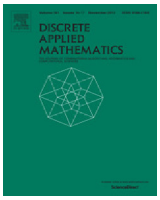
ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam



Note

Size of a phylogenetic network[☆]

Charles Semple

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

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form 23 August 2016
Accepted 4 September 2016
Available online xxxx

Keywords:
Phylogenetic networks
Size

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, [1–4,7].

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 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

---

[☆] The author was supported by the New Zealand Marsden Fund.
E-mail address: charles.semple@canterbury.ac.nz.

http://dx.doi.org/10.1016/j.dam.2016.09.004
0166-218X/© 2016 Elsevier B.V. All rights reserved.

Please cite this article in press as: C. Semple, Size of a phylogenetic network, Discrete Applied Mathematics (2016), http://dx.doi.org/10.1016/j.dam.2016.09.004

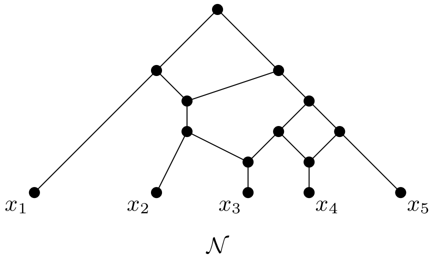


Fig. 1. A phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, x_3, x_4, x_5\}$.

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 Fig. 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

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

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 Fig. 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

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

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:

Please cite this article in press as: C. Semple, Size of a phylogenetic network, Discrete Applied Mathematics (2016), http://dx.doi.org/10.1016/j.dam.2016.09.004