Enumerative and Distributional Results for d-combining Tree-Child Networks

Yu-Sheng Chang[*] Michael Fuchs[*] Hexuan Liu[†]
Michael Wallner[‡] Guan-Ru Yu[§]

September 12, 2022

Abstract

Tree-child networks are one of the most prominent network classes for modeling evolutionary processes which contain reticulation events. Several recent studies have addressed counting questions for *bicombining tree-child networks* in which every reticulation node has exactly two parents. We extend these studies to *d-combining tree-child networks* where every reticulation node has now $d \geq 2$ parents, and we study one-component as well as general tree-child networks. Moreover, we also give results on the distributional behavior of shape parameters (e.g., number of reticulation nodes, Sackin index) of a network which is drawn uniformly at random from the set of all tree-child networks with the same number of leaves. We show phase transitions depending on d, leading to normal, Bessel, Poisson, and degenerate distributions. Some of our results are new even in the bicombining case.

Keywords: Phylogenetic network, tree-child network, exact enumeration, asymptotic enumeration, stretched exponential, limit law, phase transition

1 Introduction and Results

The evolutionary process of, e.g., chromosomes, species, and populations is not always tree-like due to the occurrence of reticulation events caused by meiotic recombination on the chromosome level, specification and horizontal gene transfer on the species level, and sexual recombination on the population level. Because of this, *phylogenetic networks* have been introduced as appropriate models for reticulate evolution. Studying the properties of these networks is at the moment one of the most active areas of research in phylogenetics; see [14, 17].

While algorithmic and combinatorial aspects of phylogenetic networks have been investigated now for a few decades, enumerating and counting phylogenetic networks as well as understanding their "typical shape" are relatively recent areas of research; see [17, page 253] where such questions are only discussed in one short paragraph. However, the last couple of years have seen a lot of progress on these questions, in particular for the class of *tree-child networks*, which is one of the most prominent subclasses amongst the many subclasses of phylogenetic networks; see [3, 9–13, 15, 16].

[*]Department of Mathematical Sciences, National Chengchi University, Taipei 116, Taiwan.
[†]Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804, Taiwan.
[‡]Institute for Discrete Mathematics and Geometry, TU Wien, 1040 Vienna, Austria.
[§]Department of Mathematics, National Kaohsiung Normal University, Kaohsiung 824, Taiwan.

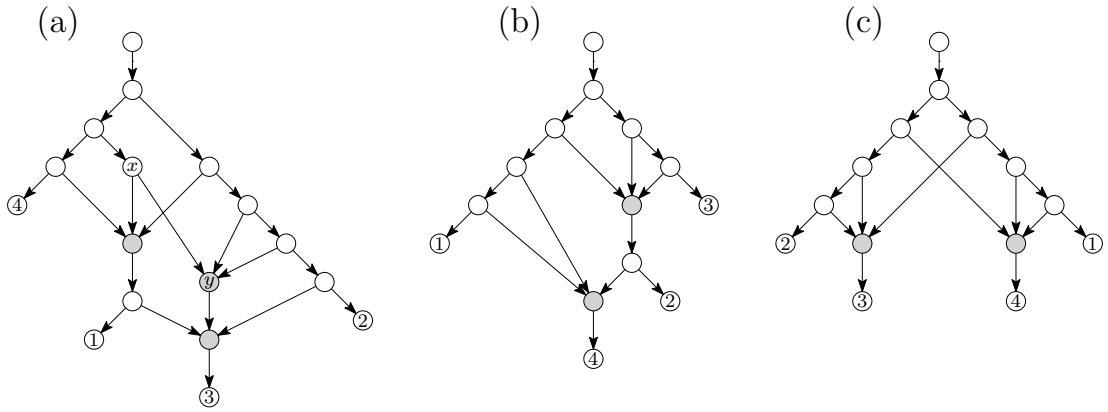

Figure 1: (a) A 3-combining phylogenetic network which is *not* a tree-child network (because both children of the tree node x are reticulation nodes and the only child of the reticulation node y is also a reticulation node); (b) a 3-combining tree-child network; (c) a 3-combining one-component tree-child network.

Most of the studies on tree-child networks have focused on *bicombining tree-child networks* which are tree-child networks where every reticulation event involves exactly two individuals. The purpose of this paper is to discuss extensions of previous results to *multicombining tree-child* networks. More precisely, we will focus on d*-combining tree-child networks* which are tree-child networks whose reticulation events involve exactly $d \geq 2$ individuals. Our main motivation for doing so, apart from scientific curiosity, is that this level of generality will ultimately result in a better understanding of the bicombining case, too (which is the most important case in applications). In particular, we will see that the two cases $d = 2$ and $d > 2$ frequently exhibit a very different behavior.

Before explaining our results, we will give precise definitions and fix some notation. We start with the definition of phylogenetic networks.

Definition 1.1 (Phylogenetic network)**.** *A (rooted) phylogenetic network with n leaves is a simple, directed acyclic graph (DAG) with no nodes of in- and out-degree 1, a (unique) node of in-degree 0 and out-degree 1 (the root) and exactly n nodes of in-degree 1 and out-degree 0 (the leaves) which are bijectively labeled with labels from the set $\{1, \ldots, n\}$.*

This definition is very general. In the sequel, we will restrict ourselves to the above mentioned d-combining networks where $d \geq 2$ is a fixed integer. ($d = 2$ is the bicombining case.)

Definition 1.2 (d-combining network)**.** *A phylogenetic networks is a d-combining network if all internal nodes (i.e. nodes which are neither leaves nor the root) have either in-degree 1 and out-degree 2 (tree nodes) or in-degree d and out-degree 1 (reticulation nodes).*

See Figure 1 for examples with $d = 3$. We next recall the definition of tree-child networks.

Definition 1.3 (Tree-child network)**.** *A d-combining network is called a tree-child network if every non-leaf node has at least one child which is not a reticulation node.*

In other words, a d-combining network is a tree-child network if (a) the root is not followed by a reticulation node; (b) a reticulation node is not followed by another reticulation node; and (c) a tree node has at least one child which is not a reticulation node; see Figure 1, (b) for an example. A simple and important subclass of tree-child networks is the class of one-component tree-child networks; see the definition below and Figure 1, (c) for an example.

Definition 1.4 (One-component tree-child network)**.** *A tree-child network is called a one-component tree-child network if every reticulation node is directly followed by a leaf.*

One-component networks are more "tree-like" than general tree-child networks. Moreover, they constitute an important building block in the construction of general tree-child networks; see [3] for the bicombining case and Appendix B for the d-combining case.

In this paper, we will give exact and asymptotic counting results for the number of one-component and general d-combining tree-child networks. Moreover, we will investigate the number of reticulation nodes (and other parameters) of a *random network* where random here means that the network is picked with the uniform distribution. We will detail some of our results below and give more results in the subsequent sections.

In order to state our results, we will need some notation. We denote by $\mathrm{OTC}_{n,k}^{(d)}$ and $\mathrm{TC}_{n,k}^{(d)}$ the number of one-component and general d-combining tree-child networks with n leaves and k reticulation nodes, respectively. Note that the tree-child property implies that $k \leq n-1$. Thus, the total number of one-component and general d-combining tree-child networks, denoted by $\mathrm{OTC}_n^{(d)}$ and $\mathrm{TC}_n^{(d)}$, satisfy

$$\mathrm{OTC}_n^{(d)} = \sum_{k=0}^{n-1} \mathrm{OTC}_{n,k}^{(d)} \qquad \text{and} \qquad \mathrm{TC}_n^{(d)} = \sum_{k=0}^{n-1} \mathrm{TC}_{n,k}^{(d)}.$$

Now, we are ready to present our results. First, for one-component tree-child networks, we will extend the exact formula for $\mathrm{OTC}_{n,k}^{(d)}$ for $d = 2$ from Theorem 13 in [3]. From this extension, we will then derive the following asymptotic counting results.

Theorem 1.5. *The following asymptotic equalities hold for one-component d-combining tree-child networks.*

(i) For $d = 2$ (bicombining case), we have

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

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

$$\begin{aligned}\mathrm{OTC}_n^{(3)} &\sim I_1(2) \cdot \mathrm{OTC}_{n,n-1}^{(3)} \\ &\sim \frac{I_1(2)\sqrt{3}}{9\pi}(n!)^3 \left(\frac{9}{2}\right)^n n^{-3},\end{aligned}$$

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

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

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

The result for the case $d = 2$ is already contained in [13]. Note that it is the only case of the three cases above in which we find a stretched exponential in the asymptotics; see [6]. The above mentioned formula for $\mathrm{OTC}_{n,k}^{(d)}$ also gives the following distributional result for the number of reticulation nodes.

Theorem 1.6. *Let $R_n^{(d)}$ be the number of reticulation nodes of a one-component d-combining tree-child network picked uniformly at random from the set of all one-component d-combining tree-child networks with n leaves. Then, we have the following limit behavior of $R_n^{(d)}$.*

(i) For $d = 2$ (bicombining case), we have the weak[1] *convergence result:*

$$\frac{R_n^{(2)} - n + \sqrt{n}}{\sqrt[4]{n/4}} \xrightarrow{w} N(0,1),$$

where $N(0,1)$ denotes the standard normal distribution.

(ii) For $d = 3$, we have the weak convergence result:

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

where Bessel(v,a) denotes the Bessel distribution, which is defined via $I_v(\alpha)$ from Theorem 1.5 (ii):

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

(iii) For $d \geq 4$, the limit law of $n-1-R_n^{(d)}$ is degenerate at 0, i.e., $n-1-R_n^{(d)} \xrightarrow{w}$ Dirac(0), where Dirac(λ) denotes the Dirac measure at λ.

Note that the above result for $d = 2$ is already contained in the proof of Theorem 3 in [13] where even a local limit theorem was proved; see also [12].

Remark 1.7. If t denotes the number of tree nodes and $\tilde{n}$ the total number of nodes, then by the hand-shaking lemma, we have

$$t = n + (d-1)k - 1 \tag{1}$$

and thus,

$$\tilde{n} = 2n + dk.$$

Therefore, we have similar limit distribution results for these numbers as well.

We next turn to general tree-child networks. Here, in contrast to one-component tree-child networks, we do not have an easy formula for $\mathrm{TC}_{n,k}^{(d)}$. However, we will introduce a way of encoding these networks by (certain) words and this encoding will lead to a recursive method for computing $\mathrm{TC}_{n,k}^{(d)}$. Using this method, the values of this sequence for small n, k and d can be computed; see Appendix A.

In addition, we will see that the growth of $\mathrm{TC}_{n,k}^{(d)}$ is dominated by $\mathrm{TC}_{n,n-1}^{(d)}$. For the latter sequence, a recurrence used for the computation of its values and the method of [6] (which needs some adaptions because of the dependence on d) will yield the following asymptotic counting result for $\mathrm{TC}_n^{(d)}$. For the bicombining case, this result was proved in [13].

Theorem 1.8. *For the number of d-combining tree-child networks with n leaves, we have*

$$\mathrm{TC}_n^{(d)} = \Theta\left(\mathrm{TC}_{n,n-1}^{(d)}\right) = \Theta\left((n!)^d \gamma(d)^n \, e^{3a_1\beta(d)n^{1/3}} n^{\alpha(d)}\right), \tag{2}$$

where $a_1 = -2.33810741\ldots$ is the largest root of the Airy function of the first kind, defined as the unique function Ai(z) *satisfying* Ai$''(z) = z\mathrm{Ai}(z)$ *such that* $\lim_{z\to\infty} \mathrm{Ai}(z) = 0$, *and*

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

The first few specific values of the asymptotic parameters $\alpha(d)$, $\beta(d)$, and $\gamma(d)$ are shown Table 1. Also, by performing a finer analysis for k close to n, we obtain the following distributional result for the number of reticulation nodes. (In the case $d \geq 3$, the above mentioned encoding will again play an important role in the proof; the case $d = 2$ will need a different treatment.)

[1] A sequence $(X_n)_{n\geq 1}$ of random variables converges weakly to a random variable X (denoted by $X_n \xrightarrow{w} X$) if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for each continuity point $x \in \mathbb{R}$ of F_X, where F_{X_n} and F_X are the respective cumulative distribution functions.

d	$\alpha(d)$	$\approx$	$\beta(d)$	$\approx$	$\gamma(d)$	$\approx$
2	$-\frac{5}{3}$	-1.67	$(\frac{1}{3})^{2/3}$	0.48	12	12.00
3	-3	-3.00	$(\frac{1}{2})^{2/3}$	0.63	32	32.00
4	$-\frac{22}{5}$	-4.40	$(\frac{3}{5})^{2/3}$	0.71	$\frac{250}{3}$	83.33
5	$-\frac{35}{6}$	-5.83	$(\frac{2}{3})^{2/3}$	0.76	216	216.00
6	$-\frac{51}{7}$	-7.29	$(\frac{5}{7})^{2/3}$	0.80	$\frac{16807}{30}$	560.23
7	$-\frac{35}{4}$	-8.75	$(\frac{3}{4})^{2/3}$	0.83	$\frac{65536}{45}$	1456.36
8	$-\frac{92}{9}$	-10.22	$(\frac{7}{9})^{2/3}$	0.85	$\frac{531441}{140}$	3796.01

Table 1: Specific values of the asymptotic parameters $\alpha(d)$, $\beta(d)$, and $\gamma(d)$ from Theorem 1.8.

Theorem 1.9. *Let $T_n^{(d)}$ be the number of reticulation nodes of a d-combining tree-child network picked uniformly at random from the set of all d-combining tree-child networks with n leaves. Then, we have the following limit behavior of $T_n^{(d)}$.*

(i) For $d = 2$ (bicombining case), we have the weak convergence result:

$$n - 1 - T_n^{(2)} \xrightarrow{w} \mathrm{Poisson}(1/2),$$

where $\mathrm{Poisson}(\alpha)$ denotes the Poisson distribution.

(ii) For $d \geq 3$, the limit distribution of $n - 1 - T_n^{(d)}$ is degenerate at 0.

This result is new even in the case $d = 2$. In fact, as far as we are aware of, it constitutes the first limit law result for a shape parameter in random tree-child networks. Also, the result for $d = 2$ improves [15, Theorem 1.5, (iii)], which states that the number of reticulation nodes of almost all bicombining tree-child networks with n leaves is asymptotic to n.

In addition, the above result can be used to improve and extend [15, Proposition 1.6, (ii)], which was concerned with the number of *twigs* of (bicombining) tree-child networks. A twig is a tree node which is contained in a *pendant subtree*, i.e., a tree node that has no reticulation node as descendant. In [15], it was proved that the number of twigs in a random bicombining tree-child networks is $o(n)$. In fact, twigs are even rarer than that.

Corollary 1.10. *Let $W_n^{(d)}$ be the number of twigs of a d-combining tree-child network picked uniformly at random from the set of all d-combining tree-child networks with n leaves.*

(i) For $d = 2$ (bicombining), we have

$$\mathbb{E}(W_n^{(d)}) = \mathcal{O}(1).$$

(ii) For $d \geq 3$, the limit law of $W_n^{(d)}$ is degenerate at 0. More precisely,

$$\mathbb{E}(W_n^{(d)}) \longrightarrow 0.$$

This result also shows that the expected number of *cherries*, i.e., tree nodes with both children leaves, of a random d-combining tree-child network is bounded, too (since cherries are clearly twigs). Note that the number of cherries is a popular parameter in phylogenetics and has been extensively studied for *phylogenetic trees* (which are bicombining networks without reticulation nodes).

Finally, as another consequence of Theorem 1.9, we obtain the following improvement of the first equality in (2).

Corollary 1.11. *The following asymptotic equalities hold for d-combining tree-child networks.*

(i) For $d = 2$ (bicombining case), we have $\mathrm{TC}_n^{(2)} \sim \sqrt{e} \cdot \mathrm{TC}_{n,n-1}^{(2)}$.

(ii) For $d \geq 3$, we have $\mathrm{TC}_n^{(d)} \sim \mathrm{TC}_{n,n-1}^{(d)}$.

Remark 1.12. Note that even with the above result, it is still not possible to give the first-order asymptotics of $\mathrm{TC}_n^{(d)}$ since the approach of [6] (which we are going to use in order to prove Theorem 1.8) gives only a Theta-result.

We next give an outline of the paper. We split our subsequent considerations into two parts: one-component and general d-combining tree-child networks.

In Section 2, we consider one-component d-combining tree-child networks. This section is divided into three subsections. In the first (Section 2.1), we state and prove an exact counting formula; the second subsection (Section 2.2) will then use this formula to deduce Theorem 1.5 (asymptotic counting) and Theorem 1.6 (limit law for the number of reticulation nodes). Finally, in Section 2.3, we will discuss another shape parameter, namely, the Sackin index. This index was recently investigated in [20] where the order of the mean was found for the bicombining case. Here, we will simplify the analysis and extend it to the d-combining case.

Next, in Section 3, we will present our results for general d-combining tree-child networks. Again this section will be split into three subsections. In Section 3.1, we will show how to encode networks by words. This encoding is similar to the one proposed in [16]; however, our encoding offers two advantages: (a) it can be rigorously proved (the encoding in [16] is just conjectural) and (b) our encoding works for all d (whereas the encoding from [16] seems to be restricted to $d = 2$). As mentioned above, this encoding will lead to a recursive method for computing $\mathrm{TC}_{n,k}^{(d)}$. This method will then be used in Section 3.2 to derive the Theta-result from Theorem 1.8. In addition, the encoding will also be helpful in Section 3.3 which will contain the proof of Theorem 1.9 (limit law for the number of reticulation nodes) in Section 3.3.1 (for $d = 2$) and Section 3.3.2 (for $d \geq 3$). Moreover, in Section 3.3.3 we will prove Corollary 1.10 for the number of twigs.

We will finish the paper with some concluding remarks and comment on further interesting generalizations in Section 4. Additionally, the paper contains two appendices: in Appendix A, we will list values of $\mathrm{TC}_{n,k}^{(d)}$ for small values of n, k, d. These values are computed with the recurrence from Section 3.1. Alternatively, the computation can be done with the method of component graphs from [3] whose extensions to d-combining networks will be explained in Appendix B. However, the computation of values via this method is more involved. Nevertheless, the method is still useful since it will allow us to find the first-order asymptotics of $\mathrm{TC}_{n,k}^{(d)}$ as $n \to \infty$ and fixed k; see [9–11] for the corresponding results in the bicombining case. Moreover, the method of component graphs also gives formulas for $\mathrm{TC}_{n,k}^{(d)}$ for (fixed) small values of k and d; see [3] and [10] for similar formulas for $d = 2$. We will list some of these formulas also in Appendix B.

We conclude the introduction by explaining the difference between this paper and the extended abstract [4] which was presented at the *33rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA2022)* which took place from June 20th to June 25th at the University of Pennsylvania, Philadelphia, USA. For one-component networks, the material of Section 2.1 and Section 2.2 was already contained in [4], while the material from Section 2.3 is new. For general networks, [4] just contained a sketch of the proof of Theorem 1.8. Here, we give the full proof in Section 3.2. Regarding Sections 3.1 and 3.3, some of its results were just mentioned (without any details) in the conclusion of [4], e.g., Theorem 1.9 (which we prove in Section 3.3) was stated only as conjecture. Finally, in Appendix B, we will prove and correct [4, Theorem 8], which was not proved previously. Also, the formulas in this appendix have been announced in [4].

2 One-Component Networks

In this section, we will prove results for one-component d-combining tree-child networks.

2.1 Exact Counting

In [3], the authors derived an exact formula for $\mathrm{OTC}^{(2)}_{n,k}$. The generalization of this formula to d-combining networks reads as follows.

Theorem 2.1. *The number of one-component d-combining tree-child networks with n leaves and k reticulation nodes for $0 \leq k \leq n-1$ is given by*

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

Remark 2.2. Since both Theorem 1.5 and Theorem 1.6 are derived from this result, we will give three different proofs of it; one below and the other two in the two remarks succeeding the proof. The first proof uses a recursion on n and k and is different from the one in [3], the second one uses a recursion on d and eventually relies on the formula for $\mathrm{OTC}^{(2)}_{n,k}$ in [3], and the third one gives a direct combinatorial interpretation of the closed form.

Proof. Suppose N is a one-component d-combining tree-child network with $n-1$ leaves and $k-1$ reticulation nodes.

Then, we can construct one-component d-combining tree-child networks with n leaves and k reticulation nodes from N by the following three steps: (i) put d new nodes into the *candidate edges* where we call an edge of N a candidate edge if it is not incident to any reticulation node; (ii) create a new reticulation node which is adjacent to the d new nodes; and (iii) add a new leaf as a child of this reticulation node and label it with a label from $\{1,\dots,n\}$; then increase all (old) labels in N which are at least as large as the new label by $+1$ (if there are any).

Now, note that in step (i), we have

$$\underbrace{n-1+(d-1)(k-1)-1}_{\substack{\#\text{ edges leading to}\\ \text{a tree node; see (1)}}} + \underbrace{n-1}_{\substack{\#\text{ edges leading}\\ \text{to a leaf}}} - \underbrace{(k-1)}_{\substack{\#\text{ edges below}\\ \text{ret. nodes}}} = 2n+(d-2)(k-1)-3$$

candidate edges and thus there are

$$\binom{2n+(d-2)k-2}{d}$$

choices of the d nodes. Moreover, in step (iii), there are n choices of the label. Finally, note that the above construction gives each network exactly k times.

Overall, the above arguments give

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

and by iteration,

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

The result follows now by the fact that

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

since $\mathrm{OTC}^{(d)}_{n-k,0}$ is the (well-known) number of phylogenetic trees with $n-k$ leaves given by the odd double factorials; e.g., see [17, Section 2.1]. □

Remark 2.3. A second way to obtain $\mathrm{OTC}_{n,k}^{(d)}$ proceeds by constructing $\mathrm{OTC}_{n,k}^{(d)}$ from $\mathrm{OTC}_{n,k}^{(d-1)}$ as follows. First, let a $(d-1)$-combining one-component tree-child networks with n leaves and k reticulation nodes be given. Then, there are $2n+(d-3)k-1$ edges whose end points are not reticulation nodes (i.e. candidate edges). We add k different nodes (each node being the dth parent of a reticulation node) to these edges. Overall there are

$$\prod_{i=0}^{k-1}(2n+(d-3)k-1+i)$$

ways to do this. Now, if we assign the first one of the k nodes to the first reticulation node, the second one to the second reticulation node, etc., we obtain every one-component d-combining tree-child network with n leaves and k reticulation nodes exactly d^k times. Thus,

$$\frac{\mathrm{OTC}_{n,k}^{(d)}}{\mathrm{OTC}_{n,k}^{(d-1)}} = \frac{\prod_{i=0}^{k-1}(2n+(d-3)k-1+i)}{d^k}.$$

From this, we can get the result for one-component d-combining networks by iteration and using the known result for the bicombining case from [3].

Remark 2.4. Alternatively, we may give the closed form for one-component networks a direct combinatorial interpretation. We construct all one-component networks as follows:

(i) Start with a phylogenetic tree with $n-k$ leaves, i.e., without reticulation nodes.

(ii) Place dk unary nodes along the $2(n-k)-1$ edges.

(iii) Attach k reticulation nodes that are each connected to d of the unary nodes. Attach to each reticulation node a leaf and label it from 1 to k in the order of creation.

(iv) Relabel the n leaves respecting the orders of the $n-k$ initial and k newly created leaves.

Each in that way created network is different and all one-component networks satisfy such a decomposition. Multiplying the number of possibilities for each step gives the claimed formula:

$$\mathrm{OTC}_{n,k}^{(d)} = \mathrm{OTC}_{n-k,0}^{(d)} \cdot \binom{2(n-k)+dk-2}{dk} \cdot \binom{dk}{d,d,\ldots,d} \cdot \binom{n}{k}, \tag{4}$$

where the result again follows from formula (3) for the total number of phylogenetic trees.

2.2 Number of Reticulation Nodes

From Theorem 2.1, we can now deduce Theorems 1.5 and 1.6 by the Laplace method. (A standard method of asymptotic analysis; see, e.g., [8, Chapter 4.7].)

Proof of Theorems 1.5 and 1.6. Since the results for $d=2$ are already contained in [13] (see also [12]), we can focus on the cases $d \geq 3$.

We start with the case $d=3$. Note that

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

and this sequence is increasing in k. (This is in contrast to $d=2$ where this sequence increases until its maximum at $k=n-\sqrt{n+1}$ and then decreases; see [13].) By replacing k by $n-1-k$ and using Stirling's formula, we obtain

$$\mathrm{OTC}_{n,n-1-k}^{(3)} = \frac{1}{k!(k+1)!} \cdot \frac{n(3n-3)!}{6^{n-1}} \left(1+\mathcal{O}\left(\frac{1+k^2}{n}\right)\right) \tag{5}$$

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

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

which is the first claim from Theorem 1.5, (ii); the second follows from this by another application of Stirling's formula. Moreover, since

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

the result from Theorem 1.6, (ii) follows from the above two expansions, too.

Next, we consider the case $d \geq 4$. The details of the proof are the same as above, with the main difference that the expansion (5) now becomes

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

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 $\mathrm{OTC}_n^{(d)}$ being now asymptotically dominated by $\mathrm{OTC}_{n,n-1}^{(d)}$ (proving Theorem 1.5, (iii)) and the limiting distribution of $n-1-R_n^{(d)}$ being degenerate (proving Theorem 1.6, (iii)). □

2.3 Sackin Index

In this section, we will investigate another shape parameter for random one-component tree-child networks, namely, the Sackin index. For phylogenetic trees, the investigation of this index has a long history and many results have been proved; see [7] for a summary of some of these results. For tree-child networks, [20] recently gave the first generalization of the Sackin index to networks. We will simplify this approach (which treated the bicombining case) and extend the main results to the d-combining case.

Until the end of this subsection, we assume that N is a one-component tree-child networks with n leaves and k reticulation nodes. Let us first state the definition of the Sackin index from [20].

Definition 2.5 (Sackin index)**.** *The Sackin index of N, in symbols $S(N)$, is defined as the sum over all leaves of the lengths of the longest paths to the leaves.*

For the analysis, following [20], we will define a second index. The *top tree component* of N, in symbols $C(N)$, is defined as the tree obtained from N by deleting all reticulation nodes together with their incident edges and their leaves below. Note that we retain the parents of all reticulation nodes and thus the top tree component is *not* a binary tree, it also has unary nodes. We denote by $P(N)$ the (total) path length of $C(N)$, i.e., the sum over all root-distances of all vertices. This index is related to the Sackin index as follows.

Lemma 2.6. *For all one-component tree-child networks N with at least two leaves, we have for $d = 2$*

$$S(N) \leq P(N) + 1 \leq 2S(N)$$

and for $d \geq 3$

$$S(N) \leq P(N) \leq dS(N). \tag{6}$$

Consequently $S(N) = \Theta(P(N))$, where the implied constants are absolute.

Proof. The result for $d = 2$ was given in [20] and the ideas of the proof in [20] can be used to handle also the case $d \geq 3$. For the readers convenience, we give some of the details.

First, we define the following sets of vertices of N:

(i) L_T collects leaves of N which are not below a reticulation node;

(ii) L_R collects leaves of N which are below a reticulation node;

(iii) P collects the parents of all reticulation nodes;

(iv) $R = V(C(N)) \setminus (L_T \cup P)$ collects the remaining vertices in $C(N)$. (Here, $V(C(N))$ denotes the set of vertices of $C(N)$.)

Then,

$$S(N) = \sum_{v \in L_T \cup L_R} \mathrm{depth}(v),$$

where the depth of v is the longest distance to the root.

Now, in order to prove the lower bound of $P(N)$, note that for $v \in L_R$, we have

$$\mathrm{depth}(v) = 2 + \max\{\mathrm{depth}(p) \ : \ p \text{ is a grandparent of } v\} \leq \sum_{p \text{ is a grandparent of } v} \mathrm{depth}(p).$$

Here, for the last inequality, we used that $d \geq 3$. Thus,

$$S(N) \leq \sum_{v \in L_T \cup P} \mathrm{depth}(v) \leq \sum_{v \in L_T \cup P \cup R} \mathrm{depth}(v) = P(N)$$

which shows the first part of the claim in (6).

Next, in order to show the upper bound of $P(N)$, we first recall that there exists a bijection ϕ from R to L_T such that w is an ancestor of $\phi(w)$ for all $w \in R$; see the appendix of [20] for details. Consequently, $\mathrm{depth}(w) \leq \mathrm{depth}(\phi(w))$ and thus,

$$\begin{aligned} P(N) &\leq \sum_{v \in L_T} \mathrm{depth}(v) + d \cdot \sum_{v \in L_R} \mathrm{depth}(v) + \underbrace{\sum_{w \in R} \mathrm{depth}(\phi(w))}_{= \sum_{v \in L_T} \mathrm{depth}(v)} \\ &= 2 \sum_{v \in L_T} \mathrm{depth}(v) + d \sum_{v \in L_R} \mathrm{depth}(l) \leq d\, S(N). \end{aligned}$$

This completes the proof. □

We will next analyze $P(N)$. Denote by $\mathcal{OC}_{n,k}^{(d)}$ the set of all one-component tree-child networks with n leaves and k reticulation nodes where the leaves below the reticulation nodes are labeled by the k *largest* labels from $\{1, \ldots, n\}$. Note that

$$\mathrm{OTC}_{n,k}^{(d)} = \binom{n}{k} |\mathcal{OC}_{n,k}^{(d)}|,$$

since all one-component tree-child networks are obtained from the networks in $\mathcal{OC}_{n,k}^{(d)}$ by re-labeling the leaves. Also, set

$$P(\mathcal{OC}_{n,k}^{(d)}) := \sum_{N \in \mathcal{OC}_{n,k}^{(d)}} P(N).$$

For this quantity, we have the following recurrence and exact formula; see [20, Theorem 3] for the bicombining case. The subsequent results simplify when using the notation of double factorials given by (see also (3))

$$n!! := \prod_{k=0}^{\lfloor n/2 \rfloor - 1} (n - 2k), \tag{7}$$

i.e., $(2n)!! = (2n) \cdot (2n-2) \cdot (2n-4) \cdots 2$ and $(2n+1)!! = (2n+1) \cdot (2n-1) \cdot (2n-3) \cdots 1$.

Proposition 2.7. *$P(\mathcal{OC}_{n,k}^{(d)})$ satisfies the recurrence*

$$P(\mathcal{OC}_{n,k}^{(d)}) = \binom{2n + (d-2)k}{d} P(\mathcal{OC}_{n-1,k-1}^{(d)})$$

with $P(\mathcal{OC}_{n,0}^{(d)}) = (2n)!! - (2n-1)!!$ as initial condition. Consequently,

$$P(\mathcal{OC}_{n,k}^{(d)}) = \frac{(2n + (d-2)k)!}{(d!)^k (2n-2k)!} \left((2n-2k)!! - (2n-2k-1)!! \right). \tag{8}$$

Proof. Let N be a network in $\mathcal{OC}_{n-1,k-1}^{(d)}$. We subsequently denote by $V(C(N))$ and $E(C(N))$ the vertex set and edge set of the top tree component, respectively. Moreover, for a $v \in V(C(N))$, we denote by $\delta_{C(N)}(v)$ the number of descendants and by $\alpha_{C(N)}(v)$ the number of ascendants of v in $C(N)$. Note that the number of ascendants of v in N and $C(N)$ is the same, i.e., $\alpha_{C(N)}(v) = \alpha_N(v)$. Finally, we denote by $\mathcal{S}$ the set of multisets of d edges of $C(N)$ (where edges are counted with repetition). Every $S \in \mathcal{S}$ corresponds to a set of edges where nodes are inserted in order to add an additional reticulation node with label n. This notion allows us to construct all networks of $\mathcal{OC}_{n,k}^{(d)}$ from those of $\mathcal{OC}_{n-1,k-1}^{(d)}$. More precisely, for $S \in \mathcal{S}$, we construct a network $N' = N'(S)$ by inserting nodes into the d edges from S and connecting them with a new reticulation node whose child has label n. (Note that this is a similar construction to the one we used in the proof of Theorem 2.1.)

We consider now $P(N')$ which by definition is given by:

$$P(N') = \sum_{v \in V(C(N'))} \alpha_{N'}(v).$$

Partitioning vertices of $C(N')$ into vertices of $C(N)$ and the d new ones which have been added to N to construct N', we have

$$P(N') = \sum_{v \in V(C(N))} \alpha_{N'}(v) + \sum_{v \in D} \alpha_{N'}(v),$$

where $D = V(C(N')) \setminus V(C(N))$ is the set of new vertices.

Now, to find a relation to $P(N)$, we replace $\alpha_{N'}$ by α_N and get

$$P(N') = \sum_{v \in V(C(N))} \alpha_N(v) + \sum_{v \in V(C(N))} \beta(v) + \sum_{v \in D} \alpha_N(v^{\downarrow}) + \sum_{v \in D} \gamma(v), \tag{9}$$

where

(i) $\beta(v)$ denotes the number of nodes in D that are ascendants of v in $C(N')$;

(ii) $v^{\downarrow}$ is closest descendant of v in $C(N')$ that belongs to $V(C(N))$;

(iii) $\gamma(v)$ denotes the number of nodes in D that are ascendants of v in $C(N')$.

Observe that the second sum can be rewritten as:

$$\sum_{v \in V(C(N))} \beta(v) = \sum_{v \in D} \delta_{C(N)}(v^{\uparrow}),$$

where $v^{\uparrow}$ is the closest ascendant of v in $C(N')$ which belongs to $V(C(N))$. Thus, (9) can be rewritten into:

$$P(N') = P(N) + \sum_{v \in D} \left(\delta_{C(N)}(v^{\uparrow}) + \alpha_N(v^{\downarrow}) \right) + \sum_{v \in D} \gamma(v).$$

Next, we sum both sides over $S \in \mathcal{S}$, which on the right-hand side gives three sums which we denote by Σ_1, Σ_2 and Σ_3, respectively.

First, for Σ_1, note that $E := |E(C(N))| = 2(n-k) - 1 + dk$. Thus,

$$\Sigma_1 = \sum_{S \in \mathcal{S}} P(N) = |\mathcal{S}| P(N) = \binom{E+d-1}{d} P(N)$$

since

$$|\mathcal{S}| = \#\{x_1 + \cdots + x_E = d \ : \ x_i \text{ non-negative integers}, \ 1 \leq i \leq E\},$$

where each x_i corresponds to an an edge from $E(C(N))$ (and counts how often that edge occurs in the set S).

Next, for Σ_2, set $S = \{e_1, \ldots, e_d\}$ and $B(e) = \delta_{C(N)}(v^{\uparrow}) + \alpha_N(v^{\downarrow})$ where e is an edge in S into which v is inserted and $v^{\uparrow}$ and $v^{\downarrow}$ are initial and end point of e, respectively. Then,

$$\Sigma_2 = \sum_{S \in \mathcal{S}} \left(B(e_1) + \cdots + B(e_d) \right).$$

We fix an edge e in $C(N)$ and count the number of times $B(e)$ occurs in the above sum. This gives,

$$\Sigma_2 = \sum_{e \in E(C(N))} \left(\sum_{x_1 + \cdots + x_E = d} x_i \right) B(e) = 2 \binom{E+d-1}{d-1} P(N),$$

where x_i inside the second sum is the term corresponding to e and for the third expression, we used that

$$\sum_{x_1 + \cdots + x_E = d} x_i = \frac{1}{E} \sum_{x_1 + \cdots + x_E = d} x_1 + \cdots + x_E = \binom{E+d-1}{d-1},$$

which holds for all $1 \leq i \leq E$ by symmetry. Moreover, we used that

$$\sum_{e \in E(C(N))} B(e) = 2\, P(N).$$

Finally, for Σ_3, we first give an explicit formula

$$\Sigma_3 = \sum_{v \in C(N) \setminus \{\rho\}} \sum_{0 \leq i+j \leq d} \left(ij + \frac{i(i-1)}{2} \right) \binom{\alpha_N(v) - 2 + j}{j} \binom{E - \alpha_N(v) - 1 + d - i - j}{d - i - j},$$

by the following combinatorial steps:

(i) Choose a vertex v in $C(N)$ which is not the root vertex ρ. We consider the incoming edge e of v;

(ii) Choose pair (i, j) with $0 \leq i + j \leq d$. Here, the meaning of i is that i nodes are inserted into e and j nodes are inserted into an edge which lies on the path from ρ to v (excluding e);

(iii) So far, the contribution of v to Σ_3 is

$$ij + \sum_{\ell=1}^{i-1} \ell = ij + \frac{i(i-1)}{2};$$

(iv) Next, there are $\binom{\alpha_N(v)-2+j}{j}$ ways of choosing the edges into which the j nodes are inserted;

(v) Finally, the second binomial coefficient counts the number of ways that the remaining $d-i-j$ new nodes are inserted into edges which are not on the path from ρ to v.

In order to simplify the formula, we set

$$\alpha := \alpha_N(v), \qquad A := \alpha - 2, \qquad B := E - \alpha - 1, \quad M := d - i.$$

Then, for the inner sum Σ_3, we have

$$\begin{aligned}
&\sum_{i=0}^{d} \left(i \sum_{j=0}^{M} j \binom{A+j}{j} \binom{B+M-j}{M-j} + \frac{i(i-1)}{2} \sum_{j=0}^{M} \binom{A+j}{j} \binom{B+M-j}{M-j} \right) \\
&= \sum_{i=0}^{d} \left(i(\alpha-1)[z^{M-1}](1-z)^{-A-2}(1-z)^{-B-1} + \frac{i(i-1)}{2}[z^M](1-z)^{-A-1}(1-z)^{-B-1} \right) \\
&= (\alpha-1) \sum_{i=0}^{d} i \binom{A+B+1+M}{M-1} + \sum_{i=0}^{d} \frac{i(i-1)}{2} \binom{A+B+1+M}{M} \\
&= (\alpha-1)[z^{d-2}](1-z)^{-A-B-5} + [z^{d-2}](1-z)^{-A-B-5} \\
&= \alpha \binom{E+d-1}{d-2}.
\end{aligned}$$

Plugging this into the above formula for Σ_3 gives

$$\Sigma_3 = \binom{E+d-1}{d-2} \sum_{v \in V(C(N)) \setminus \{\rho\}} \alpha_N(v) = \binom{E+d-1}{d-2} P(N).$$

Finally, summing the above expressions for Σ_1, Σ_2 and Σ_3 and summing over all N in $\mathcal{OC}^{(d)}_{n-1,k-1}$ gives the surprisingly simple recurrence:

$$P(\mathcal{OC}^{(d)}_{n,k}) = \binom{2n + (d-2)k}{d} P(\mathcal{OC}^{(d)}_{n-1,k-1}). \tag{10}$$

The initial condition, namely, the formula for $P(\mathcal{OC}^{(d)}_{n,0}) = (2n)!! - (2n-1)!!$ is well-known; see, e.g., [20]. From this, (8) is obtained by iteration. This concludes the proof. □

Remark 2.8. A *unary-binary tree* is a rooted tree whose nodes are either leaves (out-degree 0), unary (out-degree 1), or binary (out-degree 2). Then, we observe that the set $C(\mathcal{OC}^{(d)}_{n,k})$ of top trees obtained from all elements of $\mathcal{OC}^{(d)}_{n,k}$ is equal to the set of all unary-binary trees with $n-k$ labeled leaves and dk unary nodes. Hence, (8) nearly gives the path length of unary-binary trees, up to an overcount (which is actually a factor) that is related to reticulation nodes. Now, we adapt the construction of Remark 2.4 to build $\mathcal{OC}^{(d)}_{n,k}$ in order to determine this factor. Steps (i) and (ii) remain the same, and step (iv) is not performed as the leaves of reticulation nodes have maximal labels in $\mathcal{OC}^{(d)}_{n,k}$. It remains to

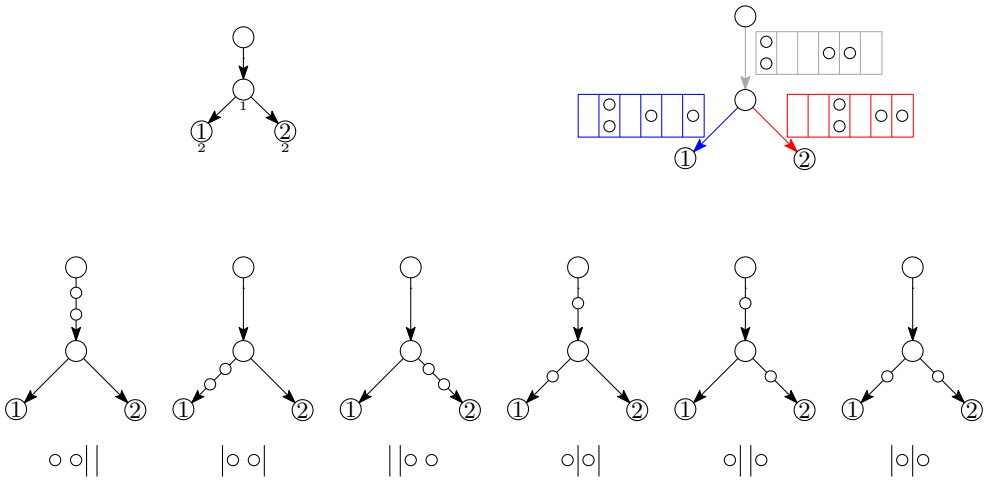

Figure 2: Construction of $\mathcal{OC}^{(2)}_{3,1}$ in Remark 2.9 ($K = 2$, $M = 3$). (Top, left) The only phylogenetic tree with 2 leaves; path lengths below each node and total path length 5. (Bottom) $\binom{4}{2} = 6$ top trees $\mathcal{OC}^{(2)}_{2,1}$ created from it after adding 2 unary nodes and the respective "balls and bars" diagrams. (Top, right) Superposition of all 6 top trees.

consider step (iii), which needs to be adapted. Observe that adding reticulation nodes and unary edges corresponds to the factor $\binom{dk}{d,d,\dots,d}$; see Figure 2. Thus, dividing (8) by this factor, we get that the path length of unary-binary trees with $n-k$ labeled leaves and dk unary nodes is equal to

$$\big((2(n-k))!! - (2(n-k)-1)!!\big) \cdot \binom{2(n-k)+dk}{dk}. \tag{11}$$

Remark 2.9. The surprising simplifications of Proposition 2.7 also allow a direct combinatorial explanation. By Remark 2.8, it suffices to show that the path length of unary-binary trees with $n-k$ labeled leaves and dk unary nodes is equal to (11).

The main idea is to use again the construction from Remark 2.4 to build the networks in $\mathcal{OC}^{(d)}_{n,k}$ bottom-up. We start in step (i) with a phylogenetic tree (i.e., without reticulation nodes) with $L := n-k$ leaves that is weighted by the path length. In the sequel, we call the nodes/edges of this phylogenetic tree, the original nodes/edges. As noted before, the total weight of such trees is $P(\mathcal{OC}^{(d)}_{L,0}) = (2L)!! - (2L-1)!!$. Then, we place $K := dk$ unary nodes along the $M := 2L-1$ edges. This gives a new structure enumerated by

$$P(\mathcal{OC}^{(d)}_{L,0}) \binom{K+M-1}{K}. \tag{12}$$

Note that this is nearly equal to what we want to prove in (11), except that the binomial coefficient should be $\binom{K+M+1}{K}$. What remains to be done, is to properly change the weights, as they do not correspond to the path lengths anymore, due to the additional unary nodes.

Let us start with a simple observation: Set all edge weights to w in a phylogenetic tree with L leaves. Then the path length is equal to w times the path length of the unweighted tree.

Having this observation in mind, we interpret unary nodes as weights on the original edges. We do this in three steps; the process is visualized in Figures 2 and 3 for $\mathcal{OC}^{(2)}_{3,1}$. The main idea is to superimpose all $\binom{K+M-1}{K}$ created trees and group them cleverly.

First, we "forget" about the unary nodes, and interpret the $\binom{K+M-1}{K}$ new instances of a phylogenetic tree as the same phylogenetic tree. Hence, the weights of all edges change to $\binom{K+M-1}{K}$ and the total

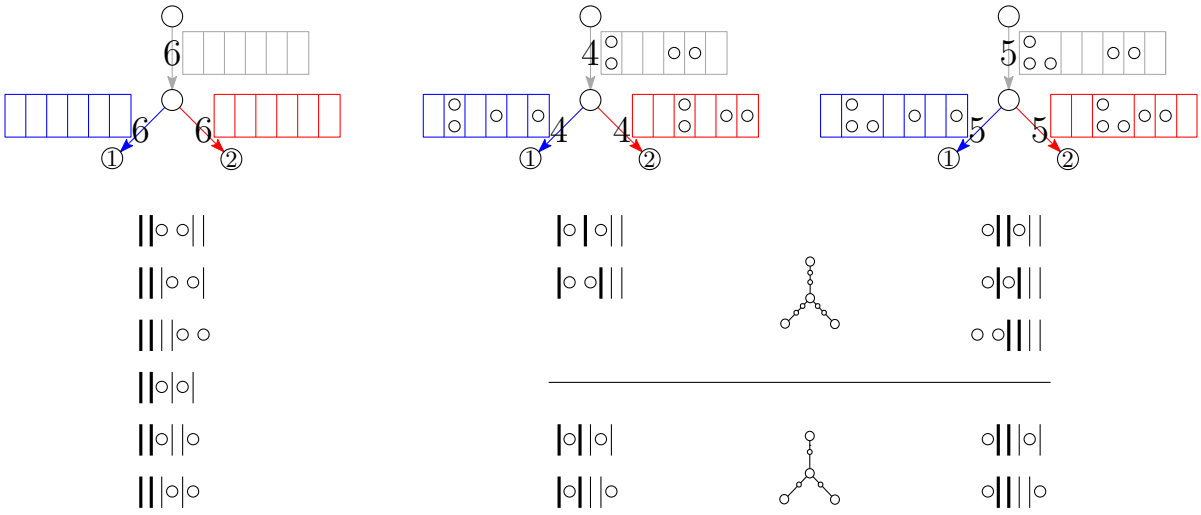

Figure 3: Enumeration of $P(\mathcal{OC}^{(2)}_{3,1})$ in Remark 2.9 using superposition ($K = 2$, $M = 3$); see Figure 2. (Left) Step 1: Each new instance increases edge weight by one; total $\binom{4}{2} = 6$. (Middle) Step 2: Correct path length of original nodes; sum unary nodes per edge. (Right) Step 3: Add path length for unary nodes; for i nodes weight $1 + 2 + \cdots + i$. (Bottom) "Balls and bars" corresponding to each step; the two added bars are shown bold.

number of these weighted phylogenetic trees is (12). In the sequel, the following direct interpretation of the binomial coefficient using "balls and bars" will be useful: The K unary nodes correspond to balls and the edges M to bins, which are modeled by $M - 1$ bars. Then each distribution of unary nodes on the edges corresponds to choosing K out of M bins with repetitions allowed; see Figure 2.

Second, we correct the weights of all original edges. Note that each unary node splits an edge into two and thereby increases the path length of any node below by one, or, equivalently, increases the weight of the associated original edge by one. Hence, the weight of an edge increases by the number of unary nodes it contains. In the superposition of all trees, all edges have the same number of unary nodes, as each edge is chosen equally often. The weight is equal to the sum of all unary nodes placed onto a fixed edge among the $\binom{K+M-1}{K}$ configurations. As there are M edges and K unary nodes (all equally distributed) we need a factor of K/M. Thus,

$$P(\mathcal{OC}^{(d)}_{L,0})\binom{K+M-1}{K}\left(1+\frac{K}{M}\right) = P(\mathcal{OC}^{(d)}_{L,0})\binom{K+M}{K}.$$

A direct interpretation of the last formula is as follows: Fix the root edge. Mark one of the unary nodes of the root edge. Like this, we create as many instances (with different markers) as there are unary nodes on the root edge in all configurations, i.e., we count the total number of unary nodes on the root edge. We model this marker, by splitting the root edge just after this marker into two parts. Hence, we have now $M + 1$ edges to distribute K unary nodes. We interpret the case when the first part of the root edge is empty, as the $\binom{K+M-1}{K}$ created weighted trees from placing unary nodes. Thus, there are $\binom{K+M}{K}$ many choices.

Third, we assign weights for the new unary nodes. Consider an original edge with i unary nodes. To give the lowest one its correct weight, we remove the original node just below and replace it by the lowest unary node. Then, we may use the idea of the previous step: The weight of the edge above is now i, as $i - 1$ unary nodes remain. In the superposition, each such sequence of i unary nodes appears on each original edge the same number of times. Hence, we collect a weight i on each edge, and the total weight gives the correct weights for all lowest unary nodes. Then, we repeat this process for the

next unary node, with an edge weight $i-1$, etc. In total, an original edge with i unary nodes gives rise to a total weight of $1+2+\cdots+i$. For an arbitrary edge, this is now, analogously to before, equal to the sum of unary nodes that are before the marked node from the previous step. Thus, this gives a factor $K/(M+1)$ and we get

$$P(\mathcal{OC}_{L,0}^{(d)})\binom{K+M}{K}\left(1+\frac{K}{M+1}\right)=P(\mathcal{OC}_{L,0}^{(d)})\binom{K+M+1}{K}.$$

Note, as before, this has a direct combinatorial interpretation, where we split the first part of the root edge again. Hence, there are now $M+2$ edges to distribute K unary nodes, as claimed. Moreover, note that the last formula also gives the path length of unary binary trees with L labeled leaves and K unary nodes ($M=2L-1$); compare with Remark 2.8.

Using the last proposition, we are now ready to complete our analysis. Let $P_n^{(d)}$ be the path length of the top tree component of a random one-component tree-child network with n leaves. Then,

$$\mathbb{E}(P_n^{(d)})=\frac{\sum_{k=0}^{n-1}\binom{n}{k}P(\mathcal{OC}_{n,k}^{(d)})}{\mathrm{OTC}_n^{(d)}}.$$

Applying the Laplace method to the numerator and using Theorem 1.5 gives the following result.

Proposition 2.10. *(i) For $d=2$ (bicombining case), we have*

$$\mathbb{E}(P_n^{(2)})\sim 2\sqrt{\pi}n^{7/4}.$$

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

$$\mathbb{E}(P_n^{(3)})\sim\frac{9(\cosh(2)-I_0(2))}{2I_1(2)}\,n^2,$$

where the constant is approximately 4.19438713.

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

$$\mathbb{E}(P_n^{(d)})\sim\frac{d^2}{2}\,n^2.$$

Proof. We start with the bicombining case ($d=2$). Using the result from the Proposition 2.7 yields

$$\begin{aligned}\sum_{k=0}^{n-1}\binom{n}{k}P(\mathcal{OC}_{n,k}^{(2)})&=\sum_{k=0}^{n-1}\binom{n}{k+1}P(\mathcal{OC}_{n,n-1-k}^{(2)})\\&=\sum_{k=0}^{n-1}\binom{n}{k+1}\frac{(2n)!}{2^{n-1-k}(2k+2)!}\left(2^{k+1}(k+1)!-\frac{(2k+1)!}{2^k k!}\right).\end{aligned}$$

We break the last sum S into two sums, i.e., $S=S_1+S_2$ according to the two terms in the bracket. Thus,

$$S_1=\frac{n!(2n)!}{2^{n-2}}\sum_{k=0}^{n-1}\frac{4^k}{(2k+2)!(n-1-k)!}$$

and we have a similar expression for S_2. Note that the terms inside the sum increase until a positive integer k^* with $k^*=\sqrt{n}+\mathcal{O}(1)$ and decrease afterwards. Moreover, by using Stirling's formula, we see that

$$\frac{4^k}{(2k+2)!(n-1-k)!}=\frac{1}{8\pi\sqrt{2e}}n^{-3/4}e^{2\sqrt{n}}e^n n^{-n}e^{-x^2/\sqrt{n}}\left(1+\mathcal{O}\left(\frac{1+|x|}{\sqrt{n}}+\frac{x^3}{n}\right)\right)$$

uniformly for $|x| \leq n^{3/10}$ where $k = \sqrt{n}+x$. Consequently, from a standard application of the Laplace method:

$$\sum_{k=0}^{n-1} \frac{4^k}{(2k+2)!(n-1-k)!} \sim \frac{1}{8\pi\sqrt{2e}} n^{-3/4} e^{2\sqrt{n}} e^n n^{-n} \int_{-\infty}^{\infty} e^{-x^2/\sqrt{n}} \mathrm{d}x$$
$$= \frac{1}{8\sqrt{2e\pi}} n^{-1/2} e^{2\sqrt{n}} e^n n^{-n}$$

and thus,

$$S_1 \sim \frac{n!(2n)!}{2^{n-2}} \cdot \frac{1}{8\sqrt{2e\pi}} n^{-1/2} e^{2\sqrt{n}} e^n n^{-n} \sim \frac{1}{2\sqrt{e}} (2n)! 2^{-n} e^{2\sqrt{n}},$$

where we again used Stirling's formula. Similarly, we can derive the asymptotics of S_2 which shows that S_2 is of a smaller asymptotic order, i.e., $S_2 = o(S_1)$. Consequently, $S \sim S_1$. Finally, dividing by the asymptotics of $\mathrm{OTC}_n^{(2)}$ from part (i) of Theorem 1.5 and using (once more) Stirling's formula gives the claimed result.

Next, for $d = 3$, we first note that

$$\binom{n}{k} P(\mathcal{OC}_{n,k}^{(3)}) = \binom{n}{k} \frac{(2n+k)!}{6^k (2n-2k)!} \left(2^{n-k}(n-k)! - \frac{(2n-2k-1)!}{2^{n-k-1}(n-k-1)!} \right)$$

is increasing in k with $0 \leq k \leq n-1$. By replacing k by $n-1-k$ and using Stirling's formula, we obtain that

$$\binom{n}{k+1} P(\mathcal{OC}_{n,n-1-k}^{(3)}) = \left(\frac{4^{k+1}}{(2k+2)!} - \frac{1}{(k+1)!^2} \right) \frac{(3n)!}{6^n} \left(1 + \mathcal{O}\left(\frac{1+k^2}{n} \right) \right),$$

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

$$\sum_{k=0}^{n-1} \binom{n}{k+1} P(\mathcal{OC}_{n,n-1-k}^{(3)}) \sim \left(\sum_{k \geq 1} \frac{4^k}{(2k)!} - \frac{1}{k!^2} \right) \frac{(3n)!}{6^n}$$
$$= \big(\cosh(2) - I_0(2) \big) \frac{(3n)!}{6^n},$$

where $I_0(2)$ is the modified Bessel function (see Theorem 1.5, (ii)). Dividing by the asymptotics of $\mathrm{OTC}_n^{(3)}$ (again see Theorem 1.5, (ii)), we have

$$\mathbb{E}(P_n^{(3)}) \sim \frac{9(\cosh(2) - I_0(2))}{2I_1(2)} n^2.$$

This proves the claim in this case. (Note the similarity of this proof to the one of Theorem 1.6, (ii).)

For $d \geq 4$, with similar arguments as in the proof of part (iii) of Theorem 1.6:

$$\sum_{k=0}^{n-1} \binom{n}{k+1} P(\mathcal{OC}_{n,n-1-k}^{(d)}) \sim n\, P(\mathcal{OC}_{n,n-1}^{(d)}) = \frac{n\,(dn-d+2)!}{2\,(d!)^{n-1}}.$$

Dividing by the asymptotics of $\mathrm{OTC}_n^{(d)}$ from part (iii) in Theorem 1.6, we have

$$\mathbb{E}(P_n^{(d)}) \sim \frac{d^2}{2} n^2,$$

which proves the claimed result also in this case. □

Remark 2.11. A (slightly) weaker version of the above result for the bicombining case was derived in [20, Proposition 2].

Denote by $S_n^{(d)}$ the Sackin index of a random one-component tree-child network with n leaves. Then, by combining the last proposition with Lemma 2.6, we obtain the main result of this subsection. (This result for $d = 2$ was also the main result of [20].)

Theorem 2.12. *(i) For $d = 2$ (bicombining case), we have*

$$\mathbb{E}(S_n^{(d)}) = \Theta(n^{7/4}).$$

(ii) For $d \geq 3$, we have

$$\mathbb{E}(S_n^{(d)}) = \Theta(n^2).$$

3 General Networks

In this section, we will consider general d-combining tree-child networks.

3.1 Encoding Networks by Words

We will start with a formula for $\mathrm{TC}_{n,k}^{(d)}$ which is not as explicit as the formula for $\mathrm{OTC}_{n,k}^{(d)}$ from Theorem 2.1. (It will, however, lead to an efficient recursive way of computing $\mathrm{TC}_{n,k}^{(d)}$.) This formula contains the number of certain words, which we describe next.

Definition 3.1. *Let $\mathcal{C}_{n,k}^{(d)}$ denote the class of words consisting of the letters $\{\omega_1, \ldots, \omega_n\}$ in which k letters occur $d+1$ times and $n-k$ letters occur 2 times and which satisfy the following condition: In every prefix of a word, either a letter has not occurred more than $d-2$ times, 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 $j > i$. Here, for the letters appearing only 2 times, we treat the 0th, 1st, and 2nd occurrence as the $(d-1)$st, dth, and $(d+1)$st occurrence, respectively.*

Remark 3.2. For $k = n$, we recover the words from [4, Definition 12] which in turn generalized the words from [13, Definition 2] from the bicombining case.

Remark 3.3. The words $\mathcal{C}_{n,k}^{(d)}$ can also be encoded by Young tableaux with walls, where a wall between two cells indicates that there are no order constraints between the respective entries; see [1, 2]. The ith column is associated to the ith letter ω_i and its size is equal to the number of occurrences of ω_i. Therefore, the corresponding Young tableaux consist of k columns with $d+1$ cells and $n-k$ columns with 2 cells placed next to each other in such a way that the top cells are all side-by-side. We put vertical walls between all cells in rows three to $d+1$. Finally, the cells are filled in increasing order from left to right and bottom to top with the numbers $\{1, 2, \ldots, 2n + k(d-1)\}$. The bijection is as follows: Read the words from left to right. A letter ω_i at position j indicates that the value j is put into column i.

This generalizes the class analyzed in [2, Section 4] consisting of only one row with walls. Our asymptotic counting result for $\mathcal{C}_{n,n}^{(d)}$ (Theorem 1.8) gives then directly the respective result for rectangular Young tableaux with vertical walls in all but the first two rows of shape $(d+1) \times n$.

The next result connects tree-child networks and the words from Definition 3.1.

Theorem 3.4. *Let $c_{n,k}^{(d)} := |\mathcal{C}_{n,k}^{(d)}|$. Then,*

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

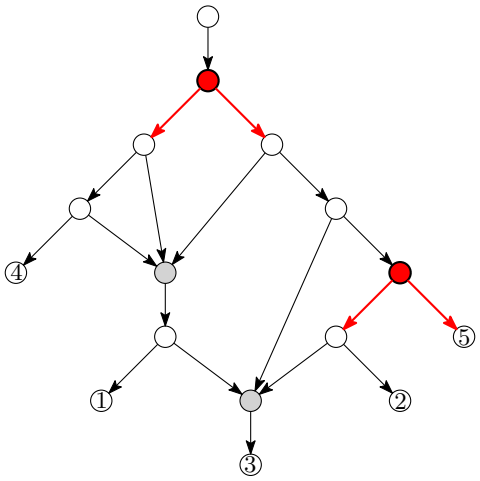

Figure 4: A 3-combining tree-child network with 5 leaves and 2 reticulation nodes (gray nodes). Thus, the number of free tree nodes is 2 (red nodes) and there are 4 free edges (red edges).

Remark 3.5. For $d = 2$, [16] proposed an encoding of tree-child networks with n leaves and k reticulation nodes by a (slightly) different class of words. This encoding led to a similar formula for their numbers. However, whereas the formula from [16] is just a conjecture, we can provide a rigorous proof of our result.

Before proving the above theorem, we recall some concepts and provide generalizations of results from [13]. First, we call a tree node in a tree-child network *free* if both its children are not reticulation nodes. The edges to the two children of a free tree node are called *free edges*.

Lemma 3.6. *A tree-child network with n leaves and k reticulation nodes has exactly $n-k-1$ free tree nodes and thus $2(n-k-1)$ free edges.*

Note that exactly the same result holds in the bicombining case; see [13, Lemma 1]. For an example demonstrating the last lemma see Figure 4.

Proof. By (1), the number of tree nodes in a tree-child network with n leaves and k reticulation nodes equals $n + (d-1)k - 1$. Now, observe that the d parents of a reticulation node are (i) different for all reticulation nodes (because of the tree-child property); (ii) tree nodes which are not free; and (iii) *all* non-free tree nodes. Thus, the number of free tree nodes equals

$$n + (d-1)k - 1 - dk = n - k - 1. \qquad \square$$

We next recall the structure of *maximally reticulated tree-child networks*, i.e., tree-child networks with n leaves and $n-1$ reticulation nodes, which was established in [13, Lemma 4] for the bicombining case. It also carries over to the d-combining case: Every maximally reticulated tree-child network admits a (unique) decomposition into *path-components*, which are maximal paths that start at a node and end at a leaf, where all its intermediate nodes are tree nodes. (See Figure 4 in [13] for an example.)

Now, we are ready to prove Theorem 3.4.

Proof of Theorem 3.4. Let N be a tree-child network with n leaves and k reticulation nodes. (We take the network from Figure 4 as our running example.) From Lemma 3.6, we know that N has $n-k-1$ free tree nodes, which each have 2 free edges. We choose one free edge from each of these pairs of free edges. Then, the number of all networks N together with choices of free edges equals

$$2^{n-k-1}\mathrm{TC}_{n,k}^{(d)} = n!c_{n-1,k}^{(d)}, \tag{13}$$

where the equality with the right-hand side is our claim. (See Figure 5-(a) for the network from Figure 4 and its 4 choices of free edges.) Thus, in order to prove the claim it suffices to find a bijection between the networks N and a choice of free edges to tuples consisting of a permutation and a word from $\mathcal{C}^{(d)}_{n-1,k}$.

In order to explain this bijection, fix a network N and a choice of free edges; see Figure 5-(a). For every free tree node, insert $d-1$ nodes on its chosen free edge and a single node on its other free edge. Connect the $d-1$ nodes to the single node, thereby turning this single node into a new reticulation node. Notice that the resulting network is a maximally reticulated tree-child network; see Figure 5-(b). The rest of the proof proceeds now as the proof of [13, Proposition 2].

First, we index the path-components as follows: the path-component containing the root gets index 0. Consider all other path-components (which must start with a reticulation node) with parents of the reticulation node already in indexed path components. Index them according to the largest index of the path-component which contains the parents, or, in case of equal largest indices according to whose last parent is encountered first when reading the nodes in the path component from the starting node to the leaf. Repeat this until all path-components have been indexed; see Figure 5-(c). Note that one path-component starts with the root, and $n-1$ path-components start with a reticulation node.

Now, label the reticulation node and all its parents of the path-component with index 1 by a, of the path component with index 2 by b, etc. Next, for each chosen free edge, treat the added $d-1$ nodes (which all have the same label) as a single node; see Figure 5-(c). Then, a word from $\mathcal{C}^{(d)}_{n-1,k}$ is obtained by reading the labels of nodes of the path-components in increasing order. Finally, a permutation is obtained by recording the labels of each leaf of the path-components when read in the above order; see Figure 5-(d).

Overall, this gives a bijection between N and a fixed choice of free edges for every free tree node to a word from $\mathcal{C}^{(d)}_{n-1,k}$ and a permutation of length n. Thus, we have proved (13) and the claim. □

Theorem 3.4 reduces the problem of counting $\mathrm{TC}^{(d)}_{n,k}$ to that of $c^{(d)}_{n,k}$. For the latter sequence, we have the following relation to the sequence $b^{(d)}_{n,k,m}$ which can be computed recursively.

Proposition 3.7. *Let* $c^{(d)}_{n,k} := |\mathcal{C}^{(d)}_{n,k}|$. *Then,*

$$c^{(d)}_{n,k} = \sum_{m\geq 1} b^{(d)}_{n,k,m},$$

where $b^{(d)}_{n,k,m}$ $(1 \leq m \leq n, 0 \leq k \leq n)$ *can be recursively computed as*

$$b^{(d)}_{n,k,m} = \sum_{j=1}^{m} b^{(d)}_{n-1,k,j} + \binom{n+m+k(d-1)-2}{d-1} \sum_{j=1}^{m} b^{(d)}_{n-1,k-1,j}, \qquad (n \geq 2) \tag{14}$$

with initial conditions $b^{(d)}_{n,k,m} = 0$ *for* $n < m$ *of* $n < k$, $b_{n,-1,m} = 0$ *and* $b^{(d)}_{1,0,1} = b^{(d)}_{1,1,1} = 1$.

Proof. First, note that because of the condition the words from $\mathcal{C}^{(d)}_{n,k}$ have to satisfy (see Definition 3.1), any word from $\mathcal{C}^{(d)}_{n,k}$ has a suffix $\omega_n\omega_m\omega_{m+1}\cdots\omega_{n-1}\omega_n$ with $1 \leq m \leq n$. Denote by $b^{(d)}_{n,k,m}$ the number of these words. We now consider two cases.

First, we assume that ω_n is a letter which occurs twice. Then, removing the 2 occurrences of ω_n from these words gives a word of $\mathcal{C}^{(d)}_{n-1,k}$ with suffix $\omega_m\omega_{m+1}\cdots\omega_{n-1}$, i.e., it has a suffix $\omega_{n-1}\omega_j\omega_{j+1}\ldots\omega_{n-1}$ for $j = 1,\ldots,m$. Reversing this procedure gives the contribution

$$\sum_{j=1}^{m} b^{(d)}_{n-1,k,j} \tag{15}$$

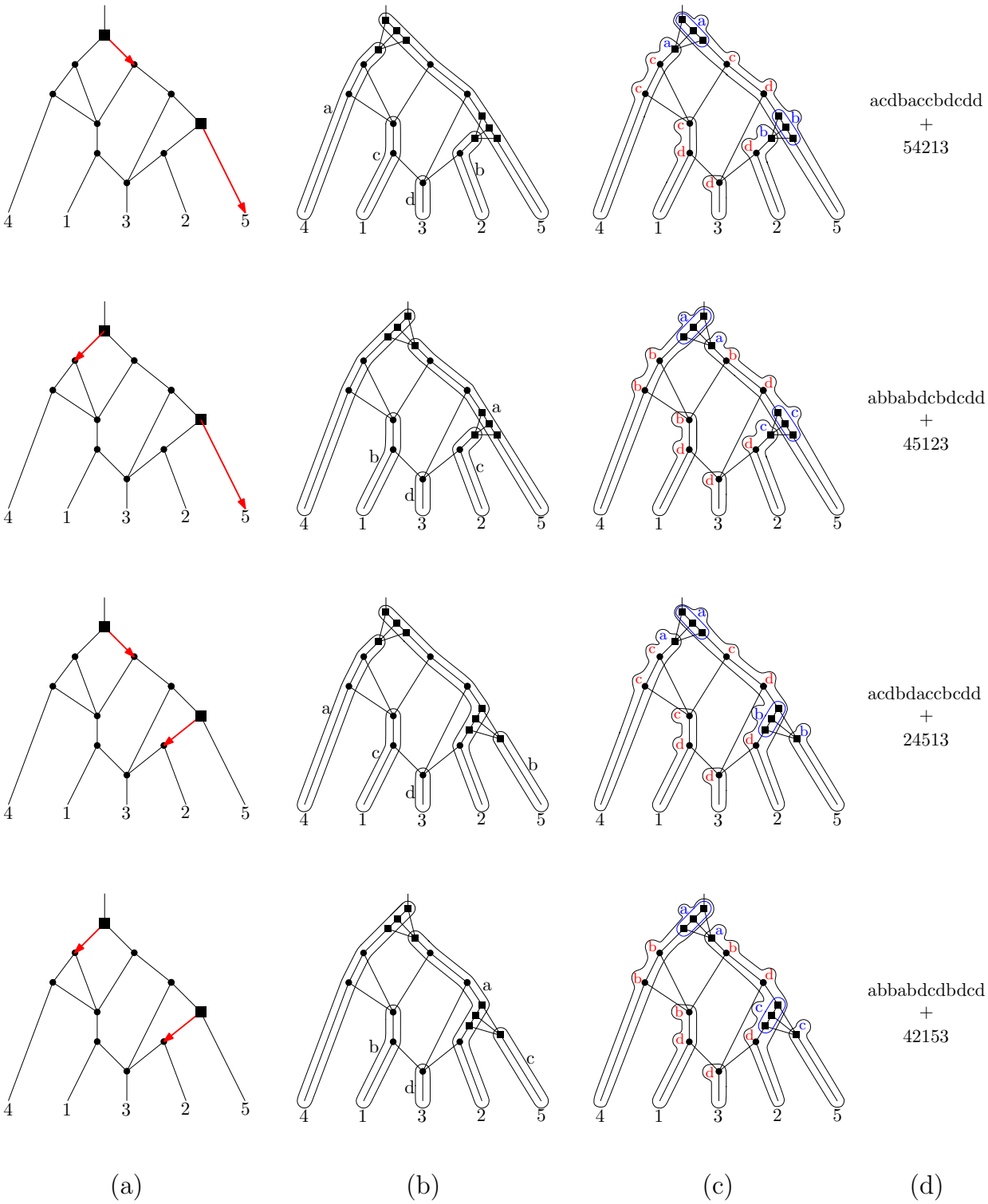

Figure 5: (a) The network from Figure 4 together with the 4 possible ways of choosing an outgoing free edge for every free tree node; (b) Replacing each free tree node by a reticulation node which results in a maximally reticulated tree-child network whose path-components are indexed; (c) Labeling all internal nodes by labeling reticulation nodes and their parents with the label of their path-component. Note that all nodes on the chosen free edges only receive one label; (d) The word from $\mathcal{C}_{4,2}^{(3)}$ and the permutation corresponding to each network.

to $b^{(d)}_{n,k,m}$, which is the first term on the right-hand side of (14).

Second, if ω_n is a letter which occurs $d+1$ times, we remove the $d+1$ occurrences of this letter. Then, with the same line of reasoning as above, we obtain the second term of the right-hand side of (14):

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

where the binomial coefficient counts the number of ways of adding back the $d-1$ occurrences of ω_n after two ω_n's have been added, one before the last ω_m and one at the end of the word. By Definition 3.1 these first $d-1$ occurrences of ω_n may be anywhere. □

The recurrence from the above proposition combined with Theorem 3.4 allows one to compute values of $\mathrm{TC}^{(d)}_{n,k}$ for small values of n, k, d; see Appendix A. Also, like this, we can recover the table for $\mathrm{TC}^{(2)}_{n,k}$ from [3] which was computed in that paper with a much more computation-intensive approach (which will be explained in Appendix B).

3.2 Asymptotic Counting

In this section, we will prove Theorem 1.8, namely, we will derive an asymptotic result for $\mathrm{TC}^{(d)}_n$. Recall that

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

The first main observation is that the last term in this sum dominates; see the first equality in (2). More precisely, we have the following.

Lemma 3.8. *For $0 \leq k \leq n-2$, we have*

$$\mathrm{TC}^{(d)}_{n,k} \leq \frac{1}{2(n-k-1)}\mathrm{TC}^{(d)}_{n,k+1} \tag{16}$$

and consequently,

$$\mathrm{TC}^{(d)}_n = \Theta\left(\mathrm{TC}^{(d)}_{n,n-1}\right).$$

Proof. Let N be a tree-child network with n leaves and k reticulation nodes. Recall that N has $2(n-k-1)$ free edges; see Lemma 3.6.

We can construct tree-child networks with n leaves and $k+1$ reticulation nodes from N by (i) inserting $d-1$ tree nodes into the root edge of N and a reticulation node into a free edge and (ii) connecting the $d-1$ new tree nodes to the new reticulation node. Note that each network built in this way is different. Thus,

$$2(n-k-1)\mathrm{TC}^{(d)}_{n,k} \leq \mathrm{TC}^{(d)}_{n,k+1},$$

which implies the first claim.

Next, by iteration of (16), we obtain

$$\mathrm{TC}^{(d)}_{n,k} \leq \frac{1}{2^{n-k-1}(n-k-1)!}\mathrm{TC}^{(d)}_{n,n-1} \tag{17}$$

and thus,

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

which proves the second claim. □

Remark 3.9. Note that for $d = 2$ and $k = n - 2$, equality holds in (16) because in this case, (a) the networks constructed from N in the above proof are maximally reticulated and (b) the child of the root of each maximally reticulated network is not free. Thus, the construction from the proof is reversible and we have a bijection. (This was first proved in [3, Proposition 17].)

As a consequence of the last result, we can now entirely concentrate on the maximal reticulated case for which we obtain from Theorem 3.4:

$$\mathrm{TC}^{(d)}_{n,n-1} = n! c^{(d)}_{n-1},$$

where we have set $c^{(d)}_n := c^{(d)}_{n,n}$. By Proposition 3.7, this sequence satisfies

$$c^{(d)}_n = \sum_{m \geq 1} b^{(d)}_{n,m},$$

where $b^{(d)}_{n,m} := b^{(d)}_{n,n,m}$ satisfies

$$b^{(d)}_{n,m} = \binom{m + nd - 2}{d - 1} \sum_{j=1}^{m} b^{(d)}_{n-1,j}. \tag{18}$$

The recurrence (18) can be brought in a slightly easier form.

Lemma 3.10. *We have,*

$$b^{(d)}_{n,m} = \frac{dn + m - 2}{dn + m - d - 1} b^{(d)}_{n,m-1} + \binom{dn + m - 2}{d - 1} b^{(d)}_{n-1,m}, \qquad (n \geq 2, 0 \leq m \leq n) \tag{19}$$

with initial conditons $b^{(d)}_{1,1} = 1$ *and* $b^{(d)}_{n,m} = 0$ *for (i)* $n \geq 2$ *and* $m = -1$*; (ii)* $n = 1$ *and* $m = 0$*; and (iii)* $n < m$*.*

Proof. The recursive structure in (18) yields

$$\frac{b^{(d)}_{n,m}}{\binom{dn+m-2}{d-1}} - \frac{b^{(d)}_{n,m-1}}{\binom{dn+m-3}{d-1}} = b^{(d)}_{n-1,m}.$$

This gives the claimed recurrence and the initial conditions are easily checked. □

To this recurrence, we apply now the method from [6]. Due to the similarities, we will only discuss the main differences. We start with the following transformation of $(b^{(d)}_{n,m})_{0 \leq m \leq n}$ to $(e^{(d)}_{i,j})_{\substack{0 \leq i \leq j \\ i-j \text{ even}}}$, which changes the indices and captures the exponential and superexponential terms coming from the binomial coefficient in (19).

Lemma 3.11. *We have*

$$b^{(d)}_{n,m} = \lambda(d)^n (n!)^{d-1} e^{(d)}_{n+m,n-m} \qquad \textit{with} \qquad \lambda(d) = \frac{(d+1)^{d-1}}{(d-1)!},$$

where $e^{(d)}_{n,m}$ *satisfies the following recurrence*

$$e^{(d)}_{n,m} = \mu^{(d)}_{n,m} e^{(d)}_{n-1,m+1} + \nu^{(d)}_{n,m} e^{(d)}_{n-1,m-1} \tag{20}$$

with

$$\mu^{(d)}_{n,m} = 1 + \frac{2(d-1)}{(d+1)n + (d-1)m - 2(d+1)} \quad \textit{and} \quad \nu^{(d)}_{n,m} = \prod_{i=2}^{d} \left(1 - \frac{2(m+i)}{(d+1)(n+m)}\right)$$

for $n \geq 3$ *and* $m \geq 0$*, where* $e^{(d)}_{n,-1} = e^{(d)}_{2,n} = 0$ *except for* $e^{(d)}_{2,0} = 1/\lambda(d)$*.*

Now, we are interested in

$$e^{(d)}_{2n,0} = \frac{b^{(d)}_{n,n}}{\lambda(d)^n (n!)^{d-1}}$$

because by the previous lemmas and (18) we have

$$\mathrm{TC}^{(d)}_n = \Theta\left(\mathrm{TC}^{(d)}_{n,n-1}\right) = \Theta\left(n! c^{(d)}_{n-1}\right) = \Theta\left(n!\, n^{1-d} b^{(d)}_{n,n}\right) = \Theta\left((n!)^d \lambda(d)^n n^{1-d} e^{(d)}_{2n,0}\right). \tag{21}$$

Moreover, observe that for the Theta-result the initial value of $e^{(d)}_{2,0}$ is irrelevant, as it creates only a constant factor. So we may set it to $e^{(d)}_{2,0} = 1$, or any convenient constant. Note that this recurrence is very similar to that of relaxed trees [6, Equation (2)], yet with more complicated factors. Observe also that this is exactly recurrence [13, Equation (10)] for $d = 2$.

Motivated by experiments for large n, we use the following ansatz

$$e^{(d)}_{n,m} \approx h(n) f\left(\frac{m+1}{n^{1/3}}\right),$$

where h and f are some "regular" functions. Next, we substitute $s(n) = h(n)/h(n-1)$ and $m = \kappa n^{1/3} - 1$ into (20). Then, for $n \to \infty$ we get the expansion

$$f(\kappa)s(n) = 2f(\kappa) + \left(f''(\kappa) - \frac{2(d-1)}{d+1}\kappa f(\kappa)\right) n^{-2/3} + \mathcal{O}\left(n^{-1}\right).$$

Hence, we may assume that

$$s(n) = 2 + c_1 n^{-2/3} + c_2 n^{-1} + \dots$$

and this implies that $f(\kappa)$ satisfies the differential equation

$$f''(\kappa) = \left(c_1 + \frac{2(d-1)}{d+1}\kappa\right) f(\kappa)$$

that is solved by the Airy function Ai of the first kind, as we have $e_{n,m} = 0$ for $m > n$ which corresponds to $\lim_{x\to\infty} f(x) = 0$. Additionally, the boundary conditions allow to compute c_1 and we get that

$$f(\kappa) = C\,\mathrm{Ai}\left(a_1 + B^{1/3}\kappa\right) \qquad \text{where} \qquad B := \frac{2(d-1)}{d+1}, \tag{22}$$

$a_1 \approx 2.338$ is the largest root of the Airy function Ai, and C is an arbitrary constant. From this we get that $c_1 = a_1 B^{1/3}$. These heuristic arguments guide us to the following results. The proofs are analogous to [5, 6, 13]; for the details we refer to the accompanying Maple worksheet [18]. Note that the next two results generalize [13, Propositions 4 and 5], whose results are recovered by setting $d = 2$.

Proposition 3.12. *For all $n, m \geq 0$ let*

$$\tilde{X}_{n,m} := \left(1 - \frac{2d-1}{3(d+1)}\frac{m^2}{n} - \frac{3d^2+12d-11}{6(d+1)}\frac{m}{n}\right) \mathrm{Ai}\left(a_1 + \frac{B^{1/3}(m+1)}{n^{1/3}}\right) \quad \textit{and}$$

$$\tilde{s}_n := 2 + \frac{a_1 B^{2/3}}{n^{2/3}} - \frac{3d^2-5d+4}{3(d+1)n} - \frac{1}{n^{7/6}}.$$

Then, for any $\varepsilon > 0$, there exists an $\tilde{n}_0$ such that

$$\tilde{X}_{n,m}\tilde{s}_n \leq \mu^{(d)}_{n,m}\tilde{X}_{n-1,m+1} + \nu^{(d)}_{n,m}\tilde{X}_{n-1,m-1}$$

for all $n \geq \tilde{n}_0$ and for all $0 \leq m < n^{2/3-\varepsilon}$, where $\mu^{(d)}_{n,m}$ and $\nu^{(d)}_{n,m}$ are as in Lemma 3.11.

Proposition 3.13. *Choose $\eta > \frac{(2d-1)^2}{18(d+1)^2}$ fixed and for all $n, m \geq 0$ let*

$$\hat{X}_{n,m} := \left(1 - \frac{2d-1}{3(d+1)}\frac{m^2}{n} - \frac{3d^2+12d-11}{6(d+1)}\frac{m}{n} + \eta\frac{m^4}{n^2}\right) \mathrm{Ai}\left(a_1 + \frac{B^{1/3}(m+1)}{n^{1/3}}\right) \quad \textit{and}$$

$$\hat{s}_n := 2 + \frac{a_1 B^{2/3}}{n^{2/3}} - \frac{3d^2-5d+4}{3(d+1)n} + \frac{1}{n^{7/6}}.$$

Then, for any $\varepsilon > 0$, there exists a constant $\hat{n}_0$ such that

$$\hat{X}_{n,m}\hat{s}_n \geq \mu^{(d)}_{n,m}\hat{X}_{n-1,m+1} + \nu^{(d)}_{n,m}\hat{X}_{n-1,m-1}$$

for all $n \geq \hat{n}_0$ and all $0 \leq m < n^{1-\varepsilon}$.

Proof of Propositions 3.12 and 3.13. The proofs follow nearly verbatim the steps of [6, Lemmas 4.2 and 4.4] as well as [5, Lemmas 7 and 9]. For the convenience of the reader, we repeat the main steps for Proposition 3.12 and point out the main differences. Full details of all computations are given in our accompanying Maple worksheet [18].

We start by defining the following sequence

$$P_{n,m} := -Z_{n,m}s_n + \mu^{(d)}_{n,m}Z_{n-1,m+1} + \nu^{(d)}_{n,m}Z_{n-1,m-1},$$

where we use the ansatz

$$s_n := \sigma_0 + \frac{\sigma_1}{n^{1/3}} + \frac{\sigma_2}{n^{2/3}} + \frac{\sigma_3}{n} + \frac{\sigma_4}{n^{7/6}}$$

$$Z_{n,m} := \left(1 + \frac{\tau_2 m^2 + \tau_1 m}{n}\right) \mathrm{Ai}\left(a_1 + \frac{B^{1/3}(m+1)}{n^{1/3}}\right),$$

with parameters $\sigma_i, \tau_j \in \mathbb{R}$, and B from (22). Note that in [5,6] we had $B = 2$, whereas here B depends on d. Then the claimed inequality is equivalent to $P_{n,m} \geq 0$ with the parameters chosen accordingly.

To prove it, we expand the Airy function $\mathrm{Ai}(z)$ in a neighborhood of

$$\alpha = a_1 + \frac{B^{1/3}m}{n^{1/3}}, \tag{23}$$

and, due to the defining differential equation $\mathrm{Ai}''(x) = x\mathrm{Ai}(x)$, we get the following expansion

$$P_{n,m} = p_{n,m}\mathrm{Ai}(\alpha) + p'_{n,m}\mathrm{Ai}'(\alpha),$$

where $p_{n,m}$ and $p'_{n,m}$ are functions of m and n^{-1} and may be expanded as power series in $n^{-1/6}$ whose coefficients are polynomials in m. As the Airy function is entire, these series converge absolutely for $n > 1$ and $m < n$.

Now we proceed with the technical analysis. First, we show that $[m^i n^j]P_{n,m} = 0$ for $i + j > 1$, $i, j \in \mathbb{Q}$. We omit this step, as it is analogous to the previous cases. Second, we use computer algebra to strengthen this result by choosing suitable values σ_i and τ_j to eliminate more terms; see Figure 6. A solid diamond at (i, j) indicates that the coefficient $[m^i n^j]P_{n,m}$ is non-zero for generic values of σ_i and τ_j; an empty diamond indicates that the specific choice of σ_i and τ_j makes it vanish. The convex hull is formed by the following three lines

$$L_1 : j = -\frac{7}{6} - \frac{7i}{18}, \qquad L_2 : j = -\frac{1}{3} - \frac{2i}{3}, \qquad L_3 : j = 1 - i.$$

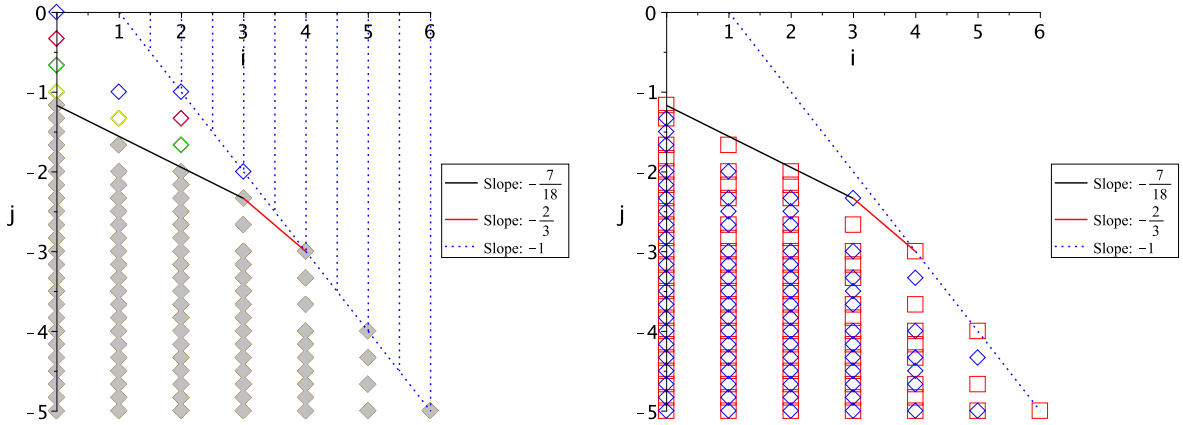

Figure 6: (Left) Non-zero coefficients of $P_{n,m} = \sum a_{i,j} m^i n^j$ shown by diamonds for $s_n := \sigma_0 + \frac{\sigma_1}{n^{1/3}} + \frac{\sigma_2}{n^{2/3}} + \frac{\sigma_3}{n} + \frac{\sigma_4}{n^{7/6}}$ and $Z_{n,m} := \left(1 + \frac{\tau_2 m^2 + \tau_1 m}{n}\right) \mathrm{Ai}\left(a_1 + \frac{2^{1/3}(m+1)}{n^{1/3}}\right)$. There are no terms in the blue dashed area. The blue terms vanish for $\sigma_0 = 2$, the red terms vanish for $\sigma_1 = 0$, the green terms vanish for $\sigma_2 = B^{2/3} a_1$, and the yellow terms vanish for $\sigma_3 = -\frac{3d^2-5d+4}{3(d+1)}$ and $\tau_2 = -\frac{2d-1}{3(d+1)}$. The black, red, and blue lines represent the parts L_1, L_2, and L_3, respectively, of the convex hull. (Right) The solid gray diamonds are decomposed into the coefficients $p_{n,m}$ of $\mathrm{Ai}(\alpha)$ (red boxes) and $p'_{n,m}$ of $\mathrm{Ai}'(\alpha)$ (blue diamonds).

In a final step, we distinguish between $p_{n,m}$ and $p'_{n,m}$; see Figure 7. The expansions for n tending to infinity start as follows, where the elements on the convex hull are written in color:

$$P_{n,m} = \mathrm{Ai}(\alpha)\left(-\frac{\sigma_4}{n^{7/6}} - \frac{B^{5/3} a_1 m}{3n^{5/3}} - \frac{(23d^2-14d+5)m^2}{9(d+1)^2 n^2} - \frac{2(2d-1)(3d-1)B^{5/3} a_1 m^3}{9(d+1)^2 n^{8/3}} - \frac{(2d-1)(23d-9)Bm^4}{18(d+1)^2 n^3} + \frac{(2d-1)(209d^2-258d+129)Bm^5}{270(d+1)^3 n^4} + \dots\right) +$$
$$\mathrm{Ai}'(\alpha)\left(-\frac{B^{1/3}\sigma_4}{n^{3/2}} - \frac{4(d-2)Ba_1 m}{9(d+1)n^2} - \frac{2(9d^3+50d^2-67d+21)B^{1/3} m^2}{9(d+1)^2 n^{7/3}} - \frac{4(2d-1)^2 B^{1/3} m^3}{9(d+1)^2 n^{7/3}} - \frac{(2d-1)(5d-1)B^{4/3} m^4}{18(d+1)^2 n^{10/3}} - \frac{(2d-1)(119d^2+32d-51)B^{4/3} m^5}{270(d+1)^3 n^{13/3}} + \dots\right).$$

We now choose $\sigma_4 = -1$ which leads to a positive term $\mathrm{Ai}(\alpha) n^{-7/6}$. Next, for fixed (large) n we prove that for all m the dominant contributions in $P_{n,m}$ are positive. Motivated by Figures 6 and 7, we consider three different regimes:

$$\text{(i)} \quad m \le C n^{1/3}; \qquad \text{(ii)} \quad C n^{1/3} < m \le n^{7/18}; \qquad \text{(iii)} \quad n^{7/18} < m < n^{2/3-\epsilon}$$

for a suitable constant $C > 0$. This part is analogous to the one of [6, Lemma 4.2], which is why we omit the technical details. In the end we get that there exists an $N > 0$ such that all terms are positive for $n > N$ and all $m < n^{2/3}$, which ends the proof of Proposition 3.12.

The proof of Proposition 3.13 follows analogously. □

Proof of Theorem 1.8. Let us start with the lower bound.

We first define a sequence $X_{n,m} := \max\{\tilde{X}_{n,m}, 0\}$ which satisfies the inequality of Proposition 3.12 for *all* $m \le n$. Then, we define an explicit sequence $\tilde{h}_n := \tilde{s}_n \tilde{h}_{n-1}$ for $n > 0$ and $\tilde{h}_0 = \tilde{s}_0$. From

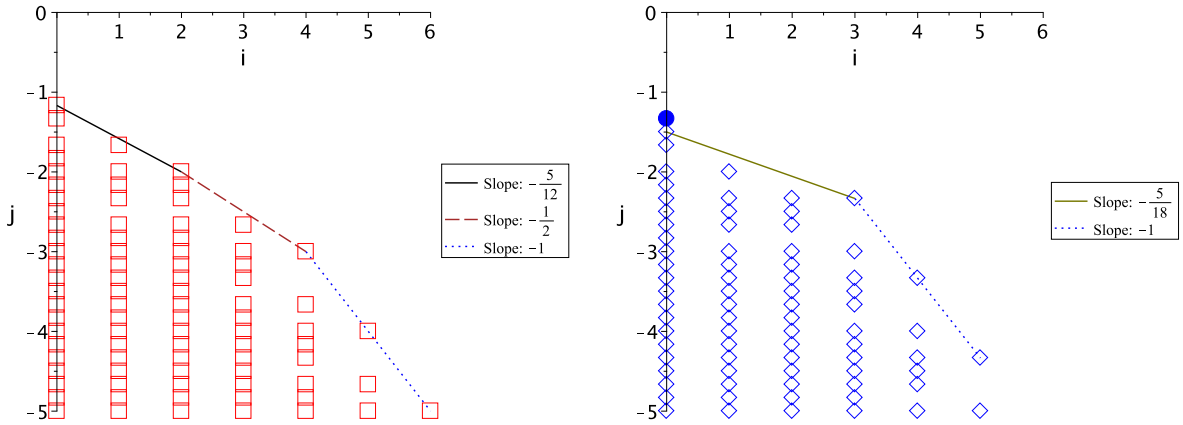

Figure 7: Non-zero coefficients $p_{n,m} = \sum \tilde{a}_{i,j} m^i n^j$ (red) and $p'_{n,m} = \sum \tilde{a}'_{i,j} m^i n^j$ (blue) of the expansion (23) for $P_{n,m}$. The coefficient of $n^{-4/3}$ in the right picture depicted as a solid blue circle disappears for $\tau_1 = \frac{3d^2+12d-11}{6(d+1)}$.

this, we get by induction that $e^{(d)}_{n,m} \geq C_0 \tilde{h}_n X_{n,m}$ for some constant $C_0 > 0$ and all $n \geq \tilde{n}_0$ and all $0 \leq m \leq n$. Hence,

$$\begin{aligned} e^{(d)}_{2n,0} &\geq C_0 \tilde{h}_{2n} X_{2n,0} \\ &\geq C_0 \prod_{i=1}^{2n} \left(2 + \frac{a_1 B^{2/3}}{i^{2/3}} - \frac{3d^2 - 5d + 4}{3(d+1)i} - \frac{1}{i^{7/6}}\right) \mathrm{Ai}\left(a_1 + \frac{B^{1/3}}{(2n)^{1/3}}\right) \\ &\geq C_1 4^n e^{3a_1 (B/2)^{2/3} n^{1/3}} n^{\frac{d^2+d-2}{2(d+1)}}. \end{aligned}$$

Finally, combining this with (21) we get the lower bound.

The upper bound is similar, yet more technical.

The starting point is Proposition 3.13 and a function $X_{n,m}$ that is valid for *all* $0 \leq m \leq n$. For this purpose we define a sequence $\hat{e}^{(d)}_{n,m}$ such that $\hat{e}^{(d)}_{n,m} := e^{(d)}_{n,m}$ for $0 \leq m \leq n^{1-\varepsilon}$ and $\hat{e}^{(d)}_{n,m} := 0$ otherwise; compare with [5, 6]. The missing key step is now to show that $e^{(d)}_{2n,0} = \mathcal{O}(\hat{e}^{(d)}_{2n,0})$. Combining this with the analogous computations performed for the lower bound we get

$$\hat{e}^{(d)}_{2n,0} \leq \hat{C}_1 4^n e^{3a_1 (B/2)^{2/3} n^{1/3}} n^{\frac{d^2+d-2}{2(d+1)}}.$$

To complete the prove we show $e^{(d)}_{2n,0} \leq 2\hat{e}^{(d)}_{2n,0}$ using lattice path theory and computer algebra. The argument will follow along the same lines as in [13, Appendix]. We start from Equation (20) of $e^{(d)}_{n,m}$, which we interpret as a recurrence counting lattice paths. They are composed of steps $(1,1)$ weighted by $\mu^{(d)}_{n,m}$ and $(1,-1)$ weighted by $\nu^{(d)}_{n,m}$ when the respective step ends at (n,m). The total weight of a path is the product of its weights. Now, we are interested in the paths never crossing $y = 0$ and ending at $(2n, 0)$. Let now $p_{\ell,k,2n}$ be the number of such paths starting at (ℓ, k) and ending at $(2n, 0)$. From (20) we directly get

$$p_{\ell,k,2n} = \mu^{(d)}_{\ell+1,k-1} p_{\ell+1,k-1,2n} + \nu^{(d)}_{\ell+1,k+1} p_{\ell+1,k+1,2n},$$

with $p_{\ell,-1,2n} = 0$ and $p_{2n,k,2n} = \delta_{k,0}$.

Now, as in [13] we are able to show that

$$\frac{p_{\ell,j,2n}}{(j+1)^2} \geq \frac{p_{\ell,k,2n}}{(k+1)^2}, \tag{24}$$

for integers $0 \leq j < k \leq \ell \leq 2n$ such that $2 \mid k - j$. For the technical details, using reverse induction on ℓ, we refer to our accompanying Maple worksheet [18].

Finally, from (24) we directly get

$$p_{2x,2y,2n} \leq (2y+1)^2 p_{2x,0,2n}, \tag{25}$$

which we need to apply [6, Lemma 4.6] together with the bound $e^{(d)}_{2x,2y} \leq \binom{2x}{x+y}$, which holds due to the same reasons as in [13]: combining the weights of up an down steps gives a weight less than one. This proves $e^{(d)}_{2n,0} \leq 2\hat{e}^{(d)}_{2n,0}$ and ends the proof of Theorem 1.8. □

Remark 3.14. Note that in [5, 6] a stronger result than (24) was proved, where the powers in the denominators are 1 instead of 2. However, any polynomial coefficient in (25) suffices to get the same result using [6, Lemma 4.6].

With the same strategy as in [5, 6], we could show the stronger result $p_{2x,2y,2n} \leq 2(2y+1)p_{2x,0,2n}$, however, then other technicalities arise: the value $j = 0$ and the range $0 \leq j < k \leq \ell \leq 2n$ have to be treated separately.

3.3 Number of Reticulation Nodes

This section will contain the proof of Theorem 1.9. Since the proof for $d = 2$ and $d \geq 3$ is different, we will split it into two subsections (Section 3.3.1 and Section 3.3.2 below). A final subsection will contain the proof of Corollary 1.10.

3.3.1 Bicombining Networks

In this section, we consider the case $d = 2$. For convenience, we drop the superindex in the notation.

We start with the following bounds for $\mathrm{TC}_{n,k}$.

Lemma 3.15. *For* $1 \leq k \leq n-1$,

$$\frac{n-k}{k(3n-k-3)}\mathrm{TC}_{n,n-k} \leq \mathrm{TC}_{n,n-1-k} \leq \frac{1}{2k}\mathrm{TC}_{n,n-k}. \tag{26}$$

Proof. The upper bound follows from (16).

For the lower bound, we generalize the argument from [13, Lemma 3]. Therefore, consider a tree-child network N with n leaves and $n-1-k$ reticulation nodes. From Lemma 3.6, we know that N has $2k$ free edges. Moreover, N has $3n-k-2$ edges which do not end in a reticulation node; see (1). Now, by inserting a node into a tree edge and connecting it to a node which is inserted into a free edge, we obtain at most $2k(3n-k-3)\mathrm{TC}_{n,n-1-k}$ tree-child networks with n leaves and $n-k$ reticulation nodes (as those with cycles have to be discarded). On the other hand, each network is created from a latter network exactly $2(n-k)$ times. Thus,

$$2(n-k)\mathrm{TC}_{n,n-k} \leq 2k(3n-k-3)\mathrm{TC}_{n,n-1-k}.$$

which proves the lower bound. □

From the above bounds, we deduce the following lemma.

Lemma 3.16. *We have,*

$$\frac{1}{3^k k!}(1+o(1))\mathrm{TC}_{n,n-1} \leq \mathrm{TC}_{n,n-1-k} \leq \frac{1}{2^k k!}\mathrm{TC}_{n,n-1}$$

uniformly in $k = o(\sqrt{n})$.

Proof. The upper bound follows from iterating the upper bound of (26); see also (17).

For the lower bound, observe that

$$\frac{n-k}{k(3n-k-3)} = \frac{1}{3k}\left(1+\mathcal{O}\left(\frac{k}{n}\right)\right).$$

Thus, by iterating the lower bound in (26):

$$\frac{1}{3^k k!}\left(1+\mathcal{O}\left(\frac{k}{n}\right)\right)^k \mathrm{TC}_{n,n-1} \leq \mathrm{TC}_{n,n-1-k}.$$

From this the result follows since for the indicated range of k, we have

$$\left(1+\mathcal{O}\left(\frac{k}{n}\right)\right)^k = 1+\mathcal{O}\left(\frac{k^2}{n}\right) = 1+o(1).$$

This concludes the proof of the lemma. □

We next denote by $\mathrm{F}_{n,k}$ resp. $\mathrm{NF}_{n,k}$ the number of tree-child networks with n leaves and k reticulation nodes whose child of the root is free resp. not free.

We start with an easy observation.

Lemma 3.17. *For* $1 \leq k \leq n-1$, *we have* $(2k)\mathrm{TC}_{n,n-1-k} = \mathrm{NF}_{n,n-k}$.

Proof. A tree-child network with n leaves and $n-1-k$ reticulation nodes has $2k$ free edges; see Lemma 3.6. Taking any of these free edges and the root edge, inserting nodes in both edges and connecting the node inserted into the root edge with the other node gives a tree-child network with n leaves and $n-k$ reticulation nodes that is not free. Moreover, this construction is clearly reversible. □

Remark 3.18. Note that $\mathrm{NF}_{n,n-1} = \mathrm{TC}_{n,n-1}$. Thus, for $k=1$, the above result shows that equality in (16) holds for $k = n-2$; compare with Remark 3.9.

Another easy observation is the following.

Lemma 3.19. *For* $1 \leq k \leq n-1$,

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

Proof. A tree-child network with n leaves and $n-1-k$ reticulation nodes whose child of the root is free has $2k$ free edges of which $2k-2$ are not the edges from the child of the root to its children. By picking one of the latter two edges, one of the remaining $2k-2$ edges, inserting nodes and connecting the former edge to the latter, we obtain a tree-child network with n leaves and $n-k$ reticulation nodes whose child of the root is again free. Conversely, every such network is obtained by this construction at most 2 times. Thus,

$$2(2k-2)\mathrm{F}_{n,n-1-k} \leq 2\mathrm{F}_{n,n-k}$$

or

$$\mathrm{F}_{n,n-1-k} \leq \frac{1}{2(k-1)}\mathrm{F}_{n,n-k}.$$

Iterating this gives the claimed result. □

The final result we need is the following.

Lemma 3.20. *We have,*

$$\mathrm{F}_{\mathrm{n,n-2}} = \mathcal{O}\left(\frac{\mathrm{TC}_{n,n-1}}{n^{2/3}}\right).$$

Proof. Let N be a network with n leaves and $n-2$ reticulation nodes whose child of the root is free. Note that the two words constructed from N in the proof of Theorem 3.4 both start with a and that this is the sole letter which occurs only twice. Conversely, all words with this property arise from networks with n leaves and $n-2$ reticulation nodes whose child of the root is free. Thus, with the same arguments as in the proof of Theorem 3.4, we have

$$\mathrm{F}_{n,n-2} = \frac{n!}{2} g_{n-1},$$

where g_{n-1} is the number of words in $\mathcal{C}_{n-1,n-2}$ which start with a and this is the sole letter which occurs twice. Next, with the same arguments as used in the proof of Proposition 3.7:

$$g_n = \sum_{m \geq 1} h_{n,m}$$

where

$$h_{n,m} = (n+m+n-4) \sum_{j=1}^{m} h_{n-1,j}.$$

We now apply to this sequence the same method as in the last section, where we only need an upper bound. This gives

$$g_n = \mathcal{O}\left(n! 12^n e^{a_1 (3n)^{1/3}} n^{-4/3}\right)$$

and thus,

$$\mathrm{F}_{n,n-2} = \mathcal{O}\left((n!)^2 12^n e^{a_1 (3n)^{1/3}} n^{-7/3}\right).$$

Comparing with the Theta-result for $\mathrm{TC}_{n,n-1}$ from Theorem 1.8 (which was also the main result of [13]) gives the claimed result. □

Now, we can prove the following proposition.

Proposition 3.21. *We have,*

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

uniformly for $0 \leq k \leq c \log n$ *where* $c = 1/(3\log(3/2))$.

Proof. First note that

$$\mathrm{TC}_{n,n-k} = \mathrm{NF}_{n,n-k} + \mathrm{F}_{n,n-k} = (2k)\mathrm{TC}_{n,n-1-k} + \mathrm{F}_{n,n-k}, \tag{27}$$

where we used Lemma 3.17.

Next, by using Lemma 3.19, Lemma 3.20 and the lower bound in Lemma 3.16, we obtain that

$$\mathrm{F}_{n,n-k} = \mathcal{O}\left(\frac{\mathrm{F}_{n,n-2}}{2^k(k-2)!}\right) = \mathcal{O}\left(\frac{\mathrm{TC}_{n,n-1}}{2^k(k-2)!n^{2/3}}\right) = \mathcal{O}\left(\left(\frac{3}{2}\right)^k \frac{k}{n^{2/3}} \times \mathrm{TC}_{n,n-k}\right)$$

for $1 \leq k \leq n^{1/4}$ (which is within the range of applicability of Lemma 3.16). Thus, for $1 \leq k \leq c \log n$,

$$\mathrm{F}_{n,n-k} = \mathcal{O}\left(\frac{\log n}{n^{1/3}} \times \mathrm{TC}_{n,n-k}\right).$$

Plugging this into (27), we obtain

$$\mathrm{TC}_{n,n-1-k} = \frac{1}{2k}\left(1 + \mathcal{O}\left(\frac{\log n}{n^{1/3}}\right)\right)\mathrm{TC}_{n,n-k}$$

and by iteration

$$\mathrm{TC}_{n,n-1-k} = \frac{1}{2^k k!}\left(1+\mathcal{O}\left(\frac{\log n}{n^{1/3}}\right)\right)^k \mathrm{TC}_{n,n-1}$$

from which the result follows since

$$\left(1+\mathcal{O}\left(\frac{\log n}{n^{1/3}}\right)\right)^k = 1+\mathcal{O}\left(\frac{\log^2 n}{n^{1/3}}\right) = 1+o(1).$$

This completes the proof. □

Remark 3.22. Note that the last result improves the bounds of Lemma 3.16, but for a smaller range of k. (The value of c in the proposition is not best possible; however, it is sufficient for our purpose.)

Proof of Theorem 1.9. We first consider the number of tree-child networks with n leaves. Recall that

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

Let $k^* = c\log n$ with c from the last proposition. Then,

$$\mathrm{TC}_n = \sum_{k\le k^*} \mathrm{TC}_{n,n-1-k} + \sum_{k^*<k\le n-1} \mathrm{TC}_{n,n-1-k}.$$

Using the upper bound in (3.16), we obtain

$$\sum_{k^*<k\le n-1} \mathrm{TC}_{n,n-1-k} \le \mathrm{TC}_{n,n-1} \sum_{k>k^*} \frac{1}{2^k k!} = o(\mathrm{TC}_{n,n-1}).$$

On the other hand, by the last proposition:

$$\begin{aligned}\sum_{k\le k^*} \mathrm{TC}_{n,n-1-k} &= (1+o(1))\mathrm{TC}_{n,n-1} \sum_{k\le k^*} \frac{1}{2^k k!}\\ &= (1+o(1))\mathrm{TC}_{n,n-1} \sum_{k=0}^{\infty} \frac{1}{2^k k!}\\ &= (1+o(1))\mathrm{TC}_{n,n-1} e^{1/2}.\end{aligned}$$

Combining the last two displays gives

$$\mathrm{TC}_n \sim e^{1/2}\mathrm{TC}_{n,n-1}, \qquad (n\to\infty). \tag{28}$$

Thus, for all fixed k:

$$P(n-1-R_n=k) = \frac{\mathrm{TC}_{n,n-1-k}}{\mathrm{TC}_n} \longrightarrow \frac{e^{-1/2}}{2^k k!}, \qquad (n\to\infty)$$

which proves the claimed Poisson limit law. □

Remark 3.23. With the same arguments as used to prove (28), we can also show that

$$\sum_{k=0}^{n-1} k\mathrm{TC}_{n-1-k} \sim \frac{1}{2} e^{1/2}\mathrm{TC}_{n,n-1}$$

and thus $\mathbb{E}(n-1-R_n)\sim 1/2$. Moreover, in a similar way, higher moments of $n-1-R_n$ can be shown as well to converge to those of $\mathrm{Poisson}(1/2)$.

3.3.2 d-combining Networks with $d \geq 3$

Here, we prove Theorem 1.9 for $d \geq 3$. For the sake of simplicity, we restrict to the case $d = 3$; the case of larger d follows along similar lines.

We start with some notation. First, we denote by $\mathcal{TC}^{(d)}_{n,k}$ the set of tree-child networks with n leaves and k reticulation nodes. Note that in Proposition 3.4, we constructed a bijection f from

$$f : \underbrace{\{0,1\}^{n-k-1} \times \mathcal{TC}^{(d)}_{n,k}}_{=: 2^{n-k-1} \times \mathcal{TC}^{(d)}_{n,k}} \longmapsto \mathcal{C}^{(d)}_{n-1,k} \times S_n,$$

where S_n denotes the symmetric group of order n.

In particular, this bijection implies that $\mathcal{TC}^{(d)}_{n,n-1}$ is in bijection with $\mathcal{C}^{(d)}_{n-1,n-1} \times \mathcal{S}_n$ for $d = 2, 3$ ($k = n-1$) and $2 \times \mathcal{TC}^{(d)}_{n,n-2}$ is in bijection with $\mathcal{C}^{(d)}_{n-1,n-2} \times \mathcal{S}_n$ for $d = 2, 3$ ($k = n-2$); see the upper half and lower half of Figure 8. Also, from Remark 3.9, we have a bijection from $\mathcal{TC}^{(2)}_{n,n-1}$ to $2 \times \mathcal{TC}^{(2)}_{n,n-2}$ (see the middle arrow in the left half of Figure 8) which gives a 2-to-1 map from $\mathcal{TC}^{(2)}_{n,n-1}$ to $\mathcal{TC}^{(2)}_{n,n-2}$ (take a maximally reticulated network and remove the reticulation node associated with the non-free edge of the child of the root, the non-free edge and its initial node; see Remark 3.9) followed by a 1-to-2 map from $\mathcal{TC}^{(2)}_{n,n-2}$ to $2 \times \mathcal{TC}^{(2)}_{n,n-2}$ (by picking the newly created free edge which contained the reticulation node in the previous construction).

Using this, we can now prove the following result.

Lemma 3.24. *We have,*

$$\mathrm{TC}^{(3)}_{n,n-2} = o(\mathrm{TC}^{(3)}_{n,n-1}).$$

Proof. We first explain a construction of $\mathcal{TC}^{(3)}_{n,n-1}$ from $\mathcal{TC}^{(2)}_{n,n-1}$: for a network from $\mathcal{TC}^{(2)}_{n,n-1}$, 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 reticulation node in the construction of the word and permutation from the proof of Theorem 3.4, where the reticulation nodes are processed consecutively (in any order). Depending on the choice of the edges, we get several networks in $\mathcal{TC}^{(3)}_{n,n-1}$ which all have essentially the same path-component structure as the network from $\mathcal{TC}^{(2)}_{n,n-1}$ (with respect to the encoding from Section 3.1). Conversely, removing the first parent (and corresponding edge) of each reticulation node of a network in $\mathcal{TC}^{(3)}_{n,n-1}$ gives a network in $\mathcal{TC}^{(2)}_{n,n-1}$. Thus, this gives a bijection between networks in $\mathcal{TC}^{(2)}_{n,n-1}$ and classes of networks from $\mathcal{TC}^{(3)}_{n,n-1}$, where these classes form a partition of $\mathcal{TC}^{(3)}_{n,n-1}$; see the second row in the top half of Figure 8.

Equivalently, when viewing networks as words and permutations (where the permutation however is here irrelevant because it does not change in the construction), any word in $\mathcal{C}^{(3)}_{n-1,n-1}$ can be obtained by adding a new ω_i at any position before the first occurrence of ω_i in a word from $\mathcal{C}^{(2)}_{n-1,n-1}$, where ω_i runs through all letters.

For example: $baaabb$ (a word in $\mathcal{C}^{(2)}_{2,2}$) leads to the class $\{bbaaaabb, babaaabb, abbaaabb\}$ (words in $\mathcal{C}^{(3)}_{2,2}$); see Figure 9 for the corresponding networks.

Next, construct $2 \times \mathcal{TC}^{(3)}_{n,n-2}$ from $2 \times \mathcal{TC}^{(2)}_{n,n-2}$ in a similar way (where we turn networks into maximally reticulation networks as in the proof of Theorem 3.4). In particular, since every network in $2 \times \mathcal{TC}^{(2)}_{n,n-2}$ corresponds to a word from $\mathcal{C}^{(2)}_{n-1,n-2}$ and a permutation (which is however again irrelevant), we apply to the words from $\mathcal{C}^{(2)}_{n-1,n-2}$ the same construction as above with the only difference that we only use the ω_i's which are repeated 3 times. Like this, we obtain all words from $\mathcal{C}^{(3)}_{n-1,n-2}$.

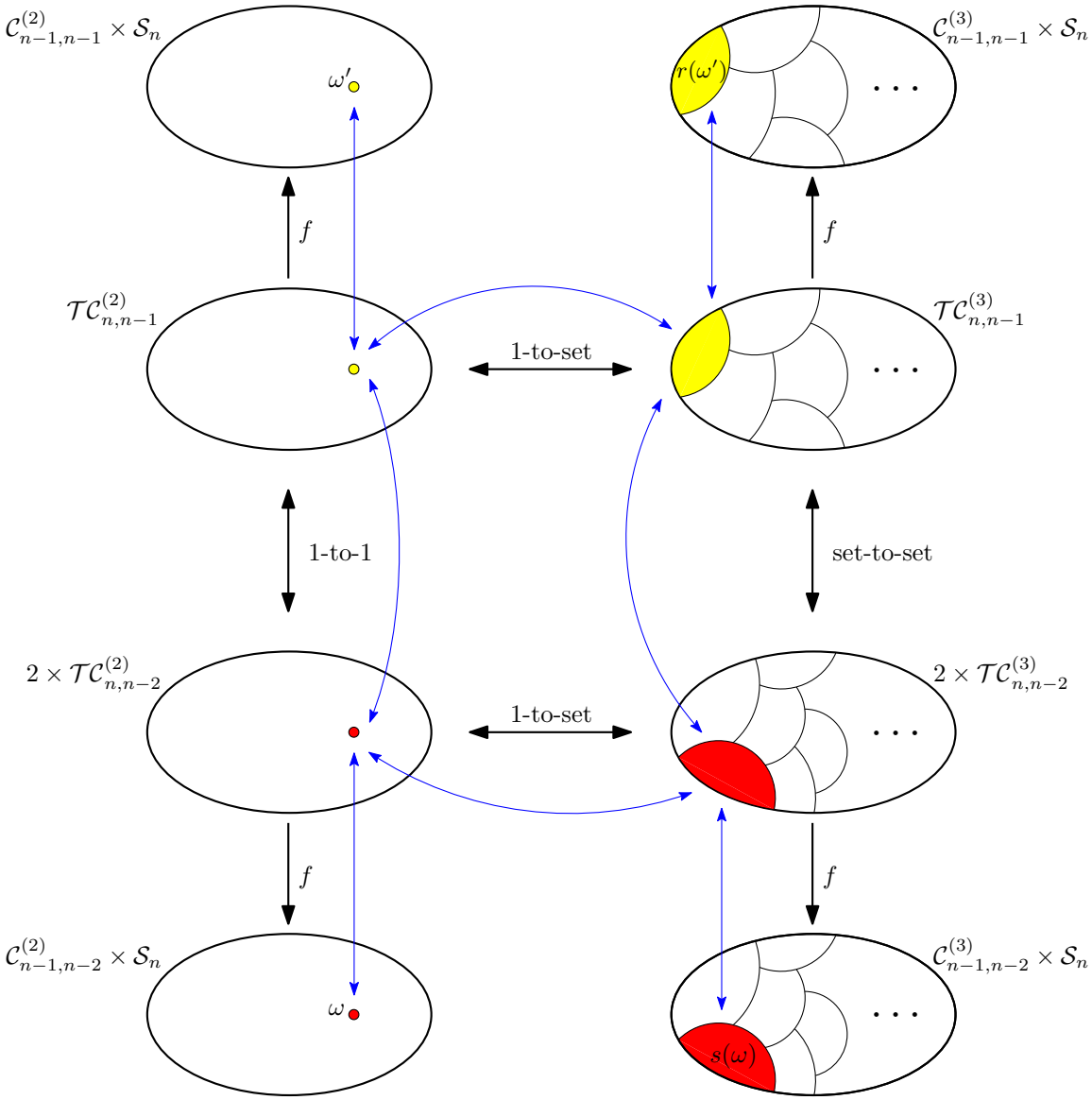

Figure 8: Visualization of the constructions and maps which are used in the proof of Lemma 3.24.

For example, take $abbab + 132$ and $aabbb + 312$ (which encode the same network in $\mathcal{TC}_{3,1}$); $abbab$ leads to the words $\{babbab, abbbab\}$ from $\mathcal{C}^{(3)}_{2,1}$ and $aabbb$ leads to the words $\{baabbbb, ababbb, aabbbb\}$ from $\mathcal{C}^{(2)}_{2,1}$; see Figure 10 for a plot of the corresponding networks from $\mathcal{TC}^{(3)}_{3,1}$.

Next, as mentioned in the paragraph before the lemma, there is a bijection between $\mathcal{TC}^{(2)}_{n,n-1}$ and $2 \times \mathcal{TC}^{(2)}_{n,n-2}$. This bijection gives rise to a bijection between $\mathcal{C}^{(2)}_{n-1,n-1} \times \mathcal{S}_n$ and $\mathcal{C}^{(2)}_{n-1,n-2} \times \mathcal{S}_n$ (which just removes the first letter from a word ω' of the former to obtain a word ω of the latter); see left half of Figure 8. (The permutation remains unchanged.) Note that ω' is bijectively mapped onto a set of words from $\mathcal{C}^{(3)}_{n,n-1}$ and ω onto a set of words from $\mathcal{C}^{(2)}_{n,n-2}$; see the top and bottom half of Figure 8. Denote the cardinality of these sets by $r(\omega')$ and $s(\omega)$, respectively. Then, to show that $\mathrm{TC}^{(3)}_{n,n-2} = o(\mathrm{TC}^{(3)}_{n,n-1})$, it suffices to show that $s(\omega) = o(r(\omega'))$ uniformly over all ω in $\mathcal{C}^{(2)}_{n,n-2}$, or equivalently, we have to find a uniform lower bound of the ratio $r(\omega')/s(\omega)$ that tends to infinity. We will consider this ratio next.

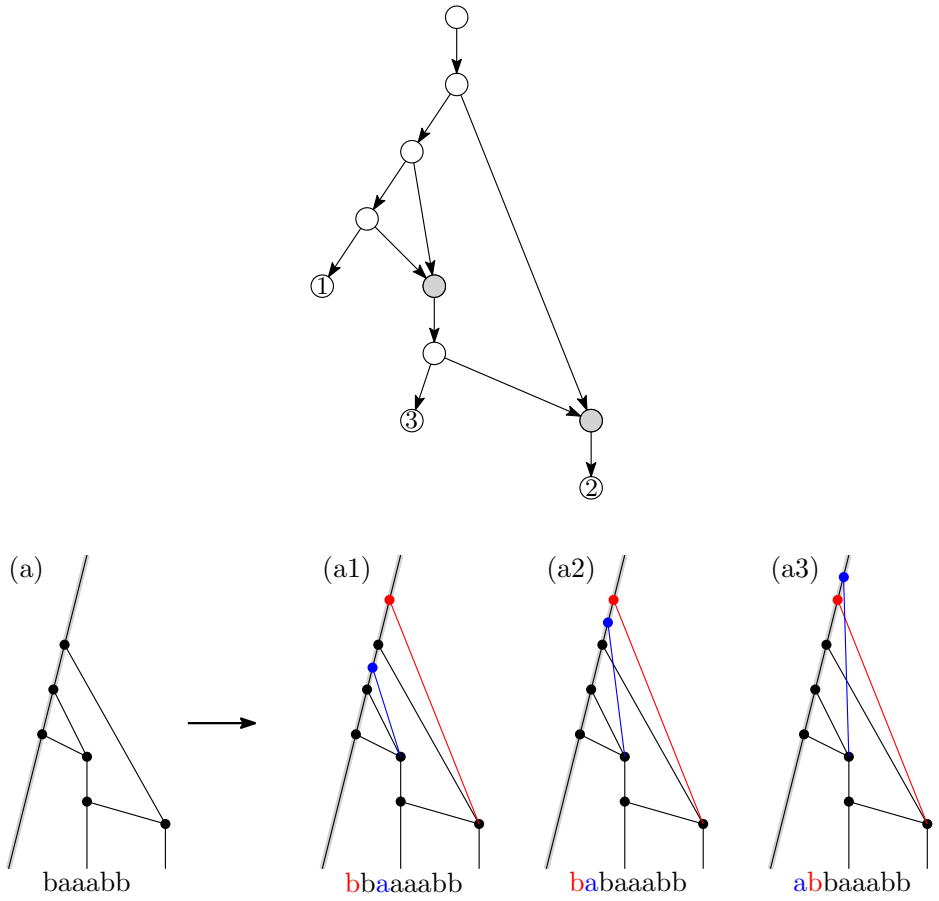

Figure 9: The construction of $\mathcal{TC}^{(3)}_{n,n-1}$ from $\mathcal{TC}^{(2)}_{n,n-1}$. Top: a network from $\mathcal{TC}^{(2)}_{n,n-1}$. Bottom: the three networks from $\mathcal{TC}^{(3)}_{n,n-1}$ constructed from the top network; the corresponding words are below each network. (The permutation in each case is 132; that is why we refrained from indicating it.)

For example, if $\omega = abbccabc$, then $\omega' = aabbccabc$ and the ratio becomes

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

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 ith 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 $k_i/(k_i - 2)$'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, i.e., the 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 $\mathrm{TC}^{(3)}_{n,n-2} = o(\mathrm{TC}^{(3)}_{n,n-1})$. This is the claimed result. □

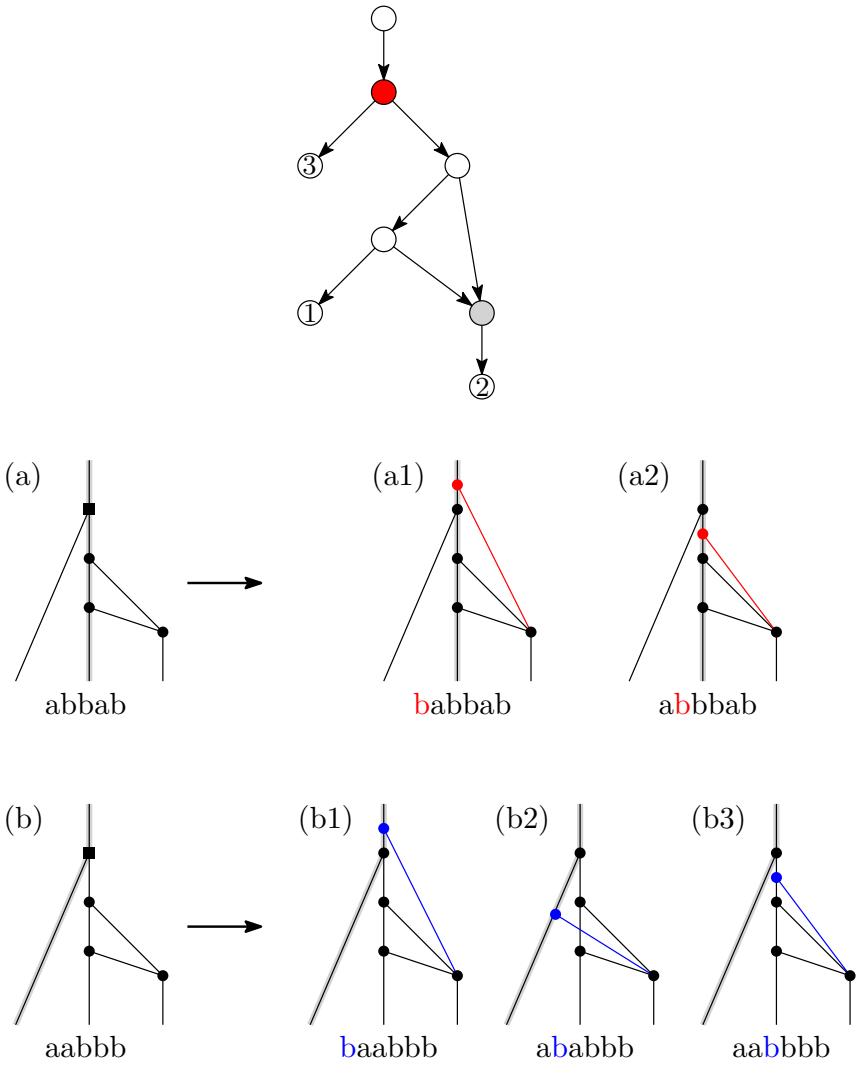

Figure 10: The construction of $\mathcal{TC}^{(3)}_{n,n-2}$ from $\mathcal{TC}^{(2)}_{n,n-2}$. Top: a network from $\mathcal{TC}^{(2)}_{n,n-2}$. Bottom left: the two corresponding networks from $2 \times \mathcal{TC}^{(2)}_{n,n-2}$ where the first path-component (using the indexing from the proof of Theorem 3.4) is in bold. Bottom right: the networks from $\mathcal{TC}^{(3)}_{n,n-2}$ constructed from each of the two networks; the corresponding words are below the networks.

We can now proof the second case of Theorem 1.9.

Proof of Theorem 1.9-(ii). First, observe that from (16) and the previous lemma, we have

$$\mathrm{TC}^{(3)}_{n,n-1-k} = o(\mathrm{TC}^{(3)}_{n,n-1}) \tag{29}$$

for all fixed $k \geq 1$.

Next, by iterating (16),

$$\mathrm{TC}^{(3)}_{n,n-1-k} \leq \frac{1}{2^{k-1}k!}\mathrm{TC}^{(3)}_{n,n-2}$$

for all $1 \leq k \leq n-1$. Consequently,

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

Thus,

$$\mathrm{TC}_n^{(3)} = \mathrm{TC}_{n,n-1}^{(3)} + \mathrm{TC}_{n,n-2}^{(3)} + \sum_{k=2}^{n-1} \mathrm{TC}_{n,n-1-k}^{(3)} \sim \mathrm{TC}_{n,n-1}^{(3)}. \tag{30}$$

Now, we can prove the claim:

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

where the last step follows from (29) and (30). □

Remark 3.25. Similar as pointed out in Remark 3.23, we can prove that also all moments of $n-1-T_n^{(3)}$ converge to that of the Dirac measure at 0.

3.3.3 Proof of Corollary 1.10

Note that

$$\mathbb{E}(W_n^{(d)}) = \sum_{\ell \geq 0} P(W_n^{(d)} > \ell) \leq \sum_{\ell \geq 0} P(n-1-T_n^{(d)} > \ell) = \mathbb{E}(n-1-T_n^{(d)}).$$

where the inequality follows from the fact that each twig is a free tree node and the number of free tree nodes is given by $n-1-k$ where k is the number of reticulation nodes; see Lemma 3.6. The result follows now from:

$$\mathbb{E}(n-1-T_n^{(d)}) \longrightarrow \begin{cases} 1/2, & \text{if } d=2; \\ 0, & \text{if } d \geq 3; \end{cases}$$

see Remark 3.23 and Remark 3.25.

4 Conclusion and Open Questions

The main purpose of this paper was to propose and investigate the class of *d-combining tree-child networks* which generalizes the class of bicombining tree-child networks. The latter class is one of the most important classes of phylogenetic networks; it has been widely studied in recent years. One of our major reasons for generalizing it was that a better understanding of its combinatorial properties was gained by placing it into a larger framework. We give a short summary of the results we obtained.

First, for one-component d-combining tree-child network, the availability of an easy counting formula (see Section 2.1) made it possible to derive asymptotic counting results and distributional results for the number of reticulation nodes of random one-component d-combining tree-child networks by standard methods (see Section 2.2). Moreover, we also investigated the Sackin index for this class (see Section 2.3) following [3] where the bicombining case was considered. However, whereas we derived convergence of all moments and limit law results for the number of reticulation nodes, we only proved a Theta-result for the mean of the Sackin index. It would be interesting to derive higher moments and prove a limit law result for the Sackin index, too. Moreover, it would be interesting to prove similar results also for other shape parameters such as the number of cherries.

For general d-combining tree-child networks, counting them turned out to be more involved. For the bicombining case, in [16] an encoding by words was proposed which led to a counting formula for this class of networks. However, this encoding was just conjectural. In Section 3.1, we proposed a slightly modified encoding which could be rigorously established (and extends from $d=2$ to $d \geq 3$). Our encoding again led to a formula for the number of networks providing a recursive way of computing these

numbers for small values of n, k, d (see Appendix A). In addition, we used our encoding to establish a Theta-result for the number of networks (see Section 3.2) and again proved convergence of moments and a limit distribution result for the number of reticulation nodes (see Section 3.3). From the latter result, we also derived a preliminary result about the distribution of cherries. Proving more detailed stochastic results for this number as well as investigating the Sackin index (and other parameters) for general d-combining tree-child networks is still a challenge and we leave these questions open for further investigations. In addition, it would be also interesting to know whether the results from Appendix B, where results for fixed k are discussed with a generalization of the method from [3] from the bicombining to the d-combining case, can be derived with our encoding from Section 3.1? Moreover, also the question whether the conjecture from [16] can be generalized to d-combining networks is open. (We might treat these questions elsewhere.)

A straightforward (and natural) further generalization of our class of d-combining networks would be to allow a finite set, say $\{d_1, \ldots, d_m\}$, of in-degrees for reticulation nodes. (Allowing an infinite set would make the counting problem meaningless.) Indeed, many of our exact results seem to generalize to this situation and asymptotic results should be doable as well. However, so far, we cannot see what can be gained from such a further generalization; the results of this paper seem to continue to hold and no new phenomena seem to arise. That is why we refrained from doing this here and unless such a generalization leads to considerable new mathematical challenges and/or different phenomena and/or additional insights about the (most important) bicombining case, this level of generality will only be explored in the PhD thesis of the first author.

5 Acknowledgments

Yu-Sheng Chang and Michael Fuchs were partially supported by NSTC (National Science and Technology Council) under the grant NSTC-111-2115-M-004-002-MY2; Michael Wallner was supported by the Austrian Science Fund (FWF): P 34142; Guan-Ru Yu was supported by NSTC under the grant NSTC-110-2115-M-017-003-MY3.

References

[1] C. Banderier, P. Marchal, M. Wallner (2018). Rectangular Young tableaux with local decreases and the density method for uniform random generation, *GASCom 2018*, CEUR Workshop Proceedings. Vol. 2113. 2018, pp. 60–68.

[2] C. Banderier and M. Wallner (2021). Young tableaux with periodic walls: counting with the density method, *Sém. Lothar. Combin.* 85B, Art. 47, 12 pp.

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

[4] 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 Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, 225, Paper 5.

[5] A. Elvey Price, W. Fang, M. Wallner (2020). Asymptotics of minimal deterministic finite automata recognizing a finite binary language, LIPICS, Proceedings of the 31st Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, 159, Paper 11.

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

[7] M. Fischer, L. Herbst, S. Kersting, L. Kühn, K. Wicke. Tree balance indices: a comprehensive survey, arXiv:2109.12281.

[8] P. Flajolet and R. Sedgewick. *An Introduction to the Analysis of Algorithms*, Addison-Wesley Professional, 2nd edition, 2013.

[9] 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.

[10] 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.

[11] M. Fuchs, E.-Y. Huang, G.-R. Yu (2022). Counting phylogenetic networks with few reticulation vertices: a second approach, *Discrete Appl. Math.*, **320**, 140–149.

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

[13] 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 pp.

[14] D. H. Huson, R. Rupp, C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge University Press, 1st edition, 2010.

[15] C. McDiarmid, C. Semple, D. Welsh (2015). Counting phylogenetic networks, *Ann. Comb.*, **19:1**, 205–224.

[16] 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, **11**, Article number: 21875.

[17] M. Steel. *Phylogeny—Discrete and Random Processes in Evolution*, CBMS-NSF Regional Conference Series in Applied Mathematics, 89, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

[18] M. Wallner, http://dmg.tuwien.ac.at/mwallner, 2022.

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

[20] L. Zhang (2022). The Sackin index of simplex networks, In: Jin, L., Durand, D. (eds) Comparative Genomics. RECOMB-CG 2022, Lecture Notes in Computer Science, **13234**, Springer, Cham.

A Tables

$n\backslash k$	0	1	2	3	4	5	6	7
2	1	2						
3	3	21	42					
4	15	228	1272	2544				
5	105	2805	30300	154500	309000			
6	945	39330	696600	6494400	31534200	63068400		
7	10395	623385	16418430	241204950	2068516800	9737380800	19474761600	
8	135135	11055240	405755280	8609378400	113376463200	920900131200	4242782275200	8485564550400

Table 2: $\mathrm{TC}^{(2)}_{n,k}$ for $2 \leq n \leq 8$ and $0 \leq k < n$; see also [3].

$n\backslash k$	0	1	2	3	4	5	6
2	1	2					
3	3	33	150				
4	15	492	7908	55320			
5	105	7725	291420	6179940	57939000		
6	945	132030	9603270	430105320	11292075000	132120450000	
7	10395	2471805	307525050	24586633890	1284266876760	40079165452200	560319972030000

Table 3: $\mathrm{TC}^{(3)}_{n,k}$ for $2 \leq n \leq 7$ and $0 \leq k < n$.

$n\backslash k$	0	1	2	3	4	5
2	1	2				
3	3	48	546			
4	15	942	45132	1243704		
5	105	18375	2394360	227116260	11351644920	
6	945	375705	107314200	23919407460	3724353682560	291451508298720

Table 4: $\mathrm{TC}^{(4)}_{n,k}$ for $2 \leq n \leq 6$ and $0 \leq k < n$.

$n\backslash k$	0	1	2	3	4
2	1	2			
3	3	66	2016		
4	15	1650	242496	28710864	
5	105	39135	17566470	7876446840	2307919133520

Table 5: $\mathrm{TC}^{(5)}_{n,k}$ for $2 \leq n \leq 5$ and $0 \leq k < n$.

$n\backslash k$	0	1	2	3	4
2	1	2			
3	3	87	7524		
4	15	2700	1246740	676431360	
5	105	76515	118491090	262058953860	483098464854720

Table 6: $\mathrm{TC}^{(6)}_{n,k}$ for $2 \leq n \leq 5$ and $0 \leq k < n$.

B The Method of Component Graphs

The purpose of this appendix is to explain that the method of component graphs from [3] also extends from the bicombining case to the d-combining case. The method yields another formula for $\mathrm{TC}^{(d)}_{n,k}$ (see Section B.1 below) which in the bicombining case was used in [3] to (a) compute values for small n and k and (b) derive formulas for $k = 1$ and $k = 2$. The same can be done with our extension in the d-combining case (see Section B.2 below), however, the computation of values is more effective with the method proposed in Section 3.4 (even in the bicombining case). Moreover, the method from [3] was used in [11] to obtain the first-order asymptotics of $\mathrm{TC}^{(2)}_{n,k}$ for fixed k as n tends to infinity; again this carries over to the d-combining case (see Section B.3 below).

B.1 Exact Counting

The main idea of [3] was to reduce each tree-child network to its component graph and then conversely built all tree-child networks from their component graphs. Before explaining this in detail, we give the definition of a component graph.

Definition B.1 (Component graph)**.** *A component graph is a (rooted) vertex-labeled DAG such that the following properties hold.*

(i) Every non-root node has in-degree equal to d;

(ii) the root has in-degree equal to 0;

(iii) multi-edges between two nodes are allowed.

We denote by $\mathcal{K}^{(d)}_m$ the set of component graphs with m nodes and by $\mathcal{K}^{(d)}_{m,s}$ the set of component graphs with m nodes of which s are leaves. Set $k^{(d)}_m = |\mathcal{K}^{(d)}_m|$ and $k^{(d)}_{m,s} = |\mathcal{K}^{(d)}_{m,s}|$. Then,

$$k^{(d)}_m = \sum_{s=1}^{m-1} k^{(d)}_{m,s},$$

since the number of leaves of a component graph is at least 1 and at most $m-1$ (for the star-component graph; see Figure 12).

This, together with the following recursive formula for $k^{(d)}_{m,s}$, makes it possible to compute $k^{(d)}_m$. (See [3, Theorem 15] for the bicombining case.)

Theorem B.2. *For $m \geq 2$,*

$$k^{(d)}_{m,s} = \sum_{1 \leq t \leq m-s-1} \binom{m}{s} \beta^{(d)}(m,s,t)\, k^{(d)}_{m-s,t}, \qquad (1 \leq s \leq m-1),$$

with initial condition $k^{(d)}_{1,1} = 1$ and

$$\beta^{(d)}(m,s,t) = \sum_{0 \leq \ell \leq t} (-1)^\ell \binom{t}{\ell} \binom{m-s-\ell+d-1}{d}.$$

Proof. The recurrence can be obtained by the following way of constructing all component graphs in $\mathcal{K}^{(d)}_{m,s}$ from those in $\mathcal{K}^{(d)}_{m-s,t}$:

(i) Choose t with $1 \leq t \leq m-s-1$ and a graph G in $\mathcal{K}^{(d)}_{m-s,t}$;

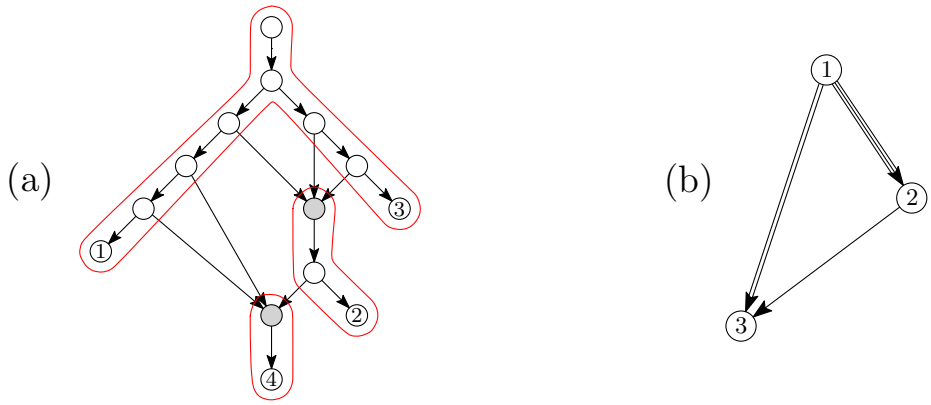

Figure 11: (a) The 3-combining network (with two reticulation nodes) from Figure 1, (a). The three tree-components when incoming edges of the reticulation nodes are removed are encircled in red. (b) The corresponding component graph.

(ii) Add s new nodes, labelled by $\{1', \cdots, s'\}$, to G such that (a) these nodes become the new leaves and (b) all old leaves have at least one out-going edge (i.e., all of them become internal nodes). By the inclusion–exclusion principle, there are

$$\beta^{(d)}(m,s,t) = \sum_{0 \leq \ell \leq t} (-1)^{\ell} \binom{t}{\ell} \binom{m-s-\ell+d-1}{d}^{s}$$

ways of choosing the d incoming edges for the new leaves; here, the counting is done such that ℓ of the old leaves are not used as parents of the new leaves;

(iii) Choose s leaves from the set of labels $\{1, \ldots, m\}$ and use them to re-label the new leaves; use the remaining labels to re-label the remaining nodes in an order-consistent way. □

Now assume that a tree-child networks N with k reticulation nodes is given. Then, its component graph is obtained as follows. First, remove all the incoming edges of the reticulation nodes. The resulting graph is a forest consisting of $k+1$ (directed) trees (called *tree-components*) which are either rooted at the network root or at a reticulation node; see Figure 11, (a). The component graph of N is then the DAG whose vertex set corresponds to the set of tree-components with an edge between vertices if there was an edge (which was deleted in the first step above) connecting the two tree-components. The vertices are labeled as follows: consider the set of leaf-labels of each tree-component which partition $\{1, \ldots, n\}$. The blocks of this partition can be indexed by the rank of the smallest element of the block; these indices are then given as label to the corresponding nodes of the component graph; see Figure 11, (b) for an example.

The above construction reduces every tree-child network with k reticulation nodes to a component graph with $k+1$ nodes. Conversely, all tree-child networks with k reticulation nodes can be obtained from the component graphs with $k+1$ nodes by a blow-up procedure: first choose a component graph with $k+1$ nodes and a set partition of $\{1, \ldots, n\}$ into $k+1$ blocks. Order the blocks according to the ranks of their smallest element and distribute them to the vertices of the (chosen) component graph so that the node with label j receives the jth block ($1 \leq j \leq k+1$). For each node, choose a phylogenetic tree whose set of leaf-labels corresponds to the block of the partition which was assigned to the node. Next, add nodes on the edges of the phylogenetic trees such that the number of added nodes equals the out-degrees of the nodes of the component graph. Finally, connect these nodes to the roots of the phylogenetic trees in the same way as the corresponding nodes in the component graph are connected. The resulting networks are all tree-child networks with k reticulation nodes.

The blow-up procedure just described immediately translates into a formula for $\mathrm{TC}^{(d)}_{n,k}$. (See [3, Theorem 16] for the bicombining case.)

Theorem B.3. *Let $\prod_{n,k+1}$ be the set of partitions of $\{1,\dots,n\}$ into $k+1$ blocks. Then,*

$$\mathrm{TC}^{(d)}_{n,k} = \frac{1}{2^{n-k-1}} \sum_{\{B_j\}_{j=1}^{k+1} \in \prod_{n,k+1}} \sum_{G \in \mathcal{K}^{(d)}_{k+1}} \prod_{j=1}^{k+1} \frac{(2b_j + g_j - 2)!}{(b_j-1)! \prod_{\ell=1}^{k+1} (g_{j,\ell})!}, \tag{31}$$

where $b_j = |B_j|$ for $1 \le j \le k+1$, $g_{j,\ell}$ is the number of edges in G which are directed from node j to node ℓ, and $g_j = \sum_\ell g_{j,\ell}$ is the out-degree of node j for $1 \le j \le k+1$.

Proof. By the description preceding the theorem, the formula is explained as follows:

(i) G is the chosen graph in $\mathcal{K}^{(d)}_{k+1}$ and $\{B_j\}_{j=1}^{k+1}$ is the chosen partition in $\prod_{n,k+1}$.

(ii) The number of possible phylogenetic trees assigned to the nodes of G is:

$$\prod_{j=1}^{k+1} (2b_j - 3)!! = \prod_{j=1}^{k+1} \frac{(2b_j-2)!}{2^{b_j-1}(b_j-1)!}, \tag{32}$$

where we assume that B_j is the block belonging to node j in G.

(iii) The number of ways of adding nodes to the phylogenetic tree of node j is:

$$(2b_j - 1) \cdots (2b_j - 1 + g_j - 1)$$

(iv) Connecting the nodes which have been added to the phylogenetic tree of node j to the root of the phylogenetic tree of node ℓ gives every tree-child networks $(g_{j,\ell})!$ times. Thus, the number of tree-child networks arising from a fixed choice of G, $\{B_j\}_{j=1}^{k+1}$ and a set of phylogenetic trees (whose number of leaves equals to b_j for $1 \le j \le k+1$) is:

$$\prod_{j=1}^{k+1} \frac{(2b_j - 1) \cdots (2b_j - 1 + g_j - 1)}{\prod_{\ell=1}^{k+1} (g_{j,\ell})!} \tag{33}$$

(v) Finally, multiplying (32) and (33) gives

$$\begin{aligned} \prod_{j=1}^{k+1} \frac{(2b_j-2)!}{2^{b_j-1}(b_j-1)!} \frac{(2b_j - 1) \cdots (2b_j - 1 + g_j - 1)}{\prod_{\ell=1}^{k+1} (g_{j,\ell})!} &= \frac{1}{2^{(\sum_j b_j)-k-1}} \prod_{j=1}^{k+1} \frac{(2b_j + g_j - 2)!}{(b_j-1)! \prod_{\ell=1}^{k+1} (g_{j,\ell})!} \\ &= \frac{1}{2^{n-k-1}} \prod_{j=1}^{k+1} \frac{(2b_j + g_j - 2)!}{(b_j-1)! \prod_{\ell=1}^{k+1} (g_{j,\ell})!}, \end{aligned}$$

which is the claimed factor (in front of and inside) of the double sum in (31). □

B.2 Values and Formulas for Small k

The Formula (31) makes it possible to compute the values of $\mathrm{TC}^{(d)}_{n,k}$ for small values of n, k, d. (In [3], this was done for the special case $d = 2$.) However, the method presented in Section 3.4 to achieve the same task is less computation-intensive since using (31) makes it necessary to generate all component

graphs with $k+1$ vertices. (The number of component graphs increases rapidly as can be seen from Theorem B.2.)

Another application of (31), which was also given in [3], is the derivation of formulas for small values of k (which hold for all n), e.g., in [3] such formulas were obtained for $d=2$ and $k=1$ and $k=2$. (See also [10, 16] for related formulas.) Indeed, with (31), we can give now generalizations for general $d \geq 2$.

We start with expressions which contain generating functions.

Proposition B.4. *Set*

$$f_d(z) := \sum_{m\geq 1} \frac{(2m+d-2)!}{(m-1)!m!} z^m.$$

(i) For $k=1$,

$$\mathrm{TC}_{n,1}^{(d)} = \frac{n!}{d!2^{n-2}}[z^n] f_d(z) f_0(z). \tag{34}$$

(ii) For $k=2$,

$$\mathrm{TC}_{n,2}^{(d)} = \frac{n!}{d!\,2^{n-3}} \sum_{\ell=0}^{d} \frac{1}{(d-\ell)!\,\ell!}\,[z^n] f_{2d-\ell}(z)\, f_\ell(z)\, f_0(z) - \frac{n!}{(d!)^2 2^{n-2}}[z^n]\, f_{2d}(z)\, f_0^2(z). \tag{35}$$

Proof. We start with $k=1$. By using (31), we have

$$\mathrm{TC}_{n,1}^{(d)} = \frac{1}{2^{n-2}} \sum_{\{B_j\}_{j=1}^2 \in \prod_{n,2}} \sum_{G\in\mathcal{K}_2^{(d)}} \prod_{j=1}^{2} \frac{(2b_j+g_j-2)!}{(b_j-1)!\prod_{\ell=1}^2 (g_{j,\ell})!}.$$

Observe that $\mathcal{K}_2^{(d)}$ contains only two graphs, namely, the graph consisting of a root to which a child is attached by d edges and either the root has label 1 or 2. Consequently,

$$\mathrm{TC}_{n,1}^{(d)} = \frac{1}{d!\,2^{n-2}} \sum_{\{B_j\}_{j=1}^2 \in \prod_{n,2}} \left(\frac{(2b_1+d-2)!}{(b_1-1)!} \cdot \frac{(2b_2-2)!}{(b_2-1)!} + \frac{(2b_1-2)!}{(b_1-1)!} \cdot \frac{(2b_2+d-2)!}{(b_2-1)!} \right).$$

Summing according to the size of the blocks in the partition $\{B_j\}_{j=1}^2$, we have

$$\begin{aligned}\mathrm{TC}_{n,1}^{(d)} &= \frac{1}{d!\,2^{n-2}} \sum_{\substack{b_1+b_2=n,\\ b_1,b_2\geq 1}} \binom{n-1}{b_1-1} \frac{(2b_1+d-2)!\,(2b_2-2)! + (2b_1-2)!\,(2b_2+d-2)!}{(b_1-1)!\,(b_2-1)!} \\ &= \frac{1}{d!\,2^{n-2}} \sum_{b=1}^{n-1} \binom{n}{b} \frac{(2b+d-2)!\,(2n-2b-2)!}{(b-1)!\,(n-b-1)!}\end{aligned}$$

which translates into (34) by using the generating function $f_d(z)$.

Next, we consider $k=2$. Here, again by (31),

$$\mathrm{TC}_{n,2}^{(d)} = \frac{1}{2^{n-3}} \sum_{\{B_j\}_{j=1}^3 \in \prod_{n,3}} \sum_{G\in\mathcal{K}_3^{(d)}} \prod_{j=1}^{3} \frac{(2b_j+g_j-2)!}{(b_j-1)!\prod_{\ell=1}^3 (g_{j,\ell})!}.$$

In contrast to $k=1$, there are now more possibilities for G: G has three vertices v_1, v_2, v_3 one of which is the root (say v_1); v_1 is connected to v_2 by d edges and to v_3 by ℓ edges with $0 \leq \ell \leq d$; moreover,

v_2 is connected to v_3 by $d-\ell$ edges; finally, there are 3 possible labelings if $\ell = d$ and $3! = 6$ possible labelings if $\ell < d$. Thus,

$$\begin{aligned}\mathrm{TC}^{(d)}_{n,2} =& \frac{1}{2^{n-3}} \sum_{b_1+b_2+b_3=n} \binom{n}{b_1, b_2, b_3} \frac{1}{3!} \\ &\times \left(\sum_{\ell=0}^{d-1} \frac{3!}{d!\ell!(d-\ell)!} \cdot \frac{(2b_1+d+\ell-2)!}{(b_1-1)!} \cdot \frac{(2b_2+d-\ell-2)!}{(b_2-1)!} \cdot \frac{(2b_3-2)!}{(b_3-1)!} \right. \\ &\left. + \frac{3}{(d!)^2} \cdot \frac{(2b_1+2d-2)!}{(b_1-1)!} \cdot \frac{(2b_2-2)!}{(b_2-1)!} \cdot \frac{(2b_3-2)!}{(b_3-1)!} \right).\end{aligned}$$

From this, by using the generating $f_d(z)$, we obtain

$$\mathrm{TC}^{(d)}_{n,2} = \frac{n!}{d!2^{n-3}} \sum_{\ell=1}^{d} \frac{1}{\ell!(d-\ell)!} [z^n] f_{2d-\ell}(z) f_\ell(z) f_0(z) + \frac{n!}{(d!)^2 2^{n-2}} [z^n] f_{2d}(z) f_0^2(z)$$

which is equivalent to (35). □

In order to simplify these expressions, we need two technical lemmas. The first gives a recursive way of computing $f_d(z)$.

Lemma B.5. *Set $X := \sqrt{1-4z}$. Then, $f_d(X)$ can be recursively computed as*

$$f_d(X) = (-X^{-1} + X) f'_{d-1}(X) + (d-2) f_{d-1}(X), \qquad (d \geq 1)$$

with initial condition $f_0(X) = 1/2 - X/2$.

Proof. First, for the initial condition, observe that

$$\frac{(2m-2)!}{(m-1)!m!} = \frac{1}{m} \binom{2m-2}{m-1}$$

and thus $f_0(z)$ is the generating function of the (shifted) Catalan numbers. Consequently,

$$f_0(z) = \frac{1-\sqrt{1-4z}}{2} = \frac{1-X}{2}.$$

Next, in order to prove the recurrence, we have

$$\begin{aligned} f_d(z) &= 2 \sum_{m \geq 1} \frac{m(2m+d-3)!}{(m-1)!m!} z^m + (d-2) \sum_{m \geq 1} \frac{(2m+d-3)!}{(m-1)!m!} z^m \\ &= 2z f'_{d-1}(z) + (d-2) f'_{d-1}(z). \end{aligned}$$

Writing this in terms of X gives the claimed result. □

From this lemma, we see that the first few values of $f_d(X)$ are:

$$f_1(X) = \frac{1}{2X} - \frac{1}{2}, \quad f_2(X) = \frac{1}{2X^3} - \frac{1}{2X}, \quad f_3(X) = \frac{3}{2X^5} - \frac{3}{2X^3}.$$

In particular, it is not hard to see that $f_d(X)$ for $d \geq 2$ is a (finite) linear combinations of terms of the form X^{-m} with m odd.

We need a second technical lemma, which will help us with the extraction of coefficients.

Lemma B.6. *Let* $X := \sqrt{1-4z}$. *Then, for odd* m

$$[z^n]\, X^{-m} = \frac{1}{\binom{m-1}{(m-1)/2}} \binom{n+(m-1)/2}{(m-1)/2} \binom{2n+m-1}{n+(m-1)/2}$$

and for even m

$$[z^n]\, X^{-m} = 4^n \binom{n+(m-2)/2}{(m-2)/2}.$$

Proof. Note that

$$[z^n]\, X^{-m} = \binom{-m/2}{n} (-4)^n = \frac{m(m+2)\cdots(m+2n-2)}{n!}\, 2^n.$$

From this both claims follow by standard manipulations. □

Now, we can simplify (34) for small values of d; recall the notation (7) of double factorials.

Corollary B.7. *The number of tree-child networks with* n *leaves and* 1 *reticulation node is*

(i) for $d = 2$*:*

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

(ii) for $d = 3$*:*

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

Proof. First, from (34) and the initial condition in Lemma B.5:

$$\mathrm{TC}^{(d)}_{n,1} = \frac{n!}{d!2^{n-1}}[z^n]f_d(z) - \frac{n!}{d!2^{n-1}}[z^n]f_d(z)X = \frac{(2n+d-2)!}{d!2^{n-1}(n-1)!} - \frac{n!}{d!2^{n-1}}[z^n]f_d(z)X. \tag{36}$$

The second term becomes for $d = 2$,

$$[z^n]f_2(z)X = [z^n]\left(\frac{1}{2\,X} - \frac{1}{2\,X^3}\right)X = -\frac{1}{2}[z^n]X^{-2} = -\frac{4^n}{2}$$

and for $d = 3$,

$$[z^n]f_3(z)X = [z^n]\left(\frac{3}{2X^5} - \frac{3}{2X^3}\right)X = \frac{3}{2}[z^n]X^{-4} - \frac{3}{2}[z^n]X^{-2} = \frac{3\cdot 4^n}{2}n.$$

Plugging this into (36) and standard manipulations give the claimed result. □

Remark B.8. The formula for $d = 2$ is known; see, e.g., [3, 10, 16].

Remark B.9. The method of proof gives the following structural result for general d and $n \geq 2$:

$$\mathrm{TC}^{(d)}_{n,1} = \binom{2n+d-2}{d}(2n-3)!! - p_d(n)(2n-2)!!,$$

where $p_d(n)$ is a polynomial of degree $d-1$. Note that $\mathrm{TC}^{(d)}_{1,1} = 0$.

Likewise, we can also simplify (35) for small values of d.

Corollary B.10. *The number of tree-child networks with* n *leaves and* 2 *reticulation node is*

(i) for $d = 2$*:*

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

(ii) for $d = 3$*:*

$$\mathrm{TC}^{(3)}_{n,2} = n(n-1)\left(\frac{70n^2+244n+177}{315}(2n+1)!! - \frac{16n+13}{48}(2n+2)!!\right).$$

Remark B.11. The formula for $d = 2$ is again known; see, e.g., [3, 10, 16].

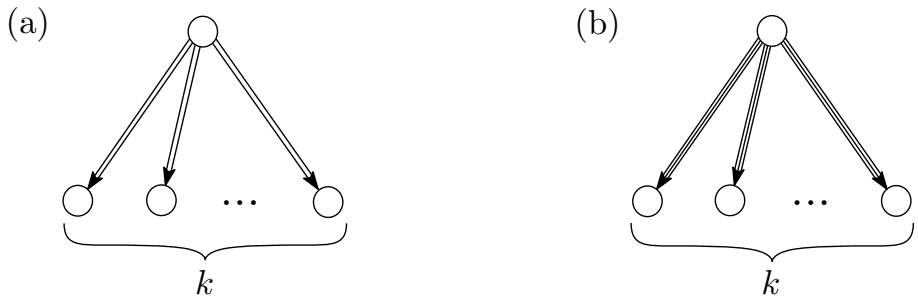

Figure 12: Star-component graphs for generating the tree-child networks whose numbers dominate the asymptotics of $\mathrm{TC}_{n,k}^{(d)}$ for fixed k as n tends to infinity. (Left: $d = 2$; Right: $d = 3$. Labels of nodes are removed.)

B.3 Asymptotics for Fixed k

In this subsection, we give a final application of the method of component graphs, namely, we derive the first-order asymptotics of $\mathrm{TC}_{n,k}^{(d)}$ for fixed k as $n \to \infty$. This extends the main result from [11] to general d.

The main observation of [11] was that the main asymptotic contribution to the asymptotics of $\mathrm{TC}_{n,k}^{(d)}$ comes from the tree-child networks constructed from the star-component graph with k leaves; see Figure 12. (In fact, since component graphs are labeled, there are $k+1$ star-component graphs depending on the label of the root vertex.)

We denote by $S_{n,k}^{(d)}$ the number of tree-child networks arising from the star-component graph(s) with k leaves. Then, we have the following formula for this number which generalizes the formula from [11, Lemma 5] for $d = 2$.

Lemma B.12. *We have,*

$$S_{n,k}^{(d)} = \frac{n!}{(d!)^k 2^{n-k-1}(k-1)!} \sum_{j=1}^{n-k} \frac{(2j+dk-2)!}{j!(j-1)!} \cdot \frac{(2n-k-2j-1)!}{(n-k-j)!(n-j)!}.$$

Proof. The proof is very similar to the one of [11, Lemma 5]. We give a short sketch.

First, by a combinatorial argument, we have

$$S_{n,k}^{(d)} = \sum_{j=1}^{n-k} \binom{n}{j} \frac{(2j+dk-2)!}{(d!)^k 2^{j-1}(j-1)!} \cdot \frac{1}{k!}(n-j)![z^{n-j}]T(z)^k,$$

where

$$T(z) = \sum_{n\geq 1} (2n-1)!! \cdot \frac{z^n}{n!} = 1 - \sqrt{1-2z}$$

is the exponential generating function of the number of phylogenetic trees.

Briefly, the construction on which this combinatorial argument is based works as follows: first, we pick a one-component tree-child networks with $j+k$ leaves where the leaves with labels $\{1,\dots,k\}$ are the ones below the reticulation nodes. Then, we replace the leaves below reticulation nodes by phylogenetic trees. (For this, we need a forest of k phylogenetic trees.) Finally, we re-label the leaves.

Next, by a standard application of the Lagrange inversion formula, we obtain

$$[z^{n-j}]T(z)^k = \frac{k}{n-j} 2^{j+k-n} \binom{2n-k-2j-1}{n-k-j}.$$

Plugging this into the expression above and straightforward manipulations yield the claimed result. □

Applying to this the Laplace method, we have the following asymptotic result.

Proposition B.13. *For fixed k, as $n \to \infty$,*

$$S^{(d)}_{n,k} \sim \frac{2^{dk-1}}{(d!)^k\, k!\sqrt{\pi}} n! 2^n n^{dk-3/2}.$$

Proof. The proof is similar to [11, Lemma 6]. Again, we just give a sketch.

First, it suffices to consider the asymptotics of the sum in the expression for $S^{(d)}_{n,k}$ from Lemma B.12 as the factor in front of it has already the right shape. Thus, we set:

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

Now, observe that the first term in the last sum is decreasing in j and has the expansion:

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

uniformly in j as $j = o(n)$. Thus, by a standard application of the Laplace method:

$$\Sigma_{n,k} \sim \frac{2^{(d-2)k}}{\sqrt{\pi}} \left(\sum_{j=0}^{\infty} \frac{(2j+k-1)!}{j!(j+k)!} 4^{-j}\right) 4^{n-1} n^{dk-3/2} = \frac{2^{(d-1)k}}{k\sqrt{\pi}} 4^{n-1} n^{dk-3/2},$$

where we used [11, Lemma 7] in the last step.

From the above asymptotics multiplied with the factor in front of the sum in the formula for $S^{(d)}_{n,k}$ from Lemma B.12, we obtain the claimed result. □

Finally, using the same arguments as in [11], we can show that also here the contribution from tree-child networks arising from the star-component graph(s) dominates; we leave details to the interested reader.

Theorem B.14. *For the number of d-combining tree-child networks with n leaves and k reticulation nodes, we have for fixed k, as $n \to \infty$,*

$$\mathrm{TC}^{(d)}_{n,k} \sim \frac{2^{dk-1}}{(d!)^k\, k!\, \sqrt{\pi}}\, n!\, 2^n\, n^{dk-3/2}.$$

Remark B.15. This result was stated in [4, Theorem 8] without proof. (Note that this also corrects two typos in the statement of [4, Theorem 8].)