Counting Phylogenetic Networks with Few Reticulation Vertices: Galled and Reticulation-Visible Networks

Yu-Sheng Chang and Michael Fuchs
Department of Mathematical Sciences
National Chengchi University
Taipei 116
Taiwan

April 25, 2024

Abstract

We give exact and asymptotic counting results for the number of galled networks and reticulation-visible networks with few reticulation vertices. Our results are obtained with the component graph method, which was introduced by L. Zhang and his coauthors, and generating function techniques. For galled networks, we in addition use analytic combinatorics. Moreover, in an appendix, we consider maximally reticulated reticulation-visible networks and derive their number, too.

1 Introduction and Results

This is the fourth of a series of papers which is concerned with the enumeration of phylogenetic networks with few reticulation vertices from a fixed class of phylogenetic networks. In the first three papers, we considered *normal* and *tree-child* networks; see [7, 8, 9]. More precisely, in [7, 9], we proposed two methods for deriving asymptotic counting results when the number of reticulation vertices is fixed and the number of leaves tends to infinity. In [8], we corrected mistakes from the approach from [7] and derived exact counting formulas. For instance, one of the results of [8] reads as follows.

Theorem ([8])**.** *For the number $N_{\ell,2}$ of normal networks with ℓ leaves and two reticulation vertices,*

$$N_{\ell,2} = \frac{(3\ell-4)(\ell^2+11\ell+6)}{3}(2\ell-1)!! - 2^\ell(\ell+2)(3\ell-4)\ell!.$$

This result solved an open problem from [2] where exact counting results for tree-child networks were obtained. However, the above result was derived via generating function techniques, whereas the results from [2] were obtained with the *component graph method*. This method as well as generating function techniques will also play an important role in the current paper.

The main purpose of this paper is to prove exact and asymptotic counting results for the classes of galled networks and reticulation-visible networks. We will start with definitions.

Definition 1 (Phylogenetic Networks)**.** A (rooted) *phylogenetic network* of size ℓ is a rooted simple directed acyclic graph (rooted DAG) whose vertices belong to the following three categories:

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

(ii) *Leaves* which have indegree 1 and outdegree 0 and which are bijectively labeled with elements from the set $\{1,\dots,\ell\}$;

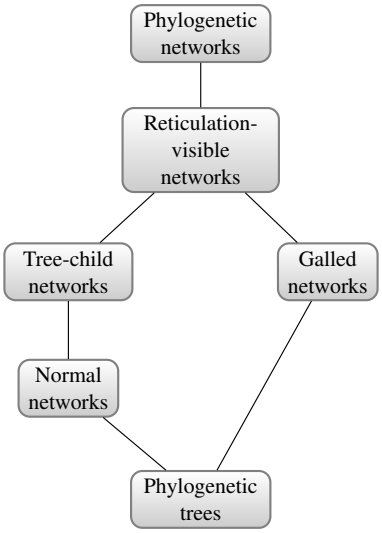


Figure 1: Hasse diagram of the network classes considered in this paper.

(iii) *Internal vertices* which have indegree and outdegree at least 1 but not both equal to 1.

Internal vertices with indegree at least 2 are called *reticulation vertices*; all other internal vertices are called *tree vertices*. Moreover, a phylogenetic network is called *binary* if all internal vertices have total degree equal to 3. If not stated otherwise, networks will subsequently always be binary.

Many subclasses of the class of phylogenetic networks have been proposed; see the recent survey [17] for most of them. The ones of relevance for this paper are defined next.

To be able to state the definitions, we need two notions: first, a *tree cycle* is a pair of edge disjoint paths from a common tree vertex to a common reticulation vertex with all remaining vertices being tree vertices; second, a vertex in a phylogenetic network is called *visible*, if there exists a leaf such that any path from the root to the leaf must contain the vertex.

Definition 2. A phylogenetic network is called:

(i) *Tree-child network* if every non-leaf vertex has at least one child which is not a reticulation vertex;

(ii) *Normal* if it is tree-child and has no *shortcuts*, i.e., the two parents of a reticulation vertex are not in an ancestor-descendant relationship.

(iii) *Galled* if each reticulation vertex is in a (necessarily unique) tree cycle;

(iv) *Reticulation-visible* if each reticulation vertex is visible.

The set-inclusion relationship of these classes is depicted in Figure 1, where the class at the bottom is the class of *phylogenetic trees* which are networks without reticulation vertices.

We are interested in enumeration results for these classes. Throughout the paper, we will use the following notation for the number of networks from a class with ℓ leaves and k reticulation vertices. (The notation for the number of normal networks already appeared in the above theorem.)

network class	number of networks
normal networks	$N_{\ell,k}$
tree-child networks	$\mathrm{TC}_{\ell,k}$
galled networks	$\mathrm{GN}_{\ell,k}$
reticulation-visible networks	$\mathrm{RV}_{\ell,k}$
phylogenetic networks	$\mathrm{PN}_{\ell,k}$

The following asymptotic result for these numbers is known. For fixed k,

$$\mathrm{PN}_{\ell,k} \sim \mathrm{TC}_{\ell,k} \sim \mathrm{N}_{\ell,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^{\ell} \ell^{\ell+2k-1}, \qquad (\ell \to \infty).$$

The first asymptotic equivalence was proved in [18] and the second and third in [7, 8] (without the closed-form expression for the leading constant) and in [9] (with the closed-form expression for the leading constant). An easy consequence of these results is the following.

Corollary 1. *For fixed k,*

$$\mathrm{RV}_{\ell,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^{\ell} \ell^{\ell+2k-1}, \qquad (\ell \to \infty).$$

Thus, for fixed k, the asymptotics of the number of networks for all classes from Definition 2 is known *except* for the class of galled networks. Here, somehow surprisingly, again the same result holds.

Theorem 2. *For fixed k,*

$$\mathrm{GN}_{\ell,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^{\ell} \ell^{\ell+2k-1}, \qquad (\ell \to \infty).$$

We will prove this result below. Our method will also allow us to obtain closed-form expressions for $\mathrm{GN}_{\ell,k}$ for small k. We state the results for $k = 2$ and $k = 3$.

Theorem 3. *We have,*

$$\mathrm{GN}_{\ell,2} = \frac{6\ell^4 + 31\ell^3 + 30\ell^2 - 7\ell - 9}{3}(2\ell-3)!! - 2^{\ell-2}(7\ell+10)(\ell+1)!$$

and

$$\begin{aligned}\mathrm{GN}_{\ell,3} =& \frac{140\ell^6 + 3184\ell^5 + 17195\ell^4 + 34125\ell^3 + 19475\ell^2 - 8599\ell - 6090}{105}(2\ell-3)!! \\ &- 2^{\ell-5}\frac{225\ell^3 + 2045\ell^2 + 5878\ell + 5448}{3}(\ell+1)!.\end{aligned}$$

Remark 1. The formula for $k = 2$ already appeared in [2] (with typos which we have corrected here).

Moreover, we will also give similar formulas for $\mathrm{RV}_{\ell,k}$ for small k.

Theorem 4. *We have,*

$$\mathrm{RV}_{\ell,2} = \frac{6\ell^4 + 7\ell^3 + 6\ell^2 - \ell - 3}{3}(2\ell-3)!! - 2^{\ell-1}(2\ell^2+2\ell+1)\ell!$$

and

$$\begin{aligned}\mathrm{RV}_{\ell,3} =& \frac{4\ell^6 + 20\ell^5 + 33\ell^4 - 32\ell^3 - 76\ell^2 + 12\ell + 12}{3}(2\ell-3)!! \\ &- 2^{\ell-4}\frac{48\ell^4 + 175\ell^3 + 99\ell^2 - 262\ell - 264}{3}\ell!.\end{aligned}$$

Remark 2. Note that for $k = 0$, all the numbers coincide, i.e.,

$$\mathrm{PN}_{\ell,0} = \mathrm{RV}_{\ell,0} = \mathrm{GN}_{\ell,0} = \mathrm{TC}_{\ell,0} = \mathrm{N}_{\ell,0} = (2\ell-3)!!,$$

where the last number is the (well-known) number of phylogenetic trees with ℓ leaves. Also, for $k = 1$, all the numbers, except the number of normal network, coincide:

$$\mathrm{PN}_{\ell,1} = \mathrm{RV}_{\ell,1} = \mathrm{GN}_{\ell,1} = \mathrm{TC}_{\ell,1} = \ell(2\ell-1)!! - 2^{\ell-1}\ell!,$$

where the last result was derived, e.g., in [2]. ($N_{\ell,1}$ is known as well; see, e.g., [20].)

Finally, we will consider maximal reticulated networks. Here, the following sharp upper bounds for the number of reticulation vertices k have been proved.

network class	maximal number of reticulation vertices
normal networks	$\ell - 2$
tree-child networks	$\ell - 1$
galled networks	$2\ell - 2$
reticulation-visible networks	$3\ell - 3$

The result for normal networks first appeared in [19]; the result for tree-child networks was discovered by many authors (and is, in fact, easy to prove). For galled networks, the optimal upper bound was proved, e.g., in [13, 16]; see also [11]. Finally, for reticulation-visible networks, in [12], the upper bound $4\ell - 4$ was established and then this upper bound was reduced to the optimal one independently in [1] and [16]. Here, we will give a simplified proof which in addition also gives the number of maximal reticulated reticulation-visible networks.

Theorem 5. *The maximum number of reticulation vertices in a reticulation-visible networks with ℓ leaves equals $3\ell - 3$. Moreover,*

$$\mathrm{RV}_{\ell,3\ell-3} = \mathrm{TC}_{\ell,\ell-1} = \Theta\left(\ell^{-2/3} e^{a_1(3\ell)^{1/3}} \left(\frac{12}{e^2}\right)^{\ell} \ell^{2\ell}\right),$$

where a_1 is the largest root of the Airy function of the first kind.

Remark 3. The asymptotic result follows from the main result in [10].

We conclude the introduction by an outline of the paper. The main method of proof will be the component graph method which was introduced by Zhang and his coauthors. We will recall this method and some related results in the next section. In Section 3, we will then combine it with generating function techniques and tools from analytic combinatorics to prove Theorem 2 and Theorem 3 for galled networks. Section 4 will contain the proof of Theorem 4 for reticulation-visible networks. Theorem 5 is not really concerned with phylogenetic networks with few reticulation vertices and therefore does not really fall into the scope of this paper. Therefore, we will prove it in the appendix. The paper will be concluded with some final remarks in Section 5.

2 The Component Graph Method

The component graph method was introduced by Zhang and used by him and his coauthors to solve algorithmic problems for phylogenetic networks; see [2, 13, 14, 15]. For instance, it was used to algorithmically solve the counting problem for tree-child networks in [2] and galled networks in [14].

The central object of the method are *component graphs*. First, for a given phylogenetic network, its *tree components* are obtained by removing the two incoming edges of every reticulation vertex. (These edges are called *reticulation edges*.) Using these objects, component graphs are defined as follows.

Definition 3 (Component Graph)**.** The *component graph* of a phylogenetic network N has a vertex for each tree-component. These vertices are then connected via edges according to how the tree-components are connected in N via the reticulation edges. Finally, we attach the leaves (with their labels) in the tree-components to their vertices in the component graph except when the tree-component has only one leaf and corresponds to a terminal vertex in the component graph with incoming edges a double edge, in which case we use the label of the leaf to label the terminal vertex.

See Figures 2 for a phylogenetic network and its component graph. Note that we have replaced double edges by single edges but have indicated that they are double edges by placing arrows on them.

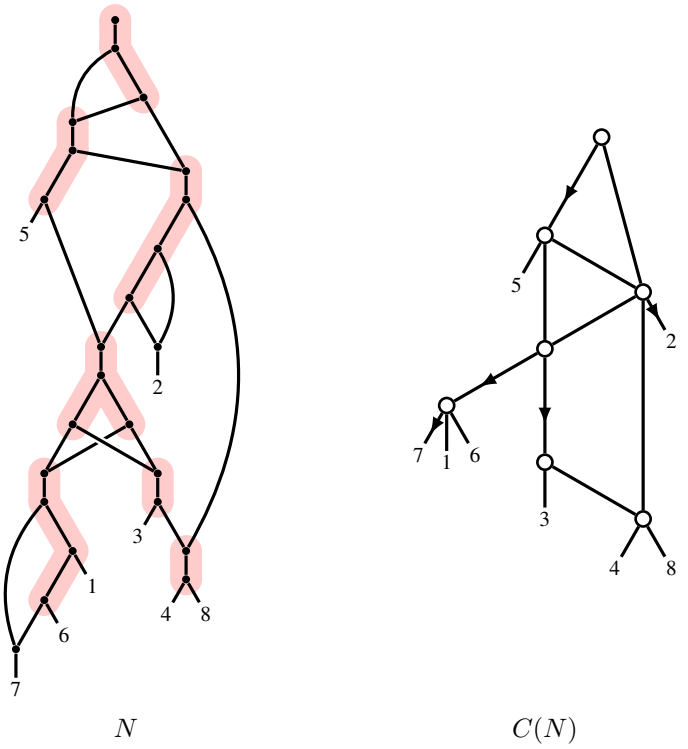


Figure 2: Left: A reticulation-visible network with 8 leaves and 8 reticulation vertices (with the tree-components highlighted). Right: The component graph of the network. Note that it is tree-child network.

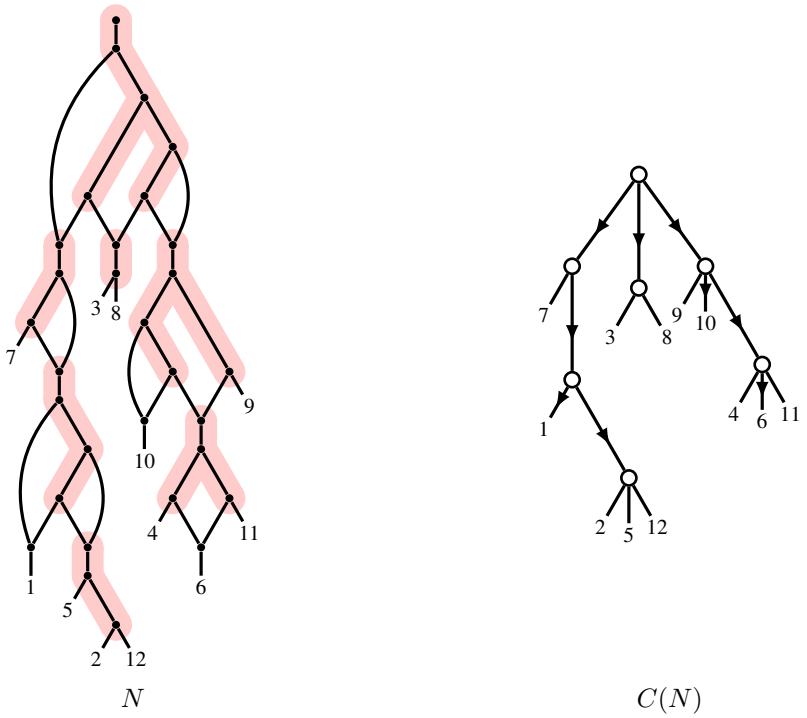


Figure 3: Left: A galled network with 12 leaves and 9 reticulation vertices (with the tree-components highlighted). Right: The component graph of the network. Note that it is a phylogenetic tree.

Remark 4. Note that different definitions of the component graph are used for different network classes. For example, for tree-child networks, the leaves of the tree-components are not attached to the nodes of the component graph but the nodes themselves are labeled; see, e.g., [2, 4, 9]. Our above definition is most suitable for galled networks and reticulation-visible networks which are the network classes considered in this paper.

If N is a network, we denote its component graph by $C(N)$. The component graph can be seen as a compression of the network. We also use the notation $\tilde{C}(N)$ to denote the component graph with the arrows on single edges removed.

Note that the phylogenetic network in Figure 2 is a reticulation-visible network and its component graph is a (non-binary) tree-child network. This, in fact, is a general phenomenon.

Theorem ([15]). *Let N be a phylogenetic network.*

(i) N is galled if and only if $\tilde{C}(N)$ is a (not necessarily binary) phylogenetic tree.

(ii) N is reticulation-visible if and only if $\tilde{C}(N)$ is a (not necessarily binary) tree-child network with all vertices of indegree at most 2 and no reticulation vertex has just one child that is, moreover, a tree vertex.

See Figure 3 for an illustration of part (i) of the above result.

In order to generate networks, we start from the component graphs and decompress them, i.e., we generate all networks whose component graph is the given component graph. For this, the notion of *one-component phylogenetic networks* is useful.

Definition 4. A phylogenetic network is called a *one-component phylogenetic network* if every reticulation vertex is directly followed by a leaf.

In other words, a network is one-component if it has only one non-trivial tree-component.

One-component galled networks were counted in [14]. (Note that, in fact, the class of one-component networks and the class of one-component galled networks coincide; see [2].) Denote by $\mathrm{OGN}_{\ell,k}$ the number of one-component galled networks with ℓ leaves and k reticulation vertices and by $M_{\ell,k}$ those one-component galled networks whose leaves below the reticulation vertices are labeled by $\{1,\dots,k\}$. Then, the following result was proved in [14]. (Note that part (i) is trivial.)

Proposition ([14]). *(i) We have,*

$$\mathrm{OGN}_{\ell,k} = \binom{\ell}{k} M_{\ell,k}.$$

(ii) For $2 \le k \le \ell$,

$$M_{\ell,k} = (\ell+k-2)M_{\ell,k-1} + (k-1)M_{\ell,k-2} + \frac{1}{2}\sum_{1\le d\le k-1}\binom{k-1}{d}(2d-1)!!\,(M_{\ell-d,k-1-d} - M_{\ell+1-d,k-1-d}) \quad (1)$$

with initial values $M_{\ell,0} = (2\ell-3)!!$ and $M_{\ell,1} = (\ell-1)(2\ell-3)!!$.

Now, consider the component graph of a galled network; see Figure 4 for an example. In order to decompress it, first pick a one-component galled network N^o for the root r of the component graph, where the number of leaves, say ℓ_r (the network over the first arrow in the example), of N^o is the outdegree of r and the number of reticulation vertices, say k_r, is the number of arrows on the outgoing edges of r. ($\ell_r = k_r = 3$ in the example.) Moreover, let the leaves below the reticulation vertices in

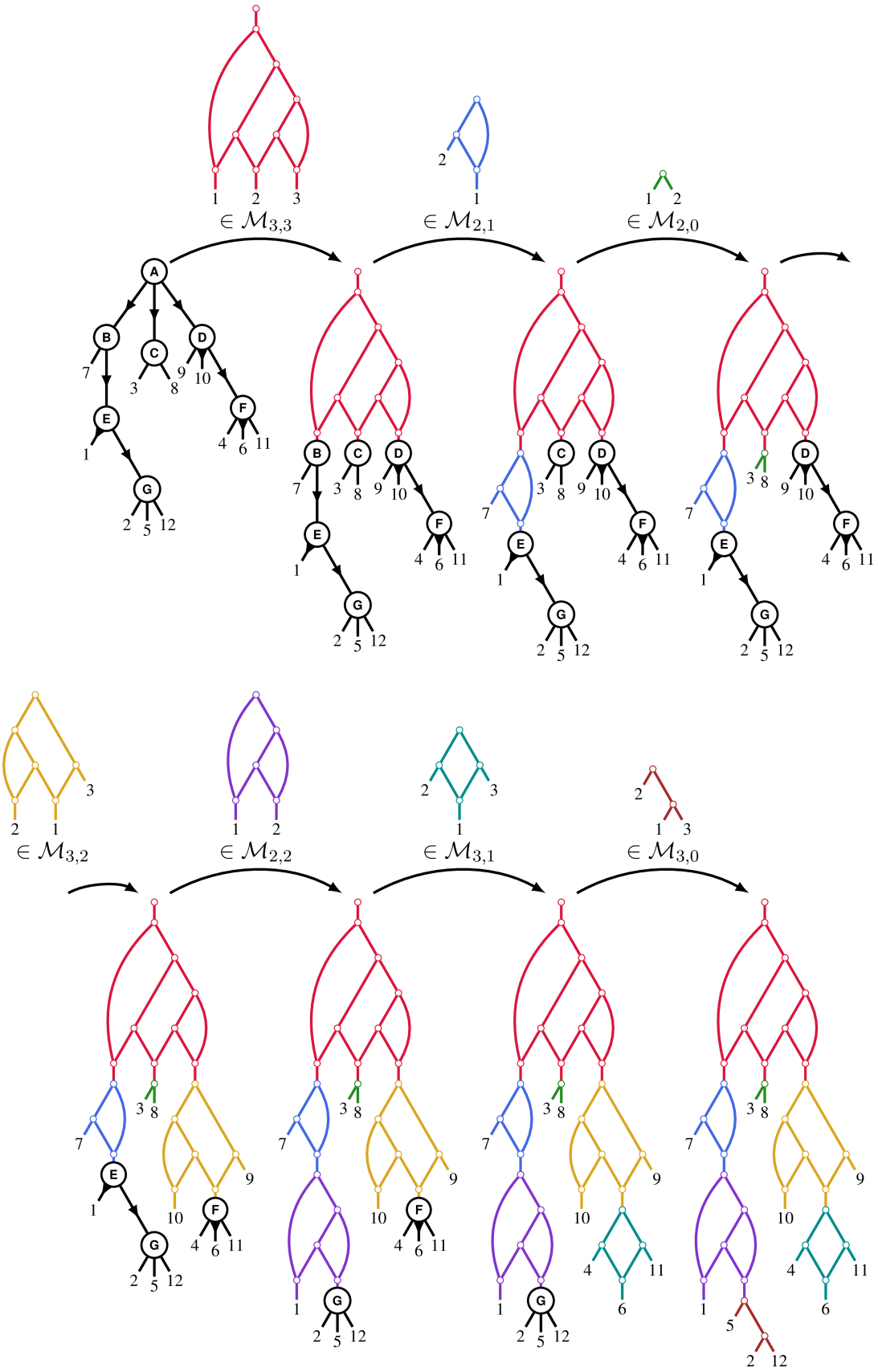


Figure 4: Step-by-step decompression of the component graph from Figure 3; the nodes are processed in the order indicated and the one-component networks replacing nodes are above the arrows. Note that except for the first of these networks, the root edge has to be removed in all others.

N^o have labels $\{1, \ldots, k_r\}$. Next, remove r from the component graph which gives a forest of ℓ_r trees. (The subtrees rooted at B, C, and D in the example.) The trees which have been attached by edges with arrows to r (all trees in the example) go to the leaves below reticulation vertices in N^o and the remaining trees (which consist of just one labeled vertex) are used the relabel the remaining leaves of N^o in an order consistent way (i.e., the smallest goes to $k_r + 1$, the second smallest to $k_r + 2$, etc.). Moreover, for the trees which go to the leaves below reticulation vertices, the order in which they are attached is the increasing order of their smallest leaf label (i.e., the one with the smallest leaf label — the subtree rooted at B in the example as the smallest label of this subtree is 1 — goes to 1, the one with the second smallest leaf label — the subtree rooted at C in the example as the smallest label of this subtree is 3 — goes to 2, etc.). Then, continue with the non-leaf vertices of the trees in a recursive way. This gives, by picking all possible one-component networks in each step, all possible galled networks whose component graph is the given component graph. Moreover, if the given component graph has no arrows (which for galled networks means it is just a phylogenetic tree), then arrows have to be placed on edges leading to non-leaf vertices, whereas on the pendant edges, we can freely decide whether we want to place an arrow or not. Overall, this procedure gives the following result from [14] for the number of galled networks with ℓ leaves, denoted by GN_ℓ.

Theorem ([14])**.** *For the number* GN_ℓ *of galled networks with* ℓ *leaves,*

$$\mathrm{GN}_\ell = \sum_{\mathcal{T}} \prod_{v} \sum_{j=0}^{c_{\mathrm{lf}}(v)} \binom{c_{\mathrm{lf}}(v)}{j} M_{c(v),c_{\mathrm{nlf}}(v)+j}, \tag{2}$$

where the first sum runs over all (not necessarily binary) phylogenetic trees $\mathcal{T}$*, the product runs over all internal vertices* v *of* $\mathcal{T}$*,* $c(v)$ *is the outdegree of* v*, and* $c_{\mathrm{lf}}(v)$ *resp.* $c_{\mathrm{nlf}}(v)$ *are the number of leaves resp. non-leaves amongst the children of* v*.*

By keeping track of the number of arrows, $\mathrm{GN}_{\ell,k}$ can be computed as well.

Next, a similar procedure can be used for reticulation-visible networks, too. First, observe that the classes of one-component reticulation-visible networks and one-component galled networks coincide (since both are, in fact, the class of one-component networks without any restrictions). Thus, we can again work with $M_{\ell,k}$ which satisfies (1).

Now, consider the component graph of a reticulation-visible network (which by the above result is a particular tree-child network; also, assume that there are arrows on the edges); see Figure 5 for an example. The decompressing works as for a galled networks, with the only difference being how the children of a node are attached to the leaves of the one-component network which replaces the node. More precisely, in which order should this be done? (Again, the children attached with edges having arrows go to the leaves below reticulation vertices, the others not.) Here, the tree-child property comes into play. It implies that that for every vertex, there exists a set of leaves which can only be reached from this vertex. All children of a node have such a set and we can order these sets (and consequently the children) according to their smallest elements. This is then used to attach the children of a node to the leaves of the one-component network which replaces the node. Overall, this gives the following result for the number of reticulation-visible networks with ℓ leaves, denoted by RV_ℓ, where we again start from component graphs without arrows.

Theorem 6. *For the number* RV_ℓ *of reticulation-visible networks with* ℓ *leaves,*

$$\mathrm{RV}_\ell = \sum_{\mathcal{TC}} \prod_{v} \sum_{j=0}^{c_{\mathrm{lf}}(v)} \binom{c_{\mathrm{lf}}(v)}{j} M_{c(v),c_1(v)+j}, \tag{3}$$

where the first sum runs over all (not necessarily binary) tree-child networks with all vertices of indegree at most 2 *and no reticulation vertex has just one child that is, moreover, a tree vertex, the product runs*

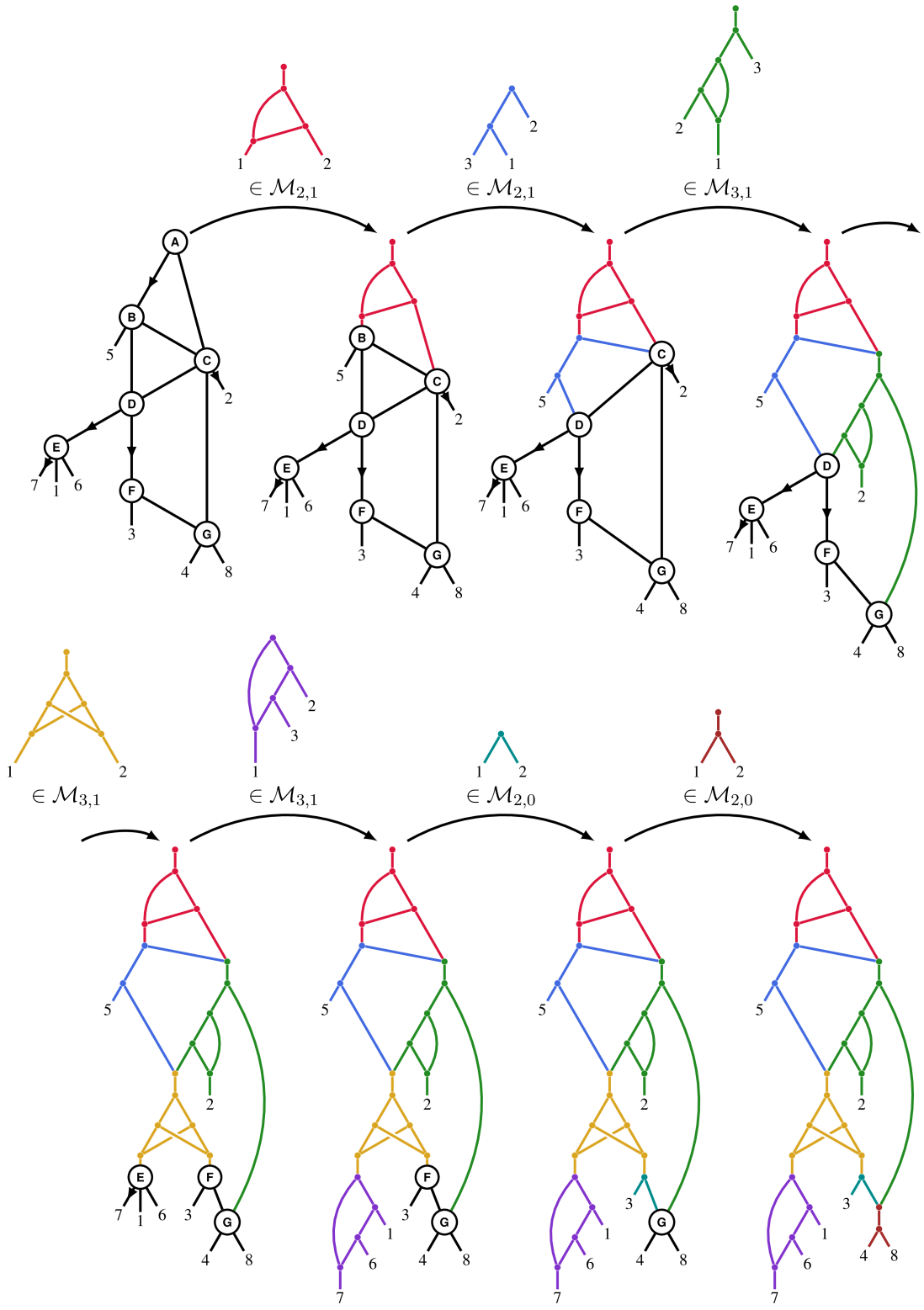


Figure 5: Step-by-step decompression of the component graph from Figure 2; again the nodes are processed in the indicated order and the one-component networks are above the arrows. Here, the root edges of the latter networks have to be removed if and only if the replaced nodes in the component graph have indegree 1 (and thus the incoming edge has an arrow on it).

over all internal vertices v of the tree-child network, $c(v)$ is the outdegree of v, $c_{\mathrm{lf}}(v)$ is the number of leaves below v, and $c_1(v)$ is the number of children of v which are not leaves and have indegree 1.

Again, by keeping track of arrows, we can compute $\mathrm{RV}_{\ell,k}$. We will do this in Section 4 to derive formulas for small values of k and any ℓ.

3 Galled Networks

In this section, we use the component graph method to count (both exactly and asymptotically) galled networks with a small number of reticulation vertices. In addition, we use generating function techniques and the method of singularity analysis which is described in Chapter VI of [6]. (We will use below some of the notation from that chapter.)

Since the component graph method is based on one-component galled networks, we start by analyzing their number. First, from the initial values and (1), $M_{\ell,k}$ for small values of k equals

$$\begin{aligned} M_{\ell,0} &= (2\ell-3)!!, \\ M_{\ell,1} &= (\ell-1)(2\ell-3)!!, \\ M_{\ell,2} &= (2\ell-1)(\ell-1)^2(2\ell-5)!!. \end{aligned}$$

From this, we observe the following pattern.

Lemma 7. *For fixed $k \geq 1$,*

$$M_{\ell,k} = p_k(\ell)(2(\ell-k)-3)!!, \qquad (\ell \geq k),$$

where $p_k(\ell)$ is a polynomial of degree $2k$ with leading coefficient 2^k.

Proof. This follows by induction on k. First, the claim holds for $k=1, k=2$, and also for $k=0$ when $\ell \geq 1$. (This is why we used $(2(\ell-k)-3)!!$ instead of $(2(\ell-k+1)-3)!!$.)

Assume now that it holds for $2 \leq k' < k$. Then, we observe that there are four terms on the right-hand side of (1) where we have to plug in the induction hypothesis. After doing this and bringing the double factorial into the form $(2(\ell-k)-3)!!$, we see that the first term ($M_{\ell,k-1}$) has as multiplicative factor a polynomial in ℓ of degree $2k-1$ which gets multiplied by $\ell+k-2$ and thus becomes a polynomial of degree $2k$, the second term ($M_{\ell,k-2}$) produces a polynomial of degree $2k-2$, the third factor ($M_{\ell-d,k-1-d}$) yields a polynomial of degree $2k-1-2d$, and the polynomial of the final factor ($M_{\ell+1-d,k-1-d}$) is of degree $2k-2d$. Thus, by collecting all these polynomials, we see that $M_{\ell,k}$ has the desired form where the degree and leading term comes from that of the first term. The latter is the leading term of the polynomial of $M_{\ell,k-1}$ multiplied by $2\ell^2$ which proves the claim. ∎

Note that from the last result, we have

$$M_{\ell+k,k} = q_k(\ell)(2\ell-3)!!, \qquad (\ell \geq 0), \tag{4}$$

where $q_k(\ell)$ is again a polynomial of degree $2k$ with leading term 2^k. We next consider the exponential generating function of this quantity:

$$F_k(z) = \sum_{\ell \geq 0} M_{\ell+k,k} \frac{z^\ell}{\ell!},$$

where $M_{0,0} = 0$. Then, from the expressions for $M_{\ell,k}$ for $k = 0, 1, 2$, we obtain that:

$$F_0(z) = 1 - \sqrt{1-2z}, \qquad F_1(z) = \frac{z}{(1-2z)^{3/2}} \tag{5}$$

and

$$F_2(z) = \frac{3 - z + 7z^2 - 4z^3}{(1-2z)^{7/2}}. \tag{6}$$

Also, we have the following asymptotic result for $F_k(z)$ for all $k \geq 1$, where we use the notion of Δ-analyticity of a function $f(z)$ at z_0, i.e., $f(z)$ is an analytic function in a domain of the form

$$\Delta := \{z \,:\, |z| < r,\ |\arg(z - z_0)| > \phi\},$$

for some $r > |z_0|$ and $0 < \phi < \pi/2$.

Lemma 8. *For $k \geq 1$, $F_k(z)$ is Δ-analytic for some Δ-domain at $1/2$ and satisfies, as $z \to 1/2$, in the Δ-domain:*

$$F_k(z) \sim \frac{(4k-3)!!}{2^k(1-2z)^{2k-1/2}}. \tag{7}$$

Remark 5. The asymptotic (7) is called *singularity expansion* of $F_k(z)$ at $z \to 1/2$. Note that $k = 0$ is not included since from (5), $F_0(z) \sim 1$, as $z \to 1/2$, which does not follow the pattern from (7).

Proof. First, set

$$P(z) := \sum_{\ell \geq 0} (2\ell - 3)!! \frac{z^\ell}{\ell!} = 2 - \sqrt{1-2z}. \tag{8}$$

From (4), we see that the function $F_k(z)$ is built from $P(z)$ as a linear combination of the power series

$$D^j P(z),$$

where $0 \leq j \leq 2k$ and D^j is the j-th iteration of $D := z\frac{d}{dz}$. Since $P(z)$ is clearly Δ-analytic, and Δ-analyticity is closed under differentiation (see Theorem VI.8 in [6]), multiplication by z, and linear combination (see Section VI.6 in [6]), we obtain that $F_k(z)$ is Δ-analytic. In addition, the singularity expansion of the derivative is obtained by differentiating the singularity expansion of a function (again see Theorem VI.8 in [6]) and obvious rules hold for multiplying with z (which introduces a factor of $1/2$, as $z \to 1/2$) and taking linear combinations (see again Section VI.6 in [6]). Thus, the dominant term in the singularity expansion expansion of $F_k(z)$ arises from the highest derivative, and so we have, as $z \to 1/2$,

$$F_k(z) \sim 2^k D^{2k} P(z) \sim \frac{(4k-3)!!}{2^k(1-2z)^{2k-1/2}},$$

where the second asymptotic equivalence follows by differentiating (8) $2k$ times, multiplying with $1/2$ after every differentiation, and multiplying the final result by 2^k. ∎

We now consider general galled networks which are built from one-component galled networks and component graphs. This was described in detail in Section 2 and the recursive method there can be translated into generating functions (because the component graphs are trees). Since we are interested in reticulation vertices, we need to keep track of them. We therefore consider the generating function

$$G(z, v) := \sum_{\ell \geq 0} \sum_{k \geq 0} \mathrm{GN}_{\ell,k} \frac{z^\ell}{\ell!} v^k.$$

Then, we have the following result.

Proposition 9. *We have,*

$$G(z, v) = \sum_{j \geq 0} F_j(z) \frac{(vG(z, v))^j}{j!}. \tag{9}$$

Proof. We use symbolic combinatorics as described in Chapter II and Chapter III of [6]. Note that the decompression procedure of component graphs from Section 2 entails that every galled network is built from a one-component galled network (which was used to replace the root in the decompression procedure) whose leaves below reticulation vertices are replaced by an unordered sequence of galled networks. If the former has j reticulation vertices, it is counted by $F_j(z)$ and the latter is then counted by $(vG(z,v))^j/j!$ where v counts reticulation vertices, $G(z,v)^j$ counts ordered sequences of galled networks and the factor $1/j!$ is needed to take away the order. Next, the product of these two generating functions counts the galled networks which are re-labelled as described in Section 2 since the product of exponential generating functions corresponds to the product of labeled combinatorial classes; see Chapter II in [6]. Finally, summing over j gives the claimed result. ∎

The exponential generating function for the number of galled networks with k reticulation vertices, i.e.,

$$E_k(z) := \sum_{\ell \geq k} \mathrm{GN}_{\ell,k} \frac{z^\ell}{\ell!}$$

is obtained from $G(z,v)$ by partial differentiation and evaluating at $v=0$:

$$E_k(z) = \frac{1}{k!} \frac{\partial^k}{\partial v^k} G(z,v)\Big|_{v=0}.$$

From Proposition 9, we obtain a recurrence.

Lemma 10. *For $k \geq 1$,*

$$E_k(z) = \sum_{j=1}^{k} \frac{F_j(z)}{j!} \sum_{\ell_1+\cdots+\ell_j=k-j} E_{\ell_1}(z)\cdots E_{\ell_j}(z). \tag{10}$$

Proof. Differentiating (9) k-times and evaluating at $v=0$ gives

$$\begin{aligned} E_k(z) &= \frac{1}{k!} \sum_{j=1}^{k} \frac{F_j(z)}{j!} \frac{d^k}{dv^k} \left(vG(z,v)\right)^j \Big|_{v=0} \\ &= \frac{1}{k!} \sum_{j=1}^{k} \frac{F_j(z)}{j!} \binom{k}{j} j! \frac{d^{k-j}}{dz^{k-j}} G(z,v)^j \\ &= \sum_{j=1}^{k} \frac{F_j(z)}{j!(k-j)!} \sum_{\ell_1+\cdots+\ell_j=k-j} \binom{k-j}{\ell_1,\cdots,\ell_j} \ell_1! E_{\ell_1}(z) \cdots \ell_j! E_{\ell_j}(z). \end{aligned}$$

The claim is obtained by writing the multinomial coefficients as factorials and canceling terms. ∎

We have now everything ready to prove Theorem 3.

Proof of Theorem 3. Note that $E_0(z) = 1 - \sqrt{1-2z}$. Then, from (10) and (5):

$$E_1(z) = F_1(z)E_0(z) = \frac{z(1-\sqrt{1-2z})}{(1-2z)^{3/2}}.$$

Moreover, by using once again (10) and (5) as well as (6):

$$E_2(z) = F_1(z)E_1(z) + \frac{F_2(z)E_0(z)^2}{2}$$

$$= \frac{12z^4 - 18z^3 + 17z^2 - 36z + 21 + (12z^3 - 10z^2 + 15z - 21)\sqrt{1-2z}}{3(1-2z)^{7/2}}.$$

Extracting coefficients gives the claimed result for $\mathrm{GN}_{\ell,2}$.

As for $\mathrm{GN}_{\ell,3}$, the same method can be used, only the resulting computation is more tedious (and therefore best done with mathematical software, e.g., Maple). ∎

What is left is to prove the asymptotic counting result for $\mathrm{GN}_{\ell,k}$ from Theorem 2. This will follow from the following result for the singularity expansion of $E_k(z)$ which is obtained from (10) and induction.

Proposition 11. *For $k \geq 1$, $E_k(z)$ is Δ-analytic for some Δ-domain at $1/2$ and satisfies, as $z \to 1/2$, in the Δ-domain:*

$$E_k(z) \sim \frac{(4k-3)!!}{k!2^k(1-2z)^{2k-1/2}}.$$

Proof. Note that

$$E_0(z) = F_0(z) = 1 - \sqrt{1-2z} \sim 1.$$

Thus, if $k = 0$ is included in the claim, then the power of $1-2z$ in the denominator is $\max\{2k-1/2, 0\}$.

We use induction on k. For $k = 1$, the claim holds since $E_1(z) \sim F_1(z)$ which has the desired form by (7). Thus, we can assume that the claim holds for $k' < k$. We need to prove it for k. Plugging the induction hypothesis into (10) and using (7), we obtain that for the terms inside the double sum of (10):

$$F_j(z)E_{\ell_1}(z)\cdots E_{\ell_j}(z) \sim c(1-2z)^{-(2j-1/2+\max\{2\ell_1-1/2,0\}+\cdots+\max\{2\ell_j-1/2,0\})},$$

where c is a suitable constant and $\ell_1 + \cdots + \ell_j = k - j$. The term inside the bracket is maximized if and only if $j = k$ and thus $\ell_1 = \cdots = \ell_k = 0$. This shows that

$$E_k(z) \sim \frac{F_k(z)}{k!}$$

which by (7) gives the claim. ∎

Now, we can prove Theorem 2 for $k \geq 1$.

Proof of Theorem 2. From Proposition 11 and Corollary VI.1 in [6]:

$$\begin{aligned}\mathrm{GN}_{\ell,k} = \ell![z^\ell]E_k(z) &\sim \ell!\frac{(4k-3)!!}{k!2^k\Gamma(2k-1/2)}[z^n](1-2z)^{-2k+1/2} \\ &\sim \ell!\frac{(4k-3)!!}{k!2^k\Gamma(2k-1/2)}2^\ell\ell^{2k-3/2}.\end{aligned}$$

Note that

$$\Gamma(2k-1/2) = 2^{-2k+1}(4k-3)!!\sqrt{\pi}.$$

Plugging this into the above expression and using Stirling's formula gives the claimed result. ∎

Remark 6. It is easily verified that Theorem 2 holds for $k = 0$, too.

4 Reticulation-Visible Networks

In this section, we prove the formulas for the number of reticulation-visible networks with ℓ leaves and k reticulation vertices with $k = 2$ and $k = 3$ to establish Theorem 4.

Let N be a reticulation-visible network. We remove the leaves and their pendant edges from $C(N)$ except those whose pendant edge has an arrow on it. Then, the resulting component graph is a (unlabeled) rooted simple DAG with every non-root vertex of indegree 2. (Note that an edge with an arrow on it is actually a double edge and thus counts as two edges.) Moreover, this DAG has exactly $k+1$ vertices. For $k=2$ and $k=3$, all these DAGs are listed in Figure 8 in [2]; see also Figure 6 and Figure 7 below.

Using these DAGs and the decompression procedure explained in Section 2, we can derive the exponential generating function of $\mathrm{RV}_{\ell,k}$.

Proposition 12. *Let $\mathcal{D}_m$ be the set of (unlabelled) rooted DAGs with m vertices in which non-root vertices have indegree 2 and double edges between two vertices are allowed. Then, for given k,*

$$\sum_{\ell\geq 1} \mathrm{RV}_{\ell,k}\,\frac{z^\ell}{\ell!} = \sum_{G\in\mathcal{D}_{k+1}} \frac{1}{\mathrm{m}(G)} \prod_{v\in G} \sum_{\ell\geq \ell_0} M_{\ell+c(v),c_1(v)}\,\frac{z^\ell}{\ell!}, \tag{11}$$

where $\mathrm{m}(G)$ counts symmetries in G, the product runs over all vertices in G, and the final sum on the right-hand side is the generating function with respect to the number of labeled leaves which are attached to v, where $c(v)$ is the outdegree of v in G, $c_1(v)$ is the number of children of v with arrow on their edges, and ℓ_0 is 0 or 1 according to whether $c_1(v)>0$ or $c_1(v)=0$, respectively.

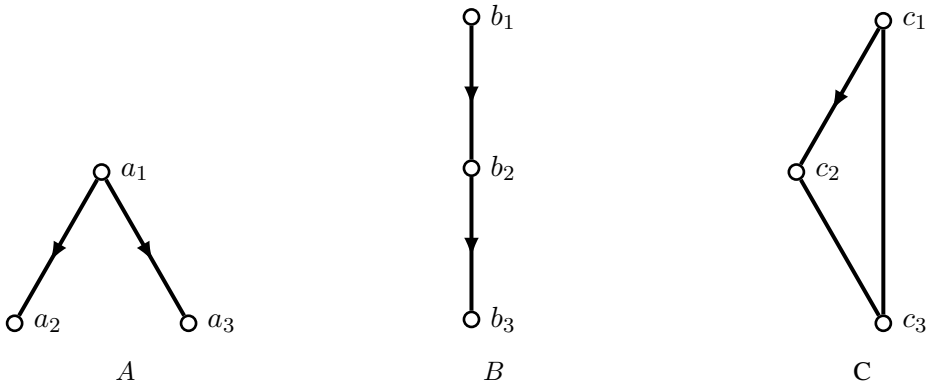


Figure 6: The 3 DAGs from the set $\mathcal{D}_3$ where for convenience, we have labeled the vertices. Call the reticulation vertices of the decompressed reticulation-visible networks r_1 and r_2. Then, A gives all the networks where r_1 and r_2 are not in an ancestor-descendant relationship; B gives all networks where r_1 is above r_2 but they are not in a tree-cycle; and C gives all networks where r_1 is in the tree cycle of r_2.

Now, to find $\mathrm{RV}_{\ell,2}$, we start from the DAGs in the set $\mathcal{D}_3$ which are listed in Figure 6. Note that $m(A)=2$ (due to the symmetry about the root), $m(B)=m(C)=1$, and for each vertex v, the exponential generating function $f_v(z)$ from Proposition 12 equals:

$$f_{a_1}(z) = F_2(z) = \sum_{\ell\geq 0} M_{\ell+2,2}\,\frac{z^n}{n!} = \frac{15}{4}\,X^{-7} - \frac{3}{2}\,X^{-5} + \frac{1}{4}\,X^{-3} + \frac{1}{2}\,X^{-1};$$

$$f_{a_2}(z) = f_{a_3}(z) = f_{b_3}(z) = f_{c_3}(z) = F_0(z) = \sum_{\ell\geq 1} M_{\ell,0}\,\frac{z^\ell}{\ell!} = 1 - X;$$

$$f_{b_1}(z) = f_{b_2}(z) = F_1(z) = \sum_{\ell\geq 0} M_{\ell+1,1}\,\frac{z^\ell}{\ell!} = \frac{1}{2}\,X^{-3} - \frac{1}{2}\,X^{-1};$$

$$f_{c_1}(z) = \sum_{\ell\geq 0} M_{\ell+2,1}\,\frac{z^\ell}{\ell!} = \frac{3}{2}\,X^{-5} - \frac{1}{2}\,X^{-3};$$

$$f_{c_2}(z) = \sum_{\ell \geq 1} M_{\ell+1,0} \frac{z^\ell}{\ell!} = X^{-1} - 1,$$

where we have used the abbreviation $X := \sqrt{1-2z}$. Thus, from Proposition 12, we have for the exponential generating function of $\mathrm{RV}_{\ell,2}$,

$$\frac{(3 - z + 7z^2 - 4z^3)(1 - z - \sqrt{1-2z})}{(1-2z)^{7/2}} = \frac{(1-X)^2(15 - 6X^2 + X^4 + 2X^6)}{8X^7}.$$

What is left is to extract coefficients which can be done with the following lemma.

Lemma 13. *We have, for n large enough,*

$$[z^n]X^d = \begin{cases} 0, & \text{if } d \geq 0 \text{ and } d \text{ is even, } k := \frac{d}{2}; \\ (-1)^{k+1}(2k+1)!!\dfrac{(2n-2k-3)!!}{n!}, & \text{if } d \geq 0 \text{ and } d \text{ is odd, } k := \frac{d-1}{2}; \\ 2^n \dbinom{n+k-1}{k-1}, & \text{if } d < 0 \text{ and } d \text{ is even, } k := -\frac{d}{2}; \\ \dfrac{1}{(2k-3)!!}\dfrac{(2n+2k-3)!!}{n!}, & \text{if } d < 0 \text{ and } d \text{ is odd, } k := -\frac{d-1}{2}. \end{cases}$$

Proof. All cases are obtained by the binomial theorem and standard computations. ∎

Applying the lemma gives the following result.

$$\mathrm{RV}_{\ell,2} = \frac{6\ell^4 + 7\ell^3 + 6\ell^2 - \ell - 3}{3}(2\ell-3)!! - 2^{\ell-1}(2\ell^2 + 2\ell + 1)\ell!.$$

We next consider $k = 3$. Here the set $\mathcal{D}_4$ has 13 elements; see Figure 7. Note that these DAGs fall into two types: the A_j's are tree structures and the B_j's are not. (The latter generate the reticulation-visible networks which are not galled networks.) Next, we note that $m(A_1) = 6, m(A_2) = m(B_1) = m(B_4) = 2$ and the values are 1 in all other cases. Thus, by Proposition 12, we obtain for the exponential generating function $f_A(z)$ arising by the A_j's,

$$f_A(z) = \frac{4z^3(29 + 12z + 29z^2 - 37z^3 + 36z^4 - 14z^5)}{(1-2z)^{11/2}\,(1+\sqrt{1-2z})^3} + \frac{6z^3(3 - z + 7z^2 - 4z^3)}{(1-2z)^5\,(1+\sqrt{1-2z})^2} + \frac{2z^4}{(1-2z)^{9/2}\,(1+\sqrt{1-2z})},$$

and for the exponential generating function $f_B(z)$ arising from the B_j's,

$$f_B(z) = \frac{(1-X)^2(258 - 105X - 153X^2 - 16X^3 + 26X^4 + 7X^5 + 3X^6 - 2X^7 - 2X^8)}{8\,X^{10}}.$$

Then, by extracting coefficients of $f_A(z) + f_B(z)$ with Lemma 13,

$$\mathrm{RV}_{\ell,3} = \frac{4\ell^6 + 20\ell^5 + 33\ell^4 - 32\ell^3 - 76\ell^2 + 12\ell + 12}{3}(2\ell-3)!! - 2^{\ell-4}\frac{48\ell^4 + 175\ell^3 + 99\ell^2 - 262\ell - 264}{3}\ell!.$$

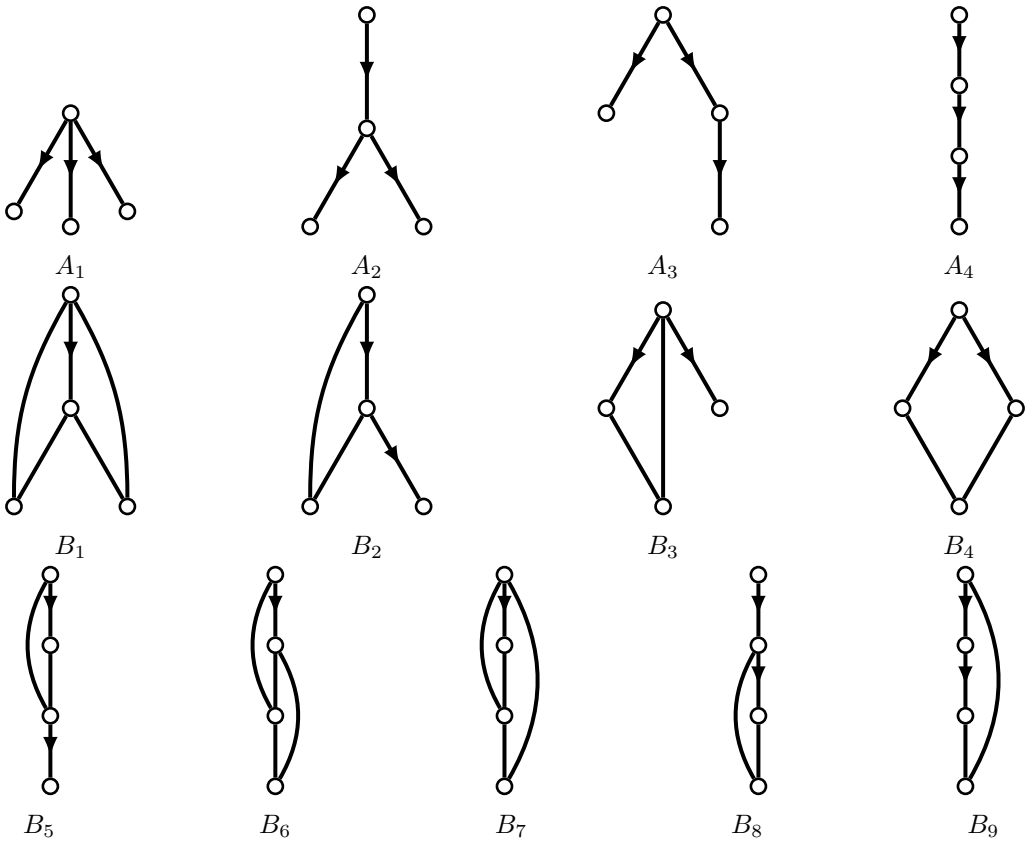


Figure 7: The 13 DAGs from the set $\mathcal{D}_4$.

5 Conclusion

In this paper, we derived exact and asymptotic counting results for the number of galled networks and reticulation-visible networks with a fixed number of reticulation vertices. For galled networks, we proposed a generating function approach which is based on the component graph method by Zhang and his coauthors. The exact results followed from it by coefficient extraction. Moreover, the asymptotic result was derived from it by methods from analytic combinatorics. For reticulation-visible networks, we also used the component graph method and generating functions to derive their exact counts. Moreover, the asymptotic result followed from previous work.

Thus, an asymptotic counting result for fixed k is now known for all network classes from Figure 1; in fact, the asymptotic main term is the same for all these classes. This is slightly surprising for galled networks, since they are on a different "branch" in the diagram in Figure 1. This suggests to consider phylogenetic networks which are both galled and satisfy the tree-child condition. This class of *galled tree-child networks* is also interesting from a theoretical point of view since very different methods have been used to determine the (asymptotic) numbers of tree-child and galled networks; see [10, 11]. We will do this in the forthcoming work [5].

All the results in this paper were proved for fixed k. How about the number of networks if one sums over all possible values of k? The first order asymptotics of this number for galled networks is known and was obtained in [11]. In fact, the latter also uses the component graph method. More precisely, the asymptotics was deduced from (2). The analysis proceeded in two steps: first, the number of one-component galled networks was asymptotically studied by using the recurrence (1). It turned out that $M_{\ell,k}$ has its maximum at $k = \ell$ (which is also the largest possible value of k) and satisfies a Poisson law. Then, in the second step, (2) was used to find the asymptotics of GN_ℓ. This was possible because

the first sum is over the set of (not necessarily binary) phylogenetic trees and the authors in [11] found the trees whose contributions dominate the asymptotics.

In view of (3) does this approach also work for reticulation-visible networks? First, we can skip the first step because, as mentioned in Section 2, the class of one-component galled networks and one-component reticulation-visible networks coincide. Thus, in order to use the approach, we need to understand the class of (not necessarily binary) tree-child networks, which appears in the first sum of (3). In particular, we need to know which of these networks contribute to the main term of the asymptotics.

For this, it would be good to have a better understanding of structural properties of this class. Binary tree-child networks have been counted in [10]. However, generalizing these results to the non-binary case is non-trivial. For example, in [3, 4], we considered d-combining tree-child networks which are tree-child networks with all internal vertices either bifurcating tree vertices or reticulation vertices with one child and exactly d parents. Extending these results to general tree-child networks seems to be a major challenge. Such a generalization (and the structural knowledge it would entail) could, however, be helpful, if one wants to use (3) to find the asymptotics of RV_ℓ.

Acknowledgment

We thank the two reviewers and the associate editor for helpful suggestions. Both authors acknowledge partially support by the National Science and Technology Council (NSTC), Taiwan under the grant NSTC-111-2115-M-004-002-MY2.

References

[1] M. Bordewich and C. Semple (2016). Reticulation-visible networks, *Adv. in Appl. Math.*, **78**, 114–141.

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

[3] 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.

[4] Y.-S. Chang, M. Fuchs, H. Liu, M. Wallner, G.-R. Yu (2022). Enumerative and distributional results for d-combining tree-child networks, Adv. in Appl. Math., to appear.

[5] Y.-S. Chang, M. Fuchs, G.-R. Yu, Galled tree-child networks, arXiv:2403.02923.

[6] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*, 1st edition, Cambridge University Press, Cambridge, 2009.

[7] 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.

[8] 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.

[9] 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.

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

[11] M. Fuchs, G.-R. Yu, L. Zhang (2022). Asymptotic enumeration and distributional properties of galled networks, *J. Comb. Theory Ser. A.*, **189**, 105599, 28 pages.

[12] P. Gambette, A. D. M. Gunawan, A. Labarre, S. Vialette, L. Zhang (2015). Locating a tree in a phylogenetic network in quadratic time, In *Proc. of the 19th Int'l Conf. Res. in Comput. Mol. Biol. (RECOMB)*, 96–107, Warsaw, Poland.

[13] A. D. M. Gunawan, B. DasGupta, L. Zhang (2017). A decomposition theorem and two algorithms for reticulation-visible networks, *Inf. Comput.*, **252**, 161–175.

[14] A. D. M. Gunawan, J. Rathin, L. Zhang (2020). Counting and enumerating galled networks, *Discrete Appl. Math.*, **283**, 644–654.

[15] A. D. M. Gunawan, H. Yan, L. Zhang (2019). Compression of phylogenetic networks and algorithm for the tree containment problem, *J. Comput. Biol.*, **26:3**, 285–294.

[16] A. D. M. Gunawan and L. Zhang (2015) Bounding the size of a network defined by visibility property, arXiv:1510.00115.

[17] S. Kong, J. C. Pons, L. Kubatko, K. Wicke (2022). Classes of explicit phylogenetic networks and their biological and mathematical significance, *J. Math. Biol.*, **84**, Paper: 47.

[18] M. Mansouri (2022). Counting general phylogenetic networks, *Australas. J. Combin.*, **83**, 40–86.

[19] S. J. Willson (2010). Properties of normal phylogenetic networks, *Bull. Math. Biol.*, **72:2**, 340–358.

[20] L. Zhang (2019). Generating normal networks via leaf insertion and nearest neighbor interchange, *BMC Bioinformatics*, **20**, Article 642.

Appendix: Proof of Theorem 5

We use the component graph method from Section 2. Since the component graph of a reticulation-visible network is a tree-child network with all vertices of indegree at most 2 and no reticulation vertex has just one child that is, moreover, a tree vertex, we first fix such a tree-child network $\tilde{C}$ (without arrows on the edges).

The maximal number of reticulation vertices a network decompressed from $\tilde{C}$ can have is obtained by placing arrows on all pendant edges except the ones directly below reticulation vertices. Let $r(\tilde{C})$ be the number of these edges plus the number of reticulation vertices of $\tilde{C}$ plus the number of internal vertices with exactly one incoming edge. Then, our goal is to find those $\tilde{C}$ which maximize $r(\tilde{C})$. Note that $r(\tilde{C})$ remains invariant if we replace vertices with indegree 2 and outdegree at least 2 by a reticulation vertex followed by a tree vertex and do not count the additional created edge. This is the set of $\tilde{C}$, we will consider in the sequel. (Thus, for the decompressing procedure, we first have to merge reticulation vertices followed by just one tree vertices, if there are any such vertices.)

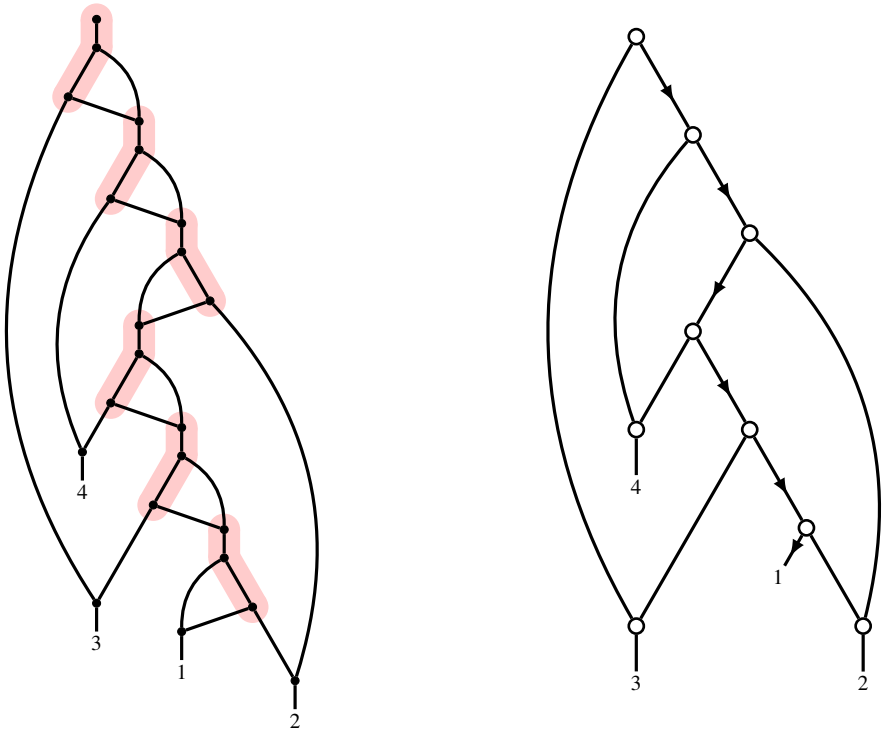
We start with the following claim.

Claim 1: $r(\tilde{C})$ is maximized only for binary tree-child networks $\tilde{C}$.

Assume that $\tilde{C}$ has at least one vertex, say v, with outdegree ≥ 3. Replacing v with a vertex of outdegree 2 and attaching to it one of the children of v and a vertex whose children are the remaining children of v clearly gives a new tree-child network with $r(\tilde{C})$ increased by 1 (since we have created a new internal vertex with indegree 1). By iterating this procedure, we end up with a tree-child network which is binary. This proves our claim.

N $C(N)$

Figure 8: Left: The smallest example of a maximal reticulated reticulation-visible network with 2 leaves (with the tree-components highlighted). Right: The component graph of the network; note that it is a maximal reticulated binary tree-child network.



N $C(N)$

Figure 9: A maximal reticulated reticulation-visible network with 4 leaves and 9 reticulation vertices. The component graph is a maximal reticulated binary tree-child network. Note that N is the only network obtained by decompressing $C(N)$.

Claim 2: The maximum of $r(\tilde{C})$ over the set of all binary tree-child networks is $3\ell - 3$ and this bound is achieved if and only if $\tilde{C}$ is a maximal reticulated binary tree-child network.

Note that in any binary tree-child network, we have $\ell + k = t + 2$, where t is the number of tree vertices; see Section 1 in [10]. Thus,

$$r(\tilde{C}) = \ell + k + t - k = 2\ell + k - 2.$$

Since $k \leq \ell - 1$ (see Section 1) with this bound achieved exactly by the maximal reticulated tree-child

networks, the claim follows.

Overall, we have proved so far that the maximal number of reticulation vertices of a reticulation-visible network is $3\ell-3$ and this bound is achieved if and only if the component graph of the network is a maximal reticulated binary tree-child network. Finally, the tree vertices of these networks have one child which is a reticulation vertex and one child which is not; see Lemma 1 in [10]. Thus, they are replaced by a one-component network which has 2 leaves exactly one of which is below a reticulation vertex. However, the number of these one-component networks is $M_{2,1} = 1$. Consequently, the decompression of every maximal reticulated binary tree-child network gives exactly one reticulation-visible network; see Figure 8 for the smallest example (see also [1]) and Figure 9 for a larger example. ∎