WHEN IS A PHYLOGENETIC NETWORK SIMPLY AN AMALGAMATION OF TWO TREES?

CHARLES SEMPLE AND JACK SIMPSON

Abstract. Phylogenetic networks generalise phylogenetic (evolutionary) trees by allowing for the representation of reticulation (non-treelike) events. The structure of such networks is often viewed by the phylogenetic trees they embed. In this paper, we determine when a phylogenetic network $\mathcal{N}$ has two phylogenetic tree embeddings which collectively contain all of the edges of $\mathcal{N}$. This determination leads to a polynomial-time algorithm for recognising such networks and an unexpected characterisation of the class of reticulation-visible networks.

1. Introduction

The presence of reticulation (non-treelike) events in evolution has meant that, for certain collections of present-day species, phylogenetic networks, rather than phylogenetic trees, provide a more accurate description of evolutionary history. Such events include hybridisation and lateral gene transfer. However, the evolution of a particular gene can generally be described without reticulation events. As a result, there has been a variety of recent investigations concerning the underlying treelike structure of phylogenetic networks. These investigations include the small maximum parsimony problem for phylogenetic networks [18], determining whether a phylogenetic network displays a tree twice [5] and the well-studied tree-containment problem (for example, see [10, 15, 17]). In this context, a natural question to ask is to what extent does a given phylogenetic network differ from a phylogenetic tree? This question is particularly relevant in evolutionary biology to the continuing debate of whether, for certain collections such as prokaryotes, evolution is treelike with some reticulations or it has no treelike similarities at all [6, 7].

Date: June 23, 2018.

1991 *Mathematics Subject Classification.* 05C85, 92D15.

Key words and phrases. Phylogenetic networks, reticulation-visible networks, stack-free networks, tree-based networks.

The first author was supported by the New Zealand Marsden Fund.

To quantify this question, Francis and Steel [9] introduced the class of phylogenetic networks called tree-based networks. Loosely speaking, a phylogenetic network $\mathcal{N}$ is tree-based if it can be obtained from a phylogenetic tree $\mathcal{T}$ by simply adding edges whose end-vertices subdivide edges of $\mathcal{T}$. Equivalently, $\mathcal{N}$ is tree-based if, up to degree-two vertices, it has an embedding of a phylogenetic tree containing all of the vertices of $\mathcal{N}$. Here, we take a different approach to this question. Dating back at least to Hein [13], phylogenetic networks are frequently viewed as amalgamations of gene trees. For example, one of the most well-known tasks in mathematical and computational phylogenetics is to find, amongst all phylogenetic networks that embed a given set of gene trees, a network with the minimum number of reticulations (see, for example, [1, 14, 21]). From this viewpoint, the simplest phylogenetic network which is not a phylogenetic tree is one that is the amalgamation of two phylogenetic trees. Informally, a phylogenetic network $\mathcal{N}$ is 'two-tree coverable' if it has two phylogenetic tree embeddings which collectively contain every edge of $\mathcal{N}$. Not surprisingly, not every phylogenetic network is two-tree coverable. In this paper, we characterise the class of phylogenetic networks that are two-tree coverable. This characterisation leads immediately to a polynomial-time algorithm for deciding if an arbitrary phylogenetic network has a two-tree covering. It turns out that the increasingly prominent class of reticulation-visible networks is a subclass of the class of phylogenetic networks that are two-tree coverable. In fact, as we shall show, a particularly special subclass. Recent studies of reticulation-visible networks include [2, 11].

The paper is organised as follows. In the next section, we state the three main results, Theorems 2.1, 2.3, and 2.4, and end with an open problem of which Theorem 2.4 is a partial solution. The proofs of Theorems 2.1, 2.3 and 2.4 are given in Sections 3, 4, and 5, respectively.

We end the introduction with a comment. The underlying purpose of the paper is to introduce some new concepts and ideas to help instigate further research into the particularly important, but poorly understood, topic of the relationship between phylogenetic networks and the sets of phylogenetic trees they display. We hope this paper goes some way towards achieving this aim.

2. Main Results

Throughout the paper, X denotes a finite non-empty set. A *phylogenetic network on* X is a rooted acyclic directed graph with no parallel arcs having the following properties:

(i) the (unique) root has 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 $|X| = 1$, we additionally allow the directed graph consisting of the single vertex in X to be a phylogenetic network. The vertices in X are called *leaves*. Furthermore, the vertices of in-degree one and out-degree two are *tree vertices*, while the vertices of in-degree two and out-degree one are *reticulations*. The arcs directed into a reticulation are called *reticulation arcs*; all other arcs are *tree arcs*. In the literature, what we have called a phylogenetic network is sometimes referred to as a *binary* phylogenetic network. A *(binary) phylogenetic X-tree* is a phylogenetic network on X with no reticulations.

Let $\mathcal{N}$ be a phylogenetic network on X and let $\mathcal{T}$ be a phylogenetic X-tree. We say that $\mathcal{N}$ *displays* $\mathcal{T}$ if, up to degree-two vertices, $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting arcs and non-root vertices, in which case, the resulting acyclic digraph is an *embedding* of $\mathcal{T}$ in $\mathcal{N}$. Note that if $\mathcal{S}$ is an embedding of $\mathcal{T}$ in $\mathcal{N}$, then the root of $\mathcal{S}$ is the root of $\mathcal{N}$ and so it may have out-degree one. Furthermore, if $\mathcal{N}$ displays $\mathcal{T}$, then an embedding of $\mathcal{T}$ in $\mathcal{N}$ is not necessarily unique.

A phylogenetic network $\mathcal{N}$ on X is *two-tree coverable* if there are embeddings $\mathcal{S}_1$ and $\mathcal{S}_2$ of phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, in $\mathcal{N}$ such that each arc of $\mathcal{N}$ is an arc of either $\mathcal{S}_1$ or $\mathcal{S}_2$. If this holds, then $\{\mathcal{T}_1, \mathcal{T}_2\}$ (as well as $\{\mathcal{S}_1, \mathcal{S}_2\}$) is a *two-tree cover* of $\mathcal{N}$. Note that $\mathcal{T}_1$ and $\mathcal{T}_2$ need not be distinct and so a phylogenetic tree $\mathcal{T}$ is two-tree coverable as the multiset $\{\mathcal{T}, \mathcal{T}\}$ is a two-tree cover of $\mathcal{N}$. To illustrate, consider the phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, \ldots, x_5\}$ as well as the two phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ shown in Fig. 1. As with all figures in this paper, arcs are directed down the page. Both $\mathcal{T}_1$ and $\mathcal{T}_2$ are displayed by $\mathcal{N}$. In particular, an embedding of $\mathcal{T}_1$ is given by the dashed reticulation arcs and all of the tree arcs of $\mathcal{N}$, while an embedding of $\mathcal{T}_2$ is given by the non-dashed reticulation arcs and all of the tree arcs of $\mathcal{N}$. Thus $\{\mathcal{T}_1, \mathcal{T}_2\}$ is a two-tree cover of $\mathcal{N}$, and so $\mathcal{N}$ is two-tree coverable. We next state the first main result of this paper; a characterisation of phylogenetic networks that are two-tree coverable.

A phylogenetic network $\mathcal{N}$ on X is a *stack-free network* if $\mathcal{N}$ has no two reticulations, u and v say, such that u is the parent of v, that is, there is no reticulation edge in which both end-vertices are reticulations. Such a pair of reticulations are called *stack reticulations*. The class of stack-free networks arise naturally amongst the various classes of phylogenetic networks as follows. One of the most well-studied classes of phylogenetic

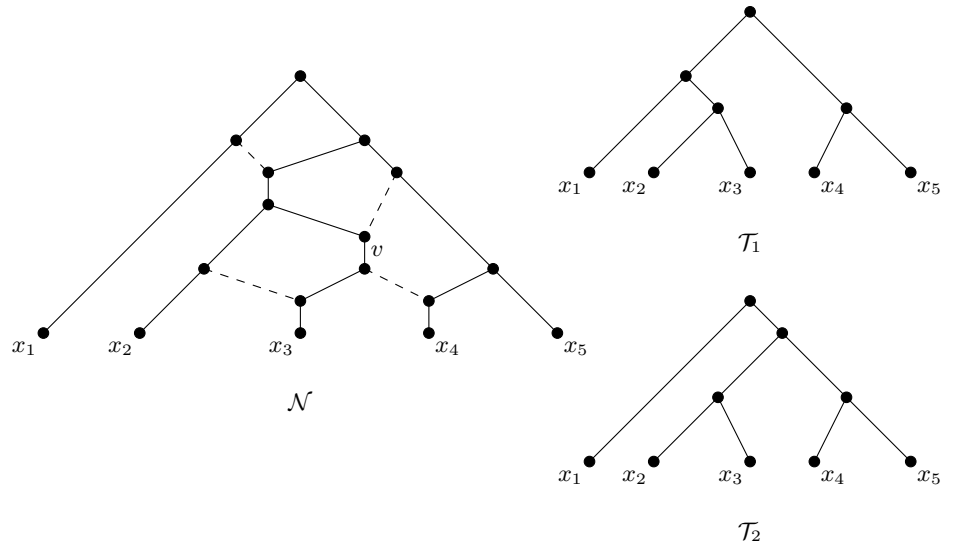


FIGURE 1. A phylogenetic network $\mathcal{N}$, and a two-tree cover $\{\mathcal{T}_1, \mathcal{T}_2\}$ of $\mathcal{N}$.

networks is the class of tree-child networks. A phylogenetic network is *tree-child* if each non-leaf vertex is the parent of a tree vertex or a leaf [4]. Tree-child networks have been characterised in a variety of ways including the characterisation that says a phylogenetic network is tree-child if and only if it has no stack or sibling reticulations [20]. Two distinct reticulations are *sibling reticulations* if they have a common parent. Thus stack-free networks generalise tree-child networks by allowing sibling reticulations. The first main result of this paper shows that the class of phylogenetic networks that are two-tree coverable coincides with the class of stack-free networks.

Theorem 2.1. *A phylogenetic network is two-tree coverable if and only if it is a stack-free network.*

By systematically checking that no reticulation arc joins two reticulations, an immediate consequence of Theorem 2.1 is the next corollary.

Corollary 2.2. *Deciding if an arbitrary phylogenetic network $\mathcal{N}$ is two-tree coverable can be done in time polynomial in the number of vertices of $\mathcal{N}$.*

In establishing Theorem 2.1, we explicitly construct, in time polynomial in the number of vertices, a two-tree cover for a stack-free network.

What is the relationship between stack-free networks and tree-based networks? A phylogenetic network $\mathcal{N}$ on X is a *tree-based network* if it has an embedding $\mathcal{S}$ of a phylogenetic X-tree $\mathcal{T}$ with the property that each vertex of $\mathcal{N}$ is a vertex of $\mathcal{S}$, in which case, $\mathcal{T}$ is a *base tree* of $\mathcal{N}$. Note that this

definition is equivalent to the original definition given in [9]. It immediately follows from Theorem 2.1 and a characterisation of Zhang [22] that stack-free networks are tree-based networks, however, the converse does not hold. For example, the phylogenetic network shown in Fig. 2(i) is tree-based but it does not have a two-tree covering.

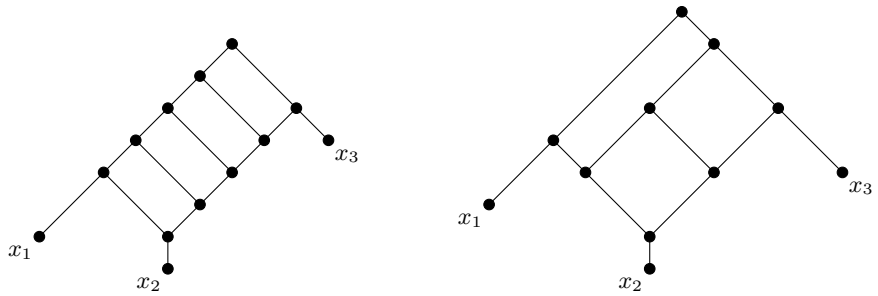
Before stating the second main result, we make two remarks. First, unlike tree-child networks, the number of vertices in a stack-free network is not bounded by the size of its leaf set (see Fig. 4 in [19]). Second, it is thus natural to ask whether, for all $n \geq 1$, there is a *universal* stack-free network on X, where $n = |X|$, that is, a stack-free network on X that (simultaneously) displays every phylogenetic X-tree. This question was originally asked in the context of tree-based networks [9], for which a positive answer was independently established in [12] and [22]. The associated constructions were recently sharpened to a 'best' possible construction using $O(n \log n)$ reticulations [3]. Curiously, all of these constructions take the same approach and all of them are stack-free networks, and so each has a two-tree cover.

To state the second main result, let $\mathcal{N}$ be a phylogenetic network on X with root ρ. A vertex u in $\mathcal{N}$ is *visible* if there is a leaf $x \in X$ such that every directed path from ρ to x traverses u. It is interesting to note that tree-child networks are precisely the phylogenetic networks in which every vertex is visible [4]. A *reticulation-visible network* is a phylogenetic network with the property that every reticulation is visible. It is easily seen that if $\mathcal{N}$ is reticulation-visible, then $\mathcal{N}$ has no stack-reticulations and so, by Theorem 2.1, $\mathcal{N}$ is a stack-free network. However, not every stack-free network is reticulation-visible. For example, in Fig. 1, $\mathcal{N}$ is a stack-free network but it is not reticulation-visible as the reticulation labelled v is not visible. To see this, observe that, for each $i \in \{1, 2, \ldots, 5\}$, there is a directed path from the root of $\mathcal{N}$ to x_i avoiding v. The next theorem characterises reticulation-visible networks in terms of two-tree coverings, thereby showing that such networks are special subclass of stack-free networks.

Let $\mathcal{N}$ be a stack-free network on X. Let $\{E_1, E_2\}$ be a partition of the reticulation arcs in $\mathcal{N}$ so that, for each reticulation v, one arc directed into v is in E_1, while the other arc directed into v is in E_2. We call $\{E_1, E_2\}$ a *complementary partition* of the set of reticulation arcs in $\mathcal{N}$. We say that $\mathcal{N}$ is *freely coverable* if, for each complementary partition of the set of reticulation arcs of $\mathcal{N}$, there is a two-tree covering $\{\mathcal{S}_1, \mathcal{S}_2\}$ of $\mathcal{N}$ such that E_1 (resp. E_2) is a subset of the arc set of $\mathcal{S}_1$ (resp. $\mathcal{S}_2$).

Theorem 2.3. *A phylogenetic network is reticulation-visible if and only if it is freely coverable.*

We now state the third main result and an open problem. Let $\mathcal{N}$ be an arbitrary phylogenetic network on X. For a positive integer k, we say that



(i) (ii)

FIGURE 2. (i) A tree-based network that has a 5-tree covering but not a 4-tree covering and (ii) a phylogenetic network that does not have a 2-tree covering, yet the longest directed path in which each vertex is a reticulation is two.

$\mathcal{N}$ is *k-tree coverable* if there is a set $\mathcal{P}$ of at most k phylogenetic X-tree embeddings in $\mathcal{N}$ with the property that each arc of $\mathcal{N}$ is an arc of some embedding in $\mathcal{P}$, in which case, $\mathcal{P}$ is a *k-tree covering* of $\mathcal{N}$. What is the smallest k for which $\mathcal{N}$ is k-tree coverable? It is easily seen that if $\mathcal{N}$ has a directed path consisting of ℓ vertices each of which is a reticulation, then $k \geq \ell + 1$. For example, the tree-based network in Fig. 2(i) has such a path consisting of four reticulations and it is easily checked that it has a 5-tree covering but not a 4-tree covering. Indeed, it is natural to conjecture that $\mathcal{N}$ has a k-tree covering if and only if $k \geq \ell + 1$, where ℓ is the number of vertices in a maximum-length directed path in $\mathcal{N}$ in which each vertex is a reticulation. However, it is easily checked that, for the phylogenetic network shown in Fig. 2(ii), the smallest k for which it has a k-tree covering is four, yet the longest directed path where each vertex is a reticulation is two.

The third main result is a resolution of the above problem for a class of phylogenetic networks that naturally generalises reticulation-visible networks but is not contained within the class of tree-based networks. Let $\mathcal{N}$ be a phylogenetic network. A reticulation of $\mathcal{N}$ is a *sink* if it is the parent of a tree vertex. For example, in each of Fig. 2(i) and (ii), the parent of x_2 is a sink but every other reticulation is not a sink. A *sink-visible network* is a phylogenetic network with the property that every sink is visible. Thus the class of reticulation-visible networks is contained in the class of sink-visible networks as every reticulation is a sink.

Let $\mathcal{N}$ be a phylogenetic network and let v be a sink of $\mathcal{N}$. A reticulation arc (t, u) of $\mathcal{N}$ is a *source arc for v* if t is a tree vertex and there is a directed path P in $\mathcal{N}$ from t to v that traverses (t, u) in which every arc in P is a reticulation arc. Observe that if there is such a path, then that path is

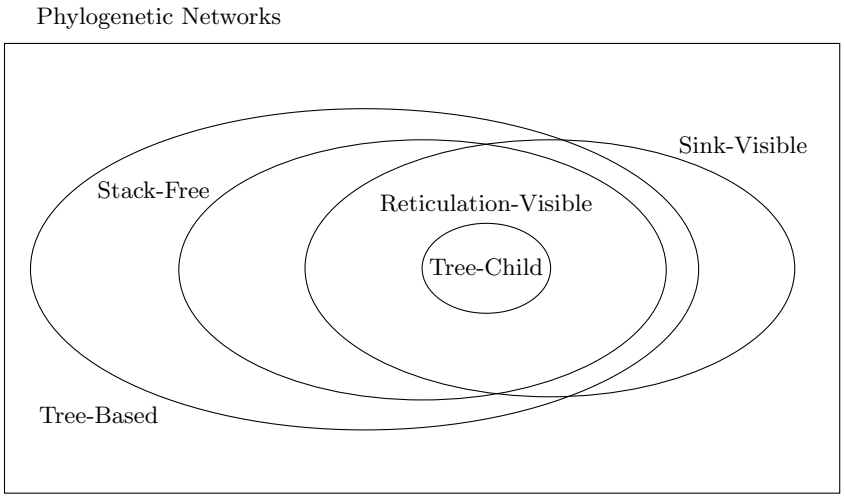


FIGURE 3. Relationships between various classes of phylogenetic networks.

unique as a reticulation has out-degree one. We denote the total number of source arcs for v by $s(v)$.

Theorem 2.4. *Let $\mathcal{N}$ be a sink-visible network. Then $\mathcal{N}$ has a k-tree covering if and only if*

$$k \geq \max\{s(v) : v \text{ is a sink of } \mathcal{N}\}.$$

The proof of Theorem 2.4 crucially relies on each sink being visible. Indeed, it is straightforward to construct an example for which the outcome of the theorem does not hold if $\mathcal{N}$ is not sink-visible. We leave it as an open problem to determine the smallest value of k for which a given arbitrary phylogenetic network has a k-covering. Since, in part, the work in this paper is motivated by tree-based networks, it worth noting that while such networks have an embedding of a phylogenetic tree containing every vertex, the smallest value of k for which tree-based networks have a k-tree covering can be arbitrarily large. The reason for this is that tree-based networks can have arbitrarily long paths in which each vertex is a reticulation.

In stating the main results of the paper, we have mentioned five classes of phylogenetic networks, namely, tree-child, reticulation-visible, stack-free, tree-based, and sink-visible networks. We end this section with a Venn diagram illustrating the relationships between these classes (see Fig. 3). It is easily checked that each of the indicated inclusions is proper and, furthermore, the intersection of the classes of stack-free and sink-visible networks is precisely the class of reticulation-visible networks.

3. Proof of Theorem 2.1

We begin with a construction which we will eventually show constructs a two-tree cover of a stack-free network. Let $\mathcal{N}$ be a stack-free network on X, and let E_R denote the set of reticulation arcs of $\mathcal{N}$. Choose a subset M of E_R satisfying the following properties:

(i) if v is a reticulation of $\mathcal{N}$, then M contains exactly one arc directed into v; and
(ii) if u is the parent of two (distinct) reticulations, then M contains exactly one arc incident with u.

To see that there exists such a subset M of reticulation arcs and how to find M, consider the following bipartite graph. Let R be the set of reticulations of $\mathcal{N}$ and let P be the set of vertices of $\mathcal{N}$ which are parents of at least one reticulation. Since $\mathcal{N}$ is stack-free, P consists of tree vertices, and so P and R are disjoint. Let B be the bipartite graph with vertex bipartition $\{P, R\}$ and edge set

$$\big\{\{p, r\} : \ p \in P,\, r \in R, \text{ and } (p, r) \text{ is an arc of } \mathcal{N}\big\}.$$

It is easily checked that each component of B consists of either a path starting and ending at vertices in P, or a cycle. It now follows that a subset of arcs in $\mathcal{N}$ satisfying (i) and (ii) exists and can be found in time polynomial in the number of vertices in $\mathcal{N}$. Observe that, viewing the arcs in M as undirected edges, M is a matching of B, that is, a subset of edges no two of which are incident with the same vertex. For the reader familiar with tree-based networks, such matchings are reminiscent of the theory underlying tree-based networks [8, 16, 22].

We refer to M as a *cover matching* of $\mathcal{N}$. Observe that the subset of reticulation arcs of $\mathcal{N}$ not in M, that is $E_R - M$, is also a cover matching, thus $\{M, E_R - M\}$ is a complementary partition of E_R. To illustrate, the dashed reticulation arcs of the stack-free network $\mathcal{N}$ shown in Fig. 1 is a cover matching of $\mathcal{N}$. The sufficient direction of Theorem 2.1 as well as the polynomial-time construction of a two-tree cover of a stack-free network follows from the next lemma.

Lemma 3.1. *Let $\mathcal{N}$ be a stack-free network on X, and let M be a cover matching of $\mathcal{N}$. Then there is an embedding in $\mathcal{N}$ of a phylogenetic X-tree containing each of the arcs in M as well as each of the tree arcs in $\mathcal{N}$.*

Proof. Let E_T denote the set of tree arcs of $\mathcal{N}$, and let $\mathcal{S}$ be the embedding in $\mathcal{N}$ induced by the arcs in $E_T \cup M$. As $\mathcal{S}$ does not contain two reticulation arcs directed into the same reticulation, $\mathcal{S}$ has no underlying cycles. Therefore, observing that at least one edge incident with the root of $\mathcal{N}$ is a tree arc,

to establish the lemma, it suffices to show that if u is a vertex of $\mathcal{S}$ with out-degree zero in the embedding, then u is an element of X.

Choose u to be a vertex of $\mathcal{S}$ with out-degree zero and suppose u is not an element of X. By construction, no arc directed out of u is a tree arc so, as $\mathcal{N}$ has no stack reticulations, u is the parent of two reticulations, v_1 and v_2 say. But then, as M is a cover matching, either (u, v_1) or (u, v_2) is an arc in $\mathcal{S}$, and so u is not a vertex of $\mathcal{S}$ with out-degree zero; a contradiction. Thus $\mathcal{S}$ is an embedding of a phylogenetic X-tree displayed by $\mathcal{N}$, thereby completing the proof of the lemma. □

Proof of Theorem 2.1. Let $\mathcal{N}$ be a phylogenetic network on X. If $\mathcal{N}$ contains a stack reticulation, then at least three phylogenetic X-tree embeddings are necessary to cover $\mathcal{N}$, that is, the smallest k for which $\mathcal{N}$ has a k-tree covering is at least three. Thus if $\mathcal{N}$ is two-tree coverable, $\mathcal{N}$ is stack-free. Conversely, suppose that $\mathcal{N}$ is a stack-free network. Choosing a cover matching of $\mathcal{N}$, it follows by Lemma 3.1 and the observation prior to it that $\mathcal{N}$ is two-tree coverable. This completes the proof of Theorem 2.1. □

4. Proof of Theorem 2.3

We start with a general lemma concerning the uniqueness of an embedding. For a phylogenetic network $\mathcal{N}$, a *tree-path* is a directed path in which each arc is a tree arc. Reversing the order of the vertices, and thus the arcs, in such a path is referred to as a *backward tree-path*.

Lemma 4.1. *Let $\mathcal{N}$ be a phylogenetic network on X, and let $\mathcal{S}$ be an embedding of a phylogenetic X-tree in $\mathcal{N}$. Let E_1 denote the subset of reticulation arcs of $\mathcal{N}$ contained in $\mathcal{S}$. Then $\mathcal{S}$ is the unique embedding in $\mathcal{N}$ of a phylogenetic X-tree containing E_1.*

Proof. Let $\mathcal{S}'$ be an embedding of a phylogenetic X-tree displayed by $\mathcal{N}$ containing each of the arcs in E_1. Let F_1 denote the set of tree arcs in $\mathcal{N}$ on a backward tree-path starting at either a leaf or a vertex that is a tail of an arc in E_1. Observe that if e is a tree arc in F_1, then $\mathcal{S}'$ contains e. But, the arcs in $E_1 \cup F_1$ have the property that, for each leaf $x \in X$, there is a directed path from the root of $\mathcal{N}$ to x using only the arcs in $E_1 \cup F_1$. Thus $E_1 \cup F_1$ is the arc set of $\mathcal{S}'$, and also the arc set of $\mathcal{S}$. It now follows that $\mathcal{S}' = \mathcal{S}$. □

Lemma 4.2. *Let $\mathcal{N}$ be a phylogenetic network on X, and let v be a reticulation of $\mathcal{N}$. If v is not visible, then $\mathcal{N}$ has an embedding of a phylogenetic X-tree avoiding v.*

Proof. Suppose v is not visible. Then, for each $x \in X$, there is a directed path from the root of $\mathcal{N}$ to x that avoids traversing v. Let $\mathcal{S}$ be the embedding of $\mathcal{N}$ induced by the subset of arcs of $\mathcal{N}$ in at least one of these paths. Up to degree-two vertices, $\mathcal{S}$ is a phylogenetic network on X. Thus a subset of arcs of $\mathcal{S}$ induces an embedding in $\mathcal{N}$ of a phylogenetic X-tree. This embedding avoids v, and so completes the proof of the lemma. □

We now combine the last two lemmas to establish Theorem 2.3.

Proof of Theorem 2.3. Let $\mathcal{N}$ be a phylogenetic network on X. Suppose that $\mathcal{N}$ is not reticulation-visible. Then, by Lemma 4.2, there is an embedding $\mathcal{S}$ in $\mathcal{N}$ of a phylogenetic X-tree that avoids a reticulation, v say. Let E_1 denote the subset of reticulation arcs of $\mathcal{N}$ in $\mathcal{S}$. Let $\{E_1', E_2\}$ be a complementary partition of the set of reticulation arcs of $\mathcal{N}$, where E_1 is a subset of E_1'. Since $\mathcal{S}$ avoids v, it follows that E_1 is a proper subset of E_1'. By Lemma 4.1, $\mathcal{S}$ is the unique embedding of a phylogenetic X-tree containing E_1, and so there is no embedding in $\mathcal{N}$ of a phylogenetic X-tree containing E_1'. It follows that $\mathcal{N}$ is not freely coverable.

Now suppose that $\mathcal{N}$ is reticulation-visible. Let $\{E_1, E_2\}$ be a complementary partition of the set of reticulation arcs in $\mathcal{N}$. We first show that $\mathcal{N}$ has an embedding of a phylogenetic X-tree containing the arcs in E_1. Let F_1 denote the set of tree arcs in $\mathcal{N}$ on a backward tree-path starting at either a leaf or a vertex that is a tail of an arc in E_1. We show that the arcs in $E_1 \cup F_1$ induces an embedding $\mathcal{S}_1$ of a phylogenetic X-tree displayed by $\mathcal{N}$.

As E_1 contains exactly one reticulation arc directed into a reticulation, $\mathcal{S}_1$ has no underlying cycles and so, by construction of $\mathcal{S}_1$, there is a (unique) path in $\mathcal{S}_1$ against the direction of the arcs from each $x \in X$ to the root ρ of $\mathcal{N}$. We next show that the vertices in $\mathcal{S}_1$ with out-degree zero are vertices in X. If not, then there is a vertex v with out-degree zero in $\mathcal{S}_1$ but with out-degree one or two in $\mathcal{N}$. By construction, v is a reticulation in $\mathcal{N}$. In turn, this implies that v is not visible in $\mathcal{N}$ as there is no $x \in X$ such that every path from ρ to x traverses v; a contradiction. Thus $\mathcal{S}_1$ is an embedding in $\mathcal{N}$ of a phylogenetic X-tree.

Similarly, let F_2 denote the set of tree arcs in $\mathcal{N}$ on a backward tree-path starting at either a leaf or a vertex that is the tail of an arc in E_2. Then, as above, the edges in $E_2 \cup F_2$ induces an embedding $\mathcal{S}_2$ of a phylogenetic X-tree displayed by $\mathcal{N}$. Now every reticulation edge of $\mathcal{N}$ is in either $\mathcal{S}_1$ or $\mathcal{S}_2$. To complete the proof, let e be a tree arc in $\mathcal{N}$ that is not an arc in either $\mathcal{S}_1$ or $\mathcal{S}_2$, that is $e \notin F_1 \cup F_2$. Then there is a tree-path consisting of arcs not in $F_1 \cup F_2$ from the head of e to a vertex u that is the parent of two reticulations, v_1 and v_2 say. But $(u, v_1), (u, v_2) \in E_1 \cup E_2$ and so $e \in F_1 \cup F_2$;

a contradiction. Hence $\{\mathcal{S}_1, \mathcal{S}_2\}$ is a two-tree cover of $\mathcal{N}$. It follows that $\mathcal{N}$ is freely coverable. □

5. Proof of Theorem 2.4

Proof of Theorem 2.4. First suppose that $\mathcal{N}$ has a k-tree covering. Let v be a sink of $\mathcal{N}$ and let $S(v)$ denote the set of source arcs for v, and assume that $\mathcal{S}$ is an embedding in $\mathcal{N}$ of a phylogenetic X-tree displayed by $\mathcal{N}$. If (t, u) is an arc in $S(v)$ and $\mathcal{S}$ contains (t, u), then $\mathcal{S}$ contains each of the arcs on the unique directed path from t to v. In particular, $\mathcal{S}$ contains no other source arc for v; otherwise, $\mathcal{S}$ has a vertex with in-degree two. Thus $\mathcal{S}$ contains at most one source arc for v. It immediately follows that $k \geq s(v)$, and so

$$k \geq \max\{s(v) : v \text{ is a sink of } \mathcal{N}\}.$$

To prove the converse, now suppose

$$k \geq \max\{s(v) : v \text{ is a sink of } \mathcal{N}\}.$$

We next construct a k-tree covering of $\mathcal{N}$. Let $\{E_1, E_2, \ldots, E_k\}$ be a collection of subsets of the set E_s of source arcs of $\mathcal{N}$ satisfying the following properties:

(i) For all $i \in \{1, 2, \ldots, k\}$, the set E_i contains exactly one source arc for each sink of $\mathcal{N}$.
(ii) The union $\bigcup_{i\in\{1,2,\ldots,k\}} E_i = E_s$.

By the choice of k, such a collection is possible.

For each i, let F_i denote the set of tree arcs in $\mathcal{N}$ on a backward tree-path starting at either a leaf or a vertex that is a tail of a source arc in E_i. Furthermore, for each i, let G_i denote the subset of reticulation arcs of $\mathcal{N}$ on a directed path from the head of a source arc in E_i to its respective sink. For all i, the arcs in

$$E_i \cup F_i \cup G_i$$

are the arcs of an embedding, $\mathcal{S}_i$ say, of a phylogenetic X-tree displayed by $\mathcal{N}$. To see this, first note that, since there is exactly one source arc in E_i for each sink, $\mathcal{S}_i$ has no underlying cycles. Therefore, for each $x \in X$, there is a unique path in $\mathcal{S}_i$ against the direction of the arcs from x to the root of $\mathcal{N}$. Moreover, if there is a vertex v in $\mathcal{S}_i$ with out-degree zero and $v \notin X$, then, by construction, v is a sink. But then v is not visible; a contradiction. It follows that $\mathcal{S}_i$ is an embedding in $\mathcal{N}$ of a phylogenetic X-tree displayed by $\mathcal{N}$.

We complete the proof by showing that $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k\}$ is a k-covering of $\mathcal{N}$. Let e be an arc of $\mathcal{N}$. If e is a tree arc, then there is a path in $\mathcal{N}$ from the head of e to either a leaf or a vertex that is the tail of a source arc, in which case, for some i, we have $e \in F_i$ and so e is an arc in $\mathcal{S}_i$. Assume e is a reticulation arc. If e is a source arc, then $e \in E_i$ for some i, so $e \in \mathcal{S}_i$. Otherwise, for some sink v of $\mathcal{N}$ and source arc f for v, we have that e lies on the unique directed path from the head of f to v, in which case, $e \in G_i$ for some i where $f \in E_i$. Thus e is an arc in one of the embeddings $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$. It now follows that $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k\}$ is a k-tree covering of $\mathcal{N}$. This completes the proof of the theorem. □

References

[1] Bordewich M, Semple C (2007) Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. IEEE/ACM Trans Comput Biol Bioinform 4: 458–466

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

[3] Bordewich M, Semple C A universal tree-based network with the minimum number of reticulations. Discrete Appl Math, in press

[4] Cardona G, Rossello F, Valiente G, (2009) Comparison of tree-child phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform 6:552–569

[5] Cordue P, Linz S, Semple C (2014) Phylogenetic networks that display a tree twice. Bull Math Biol 76:2664–2679

[6] Dagan T, Martin WF (2006) The tree of one percent. Genome Biol 7:118

[7] Doolittle WF, Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA 104:2043–2049

[8] Francis A, Semple C, Steel M (2018) New characterisations of tree-based networks and proximity measures. Adv Appl Math 93:93–107

[9] Francis AR, Steel M (2015) Which phylogenetic networks are merely trees with additional arcs? Syst Biol 64:768–777

[10] Gambette P, van Iersel L, Kelk S, Pardi F, Scornavacca C, Do branch lengths help locate a tree in a phylogenetic network? Bull Math Biol 78:1773–1795

[11] Gunawan ADM, DasGupta B, Zhang L (2017) A decomposition theorem and two algorithms for reticulation-visible networks. Inform Comput 252:161–175

[12] Hayamizu H (2016) On the existence of infinitely many universal tree-based networks. J Theor Biol 396:204–206

[13] Hein J (1990) Reconstructing evolution of sequences subject to recombination using parsimony. Math Biosci 98:185–200

[14] van Iersel L, Kelk S, Lekić N, Whidden C, Zeh N (2016) Hybridization number on three rooted binary trees is EPT. SIAM J Discrete Math 30:1607–1631

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

[16] Jetten L, van Iersel L (2018) Nonbinary tree-based phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform 15:205–217

[17] Kanj I, Nakhleh L, Than C, Xia G (2008) Seeing the trees and their branches in the network is hard. Theoret Comput Sci 401:153–164

[18] Nakhleh L, Jin G, Zhao F, Mellor-Crummey J (2005) Reconstructing phylogenetic networks using maximum parsimony. In: IEEE Computational Systems Bioinformatics Conference, pp 93–102

[19] Semple C (2017) Size of a phylogenetic network. Discrete Appl Math 217:362–367
[20] Semple C (2016) Phylogenetic networks with every embedded phylogenetic tree a base trees. Bull Math Biol 78:132–137
[21] Song Y, Hein J (2003) Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events. In: Benson G, Page R (eds) Algorithms in Bioinformatics (WABI), Lecture Notes in Bioinformatics, vol 2812, pp 287–302
[22] Zhang L (2016) On tree-based phylogenetic networks. J Comput Biol 23:553–565

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

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

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

E-mail address: jack.simpson@pg.canterbury.ac.nz