SIZE OF A PHYLOGENETIC NETWORK

CHARLES SEMPLE

Abstract. We consider the problem of when the total number n of vertices in a phylogenetic network $\mathcal{N}$ is bounded by the number ℓ of leaves in $\mathcal{N}$. The main result of the paper says that, provided $\mathcal{N}$ avoids three certain substructures, then n is at most quadratic in ℓ. Furthermore, if any of these substructures is present in $\mathcal{N}$, then ℓ does not necessarily bound n.

1. Introduction

A particularly active area of current research in phylogenetics is the mathematical study of phylogenetic networks. These networks generalise phylogenetic (evolutionary) trees as they additionally allow for the representation of non-treelike evolutionary events. These events include hybridisation and recombination, and are collectively called reticulation events. Not surprisingly, phylogenetic networks bring many new complications. For example, it is well known that the total number of vertices in a phylogenetic tree is bounded by the size of its leaf set, but the analogous result for phylogenetic networks does not necessarily hold. For phylogenetic algorithms, the typical parameter of interest is the size of the leaf set, and so this implies that it is not always possible to write the running time of phylogenetic network algorithms in terms of this parameter. However, for algorithms restricted to certain subclasses of phylogenetic networks, it is possible to write the running times in this way as the total number of vertices of a phylogenetic network that is in one of these classes is (polynomially) bounded by the size of its leaf set. See, for example, [1, 2, 3, 4, 7].

Without a predetermined class of phylogenetic networks in mind, in this paper, we investigate the problem of when the total number n of vertices of a phylogenetic network $\mathcal{N}$ is bounded by the number ℓ of leaves in $\mathcal{N}$. The main result of the paper says that, provided $\mathcal{N}$ avoids three certain substructures, then n is at most quadratic in ℓ. Moreover, as well as showing that this bound is sharp, we show that if any one of these substructures is

Date: August 19, 2016.

Key words and phrases. Phylogenetic networks; size.

The author was supported by the New Zealand Marsden Fund.

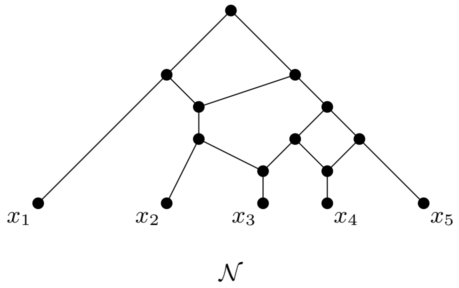

FIGURE 1. A phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, x_3, x_4, x_5\}$.

present in $\mathcal{N}$, then there is no guarantee that ℓ bounds n. The rest of the introduction formalises these results.

Throughout the paper, X denotes a nonempty finite set X, and notation and terminology follows Semple and Steel [6]. A *phylogenetic network* $\mathcal{N}$ *on* X is a rooted acyclic directed graph with no parallel edges and satisfying the following properties:

(i) the root has in-degree zero 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.

If $|X| = 1$, then, for technical reasons, we additionally allow for $\mathcal{N}$ to consist of the single vertex in X. The vertices in X are called *leaves* and X is referred to as the *leaf set* of $\mathcal{N}$. Furthermore, the vertices with in-degree one and out-degree two are called *tree vertices*, while the vertices with in-degree two and out-degree one are called *reticulations*. Thus the vertex set of $\mathcal{N}$ is partitioned into four types of vertices. Namely, the root, tree vertices, reticulations, and leaves. The edges directed into a reticulation are called *reticulation edges*. A *rooted binary phylogenetic* X*-tree* is a phylogenetic network on X with no reticulations. In the literature, phylogenetic networks, as defined here, are sometimes referred to as binary phylogenetic networks. To illustrate some of these concepts, a phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, x_3, x_4, x_5\}$ is shown in Figure 1. Here, $\mathcal{N}$ has exactly six tree vertices and three reticulations.

Let $\mathcal{N}$ be a phylogenetic network, and let u and v be distinct vertices of $\mathcal{N}$. If (u, v) is an edge of $\mathcal{N}$, then u is a *parent* of v or, equivalently, v is a child of u. More generally, if there is a directed path from u to v in $\mathcal{N}$, then u is an *ancestor* of v or, equivalently, v is a *descendant* of u.

We next describe the certain substructures of a phylogenetic network alluded to earlier in the section. Let $\mathcal{N}$ be a phylogenetic network. If (u, v) is an edge of $\mathcal{N}$ and both u and v are reticulations, we say (u, v) is a *parent-child reticulation*. This is the first of the three substructures. To describe the other two substructures, let

$$C = u_1\, v_1\, u_2\, v_2\, u_3\, \cdots u_k\, v_k\, u_{k+1}$$

be the vertices of an underlying path or cycle in $\mathcal{N}$. If u_{k+1} is a tree vertex and, for all $i \in \{1, 2, \ldots, k\}$, the vertex u_i is a tree vertex and v_i is a reticulation, we say C is a *reticulation chain*. For example, the parents of x_3 and x_4 in Figure 1 are the reticulations of a (maximal) reticulation chain in $\mathcal{N}$. Furthermore, C is *closed* if $u_1 = u_{k+1}$ and C is *overlapping* if, for some $i \neq j$, there are reticulations v_i and v_j such that v_i is an ancestor of v_j.

The main result of the paper is the following theorem.

Theorem 1.1. *Let $\mathcal{N}$ be a phylogenetic network on n vertices with ℓ leaves. Suppose that $\mathcal{N}$ has no parent-child reticulations, and no closed or overlapping reticulation chains. Then*

$$n \leq \ell^2 + 3\ell - 3.$$

Moreover, this bound is sharp.

The proof of Theorem 1.1 is given in the next section. Each of the restrictions on $\mathcal{N}$ in the statement of Theorem 1.1 are necessary for, as we show in the last section, Section 3, if $\mathcal{N}$ contains parent-child reticulations, closed reticulation chains, or overlapping reticulation chains, then ℓ does not necessarily bound n.

2. Proof of Theorem 1.1

In this section, we prove Theorem 1.1. We begin with two lemmas. The first lemma is established in [5].

Lemma 2.1. *Let $\mathcal{N}$ be a phylogenetic network on n vertices with ℓ leaves, r reticulations, and t tree vertices. Then*

$$\frac{n+1}{2} = \ell + r = t + 2.$$

Lemma 2.2. *Let $\mathcal{N}$ be a phylogenetic network with ℓ leaves, and let C be a reticulation chain in $\mathcal{N}$ that is not overlapping. If $\mathcal{N}$ has no parent-child reticulations, then $k \leq \ell$, where k is the number of reticulations in C.*

Proof. Let

$$C = u_1\, v_1\, u_2\, v_2\, u_3\, \cdots u_k\, v_k\, u_{k+1},$$

and suppose that $\mathcal{N}$ has no parent-child reticulations. Since $\ell \geq 1$, we may assume that $k \geq 2$. Let ρ denote the root of $\mathcal{N}$ and let X denote the leaf set of $\mathcal{N}$. Observing that, as C is not overlapping and so there is no directed path in $\mathcal{N}$ from a vertex in $\{u_2, u_3, \ldots, u_k\}$ to a vertex in $\{u_1, u_2, \ldots, u_{k+1}\}$, let $\mathcal{N}'$ be the phylogenetic network on X' obtained from $\mathcal{N}$ as follows:

(i) Delete every vertex that does not lie either on a path from ρ to a vertex in $\{u_2, u_3, \ldots, u_k\}$ or on a path from a vertex in $\{v_1, v_2, \ldots, v_k\}$ to a leaf.
(ii) Delete the reticulation edge (u_1, v_1) if it still remains. Similarly, delete (u_{k+1}, v_k). Denote the child vertices of v_1 and v_k as w_1 and w_k, respectively.
(iii) For each reticulation on a path from ρ to a vertex in $\{u_2, u_3, \ldots, u_k\}$ delete exactly one incident reticulation edge.
(iv) Contract any resulting non-root degree-two vertices.
(v) Lastly, if ρ has out-degree one, then contract the incident edge and relabel the identified vertex as ρ.

Since $\mathcal{N}$ has no parent-child reticulations, each of w_1 and w_k is either a tree vertex or a leaf and, by construction, $\mathcal{N}'$ has no parent-child reticulations. Furthermore, $X' \subseteq X$. Let ℓ' denote the number of leaves in $\mathcal{N}'$. Since $\ell' \leq \ell$, to complete the proof it suffices to show that $k \leq \ell'$.

Suppose, to the contrary, that $k \geq \ell' + 1$. Let n', r', and t' denote the total number of vertices, the number of reticulations, and the number of tree vertices in $\mathcal{N}'$, respectively. Using Lemma 2.2, we next count n' in two different ways. Since there are no reticulations in $\mathcal{N}'$ lying on a path from ρ to a vertex in $\{u_2, u_3, \ldots, u_k\}$, the number of tree vertices in the union of these paths is the same as the sum of the number of tree vertices and the number of leaves in a rooted binary phylogenetic tree with $k-1$ leaves. This sum is $2k-4$ as the root is not counted. Let s' denote the number of tree vertices on a path in $\mathcal{N}'$ from a vertex in $\{v_2, v_3, \ldots, v_{k-1}\} \cup \{w_1, w_k\}$ to a leaf. Then

$$t' = 2k - 4 + s',$$

and so, by Lemma 2.2,

$$\frac{n'+1}{2} = t' + 2 = s' + 2k - 2.$$

That is,

$$n' = 2s' + 4k - 5. \tag{1}$$

We next count n' in terms of ℓ' and r'. Since $\mathcal{N}'$ has no parent-child reticulations, the child of each reticulation in $\mathcal{N}'$ is either a tree vertex or a leaf. Furthermore, as each of w_1 and w_k is a tree vertex or a leaf, and its parent is not a reticulation, it follows that

$$r' \leq s' + \ell' - 2.$$

Therefore, by Lemma 2.2,

$$\frac{n'+1}{2} = \ell' + r' \leq s' + 2\ell' - 2,$$

that is,

$$n' \leq 2s' + 4\ell' - 5.$$

But, $\ell' \leq k-1$, so

$$n' \leq 2s' + 4k - 9,$$

contradicting (1). Hence $k \leq \ell'$ and so $k \leq \ell$, thereby completing the proof of the lemma. □

We now prove Theorem 1.1. A *cherry* in a phylogenetic network on X is a 2-element subset $\{a, b\}$ of X such that a and b have the same parent.

Proof of Theorem 1.1. Let n and ℓ denote the total number of vertices and the number of leaves in $\mathcal{N}$, respectively, and suppose that $\mathcal{N}$ has no parent-child reticulations, and no closed or overlapping reticulation chains. We first prove, by induction on ℓ, the inequality in the statement of the theorem. If $\ell = 1$, then $\mathcal{N}$ either consists of a single vertex, in which case the inequality holds, or the parent of the unique leaf is a reticulation, v say. Consider the latter. If both parents of v are tree vertices, then it is easily seen that $\mathcal{N}$ contains a directed cycle; a contradiction. Thus at least one parent of v is a reticulation, contradicting the assumption that $\mathcal{N}$ has no parent-child reticulations. It follows that the inequality holds when $\ell = 1$. Now suppose that $\ell \geq 2$, and that the inequality holds for all phylogenetic networks with at most $\ell - 1$ leaves, and having no parent-child reticulations, and no closed or overlapping reticulation chains.

First assume that $\mathcal{N}$ has a cherry $\{a, b\}$. Let $\mathcal{N}'$ be the phylogenetic network obtained from $\mathcal{N}$ by deleting b and contracting the resulting degree-two vertex. Note that if the parent of a and b is the root of $\mathcal{N}$, then $\mathcal{N}$ consists of three vertices, and we delete b and contract the edge incident with the root. By construction, it is clear that $\mathcal{N}'$ has no parent-child reticulation, and no closed or overlapping reticulation chains. Therefore, by induction,

$$n' \leq \left(\ell'\right)^2 + 3\ell' - 3,$$

where n' and ℓ' denote the total number of vertices and the number of leaves in $\mathcal{N}'$, respectively. But then, as $n' = n-2$ and $\ell' = \ell - 1$, we have

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

In other words,

$$n \leq \ell^2 + \ell - 3 \leq \ell^2 + 3\ell - 3,$$

and so the inequality holds.

Now assume that $\mathcal{N}$ has no cherries. In terms of the number of edges on a directed path, let w be a reticulation in $\mathcal{N}$ at maximum distance from the root. Let C be a maximal reticulation chain in $\mathcal{N}$ that contains w. Without loss of generality, we may assume that

$$C = u_1\, v_1\, u_2\, v_2\, u_3\, \cdots u_k\, v_k\, u_{k+1},$$

where, for some $i \in \{1, 2, \ldots, k\}$, we have $w = v_i$. By maximality and the assumption that $\mathcal{N}$ has no cherries, the only descendant of v_i in $\mathcal{N}$ is a leaf.

Let $\mathcal{N}'$ be the phylogenetic network obtained from $\mathcal{N}$ as follows:

(i) Delete v_i and its child leaf.
(ii) For each $j \in \{1, 2, \ldots, i-1\}$, delete (u_j, v_j) and, for each $j \in \{i+1, i+2, \ldots, k\}$, delete (u_{j+1}, v_j).
(iii) Contract each of the resulting degree-two vertices in

$$\{u_1, u_2, \ldots, u_{k+1}\} \cup \{v_1, v_2, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k\}.$$

Since $\mathcal{N}$ has no closed or overlapping reticulation chains, $\mathcal{N}'$ is indeed a phylogenetic network. Moreover, as $\mathcal{N}$ has no parent-child reticulations, it follows by the maximality of C that $\mathcal{N}'$ has no parent-child reticulations. Note that, because of maximality, one child of u_1 is a tree vertex or a leaf and one child of u_{k+1} is a tree vertex or a leaf. Furthermore, it is easily checked that $\mathcal{N}'$ has no closed or overlapping reticulation chains. Let n' and ℓ' denote the total number of vertices and the number of leaves in $\mathcal{N}'$, respectively. By induction,

$$n' \leq \left(\ell'\right)^2 + 3\ell' - 3.$$

Also, $\ell' = \ell - 1$ and, as C consists of $2k+1$ vertices,

$$n' = n - (2k+1) - 1.$$

Thus

$$n - (2k+1) - 1 \leq (\ell-1)^2 + 3(\ell-1) - 3,$$

that is,

$$n \leq \ell^2 + \ell + 2k - 3.$$

But, by Lemma 2.2, $k \leq \ell$ as C is not overlapping. Hence

$$n \leq \ell^2 + \ell + 2\ell - 3 = \ell^2 + 3\ell - 3,$$

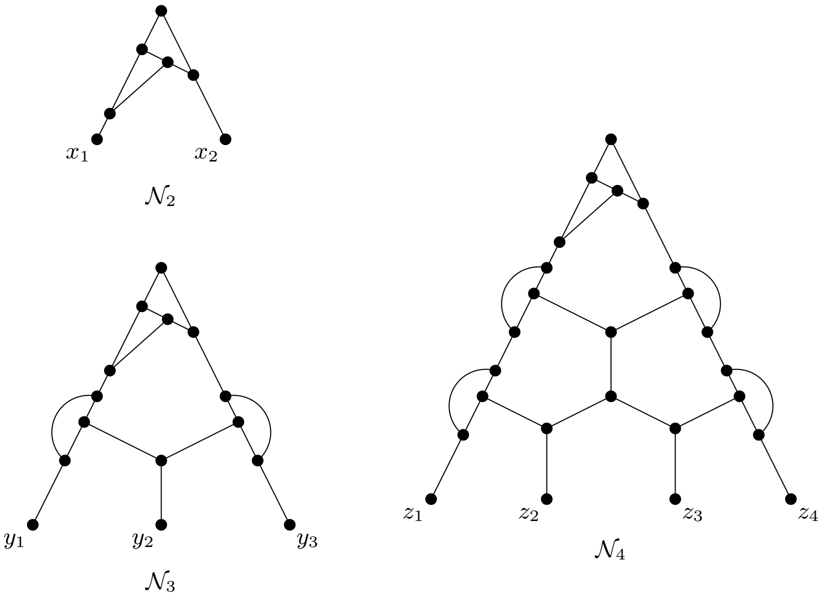

FIGURE 2. For $\ell = 2$, $\ell = 3$, and $\ell = 4$, the phylogenetic networks $\mathcal{N}_2$, $\mathcal{N}_3$, and $\mathcal{N}_4$ reach the upper bound of $\ell^2+3\ell-3$ vertices in total.

and so the inequality holds.

To see that the inequality in Theorem 1.1 is sharp, consider Figure 2. Each of the phylogenetic networks $\mathcal{N}_2$, $\mathcal{N}_3$, and $\mathcal{N}_4$ has no parent-child reticulations, and no closed or overlapping reticulation chains. Moreover, for $\ell = 2$, $\ell = 3$, and $\ell = 4$, the phylogenetic networks $\mathcal{N}_2$, $\mathcal{N}_3$, and $\mathcal{N}_4$, respectively, reach the upper bound of $\ell^2 + 3\ell - 3$ vertices in total. Note that, for $\ell = 1$, the phylogenetic network consisting of a single vertex reaches this upper bound. In general, we can recursively construct an appropriate phylogenetic network with ℓ leaves and whose total number of vertices is $\ell^2 + 3\ell - 3$ by taking the one with $\ell - 1$ leaves and adding $2\ell + 2$ vertices in a way analogous to that in which $\mathcal{N}_4$ can be constructed from $\mathcal{N}_3$. □

3. EXAMPLES

In this section, we give explicit examples to show that if $\mathcal{N}$ is a phylogenetic network that contains parent-child reticulations, closed reticulation chains, or overlapping reticulation chains, then there is no guarantee that the total number of vertices in $\mathcal{N}$ is bounded by the size of its leaf set.

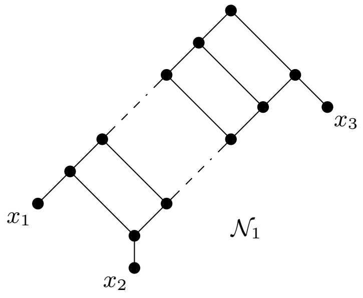

FIGURE 3. A phylogenetic network with parent-child reticulations but no closed or overlapping reticulation chains.

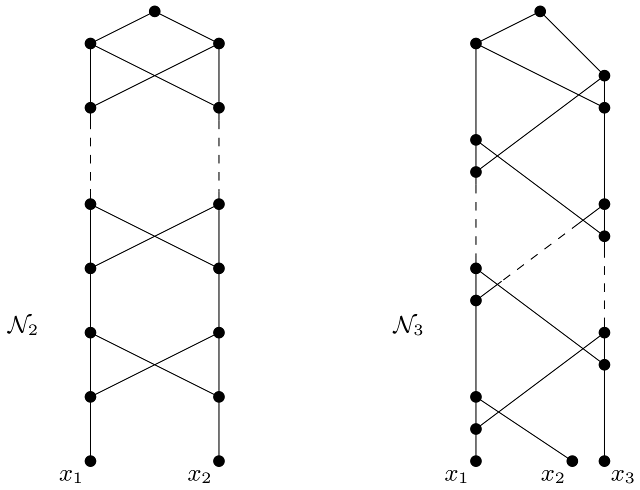

FIGURE 4. Two phylogenetic networks with no parent-child reticulations. The phylogenetic network $\mathcal{N}_2$ contains closed reticulation chains, while $\mathcal{N}_3$ contains a single maximal overlapping reticulation chain.

First consider the phylogenetic network $\mathcal{N}_1$ shown in Figure 3. The total number of vertices in $\mathcal{N}_1$ is not bounded by the size of its leaf set. It contains parent-child reticulations but no closed or overlapping reticulation chains.

Now consider the phylogenetic networks $\mathcal{N}_2$ and $\mathcal{N}_3$ shown in Figure 4. Neither $\mathcal{N}_2$ nor $\mathcal{N}_3$ contains a parent-child reticulation. Yet, for each of $\mathcal{N}_2$ and $\mathcal{N}_3$, the number of leaves does not bound the total number of vertices. The phylogenetic network $\mathcal{N}_2$ contains closed reticulations but no overlapping reticulation chain, while $\mathcal{N}_3$ contains a single maximal reticulation chain which is overlapping but not closed.

References

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

[2] P. Gambette, A.D.M. Gunawan, A. Labarre, S. Vialette, L. Zhang, Locating a tree in a phylogenetic network in quadratic time, in: Proc. 19th Ann. Inf. Conf. Res. Comp. Mol. Biol. (RECOMB'15), Lecture Notes in Computer Science, vol. 9029, 2015, pp. 96–107.

[3] A.D.M. Gunawan, B. DasGupta, L. Zhang, Locating a tree in a reticulation-visible network in cubic time, in: Proc. 20th Ann. Inf. Conf. Res. Comp. Mol. Biol. (RECOMB'16), in press.

[4] L. van Iersel, C. Semple, M. Steel, Locating a tree in a phylogenetic network, Inform. Process. Lett. 110 (2010) 1037–1043.

[5] C. McDiarmid, C. Semple, D. Welsh, Counting phylogenetic networks, Ann. Combin. 19 (2015) 205–224.

[6] C. Semple, M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.

[7] S.J. Willson, Properties of normal networks, Bull. Math. Biol. 72 (2010) 340–358.

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

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