generalize the results from the previous chapter to d-combining tree-child networks. Most of these results are new and have either appeared in the recent extended abstract [7] or are mentioned in the conclusion of [7] as open problems. These results (together with more results) will appear in the journal version of [7] which is currently under preparation.

4.1 Exact formula for $\mathrm{OTC}_{n,k}^{[d]}$

In this section, we will extend Theorem 3.1.1 to d-combining networks. The result is as follows.

Theorem 4.1.1. *The number of one-component d-combing tree-child networks with n leaves and k reticulation nodes is given by*

$$\mathrm{OTC}_{n,k}^{[d]} = \binom{n}{k} \frac{(2n+(d-2)k-2)!}{(d!)^k\, 2^{n-k-1}\, (n-k-1)!}, \qquad (0 \le k \le n-1).$$

Proof. The proof is similar to the one of Theorem 3.1.1 except that now we have to count the possible positions for d parents of the reticulation nodes instead of just two. Notice that the number of tree nodes has also changed; now it is $n-1+(d-1)(k-1)-1$ (see Proposition 1.2.7).

So, we have $2n+(d-2)(k-1)-3$ edges in each network of $\mathcal{OTC}_{n-1,k-1}^{[d]}$ where we can add parent nodes of reticulation nodes and thus there are

$$\binom{2n+(d-2)(k-1)-3+d-1}{d} = \binom{2n+(d-2)k-2}{d}.$$

DOI:10.6814/NCCU202201049

choices.

By iteration, we obtain

$$\mathrm{OTC}_{n,k}^{[d]} = \binom{n}{k} \frac{(2n+(d-2)k-2)!}{d!^k (2n-k-2)!} \mathrm{OTC}_{n-k,0}^{[d]}. \tag{4.1}$$

Moreover, by Proposition 1.1.3,

$$\mathrm{OTC}_{n-k,0}^{[d]} = (2(n-k)-3)!! = \frac{(2n-2k-2)!}{2^{n-k-1}(n-k-1)!}.$$

Inserting this into (4.1) gives the claimed result. ∎

Remark 4.1.2. A second way to count $\mathrm{OTC}_{n,k}^{[d]}$ proceeds by constructing $\mathrm{OTC}_{n,k}^{[d]}$ from $\mathrm{OTC}_{n,k}^{[d-1]}$. To give details, first, there are $2n+(d-3)k-1$ edges with end points either a tree node or a leaf in each network of $\mathcal{OTC}_{n,k}^{[d-1]}$. We add k different nodes (each node is the d-th parent of a reticulation node) to these edges. Overall there are

$$(2n-2+(d-3)k+1)(2n-2+(d-3)k+2)\cdots(2n-2+(d-3)k+k)$$

ways to do this. Now, if we assign the first of the k nodes to the first reticulation node, the second of the k nodes to the second reticulation node, etc., we obtain every network in $\mathrm{OTC}_{n,k}^{[d]}$ exactly d^k times. Thus,

$$\frac{\mathrm{OTC}_{n,k}^{[d]}}{\mathrm{OTC}_{n,k}^{[d-1]}} = \frac{(2n-2+(d-3)k+1)(2n-2+(d-3)k+2)\cdots(2n-2+(d-3)k+k)}{d^k}.$$

Since we already have $\mathrm{OTC}_{n,k}^{[2]}$ (see Theorem 3.1.1), we can get the result for d-combining networks from the above relation by iteration.

4.2 Counting $\mathrm{TC}_{n,k}^{[d]}$ by modified words

In Section 3.2, we mentioned a bijection between $\mathcal{TC}_{n,n-1}$ and $\mathcal{C}_n \times \mathcal{P}_n$ (Proposition 3.2.5). Furthermore, in Section 3.3, we presented a conjecture of Pons and Batle which generalizes this bijection to $\mathcal{TC}_{n,k}$ and $\mathcal{C}_{n,k}$ (Conjecture 3.3.3). Now, we introduce another bijection which maps $\mathcal{TC}_{n,k}^{[d]}$ onto a class of words which is different from the class of words used in Conjecture 3.3.3. Also, in contrast the the conjecture of Pons and Batle, we can prove that our map is indeed a bijection.

We start with $d=2$ since $d \geq 3$ is similar.

DOI:10.6814/NCCU202201049

Definition 4.2.1. *Let $\mathcal{W}_{n,k}^{[2]}$, or $\mathcal{W}_{n,k}$ for short, denote the class of words built from n letters $\{\omega_1, \ldots, \omega_n\}$ in which k letters occur exactly 3 times and the remaining $n-k$ letters occur 2 times and which satisfy a special property describen next. Consider the 0-th, 1-st and 2-nd occurrences of the $n-k$ letters which occur 2 times as 1-st, 2-nd and 3-rd occurrences of these letters. Then, with this convention, the property says that in every prefix of the word the letter ω_j has either not yet occurred, or, if it has, then the number of occurrences of these ω_j is at least as large as the number of occurrences of ω_j for all $j > i$.*

For example:

$$|\mathcal{W}_{2,0}| = 2: \quad \mathcal{W}_{2,0} = \{aabb, abab\};$$
$$|\mathcal{W}_{2,1}| = 7: \quad \mathcal{W}_{2,1} = \{abbab, babab, aabab, aabbb, ababb, baabb, aaabb\};$$
$$|\mathcal{W}_{2,2}| = 7: \quad \mathcal{W}_{2,2} = \{aabbab, ababab, baabab, aaabbb, aababb, abaabb, baaabb\}.$$

Note that $\mathcal{W}_{n,n}$ is exactly $\mathcal{C}_{n,n} = \mathcal{C}_n$.

We have the following relation between $\mathcal{TC}_{n,k}$ and the cardinality of $\mathcal{W}_{n,k}$ which we prove bijectively.

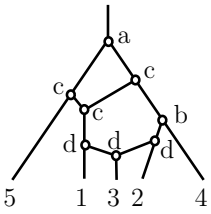
Theorem 4.2.2. *$\mathrm{TC}_{n,k}$ and $|\mathcal{W}_{n-1,k}|$ are related as follows:*

$$\mathrm{TC}_{n,k} = \frac{n!}{2^{n-k-1}} |\mathcal{W}_{n-1,k}|.$$

Proof. First, recall that we gave an order of the path-components for networks from $\mathcal{TC}_{n,n-1}$ in Propostion 3.2.5. Each tree node on the path-components in such networks is not a free tree node. However, when we replace $\mathcal{TC}_{n,n-1}$ by $\mathcal{TC}_{n,k}$, we will meet free tree nodes while reading path-components. Here, we give a new ordering of the path-components for such networks which is a generalization of the ordering for $\mathcal{TC}_{n,n-1}$. (In fact, we give several different orderings for each network when $k < n-1$.)

When we meet a free tree node, we consider the tree node as both a reticulation node and its second parent. Since a free tree node leads to two free edges, we need to decide which edge should be read first. Once this decision is taken for every free tree node, we follow the index rule used in Proposition 3.2.5 (with the path to the other free edge considered as the path starting from the free tree node considered as reticulation node). For each such ordering, we get a different word and permutation with the bijection from Proposition 3.2.5 (and each of them give an encoding of our network); see Figure 4.1 for an example.

DOI:10.6814/NCCU202201049



A network in $\mathcal{TC}_{5,2}$

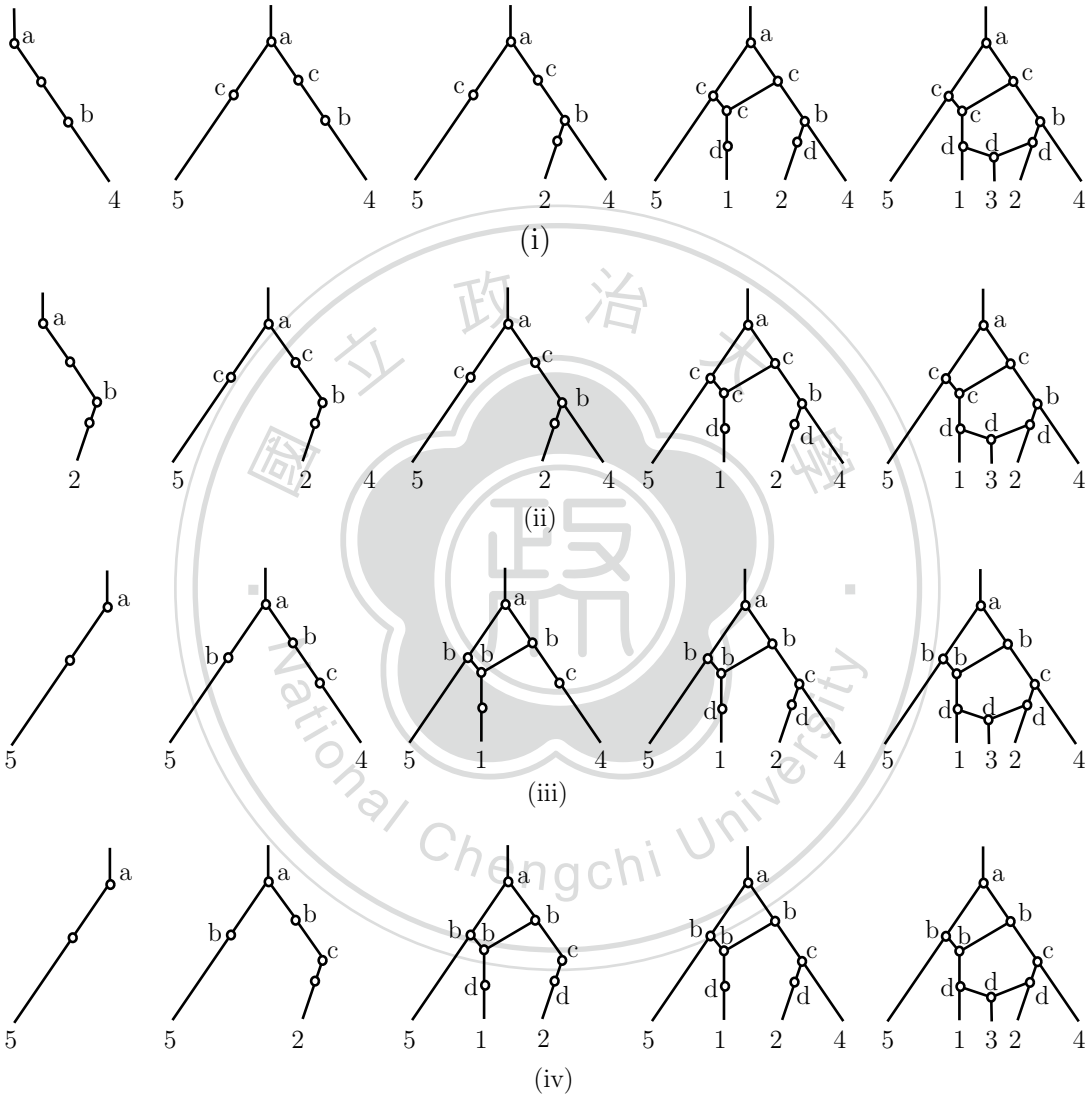


Figure 4.1: The four different ways to label the internal nodes and order the path-components of the tree-child network on the top (which is a network from $\mathcal{TC}_{5,2}$). These different ways produce the following words (with the permutation inserted into the word with each label at the end of its path-component): $acb4ac5bd2cd1d3$ for (i); $acbd2ac5b4cd1d3$ for (ii); $ab5abc4bd1cd2d3$ for (iii); and $ab5abcd2bd1c4d3$ for (iv).

DOI:10.6814/NCCU202201049

Clearly, the above gives a multivalued function between $\mathcal{TC}_{n,k}$ and $\mathcal{W}_{n-1,k} \times \mathcal{P}_n$, where $\mathcal{P}_n$ denotes the set of permutations of length n. What is left is to count the number of images of each network. Therefore, recall that each network from $\mathcal{TC}_{n,k}$ has exactly $n-k-1$ free tree nodes; see Corollary 1.2.4. Moreover, since there are two choices when we encounter a free tree node, the number of images of each network from $\mathcal{TC}_{n,k}$ under the above mentioned map is 2^{n-k-1}. This concludes the proof. ∎

Now, we extend the above result from $d = 2$ to all $d \geq 2$. We first need an extension of $\mathcal{W}_{n,k}$.

Definition 4.2.3. *Let $\mathcal{W}_{n,k}^{[d]}$ denote the class of words built from n letters $\{\omega_1, \ldots, \omega_n\}$ in which k letters occur exactly $d+1$ times and the remaining $n-k$ letters occur 2 times and which satisfy a special property describen next. Consider the 0-th, 1-st and 2-nd occurrences of the $n-k$ letters which occur 2 times as $d-1$-st, d-th and $d+1$-st occurrences of these letters. Then, with this convention, the property says that in every prefix of the word the letter ω_j has either not yet occurred more than $d-2$ times, or, if it has, then the number of occurrences of these ω_j is at least as large as the number of occurrences of ω_j for all $j > i$.*

For example:

$$
\begin{aligned}
|\mathcal{W}_{2,0}^{[3]}| &= 2: && \mathcal{W}_{2,0}^{[3]} = \{aabb, abab\};\\
|\mathcal{W}_{2,1}^{[3]}| &= 11: && \mathcal{W}_{2,1}^{[3]} = \{aaaabb, aabbbb, ababbb, baabbb, abbabb, bababb, bbaabb, abbbab,\\
& && \quad babbab, bbabab, aaabab\};\\
|\mathcal{W}_{2,2}^{[3]}| &= 25: && \mathcal{W}_{2,2}^{[3]} = \{aaabbabb, aababbab, abaabbab, baaababb, aabbaabb, ababaabb,\\
& && \quad baabaabb, abbaaabb, babaaabb, bbaaaabb, aaababbb, aabaabbb,\\
& && \quad abaaabbb, baaaabbb, aaaabbbb, abababab, aabbabab, baababab,\\
& && \quad abbaabab, bbaaabab, babaabab, aaabbbab, aababbab, abaabbab,\\
& && \quad baaabbab\}.
\end{aligned}
$$

Note that $\mathcal{W}_{n,k}^{[2]} = \mathcal{W}_{n,k}$.

Now, we can generalize Theorem 4.2.2.

Theorem 4.2.4. *$\mathrm{TC}_{n,k}^{[d]}$ and $|\mathcal{W}_{n-1,k}^{[d]}|$ are related as follows:*

$$\mathrm{TC}_{n,k}^{[d]} = \frac{n!}{2^{n-k-1}} |\mathcal{W}_{n-1,k}^{[d]}|.$$

DOI:10.6814/NCCU202201049

Proof. First, note that in a network in $\mathcal{TC}^{[d]}_{n,k}$, the number of free tree node is still $n-k-1$; see Proposition 1.2.7.

Next, generalizing the indexing procedure for path-components from Theorem 4.2.2, for a network from $\mathcal{TC}^{[d]}_{n,k}$, we index path-components by the d-th parent of reticulation nodes and consider free tree nodes as both reticulation nodes and their d-th parents. The reticulation nodes and their d parents correspond to k letters which occur $d+1$ times, the free tree nodes correspond to $n-1-k$ letters which occur 2 times. The rest is the same as in the proof of Theorem 4.2.2. ∎

Next, we introduce a method to count $\mathcal{W}^{[d]}_{n-1,k}$. More precisely, we give a recurrence with which $|\mathcal{W}^{[d]}_{n-1,k}|$ and thus $\mathrm{TC}^{[d]}_{n,k}$ can be computed for small values of n and k (via the above theorem).

Theorem 4.2.5. *For $d \geq 2$ let $b^{[d]}_{n,k,m}$ $(1 \leq m \leq n, 0 \leq k \leq n)$ be defined recursively as*

$$b^{[d]}_{n,k,m} = \sum_{j=1}^{m} b^{[d]}_{n-1,k,j} + \binom{n+m+dk-k-2}{d-1} \sum_{j=1}^{m} b^{[d]}_{n-1,k-1,j}$$

with initial conditions $b^{[d]}_{n,k,m} = 0$ for $n < m$, $b^{[d]}_{n,-1,m} = 0$ and $b^{[d]}_{1,0,1} = 1$. Then,

$$|\mathcal{W}^{[d]}_{n,k}| = \sum_{m \geq 1} b^{[d]}_{n,k,m}.$$

Remark 4.2.6. This recurrence generalizes both the recurrence from Proposition 3 in [17] (bi-combining tree-child networks and $k = n-1$) and Lemma 3.4 in [7] (d-combining tree-child networks with $k = n-1$).

Proof. Let $\mathcal{B}^{[d]}_{n,k,m}$ denote the subset of words in $\mathcal{W}^{[d]}_{n,k}$ that end with the suffix

$$\omega_n \omega_m \dots \omega_{n-1} \omega_n.$$

Moreover, set $b^{[d]}_{n,k,m} := |\mathcal{B}^{[d]}_{n,k,m}|$.

For example, $\mathcal{B}^{[3]}_{2,1,1}$ denotes the subset of words from $\mathcal{W}^{[3]}_{2,1}$ that end with bab. So,

$$b^{[3]}_{2,1,1} = 4: \quad \mathcal{B}^{[3]}_{2,1,1} = \{abbbab, babbab, bbabab, aaabab\}.$$

To prove the recurrence, we divide $\mathcal{B}^{[d]}_{n,k,m}$ into two parts according to whether ω_n is a letter which occurs twice or $d+1$ times.

DOI:10.6814/NCCU202201049

If ω_n occurs twice, then we construct such words from $\mathcal{B}^{[d]}_{n-1,k,j}$. Since the suffix of our words is $\omega_n\omega_m\dots\omega_{n-1}\omega_n$, we can collect all words in $\mathcal{B}^{[d]}_{n-1,k,j}$ with suffix $\omega_m\dots\omega_{n-1}$ and add one ω_n before the last ω_m and one ω_n after the last ω_{n-1}. Note that in order that this is possible, we must have $j \le m$. Also recall that $|\mathcal{B}^{[d]}_{n-1,k,j}| = b^{[d]}_{n-1,k,j}$. Thus, we overall obtain $\sum_{j=1}^{m} b^{[d]}_{n-1,k,j}$ words in this case.

Next, if ω_n occurs $d+1$ times, then we construct such words from $\mathcal{B}^{[d]}_{n-1,k-1,j}$. The construction is the same as above, but with the difference that we now have to add $d-1$ more ω_n's to words from $\mathcal{B}^{[d]}_{n-1,k-1,j}$. Since words from the latter set have

$$(d+1)k + 2(n-k) - (n-m+2) = n+m+dk-k-2$$

letters in front of the suffix $\omega_m\dots\omega_{n-1}$, we have

$$\binom{n+m+dk-k-2}{d-1}$$

possible choices for the positions of the first $d-1$ ω_n's. Thus, the number of words in the second part are $\binom{n+m+dk-k-2}{d-1}\sum_{j=1}^{m} b^{[d]}_{n-1,k-1,j}$.

We add the two parts together and get the result. ∎

We conclude this section with an observation which might be helpful for proving Conjecture 3.3.3. (We drop the superindex of $b^{[d]}_{n,k,m}$ since $d=2$.)

Lemma 4.2.7. *The following relation between $|\mathcal{C}_{n,m}|$ and $b_{n,k,m}$ holds:*

$$|\mathcal{C}_{n,m}| = \frac{b_{n+1,n+1,m+1}}{2n+m+1}.$$

Proof. Every letter of a word counted by $b_{n+1,n+1,m+1}$ repeats 3 times. Moreover, every such word ends with $\omega_{m+1}\dots\omega_{n+1}$. So, if we delete this suffix and the three occurrences of ω_{n+1}, we obtain a word in $\mathcal{C}_{n,m}$. In addition, since the first ω_{n+1} can appear in $2n+m+1$ places, we obtain every word in $\mathcal{C}_{n,m}$ exactly $2n+m+1$ times. This gives the claimed relation. ∎

4.3 Distributional and asymptotic results for $\mathrm{OTC}^{[d]}_{n,k}$ and $\mathrm{TC}^{[d]}_{n,k}$

In the last two sections, we obtained a formula for $\mathcal{OTC}^{[d]}_{n,k}$ and certain words to count networks from $\mathcal{TC}^{[d]}_{n,k}$ in a bijective way. Now we are going to use these results and tools in asymptotic combinatorics (such as the Laplace method) to show some distributional and asymptotic results if we pick one network from $\mathcal{OTC}^{[d]}_{n}$ and $\mathcal{TC}^{[d]}_{n}$ uniformly at random.

DOI:10.6814/NCCU202201049

4.3.1 Results for $\mathrm{OTC}_{n,k}^{[d]}$

Form Theorem (4.1.1), we obtain the following limit distribution result for the number of reticulation nodes. (Note that the result for $d = 2$ is already contained in Theorem 3.1.3.)

Theorem 4.3.1. *Let $R_n^{[d]}$ be the number of reticulation nodes of network picked uniformly at random from the set $\mathcal{OTC}_{n,k}^{[d]}$. Then, we have the following limit distribution result of $R_n^{[d]}$.*

(i) For $d = 3$, we have

$$n - 1 - R_n^{(3)} \xrightarrow{w} \mathrm{Bessel}(1, 2),$$

where Bessel(v, a) denotes the Bessel distribution, i.e.,

$$\mathbb{P}(\mathrm{Bessel}(1, 2) = k) = \frac{1}{I_1(2)k!(k+1)!}, \qquad (k \geq 0).$$

Here, $I_v(a) = \left(\frac{a}{2}\right)^v \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+v+1)} \frac{a^{2k}}{4^k}$ is the modified Bessel function of the first kind.

(ii) For $d \geq 4$, the limit law of $n - 1 - R_n^{[d]}$ is degenerate, i.e.,

$$n - 1 - R_n^{[d]} \xrightarrow{w} 0.$$

Proof.

(i) For $d = 3$, by Theorem 4.1.1, we have the formula

$$\mathrm{OTC}_{n,k}^{[3]} = \binom{n}{k} \frac{(2n + k - 2)!}{3^k 2^{n-1}(n - k - 1)!}, \qquad (0 \leq k \leq n - 1).$$

Note that this sequences is increasing in k. (This is in contrast to $d = 2$ where this sequence has a maximum at $k = n - \sqrt{n + 1}$; see the proof of Theorem 3.1.3.) By replacing k by $n - 1 - k$ and using Stirling's formula, we obtain that

$$\mathrm{OTC}_{n,n-1-k}^{[3]} = \frac{1}{k!(k+1)!} \cdot \frac{n(3n-3)!}{6^{n-1}} \left(1 + \mathcal{O}\left(\frac{1 + k^2}{n}\right)\right) \tag{4.2}$$

uniformly for k with $k = o(\sqrt{n})$. Thus, by a standard application of the Laplace method:

$$\mathrm{OTC}_n^{[3]} \sim \left(\sum_{k \geq 0} \frac{1}{k!(k+1)!}\right) \cdot \frac{n(3n-3)!}{6^{n-1}} = I_1(2) \cdot \frac{n(3n-3)!}{6^{n-1}}. \tag{4.3}$$

Now, since

$$\mathbb{P}(R_n^{[3]} = n - 1 - k) = \frac{\mathrm{OTC}_{n,n-1-k}^{[3]}}{\mathrm{OTC}_n^{[3]}},$$

the claimed result follows from the above two expansions.

DOI:10.6814/NCCU202201049

(ii) For $d \geq 4$, the details of the proof are the same as above, with the main difference that the expansion now becomes

$$\mathrm{OTC}^{[d]}_{n,n-1-k} = \left(\frac{d^2 d!}{2d^d}\right)^k \frac{1}{k!(k+1)!} \cdot n^{(3-d)k} \cdot \frac{n(dn-d)!}{d!^{n-1}} \left(1 + \mathcal{O}\left(\frac{1+k^2}{n}\right)\right)$$

uniformly for k with $k = o(\sqrt{n})$. This expansion, for $d \geq 4$, contains the (non-trivial decreasing) factor $n^{(3-d)k}$ which is responsible for $\mathrm{OTC}^{[d]}_n$ being now asymptotically dominated by $\mathrm{OTC}^{[d]}_{n,n-1}$ and the limiting distribution of $n - 1 - R^{[d]}_n$ being degenerate. This proves the claimed result also for $d \geq 4$. ∎

The above proof also contains the following first-order asymptotic result for $\mathrm{OTC}^{[d]}_n$ with $d \geq 3$. (See Corollary 3.1.5 for $d = 2$.)

Corollary 4.3.2. *(i) For $d = 3$, we have*

$$\mathrm{OTC}^{[3]}_n \sim I_1(2) \cdot \mathrm{OTC}^{[3]}_{n,n-1} \sim \frac{I_1(2)\sqrt{3}}{9\pi}(n!)^3 \left(\frac{9}{2}\right)^n n^{-3}.$$

(ii) For $d \geq 4$, we have

$$\mathrm{OTC}^{[d]}_n \sim \mathrm{OTC}^{[d]}_{n,n-1} \sim \frac{d!}{d^{d-1/2}(2\pi)^{(d-1)/2}}(n!)^d \left(\frac{d^d}{d!}\right)^n n^{3(1-d)/2}.$$

Proof. For $d = 3$, the first claim is (4.3) and the second follows from this by Stirling's formula. The proof for $d \geq 4$ is similar. ∎

4.3.2 Results for $\mathrm{TC}^{[d]}_{n,k}$

In Section 3.2, we reviewed the proof of the Θ-result for TC_n from [17]. A similar result also holds for $d \geq 3$ since the lemmas and propositions in Section 3.2 can all be extended to $d \geq 3$. For instance, Proposition 3.2.5 was already extended in Section 4.2 and Lemma 3.2.1 holds (without differences) for $d \geq 3$ as well. (We explicitly state this result because we will need it below.)

Lemma 4.3.3. *For any $1 \leq k \leq n-1$,*

$$\mathrm{TC}^{[d]}_{n,n-1-k} \leq \frac{1}{2k}\mathrm{TC}^{[d]}_{n,n-k}.$$

Proof. The proof is the same as that of Lemma 3.2.1 with the only difference that now $d - 1$ nodes are inserted in the root edge and connected to the newly created reticulation node. ∎

DOI:10.6814/NCCU202201049

The Θ-result for $\mathrm{TC}_{n,k}^{[d]}$ was stated as one of the results in [7].

Theorem 4.3.4. *We have,*

$$\mathrm{TC}_n^{[d]} = \Theta\left((n!)^d\,\gamma(d)^n\,e^{3a_1\beta(d)n^{1/3}}n^{\alpha(d)}\right).$$

where $a_1 = -2.33810741\ldots$ *is the largest root of the Airy function of the first kind and*

$$\alpha(d) = -\frac{d(3d-1)}{2(d+1)}, \qquad \beta(d) = \left(\frac{d-1}{d+1}\right)^{2/3}, \qquad \gamma(d) = 4\frac{(d+1)^{d-1}}{(d-1)!}.$$

As for the limit result for the number of reticulation nodes, the result is different from the one for $d = 2$; see Theorem 3.2.6.

Theorem 4.3.5. *For* $d \geq 3$*, let* $I_n^{[d]}$ *be the number of reticulation nodes of a network picked uniformly at random from the set* $\mathcal{TC}_n^{[d]}$*. Then, the limit law of* $n-1-I_n^{[d]}$ *is degenerate, i.e,*

$$n - 1 - I_n^{[d]} \xrightarrow{w} 0.$$

Proof. We restrict to the case $d = 3$; the case of larger d follows along similar lines.

We first explain a construction of $\mathcal{TC}_{n,n-1}^{[3]}$ from $\mathcal{TC}_{n,n-1}$: for a network from $\mathcal{TC}_{n,n-1}$, we add a new parent (and corresponding edge) for each reticulation node to an edge on the path-components we pass before we read the first parent of the recitulation node in the construction of the word from the proof of Theorem 3.2.5. Depending on the choice of the edges, we get several networks in $\mathcal{TC}_{n,n-1}^{[3]}$ which all have the same path-component indices and corresponding labels with the network of $\mathcal{TC}_{n,n-1}$ (with respect to the encoding from Section 4.2). Conversely, removing the first parent (and corresponding edge) of each reticulation node of a network in $\mathcal{TC}_{n,n-1}^{[3]}$ gives a network in $\mathcal{TC}_{n,n-1}$.

Equivalently, when viewing networks as words, any word in $\mathcal{W}_{n-1,n-1}^{[3]}$ can be obtained by adding a new ω_i at any positon before the first occurrence of ω_i in a word from $\mathcal{W}_{n-1,n-1}$, where ω_i runs through all letters.

For example: $\{baa1ab2b3\}$ (a word in $\mathcal{W}_{2,2}$ with corresponding permutation) leads to $\{bbaaa1ab2b3, babaa1ab2b3, abaa1ab2b3\}$ (words in $\mathcal{W}_{2,2}^{[3]}$ with corresponding permutations).

Next, construct $\mathcal{TC}_{n,n-2}^{[3]}$ from $\mathcal{TC}_{n,n-2}$ in a similar way. In particular, note that according to what was discussed in Section 4.2, every network in $\mathcal{TC}_{n,n-2}$ corresponds to two words from $\mathcal{W}_{n-1,n-2}$ and using for words of this set the same construction as above (but just applying it to

DOI:10.6814/NCCU202201049

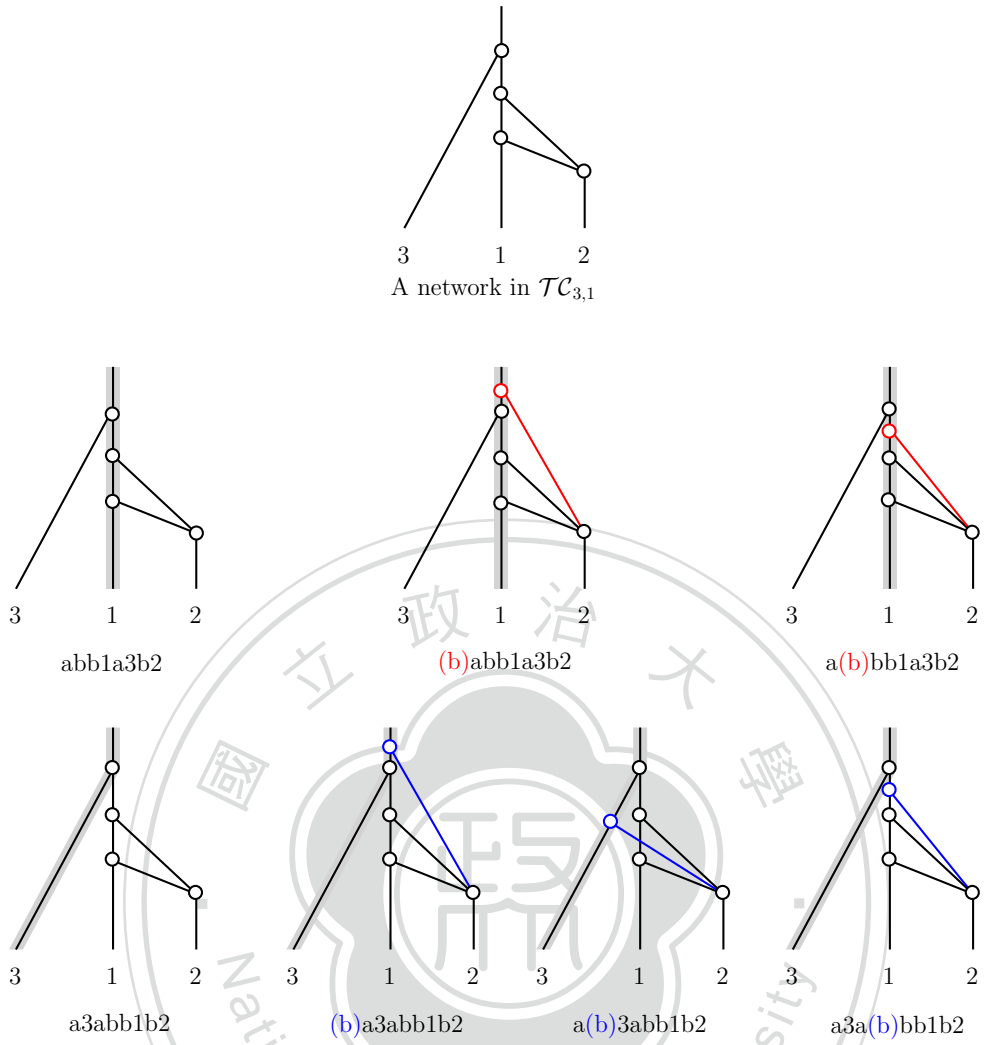


Figure 4.2: Construction of the corresponding networks in $\mathcal{TC}_{3,1}^{[3]}$ from a network in $\mathcal{TC}_{3,1}$.

ω_i which are repeated 3 times), we can obtain all words from $\mathcal{W}_{n-1,n-2}^{[3]}$. (The permutation is irrelevant because it is not changed under the construction.)

For example, $\{abb1a3b2\}$ and $\{a3abb1b2\}$ encode the same network in $\mathcal{TC}_{3,1}$; $\{abb1a3b2\}$ leads to the words $\{babb1a3b2, abbb1a3b2\}$ from $\mathcal{W}_{2,1}^{[3]}$ and $\{a3abb1b2\}$ leads to the words $\{ba3abbb1b2, ab3abb1b2, a3abbb1b2\}$ from $\mathcal{W}_{2,1}^{[3]}$; see Figure 4.2 for a plot of the corresponding networks from $\mathcal{TC}_{3,1}^{[3]}$.

Next, in the proof of Lemma 3.2.1, it was mentioned that $\mathrm{TC}_{n,n-1} = 2\mathrm{TC}_{n,n-2}$ since a network in $\mathcal{TC}_{n,n-2}$ corresponds to exactly two networks in $\mathcal{TC}_{n,n-1}$. Note that these networks can be constructed from the two words in $\mathcal{W}_{n-1,n-2}$ that represent the network in $\mathcal{TC}_{n,n-2}$ by adding the letter which occurs only twice at the beginning of the word (or conversely, by

DOI:10.6814/NCCU202201049

removing the first letter). Thus, to show that $\mathrm{TC}^{[3]}_{n,n-2} = o(\mathrm{TC}^{[3]}_{n,n-1})$, it suffices to show that the number of words in $\mathcal{W}^{[3]}_{n-1,n-2}$ generated by a word ω in $\mathcal{W}_{n-1,n-2}$ is a small o of the number of words in $\mathcal{W}^{[3]}_{n-1,n-1}$ generated by the word ω' in $\mathcal{W}_{n-1,n-1}$, where ω' is constructed from ω as above. In order to prove the latter, we denote by $s(\omega)$ the number of words which are generated from ω and by $r(\omega')$ the number of words which are generated from ω'. Moreover, we drop the permutation since, as mentioned above, it is irrelevant for our argument. For example, if $\omega = abbccabc$, then $\omega' = aabbccabc$ and

$$\frac{r(aabbccabc)}{s(abbccabc)} = \frac{1 \cdot 4 \cdot 7}{2 \cdot 5}.$$

More generally,

$$\frac{r(\omega')}{s(\omega)} = \frac{k_1 \cdot k_2 \cdots k_n}{(k_2 - 2) \cdots (k_n - 2)},$$

where i denotes the i-th first parent of the letters in ω' and k_i indicates the number of possibilities of adding an additional parent for i by the above method. Note that $k_1 = 1$ and the k_i's increase, thus, the $\frac{k_i}{k_i - 2}$'s decrease. Moreover, note that for each k_i, we have $k_i \leq 4(i-1)+1$ since the upper bound is the extremal case that each of the previous $i-1$ letters occur 4 times.

Consequently,

$$\frac{r(\omega')}{s(\omega)} = \frac{k_2 \cdots k_n}{(k_2 - 2) \cdots (k_n - 2)} \geq \frac{5 \cdot 9 \cdots (4n-3)}{3 \cdot 7 \ldots (4n-5)} = \Theta\left(\frac{\Gamma(n+\frac{1}{4})}{\Gamma(n-\frac{1}{4})}\right) = \Theta(n^{1/2}),$$

where we used Stirling's formula for the gamma function in the last step. This implies that, $s(\omega) = o(r(\omega'))$ which (as explained above) in turn implies that $\mathrm{TC}^{[3]}_{n,n-2} = o(\mathrm{TC}^{[3]}_{n,n-1})$. Next, from Lemma 4.3.3,

$$\mathrm{TC}^{[3]}_{n,n-1-k} = o(\mathrm{TC}^{[3]}_{n,n-1})$$

for all fixed $k \geq 1$. Also, again by Lemma 4.3.3,

$$\sum_{k=2}^{n-1} \mathrm{TC}^{[3]}_{n,n-1-k} = \mathcal{O}(\mathrm{TC}^{[3]}_{n,n-2}) = o(\mathrm{TC}^{[3]}_{n,n-1}).$$

Thus,

$$\mathrm{TC}^{[3]}_{n} = \mathrm{TC}^{[3]}_{n,n-1} + \mathrm{TC}^{[3]}_{n,n-2} + \sum_{k=2}^{n-1} \mathrm{TC}^{[3]}_{n,n-1-k} \sim \mathrm{TC}^{[3]}_{n,n-1}.$$

Consequently,

$$P(n-1-I^{[3]}_n = k) = P(I^{[3]}_n = n-1-k) = \frac{\mathrm{TC}^{[3]}_{n,n-1-k}}{\mathrm{TC}^{[3]}_{n}} \longrightarrow \begin{cases} 1, & \text{if } k = 0; \\ 0, & \text{if } k \geq 1 \end{cases}$$

which proves the claim. ∎

DOI:10.6814/NCCU202201049

We obtain the following corollary (which once again shows the difference with $d = 2$; see Corollary 3.2.7).

Corollary 4.3.6. *For $d \geq 3$, We have*

$$\mathrm{TC}_n^{[d]} \sim \mathrm{TC}_{n,n-1}^{[d]}.$$

4.4 Some open problems

Even though we can count tree-child networks exactly (see Section 4.2) and have found the limit law for the number of reticulation nodes of a random tree-child network (see Section 4.3.2), there are still lots of problems left.

In particular, limit law results for shape parameters of random tree-child networks (both bi-combining and d-combining) are still little explored, e.g., distributional results of reticulation nodes followed by a leaf in $\mathcal{TC}_n^{[d]}$, height of a network in $\mathcal{TC}_n^{[d]}$, Sackin index of a network in $\mathcal{TC}_n^{[d]}$ (actually we already have a results for this in the one-component case which will be included in the journal version of [7]; see also [25] for results for $d = 2$) and distributional results of the number of patterns in $\mathcal{TC}_n^{[d]}$. All this will be left for future work.

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202201049

Chapter 5

Ranked tree-child networks

We have given the definition of ranked tree-child networks (RTCNs) in Section 1.2. In this chapter, we will review some of the results from [1]. In particular, we will focus on limit distribution results for the number of occurrences of a pattern which we will improve and generalize. Most of the material in this section is contained in [16].

5.1 Previous results

First, we recall the result for the number of RTCNs from [1].

Theorem 5.1.1. *The number* $\mathrm{RTCN}_{n,k}$ *of RTCNs with* n *leaves and* k *reticulation events satisfies*

$$\mathrm{RTCN}_{n,k} = \begin{bmatrix} n-1 \\ n-1-k \end{bmatrix} \frac{n!(n-1)!}{2^{n-1}},$$

where $\left[\begin{smallmatrix} n-1 \\ n-1-k \end{smallmatrix}\right]$ *denotes the unsigned Stirling numbers of first kind.*

Note that $\left[\begin{smallmatrix} n-1 \\ n-1-k \end{smallmatrix}\right]$ counts the number of permutations of length $n-1$ with exactly $n-1-k$ cycles and $\frac{n!(n-1)!}{2^{n-1}}$ is the number of ranked trees. Indeed, there is a bijection between RTCNs with n leaves and k reticulation nodes and these objects; see [3].

Remark 5.1.2. It was proved in [1] that almost no tree-child network is rankable. Also, it is not clear how many different rankings a rankable network admits.

Next, we introduce a recursive construction (called *forward construction* in the sequel) which allow us to study distributional properties of patterns in RTCNs.

DOI:10.6814/NCCU202201049

1. Start with a branching event;

2. In the $\ell - 1$-st step pick uniformly at random a tuple (ℓ_1, ℓ_2) of lineages;

3. If $\ell_1 = \ell_2$ attach a branching event to lineage ℓ_1;

4. If $\ell_1 \neq \ell_2$ attach a reticulation event to the lineages ℓ_1, ℓ_2;

5. Stop once n lineages are obtained.

The above construction shows that when going from n lineages to $n+1$ lineages, a branching event will be created with probability $1/n$ and a reticulation event is created with probability $(n-1)/n$.

Note that the leaves of a network resulting from the above construction are not labeled and incoming edges of reticulation events are ordered. So the resulting network is not a RTCN. But the following lemma makes it possible to study the number of occurrences of a pattern in RTCNs by the corresponding number of occurrences in a network constructed as above.

Let X_n denote number of occurrence of a pattern in a network which is constructed by the procedure above.

Lemma 5.1.3. *X_n has the same distribution as the number of occurrences of the pattern in a RTCN with n leaves which is picked uniformly at random.*

Remark 5.1.4. In the sequel, we will call a RTCN with n leaves which is picked uniformly at random from all RTCNs with n leaves a *random RTCN*.

5.2 Patterns of height 1

Consider a random RTCN with n leaves. We will describe the number of occurrences of a pattern via a Markov chain (using the procedure discussed in the last section).

First, we will consider the two patterns of height 1, where the *height* represents the number of steps in the evolution process from the definition of a RTCN.

Definition 5.2.1. *A cherry is a branching event whose outgoing lineages are leaves (or alternatively, it is a tree node with both children leaves).*

A trident is a reticulation event with all three outgoing lineages leaves.

DOI:10.6814/NCCU202201049

Let C_n and T_n denote the number of cherries and tridents in a random RTCN with n leaves. The following result was proved in [1].

Theorem 5.2.2. *For cherries,*

$$C_n \xrightarrow{w} \text{Poisson}(1/4), \quad \textit{as} \quad n \to \infty$$

and for tridents,

$$\frac{T_n}{n} \xrightarrow{\mathbb{P}} \frac{1}{7}, \quad \textit{as} \quad n \to \infty.$$

The result for T_n is just a weak law of large numbers. We can improve it to a central limit law.

Theorem 5.2.3. *For the number T_n of tridents in a random ranked tree-child network with n leaves, we have*

$$\frac{T_n - n/7}{\sqrt{24n/637}} \xrightarrow{w} N(0,1), \quad \textit{as} \quad n \to \infty. \tag{5.1}$$

Note that cherry and tridents are all the patterns of height 1. Thus, the above two results clarify the distributional behavior of all patterns of height 1.

The method of moments (see Section 2.2) and induction will be the main tools we use for proving Theorem 5.2.3. But first, we recall the following Markov Chain about the variation of the number of tridents when performing one more step in the forward construction.

Lemma 5.2.4. *Define a Markov process by $T_2 = 0$ and*

$$(T_{n+1}|T_n = j) = \begin{cases} j-1, & \textit{with probability } \dfrac{3j(3j-2)}{n^2}; \\ j, & \textit{with probability } 1 - \dfrac{3j(3j-2)+(n-3j)(n-3j-1)}{n^2}; \\ j+1, & \textit{with probability } \dfrac{(n-3j)(n-3j-1)}{n^2}. \end{cases} \tag{5.2}$$

Then, T_n has the same distribution as the number of tridents in a random ranked tree-child network with n leaves.

Using this lemma, it is not hard to compute the mean of T_n.

Lemma 5.2.5. *Let $\mu_n := \mathbb{E}(T_n)$. Then, μ_n satisfies the recurrence*

$$\mu_{n+1} = \left(1 - \frac{3}{n}\right)^2 \mu_n + 1 - \frac{1}{n}$$

DOI:10.6814/NCCU202201049

whose solution is given by

$$\mu_n = \frac{(15n^3 - 85n^2 + 144n - 71)n}{105(n-1)(n-2)(n-3)}, \qquad (n \geq 4). \tag{5.3}$$

Note that this result implies that $\mu_n \sim n/7$ as $n \to \infty$.

Next, we define

$$\phi_{n,m} := \mathbb{E}(T_n - \mu_n)^m.$$

We focus on the central moments $\phi_{n,m}$ instead of $\mathbb{E}(T_n^m)$ since a random variable X_n with normal distribution satisfies

$$\mathbb{E}\left((X_n - \mathbb{E}X_n)^m\right) \sim \mathbb{E}\left(N(0,1)^m\right)\left(\mathrm{Var}(X_n)\right)^{m/2} n^{m/2}.$$

Moreover, $\phi_{n,m}$ is actually easier to compute since it already incorporates the cancellations which arise from $\mathbb{E}(T_n^m) \sim (\mathbb{E}(T_n))^m$.

The sequence $\phi_{n,m}$ satisfies the following recurrence.

Lemma 5.2.6. *For $m \geq 2$, we have*

$$\phi_{n+1,m} = \left(1 - \frac{3m}{n}\right)^2 \phi_{n,m} + \psi_{n,m}$$

with

$$\psi_{n,m} = \sum_{j=0}^{m-1} \phi_{n,j} \Lambda_j(n),$$

where $\Lambda_j(n)$ admits the complete asymptotic expansion

$$\Lambda_j(n) \sim \sum_{\ell=0}^{\infty} \frac{\lambda_{j,\ell}}{n^\ell}, \quad \textit{as} \quad n \to \infty$$

with $\lambda_{j,\ell} \in \mathbb{R}$; in particular, $\lambda_{m-1,0} = 0$ and

$$\lambda_{m-2,0} = \binom{m}{2} \times \frac{24}{49}.$$

Proof. Set $\bar{T}_n := T_n - \mu_n$. From (5.2), we have

$$\begin{aligned}
\phi_{n+1,m} = {} & \mathbb{E}(\bar{T}_n + \mu_n - \mu_{n+1} - 1)^m \frac{3(\bar{T}_n + \mu_n)(3(\bar{T}_n + \mu_n) - 2)}{n^2} \\
& + \mathbb{E}(\bar{T}_n + \mu_n - \mu_{n+1})^m \\
& \quad \times \frac{n^2 - 3(\bar{T}_n + \mu_n)(3(\bar{T}_n + \mu_n) - 2 + (n - 3(\bar{T}_n + \mu_n))(n - 3(\bar{T}_n + \mu_n) - 1)}{n^2} \\
& + \mathbb{E}(\bar{T}_n + \mu_n - \mu_{n+1} + 1)^m \frac{(n - 3(\bar{T}_n + \mu_n))(n - 3(\bar{T}_n + \mu_n) - 1)}{n^2}.
\end{aligned}$$

DOI:10.6814/NCCU202201049

From this, by expanding what is inside the means by the binomial theorem, we see that

$$\psi_{n,m} = \sum_{j=0}^{m+2} \phi_{n,j} \Lambda_j(n).$$

However, $\Lambda_{m+2}(n) = 0$ (because the probabilities in (5.2) sum up to 1). Straightforward computation (best done with a computer algebra system such as Maple) shows that $\Lambda_{m+1}(n) = \Lambda_m(n) = 0$. Next, since (5.3) implies that $\mu_n - \mu_{n+1} \sim \sum_{\ell \geq 0} \lambda_\ell / n^\ell$, we have

$$\Lambda_j(n) \sim \sum_{\ell=0}^{\infty} \frac{\lambda_{j,\ell}}{n^\ell}, \quad \text{as} \quad n \to \infty$$

and by some more computations (again best done with Maple), we obtain that $\lambda_{m-1,0} = 0$ and that $\lambda_{m-2,0}$ is as claimed. ∎

Note that the above recurrence has the (general) form

$$\phi_{n+1} = \left(1 - \frac{\kappa}{n}\right)^2 \phi_n + \psi_n, \qquad (n \geq \kappa + 1) \tag{5.4}$$

with a suitable initial value $\phi_{\kappa+1}$, where $\kappa \in \mathbb{N}$ and $\{\psi_n\}_{n \geq \kappa+1}$ is a given sequence. We need a general result for such a sequence.

Lemma 5.2.7. *Assume that ϕ_n satisfies (5.4). If $\psi_n \sim cn^\alpha$ with $\alpha > -2\kappa - 1$ a real number, then*

$$\phi_n \sim \frac{c}{2\kappa + \alpha + 1} n^{\alpha+1}, \quad \textit{as} \quad n \to \infty.$$

Remark 5.2.8. Throughout the rest of the thesis, the notation $f(n) \sim cg(n)$ (with $g(n) > 0$) means that $f(n) = cg(n) + o(g(n))$. (Note that this is the usual meaning if $c \neq 0$; however, if $c = 0$, then $f(n) = o(g(n).)$

Proof. Iterating (5.4) gives the solution

$$\phi_n = \binom{n-1}{\kappa}^{-2} \left(\psi_{\kappa+1} + \sum_{\ell=\kappa+1}^{n-1} \binom{\ell}{\kappa}^2 \psi_\ell \right). \tag{5.5}$$

Note that

$$\binom{\ell}{\kappa}^2 \sim \frac{\ell^{2\kappa}}{\kappa!^2}, \quad \text{as} \quad \ell \to \infty. \tag{5.6}$$

Thus, if $\psi_n \sim cn^\alpha$, then

$$\psi_{\kappa+1} + \sum_{\ell=\kappa+1}^{n-1} \binom{\ell}{\kappa}^2 \psi_\ell \sim \frac{c}{\kappa!^2} \sum_{\ell=\kappa+1}^{n-1} \ell^{2\kappa+\alpha} \sim \frac{c}{\kappa!^2 (2\kappa + \alpha + 1)} n^{2\kappa+\alpha+1}.$$

DOI:10.6814/NCCU202201049

Plugging this into (5.5) and using once more (5.6) gives the claimed result. ∎

Applying the last result to the recurrence for the central moments (Lemma 5.2.6) and using induction gives the following asymptotic result for all central moments of T_n. (This method is called *moment pumping*.)

Proposition 5.2.9. *As $n \to \infty$, the central moments of T_n satisfy*

$$\mathbb{E}\left((T_n - \mu_n)^m\right) \sim g_m \left(\frac{24}{637}\right)^{m/2} n^{m/2}.$$

Here, $g_m = \mathbb{E}(N(0,1)^m)$ (see (2.5)).

Proof. The claim (trivially) holds for $m = 0$ and $m = 1$. Next, we assume that the claim holds for all $m' < m$. In order to show it for m, note that from the induction hypothesis, we have for $\psi_{n,m}$ from Lemma 5.2.6:

$$\psi_{n,m} \sim \binom{m}{2} \times \frac{24}{49} \times g_{m-2} \left(\frac{24}{637}\right)^{m/2-1} n^{m/2-1}.$$

Applying Lemma 5.2.7, we obtain that

$$\phi_{n,m} \sim \binom{m}{2} \times \frac{24}{49} \times \frac{g_{m-2}}{6m + m/2} \left(\frac{24}{637}\right)^{m/2-1} n^{m/2} \sim (m-1) g_{m-2} \left(\frac{24}{637}\right)^{m/2} n^{m/2}$$

from which the claimed result follows since $(m-1)g_{m-2} = g_m$. ∎

Theorem 5.2.3 follows now from the last proposition by the Proposition 2.2.1 and Proposition 2.2.2. (See also the discussions after these propositions.)

5.3 Patterns of height 2

In this section, we study the behavior of patterns of height 2. There are a total of 9 such patterns (see Figure 5.1) and we can divide them into 3 groups according to their order of the expectation and limit distribution.

Theorem 5.3.1. *Denote by X_n the number of occurrences of a (fixed) pattern of height 2 in a random ranked tree-child networks with n leaves. Then, we have the following limit law results.*

(A) *For the patterns in Figure 5.1-(a), we have that the limit law of X_n is degenerate. More precisely,*

$$X_n \xrightarrow{L_1} 0, \quad \text{as} \quad n \to \infty.$$

DOI:10.6814/NCCU202201049

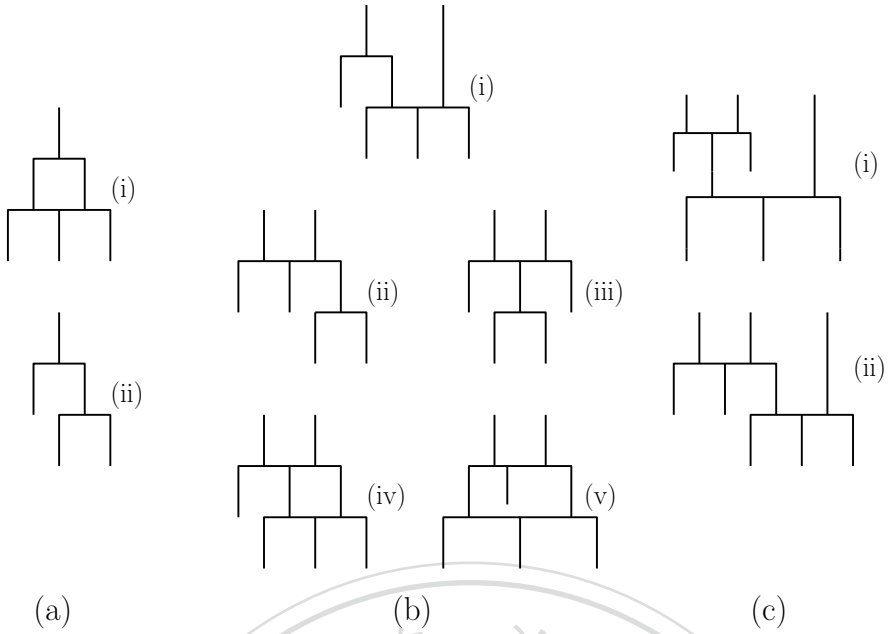


Figure 5.1: All patterns of height 2. In a RTCN with a large number of leaves, the number of occurrence of these patterns is as follows: (a) do not occur (mean tends to 0); (b) occur only sporadically (mean is constant); (c) occur frequently (mean is linear).

(B) *For the patterns in Figure 5.1-(b), we have*

$$X_n \xrightarrow{w} \text{Poisson}(\lambda), \quad \textit{as} \quad n \to \infty,$$

where

	(b-i)	*(b-ii)*	*(b-iii)*	*(b-iv)*	*(b-v)*
λ	$1/8$	$1/28$	$1/56$	$1/14$	$1/28$

(C) *For the patterns in Figure 5.1-(c), we have*

$$\frac{X_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{w} N(0,1), \quad \textit{as} \quad n \to \infty,$$

where $(\mu, \sigma^2) = (4/77, 4575916/137582445)$ *and* $(\mu, \sigma^2) = (2/77, 2930764/137582445)$ *for the patterns from Figure 5.1-(c-i) and Figure 5.1-(c-ii), respectively.*

Note that from the last section, we have the following results:

$$\mathbb{E}(T_n^m) \sim \mu_n^m \sim \frac{n^m}{7^m}, \quad \text{as} \quad n \to \infty, \tag{5.7}$$

DOI:10.6814/NCCU202201049

and

$$\mathbb{E}\left(\frac{T_n - n/7}{\sqrt{24n/637}}\right)^m \sim \mathbb{E}(N(0,1))^m, \quad \text{as} \quad n \to \infty, \tag{5.8}$$

and

$$\mathbb{E}(C_n^{\underline{m}}) := \mathbb{E}(C_n(C_n-1)\cdots(C_n-m+1)) \sim \frac{1}{4^m}, \quad \text{as} \quad n \to \infty. \tag{5.9}$$

The proof of Theorem 5.3.1 will proceed along similar lines as the proof of Theorem 5.2.3 and the results above will play an important role in the proof.

In our paper [15], we discussed several cases in detail and in this thesis, we will discuss the remaining cases as a supplement.

Degenerate Limit Laws. We will consider the pattern in Figure 5.1-(a-i) first. Our method will be more general than needed since this generality will be used in proof of the other cases.

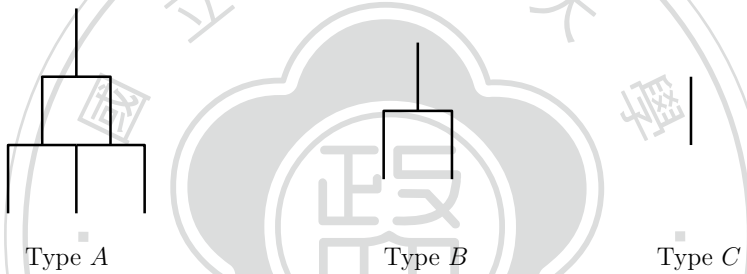


Figure 5.2: The pattern from Figure 5.1-(a-i) (Type A); A cherry (Type B); The remaining external lineages (Type C).

(a-i) Call the pattern in Figure 5.1-(a-i) a pattern of type A and a cherry a pattern of type B in the sequel. Note that each external lineage (which are lineages at the bottom of a RTCN) belongs either exactly to a pattern of type A or a pattern of type B or to neither of these patterns (such a lineage will be called a pattern of type C in the sequel). See Figure 5.2. We list what will happen with the number of patterns of type A and type B when we add a branching or reticulation event so that n lineages become $n+1$ lineages.

First, if a branching event is added, then we have the cases from Table 5.1.

For instance, if the branching event is attached to any of the three external lineages from a pattern of type A, then one patterns of type B is created whereas one pattern of type A was destroyed; this is the first row in Table 5.1 and the probability that this happens is given by the

DOI:10.6814/NCCU202201049

	type A	type B	probability
type A	-1	$+1$	$3a/n^2$
type B	0	0	$2b/n^2$
type C	0	$+1$	$(n-3a-2b)/n^2$

Table 5.1: The change of the number of patterns of type A and type B (first and second column) if a branching event is attached to an external lineage belonging to a pattern of type A, type B or type C (rows). Here, a and b denote the number of patterns of type A and B, respectively. The probability for each cases is given in the third column if one starts with a ranked tree-child network with n leaves (and thus the probability of picking a fixed external lineage is $1/n^2$).

number of possible choices of such a lineage (3a) divided by the number of choices of a pair of two external lineages (n^2). Likewise, the second row is the case where the branching event is attached to a reticulation event, thus the number of pattern of type A and type B do not change. Similarly, the remaining row is explained.

Next, if a reticulation event is added, then we have the cases listed in Table 5.2.

For instance, the sixth row of this column is explained as follows: if the reticulation event is attached to an external lineage belonging to a pattern of type A and an external lineage from a pattern of type B, then a pattern of type A and type B is destroyed (and no new pattern of type A or type B is created). The probability that this will happen is given by the number of choices of the external lineages divided by n^2; the number of choices equals $3a \cdot 2b \cdot 2$ since once the patterns of type A and B are chosen (ab choices) there are 3 choices for the external lineages in the pattern of type A and 2 in the pattern of type B; moreover, the factor 2 comes from symmetry. Similarly, the other rows are explained.

Table 5.1 and Table 5.2 give the transition probabilities of a Markov Chain for the variation of the number of patterns of type A and type B in a RTCN with n leaves. From this Markov chain, by a straightforward computation, we obtain the following.

Lemma 5.3.2. *Let X_n be the number of occurrences of the pattern from Figure 5.1-(a-i) in a random ranked tree-child network with n leaves. Then,*

$$\mathbb{E}(X_{n+1}) = \left(1 - \frac{3}{n}\right)^2 \mathbb{E}(X_n) + \frac{2\mathbb{E}(C_n)}{n^2}, \tag{5.10}$$

DOI:10.6814/NCCU202201049

	type A	type B	probability
A	-1	0	$6a/n^2$
A & A	-2	0	$9a(a-1)/n^2$
B	$+1$	-1	$2b/n^2$
B & B	0	-2	$4b(b-1)/n^2$
C & C	0	0	$(n-3a-2b)(n-3a-2b-1)/n^2$
A & B	-1	-1	$12ab/n^2$
A & C	-1	0	$6a(n-3a-2b)/n^2$
B & C	0	-1	$4b(n-3a-2b)/n^2$

Table 5.2: Columns: The change of the number of patterns of type A (first column) and type B (second column); the probabilities of these changes are contained in the last column. Rows: a single letter means that both external lineages are picked from a pattern of that type; two letters mean that the external lineages are chosen from (different) patterns of these two types.

where C_n is the number of cherries.

Using this, we can proof one case in Theorem 5.3.1.

Proof of Theorem 5.3.1-(a) for the pattern in Figure 5.1-(a-i). The recurrence (5.10) has the form (5.4) with

$$\psi_n = \frac{2\mathbb{E}(C_n)}{n^2} \sim \frac{1}{2n^2},$$

where the last step follows from (5.9). Applying Lemma 5.2.7 gives

$$\mathbb{E}(X_n) \sim \frac{1}{10n}$$

which implies the claimed result. ∎

(a-ii) Here, we consider the pattern in Figure 5.1-(a-ii). The treatment of this pattern is slightly different from the pattern (a-i). Note that patterns of type A are still the pattern in Figure 5.1-(a-i), but patterns of type B are cherries which are not contained in patterns of type A. Since this case has been explained in [15], only the tables for the transition probabilities if a branching or reticulation event is added are given here.

DOI:10.6814/NCCU202201049

	type A	type B	probability
type A	-1	$+2$	a/n^2
	0	0	$2a/n^2$
type B	$+1$	-1	$2b/n^2$
type C	0	$+1$	$(n-3a-2b)/n^2$

Table 5.3: The change of the number of patterns of type A and type B for (a-ii) if a branching event is added.

	type A	type B	probability
A	-1	0	$6a/n^2$
A & A	-2	0	$4a(a-1)/n^2$
	-2	$+1$	$4a(a-1)/n^2$
	-2	$+2$	$a(a-1)/n^2$
B	0	-1	$2b/n^2$
B & B	0	-2	$4b(b-1)/n^2$
C & C	0	0	$(n-3a-2b)(n-3a-2b-1)/n^2$
A & B	-1	-1	$8ab/n^2$
	-1	0	$4ab/n^2$
A & C	-1	0	$4a(n-3a-2b)/n^2$
	-1	$+1$	$2a(n-3a-2b)/n^2$
B & C	0	-1	$4b(n-3a-2b)/n^2$

Table 5.4: The change of the number of patterns of type A and type B of (a-ii) if a reticulation event is added.

From these tables, one can find that the number of occurrences of the pattern from (a-ii) shares the same recurrence as the pattern from (a-i). Thus, they have the same result.

Poisson Limit Laws. The pattern in Figure 5.1-(b-i) is special compared to the other patterns in Figure 5.1-(b); thus, we will discuss it later. On the other hand, the proof of remaining four patterns is the same. Consequently, we will only give details for (b-ii).

DOI:10.6814/NCCU202201049

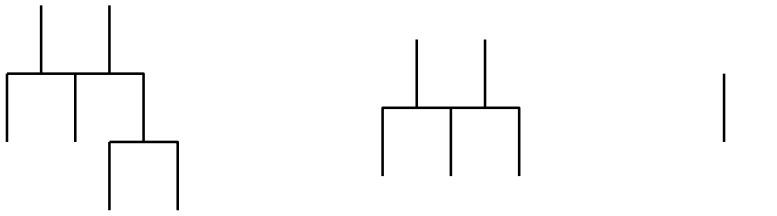


Figure 5.3: The pattern from Figure 5.1-(b-ii) (Type A); A trident (Type B); The remaining external lineages (Type C).

Similarly as above, we define patterns of type A, type B and type C: the pattern (b-ii) is of type A, a trident is a pattern of type B and patterns of type C are similarly defined as for (a-i). Then, we again can find the transition probabilities of the corresponding Markov chain; see Table 5.5 when a branching event is added and Table 5.6 when a reticulation event is added.

Then, we have the following results for mixed moments.

	type A	type B	probability
A	-1	0	$4a/n^2$
B	$+1$	-1	$2b/n^2$
	0	-1	b/n^2
C	0	0	$(n-4a-3b)/n^2$

Table 5.5: The change of the number of patterns of type A and type B for (b-ii) (see Figure 5.3) when the next event is a branching event.

Lemma 5.3.3. *Let X_n be the number of occurrences of the pattern from Figure 5.1-(b-ii) in a random ranked tree-child network with n leaves. Then, for all $r, s \geq 0$, we have*

$$\mathbb{E}(X_{n+1}^r T_{n+1}^s) = \left(1 - \frac{4r+3s}{n}\right)^2 \mathbb{E}(X_n^r T_n^s) + \frac{2r}{n^2}\mathbb{E}(X_n^{r-1} T_n^{s+1}) - \frac{2rs}{n^2}\mathbb{E}(X_n^{r-1} T_n^s) + R_n,$$

where

$$R_n = \sum_{j=0}^{s-1} \left(\lambda_{j,0} + \frac{\lambda_{j,1}}{n} + \frac{\lambda_{j,2}}{n^2}\right) \mathbb{E}(X_n^r T_n^j).$$

Here, $\lambda_{j,i}$ are constants which only depend on r and s with $\lambda_{s-1,0} = s$.

DOI:10.6814/NCCU202201049

	type A	type B	probability
A	-1	$+1$	$12a/n^2$
A & A	-2	$+1$	$16a(a-1)/n^2$
B	0	0	$6b/n^2$
B & B	0	-1	$9b(b-1)/n^2$
C & C	0	$+1$	$(n-4a-3b)(n-4a-3b-1)/n^2$
A & B	-1	0	$24ab/n^2$
A & C	-1	$+1$	$8a(n-4a-3b)/n^2$
B & C	0	0	$6b(n-4a-3b)/n^2$

Table 5.6: The change of the number of patterns of type A and type B for (b-ii) (see Figure 5.3) when the next event is a reticulation event.

Proof. Assume that $X_n = c$ and $T_n = d$. Then, we have c patterns of type A and d patterns of type B.

Using the cases from Table 5.5 and Table 5.6, we can write $X_{n+1}^{\underline{r}} T_{n+1}^{\underline{s}}$ given $X_n = c$ and $T_n = d$ as a sum of terms $(X_n + k)^{\underline{r}}(T_n + \ell)^{\underline{s}}$ which are multiplied with the probabilities of every case and where $k \in \{-2, -1, 0, 1\}$ is the value from the first column and $\ell \in \{-1, 0, 1\}$ is the value of the second column for every case.

Next, we replace $(X_n + k)^{\underline{r}}$ for $k = -2$ and $k = -1$ by

$$(X_n - 2)^{\underline{r}} = \frac{(X_n - r)(X_n - r - 1)}{X_n(X_n - 1)} X_n^{\underline{r}}, \quad (X_n - 1)^{\underline{r}} = \frac{X_n - r}{X_n} X_n^{\underline{r}}$$

and for $k = 1$ by

$$(X_n + 1)^{\underline{r}} = \frac{X_n + 1}{X_n - r + 1} X_n^{\underline{r}}.$$

Finally, we expand $(T_n + \ell)^{\underline{s}}$ by the binomial theorem and simplify the sum of all terms of the same order in this expansion with Maple. (Note that since $\ell \in \{-1, 0, 1\}$ this only has to be done for the first three terms in the expansion; the other terms are the same only multiplied with different weights). This proves the claimed result. ∎

Proof of Theorem 5.3.1-(b) for the pattern in Figure 5.1-(b-ii). As above, let X_n denote the number of occurrences of the pattern from Figure 5.1-(b-ii) in a random ranked tree-child net-

DOI:10.6814/NCCU202201049

	type A	type B	probability
A	-1	0	$4a/n^2$
B	$+1$	-1	b/n^2
	0	-1	$2b/n^2$
C	0	0	$(n-4a-3b)/n^2$

Table 5.7: The change of the number of patterns of type A and type B for (b-iii) when the next event is a branching event.

	type A	type B	probability
A	-1	$+1$	$12a/n^2$
B	0	0	$6b/n^2$
A & A	-2	$+1$	$16a(a-1)/n^2$
B & B	0	-1	$9b(b-1)/n^2$
C & C	0	$+1$	$(n-4a-3b)(n-4a-3b-1)/n^2$
A & B	-1	0	$24ab/n^2$
A & C	-1	$+1$	$8a(n-4a-3b)/n^2$
B & C	0	0	$6b(n-4a-3b)/n^2$

Table 5.8: The change of the number of patterns of type A and type B for (b-iii) when the next event is a reticulation event.

work with n leaves. We will use induction to show that for all $r, s \geq 0$:

$$\mathbb{E}(X_n^r T_n^s) \sim \frac{n^s}{28^r 7^s}, \quad \text{as} \quad n \to \infty, \tag{5.11}$$

where the induction is with respect to the lexicographic order of (r, s). Note that the base case, namely $(0, s)$ with $s \geq 0$, is implied by (5.7). Next, by the above lemma $\mathbb{E}(X_n^r T_n^s)$ satisfies a recurrence of the form (5.4) with all terms in ψ_n being of a smaller lexicographic order. Thus, by using the induction hypothesis,

$$\psi_n \sim \frac{2r}{n^2}\mathbb{E}(X_n^{r-1}T_n^{s+1}) + \lambda_{s-1,0}\mathbb{E}(X_n^r T_n^{s-1}) \sim \frac{8r+7s}{28^r 7^s} n^{s-1}.$$

The induction claim follows now from this by applying Lemma 5.2.7. ∎

DOI:10.6814/NCCU202201049

	type A	type B	probability
A	-1	0	$3a/n^2$
	-1	$+1$	a/n^2
B	0	-1	$3b/n^2$
C	0	0	$(n-4a-3b)/n^2$

Table 5.9: The change of the number of patterns of type A and type B for (b-iv) when the next event is a branching event.

	type A	type B	probability
A	-1	$+1$	$8a/n^2$
	0	0	$4a/n^2$
A & A	-2	$+1$	$9a(a-1)/n^2$
	-2	$+2$	$6a(a-1)/n^2$
	-2	$+3$	$a(a-1)/n^2$
B	0	0	$2b/n^2$
	$+1$	-1	$4b/n^2$
B & B	0	-1	$9b(b-1)/n^2$
C & C	0	$+1$	$(n-4a-3b)(n-4a-3b-1)/n^2$
A & B	-1	0	$18ab/n^2$
	-1	$+1$	$6ab/n^2$
A & C	-1	$+1$	$6a(n-4a-3b)/n^2$
	-1	$+2$	$2a(n-4a-3b)/n^2$
B & C	0	0	$6b(n-4a-3b)/n^2$

Table 5.10: The change of the number of patterns of type A and type B for (b-iv) when the next event is a reticulation event.

For the remaining three patterns, we will only give their tables and asymptotic recurrence. First, by Table 5.7 and Table 5.8, we have the recurrence for (b-iii):

$$\mathbb{E}(X_{n+1}^r T_{n+1}^s) \sim \left(1-\frac{4r+3s}{n}\right)^2 \mathbb{E}(X_n^r T_n^s) + \frac{r}{n^2}\mathbb{E}(X_n^{r-1} T_n^{s+1}) + s\mathbb{E}(X_n^r T_n^{s-1}).$$

DOI:10.6814/NCCU202201049

	type A	type B	probability
A	-1	0	$3a/n^2$
	-1	$+1$	a/n^2
B	0	-1	$3b/n^2$
C	0	0	$(n-4a-3b)/n^2$

Table 5.11: The change of the number of patterns of type A and type B for (b-v) when the next event is a branching event.

	type A	type B	probability
A	-1	$+1$	$10a/n^2$
	0	0	$2a/n^2$
A & A	-2	$+1$	$9a(a-1)/n^2$
	-2	$+2$	$6a(a-1)/n^2$
	-2	$+3$	$a(a-1)/n^2$
B	0	0	$4b/n^2$
	$+1$	-1	$2b/n^2$
B & B	0	-1	$9b(b-1)/n^2$
C & C	0	$+1$	$(n-4a-3b)(n-4a-3b-1)/n^2$
A & B	-1	0	$18ab/n^2$
	-1	$+1$	$6ab/n^2$
A & C	-1	$+1$	$6a(n-4a-3b)/n^2$
	-1	$+2$	$2a(n-4a-3b)/n^2$
B & C	0	0	$6b(n-4a-3b)/n^2$

Table 5.12: The change of the number of patterns of type A and type B for (b-v) when the next event is a reticulation event.

Thus, $X_n \sim \text{Poisson}(\frac{1}{28})$.

Next, by Table 5.9 and Table 5.10, we have the recurrence for (b-iv):

$$\mathbb{E}(X_{n+1}^r T_{n+1}^s) \sim \left(1 - \frac{4r+3s}{n}\right)^2 \mathbb{E}(X_n^r T_n^s) + \frac{4r}{n^2}\mathbb{E}(X_n^{r-1} T_n^{s+1}) + s\mathbb{E}(X_n^r T_n^{s-1}).$$

DOI:10.6814/NCCU202201049

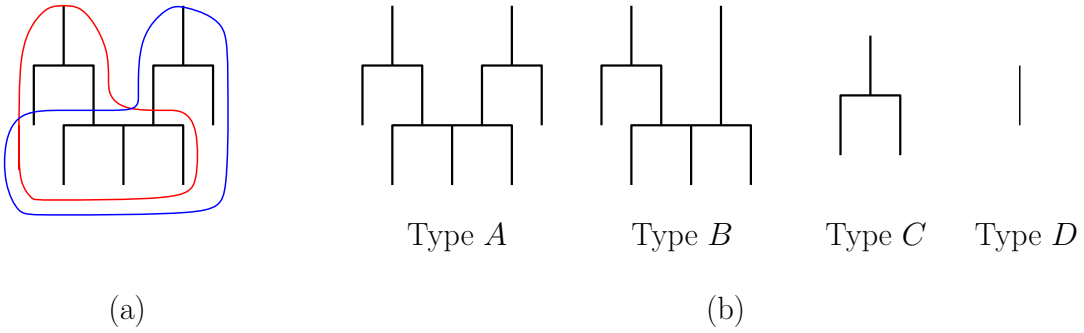


Figure 5.4: (a) The overlapping pattern containing two versions of (b-i). (b) Classification of types of patterns for (b-i): The pattern from (a) (Type A); The pattern (b-i) which is not contained in a pattern of type A (Type B); A cherry (Type C); The remaining external lineages (Type D).

Thus, $X_n \sim \text{Poisson}(\frac{1}{14})$.

Finally, by Table 5.11 and Table 5.12, we have for the recurrence for (b-iv):

$$\mathbb{E}(X_{n+1}^r T_{n+1}^s) \sim \left(1 - \frac{4r+3s}{n}\right)^2 \mathbb{E}(X_n^r T_n^s) + \frac{2r}{n^2}\mathbb{E}(X_n^{r-1} T_n^{s+1}) + s\mathbb{E}(X_n^r T_n^{s-1}).$$

Thus, $X_n \sim \text{Poisson}(\frac{1}{28})$.

Now, we return to (b-i). This pattern is more difficult to study than the other four Poisson patterns, since two occurrences of this pattern in a RTCN might overlap; see (a) in Figure 5.4. (In this figure, we do not care about the ranks of the first two events.) So in order to set up the correct Markov chain, we now have one more type to consider than before, which is the overlap case; see type A in Figure 5.4. In addition a pattern of type B is a pattern (b-i) which is not contained in a pattern of type A, a pattern of type B is a cherry and patterns of type C are all remaining external lineages.

Assume that the number of patterns of type A, type B and type C in a ranked tree-child network with n leaves is given by a, b and c, respectively. Table 5.13 and Table 5.14 give the transition probabilities of the Markov chain for the number of patterns of type A, type B and type C. Note that the number of (b-i) is actually $2a + b$.

We now have the following result.

Lemma 5.3.4. *Denote by Y_n and $\tilde{X}_n$ the number of occurrences of patterns of type A and type B, respectively, in a random ranked tree-child network with n leaves. Then, for $r, s, t \geq 0$, we*

DOI:10.6814/NCCU202201049

	type A	type B	type C	probability
A	-1	0	$+1$	$3a/n^2$
	-1	$+1$	$+1$	$2a/n^2$
B	0	-1	$+1$	$4b/n^2$
C	0	0	0	$2c/n^2$
D	0	0	$+1$	$(n-5a-4b-2c)/n^2$

Table 5.13: The change of the number of patterns of type A, type B and type C for (b-i) when the next event is a branching event.

have

$$\mathbb{E}(Y_{n+1}^{\underline{r}}\tilde{X}_{n+1}^{\underline{s}}C_{n+1}^{\underline{t}}) = \left(1-\frac{5r+4s+2t}{n}\right)^2\mathbb{E}(Y_n^{\underline{r}}\tilde{X}_n^{\underline{s}}C_n^{\underline{t}}) + \frac{t}{n}\mathbb{E}(Y_n^{\underline{r}}\tilde{X}_n^{\underline{s}}C_n^{\underline{t-1}}) \tag{5.12}$$
$$+\frac{4s}{n}\mathbb{E}(Y_n^{\underline{r+1}}\tilde{X}_n^{\underline{s-1}}C_n^{\underline{t}}) + \frac{4s}{n}\mathbb{E}(Y_n^{\underline{r}}\tilde{X}_n^{\underline{s-1}}C_n^{\underline{t+1}}) + \frac{R_n}{n^2},$$

where R_n is given by

$$\begin{aligned}
&t(2-5r-4s-2t)\mathbb{E}(Y_n^{\underline{r}}\tilde{X}_n^{\underline{s}}C_n^{\underline{t-1}}) - 2s(1+10r+8s+4t)\mathbb{E}(Y_n^{\underline{r+1}}\tilde{X}_n^{\underline{s-1}}C_n^{\underline{t}})\\
&+4s(2-5r-4s-2t)\mathbb{E}(Y_n^{\underline{r}}\tilde{X}_n^{\underline{s-1}}C_n^{\underline{t+1}}).\\
&+4r\mathbb{E}(Y_n^{\underline{r-1}}\tilde{X}_n^{\underline{s}}C_n^{\underline{t+2}}) + 2st\mathbb{E}(Y_n^{\underline{r+1}}\tilde{X}_n^{\underline{s-1}}C_n^{\underline{t-1}}) - 8s\mathbb{E}(Y_n^{\underline{r}}\tilde{X}_n^{\underline{s-1}}C_n^{\underline{t+2}})\\
&+4s(s-1)\mathbb{E}(Y_n^{\underline{r+2}}\tilde{X}_n^{\underline{s-2}}C_n^{\underline{t}}) + 8s(s-1)\mathbb{E}(Y_n^{\underline{r+1}}\tilde{X}_n^{\underline{s-2}}C_n^{\underline{t+1}}).
\end{aligned}$$

Proof. Using the results from Table 5.13 and Table 5.14, $\mathbb{E}(Y_{n+1}^{\underline{r}}\tilde{X}_{n+1}^{\underline{s}}C_{n+1}^{\underline{t}})$ given $Y_n, \tilde{X}_n$ and C_n can be written as a sum of terms of the form

$$(Y_n+k)^{\underline{r}}(\tilde{X}_n+\ell)^{\underline{s}}(C_n+m)^{\underline{t}}$$

which are multiplied with the probabilities. Here, k, ℓ and m are suitable integers, e.g., for the contribution from the third sub-row of column A & A in Table 5.14, we have $k=-2, \ell=2, m=0$. Then, we rewrite, e.g., this term as

$$(Y_n-2)^{\underline{r}}(\tilde{X}_n+2)^{\underline{s}}C_n^{\underline{t}} = \frac{(Y_n-r)(Y_n-r-1)(\tilde{X}_n+2)(\tilde{X}_n+1)}{Y_n(Y_n-1)(\tilde{X}_n-s)(\tilde{X}_n-s-1)}Y_n^{\underline{r}}\tilde{X}_n^{\underline{s}}C_n^{\underline{t}}$$

and similar for the other terms. The rest of the proof is just a long computation (which is best done with the help of Maple). ∎

The last lemma implies the following result.

DOI:10.6814/NCCU202201049

	type A	type B	type C	probability
A	-1	0	0	$20an^2$
A & A	-2	0	0	$9a(a-1)/n^2$
	-2	$+1$	0	$12a(a-1)/n^2$
	-2	$+2$	0	$4a(a-1)/n^2$
B	0	-1	0	$12b/n^2$
B& B	0	-2	0	$16b(b-1)/n^2$
C	0	0	-1	$2c/n^2$
C & C	$+1$	0	-2	$4c(c-1)/n^2$
D & D	0	0	0	$(n-5a-4b-2c)(n-5a-4b-2c-1)/n^2$
A & B	-1	-1	0	$24ab/n^2$
	-1	0	0	$16ab/n^2$
A & C	-1	$+1$	-1	$12ac/n^2$
	-1	$+2$	-1	$8ac/n^2$
A & D	-1	0	0	$6a(n-5a-4b-2c)/n^2$
	-1	$+1$	0	$4a(n-5a-4b-2c)/n^2$
B & C	0	0	-1	$16bc/n^2$
B & D	0	-1	0	$8b(n-5a-4b-2c)/n^2$
C & D	0	$+1$	-1	$4c(n-5a-4b-2c)/n^2$

Table 5.14: The change of the number of patterns of type A and type B for (b-i) when the next event is a reticulation event.

Proposition 5.3.5. *(a) Let Y_n be the number of occurrences of the pattern from Figure 5.4-(a) in a random ranked tree-child network with n leaves. Then,*

$$Y_n \xrightarrow{L_1} 0, \quad \textit{as} \quad n \to \infty.$$

(b) Define $\tilde{X}_n$ as in Lemma 5.3.4. Then,

$$(\tilde{X}_n, C_n) \xrightarrow{w} (\tilde{X}, C), \quad \textit{as} \quad n \to \infty,$$

where $\tilde{X}$ and C are independent Poisson random variables with parameters $1/8$ and $1/4$, respectively.

DOI:10.6814/NCCU202201049

Proof. We use induction with respect to the lexicographic order on (s, r, t) to show that for all $r, s, t \geq 0$:

$$\mathbb{E}(Y_n^r \tilde{X}_n^s C_n^t) \sim 0^r \cdot \left(\frac{1}{8}\right)^s \cdot \left(\frac{1}{4}\right)^t, \tag{5.13}$$

where we use the convention that $0^0 := 1$; both claims then follow from this result.

First, note that the second, third and fourth term on the right hand side of (5.3.4) and all terms in R_n are of a smaller lexicographic order than (s, r, t). Also note that (5.3.4) has the form (5.4).

Now, the induction base holds because of (5.9). Next, assume that the claim holds for all sequences (s', r', t') which are lexicographic smaller than (s, r, t).

If $r > 0$ then the induction hypothesis implies that (5.3.4) satisfies (5.4) with $\psi_n = o(1/n)$. Thus, from Lemma 5.2.7, we obtain that $\mathbb{E}(Y_n^r \tilde{X}_n^s C_n^t) = o(1)$ which proves the claim in this case.

Let $r = 0$, then again by the induction claim, (5.3.4) is of the form (5.4) with

$$\psi_n \sim \frac{t}{n}\mathbb{E}(\tilde{X}_n^s C_n^{t-1}) + \frac{4s}{n}\mathbb{E}(\tilde{X}_n^{s-1} C_n^{t+1}) \sim \frac{8s + 4t}{8^s 4^t n}.$$

Thus, by Lemma 5.2.7

$$\mathbb{E}(\tilde{X}_n^s C_n^t) \sim \frac{1}{8^s 4^t}.$$

which also proves the claim in this case. ∎

Now, we can prove the limit law result for (b-i).

Proof of Theorem 5.3.1-(b) for the pattern in Figure 5.1-(b-i). First note that $X_n = \tilde{X}_n + 2Y_n$.

Since by Proposition 5.3.5, we have $Y_n \xrightarrow{L_1} 0$ which implies $Y_n \xrightarrow{P} 0$ and $\tilde{X}_n \sim \text{Poisson}(1/8)$, by Theorem 2.2.6, we obtain that $X_n \sim \text{Poisson}(1/8)$.

This proves the claimed result. ∎

Remark 5.3.6. Note that it is also feasible to prove directly:

$$\mathbb{E}(Y_n^r X_n^{\tilde{s}} C_n^t) \sim 0^r \cdot \left(\frac{1}{8}\right)^s \cdot \left(\frac{1}{4}\right)^t. \tag{5.14}$$

However, computations will be more messy.

Normal Limit Laws. In this paragraph, we consider the patterns in Figure 5.1-(c) which according to Theorem 5.3.1 are both normal distributed. We will only give details for the

DOI:10.6814/NCCU202201049

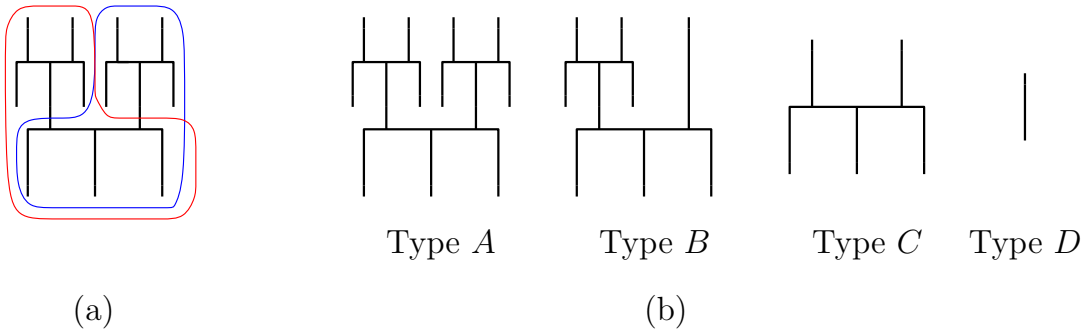


Figure 5.5: (a) The overlapping pattern containing two versions of (c-i). (b) Classification of types of patterns for (c-i): The pattern from (a) (Type A); The pattern (c-i) which is not contained in a pattern of type A (Type B); A trident which is not contained in a pattern of type A or type B (Type C); The remaining external lineages (Type D).

pattern in Figure 5.1-(c-i) since the pattern in Figure 5.1-(c-ii) can be studied in the same way; see [15].

Note that the pattern in Figure 5.1-(c-i) is similar to the pattern in Figure 5.1-(b-i), namely, both can overlap; see Figure 5.5-(a). Thus, we will consider again four types of patterns in the sequel; see Figure 5.5-(b). Assume that a RTCN with n leaves contains a, b and c patterns of type A, B and C, respectively. We will list the transition probabilities if a reticulation or branching event is attached to an RTCN with n leaves. However, this time there are a lot more cases than before. We list all of them in Tables 5.15-5.17.

Using these tables, we can first derive the means of the number of occurrences of the patterns from Figure 5.1-(c-i) and Figure 5.5-(a).

Lemma 5.3.7. *Denote by X_n and Y_n the number of occurrences of the patterns from Figure 5.1-(c-i) and Figure 5.5-(a), respectively. Let $\rho_n := \mathbb{E}(X_n)$ and $\tau_n := \mathbb{E}(Y_n)$. Then,*

$$\rho_{n+1} = \left(1 - \frac{5}{n}\right)^2 \rho_n + \frac{2}{n}\mathbb{E}(T_n) - \frac{6}{n^2}\mathbb{E}(T_n)$$

and

$$\tau_{n+1} = \left(1 - \frac{7}{n}\right)^2 \tau_n - \frac{1}{n^2}\mathbb{E}(T_n) + \frac{1}{n^2}\mathbb{E}(T_n^2).$$

The solutions of the above two recurrences are given by

$$\rho_n = \frac{(1080n^5 - 16668n^4 + 96992n^3 - 261735n^2 + 319471n - 135654)n}{41580(n-1)(n-2)(n-3)(n-4)(n-5)}, \qquad (n \geq 6) \tag{5.15}$$

DOI:10.6814/NCCU202201049

	type A	type B	type C	probability
A	-1	0	0	$3a/n^2$
	-1	$+1$	0	$4a/n^2$
B	0	-1	0	$3b/n^2$
	0	-1	$+1$	$2b/n^2$
C	0	0	-1	$3c/n^2$
D	0	0	0	$(n-7a-5b-3c)/n^2$

Table 5.15: The change of the number of patterns of type A, B and C (see Figure 5.5-(b)) when the next event is a branching event.

and, for $n \geq 7$,

$$\tau_n = \frac{(4290n^7-125730n^6+1509970n^5-9550275n^4+33968326n^3-66128140n-24510098)n}{3153150(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)(n-7)}. \quad (5.16)$$

Proof. Assume that $Y_n = d, X_n = e$ and $T_n = f$. Then, we have d patterns of type A, $e-2d$ patterns of type B and $f+e-d$ patterns of type C.

Next, we use the transition probabilities from the Tables 5.15-5.17 to get the recurrence for μ_n and ρ_n. Note that both recurrences are of type (5.4) whose exact solution is given by (5.5). Using the result from Lemma 5.2.5, a corresponding result for $\mathbb{E}(T_n^2)$, namely,

$$\mathbb{E}(T_n)^2 = \frac{(32175n^7-594825n^6+3960110n^5-10054065n^4-2875852n^3+58833733n^2-92961907n+42401933)n}{1576575(n-1)(n-2)(n-3)(n-4)(n-5)(n-6)}$$

and straightforward computations (best done with Maple) gives the claimed results for ρ_n and τ_n. ∎

We next derive the recurrence for the mixed moments.

Lemma 5.3.8. *Denote by* X_n *and* Y_n *the number of occurrences of the patterns from Figure 5.1-(c-i) and Figure 5.5-(a), respectively, in a random RTCN with* n *leaves. Moreover, set* $\mu_n := \mathbb{E}(T_n), \rho_n := \mathbb{E}(X_n), \tau_n := \mathbb{E}(Y_n)$ *and* $\bar{T}_n := T_n - \mu_n, \bar{X}_n := X_n - \rho_n, \bar{Y}_n := Y_n - \tau_n$. *Then, for all* $r, s, t \geq 0$, *we have*

$$\mathbb{E}(\bar{Y}_{n+1}^r \bar{X}_{n+1}^s \bar{T}_{n+1}^t) = \left(1 - \frac{7r+5s+3t}{n}\right)^2 \mathbb{E}(\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^t) + R_n \quad (5.17)$$

DOI:10.6814/NCCU202201049

	type A	type B	type C	probability
A	-1	0	$+1$	$22a/n^2$
	-1	0	$+2$	$8a/n^2$
	-1	$+1$	0	$8a/n^2$
	-1	$+1$	$+1$	$4a/n^2$
A & A	-1	0	0	$a(a-1)/n^2$
	-2	0	$+1$	$4a(a-1)/n^2$
	-2	$+1$	0	$4a(a-1)/n^2$
	-2	$+1$	$+1$	$16a(a-1)/n^2$
	-2	$+2$	0	$8a(a-1)/n^2$
	-2	$+2$	$+1$	$16a(a-1)/n^2$
B	0	0	0	$4b/n^2$
	0	-1	$+1$	$14b/n^2$
	0	-1	$+2$	$2b/n^2$
B & B	0	-1	0	$4b(b-1)/n^2$
	0	-1	$+1$	$4b(b-1)/n^2$
	0	-2	$+1$	$4b(b-1)/n^2$
	0	-2	$+2$	$8b(b-1)/n^2$
	0	-2	$+3$	$4b(b-1)/n^2$
	$+1$	-2	0	$b(b-1)/n^2$
C	0	0	0	$6c/n^2$
C & C	0	0	-1	$4c(c-1)/n^2$
	0	$+1$	-2	$4c(c-1)/n^2$
	$+1$	0	-2	$c(c-1)/n^2$
D & D	0	0	$+1$	$(n-7a-5b-3c)(n-7a-5b-3c-1)/n^2$

Table 5.16: The variation of the number of patterns of type A, B and C (see Figure 5.5-(b)) when attaching the next reticulation event to one or two patterns of type X with $X \in \{A, B, C, D\}$.

DOI:10.6814/NCCU202201049

	type A	type B	type C	probability
A & B	0	-1	0	$2ab/n^2$
	-1	-1	$+1$	$8ab/n^2$
	-1	-1	$+2$	$8ab/n^2$
	-1	0	0	$8ab/n^2$
	-1	0	$+1$	$20ab/n^2$
	-1	0	$+2$	$16ab/n^2$
	-1	$+1$	0	$8ab/n^2$
A & C	0	0	-1	$2ac/n^2$
	-1	0	0	$8ac/n^2$
	-1	$+1$	-1	$8ac/n^2$
	-1	$+1$	0	$16ac/n^2$
	-1	$+2$	-1	$8ac/n^2$
A & D	-1	0	$+1$	$4a(n-7a-5b-3c)/n^2$
	-1	$+1$	0	$2a(n-7a-5b-3c)/n^2$
	-1	$+1$	$+1$	$8a(n-7a-5b-3c)/n^2$
B & C	0	-1	0	$8bc/n^2$
	0	-1	$+1$	$8bc/n^2$
	0	0	-1	$8bc/n^2$
	0	0	0	$4bc/n^2$
	$+1$	-1	-1	$2bc/n^2$
B & D	0	-1	$+1$	$4b(n-7a-5b-3c)/n^2$
	0	-1	$+2$	$4b(n-7a-5b-3c)/n^2$
	0	0	0	$2b(n-7a-5b-3c)/n^2$
C & D	0	0	0	$4c(n-7a-5b-3c)/n^2$
	0	$+1$	-1	$2c(n-7a-5b-3c)/n^2$

Table 5.17: The change of the number of patterns of type A, B and C (see Figure 5.5-(b)) when attaching the next reticulation event to a pattern of type X and a pattern of type Y with $(X, Y) \in \{(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)\}$.

DOI:10.6814/NCCU202201049

with

$$R_n = \sum_{(s',r',t')} \mathbb{E}(\bar{Y}_n^{r'} \bar{X}_n^{s'} \bar{T}_n^{t'}) \Lambda_{r',s',t'}(n), \tag{5.18}$$

where the sum runs over (s', r', t') *which are of a smaller lexicographic order than* (s, r, t) *and* $\Lambda_{r',s',t'}(n)$ *admits the complete asymptotic expansion:*

$$\Lambda_{r',s',t'}(n) \sim \sum_{\ell=0}^{\infty} \frac{\lambda_{r',s',t',\ell}}{n^{\ell}}, \qquad n \to \infty. \tag{5.19}$$

Moreover, all terms in (5.18) with $(r' + s' + t')/2 - \ell \geq (r + s + t)/2 - 1$ *are given by*

$$\begin{aligned} &\frac{2s}{n}\mathbb{E}(\bar{Y}_n^r \bar{X}_n^{s-1} \bar{T}_n^{t+1}) + \frac{2r}{7n}\mathbb{E}(\bar{Y}_n^{r-1} \bar{X}_n^s \bar{T}_n^{t+1}) \\ &\quad + \binom{r}{2}\frac{20992}{540225}\mathbb{E}(\bar{Y}_n^{r-2} \bar{X}_n^s \bar{T}_n^t) + \binom{s}{2}\frac{13564}{29645}\mathbb{E}(\bar{Y}_n^r \bar{X}_n^{s-2} \bar{T}_n^t) + \binom{t}{2}\frac{24}{49}\mathbb{E}(\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^{t-2}) \\ &\quad - \frac{64}{539}st\mathbb{E}(\bar{Y}_n^r \bar{X}_n^{s-1} \bar{T}_n^{t-1}) - \frac{8}{343}rt\mathbb{E}(\bar{Y}_n^{r-1} \bar{X}_n^s \bar{T}_n^{t-1}) + \frac{652}{11319}rs\mathbb{E}(\bar{Y}_n^{r-1} \bar{X}_n^{s-1} \bar{T}_n^t). \end{aligned} \tag{5.20}$$

Proof. Assume that Y_n, X_n and T_n are given. Then, using the cases listed in the Tables 5.15-5.17, we can write the conditional expectation of $\bar{Y}_{n+1}^r \bar{X}_{n+1}^s \bar{T}_{n+1}^t$ given Y_n, X_n and T_n as a sum of terms of the form

$$(\bar{Y}_n + \tau_n - \tau_{n+1} + k)^r (\bar{X}_n + \rho_n - \rho_{n+1} + \ell)^s (\bar{T}_n + \mu_n - \mu_{n+1} + m)^t \tag{5.21}$$

multiplied with the probabilities from the Tables 5.15-5.17 (where $a = \bar{Y}_n + \tau_n, b = \bar{X}_n + \rho_n - 2(\bar{Y}_n + \tau_n), c = \bar{T}_n + \mu_n + \bar{X}_n + \rho_n - \bar{Y}_n - \tau_n$). Here, k, ℓ, m are integers depending on the case considered. For convenience, we will call this sum S in the sequel.

Now, use the expansion:

$$(\bar{X}_n + \rho_n - \rho_{n+1} + \ell)^s = \sum_{j=0}^{s} \binom{s}{j} \bar{X}_n^j (\rho_n - \rho_{n+1} + \ell)^{s-j}. \tag{5.22}$$

First, by replacing the middle term in (5.21) by $\bar{X}_n^s$ (the first term in the expansion (5.22)) and using Maple to sum up all these terms multiplied with their probabilities (which contain at most $\bar{X}_n^2$), we find that $\bar{X}_n^{s+2}$ and $\bar{X}_n^{s+1}$ do not appear in this sum. Likewise, by replacing the middle term in (5.21) by $s\bar{X}_n^{s-1}(\rho_n - \rho_{n+1} + \ell)$ (the second term in the expansion (5.22)), we see that $\bar{X}_n^{s+1}$ does not appear. Thus, the highest power of $\bar{X}_n$ in the sum S is $\bar{X}_n^s$. Next, by collecting these highest terms and repeating the above argument with the first term in (5.21), we see that terms with $\bar{Y}_n^{r+2} \bar{X}_n^s$ and $\bar{Y}_n^{r+1} \bar{X}_n^s$ do not occur in S. Finally, a similar line of reasoning

DOI:10.6814/NCCU202201049

shows that terms with $\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^{t+2}$ and $\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^{t+1}$ do not appear in S as well. Overall, this shows that sum in (5.18) is over the indicated range.

Next, the claimed expansion for $\Lambda_{r',s',t'}(n)$ follows by expanding $\tau_n - \tau_{n+1}, \rho_n - \rho_{n+1}$ and $\mu_n - \mu_{n+1}$ (see (5.3), (5.15) and (5.16), respectively; note that these expansions are all of the form (5.19)) and pointing out that the probabilities might contain factors of the form $\tau_n^2, \rho_n^2, \mu_n^2, \tau_n\rho_n, \tau_n\mu_n$, or $\rho_n\mu_n$ which are however divided by n^2 and thus expanding them gives also terms of the form (5.19).

Finally, in order to find all terms with $(r'+s'+t')/2 - \ell \geq (r+s+t)/2 - 1$, we proceed as follows: if we expand the first factor in (5.21) and retain only the terms which contain $\bar{Y}_n^r, \bar{Y}_n^{r-1}, \bar{Y}_n^{r-2}, \bar{Y}_n^{r-3}$ and $\bar{Y}_n^{r-4}$, then we see that we only loose terms $\bar{Y}_n^{r'} \bar{X}_n^{s'} \bar{T}_n^{t'}$ in S with

$$(r'+s'+t')/2 - \ell \leq (r'+s'+t')/2 \leq (r-5+s+t+2)/2 < (r+s+t)/2 - 1.$$

Similarly, we only loose smaller order terms when we keep the terms with $\bar{X}_n^s, \ldots, \bar{X}_n^{s-4}$ and the terms with $\bar{T}_n^t, \ldots, \bar{T}_n^{t-4}$ in the expansion of the second and third factor in (5.21), respectively. Thus, we only need to retain a fixed number (which does not depend on r, s and t) of terms in each of the terms of (5.21) from S. Then, the rest of the computation can then be done with Maple. ∎

The recurrence from the above lemma can now be used to prove the normal limit law of the pattern in Figure 5.1-(c-i) (which completes the proof of Theorem 5.3.1). In fact, we have a more general result.

Proposition 5.3.9. *Let the notation be as in Lemma 5.3.8. Then, as $n \to \infty$,*

$$\frac{1}{n}(Y_n - \mathbb{E}(Y_n), X_n - \mathbb{E}(X_n), T_n - \mathbb{E}(T_n)) \xrightarrow{w} N(\mathbf{0}, \Sigma),$$

where $N(\mathbf{0}, \Sigma)$ denotes a trivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix

$$\Sigma = \begin{pmatrix} \frac{270496}{203664825} & \frac{139276}{62537475} & -\frac{8}{13377} \\ \frac{139276}{62537475} & \frac{2930764}{137582445} & -\frac{304}{119119} \\ -\frac{8}{13377} & -\frac{304}{119119} & \frac{24}{637} \end{pmatrix}.$$

Proof of Proposition 5.3.9. We use induction with respect to the lexicographic order of (s, r, t) to show that for all $r, s, t \geq 0$:

$$\mathbb{E}(\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^t) \sim c_{r,s,t} n^{(r+s+t)/2},$$

DOI:10.6814/NCCU202201049

where $c_{r,s,t} := \mathbb{E}(N_1^r N_2^s N_3^t)$ with $N = (N_1, N_2, N_3)$ the trivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ.

First, the base case is implied by Proposition 5.2.9. Thus, we may assume that the claim holds for all sequences (s', r', t') which are lexicographic smaller than (s, r, t).

In order to prove the claim for (s, r, t), we observe that (5.17) has the form (5.4) with the sequence ψ_n satisfying $\psi_n \sim \tilde{c}_{r,s,t} n^{(r+s+t)/2-1}$ where

$$\begin{aligned}\tilde{c}_{r,s,t} :=& 2sc_{r,s-1,t+1} + \frac{2r}{7}c_{r-1,s,t+1} \qquad (5.23)\\ &+ \binom{r}{2}\frac{20992}{540225}c_{r-2,s,t} + \binom{s}{2}\frac{13564}{29645}c_{r,s-2,t} + \binom{t}{2}\frac{24}{49}c_{r,s,t-2}\\ &- \frac{64}{539}stc_{r,s-1,t-1} - \frac{8}{343}rtc_{r-1,s,t-1} + \frac{652}{11319}rsc_{r-1,s-1,t}. \qquad (5.24)\end{aligned}$$

Now, by Isserlis' theorem (see Thorem 2.2.4) we obtain that

$$c_{r,s-1,t+1} = t\Sigma_{3,3}c_{r,s-1,t-1} + (s-1)\Sigma_{3,2}c_{r,s-2,t} + r\Sigma_{3,1}c_{r-1,s-1,t}$$

and

$$c_{r-1,s,t+1} = t\Sigma_{3,3}c_{r-1,s,t-1} + s\Sigma_{3,2}c_{r-1,s-1,t} + (r-1)\Sigma_{3,2}c_{r-2,s,t}.$$

Inserting the two equations above into (5.24), we obtain that

$$\begin{aligned}\tilde{c}_{r,s,t} =& \binom{r}{2}\frac{270496}{7022925}c_{r-2,s,t} + \binom{s}{2}\frac{2930764}{6551545}c_{r,s-2,t} + \binom{t}{2}\frac{24}{49}c_{r,s,t-2}\\ &- \frac{304}{7007}stc_{r,s-1,t-1} - \frac{8}{637}rtc_{r-1,s,t-1} + \frac{139276}{2501499}rsc_{r-1,s-1,t}. \qquad (5.25)\end{aligned}$$

Next, again by Isserlis' theorem, we have the recurrences:

$$\begin{aligned}c_{r,s,t} &= (r-1)\Sigma_{1,1}c_{r-2,s,t} + s\Sigma_{1,2}c_{r-1,s-1,t} + t\Sigma_{1,3}c_{r-1,s,t-1};\\ c_{r,s,t} &= r\Sigma_{2,1}c_{r-1,s-1,t} + (s-1)\Sigma_{2,2}c_{r,s-2,t} + t\Sigma_{2,3}c_{r,s-1,t-1};\\ c_{r,s,t} &= r\Sigma_{3,1}c_{r-1,s,t-1} + s\Sigma_{3,2}c_{r,s-1,t-1} + (t-1)\Sigma_{3,3}c_{r,s,t-2}.\end{aligned}$$

Multiplying the first by $29r/2$, the second by $21s/2$ and the third by $13t/2$ gives as right-hand side exactly (5.25). Thus,

$$\tilde{c}_{r,s,t} = \left(\frac{29}{2}r + \frac{21}{2}s + \frac{13}{2}t\right)c_{r,s,t}.$$

Overall, we have shown that (5.17) has the form (5.4) with

$$\psi_n \sim \left(\frac{29}{2}r + \frac{21}{2}s + \frac{13}{2}t\right)c_{r,s,t}n^{(r+s+t)/2-1}.$$

DOI:10.6814/NCCU202201049

The induction claim follows now from this by applying Lemma 5.2.7. This completes the proof. ∎

5.4 A conjecture for patterns of any height

From Theorem 5.2.2, Theorem 5.2.3 and Theorem 5.3.1, we observe that there are only three limit laws for patterns of height 1 and 2: degenerate (such degenerate patterns occur not at all), Poisson (such Poisson patterns occur only sporadically) and normal (such normal patterns occur frequently). In fact, we believe this behavior will apply to patterns of any height.

In order to formulate this as a conjecture, first define a *fringe pattern* as a connected substructure of a RTCN which has entirely evolved from a fixed set of lineages by consecutively adding branching and reticulation events. Then, our conjecture for the limit law for number of occurrences of any such fringe pattern reads as follows.

Conjecture 5.4.1. *Let F be a general fringe pattern. Denote by P resp. P_1 and P_2 the patterns obtained by removing the last event. Then we have the following cases.*

(a) If P is a normal pattern, then F is a Poisson pattern; in all other cases, F is a degenerate pattern.

(b) If P_1, P_2 are both normal patterns, then F is a normal pattern; if P_1 is a normal pattern and P_2 is a Poisson pattern or vice versa, then F is a Poisson pattern; in all other cases, F is a degenerate pattern.

It is not difficult to see that this conjecture holds for all patterns studied in the previous two subsections, including the patterns of height 2 in Figure 5.1 and the two patterns of height 3 from Figure 5.4-(a) and Figure 5.5-(a).

The patterns in Figure 5.4-(a) was shown to be degenerate; see Proposition 5.3.5. Indeed, if we remove the last reticulation event of this pattern, then it splits into two Poisson patterns. Likewise, the pattern in Figure 5.5-(a) was shown to be normal; see Proposition 5.3.9; it splits into two normal patterns, namely, tridents, when the last reticulation event is removed.

Our method of proof of the results in this chapter (coupling with a Markov chain and using the method of moments) is rather computation-intensive; we believe that a different method is needed to be able to prove the above conjecture.

DOI:10.6814/NCCU202201049

Chapter 6

Conclusion

Here, we will review the whole thesis and summarize our main contributions.

We started the thesis by recalling the definition of *phylogenetic trees* which have been used in phylogenetics since Darwin and which have been extensively studied from an algorithmic, combinatorial and probabilistic point of view. We next introduced the main objects of this thesis, namely, *phylogenetic networks* which have been proposed since the evolutionary process is not always tree-like. Furthermore, of the many subclasses of phylogenetic networks, we concentrated on *tree-child network* (first introduced in [4]) and their variants as the main research subject of this thesis, because they have become one of the most prominent network classes in the last decade.

Most of the progress on enumeration and distributional properties of tree-child networks has happened the last few years; see [1, 5, 12–14, 17, 20, 22, 25]. In particular, in this thesis, we have further investigated results presented by Fuchs, Yu and Zhang in [17] and by Bienvenu, Lambert and Steel in [1]. More specifically, the main contributions of the thesis are in Chapter 3, Chapter 4, and Chapter 5.

In Chapter 3, we first reviewed the exact and asymptotic counting formulas for OTC_n and the asymptotic result for TC_n. In addition, we also explained a recent conjecture of Pons and Batle (see [22]) for $\mathrm{TC}_{n,k}$. Our main contribution in this section is a proof of the conjecture of Pons and Batle for one-component tree-child networks. (This was recently submitted; see [15].) Moreover, we also announced a Poisson limit law result for the number of reticulation nodes of a random tree-child network which will be contained in the journal version of [7] (which is currently under preparation).

DOI:10.6814/NCCU202201049

Next, in Chapter 4, we extended previous results to d-combining tree-child networks. One of the main new contributions in this part is an encoding of tree-child networks by certain words (which are different than the ones used in the conjecture of Pons and Batle). This encoding lead to recurrences (which extend the recurrence for $\mathrm{TC}_{n,n-1}$ from [17]) and in particular allowed us to prove that the limit law of the number of reticulation nodes of a random d-combining tree-child network is degenerate when $d \geq 3$. The results from this chapter have been announced in the recent extended abstract [7] or are stated there as conjectures. They (together with additional results) will be included into the journal version of [7].

Finally, in Chapter 5, we reviewed previous results on the limit behavior of the number of patterns of height 1 in random RTCNs. Moreover, we used an extension of the method in [1] (which was based on coupling with a Markov chain and the method of moments) to do further research on limit laws for the number of occurrences of patterns of height 1 and also height 2. In addition, we also gave a conjecture for patterns of any height. This is actually the first general study on statistical properties of patterns in a class of phylogenetic networks. (Corresponding studies for phylogenetic trees are plentiful; for instance see [6]). This part was recently submitted; see [16].

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202201049

Bibliography

[1] F. Bienvenu, A. Lambert, M. Steel (2022). Combinatorial and stochastic properties of ranked tree-child networks, *Random Struct. Algor.*, **60(4)**, 653–689.

[2] P. Billingsley (1995). *Probability and Measure*, third edition, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1995.

[3] A. Caraceni, M. Fuchs, G.-R. Yu (2022). Bijections for ranked tree-child networks, *Discrete Math.*, **345:9**, 112944, 10 pages.

[4] G. Cardona, G. Rossello, F. Valiente (2009). Comparison of tree-child phylogenetic networks. IEEE/ACM Trans. *Comput. Biol. Bioinform.* **6(4)**, 552–569.

[5] G. Cardona and L. Zhang (2020). Counting tree-child networks and their subclasses, *J. Comput.Syst. Sci.*, **114**, 84–104.

[6] H. Chang and M. Fuchs (2010). Limit theorems for patterns in phylogenetic trees, *J. Math. Biol.*, **60:4**, 481–512.

[7] Y.-S. Chang, M. Fuchs, H. Liu, M. Wallner, G.-R. Yu (2022). Enumeration of d-combining Tree-Child Networks, LIPICS, Proceedings of the 33rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, 225.

[8] K. P. Choi, G. Kaur, T. Wu (2021). On asymptotic joint distributions of cherries and pitchforks for random phylogenetic trees, *J. Math. Biol.*, **83:4**, Paper No. 40, 34 pp.

[9] K. P. Choi, A. Thompson, T. Wu (2020). On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees, *Theor. Popul. Biol.*, **132**, 92–104.

DOI:10.6814/NCCU202201049

[10] F. Disanto and T. Wiehe (2013). Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model, *Math. Biosci.*, **242:2**, 195–200.

[11] A. Elvey Price, W. Fang, M. Wallner (2021). Compacted binary trees admit a stretched exponential, *J. Comb. Theory Ser. A*, **177**, Article 105306.

[12] M. Fuchs, B. Gittenberger, M. Mansouri (2019). Counting phylogenetic networks with few reticulation vertices: tree-child and normal networks, *Australas. J. Combin.*, **73:2**, 385–423.

[13] M. Fuchs, B. Gittenberger, M. Mansouri (2021). Counting phylogenetic networks with few reticulation vertices: exact enumeration and corrections, *Australas. J. Combin.*, **82:2**, 257–282.

[14] M. Fuchs, E.-Y. Huang, G.-R. Yu, E.-Y. Huang (2022). Counting phylogenetic networks with few reticulation vertices: a second approach, *Discrete Appl. Math.*, in press.

[15] M. Fuchs, H. Liu, G.-R. Yu. A short note on the exact counting of tree-child networks, arXiv:2110.03842.

[16] M. Fuchs, H. Liu, T.-C. Yu. Limit theorems for patterns in ranked tree-child networks, arXiv:2204.07676.

[17] M. Fuchs, G.-R. Yu, L. Zhang (2021). On the asymptotic growth of the number of tree-child networks, *European J. Combin.*, **93**, 103278, 20 pages.

[18] C. Holmgren and S. Janson (2015). Limit laws for functions of fringe trees for binary search trees and recursive trees, Electron. *J. Probab.*, **20**, 1–51.

[19] G. Kaur, K. P. Choi, T. Wu. Distributions of cherries and pitchforks for the Ford model, arXiv:2110.02850.

[20] C. McDiarmid, C. Semple, D. Welsh (2015). Counting Phylogenetic Networks, *Ann. Comb.*, **19:1**,205–224.

[21] A. McKenzie and M. A. Steel (2000). Distributions of cherries for two models of trees, *Math. Biosci.*, **164:1**, 81–92.

DOI:10.6814/NCCU202201049

[22] M. Pons and J. Batle (2021). Combinatorial characterization of a certain class of words and a conjectured connection with general subclasses of phylogenetic tree-child networks, *Scientific Reports*, **11**, Article number: 21875.

[23] N. A. Rosenberg (2006). The mean and variance of the numbers of r-pronged nodes and rcaterpillars in Yule generated genealogical trees, *Ann. Comb.*, **10:1**, 129–146.

[24] T. Wu and K. P. Choi (2016). On joint subtree distributions under two evolutionary models, *Theor. Popul. Biol.*, **108**, 13–23.

[25] L. Zhang. The Sackin index of simplex networks, arXiv:2112.15379.

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202201049