Enumeration of d-Combining Tree-Child Networks

Yu-Sheng Chang
Department of Mathematical Sciences, National Chengchi University, Taipei, 116, Taiwan

Michael Fuchs
Department of Mathematical Sciences, National Chengchi University, Taipei, 116, Taiwan

Hexuan Liu
Department of Mathematical Sciences, National Chengchi University, Taipei, 116, Taiwan

Michael Wallner
Institute of Discrete Mathematics and Geometry, TU Wien, Vienna, 1040, Austria

Guan-Ru Yu
Department of Mathematics, National Kaohsiung Normal University, Kaohsiung, 824, Taiwan

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* which are tree-child networks with every reticulation node having exactly two parents. In this paper, we extend these studies to *d-combining tree-child networks* where every reticulation node has now $d \geq 2$ parents. Moreover, we also give results and conjectures on the distributional behavior of the number of reticulation nodes of a network which is drawn uniformly at random from the set of all tree-child networks with the same number of leaves.

2012 ACM Subject Classification Mathematics of computing → Discrete mathematics

Keywords and phrases Phylogenetic network, tree-child network, d-combining tree-child network, exact enumeration, asymptotic enumeration, reticulation node, limit law, stretched exponential

Digital Object Identifier 10.4230/LIPIcs.AofA.2022.5

Funding *Michael Fuchs*: Partially supported by MOST-109-2115-M-004-003-MY2.
Michael Wallner: Supported by the Austrian Science Fund (FWF): P 34142.
Guan-Ru Yu: Partially supported by MOST-110-2115-M-017-003-MY3.

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 [10, 13].

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 [13, 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 [1, 5–9, 11, 12].

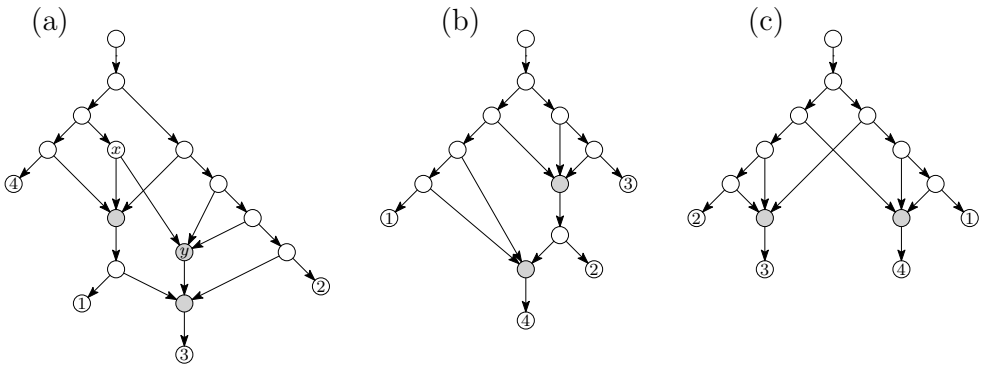

Figure 1 (a) A 3-combining phylogenetic network which is *not* a tree-child network (because both children of the tree node x are reticulation nodes and the only child of the reticulation node y is also a reticulation node); (b) a 3-combining tree-child network; (c) a 3-combining one-component tree-child 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 extended abstract 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. We will highlight similarities and differences between the two cases $d = 2$ and $d > 2$.

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

▶ **Definition 1** (Phylogenetic network). *A (rooted) phylogenetic network with n leaves is a rooted, simple, directed acyclic graph (DAG) with no nodes of in- and outdegree 1 and exactly n nodes of indegree 1 and outdegree 0 (i.e., leaves) which are bijectively labeled with labels from the set $\{1, \ldots, n\}$.*

In this work, we will only consider phylogenetic networks whose nodes all have either indegree 1 *or* outdegree 1 and whose internal nodes (i.e., nodes that are neither leaves nor the root) with indegree 1 all have outdegree 2 (bifurcating case); the latter nodes will be called *tree nodes*. Finally, we will assume that all internal nodes with outdegree 1 have indegree $d \geq 2$ and these nodes will be called *reticulation nodes*; see Figure 1 for examples with $d = 3$. Note that $d = 2$ is the above mentioned bicombining case.

We next recall the definition of tree-child networks.

▶ **Definition 2** (Tree-child network). *A phylogenetic 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 phylogenetic 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 3** (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 [1] for the bicombining case.

In the following, 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 $k \le n-1$. Thus, the total numbers 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 have the following formula which extends the one for $d = 2$ from Theorem 13 in [1].

▶ **Theorem 4.** *The numbers of one-component d-combining tree-child networks with n leaves and k reticulation nodes for $0 \le k \le n-1$ are given by*

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

and equal to 0 otherwise.

Using this formula, we obtain the following limit law result, where (here and throughout the work) we use $\xrightarrow{w}$ to denote weak convergence.

▶ **Corollary 5.** *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*

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

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

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

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

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

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

(iii) *For $d \ge 4$, the limit law of $n - 1 - R_n^{(d)}$ is degenerate, i.e.,*

$$n - 1 - R_n^{(d)} \xrightarrow{w} 0.$$

▶ Remark 6. If t denotes the number of tree nodes and N the total number of nodes, then by the handshaking lemma, we have

$$t = n + (d-1)k - 1 \qquad \text{and} \qquad N = 2n + dk. \tag{1}$$

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

Note that the above result for $d = 2$ is already contained in the proof of Theorem 3 in [9] where even a local limit theorem was proved; see also [8]. Using the above corollary, we now obtain the first order asymptotics of the total number of one-component tree-child networks.

▶ **Corollary 7.**

(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*

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

where $I_v(\alpha)$ is as in Corollary 5, (ii).

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

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

Again, the result for the case $d = 2$ is already contained in [9]; this is also the only case of the three above in which we find a stretched exponential in the asymptotics (see [3]).

We next turn to general tree-child networks. Here, in contrast to one-component tree-child networks, we only understand the behavior of $\mathrm{TC}_{n,k}^{(d)}$ for fixed k and for $k = n - 1$ (i.e., for maximally reticulated networks).

First, for fixed k, the first order asymptotics in the bicombining case ($d = 2$) was derived in [5, 6] and with a different method in [7]; see also [12] where the result was re-derived with yet another method which is however based on a (yet) unproven conjecture. The approach from [7] can also be used in the d-combining case leading to the following result.

▶ **Theorem 8.** *For the number of d-combining tree-child networks with n leaves and k reticulation nodes, we have for fixed k, as $n \to \infty$,*

$$\mathrm{TC}_{n,k}^{(d)} \sim \frac{2^{(4-d)k-1}}{d!^k\, k!\sqrt{\pi}} n! 2^n n^{(4-d)k-3/2}.$$

▶ Remark 9. Our approach can also be used to compute tables of $\mathrm{TC}_{n,k}^{(d)}$ for small values of n, k, and d; see the Appendix. Moreover, the approach is also capable of giving exact formulas for $\mathrm{TC}_{n,k}^{(d)}$ and small values of k; see [1, 6] for such formulas in the bicombining case.

Note that the asymptotic order in the above theorem is much smaller than the one obtained for one-component tree-child networks. Thus, the majority of tree-child networks do not have a bounded number of reticulation nodes. In fact, the number of reticulation nodes of a "typical" tree-child network is close to the maximum $n - 1$. More precisely, the following result holds.

▶ **Theorem 10.** *For the number of d-combining tree-child networks with n leaves, we have*

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

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

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

For the bicombining case, this result was proved in [9] by encoding tree-child networks with n leaves and $n-1$ reticulation nodes by certain words and (asymptotically) counting these words with the method from [3]. In the more general d-combining case, we will use a similar strategy, however, details will be more demanding due to the dependence on the parameter d.

As for the stochastic behavior of the number of reticulation nodes for general tree-child networks, we have a conjecture for the limit laws which we are going to present in Section 4. Note that even for $d = 2$, no limit law for any shape parameter of random tree-child networks has been established yet.

Our conjecture will clarify the behavior of $\mathrm{TC}_{n,k}^{(d)}$ for k close to n. Thus, the behavior of $\mathrm{TC}_{n,k}^{(d)}$ for small and large k is clear. For the remaining range, there is an interesting recent conjecture for the bicombining case in [12], which has been proved for the special case of one-component tree-child networks in [8]. Whether this conjecture can be extended to d-combining tree-child networks is not clear yet; see the comments in Section 4.

We conclude the introduction with an outline of this extended abstract. In the next section, we will consider one-component networks and prove Theorem 4 and Corollaries 5 and 7. In Section 3, we will discuss our results for general networks. Finally, Section 4 will contain the above mentioned conjecture for the limit laws of the number of reticulation nodes of a random d-combining tree-child network and some concluding remarks.

2 One-Component Networks

In this section, we will prove our results for one-component tree-child networks. We start with Theorem 4.

Proof of Theorem 4. 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; moreover, label it with a label from $\{1,\dots,n\}$ and 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

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

since $\mathrm{OTC}^{(d)}_{n-k,0}$ is the number of phylogenetic trees with $n-k$ leaves; see, e.g., [13, Section 2.1]. ◀

Now, we can prove the two corollaries from above.

Proof of Corollaries 5 and 7. Since the results for $d = 2$ are already contained in [9] (see also [8]), we can focus on the cases $d \geq 3$.

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

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

and this sequence is increasing in k. (This is in contrast to $d = 2$ where this sequence has a maximum at $k = n - \sqrt{n+1}$; see [9].) By replacing k by $n-1-k$ and using Stirling's formula, we obtain that

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

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

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

which is the first claim from Corollary 7, (ii); the second follows from this by another application of Stirling's formula. Moreover, since

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

the result from Corollary 5, (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 (2) now becomes

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

uniformly for k with $k = o(\sqrt{n})$. This expansion, for $d \geq 4$, contains the (non-trivial decreasing) factor $n^{(3-d)k}$ which is responsible for $\mathrm{OTC}^{(d)}_n$ being now asymptotically dominated by $\mathrm{OTC}^{(d)}_{n,n-1}$ (proving Corollary 7, (iii)) and the limiting distribution of $n-1-R^{(d)}_n$ being degenerate (proving Corollary 5, (iii)). ◀

3 General Networks

In this section, we will discuss the asymptotic enumeration of general d-combining tree-child networks with a fixed (Theorem 8) and a maximal number of reticulation nodes (Theorem 10). Note that the latter dominates asymptotically all networks of a given size.

We start with Theorem 8 on d-combining networks with a fixed number of reticulation nodes. It can be proved by generalizing the approach from [7], which was based on the classification of tree-child networks via *component graphs* from [1]. Component graphs can also be defined for d-combining tree-child networks and then be used to prove Theorem 8; details will be given in the journal version of this work. Moreover, component graphs can also be used as in [1] to (a) compute $\mathrm{TC}_{n,k}^{(d)}$ for small values of n, k, and d (see the Appendix); and (b) give explicit formulas for small values of k.

The remainder of this section is devoted to the proof of Theorem 10, which extends the approaches from [3, 9]. We start with some lemmas which are generalizations of the corresponding results from [9] (and are proved with similar arguments).

The first lemma shows that the asymptotic growth of $\mathrm{TC}_n^{(d)}$ is, up to a constant, determined by the asymptotics of $\mathrm{TC}_{n,n-1}^{(d)}$.

▶ **Lemma 11.** *For $n \to \infty$, we have*

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

Proof. Let N be a d-combining tree-child network with n leaves and k reticulation nodes. A *free tree node* is a tree node whose children are both not reticulation nodes; the edges to these children are called *free edges*. Using a simple counting argument, it is easy to see that N has $2(n-k-1)$ free edges; see [9, Lemma 1] for the case $d = 2$.

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

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

Iterating this construction yields

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

and thus,

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

which proves the claim. ◀

Next, we define a generalization of the class of words from [9] which is used to encode d-combining tree-child networks with a maximal number of reticulation nodes.

▶ **Definition 12.** *Let $\mathcal{C}_n^{(d)}$ denote the class of words built from n letters $\{\omega_1, \ldots, \omega_n\}$ in which each letter occurs exactly $d+1$ times such that in every prefix the letter ω_i has either not yet 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 $j > i$.*

In [9], a bijection between bicombining tree-child networks with n leaves and k reticulationd nodes whose labels are removed and $\mathcal{C}_{n-1}^{(2)}$ was proved. In fact, this bijection can be extended to d-combining networks. Then, because the networks are all different when labeling the (now empty) leaves by a random permutation, we obtain the following lemma.

▶ **Lemma 13.** *Let $c_n^{(d)}$ be the cardinality of $\mathcal{C}_n^{(d)}$. Then*

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

Now the recursive nature of this encoding allows us to derive the following counting result.

▶ **Lemma 14.** *We have*

$$c_n^{(d)} = \sum_{m\geq 1} b_{n,m}^{(d)}, \tag{4}$$

where $b_{n,m}^{(d)}$ satisfies the recurrence

$$b_{n,m}^{(d)} = \frac{dn+m-2}{dn+m-d-1} b_{n,m-1}^{(d)} + \binom{dn+m-2}{d-1} b_{n-1,m}^{(d)}, \qquad (n\geq 2, 0\leq m\leq n) \tag{5}$$

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

Proof. First, note that any word in $\mathcal{C}_n^{(d)}$ has a suffix $\omega_n\omega_m\omega_{m+1}\cdots\omega_{n-1}\omega_n$ with $1 \leq m \leq n$. Denote by $b_{n,m}^{(d)}$ the number of these words. Removing the d occurrences of ω_n from these words gives a word of $\mathcal{C}_{n-1}^{(d)}$ 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

$$b_{n,m}^{(d)} = \binom{dn+m-2}{d-1} \sum_{j=1}^{m} b_{n-1,j}^{(d)}, \tag{6}$$

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 12 these first $d-1$ occurrences of ω_n may be anywhere. Differencing yields

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

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

The advantage of the recurrence for $b_{n,m}^{(d)}$ is that we are actually only interested in the asymptotics of $b_{n,n}^{(d)}$ as by (4) and (6) we have

$$b_{n,n}^{(d)} = \binom{(d+1)n-2}{d-1} c_{n-1}^{(d)}.$$

Now we are ready to use the method from [3]. Due to the similarities, we will only discuss the main differences. We start with the following transformation of $(b_{n,m})_{0\leq m\leq n}$ to $(e_{i,j}^{(d)})_{\substack{0\leq i\leq j \\ i-j \text{ even}}}$, which changes the indices and captures the exponential and superexponential terms coming from the binomial coefficient in (5).

▶ **Lemma 15.** *We have*

$$b_{n,m}^{(d)} = \lambda(d)^n (n!)^{d-1} e_{n+m,n-m}^{(d)} \qquad \text{with} \qquad \lambda(d) = \frac{(d+1)^{d-1}}{(d-1)!},$$

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

$$e_{n,m}^{(d)} = \mu_{n,m}^{(d)} e_{n-1,m+1}^{(d)} + \nu_{n,m}^{(d)} e_{n-1,m-1}^{(d)} \tag{7}$$

with

$$\mu_{n,m}^{(d)} = 1 + \frac{2(d-1)}{(d+1)n + (d-1)m - 2(d+1)} \quad \text{and} \quad \nu_{n,m}^{(d)} = \prod_{i=2}^{d} \left(1 - \frac{2(m+i)}{(d+1)(n+m)}\right)$$

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

Note that we are interested in $e_{2n,0}^{(d)} = \frac{b_{n,n}^{(d)}}{\lambda(d)^n (n!)^{d-1}}$ because by the previous lemmas we have

$$\mathrm{TC}_n^{(d)} = \Theta\left(\mathrm{TC}_{n,n-1}^{(d)}\right) = \Theta\left(n! c_{n-1}^{(d)}\right) = \Theta\left(n!\, n^{1-d} b_{n,n}^{(d)}\right) = \Theta\left((n!)^d \lambda(d)^n n^{1-d} e_{2n,0}^{(d)}\right). \tag{8}$$

Moreover, observe that for the Theta-result the initial value of $e_{2,0}^{(d)}$ is irrelevant, as it creates only a constant factor. So we may set it to $e_{2,0}^{(d)} = 1$, or any convenient constant. Note that this recurrence is very similar to that of relaxed trees [3, Equation (2)], yet with more complicated factors. Observe also that this is exactly recurrence [9, Equation (10)] for $d = 2$.

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

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

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 (7). 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. Additionally, the boundary conditions allow to compute c_1 and we get that

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

$a_1 \approx 2.338$ 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 [3, 4, 9]; for the details we refer to the accompanying Maple worksheet [2].

▶ **Proposition 16.** *For all $n, m \geq 0$ let*

$$\tilde{X}_{n,m} := \left(1 - \frac{2d-1}{3(d+1)}\frac{m^2}{n} - \frac{3d^2+12-11}{6(d+1)}\frac{m}{n}\right) \mathrm{Ai}\left(a_1 + \frac{B^{1/3}(m+1)}{n^{1/3}}\right) \quad \text{and}$$
$$\tilde{s}_n := 2 + \frac{a_1 B^{2/3}}{n^{2/3}} - \frac{3d^2-5d+4}{3(d+1)n} - \frac{1}{n^{7/6}}.$$

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 15.

▶ **Proposition 17.** *Choose $\eta > \frac{(2d-1)^2}{18(d+1)^2}$ fixed and for all $n, m \geq 0$ let*

$$\hat{X}_{n,m} := \left(1 - \frac{2d-1}{3(d+1)}\frac{m^2}{n} - \frac{3d^2+12-11}{6(d+1)}\frac{m}{n} + \eta\frac{m^4}{n^2}\right) \mathrm{Ai}\left(a_1 + \frac{B^{1/3}(m+1)}{n^{1/3}}\right) \quad \text{and}$$
$$\hat{s}_n := 2 + \frac{a_1 B^{2/3}}{n^{2/3}} - \frac{3d^2-5d+4}{3(d+1)n} + \frac{1}{n^{7/6}}.$$

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 Theorem 10. 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 16 for *all* $m \leq n$. 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 this, we show 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,

$$\begin{aligned} e^{(d)}_{2n,0} &\geq C_0\tilde{h}_{2n}X_{2n,0} \\ &\geq C_0\prod_{i=1}^{2n}\left(2 + \frac{a_1B^{2/3}}{i^{2/3}} - \frac{3d^2-5d+4}{3(d+1)i} - \frac{1}{i^{7/6}}\right)\mathrm{Ai}\left(a_1 + \frac{B^{1/3}}{n^{1/3}}\right) \\ &\geq C_1(n!)^{d-1}4^n e^{3a_1B^{2/3}n^{1/3}} n^{\frac{d^2+d-2}{2(d+1)}}. \end{aligned}$$

Finally, combining this with (8) we get the lower bound.

The upper bound is similar, yet more technical. The starting point is Proposition 17 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 [3, 4]. Then, using tools from lattice path theory and computer algebra, we show that $e^{(d)}_{2n,0} = \mathcal{O}(\hat{e}^{(d)}_{2n,0})$ and that

$$\hat{e}^{(d)}_{2n,0} \leq \hat{C}_1(n!)^{d-1}4^n e^{3a_1B^{2/3}n^{1/3}} n^{\frac{d^2+d-2}{2(d+1)}}.$$

For more details, see the journal version of this work. ◀

4 Conjectures and Remarks

The main purpose of this paper was to extend recent results on bicombining tree-child networks to d-combining tree-child networks. We did this for both one-component tree-child networks as well as general tree-child networks. For one-component d-combining tree-child networks, we proved an exact counting formula for their number with n leaves and k reticulation nodes. As a consequence of this formula, we obtained limit laws for the number of reticulation nodes of a random network and (asymptotic) counting results for their total number. For general d-combining tree-child networks, our knowledge of their counts is less complete. We derived (asymptotic) results for a fixed number of reticulation nodes and the maximal number of reticulation nodes. The latter implied also an (asymptotic) counting result for their total number.

How about limit laws for the number of reticulation nodes of general d-combining tree-child networks? We think that the upper bound (3) is sharp for $d = 2$ and far away from being sharp for $d \geq 3$. More precisely, we believe that the following conjecture holds.

▶ **Conjecture 18.** *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.*

Moreover, the proof of this conjecture should also give the following result.

▶ **Corollary 19.**

(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 20. 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 [3] is only capable of giving a Theta-result.

In fact, we recently found a method which should allow us to prove these results; details are currently checked. The proofs (if correct) will be presented in the journal version of this paper.

The above limit distribution result would clarify the behavior of the number of general d-combining tree-child networks for a large number, i.e., a number close to n, of reticulation nodes. So, how about the remaining range? (Recall that the number of networks for a small number of reticulation nodes is covered by Theorem 8.)

In this regard, there is a recent interesting conjecture for the bicombining case; see [12]. To give details, denote by $\mathcal{C}_{n,k}^{(2)}$ a class of words which is similar defined as in Definition 12 but with the difference that only k letters appear 3 times while the remaining $n - k$ letters appear 2 times. Let $c_{n,k}^{(2)}$ be their number. Then, it was conjectured in [12], together with some striking consequences, that

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

Can this be extended to d-combining networks? (The obvious generalization by replacing 2 by d does not work.) We do not know the answer to this question yet. However, we have recently managed to define a modification of $\mathcal{C}_{n,k}^{(d)}$ which can be used to encode d-combining tree-child networks. This encoding seems to be useful for the proof of Conjecture 18 and might also shed further light on [12]. Details will be again discussed in the journal version.

Finally, how about extension of our results to multicombining tree-child networks, i.e., tree-child networks where different reticulation nodes may have different number of parents? We think that many of the results of this extended abstract can be generalized to this case, however, notation becomes more cumbersome. We might also include a discussion on this in the journal version of the current paper.

References

1 Gabriel Cardona and Louxin Zhang. Counting and enumerating tree-child networks and their subclasses. *J. Comput. System Sci.*, 114:84–104, 2020.

2 Yu-Sheng Chang, Michael Fuchs, Hexuan Liu, Michael Wallner, and Guan-Ru Yu. Maple worksheet, 2021. `arXiv:2203.07619`.

3 Andrew Elvey Price, Wenjie Fang, and Michael Wallner. Asymptotics of minimal deterministic finite automata recognizing a finite binary language. In *31st International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, volume 159:Art. No. 11, 13 pages of *LIPIcs. Leibniz Int. Proc. Inform.* Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2020.

4 Andrew Elvey Price, Wenjie Fang, and Michael Wallner. Compacted binary trees admit a stretched exponential. *J. Combin. Theory Ser. A*, 177:Paper No. 105306, 40 pages, 2021.

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

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

7 Michael Fuchs, En-Yu Huang, and Guan-Ru Yu. Counting phylogenetic networks with few reticulation vertices: a second approach. `arXiv:2104.07842`.

8 Michael Fuchs, Hexuan Liu, and Guan-Ru Yu. A short note on the exact counting of tree-child networks. arXiv:2110.03842. `arXiv:2110.03842`.

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

10 Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 1st edition, 2010.

11 Colin McDiarmid, Charles Semple, and Dominic Welsh. Counting phylogenetic networks. *Ann. Comb.*, 19(1):205–224, 2015.

12 Miquel Pons and Josep Batle. Combinatorial characterization of a certain class of words and a conjectured connection with general subclasses of phylogenetic tree-child networks. *Sci. Rep.*, 11:Article number: 21875, 2021.

13 Mike Steel. *Phylogeny – discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

A Appendix

Table 1 $\mathrm{TC}^{(2)}_{n,k}$ for $2 \le n \le 8$ and $0 \le k < n$; see also [1].

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

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

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

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

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