Galled 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

Guan-Ru Yu
Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804, Taiwan

Abstract

We propose the class of *galled tree-child networks* which is obtained as intersection of the classes of galled networks and tree-child networks. For the latter two classes, (asymptotic) counting results and stochastic results have been proved with very different methods. We show that a counting result for the class of galled tree-child networks follows with similar tools as used for galled networks, however, the result has a similar pattern as the one for tree-child networks. In addition, we also consider the (suitably scaled) numbers of reticulation nodes of random galled tree-child networks and show that they are asymptotically normal distributed. This is in contrast to the limit laws of the corresponding quantities for galled networks and tree-child networks which have been both shown to be discrete.

2012 ACM Subject Classification Mathematics of computing → Enumeration; Mathematics of computing → Distribution functions

Keywords and phrases Phylogenetic Network, galled Network, tree-child Network, asymptotic Enumeration, Limit Law, Lagrange Inversion

Digital Object Identifier 10.4230/LIPIcs.AofA.2024.2

Funding *Yu-Sheng Chang*: NSTC-111-2115-M-004-002-MY2
Michael Fuchs: NSTC-111-2115-M-004-002-MY2
Guan-Ru Yu: NSTC-110-2115-M-017-003-MY3

Acknowledgements We thank the three reviewers for positive feedback and useful suggestions. This paper is dedicated to Hsien-Kuei Hwang on the occasion of his 60th birthday.

1 Introduction

Phylogenetic networks are used to visualize, model, and analyze the ancestor relationship of taxa in reticulate evolution. To make them more relevant for biological applications as well as devise algorithms for them, many subclasses of the class of phylogenetic networks have been proposed; see the comprehensive survey [14]. A lot of recent research work was concerned with fundamental questions such as counting them and understanding the shape of a network drawn uniformly at random from a given class; see, e.g., [2, 3, 4, 8, 9, 11, 12, 10, 13, 15, 16]. Despite this, even counting results are still missing for most of the major classes of phylogenetic networks. Two notable exceptions are tree-child networks and galled networks for which such results have been proved in [11, 12]. In this work, we consider the intersection of these two network classes. We start with some basic definitions and then explain why we find this class interesting.

First, a phylogenetic network is defined as follows.

▶ **Definition 1** (Phylogenetic Network). *A (rooted) phylogenetic network of size n is a rooted, simple, directed, acyclic graph whose nodes fall into the following three (disjoint) categories:*

(a) *A unique root which has indegree 0 and outdegree 1;*

35th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2024).
Editors: Cécile Mailler and Sebastian Wild; Article No. 2; pp. 2:1–2:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

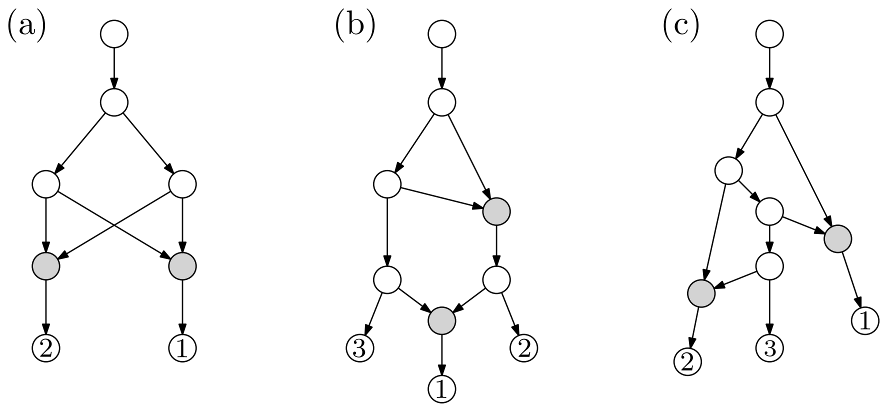


Figure 1 (a) A galled network which is not tree-child; (b) A tree-child network which is not galled; (c) A galled tree-child network.

(b) *Leaves which have indegree* 1 *and outdegree* 0 *and are bijectively labeled with labels from the set* $\{1, \ldots, n\}$*;*

(c) *Internal nodes which have indegree and outdegree at least* 1 *and total degree at least* 3*.*

Moreover, a phylogenetic network is called binary if all internal nodes have either indegree 1 *and outdegree* 2 *(tree nodes) or indegree* 2 *and outdegree* 1 *(reticulation nodes).*

▶ Remark 2. **(i)** Phylogenetic networks with all internal nodes having indegree equal to 1 are called *phylogentic trees*. They have been used as visualization tool in evolutionary biology at least since Darwin.

(ii) If not explicitly mentioned, phylogenetic networks are always binary in the sequel.

We next define galled networks and tree-child networks which are two of the major classes of phylogenetic networks. (The former has been introduced for computational reasons, the latter because of its biological relevance; see [14].) For the definition, we need the notion of a *tree cycle* which is a pair of edge-disjoint paths in a phylogenetic network that start at a common tree node and end at a common reticulation node with all other nodes being tree nodes.

▶ **Definition 3.** **(a)** *A phylogenetic network is called a tree-child network if every non-leaf node has at least one child which is either a tree node or a leaf.*

(b) *A phylogenetic network is called a galled network if every reticulation node is in a (necessarily unique) tree cycle.*

▶ Remark 4. Note that neither the class of tree-child networks is contained in the class of galled networks nor vice versa; see Figure 1.

Let $\mathrm{TC}_{n,k}$ and $\mathrm{GN}_{n,k}$ denote the number of tree-child networks and galled networks of size n with k reticulation nodes, respectively. It is not hard to see that $k \le n-1$ for tree-child networks and $k \le 2n-2$ for galled networks where both bounds are sharp; see, e.g., [11, 12]. Thus, the total numbers are given by:

$$\mathrm{TC}_n := \sum_{k=0}^{n-1} \mathrm{TC}_{n,k} \qquad \text{and} \qquad \mathrm{GN}_n := \sum_{k=0}^{2n-2} \mathrm{GN}_{n,k}. \tag{1}$$

The asymptotic growth of both of these sequences is known. First, in [11], it was proved that for the number of tree-child networks, as $n \to \infty$,

$$\mathrm{TC}_n = \Theta\left(n^{-2/3} e^{a_1(3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\right), \tag{2}$$

where a_1 is the largest root of the Airy function of the first kind. The surprise here was the presence of a *stretched exponential* in the asymptotic growth term. On the other hand, no stretched exponential is contained in the asymptotics of the number of galled networks. More precisely, it was proved in [12] that, as $n \to \infty$,

$$\mathrm{GN}_n \sim \frac{\sqrt{2e\sqrt[4]{e}}}{4} n^{-1} \left(\frac{8}{e^2}\right)^n n^{2n}. \tag{3}$$

The tools used to establish (2) and (3) were very different: for (2), a bijection to a class of words was proved and a recurrence for these word was found which could be (asymptotically) analyzed with the approach from [6]; for (3), the component graph method introduced in [13] together with the Laplace method and a result from [1] was used.

Another difference was the location in (1) of the terms which dominate the two sums. For tree-child networks, the main contribution comes from networks with k close to $n-1$ (the maximally reticulated networks), whereas for galled networks, the main contributions comes from networks with $k \approx n$. In fact, the limit law of the number of reticulation nodes, say R_n, was derived in [5, 12] for both network classes if a network of size n is sampled uniformly at random. More precisely, for tree-child networks, it was shown in [5] that, as $n \to \infty$,

$$n - 1 - R_n \xrightarrow{d} \mathrm{Poisson}(1/2),$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\mathrm{Poisson}(\lambda)$ is a Poisson law with parameter λ. A similar discrete limit law was proved in [12] for galled networks. More precisely, it was shown that, as $n \to \infty$,

$$\mathbb{E}(R_n) = n - \frac{3}{8} + o(1)$$

and that the limit law of $n - R_n$ is not Poisson but a mixture of Poisson laws; see Theorem 2 in [12] for more details.

Due to the above results and differences, one wonders how the intersection of the class of tree-child networks and galled networks behaves?

▶ **Definition 5** (Galled Tree-Child Network)**.** *A galled tree-child network is a network which is both a galled network and a tree-child network.*

Let $\mathrm{GTC}_{n,k}$ denote the number of galled tree-child networks of size n with k reticulation nodes. We show below that again k has the sharp upper bound $n-1$. (See Lemma 19 in Section 3.) Set:

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

Then, this sequence has the following first-order asymptotics.

▶ **Theorem 6.** *For the number of galled tree-child networks, we have, as $n \to \infty$,*

$$\mathrm{GTC}_n \sim \frac{1}{2\sqrt[4]{e}} n^{-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n n^{2n}.$$

▶ Remark 7. Note that the asymptotic expansion contains a stretched exponential as does the expansion (2) for tree-child networks, however, the proof will use the tools which were developed in [12] to derive (3) for galled networks.

We next consider the number of reticulation nodes R_n of a *random galled tree-child network* which is a galled tree-child network of size n that is sampled uniformly at random from the set of all galled tree-child networks of size n. In contrast to tree-child networks and galled networks, the limit law of R_n (suitably scaled) is continuous.

▶ **Theorem 8.** *The number of reticulation nodes R_n of a random galled tree-child networks satisfies, as $n \to \infty$,*

$$\frac{R_n - \mathbb{E}(R_n)}{\sqrt{\mathrm{Var}(R_n)}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ denotes the standard normal distribution. Moreover, as $n \to \infty$,

$$\mathbb{E}(R_n) = n - \sqrt{n} + o(\sqrt{n}) \qquad \text{and} \qquad \mathrm{Var}(R_n) \sim \sqrt{n}/2.$$

The above results show that galled tree-child networks behave quite different from both tree-child networks and galled networks. That is one reason why we find them interesting.

Another reason stems from a recent result which was proved in [4]. In the latter paper, the asymptotics of $\mathrm{GN}_{n,k}$ for fixed k was derived. Let $\mathrm{PN}_{n,k}$ denote the number of phylogenetic networks of size n and k reticulation nodes. (Note that this number is finite, whereas it becomes infinite when summing over k.) Then, one of the main results from [4] implies that for fixed k, as $n \to \infty$,

$$\mathrm{PN}_{n,k} \sim \mathrm{TC}_{n,k} \sim \mathrm{GN}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}. \tag{4}$$

(The first two asymptotic equivalences were proved in [10, 15].) That $\mathrm{TC}_{n,k}$ and $\mathrm{GN}_{n,k}$ have the same first-order asymptotics for fixed k was a surprise since the classes of tree-child networks and galled networks are quite different, e.g., neither contains the other; see Remark 4. However, the above result can be explained via the class of galled tree-child networks as will be seen in Section 3 below.

We conclude the introduction with a short sketch of the paper. The proofs of Theorem 6 and Theorem 8 follow with a similar approach as used for galled networks in [11]. This approach is based on the component graph method from [13] which we recall in the next section. Then, in Section 3, we consider $\mathrm{GTC}_{n,k}$ for small and large values of k. Finally, Section 4 contains the proofs of our main results (Theorem 6 and Theorem 8). We conclude the paper with some final remarks in Section 5.

2 The Component Graph Method

The component graph method for galled networks was introduced in [13] and used in [4, 12] to prove asymptotic results. It is explained in detail in all these papers. However, to make the current paper more self-contained, we briefly recall it.

Let N be a galled network. Then, by removing all the edges leading to reticulation vertices (these are the so-called *reticulation edges*), we obtain a forest whose trees are called the *tree-components* of N.

The *component graph* of N, denoted by $C(N)$, is now a directed, acyclic graph which has a vertex for every tree-component. Moreover, the vertices are connected by the removed

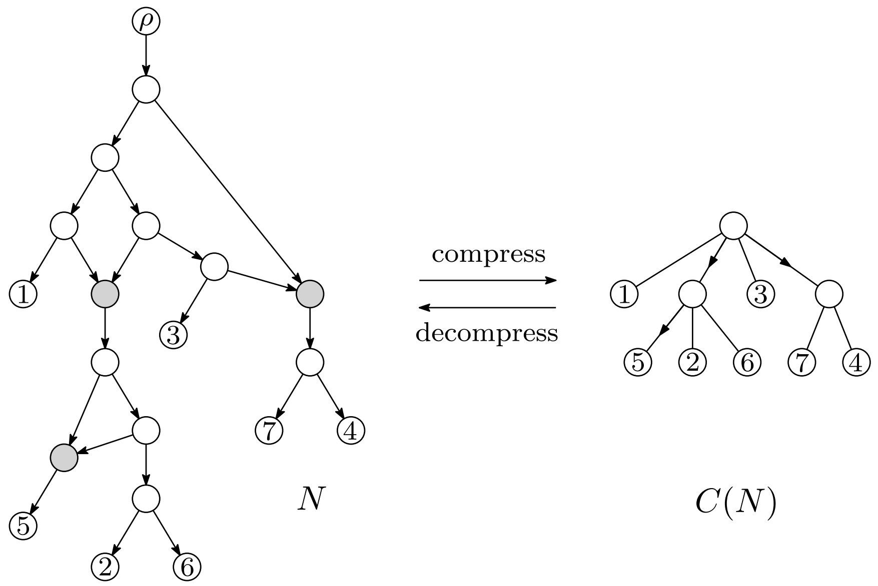


Figure 2 A galled network N and its component graph $C(N)$ which is a phylogenetic tree.

reticulation edges in the same way as the tree-components have been connected by them. Finally, we attach the leaves in the tree-components to the corresponding vertices in $C(N)$ unless a vertex v of $C(N)$ is a terminal vertex and its corresponding tree-component has exactly one leaf, in which case we use the label of that leaf to label v. Note that $C(N)$ may contain double edges. We replace such a double edge by a single edge and indicate that it was a double edge by placing an arrow on it; see Figure 2 for a galled network together with its component graph. Also, denote by $\tilde{C}(N)$ the component graph of $C(N)$ with all arrows on edges removed. Then, the authors of [13] made the following important observation.

▶ **Proposition 9** ([13]). *N is a galled network if and only if $\tilde{C}(N)$ is a (not necessarily binary) phylogenetic tree.*

▶ Remark 10. By this result, for a galled network N, $C(N)$ must have arrows on all internal edges (i.e., all edges whose two endpoints are both internal nodes).

The component graph can be seen as a kind of compression of N that retains some but not all structural properties of N. Indeed, different networks N might share the same component graph. However, we can generate all galled networks of size n from a list of all component graphs (i.e., phylogenetic trees) with n labeled leaves by a decompression procedure which is explained below.

First, we need the notion of *one-component networks.*

▶ **Definition 11** (One-component Network). *A phylogenetic network is called a one-component network if every reticulation node has a leaf as its child.*

▶ Remark 12. The name comes from the fact that one-component networks only have one non-trivial tree-component.

Now, let a component graph C of a galled tree-child network be given. We do a breadth-first traversal of the internal vertices of C and replace these vertices v by a one-component galled network O_v whose leaves below reticulation vertices are labeled with the first k labels, where k is the number of outgoing edges of v in C that have an arrow on them, and whose size is equal to the outdegree $c(v)$ of v. (In order to avoid confusion, the labels of O_v are subsequently assumed to be from the set $\{\overline{1},\dots,\overline{c(v)}\}$.) Then, attach the subtrees rooted at the children of v which are connected to v by edges with arrows on them to the leaves of O_v with labels $\{\overline{1},\dots,\overline{k}\}$, where the subtree with the smallest label is attached to $\overline{1}$, the subtree with the second smallest label is attached to $\overline{2}$, etc. Moreover, relabel the remaining leaves of O_v, namely the ones with the labels $\{\overline{k+1},\dots,\overline{c(v)}\}$, by the remaining labels of the subtrees of v (which are all of size 1, i.e., they are leaves in C) in an order-consistent way. By using all possible one-component galled networks in every step, this gives all possible galled networks with C as component graph. Moreover, if we start from $\tilde{C}$, then we first have to place arrows on all edges whose heads are internal nodes of $\tilde{C}$ (see Remark 10) and for all remaining edges, we can freely decide if we want to place an arrow on them or not. Overall, this gives the following result which was one of the main results in [13].

▶ **Proposition 13** ([13]). *We have,*

$$\mathrm{GN}_n = \sum_{\mathcal{T}} \prod_{v} \sum_{j=0}^{c_{\mathrm{lf}}(v)} \binom{c_{\mathrm{lf}}(v)}{j} M_{c(v),c(v)-c_{\mathrm{lf}}(v)+j},$$

where the first sum runs over all (not necessarily binary) phylogenetic trees $\mathcal{T}$ of size n, the product runs over all internal nodes of $\mathcal{T}$, $c(v)$ is the outdegree of v, $c_{\mathrm{lf}}(v)$ is the number of children of v which are leaves, and $M_{n,k}$ denotes the number of one-component galled networks of size n with k reticulation vertices, where the leaves below the reticulation vertices are labeled with labels from the set $\{1,\dots,k\}$.

For galled tree-child networks, it is now clear that the same formula holds with the only difference that $M_{n,k}$ has to be replaced by the corresponding number of one-component galled tree-child networks. However, this number is the same as the number of one-component tree-child networks.

▶ **Lemma 14.** *Every one-component tree-child network is a one-component galled tree-child network.*

Proof. Let v be a reticulation vertex and consider a pair of edge-disjoint paths from a common tree vertex to v. (Note that such a pair trivially exists.) Then, no internal vertex can be a reticulation vertex because such a reticulation vertex would not be followed by a leaf. Thus, v is in a tree cycle which shows that the network is indeed galled. ◀

Denote by $B_{n,k}$ the number of one-component tree-child networks of size n and k reticulation vertices, where the labels of the leaves below the reticulation vertices are $\{1,\dots,k\}$. Then, we have the following analogous result to Proposition 13.

▶ **Proposition 15.** *We have,*

$$\mathrm{GTC}_n = \sum_{\mathcal{T}} \prod_{v} \sum_{j=0}^{c_{\mathrm{lf}}(v)} \binom{c_{\mathrm{lf}}(v)}{j} B_{c(v),c(v)-c_{\mathrm{lf}}(v)+j}, \tag{5}$$

where notation is as in Proposition 13 and $B_{n,k}$ was defined above.

▶ Remark 16. Using this result, by systematically generating all (not necessarily binary) phylogenetic trees of size n and computing $B_{n,k}$ with the closed-form expression below, we obtain the following table for small values of n:

n	GTC_n
1	1
2	3
3	48
4	1,611
5	87,660
6	6,891,615
7	734,112,540
8	101,717,195,895
9	17,813,516,259,420
10	3,857,230,509,496,875

Table 1 The values of GTC_n for $1 \le n \le 10$.

We will deduce all our results from (5). In addition, we make use of the following results for $B_{n,k}$ which were proved in [3] and [11]. To state them, denote by $\mathrm{OTC}_{n,k}$ the number of one-component tree-child networks of size n with k reticulation vertices and by OTC_n the (total) number of one-component tree-child networks of size n. Then,

$$\mathrm{OTC}_{n,k} = \binom{n}{k} B_{n,k} \tag{6}$$

and

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

(Note that the tree-child property implies the $k \le n-1$ and this bound is sharp.)

▶ **Proposition 17** ([3, 11]). **(i)** *We have,*

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

(ii) *As $n \to \infty$,*

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

where $k = n - \sqrt{n} + x$ and $x = o(n^{1/3})$.

The second result above gives a local limit theorem (see, e.g., Section IX.9 in [7]) for the (random) number of reticulation vertices of a one-component tree-child network of size n which is picked uniformly at random from all one-component tree-child networks of size n. It implies the following (asymptotic) counting result for OTC_n.

▶ **Corollary 18** ([11]). *As $n \to \infty$,*

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

3 Networks with Few and Many Reticulation Nodes

In this section, we consider $\mathrm{GTC}_{n,k}$ for small and large k. We start with large k.

As mentioned in the last section, for tree-child networks, we have that $k \le n-1$ and this bound is sharp. Clearly, this implies that $k \le n-1$ also holds for galled tree-child networks. Again this bound is sharp. We summarize this in the following lemma.

▶ **Lemma 19.** *The number of reticulation vertices of a galled tree-child network of size n is at most $n-1$ where this bound is sharp.*

Proof. Let $\tilde{C}$ be the component graph of a galled tree-child network of size n which by Proposition 9 is a phylogenetic tree. The maximal number of reticulation vertices of a network decompressed from $\tilde{C}$ is achieved by placing the maximal number of arrows at all outgoing edges of internal vertices v of $\tilde{C}$. Note that this number is $c(v)-1$, where $c(v)$ denotes the degree of v, since placing arrows on all outgoing edges is not possible because $B_{c(v),c(v)} = 0$ (as $B_{n,k}$ denotes the number of certain one-component tree-child networks and $k \le n-1$). Thus, the maximal number of reticulation vertices equals

$$\sum_v (c(v)-1) = \sum_v c(v) - (\#\text{ internal nodes of } \tilde{C}), \tag{7}$$

where the sums run over all internal vertices of $\tilde{C}$. By the handshake lemma,

$$\sum_v c(v) = (\#\text{ internal nodes of } \tilde{C} - 1) + n$$

which, by plugging into (7), gives the claimed result. ◀

The proof of the last lemma also reveals the structure of maximally reticulated galled tree-child networks of size n: They are obtained by decompressing component graphs $\tilde{C}$ that are phylogenetic trees of size n with at least one leaf ℓ attached to every internal vertex v by placing arrows on all outgoing edges of v except the one leading to ℓ. This can be translated into generating functions. Set:

$$M(z) := \sum_{n\ge1} \mathrm{GTC}_{n,n-1}\frac{z^n}{n!}, \qquad B(z) := \sum_{n\ge1} B_{n,n-1}\frac{z^n}{n!} = \sum_{n\ge1}\frac{(2n-2)!}{2^{n-1}n!}z^n,$$

where the last line follows from (6) and Proposition 17-(i). Then, we have the following result.

▶ **Lemma 20.** *We have,*

$$M(z) = z + zB'(M(z)). \tag{8}$$

Proof. According to the explanation in the paragraph preceding the lemma, a maximally reticulated galled tree-child network is either a leaf or obtained from a maximally reticulated one-component tree-child network with the leaves below the reticulation vertices replaced by maximally reticulated galled tree-child networks. This translates into

$$M(z) = z + \sum_{n\ge1} B_{n,n-1}\frac{zM(z)^{n-1}}{(n-1)!},$$

where the z inside the sum counts the leaf which is not below the reticulation vertex and the factor $1/(n-1)!$ discards the order of the maximally reticulated galled tree-child networks (counted by $M(z)^{n-1}$) which are attached to the children below the reticulation vertices. The claimed result follows from this. ◀

Note that (8) is of *Lagrangian type*. Thus, we can obtain the asymptotics of $\mathrm{GTC}_{n,n-1}$ by applying Lagrange's inversion formula and the following result from [1].

▶ **Theorem 21** ([1]). *Let $S(z)$ be a formal power series with $s_0 = 0, s_1 \neq 0$ and $ns_{n-1} \sim \gamma s_n$. Then, for $\alpha \neq 0$ and β real numbers,*

$$[z^n](1 + S(z))^{\alpha n+\beta} \sim \alpha e^{\alpha s_1 \gamma} n s_n.$$

▶ **Theorem 22.** *The number of maximally reticulated galled tree-child networks $\mathrm{GTC}_{n,n-1}$ satisfies, as $n \to \infty$,*

$$\mathrm{GTC}_{n,n-1} \sim \sqrt{e\pi} n^{-1/2} \left(\frac{2}{e^2}\right)^n n^{2n}.$$

▶ Remark 23. For tree-child networks, it was proved in [11] that $\mathrm{TC}_n = \Theta(\mathrm{TC}_{n,n-1})$. (This was a main step in the proof of (2).) The above result together with Theorem 6 shows that the same is not true for galled tree-child networks.

Proof. Applying the Lagrange inversion formula to (8) gives

$$\mathrm{GTC}_{n,n-1} = n![z^n]M(z) = (n-1)![\omega^{n-1}](1 + B'(\omega))^n. \tag{9}$$

Next, by Stirling's formula, as $n \to \infty$,

$$[z^n]B'(z) = \frac{B_{n+1,n}}{n!} = \frac{(2n)!}{2^n n!} \sim \sqrt{2}\left(\frac{2}{e}\right)^n n^n.$$

Thus, we can apply Theorem 21 to (9) with $\gamma = 1/2$ and obtain that, as $n \to \infty$,

$$\mathrm{GTC}_{n,n-1} \sim \sqrt{e} n B_{n,n-1} = \sqrt{e} n \frac{(2n-2)!}{2^{n-1}} \sim \sqrt{e\pi} n^{-1/2} \left(\frac{2}{e^2}\right)^n n^{2n}.$$

This is the claimed result. ◀

We next consider $\mathrm{GTC}_{n,k}$ with k small, i.e., the other extreme case of the number of reticulation vertices. Here, we have the following result which shows that the distribution of a uniformly chosen phylogenetic network with n leaves and k reticulation nodes concentrates on the set of galled tree-child networks. This explains why the asymptotic expansions of $\mathrm{TC}_{n,k}$ and $\mathrm{GN}_{n,k}$ in (4) are the same. (It would be interesting to know whether or not this distribution concentrates on an even smaller set.)

▶ **Theorem 24.** *For fixed k, as $n \to \infty$,*

$$\mathrm{GTC}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!} \left(\frac{2}{e}\right)^n n^{n+2k-1}. \tag{10}$$

The proof of this result uses ideas from [10].

Proof. First consider galled tree-child networks of size n which are obtained by decompressing phylogenetic trees of size n which have all k arrows on the edges from the root, i.e., the root has at least one leaf and all other children are either internal nodes or leaves (with at most k internal nodes) and all internal nodes have just leaves as children. By Proposition 8 in [10], the number of these galled tree-child network has the same asymptotics as the one on the right-hand side of (10). Moreover, these networks also dominate the asymptotics in the case of tree-child networks. Thus, the remaining galled tree-child networks are asymptotically negligible as their number is bounded above by the number of the remaining tree-child networks. ◀

▶ Remark 25. Note that this re-proves the (surprising) asymptotic result for $\mathrm{GN}_{n,k}$ in (4) from [4]. On the other hand, the above asymptotic result could be also deduced from (4). In order to explain this, denote by $\mathcal{PN}_{n,k}$ (resp. $\mathcal{TC}_{n,k}/\mathcal{GN}_{n,k}/\mathcal{GTC}_{n,k}$) the set of all phylogenetic networks (resp. tree-child networks/galled networks/galled tree-child networks) with n leaves and k reticulation nodes. Then,

$$\begin{aligned}|\mathcal{TC}_{n,k}\cup\mathcal{GN}_{n,k}| &= |\mathcal{TC}_{n,k}|+|\mathcal{GN}_{n,k}|-|\mathcal{TC}_{n,k}\cap\mathcal{GN}_{n,k}|\\ &= \mathrm{TC}_{n,k}+\mathrm{GN}_{n,k}-\mathrm{GTC}_{n,k}\end{aligned}$$

and $|\mathcal{TC}_{n,k}\cup\mathcal{GN}_{n,k}|\le \mathrm{PN}_{n,k}$. From this the asymptotic result for $\mathrm{GTC}_{n,k}$ follows from those of (4). (We are thankful to one of the reviewers for this remark.)

4 Proof of the Main Results

In this section, we first prove Theorem 6 and then state a result which implies Theorem 8.

For the proof of Theorem 6, we closely follow the method of proof of (3) from [12]. The main idea is to use (5) to find asymptotic matching upper and lower bounds for GTC_n.

First, for an upper bound, we pick a (not necessarily binary) phylogenetic tree $\mathcal{T}$ of size n (which is considered to be a component graph of a galled tree-child network of size n) and decompress it by picking for internal vertices v of $\mathcal{T}$ *any* one-component tree-child network of size $c(v)$ (where the notation is as in Proposition 13). Since, as explained in Section 2, actually only certain one-component tree-child networks are permissible, this modified decompression procedure overcounts the number of galled tree-child networks of size n. More precisely, we consider

$$U_n := \sum_{\mathcal{T}}\prod_{v}\mathrm{OTC}_{c(v)},$$

where the first sum runs over all phylogenetic trees $\mathcal{T}$ of size n and the product runs over internal vertices of $\mathcal{T}$. Then, we have $\mathrm{GTC}_n\le U_n$. Next, set

$$U(z):=\sum_{n\ge1}U_n\frac{z^n}{n!},\qquad A(z):=\sum_{n\ge1}\mathrm{OTC}_{n+1}\frac{z^n}{(n+1)!}.$$

Then, the definition of U_n implies the following result.

▶ **Lemma 26.** *We have,*

$$U(z)=z+U(z)A(U(z)).$$

Proof. The networks counted by U_n are either a leaf or a one-component tree-child network with n leaves which are replaced by an unordered sequence of networks of the same type. This gives

$$U(z)=z+\sum_{n\ge2}\mathrm{OTC}_n\frac{U(z)^n}{n!}$$

from which the claimed result follows. ◀

Now, we can proceed as in the proof of Theorem 22 to obtain the following asymptotic result for U_n.

▶ **Proposition 27.** *As $n \to \infty$,*

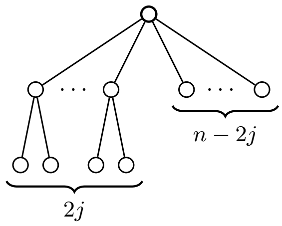
$$U_n \sim \frac{1}{2\sqrt[4]{e}} n^{-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n n^{2n}.$$

Proof. From Lemma 26 and the Lagrange inversion formula,

$$U_n = n![z^n]U(z) = (n-1)![\omega^{n-1}](1 - A(\omega))^{-n}.$$

The result follows from this by applying Theorem 21 and Corollary 18. ◀

Next, we need a matching lower bound. Therefore, we consider (5) with the first sum restricted to phylogenetic trees of the shape (where we have removed the leaf labels):



We denote the resulting term by L_n. The decompression procedure from Section 2 then gives the following result.

▶ **Lemma 28.** *We have,*

$$\begin{aligned} L_n &= \sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n}{2j} \frac{(2j)!}{j!2^j} \sum_{\ell=0}^{n-2j} \binom{n-2j}{\ell} L_{n-j,j+\ell} \\ &= \sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n}{2j} \frac{(2j)!}{j!2^j} \sum_{\ell=0}^{n-2j} \binom{n-2j}{\ell} \frac{(2n-2j-2)!}{2^{n-j-1}(n-2j-\ell-1)!}. \end{aligned} \tag{11}$$

Proof. The first equality is explained as in the proof of Lemma 9 in [12] and the second equality follows from (6) and Proposition 17-(i). ◀

From this result, we can deduce (matching) first-order asymptotics for L_n which then together with the asymptotics of the upper bound (Proposition 27) concludes the proof of Theorem 6.

▶ **Proposition 29.** *As $n \to \infty$,*

$$L_n \sim \frac{1}{2\sqrt[4]{e}} n^{-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n n^{2n}.$$

Sketch of the proof. From Stirling's formula (similar to the proof of Proposition 17-(ii)),

$$\binom{n-2j}{\ell} \frac{(2n-2j-2)!}{2^{n-j-1}(n-2j-\ell-1)!} \sim \frac{1}{2^{j+1}\sqrt{e\pi}} n^{-3/2} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^j n^{2n-2j} e^{-x^2/\sqrt{n}},$$

where $k = n - \sqrt{n} + x$ and this holds uniformly for $|x| \leq n^{1/2+\epsilon}$ and $j \leq n^\epsilon$ with $\epsilon > 0$ arbitrarily small. Using the Laplace method then gives,

$$\sum_{\ell=0}^{n-2j} \binom{n-2j}{\ell} \frac{(2n-2j-2)!}{2^{n-j-1}(n-2j-\ell-1)!} \sim \frac{1}{2^{j+1}\sqrt{e}} n^{-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n n^{2n-2j}$$

uniformly for $j \leq n^{\epsilon}$ for arbitrarily small $\epsilon > 0$. Finally, by plugging the last relation into (11),

$$L_n \sim \frac{1}{2\sqrt{e}}\left(\sum_{j\geq 0}\frac{1}{j!4^j}\right) n^{-5/4} e^{2\sqrt{n}} \left(\frac{2}{e^2}\right)^n n^{2n}$$

which gives the claimed result. ◀

▶ Remark 30. Note that this proposition shows that a "typical" galled tree-child network of size n is obtained by decompressing component graphs of the form given before Lemma 28. This implies, e.g., that the Sackin index defined in [17] of a galled tree-child network has the unusual expected order $n^{7/4}$.

Finally, by refining the above method (see Section 6 of [12] where the same was done for galled networks), we obtain the following result which implies our second main result (Theorem 8).

▶ **Theorem 31.** *Let I_n be the number of reticulation vertices of a random galled tree-child network of size n which are not followed by a leaf and R_n be the total number of reticulation vertices. Then, as $n \to \infty$,*

$$\left(I_n, \frac{R_n - n + \sqrt{n}}{\sqrt[4]{n/4}}\right) \xrightarrow{d} (I, R),$$

where I and R are independent with $I \stackrel{d}{=} \mathrm{Poisson}(1/4)$ and $R \stackrel{d}{=} N(0,1)$.

5 Conclusion

In this paper, we introduced the class of *galled tree-child networks* which is obtained as intersection of the classes of galled networks and tree-child networks. Our reason for doing so was two-fold: (i) Different tools have been used to prove results for galled networks and tree-child networks [11, 12]; consequently, we were curious about which tools apply to the combination of these classes? (ii) It was recently proved that the number of galled networks and tree-child networks have the same first-order asymptotics when the number of reticulation vertices is fixed [4, 10]. Why is that the case?

As for (i), we showed that an asymptotic counting result for galled tree-child networks (Theorem 6) can be obtained with the methods for galled networks, however, the result contains a stretched exponential as does the asymptotic result for tree-child networks. In addition, we showed that the number of reticulation vertices for a random galled tree-child networks is asymptotically normal (Theorem 8), whereas the limit laws of the same quantities for galled networks and tree-child networks were discrete. As for (ii), we showed that the number of galled tree-child networks also satisfies the same first order asymptotics when the number of reticulation vertices is fixed. This explains the previous results from [4, 10].

Overall, the class of galled tree-child networks is interesting and thus merits further examination. In particular, due to Remark 30, studying the shape of random galled tree-child networks seems to be more feasible than studying the shape of random networks from other network classes because such a study boils down to the easier task of studying the shape of one-component tree-child networks which have a straightforward recursive decomposition that, e.g., resulted in a closed-form expression for their numbers; see [17]. The latter paper, where one-component tree-child networks are called *simplex networks*, e.g., asks for properties of the height and such results would immediately entail corresponding results for random galled tree-child networks. (Studying the height is an open problem for most classes of phylogenetic networks.) We may come back to this question in future work.

References

1 Edward A. Bender and L. Bruce Richmond. An asymptotic expansion for the coefficients of some power series. II. Lagrange inversion. *Discrete Math.*, 50(2-3):135–141, 1984. doi:10.1016/0012-365X(84)90043-8.

2 Mathilde Bouvel, Philippe Gambette, and Marefatollah Mansouri. Counting phylogenetic networks of level 1 and 2. *J. Math. Biol.*, 81(6-7):1357–1395, 2020. doi:10.1007/s00285-020-01543-5.

3 Gabriel Cardona and Louxin Zhang. Counting and enumerating tree-child networks and their subclasses. *J. Comput. System Sci.*, 114:84–104, 2020. doi:10.1016/j.jcss.2020.06.001.

4 Yu-Sheng Chang and Michael Fuchs. Counting phylogenetic networks with few reticulation vertices: Galled and recticulation-visible networks. *Bull. Math. Biol.* to appear.

5 Yu-Sheng Chang, Michael Fuchs, Hexuan Liu, Michael Wallner, and Guan-Ru Yu. Enumerative and distributional results for d-combining tree-child networks. *Adv. in Appl. Math.*, 157:Paper No. 102704, 2024. doi:10.1016/j.aam.2024.102704.

6 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, 2021. doi:10.1016/j.jcta.2020.105306.

7 Philipp Flajolet and Robert Sedgewick. *Analytic Combinatorics.* Cambridge University Press, 2009.

8 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.

9 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.

10 Michael Fuchs, En-Yu Huang, and Guan-Ru Yu. Counting phylogenetic networks with few reticulation vertices: a second approach. *Discrete Appl. Math.*, 320:140–149, 2022. doi:10.1016/j.dam.2022.03.026.

11 Michael Fuchs, Guan-Ru Yu, and Louxin Zhang. On the asymptotic growth of the number of tree-child networks. *European J. Combin.*, 93:Paper No. 103278, 20, 2021. doi:10.1016/j.ejc.2020.103278.

12 Michael Fuchs, Guan-Ru Yu, and Louxin Zhang. Asymptotic enumeration and distributional properties of galled networks. *J. Combin. Theory Ser. A*, 189:Paper No. 105599, 28, 2022. doi:10.1016/j.jcta.2022.105599.

13 Andreas D. M. Gunawan, Jeyaram Rathin, and Louxin Zhang. Counting and enumerating galled networks. *Discrete Appl. Math.*, 283:644–654, 2020. doi:10.1016/j.dam.2020.03.005.

14 Sungsik Kong, Joan Carles Pons, Laura Kubatko, and Kristina Wicke. Classes of explicit phylogenetic networks and their biological and mathematical significance. *J. Math. Biol.*, 84(6):Paper No. 47, 44, 2022. doi:10.1007/s00285-022-01746-y.

15 Marefatollah Mansouri. Counting general phylogenetic networks. *Australas. J. Combin.*, 83:40–86, 2022.

16 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.

17 Louxin Zhang. The Sackin index of simplex networks. In *Comparative genomics*, volume 13234 of *Lecture Notes in Comput. Sci.*, pages 52–67. Springer, Cham, [2022] ©2022. URL: https://doi.org/10.1007/978-3-031-06220-9_4, doi:10.1007/978-3-031-06220-9_4.