COUNTING PHYLOGENETIC NETWORKS

COLIN MCDIARMID, CHARLES SEMPLE, AND DOMINIC WELSH

For our friend and colleague James Oxley

Abstract. We give approximate counting formulae for the numbers of labelled general, tree-child, and normal (binary) phylogenetic networks on n vertices. These formulae are of the form $2^{\gamma n \log n + O(n)}$, where the constant γ is $\frac{3}{2}$ for general networks, and $\frac{5}{4}$ for tree-child and normal networks. We also show that the number of leaf-labelled tree-child and normal networks with ℓ leaves are both $2^{2\ell \log \ell + O(\ell)}$. Further we determine the typical numbers of leaves, tree vertices, and reticulation vertices for each of these classes of networks.

1. Introduction

Ever since Darwin's publication of the *Origin of Species* in 1859, phylogenetic (evolutionary) trees have been used to represent the ancestral history of a collection of present-day species. However, it is now well-known that the ancestral history for certain collections of species is more realistically represented by a phylogenetic network rather than a phylogenetic tree because of evolutionary processes such as recombination and hybridisation. Mathematically, phylogenetic networks provide a much more significant challenge and, indeed, relatively little is known about these objects. For example, the number of leaf-labelled binary phylogenetic trees with ℓ leaves has been known since Schröder's work in 1870, and this also gives the number of such trees on n labelled vertices—see (1) and (3) below. In contrast, the number of binary phylogenetic networks on n labelled vertices is unknown, and similarly for subclasses like tree-child networks.

Date: 2 April 2014.

Key words and phrases. Phylogenetic networks, tree-child networks, normal networks.

The second author was supported by a Canterbury Fellowship at the University of Oxford and the New Zealand Marsden Fund.

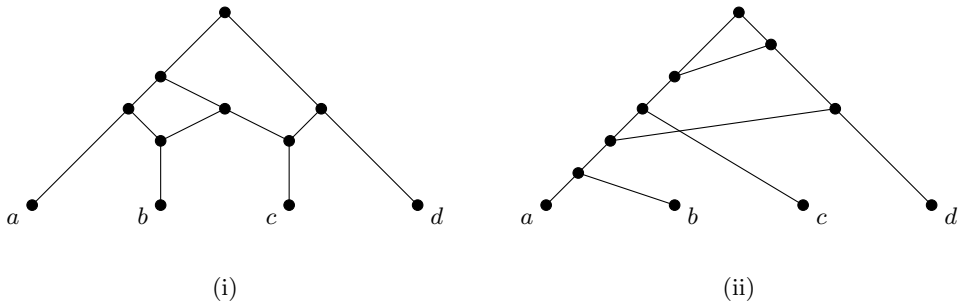

FIGURE 1. Two phylogenetic networks, where (i) is a general network and (ii) is a tree-child network. Edges are directed down the page.

The purpose of this paper is to investigate some of the combinatorial properties of phylogenetic networks. In particular, we provide some answers to the problems of counting the numbers of phylogenetic networks and of determining the typical proportions of vertices of different kinds. The rest of the introduction contains some necessary preliminaries and the statements of the main results.

For a finite set X, a *phylogenetic network on* X is a rooted acyclic directed graph with the following properties:

(i) the (unique) root is a vertex with in-degree 0 and out-degree two;
(ii) a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is X; and
(iii) all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one.

We do not allow parallel edges. However, for technical reasons, we do allow a single-root vertex to be a phylogenetic network. From now on, a *network* will always mean a phylogenetic network.

For a network $\mathcal{N}$ on X, we refer to the vertices of out-degree zero as the *leaves* of $\mathcal{N}$ and the set X as the *leaf-label set* of $\mathcal{N}$. As an example, two phylogenetic networks are shown in Figure 1. For both networks, the leaf-label set is $\{a, b, c, d\}$. Since there are no directed cycles, there is always a directed path from the root to any vertex and from any vertex to some leaf. Vertices with in-degree one and out-degree two are *tree vertices*, while vertices with in-degree two and out-degree one are *reticulation vertices*. Biologically, the leaves represent present-day species,

while all other vertices represent (hypothetical) ancestral species. A reticulation vertex represents, for example, a hybrid species.

A directed edge uv is a *reticulation edge* if v is a reticulation vertex; otherwise uv is a *tree edge*. Strictly speaking, $\mathcal{N}$ is a *binary* phylogenetic network as we are not allowing the out-degree of a tree vertex to be more than two or the in-degree of a reticulation vertex to be more than two. As a comparison, a *binary phylogenetic tree on* X is a network with no reticulation vertices.

A *leaf-labelled* phylogenetic network is a phylogenetic network in which the leaves are labelled but non-leaf vertices are unlabelled. In evolutionary biology, it is leaf-labelled phylogenetic networks that are typically of interest.

Counting networks. As an introduction to our new counting results, let us briefly discuss binary phylogenetic trees. For each $\ell \geq 2$, let $\widetilde{\mathcal{T}}_\ell$ be the set of leaf-labelled binary phylogenetic trees with leaf label set $[\ell] = \{1, 2, \ldots, \ell\}$; and, for each odd integer $n \geq 1$, let $\mathcal{T}_n$ be the set of binary phylogenetic trees on set of vertices $[n]$. Schröder [11] showed in 1870 that

$$\left|\widetilde{\mathcal{T}}_\ell\right| = 1 \times 3 \times 5 \times \cdots \times (2\ell - 3) = \frac{(2\ell-2)!}{(\ell-1)!\,2^{\ell-1}} \tag{1}$$

and so, by Stirling's approximation,

$$\left|\widetilde{\mathcal{T}}_\ell\right| \sim \frac{1}{\sqrt{2}} \left(\frac{2}{e}\right)^\ell \ell^{\ell-1} \tag{2}$$

as $\ell \to \infty$. A binary tree with ℓ leaves has $n = 2\ell - 1$ vertices in total. Thus, for an odd positive integer n, if we let $\ell = (n+1)/2$, then, by (1), we have

$$|\mathcal{T}_n| = \binom{n}{\ell} (\ell-1)! \cdot \left|\widetilde{\mathcal{T}}_\ell\right| = \binom{n}{\ell} (n-1)!\, 2^{1-\ell}. \tag{3}$$

(For the first equality, note that there are $\binom{n}{\ell}$ choices for the leaf-label set; then $\left|\widetilde{\mathcal{T}}_\ell\right|$ choices for the tree; and finally $(\ell-1)!$ ways to label the non-leaf vertices, each giving a distinct labelled tree.) From (3), it follows by Stirling's approximation that

$$|\mathcal{T}_n| \sim 2\sqrt{2} \left(\frac{\sqrt{2}}{e}\right)^n n^{n-1} \tag{4}$$

as $n \to \infty$.

Indeed, for phylogenetic trees we can use the methods of analytic combinatorics to obtain precise estimates of various quantities. For example, Bona and Flajolet [3] find the asymptotic probability that two random trees from $\tilde{\mathcal{T}}_\ell$ are isomorphic when we remove the labels on the leaves, and Gill [6] estimates parameters of X-trees and X-forests (which are related to phylogenetic trees with leaf-set X).

However, in this paper we move in a different direction. By (2) and (4),

$$\left|\widetilde{\mathcal{T}}_\ell\right| = 2^{\ell \log \ell + O(\ell)}$$

and

$$|\mathcal{T}_n| = 2^{n \log n + O(n)}$$

(all our logarithms are to the base 2). These are much cruder versions of (2) and (4), but they still show the main terms. Our counting results for phylogenetic networks are at this level of precision.

Now for our new results. Every network has an odd number of vertices, as we shall see shortly. For all odd integers $n \geq 3$, let $\mathcal{GN}_n$ denote the set of all (labelled) networks with vertex set $[n]$. Here the $\mathcal{G}$ is for **g**eneral and the $\mathcal{N}$ is for **n**etwork. The first main result of the paper is the following.

Theorem 1.1. *There exist positive constants c_1 and c_2 such that, for all odd integers $n \geq 3$,*

$$(c_1 n)^{\frac{3}{2}n} \leq |\mathcal{GN}_n| \leq (c_2 n)^{\frac{3}{2}n}.$$

An essentially equivalent result is that

$$|\mathcal{GN}_n| = 2^{\frac{3}{2}n \log n + O(n)}$$

for odd $n \geq 3$.

Tree-child and normal networks are two classes of networks that are prominent in the literature. One reason for their introduction [4, 14] and subsequent prominence is that they provide enough additional structure so that if a network $\mathcal{N}$ is in one of these two classes, then $\mathcal{N}$ is determined by certain local bits of structural information. Moreover, biologically, such networks guarantee that all species arising from a speciation event (represented by a tree vertex) or a reticulation event exist for a certain period of time before (possibly) going extinct. This is well-motivated as the process of extinction takes at least several

generations. The second main result of this paper is the analogue of Theorem 1.1 for tree-child and normal networks.

A network $\mathcal{N}$ is *tree-child* if, for each non-leaf vertex v of $\mathcal{N}$, at least one of its children is a tree vertex or a leaf. Equivalently, $\mathcal{N}$ is tree-child if, for each vertex v of $\mathcal{N}$, there is a directed path from v to a leaf in which each edge is a tree edge. A network $\mathcal{N}$ is *normal* if it is tree-child and has the additional property that if there is a directed path from vertex u to vertex v with at least two edges, then there is no directed edge uv. To illustrate, the network in Figure 1(ii) is tree-child, but it is not normal. The network in Figure 1(i) is not tree-child. For odd integers $n \geq 3$, let $\mathcal{TC}_n$ and $\mathcal{NL}_n$ denote, respectively, the sets of **t**ree-**c**hild and **n**orma**l** networks with vertex set $[n]$.

Theorem 1.2. *There exist positive constants c_1 and c_2 such that, for all odd integers $n \geq 3$,*

$$(c_1 n)^{\frac{5}{4}n} \leq |\mathcal{NL}_n| \leq |\mathcal{TC}_n| \leq (c_2 n)^{\frac{5}{4}n}.$$

For each integer $\ell \geq 2$, let $\widetilde{\mathcal{TC}}_\ell$ and $\widetilde{\mathcal{NL}}_\ell$ denote, respectively, the set of leaf-labelled tree-child and normal networks with leaf label set $[\ell]$.

Theorem 1.3. *There exist positive constants c_1 and c_2 such that, for all integers $\ell \geq 2$,*

$$(c_1 \ell)^{2\ell} \leq |\widetilde{\mathcal{NL}}_\ell| \leq |\widetilde{\mathcal{TC}}_\ell| \leq (c_2 \ell)^{2\ell}.$$

With regards to Theorem 1.3, Bickner [1, Corollary 2] established a non-asymptotic upper bound for $|\widetilde{\mathcal{NL}}_\ell|$. Writing this upper bound at the level of precision of the results in this paper, Bickner showed that there is a constant c such that $|\widetilde{\mathcal{NL}}_\ell| \leq (c\ell)^{4\ell}$.

It would be natural to expect that there are many more tree-child networks than normal networks, and indeed this is the case as stated in the next theorem.

Theorem 1.4.

(i) *As $n \to \infty$,*
$$|\mathcal{NL}_n|/|\mathcal{TC}_n| \to 0$$
and,
(ii) *as $\ell \to \infty$,*
$$|\widetilde{\mathcal{NL}}_\ell|/|\widetilde{\mathcal{TC}}_\ell| \to 0.$$

Typical parameters of networks. Given a network $\mathcal{N}$, let $n(\mathcal{N})$ be the number of vertices, $\ell(\mathcal{N})$ the number of leaves, $r(\mathcal{N})$ the number of reticulation vertices, and $t(\mathcal{N})$ the number of tree vertices. In the next section, we shall see that always

$$\ell + r = t + 2 = (n+1)/2. \tag{5}$$

Thus, in particular, except for t and n, any two of ℓ, r, t, and n determine the other parameters and, for large n, both $\ell + r$ and t are about $n/2$.

What values do these parameters usually take? Let us say that *almost all* networks in $\mathcal{GN}_n$ have some property if the proportion of networks in $\mathcal{GN}_n$ which have the property tends to 1 as $n \to \infty$; and similarly for the other classes of networks.

Theorem 1.5.

(i) *Almost all networks in $\mathcal{GN}_n$ have $o(n)$ leaves and $(\frac{1}{2} + o(1))n$ reticulation vertices.*
(ii) *Almost all networks in $\mathcal{TC}_n$ and almost all networks in $\mathcal{NL}_n$ have $(\frac{1}{4} + o(1))n$ leaves and $(\frac{1}{4} + o(1))n$ reticulation vertices.*
(iii) *Almost all leaf-labelled networks in $\widetilde{\mathcal{TC}}_\ell$ and almost all leaf-labelled networks in $\widetilde{\mathcal{NL}}_\ell$ have $(1 + o(1))\ell$ reticulation vertices and $(4 + o(1))\ell$ vertices in total.*

A *cherry* in a network consists of two sibling leaves and their common tree-vertex parent. Are cherries common? We shall see below that almost all networks in $\mathcal{TC}_n$ have $o(n)$ cherries and, similarly for other kinds of networks.

Let us extend this discussion. Let v be a vertex in a network $\mathcal{N}$. If v and each of its descendants (if any) are tree vertices or leaves then they form a tree network T with root v, which we call a *pendant subtree* of $\mathcal{N}$. A non-leaf vertex which is in some pendant subtree is called a *twig* of $\mathcal{N}$. Each cherry contains one twig, so the number of cherries in $\mathcal{N}$ is at most the number of twigs.

Since a binary tree has one more leaf than non-leaf, every network has fewer twigs than leaves. Thus almost all n-vertex general networks have $o(n)$ twigs, by Theorem 1.5 (i). But there are many leaves in tree-child and normal networks, so it is not immediate whether twigs are common in these types of networks.

Proposition 1.6.

(i) *Almost all networks in $\mathcal{TC}_n$ and almost all networks in $\mathcal{NL}_n$ have $o(n)$ twigs, and*
(ii) *almost all leaf-labelled networks in $\widetilde{\mathcal{TC}}_\ell$ and almost all leaf-labelled networks in $\widetilde{\mathcal{NL}}_\ell$ have $o(\ell)$ twigs.*

Thus, for example, almost all n-vertex tree-child networks contain about $n/4$ leaves but only $o(n)$ twigs.

Plan of the proofs. The next section contains some elementary lemmas which we shall need relating the numbers of vertices, leaves, reticulation vertices, tree-vertices, and twigs in a network. Proposition 1.6 will follow directly from Lemma 2.4, and parts (ii) and (iii) of Theorem 1.5.

In Section 3 on general networks, we first prove an upper bound, Lemma 3.1, on the number of networks in $\mathcal{GN}_n$ with a given number of leaves. Then we give a lower bound on $|\mathcal{GN}_n|$ based on the fact that almost all cubic graphs have a Hamilton circuit. Together these bounds establish Theorem 1.1 on the number of general networks, and yield part (i) of Theorem 1.5 on typical behaviour.

In Section 4 on tree-child and normal networks, we first note an upper bound from Lemma 3.1 on numbers of tree-child networks. The bulk of the section is devoted to a construction yielding many normal networks, and thus completing the proof of Theorem 1.2. These results also yield part (ii) of Theorem 1.5.

Leaf-labelled networks are considered in Section 5, where we prove Theorem 1.3 and part (iii) of Theorem 1.5. In Section 6, we prove Theorem 1.4. Then, in the final section, we make some concluding remarks and mention some natural open questions.

2. Parameters of Networks

This section consists of several elementary results. The first result establishes (5) and shows that a network always has an odd number of vertices.

Lemma 2.1. *Let $\mathcal{N}$ be a network on n vertices with ℓ leaves, r reticulation vertices, and t tree vertices. Then $t = \ell + r - 2$ and $n = 2t + 3$. Also, $\mathcal{N}$ has $3r + 2\ell - 2$ edges.*

Proof. Note first that

$$n = r + \ell + t + 1.$$

Since the sum of the out-degrees equals the number e of edges which, in turn, equals the sum of the in-degrees, we have

$$r + 2t + 2 = e = 2r + t + \ell.$$

Hence $t = r + \ell - 2$, and now the lemma follows easily. □

With one exception, the equations in (5), established in Lemma 2.1, characterise the possible parameters of a network; that is, if there are integers $r \geq 0$ and $\ell \geq 1$ with $r + \ell \geq 2$, then there is a network with these parameters.

To see this, first observe that, for $r = \ell = 1$, in which case, $t = 0$ and $n = 3$, there is no network with these parameters as we do not allow parallel edges. However, for each ordered pair (r, ℓ) of integers with $r \geq 0$ and $\ell \geq 1$ other than $(1, 1)$, there is a corresponding network on $[n]$ satisfying (5).

For, it is easily seen that there are networks for $(0, 2)$, $(1, 2)$ and $(2, 1)$. Also, for all $r \geq 0$ and $\ell \geq 1$, if there is a network for (r, ℓ), then there is one for $(r, \ell + 1)$. We can see this by using a new vertex v to subdivide an edge incident with a leaf and making v incident to a new leaf. Further, for all $r \geq 0$, if there is a network for $(r, 1)$, then there is one for $(r + 1, 1)$. We can see this by adjoining a new root via two new edges, one to the original root and the other to the leaf v, and making v adjacent to a new leaf.

The two parts of the next lemma are established in [4] and [1], respectively. However, we include its short proof for completeness.

Lemma 2.2. *Let $\mathcal{N}$ be a network with ℓ leaves and r reticulation vertices. If $\mathcal{N}$ is tree-child, then $r \leq \ell - 1$. Furthermore, if $\mathcal{N}$ is normal, then $r \leq \ell - 2$.*

Proof. Let A be the set of reticulation vertices of $\mathcal{N}$ together with the root ρ. Suppose that $\mathcal{N}$ is tree-child. Then, for each vertex $v \in A$, there is a directed path P_v from v to a leaf which contains no reticulation

edge. Since these paths must be vertex disjoint, $r + 1 \leq \ell$. If, in addition, $\mathcal{N}$ is normal, then it is easily seen that both children of ρ are tree vertices. Replacing ρ in A by these two children and applying the same argument, we deduce that $r + 2 \leq \ell$. □

Note that the bounds in Lemma 2.2 are best possible [4, 1].

Lemma 2.2 and (5) characterise the possible parameters of a tree-child and normal network; that is, given integers $r \geq 0$ and $\ell \geq 1$ with $r+\ell \geq 2$, there is a tree-child (respectively, normal) network with these parameters provided $r \leq \ell - 1$ (respectively, $r \leq \ell - 2$). We show this for normal networks and omit the similar, but simpler, argument for tree-child networks.

Clearly, there is a normal network for $(0, 2)$. Further, using the same construction as that for general networks it follows that, for all $r \geq 0$ and $\ell \geq 1$, if there is a normal network for (r, ℓ), then there is one for $(r, \ell + 1)$. Now suppose that there is a normal network $\mathcal{N}$ for (r, ℓ), where $r \leq \ell - 3$. We next show that there is a normal network for $(r + 1, \ell)$.

As in the proof of Lemma 2.2, for each vertex $v \in A$, there is a path P_v from v to a leaf. Since $|A| \leq \ell - 1$, there is a vertex u in A with the property that there are two paths P_1 and P_2 starting at u, ending at distinct leaves, and containing no reticulation edge. Let w be the last vertex in $\mathcal{N}$ common to P_1 and P_2, and let w' be a child of w. Let $v' \in A - \{u\}$ and let l be the leaf at the end of $P_{v'}$. Now, add two new vertices to $\mathcal{N}$, the first subdividing the edge incident with l and the second subdividing the edge ww', and add an edge directed from the first new vertex to the second new vertex. It is easily checked that the resulting network is normal.

Combining (5) and Lemma 2.2, we get the following corollary.

Corollary 2.3. *If a tree-child network has n vertices, ℓ leaves, and r reticulation vertices, then*

$$r < \frac{n}{4} < \ell.$$

Proof. By (5) and Lemma 2.2,

$$4r < 2(r + (r + 1)) - 1 \leq n = 2(r + \ell) - 1 \leq 2((\ell - 1) + \ell) - 1 < 4\ell.$$

□

Now we consider the case in a tree-child network when the number of leaves is not much more than $n/4$.

Lemma 2.4. *Let $\mathcal{N}$ be a tree-child network with n vertices, ℓ leaves, and r reticulation vertices, and suppose that $\ell \leq n/4 + x$. Then $r > n/4 - x$ and $\ell - r < 2x$. Further, $\mathcal{N}$ has less than $2x$ twigs.*

Proof. By (5), $\ell + r = (n+1)/2$, so $r \geq n/4 + 1/2 - x$ and $\ell - r \leq 2x - 1/2$.

For the second part of the lemma, let $\mathcal{T}$ be the set of maximal pendant subtrees in $\mathcal{N}$. For each $T \in \mathcal{T}$, denote the number of leaves in T by $\ell(T)$, so the number of non-leaves is $\ell(T) - 1$, and the number of twigs in $\mathcal{N}$ is $\sum_{T \in \mathcal{T}} (\ell(T) - 1)$. The paths P_v in the proof of Lemma 2.2 must end at leaves in distinct maximal pendant subtrees. But each leaf of $\mathcal{N}$ is in exactly one tree in $\mathcal{T}$, so

$$r + 1 \leq |\mathcal{T}| = \sum_{T \in \mathcal{T}} (\ell(T) - (\ell(T) - 1)) = \ell - \sum_{T \in \mathcal{T}} (\ell(T) - 1).$$

Thus the number of twigs in $\mathcal{N}$ is

$$\sum_{T \in \mathcal{T}} (\ell(T) - 1) \leq \ell - r - 1 < 2x.$$

□

Note that Proposition 1.6 follows directly from the last lemma, and parts (ii) and (iii) of Theorem 1.5.

3. Proofs for General Networks

Let $g(n, \ell)$ be the number of networks in $\mathcal{GN}_n$ which have ℓ leaves. Thus $|\mathcal{GN}_n| = \sum_\ell g(n, \ell)$. In this section, we first prove an upper bound on $g(n, \ell)$, and then we give a lower bound on $|\mathcal{GN}_n|$ based on the fact that almost all cubic graphs have a Hamilton circuit. Together these bounds establish Theorem 1.1 on the number of general networks. Further, from these results we quickly prove part (i) of Theorem 1.5.

Lemma 3.1. *There exists a positive constant c such that, for all integers $\ell \geq 1$ and all odd integers $n \geq 3$,*

$$g(n, \ell) \leq c^n n^{\frac{3}{2}n - \ell}$$

and

$$|\mathcal{GN}_n| \leq c^n n^{\frac{3}{2}n}.$$

Proof. Let $f(n,\ell)$ be the number of (simple, undirected) graphs on vertex set $[n]$ with ℓ vertices of degree one, 1 vertex of degree two, and the remaining vertices of degree three. Note that each such graph has degree sum

$$\ell + 2 + 3(n - \ell - 1) = 3n - 2\ell - 1.$$

We use a configuration model, see for example [2, Section 2.1] or [7, Section 9.1]. Consider such a model with $3n - 2\ell - 1$ labelled points partitioned into $\ell + 1 + (n - \ell - 1)$ parts with ℓ parts containing a single point, 1 part containing two points, and each of the remaining parts containing three points. The number of perfect matchings is $(3n - 2\ell - 2)!! \leq (3n)^{\frac{3}{2}n-\ell}$. Since there are n choices for the single vertex of degree two and $\binom{n}{\ell}$ choices for the ℓ vertices of degree one, it follows that

$$f(n,\ell) \leq n \cdot \binom{n}{\ell} \cdot (3n)^{\frac{3}{2}n-\ell}.$$

Since $\binom{n}{\ell} \leq 2^n$ and there are at most 2^{3n} choices of orientation for the edges,

$$\begin{aligned} g(n,\ell) &\leq 2^{3n} \cdot n \cdot 2^n \cdot (3n)^{\frac{3}{2}n-\ell} \\ &\leq c^n n^{\frac{3}{2}n-\ell} \end{aligned}$$

for a suitable constant c. The second part follows by summing over $\ell \geq 1$. □

Now let us establish the lower bound part of Theorem 1.1. A graph G is *cubic* if each vertex of G has degree three. Note that if a graph is cubic, then it has an even number of vertices. Robinson and Wormald [10] showed in 1992 that almost all labelled cubic graphs are Hamiltonian; that is, for even n, the proportion of cubic graphs on vertex set $[n]$ which are Hamiltonian tends to 1 as $n \to \infty$. In fact, they showed [9] in 1984, that, for sufficiently large even n, at least 98% of cubic graphs on n vertices are Hamiltonian, and that is sufficient for our needs. We make use of this fact in the proof of the next result, which establishes the lower bound part of Theorem 1.1.

Lemma 3.2. *There exists a constant $c > 0$ such that, for all odd positive integers n,*

$$|\mathcal{GN}_n| \geq (cn)^{\frac{3}{2}n}.$$

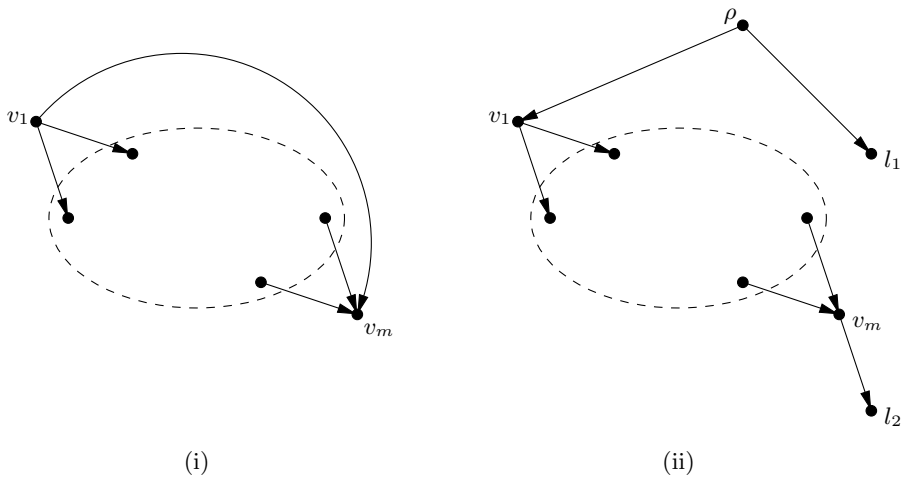

FIGURE 2. (i) Cubic graph G under orientation. (ii) Network $\mathcal{N}$ resulting from G and its orientation.

Proof. Let $m \geq 4$ be an integer and let G be a cubic graph on $[m]$. Suppose that G has a Hamiltonian cycle $C = v_1v_2 \cdots v_mv_1$. Orient G by directing each edge $\{v_i, v_j\}$ from v_i to v_j if $i < j$. Except for v_1 and v_m, each vertex of G under this orientation has either in-degree one and out-degree two or in-degree two and out-degree one. Now let $\mathcal{N}$ be obtained by deleting the edge v_1v_m, and adding new vertices ρ, l_1, l_2 and new directed edges ρv_1, ρl_1, and $v_m l_2$. An illustration of this construction is shown in Figure 2. It is easily checked that $\mathcal{N}$ is a network with root ρ. Furthermore, under this construction, each labelled cubic graph on $[m]$ with a Hamiltonian cycle yields a distinct network on $[m] \cup \{\rho, l_1, l_2\}$. There is a constant $d > 0$ such that, for all sufficiently large even m, the number of cubic graphs on $[m]$ is at least $(dm)^{\frac{3}{2}m}$ (see, for example, [2, Corollary 2.17] or [7, Corollary 9.8]). By the comments above on the proportion of cubic graphs which are Hamiltonian it now follows using the above construction, that there is a positive constant c such that, for all sufficiently large odd integers n,

$$|\mathcal{GN}_n| \geq (cn)^{\frac{3}{2}n}.$$

Finally, since $|\mathcal{GN}_n| \geq 1$ for each odd positive integer n, we may drop the qualification that n be sufficiently large. This completes the proof of the lemma. □

The above two lemmas immediately give Theorem 1.1. Part (i) of Theorem 1.5 follows from (5) and the next lemma.

Lemma 3.3. *There is a constant $C > 0$ such that, for almost all networks $\mathcal{N}$ in $\mathcal{GN}_n$,*

$$\ell(\mathcal{N}) \leq C\frac{n}{\log n}.$$

Proof. By Lemma 3.1, the number of networks in $\mathcal{GN}_n$ with at least $Cn/\log n$ leaves is at most

$$\sum_{\ell \geq Cn/\log n} g(n,\ell) \leq nc^n n^{\frac{3}{2}n - Cn/\log n} = n(c2^{-C})^n n^{\frac{3}{2}n}.$$

By Lemma 3.2, this upper bound is much smaller than $|\mathcal{GN}_n|$ if C is sufficiently large. □

4. Proofs for Tree-Child and Normal Networks

In this section, we shall prove Theorem 1.2 and part (ii) of Theorem 1.5. We start with an upper bound on the number of tree-child networks, which is an easy consequence of Lemma 3.1.

Lemma 4.1. *Let $c > 0$ be the constant from Lemma 3.1. Then, for all odd integers n,*

$$|\mathcal{TC}_n| \leq c^n n^{\frac{5}{4}n}.$$

Also, for any $C > 0$, the number of tree-child networks on $[n]$ with at least $n/4 + Cn/\log n$ leaves is at most $n(c2^{-C})^n n^{\frac{5}{4}n}$. deleted 'fixed'

Proof. By Corollary 2.3, a tree-child network with n vertices has $\ell > n/4$ leaves. Thus, by Lemma 3.1, $|\mathcal{TC}_n| \leq c^n n^{\frac{5}{4}n}$, where c is the constant in Lemma 3.1. Further, by the same lemma, the number of tree-child networks on $[n]$ with at least $n/4 + Cn/\log n$ leaves is at most

$$nc^n n^{\frac{3}{2}n - \frac{1}{4}n - Cn/\log n} = n(c2^{-C})^n n^{\frac{5}{4}n}.$$

□

The next lemma gives the main step in the constructions we shall use to establish a lower bound on the number of normal networks.

Lemma 4.2. *For each integer $k \geq 3$, let $\ell_k = 2^{k-2}$ and $n_k = 2^k - 2k + 1$. There is a constant $c > 0$ such that, for each k, there are at least $(c\ell_k)^{5\ell_k}$ normal networks on vertex set $[n_k]$ with ℓ_k leaves.*

Proof. Fix an integer $k \geq 3$ and write ℓ for ℓ_k and n for n_k. Consider the complete binary tree of depth $k-2$, not embedded in the plane. It has $t = 2^{k-1} - 1 \geq \frac{1}{2}n$ vertices, of which $s = 2^{k-2} - 1$ are non-leaves. We claim that, for a suitable constant $d_1 > 0$, there are at least $(d_1 n)^{\frac{1}{2}n}$ distinct such labelled trees with vertex labels from $[t]$.

To see this, consider the set Ω of rooted complete binary trees of depth $k-2$ on vertex set $[t]$, and the set Ω' of such trees which are embedded in the plane (with the root at the top). Clearly, $|\Omega'| = t!$. Each tree T in Ω has exactly 2^s embeddings in the plane, since at each non-leaf we choose one edge to a child to be the left edge, and so the other is the right edge. Hence $|\Omega| = t!/2^s$. Now, since $s \leq t$ and $t! \geq (t/e)^t$, we have

$$|\Omega| = \frac{t!}{2^s} \geq \left(\frac{t}{2e}\right)^t \geq \left(\frac{n}{4e}\right)^{\frac{1}{2}n}.$$

Thus we may take $d_1 = \frac{1}{4e}$.

Fix one of these labelled trees and call it T. The *depth* of a vertex v in T is the number of edges on the path from the root to v. For all $j \in \{0, 1, \ldots, k-2\}$, let V_j be the set of 2^j vertices at depth j and, for all $j \in \{1, 2, \ldots, k-2\}$, let E_j be the set of 2^j edges joining vertices at depth $j-1$ with vertices at depth j. Within each layer E_j, we shall carefully add $2^{j-1} - 1$ 'cross-edges' and their end vertices.

Let M be the set of non-leaf vertices of T. Observe that

$$|M| = \tfrac{1}{2}(t-1) = 2^{k-2} - 1 = \tfrac{1}{4}n + O(\log n) \approx \tfrac{1}{4}n.$$

Let π be an ordering on M with the property that if u has depth strictly less than that of v, then u comes before v in the ordering (that is, π is a linear extension of the depth partial order on M). We claim that, for a suitable constant $d_2 > 0$, there are at least $(d_2 n)^{\frac{1}{4}n}$ choices for π.

To see this, observe first that for each $j \in \{3, 4, \ldots, k-2\}$, the number of orderings of V_{k-j} is $(2^{k-j})! \geq (2^{k-j}/e)^{2^{k-j}}$. So, the log of the

number of choices for π is at least

$$\begin{aligned}\sum_{j=3}^{k-2} 2^{k-j} \log\left(\frac{2^k}{2^j e}\right) &= 2^k \sum_{j=3}^{k-2} 2^{-j} \left(\log(2^k) - \log(2^j e)\right) \\ &= 2^k \left(k \sum_{j=3}^{k-2} 2^{-j} + O(1)\right) \\ &= 2^k \left(\tfrac{1}{4}k + O(1)\right) \\ &= \tfrac{1}{4} n \log n + O(n).\end{aligned}$$

Thus there is a constant $d_2 > 0$ such that there are at least $(d_2 n)^{\frac{1}{4}n}$ choices for π.

Also, let λ be an ordering on the set $R = [n] \backslash [t] = \{t+1, t+2, \ldots, n\}$. Note that

$$|R| = n - t = 2^{k-1} - 2k + 2 = \tfrac{1}{2} n + O(\log n).$$

Thus, for n sufficiently large,

$$|R|! \geq (|R|/e)^{|R|} \geq (n/6)^{\frac{1}{2}n + O(\log n)} \geq (n/7)^{\frac{1}{2}n}.$$

Hence there is a constant $d_3 > 0$ such that there are at least $(d_3 n)^{\frac{1}{2}n}$ choices for λ.

For each choice of T, π, and λ, we shall construct a distinct normal network on vertex set $[n]$. Note that, in terms of n, the number of choices of T, π, and λ is at least

$$(d_1 n)^{\frac{1}{2}n} \cdot (d_2 n)^{\frac{1}{4}n} \cdot (d_3 n)^{\frac{1}{2}n} = (d_4 n)^{\frac{5}{4}n}$$

for an appropriate constant $d_4 > 0$. Now $n = 4\ell - O(k) = 4\ell - O(\log \ell)$, so, once $n \geq 3\ell$,

$$(d_4 n)^{\frac{5}{4}n} \geq (3 d_4 \ell)^{\frac{5}{4}n} = (3 d_4 \ell)^{5\ell - O(\log \ell)} \geq (d_4 \ell)^{5\ell}$$

for ℓ sufficiently large. Thus, in terms of ℓ, the number of choices of T, π, and λ is at least $(c\ell)^{5\ell}$ for a suitable constant $c > 0$.

For the construction, we work through T level by level starting at the root. For level E_1, we do nothing. Now suppose that, for some $1 < j \leq k-2$, we have reached the set E_j of edges joining vertices in V_{j-1} with vertices in V_j. Suppose that π orders the vertices in V_{j-1} as $w_1, w_2, \ldots, w_s$, where $s = 2^{j-1}$. For each $i = 1, 2, \ldots, s-1$ in turn, we add an edge (and its two end vertices) as follows:

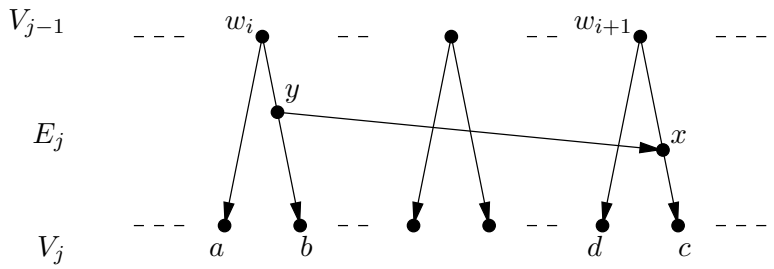

FIGURE 3. An illustration of the construction in the proof of Lemma 4.2.

(i) Suppose that the two edges from w_i to vertices in V_j are w_ia and w_ib, where the label of a is less than that of b, and the two edges from w_{i+1} to vertices in V_j are $w_{i+1}c$ and $w_{i+1}d$, where the label of c is less than that of d.
(ii) Let y and x be the next two unused vertices in R in the ordering λ. Subdivide w_ib with y and subdivide $w_{i+1}c$ with x, and then add an edge directed from y to x. Note that y is a tree vertex and x is a reticulation vertex.

This construction is shown in Figure 3.

We next observe some properties of the construction. Firstly, we do not create any cycles. To see this, we may think of the y vertices in (ii) being inserted high up (near a vertex in V_{j-1}) and the x vertices being inserted low down (near a vertex in V_j), and so all directed edges slope downwards. Secondly, for each vertex in V_{j-1}, at most one of the original edges directed out of it is subdivided with a resulting reticulation vertex. Thus this construction yields a tree-child network. In fact, it is easy to see that it is a normal network.

Now, the construction adds

$$\sum_{i=1}^{k-2}(2^{i-1}-1) = 2^{k-2}-1-(k-2) = 2^{k-2}-k+1$$

edges and, therefore, $2^{k-1}-2k+2$ vertices. Since $t = 2^{k-1}-1$, the resulting network has in total

$$(2^{k-1}-1)+(2^{k-1}-2k+2) = 2^k-2k+1 = n$$

vertices, including 2^{k-2} leaves.

We have given a lower bound on the number of constructions. To see that the networks constructed are all distinct, it remains to check that, from each network $\mathcal{N}$ constructed, we can recover the original labelled tree T, and the two linear orders π and λ.

Given the normal network $\mathcal{N}$, it is clear that we can 'see' (determine) T as its vertices are labelled with elements in $[t]$. Furthermore, for each $j \in \{2, 3, \ldots, k-2\}$, we can see the edges added at level E_j and therefore determine π on V_{j-1}, and thus determine π completely. But knowing π means that we also know λ. This completes the proof. □

From Lemma 4.2, we may deduce corresponding results for general numbers of leaves and general numbers of vertices in normal networks. We first consider leaves. Recall that if a normal network has n vertices and ℓ leaves, then $\ell > n/4$.

Lemma 4.3. *There is a constant $c > 0$ such that, for each integer $\ell \geq 2$, there exists $n = n(\ell)$ for which there are at least $(c\ell)^{5\ell}$ normal networks with ℓ leaves on vertex set $[n(\ell)]$.*

Proof. Let $\ell \geq 2$ be an integer. Assume for now that ℓ is even. By expressing ℓ in binary, we see that there exist $1 \leq m \leq \log_2 \ell$ and $k_1 > k_2 > \cdots > k_m \geq 1$ such that $\ell = \sum_{i=1}^{m} 2^{k_i}$. If ℓ is a power of 2, then we are done by the last lemma, so suppose that $m \geq 2$.

Now, take a binary tree with $2m-1$ vertices including m leaves, and identify the leaves with the roots of m normal networks $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_m$, where $\mathcal{N}_i$ has $\ell_i = 2^{k_i}$ leaves as in the last lemma, to obtain a normal network $\mathcal{N}$ with $\ell = \sum_i \ell_i$ leaves. Note that each $\mathcal{N}_i$ has $n_i = 4\ell_i - 2k_i - 3$ vertices in total. Thus $\mathcal{N}$ has

$$\sum_i n_i + m - 1 = 4\ell - \sum_i (2k_i + 3) + m - 1 = 4\ell - \sum_i (2k_i + 2) - 1$$

vertices in total, and we set $n(\ell)$ to be this number.

So far, we have assumed that ℓ is even. If ℓ is odd, then, as we can always replace a leaf by a tree vertex and two leaves, we may set $n(\ell) = n(\ell - 1) + 2$.

By Lemma 4.2, it suffices now to show that $\prod_i \ell_i^{5\ell_i} \geq (c\ell)^{5\ell}$. Let $k = k_1$. Then it suffices to show that

$$\prod_i \left(2^{k_i}\right)^{5\cdot 2^{k_i}} \geq 2^{5k\ell + O(\ell)},$$

that is,

$$\sum_i k_i 2^{k_i} \geq k\ell + O(\ell).$$

Let $I = \{k_1, k_2, \ldots, k_m\}$. Then

$$\sum_i k_i 2^{k_i} = \sum_{j=0}^{k} \mathbf{1}_{\{k-j \in I\}}(k-j)2^{k-j} \geq k\sum_i 2^{k_i} - 2^k \sum_{j=0}^{k} j2^{-j} = k\ell + O(\ell),$$

as required. □

From the last lemma, we now deduce a result focussing on the total number of vertices rather than on the number of leaves.

Lemma 4.4. *There is a constant $d > 0$ such that, for each odd integer $n \geq 3$, there are at least $(dn)^{\frac{5}{4}n}$ normal networks on vertex set $[n]$.*

Proof. Let $n > 8$ be an odd integer. Set $\ell = \left\lfloor \frac{n}{4} \right\rfloor$, and apply the last lemma. Thus there are at least $(c\ell)^{5\ell}$ normal networks with vertex set $[n(\ell)]$. But $n(\ell) \leq 4\ell \leq n$, and we can always add two vertices to a normal network (by replacing a leaf with a tree vertex and two leaf children), so there are at least $(c\ell)^{5\ell}$ normal networks with vertex set $[n]$. Further, since $\ell \geq (n-3)/4$,

$$(c\ell)^{5\ell} \geq ((c/4)(n-3))^{\frac{5}{4}(n-3)} \geq (dn)^{\frac{5}{4}n}$$

for a suitable constant $d > 0$. □

The last lemma together with Lemma 4.1 immediately yield Theorem 1.2. Consider part (ii) of Theorem 1.5. The second part of Lemma 4.1, together with Corollary 2.3 and (5), now yields the results on number $\ell(\mathcal{N})$ of leaves and $r(\mathcal{N})$ reticulation vertices in part (ii) of Theorem 1.5, in the following stronger form. There is a constant $C > 0$ such that, for almost all networks $\mathcal{N}$ in $\mathcal{TC}_n$ and for almost all networks $\mathcal{N}$ in $\mathcal{NL}_n$,

$$\tfrac{1}{4}n - Cn/\log n < r(\mathcal{N}) < \tfrac{1}{4}n < \ell(\mathcal{N}) < \tfrac{1}{4}n + Cn/\log n.$$

5. Leaf-Labelled Tree-Child and Normal Networks

In this section, we prove Theorem 1.3 and part (iii) of Theorem 1.5. First let us prove Theorem 1.3.

Consider a (labelled) network. For each vertex v, let $D(v)$ denote the set of strict descendants of v, that is, the set of vertices other than v that can be reached by a directed path from v. We will need the following observation, a consequence of results in [4].

Lemma 5.1. *In a tree-child network, let v and v' be distinct vertices. Then $D(v) \neq D(v')$.*

Proof. If $v \in D(v')$, then $v' \notin D(v)$ since the network is acyclic, so $D(v) \neq D(v')$. Suppose $v \notin D(v')$, and consider a leaf w which can be reached from v by a path of tree edges. Then we must have $w \in D(v) \setminus D(v')$, and so again $D(v) \neq D(v')$. □

Although we do not use the result, in a normal network, distinct vertices have distinct sets of leaf descendants [15].

Now consider a tree-child network, and suppose that there is a non-trivial automorphism f which fixes each leaf. Then there is a non-leaf vertex v which is moved under f and all of $D(v)$ is fixed. But, as f is an automorphism $D(f(v)) = \{f(w) : w \in D(v)\} = D(v)$ and so, by Lemma 5.1, $f(v) = v$; a contradiction. Thus a tree-child network has no non-trivial automorphism which fixes each leaf.

Let $\mathcal{TC}_{n,\ell}$ be the set of networks $\mathcal{N}$ in $\mathcal{TC}_n$ in which there are ℓ leaves and the leaf-label set is $[\ell]$. Call two networks $\mathcal{N}$ and $\mathcal{N}'$ in $\mathcal{TC}_{n,\ell}$ *equivalent* if there is an isomorphism ϕ from $\mathcal{N}$ to $\mathcal{N}'$ which fixes all leaves (that is, $\phi(i) = i$ for each $i \in [\ell]$). We may identify the equivalence classes with the leaf-labelled n-vertex networks with leaf label set $[\ell]$.

By the remarks above, each equivalence class contains exactly $(n-\ell)!$ networks. It follows that if $\mathcal{A}$ is a class of tree-child networks which is closed under automorphisms (for example, the class of all tree-child networks or the class of all normal networks), then the number of distinct leaf-labelled networks in $\mathcal{A}$ which have leaf label set $[\ell]$ and n vertices in total equals $1/(n-\ell)!$ times the number of networks in $\mathcal{A}$ with leaf label set $[\ell]$ and vertex set $[n]$.

With these preliminaries, we may now complete the proof of Theorem 1.3.

Proof of Theorem 1.3. Consider first the upper bound. Let $\widetilde{tc}(\ell, n)$ denote the number of leaf-labelled tree-child networks on leaf-label set $[\ell]$ with n vertices in total. Let $m = n - \ell$, the number of non-leaf vertices in such a network. Then, combining the last observation with Lemma 3.1 and the inequality $m! \geq (m/e)^m$, we have

$$\widetilde{tc}(\ell, n) \;\leq\; \frac{c^n n^{\frac{3}{2}n-\ell}}{m!} \;\leq\; c^n n^{\frac{1}{2}n} n^m \left(\frac{e}{m}\right)^m = c^n n^{\frac{1}{2}n} \left(\frac{en}{m}\right)^m,$$

where c is the constant in Lemma 3.1. But if $a > 0$ and $f(x) = \left(\frac{ea}{x}\right)^x$ for $x > 0$, then $f(x) \leq e^a$. Thus $\left(\frac{en}{m}\right)^m \leq e^n$. Also, by Corollary 2.3, we have $n < 4\ell$. Therefore

$$\widetilde{tc}(\ell, n) \leq (ce)^n n^{\frac{1}{2}n} \leq (ce)^{4\ell} (4\ell)^{2\ell}$$

assuming, as we may, that $ec \geq 1$. Thus

$$|\widetilde{\mathcal{TC}}_\ell| = \sum_{n<4\ell} \widetilde{tc}(\ell, n) \leq 4\ell \cdot (ce)^{4\ell} (4\ell)^{2\ell} \leq (c_2\ell)^{2\ell}$$

for some constant $c_2 > 0$.

Now consider the lower bound. By Lemma 4.3, for each $\ell \geq 2$, there is an $n = n(\ell) < 4\ell$ such that the number of normal networks with ℓ leaves and with vertex set $[n(\ell)]$ is at least $(c\ell)^{5\ell}$ for some positive constant c. Hence the number of normal networks with leaf-label set $[\ell]$ and vertex set $[n(\ell)]$ is at least

$$\binom{n(\ell)}{\ell}^{-1} (c\ell)^{5\ell} \geq 2^{-4\ell} (c\ell)^{5\ell} \geq (c_3\ell)^{5\ell}$$

for a suitable constant $c_3 > 0$.

Each such network has at most 3ℓ non-leaves, and so the number of leaf-labelled normal networks with leaf label set $[\ell]$ is at least this quantity divided by $(3\ell)!$, and so is at least $(c_3\ell)^{5\ell}/(3\ell)^{3\ell} = (c_1\ell)^{2\ell}$ where the constant $c_1 > 0$. This completes the proof of the theorem. □

Part (iii) of Theorem 1.5 immediately follows from the next lemma.

Lemma 5.2. *There is a constant $C > 0$ such that, for almost all leaf-labelled networks $\mathcal{N}$ in $\widetilde{\mathcal{TC}}_\ell$ and almost all leaf-labelled networks $\mathcal{N}$ in $\widetilde{\mathcal{NL}}_\ell$,*

$$\ell - C\frac{\ell}{\log \ell} < r(\mathcal{N}) < \frac{1}{4}n(\mathcal{N}) < \ell.$$

Proof. By the inequality $\widetilde{tc}(\ell, n) \leq (ce)^n n^{\frac{1}{2}n}$ in the proof of Theorem 1.3 (with $ce \geq 1$), and recalling that $n < 4\ell$, we see that, for any $C > 0$, the number of leaf-labelled tree-child networks on leaf-label set $[\ell]$ with at most $4\ell - 4C\ell/\log \ell$ vertices in total is at most

$$(ce)^{4\ell}(4\ell)^{2\ell - 2C\ell/\log \ell} \leq (4(ce)^2 2^{-C}\ell)^{2\ell}.$$

Thus, if C is sufficiently large, then, by Theorem 1.3, almost all networks in $\widetilde{\mathcal{TC}}_\ell$ and almost all networks in $\widetilde{\mathcal{NL}}_\ell$ have at least $4\ell - 4C\ell/\log \ell$ vertices in total. This gives the lemma for the number of vertices. The number of reticulation vertices follows using (5). □

6. Almost All Tree-Child Networks are Not Normal

In this section, we prove Theorem 1.4. An edge uv in a tree-child network is a *shortcut* if there is a directed path from u to v with at least two edges. The idea of the proof for (labelled) networks is that from almost all networks $\mathcal{N}$ in $\mathcal{NL}_n$ we can construct many networks $\mathcal{N}'$ in $\mathcal{TC}_n$ with a unique shortcut edge; and such networks $\mathcal{N}'$ cannot be constructed many times. Thus the set of distinct networks in $\mathcal{TC}_n$ constructed is much larger than $\mathcal{NL}_n$. We begin with a construction and subsequent lemma.

Let $\mathcal{N}$ be a normal network and let v be a reticulation vertex with parents x and y. Both x and y must be tree vertices; for neither can be a reticulation vertex since $\mathcal{N}$ is tree-child, and if say x were the root, then there would be a path from x to y and on to v, and the edge xv would contradict normality. Let z be the parent of y and let z' be the child of y other than v. Observe that v, x, y, z, z' are all distinct, and $x \notin D(y)$ (since yv is not a shortcut). Let ab be an edge on a path P from the root to x, such that b is a tree-vertex. Note we could have $b = x$. A *subtree, prune, and regraft operation* (SPR) on yv and ab is performed as follows: (prune) detach yv at y leaving the new edge zz', and (regraft) insert y in the middle of edge ab, so that ab is replaced by two new edges ay and yb.

Lemma 6.1. *Consider the normal network and SPR operation on yv and ab as described above. Let $\mathcal{N}'$ be the resulting directed graph. Then $\mathcal{N}'$ is a tree-child network with exactly one shortcut edge, namely yv.*

Proof. Observe that no multiple edges have been formed, since zz' is not an edge of $\mathcal{N}$; and each vertex degree is maintained. Observe also that in $\mathcal{N}'$ there is a path Q from y to v which starts with the edge yb, and then follows the path P to x and then to v. If $\mathcal{N}'$ has a cycle, then it must contain the edge yv, and by replacing yv by Q we see that there is a cycle in $\mathcal{N}$; a contradiction. Thus $\mathcal{N}'$ is acyclic.

Now let us check that $\mathcal{N}'$ is a tree-child network. When we detach yv at y, the vertex z still has a child z' which is either a tree vertex or a leaf. Further, when we insert y in ab, then a still has a tree-vertex child (namely y) and y now has a tree-vertex child (namely b). It follows that $\mathcal{N}'$ is a tree-child network.

The path Q shows that edge yv is a shortcut. Suppose that some edge pq other than yv is a shortcut. Note that none of zz', ap, and pb are shortcuts, since their terminal vertices z', p and b are not reticulation vertices. Therefore pq must be an edge in $\mathcal{N}$. Now there must be a path Q' from p to q in $\mathcal{N}'$ other than the edge pq. At least one of the two edges zz' and yv must be on Q', since otherwise pq would be a shortcut in $\mathcal{N}$.

Choose such a path Q' with as few as possible of the edges zz' and yv. Then yv is not on Q'; otherwise, we could replace yv by the path Q and contradict our choice of Q'. Further, zz' is not on Q'; otherwise, we could replace zz' by the two edges zy and yz', thus realising that pq is a shortcut in $\mathcal{N}$. Hence yv is the only shortcut in $\mathcal{N}'$ and we have established the lemma. □

Proof of Theorem 1.4(i). Let $\mathcal{A}_n$ be the set of networks in $\mathcal{NL}_n$ which have at least $\frac{1}{5}n+2n^{\frac{1}{2}}$ reticulation vertices. By Theorem 1.5(ii), almost all networks in $\mathcal{NL}_n$ are in $\mathcal{A}_n$. Let $\mathcal{B}_n$ be the set of networks in $\mathcal{TC}_n$ which contain exactly one shortcut edge.

Let $\mathcal{N}$ be a network in $\mathcal{A}_n$. There are at most $4n^{\frac{1}{2}}$ vertices at distance from the root less than $1+\frac{1}{2}\log n$ in $\mathcal{N}$. Further, as the parents of reticulation vertices are either tree vertices or the root, and these

parents are all distinct, at most a third of these vertices are reticulation vertices. So there are at most $\frac{4}{3}n^{\frac{1}{2}}$ reticulation vertices at distance from the root less than $1+\frac{1}{2}\log n$, and hence at least $\frac{1}{5}n$ reticulation vertices at distance from the root at least $1+\frac{1}{2}\log n$ in $\mathcal{N}$. Let v be such a reticulation vertex with parents x and y. Let P be a path from the root to x. There are at least $\frac{1}{2}\log n$ edges in P. Further, x is a tree vertex, and no two successive vertices on P are reticulation vertices. Thus there are at least $\frac{1}{4}\log n$ edges ab in P with b a tree vertex.

By Lemma 6.1, we can construct from $\mathcal{N}$, using an SPR operation, at least

$$\tfrac{1}{5}n \cdot 2 \cdot \tfrac{1}{4}\log n = \tfrac{1}{10}n\log n$$

distinct networks $\mathcal{N}'$; and each is in $\mathcal{B}_n$. Let $\mathcal{B}'_n$ be the resulting collection of networks and let $\mathcal{N}'$ be a network in $\mathcal{B}'_n$. Now, $\mathcal{N}'$ has a unique shortcut edge yv. To recreate the original network $\mathcal{N}$, we need to choose a (suitable) edge zz' to perform the reverse SPR operation. But the number of such edges is at most $\frac{3}{2}n$, so each such $\mathcal{N}'$ is constructed at most $\frac{3}{2}n$ times. It follows that $|\mathcal{A}_n| \cdot \frac{1}{10}n\log n \leq |\mathcal{B}_n| \cdot \frac{3}{2}n$. Now, once n is sufficiently large, $|\mathcal{A}_n| \geq \frac{1}{2}|\mathcal{NL}_n|$ and so

$$|\mathcal{NL}_n| \leq 2|\mathcal{A}_n| \leq (30/\log n) \cdot |\mathcal{B}_n| \leq (30/\log n) \cdot |\mathcal{TC}_n|,$$

completing the proof of Theorem 1.4(i). □

To prove part (ii) of Theorem 1.4, let $\mathcal{NL}_{n,\ell}$ denote the set of networks in $\mathcal{NL}_n$ with ℓ leaves and let $\widetilde{\mathcal{NL}}_{n,\ell}$ denote the set of normal networks with n vertices and with leaf-label set $[\ell]$. Similarly, let $\mathcal{TC}_{n,\ell}$ denote the set of networks in $\mathcal{TC}_n$ with ℓ leaves and let $\widetilde{\mathcal{TC}}_{n,\ell}$ denote the set of tree-child networks with n vertices and with leaf-label set $[\ell]$. As we noted before, a tree-child network has no non-trivial automorphism which fixes each leaf. Thus

$$|\widetilde{\mathcal{NL}}_{n,\ell}| = \frac{|\mathcal{NL}_{n,\ell}|}{\binom{n}{\ell}\,(n-\ell)!}$$

and

$$|\widetilde{\mathcal{TC}}_{n,\ell}| = \frac{|\mathcal{TC}_{n,\ell}|}{\binom{n}{\ell}\,(n-\ell)!},$$

so, assuming $\mathcal{TC}_{n,\ell}$ is non-empty,

$$\frac{|\widetilde{\mathcal{NL}}_{n,\ell}|}{|\widetilde{\mathcal{TC}}_{n,\ell}|} = \frac{|\mathcal{NL}_{n,\ell}|}{|\mathcal{TC}_{n,\ell}|}. \tag{6}$$

Lemma 6.2. *Let n and ℓ be positive integers satisfying $\ell \leq \frac{3}{10}n - 2n^{\frac{1}{2}}$. Then*

$$|\mathcal{NL}_{n,\ell}| \leq \frac{8}{\log n}\,|\mathcal{TC}_{n,\ell}|.$$

Proof. We may assume that $\mathcal{NL}_{n,\ell}$ is non-empty. Let $\mathcal{N} \in \mathcal{NL}_{n,\ell}$. Then, by (5), the number r of reticulation vertices in $\mathcal{N}$ is

$$r = \frac{n+1}{2} - \ell \geq \frac{1}{5}n + 2n^{\frac{1}{2}}.$$

Now, arguing as in the proof of part (i) of Theorem 1.4 we have

$$|\mathcal{NL}_{n,\ell}| \leq \frac{15}{2\log n} \cdot |\mathcal{TC}_{n,\ell}|.$$

□

Proof of Theorem 1.4(ii). Let n_0 be sufficiently large that $\frac{3}{10}n - 2n^{\frac{1}{2}} \geq \frac{2}{7}n$ for each $n \geq n_0$. By Theorem 1.5(iii), there is an $\ell_0 \geq n_0$ such that, for each $\ell \geq \ell_0$,

$$\left|\left\{\mathcal{N} \in \widetilde{\mathcal{NL}}_\ell : n(\mathcal{N}) \geq \tfrac{7}{2}\ell\right\}\right| \geq \tfrac{1}{2}\left|\widetilde{\mathcal{NL}}_\ell\right|.$$

Hence, by (6) and Lemma 6.2, for each $\ell \geq \ell_0$,

$$\begin{aligned}
|\widetilde{\mathcal{NL}}_\ell| &\leq 2\sum_{n \geq \frac{7}{2}\ell} |\widetilde{\mathcal{NL}}_{n,\ell}| \\
&= 2\sum_{n \geq \frac{7}{2}\ell} |\widetilde{\mathcal{TC}}_{n,\ell}| \cdot \frac{|\mathcal{NL}_{n,\ell}|}{|\mathcal{TC}_{n,\ell}|} \\
&\leq 2\sum_{n \geq \frac{7}{2}\ell} |\widetilde{\mathcal{TC}}_{n,\ell}| \cdot \frac{8}{\log n} \\
&\leq \frac{16}{\log n}|\widetilde{\mathcal{TC}}_\ell|.
\end{aligned}$$

This completes the proof of part (ii) of Theorem 1.4. □

7. Concluding Remarks

We have established a range of enumerative results on general, tree-child, and normal networks. These results raise further natural questions. Let us mention three topics: improved counting, numbers of leaves in general networks, and typical depth of networks.

It is natural to hope for a more refined version of Theorem 1.1, and in particular that there is a constant $c_1 > 0$ such that $|\mathcal{GN}_n| = (c_1 + o(1))^n n^{\frac{3}{2}n}$. Similarly, are there constants $c_2 > 0$ and $c_3 > 0$ such that $|\mathcal{TC}_n| = (c_2 + o(1))^n n^{\frac{5}{4}n}$ and $|\mathcal{NL}_n| = (c_3 + o(1))^n n^{\frac{5}{4}n}$? Is $c_2 = c_3$? There are corresponding questions for leaf-labelled networks.

Concerning leaves, we have seen that almost all networks in $\mathcal{GN}_n$ have at most $O(n/\log n)$ leaves. Is that about the right order? Perhaps there are far fewer leaves, and indeed perhaps it is even true that the expected number of leaves is bounded?

The *depth* of a network is the maximum length (number of edges) of a directed path, from the root to a leaf. Since our networks are binary, the depth of an n-vertex network must be at least $\log n - 1$. Our constructions suggest that typical tree-child and normal networks have small depth, and typical general networks have much greater depth. But how large are these depths?

In the introduction, we briefly met (binary) phylogenetic trees (networks with no reticulation vertices), and perhaps we should again be guided by them. Consider such a tree $\mathcal{T}_n$ with n vertices (n must be odd) each of which is labeled, sampled uniformly at random. By work of Flajolet and Odlyzko [5], $\mathcal{T}_n$ has expected depth $\sim \sqrt{4\pi n}$, and similarly a random leaf-labeled phylogenetic tree with ℓ leaves has expected depth $\sim \sqrt{8\pi\ell}$. (Also, the expected depth of a random vertex in $\mathcal{T}_n$ is $\sim \sqrt{\pi n}$ [13].) In contrast, a random acyclic digraph typically has depth $\Theta(n)$, McKay [8].

Do random tree-child and normal networks on n vertices have expected depth $O(\sqrt{n})$? Is the expected depth of a general network larger?

Acknowledgements. We would like to thank Brendan McKay, Nick Wormald, and the referee for helpful comments.

References

[1] D. R. Bickner, On normal networks, PhD thesis, Iowa State University, 2012.
[2] B. Bollobás, Random Graphs, Second Edition, Cambridge University Press, 2001.
[3] M. Bona, P. Flajolet, Isomorphism and symmetries in random phylogenetic trees, J. Appl. Prob. 46 (2009) 1005–1019.

[4] G. Cardona, F. Rossello, G. Valiente, Comparison of tree-child phylogenetic networks, IEEE/ACM Trans. Comput. Biol. Bioinform. 6 (2009) 552–569.
[5] P. Flajolet and A. Odlyzko, The average height of binary trees and other simple trees, J. Comput. System Sci. 25 (1982) 171–213.
[6] J. Gill, The k-assignment polytope, phylogenetic trees, and permutation patterns, PhD thesis, Linköping University, Sweden, 2013.
[7] S. Janson, T. Łuczak, A. Ruciński, Random Graphs, Wiley Interscience, 2000.
[8] B. McKay, The shape of a random acyclic digraph, Math. Proc. Camb. Phil. Soc. 106 (1989) 459–465.
[9] R. W. Robinson, N. C. Wormald, Existence of long cycles in random cubic graphs, in: D. M. Jackson, S. A. Vanstone (Eds.), Enumeration and Design, Academic Press, Toronto, 1984, pp. 251–270.
[10] R. W. Robinson, N. C. Wormald, Almost all cubic graphs are Hamiltonian, Random Structures Algorithms 3 (1992) 117–125.
[11] E. Schröder, Vier combinatorische probleme, Zeitschrift für Mathematik und Physik 15 (1870) 361–376.
[12] C. Semple, M. Steel, Phylogenetics, Oxford University Press, 2003.
[13] J. S. Vitter, P. Flajolet, Average-case analysis of algorithms and data structures, in: J. van Leeuwen (Ed.), Handbook of Theoretical Computer Science, Vol A: Algorithms and Complexity, MIT Press, Amsterdam, 1990, pp. 431–524.
[14] S. J. Willson, Unique determination of some homoplasies at hybridization events, Bull. Math. Biol. 69 (2007) 1709–1725.
[15] S. J. Willson, Properties of normal networks, Bull. Math. Biol. 72 (2010) 340–358.

Department of Statistics, University of Oxford, UK

E-mail address: cmcd@stats.ox.ac.uk

Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: charles.semple@canterbury.ac.nz

Mathematical Institute, University of Oxford, UK

E-mail address: Dominic.Welsh@maths.ox.ac.uk