Asymptotic enumeration of normal and hybridization networks via tree decoration

Michael Fuchs[1], Mike Steel[2], and Qiang Zhang[3]

[1] *Department of Mathematical Sciences, National Chengchi University, Taipei 116, Taiwan*

[2,3] *Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand*

December 10, 2024

Abstract

Phylogenetic networks provide a more general description of evolutionary relationships than rooted phylogenetic trees. One way to produce a phylogenetic network is to randomly place k arcs between the edges of a rooted binary phylogenetic tree with n leaves. The resulting directed graph may fail to be a phylogenetic network, and even when it is (and thereby a 'tree-based' network), it may fail to be a tree-child or normal network. In this paper, we first show that if k is fixed, the proportion of arc placements that result in a normal network tends to 1 as n grows. From this result, the asymptotic enumeration of normal networks becomes straightforward and provides a transparent meaning to the combinatorial terms that arise. Moreover, the approach extends to allow k to grow with n (at the rate $o(n^{\frac{1}{3}})$), which was not handled in earlier work. We also investigate a subclass of normal networks of particular relevance in biology (hybridization networks) and establish that the same asymptotic results apply.

Keywords: Phylogenetic network, trees, asymptotic enumeration, generating function

1 Introduction

Rooted phylogenetic networks provide a precise way to represent the evolution of objects (species, viruses, languages etc.) under the twin processes of speciation and reticulation [9]. The leaves (vertices of out-degree 0) of these networks typically correspond to observed individuals at the present, and the other vertices correspond to ancestral species. Over the last two decades, the mathematical, statistical, and computational properties of phylogenetic networks have become an active area of research. Various classes of networks with particular properties have been identified and enumerated, and the relationships between various classes of networks has been investigated (for a recent survey, see [10]).

The simplest phylogenetic network is a rooted tree, which models speciation (only). A slightly more general class is that of hybridization networks, which also allow pairs of species in the past to combine to form new (hybrid) species. More general classes of networks allow for a greater variety of reticulate evolution; for example, lateral gene transfer.

In this paper, we focus on a class of networks that includes hybridization networks, but is slightly more general, namely the class of *normal* networks. Such networks enjoy a number of desirable properties (see e.g., [5]). We describe a new way to asymptotically count the class of normal networks with n leaves and k reticulation vertices as n becomes large (with k fixed), a topic that has been investigated by quite different methods in [6], [7] and [8]. We then show how the results can be extended to allow k to grow (slowly) with n, and then extend this approach to the subclass of hybridization networks. We begin with some definitions.

1.1 Definitions: Networks, decorated trees, and induced subdivision trees

In this paper, all trees and networks are directed away from an ancestral root edge and are binary, i.e., all edges that have in-degree 1 either have out-degree 0 (i.e., they are the leaves of the network) or out-degree 2.

Throughout, we let $[n]$ denote the set $\{1, \ldots, n\}$. A *phylogenetic network* on $[n]$ is a directed acyclic graph with n leaves (vertices of out-degree 0) labelled bijectively by the elements of $[n]$, and with each non-leaf vertex having in-degree 1 and out-degree 2 (tree vertices) or in-degree 2 and out-degree 1 (reticulation vertices), or in-degree 0 and out-degree 1 (the root of the network at the top of a 'stem' edge). All edges are directed away from the root. Edges which contain a reticulation vertex are called *reticulation edges*; all others are called *tree edges*.

Two phylogenetic networks are regarded as equivalent if there is a directed graph isomorphism between them that maps i to i for each $i \in [n]$. Three important classes of phylogenetic networks are the following:

- A *tree-child network* is a phylogenetic network for which each non-leaf vertex has at least one of its outgoing edges directed to a tree vertex or a leaf.
- A *normal network* is a tree-child network that has no 'shortcut' edge (i.e., no edge (u, v) for which there is another path from u to v).
- A *phylogenetic tree* is a phylogenetic network with no reticulation vertices.

Thus, tree-child networks include normal networks which include phylogenetic trees. For more background and details on phylogenetic networks, see [9].

Let $\mathcal{T}_n$ denote the set of phylogenetic trees on leaf set $[n]$. For $T \in \mathcal{T}_n$ and $k \geqslant 1$, let $\mathcal{S}(T, k)$ denote the set of all possible ordered pairs (T_k, ω_k), where T_k and ω_k are defined recursively as follows: For $k = 1$, set $\omega_1 = (p_1, p_1')$, where p_1 subdivides some edge of T and p_1' subdivides some edge of the resulting tree. Let T_1 be the resulting subdivided tree (with two subdivision vertices).

For $k > 1$, let $\omega_k = \omega_{k-1} \cup \{(p_k, p_k')\}$ where p_k subdivides some edge of T_{k-1} and p_k' subdivides some edge of the resulting tree. Let T_k be the resulting subdivided tree (with $2k$ subdivision vertices).

We call (T_k, ω_k) a *k-fold decorated tree on* $[n]$ with base tree T. Thus, $S(T, k)$ is the set of k-fold decorated trees with base tree T.

Let $\mathcal{G}((T_k, \omega_k))$ be the directed graph obtained from T_k by adding an arc from p_i to p_i' for each $i = 1, \ldots, k$. Note that $\mathcal{G}((T_k, \omega_k))$ may contain directed cycles (in particular, it need not be a phylogenetic network). However, if $\mathcal{G}((T_k, \omega_k))$ has no cycles, then it is a tree-based network [4], and every tree-based network on $[n]$ can be generated this way.

We also introduce a further notion associated with any k-fold decorated tree (T_k, ω_k). Consider the tree induced by the $2k$ subdivision vertices, which we call the *induced subdivision tree.* Note that every leaf of the induced subdivision tree is a subdivision vertex; in addition, there might also be subdivision vertices which are non-leaf vertices.

These concepts are illustrated by the example shown in Fig. 1.

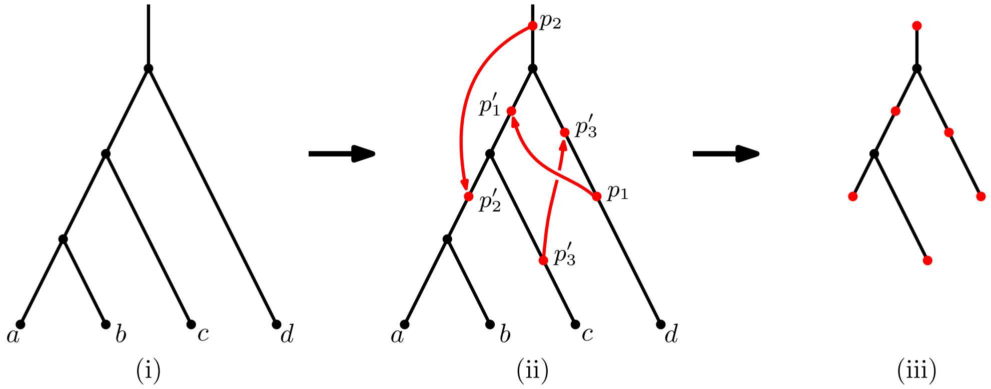


Figure 1: (i): A phylogenetic tree T with four leaves ($a, b, c, d \in [4]$), with all edges directed vertically downwards. (ii) The directed graph $\mathcal{G}((T_3, \omega_3))$ obtained from T by adding $k = 3$ pairs of subdivision points to produce a 3-fold decorated tree, and then joining p_i to p_i' for each i. In this example, the graph contains a directed cycle and thus does not correspond to a phylogenetic network. (iii) The associated induced subdivision tree which contains the six vertices $(p_1, p_2, p_3, p_1', p_2', p_3')$ (indicated in red) and two other vertices of T (indicated in black).

We also employ standard asymptotic notation: $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$, $f(n) = \mathcal{O}(g(n))$ if $f(n) \leqslant Cg(n)$ for a constant C, and $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. Moreover, we use $[z^n]f(z)$ for the n-th coefficient of a generating function $f(z)$, and for any odd integer $n > 1$, we let $n!!$ denote the product of the odd numbers from 1 to n.

2 Results

We begin by counting the set $\mathcal{S}(T, k)$.

Lemma 1. *The number of k-fold decorated trees (T_k, ω_k) on $[n]$ with base tree T is given by:*

$$|\mathcal{S}(T, k)| = \frac{(2n-1)\cdot(2n)\cdots(2n+2k-2)}{k!} \sim \frac{4^k n^{2k}}{k!}.$$

Proof. The number of arcs in the tree T (with a stem edge) is $(2n-1)$ so there are these many choices for p_1. Placing p_1 creates a tree with $2n$ arcs, so there are

$2n$ choices for p'_1. Continuing in this way for the k pairs of points (p_i, p'_i) gives the term on the numerator, and the $k!$ in the denominator accounts for the fact that the same k-fold decorated tree can be obtained by placing the pairs of points (p_i, p'_i) onto the arcs of T in any order.

The asymptotic part of the result is obtained by noting that the numerator is a polynomial of degree $2k$ in $2n$. □

Next, we define the set:

$$\mathcal{S}(n,k) := \bigcup_{T \in \mathcal{T}_n} \mathcal{S}(T,k).$$

Thus, by Lemma 1,

$$|\mathcal{S}(n,k)| = \sum_{T \in \mathcal{T}_n} |\mathcal{S}(T,k)| = \frac{(2n-1)(2n)\cdots(2n+2k-2)}{k!} r(n).$$

where

$$r(n) = |\mathcal{T}_n| = (2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!} \sim \frac{\sqrt{2}}{2}\left(\frac{2}{e}\right)^n n^{n-1}.$$

Consequently,

$$|\mathcal{S}(n,k)| \sim \frac{2^{2k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}. \tag{1}$$

Remark 2. For later purposes, we point out that (1) also holds for $k = o(\sqrt{n})$ as

$$\begin{aligned}
(2n-1)(2n)\cdots(2n+2k-2) &= (2n)^{2k}\left(1 - \frac{1}{2n}\right)\prod_{j=0}^{2k-2}\left(1 + \frac{j}{2n}\right) \\
&= (2n)^{2k}\left(1 + \mathcal{O}(n^{-1})\right) e^{\sum_{j=0}^{2k-2}\log\left(1+\frac{j}{2n}\right)} \\
&= (2n)^{2k}\left(1 + \mathcal{O}(n^{-1})\right) e^{\sum_{j=0}^{2k-2}\mathcal{O}(j/n)} \\
&= (2n)^{2k}\left(1 + \mathcal{O}(n^{-1})\right) e^{\mathcal{O}(k^2/n)} \\
&= (2n)^{2k}\left(1 + \mathcal{O}\left(\frac{k^2}{n}\right)\right),
\end{aligned}$$

where we used:

$$\sum_{j=0}^{2k-2} j = \frac{(2k-2)(2k-1)}{2} = \mathcal{O}(k^2).$$

Next, we partition $\mathcal{S}(n,k)$ into three disjoint subsets:

$$\mathcal{S}(n,k) = \mathcal{S}_c(n,k) \sqcup \mathcal{S}_{no}(n,k) \sqcup \mathcal{S}_{\neg no}(n,k).$$

The first set on the right ($\mathcal{S}_c(n,k)$) consists of all k-fold decorated trees (T_k, ω_k) on $[n]$ such that $\mathcal{G}((T_k,\omega_k))$ contains a cycle. Thus the remaining k-fold decorated trees on $[n]$ are such that $\mathcal{G}((T_k,\omega_k))$ is a phylogenetic network. We then partition this set of phylogenetic networks into two disjoint subsets: those networks for which $\mathcal{G}((T_k,\omega_k))$ is or is not a normal network ($\mathcal{S}_{no}(n,k)$ and $\mathcal{S}_{\neg no}(n,k)$, respectively).

Our goal is to show that the contributions of $\mathcal{S}_c(n,k)$ and $\mathcal{S}_{\neg no}(n,k)$ are asymptotically negligible in (1). We start by observing the following.

Lemma 3. *The induced subdivision tree has $4k-2-\ell$ edges where ℓ is the number of non-leaf subdivision vertices.*

Proof. If the induced subdivision tree has ℓ non-leaf subdivision vertices, then it is a tree with $2k-\ell$ leaves, ℓ unary vertices (including possibly the root; see, for example, the third panel in Figure 1), and the remaining $2k-\ell-1$ vertices are the binary vertices. Every vertex except the root has an incoming edge and this is the total number of edges. Thus, the number of edges is:

$$\underbrace{2k-\ell}_{\text{leaves}}+\underbrace{\ell}_{\substack{\text{unary}\\ \text{vertices}}}+\underbrace{2k-\ell-1}_{\text{binary vertices}}-\underbrace{1}_{\text{root}}=4k-2-\ell.$$

□

In addition, we have the following.

Lemma 4. *Denote by $S_{n,k,\ell}$ the number of k-fold decorated trees (T_k,ω_k) on $[n]$ such that the induced subdivision tree has exactly ℓ non-leaf subdivision vertices. Then, for fixed k and ℓ, as $n\to\infty$,*

$$S_{n,k,\ell}=\mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-1-\ell/2}\right).$$

Proof. Consider induced subdivision trees with $2k-\ell$ leaves and ℓ non-leaf subdivision vertices. The k-fold decorated trees (T_k,ω_k) which are counted by $S_{n,k,\ell}$ are obtained from these induced subdivision trees by attaching phylogenetic trees below the leaves and replacing edges by a sequence of phylogenetic trees, including a stem edge into the root, which has to be added to the induced subdivision tree. Note that, by Lemma 3, there are $4k-1-\ell$ such sequences in total.

The construction above, in terms of exponential generating function, gives the following for a fixed induced subdivision tree:

$$\underbrace{r(z)^{2k-\ell}}_{\substack{\text{trees below}\\ \text{leaves}}}\cdot\underbrace{\frac{1}{(1-r(z))^{4k-1-\ell}}}_{\substack{\text{sequences of trees}\\ \text{on edges}}},\tag{2}$$

where

$$r(z):=\sum_{n\geqslant 1}r(n)\frac{z^n}{n!}=1-\sqrt{1-2z}$$

counts phylogenetic trees and thus $1/(1-r(z))$ counts sequences of phylogenetic trees.

To obtain the coefficient of (2), we use singularity analysis; see Chapter VI of [3]. First, as $z\to 1/2$,

$$\frac{r(z)^{2k-\ell}}{(1-r(z))^{4k-1-\ell}}\sim\frac{1}{(1-2z)^{2k-1/2-\ell/2}}.$$

By Corollary VI.1 in [3], this gives (up to a constant) the upper bound for $S_{n,k,\ell}$

$$n![z^n]\frac{r(z)^{2k-\ell}}{(1-r(z))^{4k-1-\ell}}\sim n!2^n\frac{n^{2k-3/2-\ell/2}}{\Gamma(2k-1/2-\ell/2)}=\mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-1-\ell/2}\right)$$

as claimed. □

We can now show that the contributions of $\mathcal{S}_c(n,k)$ are asymptotically negligible in (1).

Proposition 5. *We have,*

$$|\mathcal{S}_c(n,k)| = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-3/2}\right).$$

Proof. If a k-fold decorated tree (T_k, ω_k) is in $\mathcal{S}_c(n,k)$, then $\mathcal{G}((T_k,\omega_k))$ contains a cycle. Thus, the induced subdivision tree has at least one non-leaf subdivision vertex. Applying Lemma 4 gives the claim. □

Next, we show that the contribution of $\mathcal{S}_{\neg no}(n,k)$ is also negligible.

Proposition 6. *We have,*

$$|\mathcal{S}_{\neg no}(n,k)| = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-3/2}\right).$$

Proof. By Lemma 4, we can restrict ourselves to k-fold decorated trees (T_k, ω_k) whose induced subdivision tree has $2k$ leaves. Note that in these decorated trees, the network $\mathcal{G}((T_k,\omega_k))$ has trees below the reticulation vertices and thus does not contain a reticulation vertex followed by a reticulation vertex. In addition, if $\mathcal{G}((T_k,\omega_k))$ contains a tree vertex followed by two reticulation vertices, then the number of such decorated trees is bounded by

$$n![z^n]\frac{r(z)^{2k}}{(1-r(z))^{4k-1-2}};$$

see the explanation in the proof of Lemma 4 (where we now have $\ell = 0$). Here, the additional -2 in the exponent of the denominator arises from the two edges below the above tree vertex in $\mathcal{G}((T_k,\omega_k))$ as empty sequences of phylogenetic trees are attached to these edges in the induced subdivision tree of (T_k,ω_k). This bound is

$$n![z^n]\frac{r(z)^{2k}}{(1-r(z))^{4k-1-2}} = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-2}\right). \tag{3}$$

Likewise, if we have a shortcut in $\mathcal{G}((T_k,\omega_k))$ (which violates the normal condition), then we obtain the upper bound

$$n![z^n]\frac{r(z)^{2k}}{(1-r(z))^{4k-1-1}} = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-3/2}\right), \tag{4}$$

where the additional -1 comes from the empty sequence attached to the shortcut in the induced subdivision tree of (T_k,ω_k) (which must be part of the induced subdivision tree). Combining these two upper bounds gives the claimed result. □

Propositions 5 and 6 provide an alternative and immediate way to asymptotically count normal networks with a given number of reticulations (k). Although this result is known (from [6], [7], [8]) our proof here provides a more transparent way to see why the asymptotic result holds; more importantly, it can be extended to allow k to grow (slowly) with n (as we describe in the next section), unlike the earlier approaches.

Let $N_{n,k}$ denote the number of normal phylogenetic networks with n leaves and k reticulation vertices.

Corollary 7. *For fixed k, as $n \to \infty$,*

$$N_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}. \tag{5}$$

Proof. Let L_n be the number of pairs (N, T) where N is a normal network with leaf set $[n]$ and k reticulation vertices, and $T \in \mathcal{T}_n$ is displayed by N. Since any normal network with k reticulation vertices displays exactly 2^k distinct phylogenetic trees (Corollary 3.4 of [15]), we have:

$$L_n = 2^k \cdot N_{n,k}. \tag{6}$$

In addition, by the definition of $\mathcal{S}(n,k)$, we have

$$L_n = |\mathcal{S}_n(n,k)|.$$

Now, from (1) and Propositions 5 and 6, we have:

$$L_n \sim \frac{2^{2k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}$$

and so, by (6):

$$N_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1},$$

which establishes (5), as required. □

3 Allowing k to grow (slowly) with n

In order to understand the range of validity of (5) when k is allowed to grow with n, we have to make the dependence on k in the $\mathcal{O}$ term in Proposition 5 and Proposition 6 explicit. Since both of these propositions crucially depend on Lemma 4, we first revisit the proof of this lemma.

From now on, we assume that $k = o(n^{1/3})$. This will turn out to be the range of k for which we can show that (5) is still valid.

By the proof of Lemma 4, the number of k-fold decorated trees (T_k, ω_k) on $[n]$ with a *fixed* induced subdivision tree having $2k - \ell$ leaves and ℓ non-leaf subdivision vertices is bounded by

$$n![z^n]\frac{r(z)^{2k-\ell}}{(1-r(z))^{4k-1-\ell}} = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-1-\ell/2}\right). \tag{7}$$

Our first goal is to sharpen this to

$$n![z^n]\frac{r(z)^{2k-\ell}}{(1-r(z))^{4k-1-\ell}} = \mathcal{O}\left(\left(\frac{2}{e}\right)^n \frac{n^{n+2k-1-\ell/2}}{\Gamma(2k-1/2-\ell/2)}\right), \tag{8}$$

where the implied constant in $\mathcal{O}$ is absolute (i.e., it does not dependent on ℓ, k, and n). In this section, $\mathcal{O}$ will always be assumed to have this property.

In order to prove this, we start with a lemma.

Lemma 8. *For $\alpha > 0$ with $\alpha = o(\sqrt{n})$, we have the following uniform bound*

$$[z^n](1-z)^{-\alpha} = \mathcal{O}\left(\frac{n^{\alpha-1}}{\Gamma(\alpha)}\right).$$

Proof. By the binomial theorem,

$$[z^n](1-z)^{-\alpha} = \binom{n+\alpha-1}{n} = \frac{\Gamma(n+\alpha)}{\Gamma(n+1)\Gamma(\alpha)}. \tag{9}$$

Next, by Stirling's formula for the Gamma function

$$\Gamma(x) = \left(\frac{x}{e}\right)^x \sqrt{\frac{2\pi}{x}}\left(1+\mathcal{O}\left(\frac{1}{x}\right)\right), \qquad (x \to \infty),$$

we have:

$$\begin{aligned}\log\frac{\Gamma(n+\alpha)}{\Gamma(n+1)} &= \log\Gamma(n+\alpha) - \log\Gamma(n+1)\\ &= (n+\alpha)\log(n+\alpha) - (n+\alpha) - \frac{1}{2}\log(n+\alpha) + \frac{1}{2}\log(2\pi)\\ &\quad - (n+1)\log(n+1) + n + 1 + \frac{1}{2}\log(n+1) - \frac{1}{2}\log(2\pi) + \mathcal{O}(n^{-1})\\ &= (\alpha-1)\log n + (n+\alpha-1/2)\log(1+\alpha/n) - \alpha + \mathcal{O}(n^{-1})\\ &= (\alpha-1)\log n + \mathcal{O}\left(\frac{\alpha^2+1}{n}\right).\end{aligned}$$

Thus,

$$\frac{\Gamma(n+\alpha)}{\Gamma(n+1)} = n^{\alpha-1}\left(1+\mathcal{O}\left(\frac{\alpha^2+1}{n}\right)\right)$$

and combining this with (9) yields the claim. □

Now, to establish (8), we use $r(z) = 1-\sqrt{1-2z}$ and the binomial theorem:

$$[z^n]\frac{r(z)^{2k-\ell}}{(1-r(z))^{4k-1-\ell}} = \sum_{j=0}^{2k-\ell}\binom{2k-\ell}{j}(-1)^j[z^n](1-2z)^{-2k+1/2+\ell/2+j/2}.$$

By the lemma, the j-th term becomes

$$\begin{aligned}&\mathcal{O}\left(\binom{2k-\ell}{j}2^n\frac{n^{2k-3/2-\ell/2-j/2}}{\Gamma(2k-1/2-\ell/2-j/2)}\right)\\ &\qquad= \mathcal{O}\left(\frac{2^n n^{2k-3/2-\ell/2}}{\Gamma(2k-1/2-\ell/2)}\frac{\Gamma(2k-1/2-\ell/2)}{\Gamma(2k-1/2-\ell/2-j/2)}\frac{(2k/\sqrt{n})^j}{j!}\right)\\ &\qquad= \mathcal{O}\left(\frac{2^n n^{2k-3/2-\ell/2}}{\Gamma(2k-1/2-\ell/2)}\frac{(2k/\sqrt[3]{n})^{3j/2}}{j!}\right),\end{aligned}$$

where, in the second-last step, we used:

$$\frac{\Gamma(2k-1/2-\ell/2)}{\Gamma(2k-1/2-\ell/2-j/2)} = \mathcal{O}((2k)^{j/2}),$$

which follows from standard estimates for the ratio of gamma functions. Summing the above estimate over j gives

$$\begin{aligned}[z^n]\frac{r(z)^{2k-\ell}}{(1-r(z))^{4k-1-\ell}} &= \mathcal{O}\left(\frac{2^n n^{2k-3/2-\ell/2}}{\Gamma(2k-1/2-\ell/2)}e^{(2k/\sqrt[3]{n})^{3/2}}\right)\\ &= \mathcal{O}\left(\frac{2^n n^{2k-3/2-\ell/2}}{\Gamma(2k-1/2-\ell/2)}\right).\end{aligned}$$

Multiplying this by the asymptotics of $n!$ gives (8).

Next, we have to multiply (8) by the number of induced subdivision trees with $2k-\ell$ leaves and ℓ non-leaf subdivision vertices which is given by:

$$\binom{2k}{\ell}r(2k-\ell)\frac{(4k-2\ell-1)\cdots(4k-\ell-2)}{k!}$$

as they are enumerated by starting with a phylogenetic tree on $2k-\ell$ leaves, choosing ℓ non-leaf subdivision vertices on the edges of this tree, and redistributing the labels. Note that

$$\begin{aligned}&\binom{2k}{\ell}r(2k-\ell)\frac{(4k-2\ell-1)\cdots(4k-\ell-2)}{\Gamma(2k-1/2-\ell/2)k!}\\ &\qquad\leqslant \frac{4^\ell k^\ell}{\ell!k!}\frac{\Gamma(2k+1)}{\Gamma(2k-1/2-\ell/2)}\frac{C_{2k-\ell-1}}{2^{2k-\ell-1}},\end{aligned}$$

where C_n denotes the n-th Catalan number:

$$C_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n^3}}.$$

Therefore, the above bound becomes:

$$\frac{4^\ell k^\ell}{\ell!k!}\frac{\Gamma(2k+1)}{\Gamma(2k-1/2-\ell/2)}\frac{C_{2k-\ell-1}}{2^{2k-\ell-1}} = \mathcal{O}\left(\frac{4^k(2k)^{3\ell/2}k^{3/2}}{\ell!k!(2k-\ell)^{3/2}}\right), \tag{10}$$

where we used

$$\frac{\Gamma(2k+1)}{\Gamma(2k-1/2-\ell/2)} = \mathcal{O}\left(2^{\ell/2}k^{(\ell+3)/2}\right).$$

Combining (8) and (10), the bound in Lemma 4 can be sharpened to

$$S_{n,k,\ell} = \mathcal{O}\left(\frac{(2k)^{3\ell/2}k^{3/2}}{\ell!(2k-\ell)^{3/2}}\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1-\ell/2}\right).$$

The bound in Proposition 5 for $|\mathcal{S}_c(n,k)|$ is obtained by summing the bound for $S_{n,k,\ell}$ for ℓ from 1 to $2k-1$. Consequently,

$$|\mathcal{S}_c(n,k)| = \mathcal{O}\left(c(n,k)\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right),$$

where

$$c(n,k) := k^{3/2}\sum_{\ell=1}^{2k-1}\frac{(2k/\sqrt[3]{n})^{3\ell/2}}{\ell!(2k-\ell)^{3/2}}$$

We break the above sum into two parts according to whether $\ell < k$ or $\ell \geqslant k$. For the first part, we get

$$k^{3/2}\sum_{\ell=1}^{k-1}\frac{(2k/\sqrt[3]{n})^{3\ell/2}}{\ell!(2k-\ell)^{3/2}} = \mathcal{O}\left(\sum_{\ell=1}^{k-1}(2k/\sqrt[3]{n})^{3\ell/2}\right) = \mathcal{O}\left((2k/\sqrt[3]{n})^{3/2}\right) = o(1),$$

where, in the last step, we used that $k = o(n^{1/3})$. For the second part, we have

$$k^{3/2}\sum_{\ell=k}^{2k-1}\frac{(2k/\sqrt[3]{n})^{3\ell/2}}{\ell!(2k-\ell)^{3/2}} = \mathcal{O}\left(k^{3/2}(2k/\sqrt[3]{n})^{3k/2}\right) = o(1),$$

again by $k = o(n^{1/3})$. Thus, we have

$$|\mathcal{S}_c(n,k)| = o\left(\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right) \tag{11}$$

for our range of k.

Next, we consider the estimate of Proposition 6 which, as explained in the proof, was obtained by estimating separately the number of k-fold decorated trees (T_k, ω_k) on $[n]$ with $\mathcal{G}((T_k,\omega_k))$ a phylogenetic network whose induced subdivision tree has (i) at least one non-leaf subdivision vertex or otherwise (ii) two outgoing edges of a tree node where no sequence is attached to or (iii) one edge where no sequence is attached to in the induced subdivision tree.

The first case is treated as above. For the second and third cases, we use the estimate (3) and (4) instead of (7), where both bounds have to be multiplied by k which is the upper bound of the number of tree nodes and edges in the induced subdivision tree, respectively. Moreover, we can use the same constant as in (10) but with $\ell = 0$ and multiplied by $\Gamma(2k-1/2)/\Gamma(2k-3/2) = \mathcal{O}(k)$ in the second case and by $\Gamma(2k-1/2)/\Gamma(2k-1) = \mathcal{O}(k^{1/2})$ in the third case. Overall, for the second case, this gives

$$\mathcal{O}\left(\frac{k^2}{n}\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right) = o\left(\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right)$$

and for the third case

$$\mathcal{O}\left(\frac{k^{3/2}}{\sqrt{n}}\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right) = o\left(\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right) \tag{12}$$

Combining these bounds with the bound for the first case, we can improve the result of Proposition 6 to

$$|\mathcal{S}_{\neg no}(n,k)| = o\left(\frac{4^k}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}\right) \tag{13}$$

for $k = o(n^{1/3})$.

Now, by combining (11), (13), and (1) (which holds for our range of k; see Remark 2), we see that (5) holds even if k is allowed to grow moderately with n, namely, for $k = o(n^{1/3})$.

4 Hybridization networks

A *hybridization network* is a tree-child network on leaf set X, which has at least one temporal ordering (or 'ranking'). This means that one can assign a real-valued temporal date ($T(v)$) to each vertex v of the networks so that (i) if (u, v) is a tree edge then $T(u) < T(v)$ and (ii) if v is a reticulation vertex with parents u and w then $T(u) = T(v) = T(w)$. Hybridization networks are particularly relevant to biology, since they model species' evolution that comprises two processes: binary speciation events (as in phylogenetic trees), and events where two contemporaneous species hybridize to give rise to a new (hybrid) species (see e.g., [11]).

It can easily be shown that every hybridization network is normal (see e.g., Proposition 10.12 of [14]); however, the converse does not hold, as Fig. 2 shows.

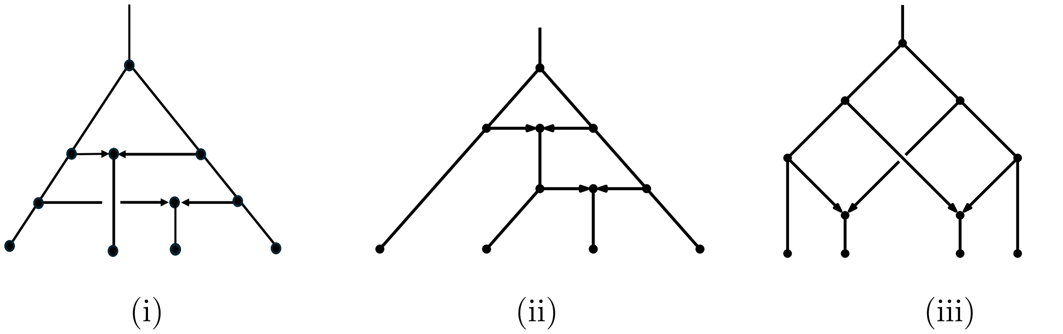


Figure 2: The three shapes of normal networks on $n = 4$ leaves with $k = 2$ reticulation vertices. Cases (i) and (ii) correspond to hybridization networks, and Case (iii) corresponds to a normal network that is not a hybridization network. There are 12 distinct phylogenetic networks of shape (i), 24 of shape (ii), and 12 of shape (iii).

Hybridization networks also correspond to the class of tree-child networks that can be 'ranked' in the sense described in [1]. However, although there is a simple, exact, and explicit formula for counting ranked tree-child networks with n leaves and k reticulation vertices, the exact enumeration of hybridization networks is more complex.

Let $H_{n,k}$ denote the number of hybridization networks with n leaves and k reticulation vertices. $H_{n,1}$ is just the number of normal networks with one reticulation vertex, and so $H_{4,1} = 54$; see (14). In addition, $H_{4,2} = 36$, compared with the 48 possible normal networks with $n = 4, k = 2$ (see Fig. 2).

4.1 Enumeration of $H_{n,k}$ for $k = 1, 2$

$H_{n,1}$ coincides with the number of normal networks with one reticulation vertex and thus:

$$H_{n,1} = N_{n,1} = \frac{1}{2}n![z^n]\left(\frac{r(z)}{1-r(z)}\right)^3 = \frac{1}{2}[(2n+1)!! + 3(2n-1)!!] - 3n!2^{n-1}, \quad (14)$$

where $r(z) = 1 - \sqrt{1-2z}$; see [16] or [7].

For $k = 2$, we have the following result.

Proposition 9. *We have,*

$$H_{n,2} = (2n-1)!!(n^3 + 9n^2 - 16n - 12) - 3n!2^n(n^2 - 4).$$

Proof. The proof uses the decomposition of a network into *bridgeless components* from [2]. First, we recall the definition of a bridgeless component from graph theory: a bridgeless component of a graph G is a maximal induced subgraph of G without cut edges (bridges). Let $\mathcal{H}_2$ denote the set of hybridization networks with exactly 2 reticulation vertices. Consider the exponential generating function for $H_{n,k}$ defined by:

$$H_k(z) = \sum_{n=1}^{\infty} H_{n,k} \frac{z^n}{n!}.$$

Any hybridization network $N \in \mathcal{H}_2$ satisfies exactly one of the following cases (for details, we refer to the Supplementary material):

Case 1: The root of N belongs to a bridgeless component which contains 0 reticulation vertices. This contributes $H_2(z)r(z) + \frac{H_1(z)^2}{2}$ to the exponential generating function of $\mathcal{H}_2$.

Case 2: The root of N belongs to a bridgeless component which contains exactly one reticulation vertex. This contributes $H_1(z)\frac{r(z)^2}{(1-r(z))^3} + \frac{1}{2}H_1(z)\frac{r(z)^2}{(1-r(z))^2}$ to the exponential generating function of $\mathcal{H}_2$.

Case 3: The root of N belongs to a bridgeless component which contains exactly two reticulation vertices. This contributes $\frac{1}{2}\frac{r(z)^4}{(1-r(z))^4} + \frac{r(z)^4}{(1-r(z))^5} + \frac{7}{2}\frac{r(z)^5}{(1-r(z))^5} + \frac{5}{4}\frac{r(z)^6}{(1-r(z))^6}$ to the exponential generating function of $\mathcal{H}_2$.

Then, for $k = 2$, we have:

$$\begin{aligned} H_2(z) = H_2(z)r(z) + \frac{H_1(z)^2}{2} + H_1(z)\frac{r(z)^2}{(1-r(z))^3} + \frac{1}{2}H_1(z)\frac{r(z)^2}{(1-r(z))^2} \\ + \frac{1}{2}\frac{r(z)^4}{(1-r(z))^4} + \frac{r(z)^4}{(1-r(z))^5} + \frac{7}{2}\frac{r(z)^5}{(1-r(z))^5} + \frac{5}{4}\frac{r(z)^6}{(1-r(z))^6} \end{aligned}$$

and consequently,

$$\begin{aligned} H_2(z) = \frac{H_1(z)^2}{2(1-r(z))} + H_1(z)\frac{r(z)^2}{(1-r(z))^4} + \frac{1}{2}H_1(z)\frac{r(z)^2}{(1-r(z))^3} \\ + \frac{1}{2}\frac{r(z)^4}{(1-r(z))^5} + \frac{r(z)^4}{(1-r(z))^6} + \frac{7}{2}\frac{r(z)^5}{(1-r(z))^6} + \frac{5}{4}\frac{r(z)^6}{(1-r(z))^7}, \end{aligned}$$

where (from (14)):

$$H_1(z) = \frac{1}{2}\frac{r(z)^3}{(1-r(z))^3}.$$

If we now let $\omega = \sqrt{1-2z} = 1 - r(z)$, then $H_2(z)$ can be rewritten as:

$$\begin{aligned} H_2(z) &= \frac{15}{8}\omega^{-7} - 6\omega^{-6} + \frac{27}{8}\omega^{-5} + 9\omega^{-4} - \frac{123}{8}\omega^{-3} + 9\omega^{-2} - \frac{15}{8}\omega^{-1} \\ &= \frac{15}{8}(1-r(z))^{-7} - 6(1-r(z))^{-6} + \frac{27}{8}(1-r(z))^{-5} + 9(1-r(z))^{-4} \\ &\quad - \frac{123}{8}(1-r(z))^{-3} + 9(1-r(z))^{-2} - \frac{15}{8}(1-r(z))^{-1}. \end{aligned}$$

By using this equation, it can be shown that:

$$H_{n,2} = n![z^n]H_2(z) = (2n-1)!!(n^3 + 9n^2 - 16n - 12) - 3n!2^n(n^2 - 4).$$

This proves the claim. □

By Proposition 9, we have $H_{4,2} = 36, H_{5,2} = 1890$, and $H_{6,2} = 66960$.

By comparison, the number $N_{n,2}$ of normal networks with two reticulation vertices is given (from [7]) by:

$$N_{n,2} = \frac{1}{3}(2n-1)!!(3n-4)(n^2+11n+6) - n!2^n(3n^2+2n-8).$$

It is readily verified that $H_{n,2}/N_{n,2} \sim 1$ as $n \to \infty$. Thus it is of interest to consider the (asymptotic) number of normal networks with two reticulation vertices that are not hybridization networks.

From the above expressions we have:

$$N_{n,2} - H_{n,2} \sim \frac{2\sqrt{2}}{3}\left(\frac{2}{e}\right)^n n^{n+2}, \quad \text{as} \quad n \to \infty.$$

4.2 Asymptotics of $H_{n,k}$

We start with a useful notion. We call two reticulation edges of a phylogenetic network *collinear* if a vertex of one of them is connected by a tree path (i.e. a path consisting just of tree edges) to the vertex of the other one. For example, in Fig. 2, the two parallel reticulation edges that slope downwards to the right are collinear.

Using this notion, we have the following result.

Lemma 10. *Let (T_k, ω_k) be a k-fold decorated tree on $[n]$ such that $\mathcal{G}((T_k, \omega_k))$ is a normal network which has no collinear reticulation edges. Then, $\mathcal{G}((T_k, \omega_k))$ is a hybridization network.*

Proof. The proof proceeds by induction on k. The claim obviously holds for $k = 0$, as, in this case, $\mathcal{G}((T_k, \omega_k))$ is just a tree, and any tree is a hybridization network.

Now suppose that the claim holds for $k-1$; we are going to establish it for k. Pick a reticulation vertex of $\mathcal{G}((T_k, \omega_k))$. Note that, by assumption, the three subgraphs induced by the descendant set of the reticulation vertex as well as the descendant sets of the two parents of the reticulation vertex are all trees. Remove them together with the reticulation vertex from $\mathcal{G}((T_k, \omega_k))$. The remaining structure corresponds to a $(k-1)$-fold decorated tree $(T'_{k-1}, \omega'_{k-1})$ on a set of leaves that consist of a subset of $[n]$ together with two new leaves that correspond to the parents of the removed reticulation vertex and which receive new labels (say, $n+1, n+2$). By applying the induction hypothesis, we obtain that $\mathcal{G}((T'_{k-1}, \omega'_{k-1}))$ is a hybridization network, i.e., it can be drawn from top to bottom such that its events are in chronological order. Next, find the two extant lineages leading to $n+1$ and $n+2$ and attach the deleted reticulation event, which becomes the next event in the temporal order. Moreover, attach to the leaves with label $n+1$ and $n+2$ the two deleted subtrees and below the re-attached reticulation vertex the third deleted subtree. Clearly, the events of the three subtrees can be ordered such that $\mathcal{G}((T_k, \omega_k))$ becomes a hybridization network. This proves the claim. □

Now, consider the set $\mathcal{S}_{no}(n,k)$, which we partition into the set which contains hybridization networks and the set which does not:

$$\mathcal{S}_{no}(n,k) = \mathcal{S}_{hyb}(n,k) \sqcup \mathcal{S}_{\neg hyb}(n,k).$$

The cardinality of the set of hybridization networks again satisfies the asymptotics in (1) since we have the following result.

Proposition 11. *We have,*

$$|\mathcal{S}_{\neg hyb}(n,k)| = \mathcal{O}\left(\left(\frac{2}{e}\right)^n n^{n+2k-3/2}\right).$$

Proof. As in the proof of Proposition 6, we can restrict ourselves to k-fold decorated trees (T_k, ω_k) whose induced subdivision tree has $2k$ leaves. In addition, if $(T_k, \omega_k) \in \mathcal{S}_{\neg hyb}(n,k)$, then, by Lemma 10, $\mathcal{G}((T_k, \omega_k))$ has collinear reticulation edges. Clearly, if one starts from an induced subdivision tree with $2k$ leaves, such a network can only be obtained from it (using the procedure from the proof of Lemma 5) by leaving at least one edge empty (i.e., by not attaching a sequence of subtrees to it). Thus, this situation is akin to the case of shortcuts in the proof of Proposition 6, which gave the bound (4). Since this is also the bound claimed in the current result, we are finished. □

In summary, we have the following result.

Theorem 12. *For fixed k, as $n \to \infty$,*

$$H_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}. \tag{15}$$

In addition, this asymptotic result is still valid in the range $k = o(n^{1/3})$.

Proof. The claimed expansion (15) follows from (1) combined with Proposition 5, Proposition 6, and Proposition 11. Moreover, that (15) holds for $k = o(n^{1/3})$ is proved by making the dependence on k in the constant of Proposition 11 explicit, which is handled as in Section 3 (as explained in the proof of Proposition 11, the situation is akin to the case of shortcuts in the proof of Proposition 6, which gave the bound (12)). □

5 Concluding comments

In this paper, we have investigated the enumerative aspects of constructing a phylogenetic network by placing arcs between the edges of the tree. Provided that the number of arcs placed (k) is constant or grows slowly enough with the number of leaves (n), the directed graph we construct is almost surely a (tree-based) phylogenetic network and, in addition, it (almost surely) belongs to the much smaller class of normal networks. This result, combined with a combinatorial property of normal networks, allows an asymptotic enumeration of this class. Our approach provides an explicit interpretation of the various terms in the asymptotic formula, and extends earlier results by allowing k to depend on n. We also show that the same asymptotic results apply for the even smaller subclass of hybridization networks.

Our analysis requires k to grow no faster than $o(n^{1/3})$ and a natural question is whether the bound $k = o(n^{1/3})$ might be improved. For example, $k = o(n^{1/2})$ seems to be the range where the corresponding result for tree-child networks holds;

see [13]. Is the same true for normal and/or hybridization networks? Or do they behave differently?

We end with some general observations. First, it is well known that any normal network has at most $n-2$ reticulation vertices, and by Theorem 5.1 of [12], almost all normal networks with n leaves have $(1+o(1))n$ reticulation vertices. Thus the two classes of networks we are enumerating are not representative of normal networks selected uniformly at random. Nor are our normal networks representative of a randomly chosen tree-child network with n leaves, since the proportion of the latter that are normal tends to 0 as n grows (again by results in [12]). Nevertheless, almost all of the tree-child networks (with k (fixed) reticulation vertices) that arise under the process we study will be normal networks (since this class of tree-child networks follows the same asymptotic law [6]).

6 Acknowledgments

This research was carried out during a sabbatical stay of MF at the Biomathematics Research Center, University of Canterbury, Christchurch. He thanks the department and MS for hospitality, and the National Science and Technology Council, Taiwan (NSTC-113-2918-I-004-001 and NSTC-113-2115-M-004-004-MY3) for financial support. MS and QZ thank the NZ Marsden Fund for research support (23-UOC-003).

References

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

[2] Bouvel, M., Gambette, P. and Mansouri, M., Counting phylogenetic networks of level 1 and 2, J. Math. Biol. 81 (6-7): 1357–1395 (2022).

[3] Flajolet, P. and Sedgewick, S., Analytic Combinatorics, Cambridge University Press, NY, 2009.

[4] Francis, A. and Steel, M., Which phylogenetic networks are merely trees with additional arcs? Syst. Biol. 64 (5):768-777 (2015).

[5] Francis, A., “Normal" phylogenetic networks emerge as the leading class, 2021. Arxiv: 2107.10414.

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

[7] Fuchs, M., Gittenberger, B. and Mansouri, M., Counting phylogenetic networks with few reticulation vertices: Exact enumeration and corrections, Australas. J. Combin. 81 (2):257–282 (2021).

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

[9] Huson, D., Rupp, R. and Scornavacca, C., Phylogenetic Networks: Concepts, Algorithms and Applications, Cambridge University Press, NY, 2010.

[10] Kong, S., Pons, J.C., Kubatko, L. and Wicke, K., Classes of explicit phylogenetic networks and their biological and mathematical significance, J. Math. Biol. 84 (6):47.

[11] Marcussen, T., Sandve, S.R., Heier. L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, Jakobsen, K.S., Wulff, B.B., Steuernagel, B., Mayer, K.F. and Olsen, O. A., Ancient hybridizations among the ancestral genomes of bread wheat, Science (6194):1250092 (2014).

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

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

[14] Steel, M., Phylogeny: Discrete and Random Processes in Evolution, SIAM, 2016.

[15] Willson, S. J., Tree-average distances on certain phylogenetic networks have their weights uniquely determined, Algorithms Mol. Biol. 7:Paper 13 (2012).

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

7 Supplementary material

It was mentioned in Section 4.1 that any hybridization network $N \in \mathcal{H}_2$ satisfies exactly one of the following three cases: (1) The root of N belongs to a bridgeless component which contains 0 reticulation vertices; (2) The root of N belongs to a bridgeless component which contains exactly 1 reticulation vertex; (3) The root of N belongs to a bridgeless component which contains exactly 2 reticulation vertices. Each case makes corresponding contributions to the exponential generating function of $\mathcal{H}_2$.

In this section, we will explain more carefully how to derive the exponential generating function of $\mathcal{H}_2$ by considering these three cases; we will follow the approach from [2].

A hybridization network N is said to be *level-k* if the number of reticulation vertices contained in any bridgeless component of N is at most k.

For any bridgeless component B with $k_B \leqslant k$ reticulation vertices of a level-k hybridization network N, there exist at least two bridges of N attached to B because without this restriction, such networks would have an unbounded number of internal vertices resulting in an infinite number of these networks.

For an arbitrary hybridization network $N \in \mathcal{H}_2$, considering the bridgeless component containing the root ρ denoted as B_ρ, suppressing any vertices of in-degree and out-degree 1, the resulting directed multi-graph is called a *generator*. A generator induced from B_ρ may have 0, 1, or 2 reticulation vertex (vertices) and is called a *level*-0, *level*-1 or *level*-2 generator. We call non-terminal edges and vertices of out-degree 0 of a level-k generator *sides*. For each level, there is a finite number of generator(s). Therefore, the three cases can be considered as: (1) The root of N belongs to a level-0 generator; (2) The root of N belongs to a level-1 generator; (3) The root of N belongs to a level-2 generator.

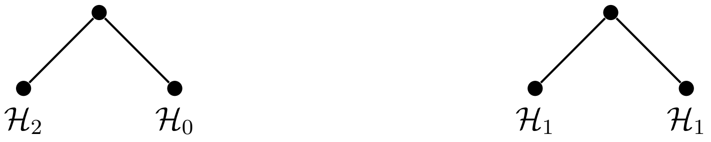
We will now consider these three cases. Let $\mathcal{H}_1$ denote the set of hybridization networks with exactly 1 reticulation vertex, and $\mathcal{H}_0$ denote the set of hybridization networks with 0 reticulation vertices.

7.1 Case 1

For any $N \in \mathcal{H}_2$ such that the root of N belongs to a level-0 generator, the root has two children. They satisfy one of two subcases:

- One child is the root of a network from $\mathcal{H}_2$ and the other child is the root of a network from $\mathcal{H}_0$, see case 1.1 in Fig. 3. This contributes $H_2(z)r(z)$ to the exponential generating function of $\mathcal{H}_2$.
- Both children are the root of networks of $\mathcal{H}_1$; see case 1.2 in Fig. 3. This contributes $\frac{H_2(z)^2}{2}$ to the exponential generating function of $\mathcal{H}_2$. Here, the factor 1/2 is because the left-to-right order is irrelevant.

Overall, Case 1 contributes $H_2(z)r(z) + \frac{H_2(z)^2}{2}$ to the exponential generating function of $\mathcal{H}_2$.



(case 1.1)

(case 1.2)

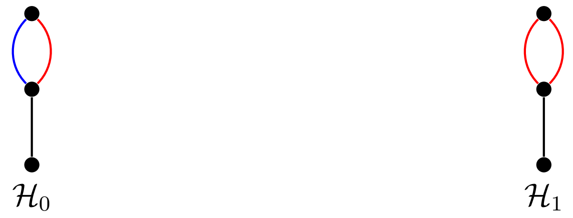
Figure 3: (case 1.1): A binary root vertex with two children that are roots of networks of $\mathcal{H}_2$ and $\mathcal{H}_0$. (case 1.2): A binary root vertex with two children that are roots of networks from $\mathcal{H}_1$ whose left-to-right order is irrelevant.

7.2 Case 2

Any $N \in \mathcal{H}_2$ such that the root of N belongs to a level-1 generator satisfies one of two subcases:

- The child of the reticulation vertex is the root of a network from $\mathcal{H}_0$ as illustrated in case 2.1 of Fig. 4.

  Exactly one level-1 network is attached to the blue side and there may or may not be level-0 network(s) attached to the blue side. Suppose that there are $k \geqslant 1$ network(s) that are attached to the blue side, exactly one of them is the level-1 network and the rest are level-0 networks. The level-1 network has k options to be placed, therefore the networks attached to the blue side contribute $\sum_{k=1}^{\infty} kH_1(z)r(z)^{k-1}$. A sequence of at least one network of $\mathcal{H}_0$ is



(case 2.1) (case 2.2)

Figure 4: (case 2.1): The child of the reticulation vertex of the level-1 generator containing the root is attached to $\mathcal{H}_0$. A sequence of at least one network from $\mathcal{H}_0$ is attached to the red side. There exists exactly one network from $\mathcal{H}_1$ attached to the blue side which also may or may not contain level-0 networks attached to it. (case 2.2): The child fo the reticulation vertex of the level-1 generator containing the root is attached to $\mathcal{H}_1$. A sequence of at least one networks from $\mathcal{H}_0$ is attached to each red side whose left-to-right order is irrelevant.

attached to the red side (otherwise this is a shortcut) and thus contributes $\frac{r(z)}{1-r(z)}$. Case 2.1 contributes $\sum_{k=1}^{\infty} kH_1(z)r(z)^{k-1}r(z)\frac{r(z)}{1-r(z)} = H_1(z)\frac{r(z)^2}{(1-r(z))^3}$.

- A sequence of at least one network from $\mathcal{H}_0$ is attached to each red side, whose left-to-right order is irrelevant and contributes $\frac{1}{2}H_1(z)\left(\frac{r(z)}{1-r(z)}\right)^2$ as seen in case 2.2 of Fig. 4.

Overall, Case 2 contributes $H_1(z)\frac{r(z)^2}{(1-r(z))^3} + \frac{1}{2}H_1(z)\left(\frac{r(z)}{1-r(z)}\right)^2$ to the exponential generating function of $\mathcal{H}_2$.

7.3 Case 3

Any $N \in \mathcal{H}_2$ such that the root of N belongs to a level-2 generator satisfies one of the four subcases:

- Case 3.1: The child of the bottom reticulation vertex is the root of a network from $\mathcal{H}_0$ as seen in case 3.1 of Fig. 5.

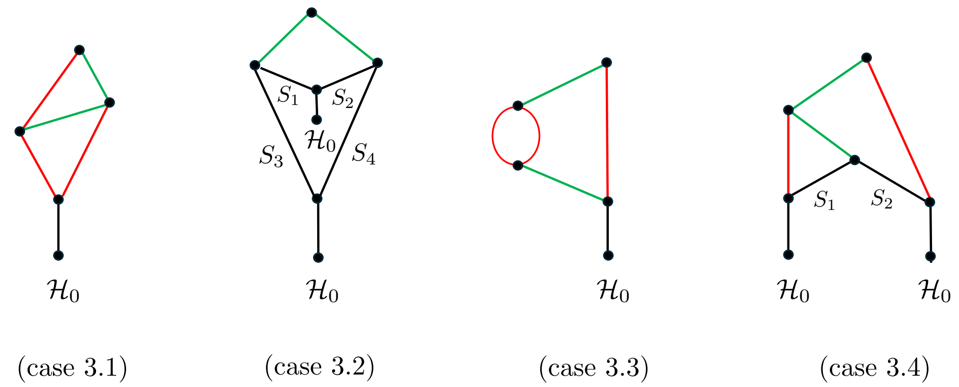
  All the three red sides cannot be empty, which means that at least one network from $\mathcal{H}_0$ is attached to them, otherwise the tree-child network and normal network conditions are violated. The two green sides can be empty, which means there may or may not be networks from $\mathcal{H}_0$ attached to them. This contributes $r(z)\left(\frac{r(z)}{1-r(z)}\right)^3\left(\frac{1}{1-r(z)}\right)^2$.

- Case 3.2: The children of the reticulation vertices are the roots of networks from $\mathcal{H}_0$ as seen in case 3.2 of Fig. 5 whose left-to-right and top-to-bottom orders are irrelevant. We consider the following situations:

  - S_1, S_2, S_3, S_4 are all non-empty. The two green sides can be empty. This contributes

    $$\underbrace{\frac{1}{4}}_{\text{symmetry}} r(z)^2\left(\frac{r(z)}{1-r(z)}\right)^4\left(\frac{1}{1-r(z)}\right)^2.$$

  - Exactly one of S_1, S_2, S_3, S_4 is empty. The two green sides can be empty.



(case 3.1) (case 3.2) (case 3.3) (case 3.4)

Figure 5: (case 3.1): The child of the bottom reticulation vertex is attached to $\mathcal{H}_0$. The three red sides are non-empty while the two green sides can be empty. (case 3.2): Each child of the reticulation vertices is attached to $\mathcal{H}_0$. The two green sides can be empty. There are three subcases which depend on whether there are networks attached to S_1, S_2, S_3, S_4. The top-to-bottom and left-to-right orders are irrelevant. (case 3.3): The child of the bottom reticulation vertex is attached to $\mathcal{H}_0$. The three red sides are non-empty. The two green sides can be empty. The left-to-right order is irrelevant. (case 3.4): Each child of the reticulation vertices is attached to $\mathcal{H}_0$. The two red sides are non-empty. The two green sides can be empty. There are two subcases which depend on whether there are networks attached to S_1 and S_2.

This contributes

$$\begin{aligned}\underbrace{4}_{\text{4 cases}} \cdot \underbrace{\frac{1}{4}}_{\text{symmetry}} \; r(z)^2 \left(\frac{r(z)}{1-r(z)}\right)^3 \left(\frac{1}{1-r(z)}\right)^2 \\ = r(z)^2 \left(\frac{r(z)}{1-r(z)}\right)^3 \left(\frac{1}{1-r(z)}\right)^2.\end{aligned}$$

- S_1, S_2 are empty, while S_3, S_4 are non-empty and vice versa. The two blue sides can be empty. This contributes

$$\begin{aligned}\underbrace{2}_{\text{2 cases}} \cdot \underbrace{\frac{1}{4}}_{\text{symmetry}} \; r(z)^2 \left(\frac{r(z)}{1-r(z)}\right)^2 \left(\frac{1}{1-r(z)}\right)^2 \\ = \frac{1}{2}\left(\frac{r(z)}{1-r(z)}\right)^4.\end{aligned}$$

- Case 3.3: The child of the bottom reticulation vertex is the roots of networks from $\mathcal{H}_0$ as seen in case 3.3 of Fig. 5. All three red sides cannot be empty. The two green sides can be empty. The left-to-right order is irrelevant. This contributes

$$\underbrace{\frac{1}{2}}_{\text{symmetry}} \; r(z) \left(\frac{r(z)}{1-r(z)}\right)^4 \left(\frac{1}{1-r(z)}\right).$$

- Case 3.4: The children of the reticulation vertices are the root of networks from $\mathcal{H}_0$ as seen in case 3.4 of Fig. 5. We consider two situations:
  - Either S_1 or S_2 is empty. The two red sides cannot be empty otherwise they are shortcuts and the two green sides can be empty. This

contributes

$$\underbrace{2}_{\text{2 cases}} r(z)^2 \left(\frac{r(z)}{1-r(z)}\right)^3 \left(\frac{1}{1-r(z)}\right)^2.$$

- S_1 and S_2 are both non-empty. The two red sides cannot be empty. The two green sides can be empty. This contributes

$$r(z)^2 \left(\frac{r(z)}{1-r(z)}\right)^4 \left(\frac{1}{1-r(z)}\right)^2.$$

In general, Case 3 contributes $\frac{1}{2}\frac{r(z)^4}{(1-r(z))^4} + \frac{r(z)^4}{(1-r(z))^5} + \frac{7}{2}\frac{r(z)^5}{(1-r(z))^5} + \frac{5}{4}\frac{r(z)^6}{(1-r(z))^6}$ to the exponential generating function of $\mathcal{H}_2$.