Attaching leaves and picking cherries to characterise the hybridisation number for a set of phylogenies

Simone Linz[a], Charles Semple[b]

[a]*School of Computer Science, University of Auckland, New Zealand*
[b]*School of Mathematics and Statistics, University of Canterbury, New Zealand*

Abstract

Throughout the last decade, we have seen much progress towards characterising and computing the minimum hybridisation number for a set $\mathcal{P}$ of rooted phylogenetic trees. Roughly speaking, this minimum quantifies the number of hybridisation events needed to explain a set of phylogenetic trees by simultaneously embedding them into a phylogenetic network. From a mathematical viewpoint, the notion of agreement forests is the underpinning concept for almost all results that are related to calculating the minimum hybridisation number for when $|\mathcal{P}| = 2$. However, despite various attempts, characterising this number in terms of agreement forests for $|\mathcal{P}| > 2$ remains elusive. In this paper, we characterise the minimum hybridisation number for when $\mathcal{P}$ is of arbitrary size and consists of not necessarily binary trees. Building on our previous work on cherry-picking sequences, we first establish a new characterisation to compute the minimum hybridisation number in the space of tree-child networks. Subsequently, we show how this characterisation extends to the space of all rooted phylogenetic networks. Moreover, we establish a particular hardness result that gives new insight into some of the limitations of agreement forests.

Keywords: agreement forest, cherry-picking sequence, minimum hybridisation, phylogenetic networks, reticulation, tree-child networks

Email addresses: s.linz@auckland.ac.nz (Simone Linz), charles.semple@canterbury.ac.nz (Charles Semple)

1. Introduction

In our quest for faithfully describing evolutionary histories, we are currently witnessing a shift from the representation of ancestral histories by phylogenetic (evolutionary) trees towards phylogenetic networks. The latter not only represent speciation events but also non-tree like events such as hybridisation and horizontal gene transfer that have played an important role throughout the evolution of certain groups of organisms as for example in plants and fish [10, 17, 18, 22].

In this paper, we focus on a problem that is related to the reconstruction of phylogenetic networks. Called Minimum Hybridisation and formally stated at the end of this section, this problem was first introduced by Baroni et al. [2]. While Minimum Hybridisation was historically motivated by attempting to quantify hybridisation events, it is now more broadly regarded as a tool to quantify all non-tree like events to which we collectively refer to as reticulation events. Pictorially speaking, Minimum Hybridisation aims at the reconstruction of a phylogenetic network that simultaneously embeds a given set of phylogenetic trees while minimising the number of reticulation events that are represented by vertices in the network whose in-degree is at least two. More formally, the problem is based on the following underlying question. Given a collection $\mathcal{P}$ of rooted phylogenetic trees on the same set of taxa that have correctly been reconstructed for different parts of the species' genomes, what is the smallest number of reticulation events that is needed to explain $\mathcal{P}$? Over the last ten years, we have seen significant progress in characterising and computing this minimum number for when $|\mathcal{P}| = 2$ (e.g. see [1, 3, 4, 8, 16, 24]). However, except for some heuristic approaches [7, 25], less is known for when $|\mathcal{P}| \geq 3$. This is due to the fact that the notion of agreement forests, which underlies almost all results that are related to Minimum Hybridisation, appears to be ungeneralisable to more than two trees.

Previously, together with Humphries, we introduced cherry-picking sequences and characterised a restricted version of Minimum Hybridisation for $\mathcal{P}$ being binary and of arbitrary size [12]. Instead of minimising the number of reticulation events needed to explain $\mathcal{P}$ over the space of all rooted phylogenetic networks, this restricted version only considers binary temporal tree-child networks. Such networks are the binary intersection of the classes of temporal networks and tree-child networks introduced by Moret et al. [19] and Cardona et al. [6], respectively. Disadvantageously, this restriction is so

strong that not even if $|\mathcal{P}| = 2$ are we guaranteed to have a solution, i.e. there may be no such network explaining $\mathcal{P}$ [11, Figure 2].

Here, we advance our work on cherry-picking sequences and establish two new characterisations to quantify the amount of reticulation events that are needed to explain a set of (not necessarily binary) phylogenetic trees. The first characterisation solves the problem over the space of tree-child networks. Unlike temporal networks, we show that every collection $\mathcal{P}$ of rooted phylogenetic trees has a solution, i.e. the trees in $\mathcal{P}$ can simultaneously be embedded into a tree-child network. Subsequently, we extend this characterisation to the space of all rooted phylogenetic networks and, hence, provide the first characterisation for Minimum Hybridisation in its most general form. Both characterisations are based on computing a cherry-picking sequence for $\mathcal{P}$, while the latter characterisation makes also use of an operation that attaches auxiliary leaves to the trees in $\mathcal{P}$.

In addition to the two new characterisations, we return back to agreement forests and investigate why they seem to be of limited use to solve Minimum Hybridisation for an arbitrary size set $\mathcal{P}$ of rooted phylogenetic trees. Roughly speaking, given $\mathcal{P}$, one can compute a particular type of agreement forest $\mathcal{F}$ of smallest size and, if $|\mathcal{P}| = 2$, then each but one component in $\mathcal{F}$ contributes exactly one to the minimum number of reticulation events that is needed to explain $\mathcal{P}$. On the other hand, if $|\mathcal{P}| > 2$, the contribution of each component in $\mathcal{F}$ to this minimum number is much less clear. Motivated by this drawback of agreement forests, we consider a set $\mathcal{P}$ of rooted binary phylogenetic trees as well as the agreement forest $\mathcal{F}$ *induced* (formally defined in Section 5) by a phylogenetic network that explains $\mathcal{P}$ and minimises the number of reticulations events and ask whether or not, it is computationally hard to calculate the minimum number of reticulation events that is needed to explain $\mathcal{P}$. We call the associated decision problem Scoring Optimum Forest. This problem was first mentioned in [13], where the authors conjecture that Scoring Optimum Forest is NP-complete. Using the machinery of cherry-picking sequences, we show that Scoring Optimum Forest is NP-complete for when one considers the smaller space of tree-child networks.

The paper is organised as follows. The remainder of the introduction contains some definitions and preliminaries on phylogenetic networks. In Section 2, we state the two new characterisations in terms of cherry-picking sequences. The first optimises Minimum Hybridisation within the space of tree-child networks and the second optimises Minimum Hybridisation within the space of all phylogenetic networks. The second characterisation is

an extension of the first by additionally allowing the attachment of auxiliary leaves. We then establish proofs for both characterisations in Section 3 as well as a formal description of the analogous algorithm. In Section 4, we establish an upper bound on the number of auxiliary leaves that, given a collection of phylogenetic trees, are needed to characterise Minimum Hybridisation over the space of all rooted phylogenetic networks. Lastly, in Section 5, we formally state the problem Scoring Optimum Forest and show that it is NP-complete. We finish the paper with some concluding remarks in Section 6.

Throughout the paper, X denotes a non-empty finite set. A *phylogenetic network $\mathcal{N}$ on X* is a rooted acyclic digraph with no parallel edges that satisfies the following properties:

(i) the (unique) root has out-degree two,

(ii) the set X is the set of vertices of out-degree zero, each of which has in-degree one, and

(iii) all other vertices either have in-degree one and out-degree two, or in-degree at least two and out-degree one.

For technical reasons, if $|X| = 1$, we additionally allow $\mathcal{N}$ to consist of the single vertex in X. The set X is the *leaf set* of $\mathcal{N}$ and the vertices in X are called *leaves*. We sometimes denote the leaf set of $\mathcal{N}$ by $\mathcal{L}(\mathcal{N})$. For two vertices u and v in $\mathcal{N}$, we say that u is a *parent* of v and v is a *child* of u if (u, v) is an edge in $\mathcal{N}$. Furthermore, the vertices of in-degree at most one and out-degree two are *tree vertices*, while the vertices of in-degree at least two and out-degree one are *reticulations*. An edge directed into a reticulation is called a *reticulation edge* while each non-reticulation edge is called a *tree edge*. We say that $\mathcal{N}$ is *binary* if each reticulation has in-degree exactly two. Lastly, a directed path P in $\mathcal{N}$ ending at a leaf is a *tree path* if every intermediate vertex in P is a tree vertex.

A phylogenetic network $\mathcal{N}$ on X is *tree-child* if each non-leaf vertex in $\mathcal{N}$ is the parent of at least one tree vertex or leaf. An example of two tree-child networks $\mathcal{N}$ and $\mathcal{N}'$ is given at the bottom of Figure 1. Note that the phylogenetic network obtained from $\mathcal{N}$ by deleting the leaf labelled 4 and suppressing the resulting degree-two vertex v results in a network that is not tree-child.

A *rooted phylogenetic X-tree* $\mathcal{T}$ is a rooted tree with no degree-two vertices except possibly the root which has degree at least two, and with leaf set X. If $|X| = 1$, then $\mathcal{T}$ consists of the single vertex in X. As for phylogenetic networks, the set X is called the *leaf set* of $\mathcal{T}$ and is denoted by $\mathcal{L}(\mathcal{T})$. In addition, $\mathcal{T}$ is *binary* if $|X| = 1$ or, apart from the root which has degree two, all interior vertices have degree three. Since we are only interested in *rooted* phylogenetic trees and *rooted* binary phylogenetic trees in this paper, we will refer to such trees simply as phylogenetic trees and binary phylogenetic trees, respectively. For a phylogenetic X-tree $\mathcal{T}$, we consider two types of subtrees. Let X' be a subset of X. The *minimal subtree* of $\mathcal{T}$ that connects all the leaves in X' is denoted by $\mathcal{T}(X')$. Moreover, the *restriction of* $\mathcal{T}$ *to* X', denoted by $\mathcal{T}|X'$, is the phylogenetic X'-tree obtained from $\mathcal{T}(X')$ by suppressing all degree-two vertices apart from the root. Lastly, for two phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, we say that $\mathcal{T}'$ is a *refinement* of $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by contracting a possibly empty set of internal edges in $\mathcal{T}'$. In addition, $\mathcal{T}'$ is a *binary refinement* of $\mathcal{T}$ if $\mathcal{T}'$ is binary.

Let $\mathcal{T}$ be a phylogenetic X'-tree. A phylogenetic network $\mathcal{N}$ on X with $X' \subseteq X$ *displays* $\mathcal{T}$ if, up to suppressing vertices with in-degree one and out-degree one, there exists a binary refinement of $\mathcal{T}$ that can be obtained from $\mathcal{N}$ by deleting edges, leaves not in X', and any resulting vertices of out-degree zero, in which case we call the resulting acyclic digraph an *embedding* of $\mathcal{T}$ in $\mathcal{N}$. If $\mathcal{P}$ is a collection of phylogenetic X-trees, then $\mathcal{N}$ *displays* $\mathcal{P}$ if each tree in $\mathcal{P}$ is displayed by $\mathcal{N}$. For example, the two phylogenetic networks at the bottom of Figure 1 both display each of the four trees shown in the top part of the same figure.

Let $\mathcal{N}$ be a phylogenetic network with vertex set V and root ρ. The *hybridisation number* of $\mathcal{N}$, denoted $h(\mathcal{N})$, is the value

$$h(\mathcal{N}) = \sum_{v \in V - \{\rho\}} \left(d^-(v) - 1\right),$$

where $d^-(v)$ denotes the in-degree of v. For example, the phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ that are shown in Figure 1 have hybridisation number 3 and 4, respectively. Observe that each tree vertex and each leaf contributes zero to this sum, but each reticulation v contributes $d^-(v) - 1$. Furthermore, for a set $\mathcal{P}$ of phylogenetic X-trees, we denote by $h_{\text{tc}}(\mathcal{P})$ and $h(\mathcal{P})$, respectively, the values

$$\min\{h(\mathcal{N}) : \mathcal{N} \text{ is a tree-child network on } X \text{ that displays } \mathcal{P}\}$$

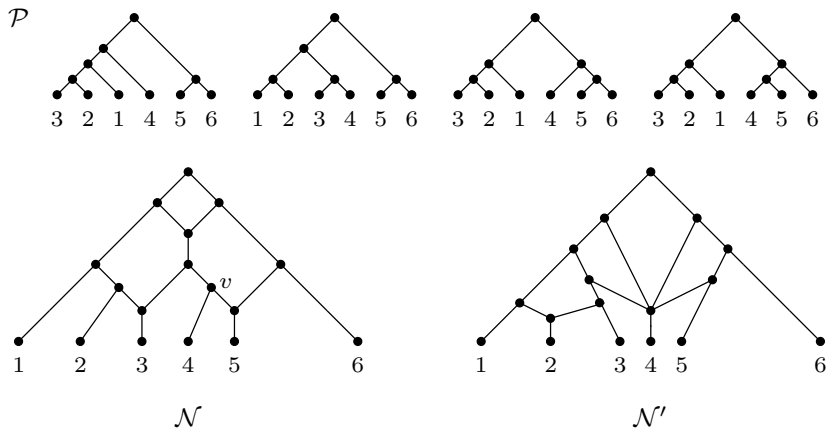

Figure 1: Top: A set $\mathcal{P}$ of four phylogenetic X-trees with $X = \{1, 2, \ldots, 6\}$. Bottom: Two tree-child networks displaying $\mathcal{P}$ with $h(\mathcal{N}) = 3$ and $h(\mathcal{N}') = 4$.

and

$$\min\{h(\mathcal{N}) : \mathcal{N} \text{ is a phylogenetic network on } X \text{ that displays } \mathcal{P}\}.$$

Remark. While the above definition of a phylogenetic network is restricted to networks whose tree vertices have out-degree exactly two, we note that the results in this paper also hold for networks with tree vertices whose out-degree is at least two. More particularly, if a set $\mathcal{P}$ of phylogenetic X-trees is displayed by a phylogenetic network $\mathcal{N}$ whose tree vertices have out-degree at least two, then, by "refining" such vertices, we can obtain a phylogenetic network $\mathcal{N}'$ whose tree vertices have out-degree exactly two, displays $\mathcal{P}$, and $h(\mathcal{N}') = h(\mathcal{N})$. Thus no generality is lost with this restriction.

We next formally state the two decision problems that this paper is centred around.

Minimum Tree-Child Hybridisation
Instance. A set $\mathcal{P}$ of phylogenetic X-trees and a positive integer k.
Question. Does there exist a tree-child network $\mathcal{N}$ on X that displays $\mathcal{P}$ such that $h(\mathcal{N}) \leq k$?

Minimum Hybridisation
Instance. A set $\mathcal{P}$ of phylogenetic X-trees and a positive integer k.

Question. Does there exist a phylogenetic network $\mathcal{N}$ on X that displays $\mathcal{P}$ such that $h(\mathcal{N}) \leq k$?

We will see at the end of this section that, for any given set $\mathcal{P}$ of phylogenetic X-trees, Minimum Tree-Child Hybridisation has a solution, i.e. there exists a tree-child network that displays $\mathcal{P}$.

It was shown in [3] that Minimum Hybridisation is NP-hard, even for when $\mathcal{P}$ consists of two rooted binary phylogenetic X-trees. To see that Minimum Tree-Child Hybridisation is also computationally hard, we again consider this restricted version of the problem and recall the following observation that was first mentioned in [12] and can be derived by slightly modifying the proof of [2, Theorem 2].

Observation 1.1. *Let $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$ be a collection of two binary phylogenetic X-trees. If there exists a phylogenetic network $\mathcal{N}$ that displays $\mathcal{P}$ with $h(\mathcal{N}) = k$, then there also exists a tree-child network $\mathcal{N}'$ that displays $\mathcal{P}$ with $h(\mathcal{N}') \leq k$.*

The next theorem, whose straightforward proof is omitted, follows from Observation 1.1 and the fact that, given a tree-child network $\mathcal{N}$ and a binary phylogenetic tree $\mathcal{T}$, it can be checked in polynomial time whether or not $\mathcal{N}$ displays $\mathcal{T}$ [14, 21].

Theorem 1.2. *The decision problem* Minimum-Tree-Child Hybridisation *is* NP*-complete.*

We end this section by showing that every collection of phylogenetic X-trees can be displayed by a tree-child network on X. For $n = 2$, let $\mathcal{U}_2$ be the unique binary phylogenetic tree on two leaves, x_1 and x_2 say. Now, for a positive integer $n > 2$, obtain $\mathcal{U}_n$ from $\mathcal{U}_{n-1}$ as follows. Viewing the root ρ of $\mathcal{U}_{n-1}$ as a vertex of in-degree zero and out-degree one adjoined to the original root, add an edge that joins a new vertex v and a new leaf x_n and, for each tree edge e in $\mathcal{U}_{n-1}$, subdivide e with a vertex u_e, and add the edge (u_e, v). The resulting phylogenetic network without viewing the root as a vertex of in-degree zero and out-degree one is $\mathcal{U}_n$.

Theorem 1.3. *Let $\mathcal{U}_n$ be the universal network on $X = \{x_1, x_2, \ldots, x_n\}$ with $n \geq 2$. Then $\mathcal{U}_n$ is tree-child and displays all binary phylogenetic X-trees.*

PROOF. By construction of $\mathcal{U}_n$ from $\mathcal{U}_{n-1}$ it is straightforward to check that, as $\mathcal{U}_2$ is tree-child, $\mathcal{U}_n$ is tree-child. To see that $\mathcal{U}_n$ displays all binary phylogenetic X-trees, we use induction on n. Clearly, $\mathcal{U}_2$ displays the unique binary phylogenetic tree on two leaves. For $n \geq 3$, assume that the universal network $\mathcal{U}_{n-1}$ on $X' = \{x_1, x_2, \ldots, x_{n-1}\}$ displays all binary phylogenetic X'-trees. Observe that $\mathcal{U}_{n-1}$ can be obtained from $\mathcal{U}_n$ by deleting x_n, the parent of x_n and all their incident edges, and suppressing all resulting vertices with in-degree one and out-degree one. Now, let $\mathcal{T}_n$ be a binary phylogenetic X-tree, and let $\mathcal{T}_{n-1}$ be $\mathcal{T}_n|X'$. Furthermore, let $\mathcal{C}$ be the subset of X' that consists of the descendant leaves of the parent of x_n in $\mathcal{T}_n$. As $\mathcal{U}_{n-1}$ displays $\mathcal{T}_{n-1}$, there exist an embedding $\mathcal{E}$ of $\mathcal{T}_{n-1}$ in $\mathcal{U}_{n-1}$ and an edge (u, v) in $\mathcal{E}$ such that the set of descendants of v in $\mathcal{E}$ is precisely $\mathcal{C}$. If (u, v) is a tree edge in $\mathcal{U}_{n-1}$, then it is easily checked that $\mathcal{U}_n$ displays $\mathcal{T}_n$ by construction. On the other hand, if (u, v) is a reticulation edge in $\mathcal{U}_{n-1}$, then v has out-degree one in $\mathcal{E}$. Let (v, w) be the unique edge in $\mathcal{E}$ that is directed out of v. Note that, as $\mathcal{U}_{n-1}$ is tree-child, w is a tree vertex in $\mathcal{U}_{n-1}$. Then, as (v, w) is a tree edge in $\mathcal{U}_{n-1}$ that is subdivided by a new vertex in the construction of $\mathcal{U}_n$ from $\mathcal{U}_{n-1}$, it again follows that $\mathcal{U}_n$ displays $\mathcal{T}_n$. This completes the proof of the theorem. □

The next corollary is an immediate consequence of Theorem 1.3 and the fact that every phylogenetic tree has a binary refinement on the same leaf set.

Corollary 1.4. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. There exists a tree-child network on X that displays $\mathcal{P}$.*

While every collection of phylogenetic X-trees can be displayed by a tree-child network on X, a simple counting argument shows that the analogous result is not true for *binary* tree-child networks. Specifically, a binary tree-child network on X has at most $|X| - 1$ reticulations [6, Proposition 1] and so displays at most $2^{|X|-1}$ distinct binary phylogenetic X-trees. But for large enough X, there are many more distinct binary phylogenetic X-trees than $2^{|X|-1}$. For related results, we refer the interested reader to [21].

2. Cherry-picking characterisations

In this section, we state the two cherry-picking characterisations whose proofs are given in the next section. Let $\mathcal{T}$ be a phylogenetic X-tree with

root ρ, where $|X| \geq 2$. If x is a leaf of $\mathcal{T}$, we denote by $\mathcal{T}\backslash x$ the operation of deleting x and its incident edge and, if the parent of x in $\mathcal{T}$ has out-degree two, suppressing the resulting degree-two vertex. Note that if the parent of x is ρ and ρ has out-degree two, then $\mathcal{T}\backslash x$ denotes the operation of deleting x and its incident edge, and then deleting ρ and its incident edge. Observe that $\mathcal{T}\backslash x$ is a phylogenetic tree on $X - \{x\}$. A 2-element subset $\{x, y\}$ of X is a *cherry* of $\mathcal{T}$ if x and y have the same parent. Clearly, every phylogenetic tree with at least two leaves contains a cherry. In this paper, we typically distinguish the leaves in a cherry, in which case we write $\{x, y\}$ as the ordered pair (x, y) depending on the roles of x and y.

Let $\mathcal{T}$ be a phylogenetic X-tree and let (x, y) be an ordered pair of leaves in X. If (x, y) is a cherry of $\mathcal{T}$, then let $\mathcal{T}' = \mathcal{T}\backslash x$; otherwise, let $\mathcal{T}' = \mathcal{T}$. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by *cherry picking* (x, y). Now, let $\mathcal{P}$ be a set of phylogenetic X-trees, and let

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s), (x_{s+1}, -)$$

be a sequence of ordered pairs in $X \times (X \cup \{-\})$ such that the following property is satisfied.

(P) For all $i \in \{1, 2, \ldots, s\}$, we have $x_i \notin \{y_{i+1}, y_{i+2}, \ldots, y_s\}$.

Setting $\mathcal{P}_0 = \mathcal{P}$ and, for all $i \in \{1, 2, \ldots, s\}$, setting $\mathcal{P}_i$ to be the set of phylogenetic trees obtained from $\mathcal{P}_{i-1}$ by cherry picking (x_i, y_i) in each tree in $\mathcal{P}_{i-1}$, we call σ a *cherry-picking sequence of* $\mathcal{P}$ if each tree in $\mathcal{P}_s$ consists of the single vertex x_{s+1}. Furthermore, for all $i \in \{1, 2, \ldots, s\}$, we say that $\mathcal{P}_i$ is obtained from $\mathcal{P}$ by *picking* $x_1, x_2, \ldots, x_i$. Additionally, if $\mathcal{P}_i \neq \mathcal{P}_{i+1}$, then we refer to (x_i, y_i) as being *essential*. Moreover, if σ is a cherry-picking sequence for $\mathcal{P}$, then the *weight* of σ, denoted $w(\sigma)$, is the value $s + 1 - |X|$. Observe that, if σ is a cherry-picking sequence of $\mathcal{P}$, then

$$s + 1 - |X| \geq 0$$

as each element in X must appear as the first element in an ordered pair in σ.

Now, let σ be a cherry-picking sequence for $\mathcal{P}$. We call σ a *minimum cherry-picking sequence* of $\mathcal{P}$ if $w(\sigma)$ is of smallest value over all cherry-picking sequences of $\mathcal{P}$. This smallest value is denoted by $s(\mathcal{P})$. It will follow from the results in the next section (Lemma 3.4) that every collection $\mathcal{P}$ of

phylogenetic trees has a cherry-picking sequence and so $s(\mathcal{P})$ is well defined. Referring to Figure 1,

$$\sigma = (3,2),(3,4),(5,6),(5,4),(1,2),(4,2),(4,6),(2,6),(6,-)$$

is a cherry-picking sequence with weight $w(\sigma) = 9 - 6 = 3$ for the four trees shown at the top of this figure.

Remark. As noted in the introduction, cherry-picking sequences were introduced in [12]. In the set-up of this paper, the difference is as follows. Instead of a cherry-picking sequence consisting of a set of ordered pairs, a cherry-picking sequence in [12] consists of an ordering of the elements in X. Moreover, this ordering has the additional property that, for each $i \in \{1, 2, \ldots, s\}$, x_i is part of a cherry of *every* tree in $\mathcal{P}_{i-1}$. Subsequently, x_i is deleted from each tree in $\mathcal{P}_{i-1}$, and the iterative process continues. The weighting of such a sequence is based, across all i, on the number of different cherries of which x_i is part of. It is not difficult to see how this could be interpreted as a special type of cherry-picking sequence as defined in this paper.

The first of our new characterisations is the next theorem. For a given set $\mathcal{P}$ of phylogenetic X-trees, it writes $h_{\mathrm{tc}}(\mathcal{P})$ in terms of cherry-picking sequences for $\mathcal{P}$.

Theorem 2.1. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Then*

$$h_{\mathrm{tc}}(\mathcal{P}) = s(\mathcal{P}).$$

To state the second characterisation, we require an additional concept. Let $\mathcal{T}$ be a phylogenetic X-tree. Consider the operation of adjoining a new leaf z to $\mathcal{T}$ in one of the following three ways.

(i) Subdivide an edge of $\mathcal{T}$ with a new vertex, u say, and add the edge (u, z).

(ii) View the root ρ of $\mathcal{T}$ as a degree-one vertex adjacent to the original root and add the edge (ρ, z).

(iii) Add the edge (v, z), where v is an interior vertex of $\mathcal{T}$.

We refer to this operation as *attaching a new leaf* z to $\mathcal{T}$. More generally, if Z is a finite set of elements such that $X \cap Z$ is empty, then *attaching* Z to $\mathcal{T}$ is the operation of attaching, in turn, each element in Z to $\mathcal{T}$ to eventually obtain a phylogenetic tree on $X \cup Z$. We refer to Z as a set of *auxiliary leaves*. Lastly, *attaching* Z to a set $\mathcal{P}$ of phylogenetic X-trees is the operation of attaching Z to each tree in $\mathcal{P}$.

Let $\mathcal{P}$ be a set of phylogenetic X-trees. A sequence

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s), (x_{s+1}, -)$$

of ordered pairs in $(X \cup Z) \times (X \cup Z \cup \{-\})$ that satisfies (P) is a *leaf-added cherry-picking sequence* for $\mathcal{P}$ if it is a cherry-picking sequence of a set of phylogenetic trees obtained from $\mathcal{P}$ by attaching Z. As for cherry-picking sequences, the *weight* of σ, denoted $w(\sigma)$, is the value $s+1-(|X|+|Z|)$. We denote the minimum weight amongst all leaf-added cherry-picking sequences of $\mathcal{P}$ by $s_+(\mathcal{P})$. Of course, $s_+(\mathcal{P}) \leq s(\mathcal{P})$, but this inequality can also be strict. To illustrate, consider the two sets $\mathcal{P}$ and $\mathcal{P}'$ of phylogenetic trees shown in Figure 2. Now

$$\begin{aligned} \sigma &= (4,5), (4,1), (4,3), (5,6), (5,3), (5,8), \\ &\quad (2,3), (3,1), (6,7), (7,8), (1,8), (8,-) \end{aligned}$$

is a cherry-picking sequence for $\mathcal{P}$ of weight $w(\sigma) = 12 - 8 = 4$. In fact, it follows from [13, 15] that $h_{\text{tc}}(\mathcal{P}) = 4$ (see Section 6 for details). On the other hand,

$$\begin{aligned} \sigma' &= (5,z), (5,8), (4,z), (4,1), (z,3), (z,6), \\ &\quad (2,3), (3,1), (6,7), (7,8), (1,8), (8,-) \end{aligned}$$

is a cherry-picking sequence for $\mathcal{P}'$ of weight $w(\sigma') = 12 - 9 = 3$. Since $\mathcal{P}'$ can be obtained by attaching z to $\mathcal{P}$, it follows that σ' is a leaf-added cherry-picking sequence for $\mathcal{P}$ and $s_+(\mathcal{P}) \leq 3$.

For a given set $\mathcal{P}$ of phylogenetic X-trees, the next theorem characterises $h(\mathcal{P})$ in terms of leaf-added cherry-picking sequences.

Theorem 2.2. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Then*

$$h(\mathcal{P}) = s_+(\mathcal{P}).$$

It is worth noting that, for a set $\mathcal{P}$ of phylogenetic X-trees, it follows from Theorems 2.1 and 2.2 that $h_{\text{tc}}(\mathcal{P})$ and $h(\mathcal{P})$ can be determined without constructing a phylogenetic network.

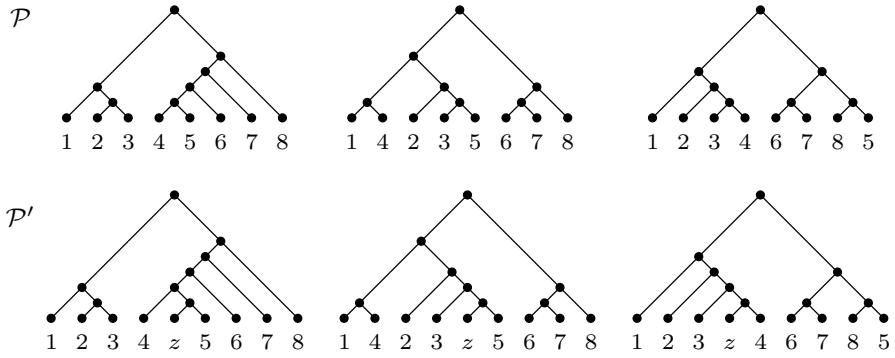

Figure 2: Two sets $\mathcal{P}$ and $\mathcal{P}'$ of phylogenetic trees, where $\mathcal{P}'$ is obtained by attaching z to $\mathcal{P}$. (In parts, adapted from [13, Figure 1].)

3. Proofs of Theorems 2.1 and 2.2

In this section, we prove Theorems 2.1 and 2.2. Most of the work is in proving Theorem 2.1. We begin by showing that $h_{\mathrm{tc}}(\mathcal{P}) \leq s(\mathcal{P})$.

Lemma 3.1. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Let σ be a cherry-picking sequence for $\mathcal{P}$. Then there exists a tree-child network $\mathcal{N}$ on X that displays $\mathcal{P}$ with $h(\mathcal{N}) \leq w(\sigma)$ satisfying the following properties:*

(i) *If u is a tree vertex in $\mathcal{N}$ and not a parent of a reticulation, then there are leaves ℓ_1 and ℓ_2 at the end of tree paths starting at the children v_1 and v_2 of u, respectively, such that (ℓ_1, ℓ_2) is an element in σ.*

(ii) *If u is a tree vertex in $\mathcal{N}$ and a parent of a reticulation v, then there are leaves ℓ_u and ℓ_v at the end of tree paths starting at u and v, respectively, such that (ℓ_v, ℓ_u) is an element in σ.*

Proof. Let

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s), (x_{s+1}, -)$$

be a cherry-picking sequence for $\mathcal{P}$. The proof is by induction on s. If $s = 0$, then $|X| = 1$ and each tree in $\mathcal{P}$ consists of the single vertex in X. It immediately follows that choosing $\mathcal{N}$ to be the phylogenetic network consisting of the single vertex in X establishes the lemma for $s = 0$.

Now suppose that $s \geq 1$, and that the lemma holds for all cherry-picking sequences for sets of phylogenetic trees on the same leaf set whose length is at most s. Let

$$\sigma' = (x_2, y_2), (x_3, y_3), \ldots, (x_s, y_s), (x_{s+1}, -),$$

and let $\mathcal{P}'$ be the set of phylogenetic trees obtained from $\mathcal{P}$ by picking x_1.

First assume that each tree in $\mathcal{P}'$ has the same leaf set, namely $X' = X - \{x_1\}$. Then σ' is a cherry-picking sequence for $\mathcal{P}'$. By induction, there is a tree-child network $\mathcal{N}'$ on X' that displays $\mathcal{P}'$ with $h(\mathcal{N}') \leq w(\sigma')$ and satisfies (i) and (ii). Since each tree in $\mathcal{P}'$ has the same leaf set, $\{x_1, y_1\}$ is a cherry in each tree in $\mathcal{P}$. Therefore, as $\mathcal{N}'$ displays a binary refinement of each tree in $\mathcal{P}'$, the tree-child network obtained from $\mathcal{N}'$ by subdividing the edge directed into y_1 with a new vertex u and adding the edge (u, x_1) displays $\mathcal{P}$. Furthermore, as $h(\mathcal{N}') \leq w(\sigma')$ and $\mathcal{N}'$ satisfies (i) and (ii) relative to σ', we have $h(\mathcal{N}) = h(\mathcal{N}') \leq w(\sigma') = w(\sigma)$ and it is easily seen that $\mathcal{N}$ satisfies (i) and (ii) relative to σ.

Now assume that not every tree in $\mathcal{P}'$ has the same leaf set. Let $\mathcal{P}'_1$ denote the subset of trees in $\mathcal{P}'$ whose leaf set is $X - \{x_1\}$. Since $\mathcal{P}' - \mathcal{P}'_1$ is non-empty, there exists some i with $i \in \{2, 3, \ldots, s+1\}$ such that $x_i = x_1$. Note that $(x_1, -)$ is not in σ; otherwise there is an ordered pair in σ whose second coordinate is x_1 and so σ is not a cherry-picking sequence for $\mathcal{P}$. Let (x_1, y_i) be the first ordered pair in σ' whose first coordinate is x_1. Let $\mathcal{T}_1$ be a tree in $\mathcal{P}'_1$. Consider the process of picking, in order, $(x_2, y_2), (x_3, y_3), \ldots, (x_{i-1}, y_{i-1})$ from $\mathcal{T}_1$. Let X_1 denote the subset of leaves in $X - \{x_1\}$ that are deleted from $\mathcal{T}_1$ in this process. Observe that, as y_i is the second coordinate in (x_i, y_i), we have $y_i \notin X_1$.

We next add x_1 to $\mathcal{T}_1$ to obtain a phylogenetic X-tree for which σ' is a cherry-picking sequence. Let w be the (unique) vertex of $\mathcal{T}_1$ that is closest to the root with the property that y_i is a descendant leaf of w, and the child of w on the path from w to y_i has all its descendant leaves in $X_1 \cup \{y_i\}$. Let $\mathcal{T}'_1$ be the phylogenetic X-tree obtained from $\mathcal{T}_1$ by adding the edge (w, x_1). We now show that σ' is a cherry-picking sequence for $\mathcal{T}'_1$. Suppose that σ' is not a cherry-picking sequence for $\mathcal{T}'_1$. Let u be the parent of w in $\mathcal{T}'_1$. Then amongst the first $i-2$ ordered pairs in σ' is an ordered pair of the form (x_j, y_i) that is essential when, the ordered pairs in σ' are (in order) picked from $\mathcal{T}_1$, where x_j is a descendant leaf of u in $\mathcal{T}'_1$. But then, each descendant leaf of w is in $X_1 \cup \{y_i\}$, contradicting the choice of w.

Repeating this placement of x_1 for each tree in $\mathcal{P}'_1$, we obtain a set $\mathcal{P}''_1$ of phylogenetic X-trees from $\mathcal{P}'_1$. Let $\mathcal{P}'' = \mathcal{P}''_1 \cup (\mathcal{P}' - \mathcal{P}'_1)$ and observe that σ' is a cherry-picking sequence for $\mathcal{P}''$. Therefore, by induction, there is a tree-child network $\mathcal{N}'$ on X that displays $\mathcal{P}''$ with $h(\mathcal{N}') \leq w(\sigma')$ and satisfies (i) and (ii).

Let p denote the parent of x_1 in $\mathcal{N}'$. If p is a reticulation, let $\mathcal{N}$ be the phylogenetic network obtained from $\mathcal{N}'$ by subdividing the edge directed into y_1 with a new vertex u and adding the edge (u, p). Since $\mathcal{N}'$ is tree-child and displays $\mathcal{P}'$, it follows that $\mathcal{N}$ is tree-child and displays $\mathcal{P}$. Furthermore,

$$h(\mathcal{N}) = h(\mathcal{N}') + 1 \leq w(\sigma') + 1 = w(\sigma).$$

Additionally, as $(x_1, y_1) \in \sigma$, it also follows that, as $\mathcal{N}'$ satisfies (i) and (ii) relative to σ', we have $\mathcal{N}$ satisfies (i) and (ii) relative to σ.

Thus we may assume that p is a tree vertex. Let w denote the child of p that is not x_1 in $\mathcal{N}'$. If w is a reticulation, then, as $\mathcal{N}'$ satisfies (ii), σ' contains a cherry in which x_1 is the second coordinate. But (x_1, y_1) is the first ordered pair in σ and so, as σ satisfies (P), x_1 is never the second coordinate in an ordered pair in σ; a contradiction. Therefore w is either a tree vertex or a leaf in $\mathcal{N}'$. So, as $\mathcal{N}'$ satisfies (i) and no ordered pair has x_1 as the second coordinate, it follows that σ' contains an ordered pair, (x_1, y_j) say, where y_j is the leaf at the end of a tree path in $\mathcal{N}'$ starting at w. Now let $\mathcal{N}$ be the phylogenetic network obtained from $\mathcal{N}'$ by subdividing the edges directed into y_1 and x_1 with new vertices u and v, respectively, and adding the edge (u, v). Since $\mathcal{N}'$ is tree-child and $h(\mathcal{N}') \leq w(\sigma')$, it is easily seen that $\mathcal{N}$ is tree-child and $h(\mathcal{N}) = h(\mathcal{N}') + 1 \leq w(\sigma') + 1 = w(\sigma)$. Furthermore, $\mathcal{N}'$ displays $\mathcal{P}' - \mathcal{P}'_1$ as well as $\mathcal{P}''_1$, and therefore $\mathcal{P}'_1 | (X - \{x_1\})$. Thus $\mathcal{N}$ displays $\mathcal{P}$. To see that $\mathcal{N}$ satisfies (i) and (ii) relative to σ, it suffices to show that $\mathcal{N}$ satisfies (ii) for p and u. Indeed, the two ordered pairs (x_1, y_j) and (x_1, y_1) in σ verify (ii) for p and u, respectively. This completes the proof of the lemma. □

The next corollary immediately follows from Lemma 3.1.

Corollary 3.2. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Then $h_{\mathrm{tc}}(\mathcal{P}) \leq s(\mathcal{P})$.*

For the proof of the converse of Corollary 3.2, we begin with an additional lemma. Let $\mathcal{N}$ be a phylogenetic network, and let x and y be two leaves in $\mathcal{N}$.

Generalising cherries to phylogenetic networks, we say that $\{x, y\}$ is a *cherry* in $\mathcal{N}$ if x and y have a common parent. Moreover, we call $\{x, y\}$ a *reticulated cherry* if the parent of x, say p_x, and the parent of y, say p_y, are joined by a reticulation edge (p_y, p_x) in which case we say that x is the *reticulation leaf* relative to $\{x, y\}$. We next define two operations on $\mathcal{N}$. First, *reducing a cherry* $\{x, y\}$ is the operation of deleting one of the two leaves in $\{x, y\}$, and suppressing the resulting degree-two vertex. Second, *reducing a reticulated cherry* $\{x, y\}$ is the operation of deleting the reticulation edge joining the parents of x and y and suppressing any resulting degree-two vertices. The proof of the next lemma is similar to the analogous result for binary tree-child networks [5, Lemma 4.1] and is omitted.

Lemma 3.3. *Let $\mathcal{N}$ be a tree-child network on X. Then the following hold.*

(i) *If $|X| \geq 2$, then $\mathcal{N}$ contains either a cherry or a reticulated cherry.*

(ii) *If $\mathcal{N}'$ is obtained from $\mathcal{N}$ by reducing either a cherry or a reticulated cherry, then $\mathcal{N}'$ is a tree-child network.*

Lemma 3.4. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Then $h_{\mathrm{tc}}(\mathcal{P}) \geq s(\mathcal{P})$.*

Proof. Let $\mathcal{N}$ be a tree-child network on X that displays $\mathcal{P}$. By Corollary 1.4, such a network exists. We establish the lemma by explicitly constructing a cherry-picking sequence σ for $\mathcal{P}$ such that $w(\sigma) \leq h(\mathcal{N})$.

Let ρ denote the root of $\mathcal{N}$, and let $v_1, v_2, \ldots, v_r$ denote the reticulations of $\mathcal{N}$. Let $\ell_\rho, \ell_1, \ell_2, \ldots, \ell_r$ denote the leaves at the end of tree paths $P_\rho, P_1, P_2, \ldots, P_r$ in $\mathcal{N}$ starting at $\rho, v_1, v_2, \ldots, v_r$, respectively. Observe that these paths are pairwise vertex disjoint. We now construct a sequence of ordered pairs as follows:

Step 1. Set $\mathcal{N} = \mathcal{N}_0$ and σ_0 to be the empty sequence. Set $i = 1$.

Step 2. If $\mathcal{N}_{i-1}$ consists of a single vertex x_i, then set σ_i to be the concatenation of σ_{i-1} and $(x_i, -)$, and return σ_i.

Step 3. If $\{x_i, y_i\}$ is a cherry in $\mathcal{N}_{i-1}$, then

(a) If one of x_i and y_i, say x_i, equates to ℓ_j for some $j \in \{1, 2, \ldots, r\}$ and v_j is not a reticulation in $\mathcal{N}_{i-1}$, then set σ_i to be the concatenation of σ_{i-1} and (x_i, y_i).

(b) Otherwise, set σ_i to be the concatenation of σ_{i-1} and (x_i, y_i), where $x_i \notin \{\ell_\rho, \ell_1, \ell_2, \ldots, \ell_r\}$.

(c) Set $\mathcal{N}_i$ to be the tree-child network obtained from $\mathcal{N}_{i-1}$ by deleting x_i, thereby reducing the cherry $\{x_i, y_i\}$.

(d) Increase i by one and go to Step 2.

Step 4. Else, there is a reticulated cherry $\{x_i, y_i\}$ in $\mathcal{N}_{i-1}$, where x_i say is the reticulation leaf.

(a) Set σ_i to be the concatenation of σ_{i-1} and (x_i, y_i).

(b) Set $\mathcal{N}_i$ to be the tree-child network obtained from $\mathcal{N}_{i-1}$ by reducing the reticulated cherry $\{x_i, y_i\}$.

(c) Increase i by one and go to Step 2.

First note that it is easily checked that the construction is well defined, that is, it returns a sequence of ordered pairs. Moreover, in each iteration i of the above construction, it follows from Lemma 3.3 that $\mathcal{N}_i$ is tree-child. We next show that, if $\{x_i, y_i\}$ is a cherry in $\mathcal{N}_{i-1}$, and x_i and y_i equate to ℓ_j and $\ell_{j'}$, respectively, where ℓ_j and $\ell_{j'}$ are elements in $\{\ell_1, \ell_2, \ldots, \ell_r\}$, then exactly one of v_j and $v_{j'}$ is a reticulation in $\mathcal{N}_{i-1}$. To see this, if v_j and $v_{j'}$ are both reticulations in $\mathcal{N}_{i-1}$, then P_j and $P_{j'}$ are not vertex disjoint in $\mathcal{N}$; a contradiction. On the other hand, suppose neither v_j and $v_{j'}$ are reticulations in $\mathcal{N}_{i-1}$. Without loss of generality, we may assume $\{x_i, y_i\}$ is the first such cherry for which this holds. Since $\mathcal{N}$ is tree-child, and therefore has no tree vertex that is the parent of two reticulations, there is an iteration $i' < i$, in which the cherry $(x_{i'}, y_{i'})$ is concatenated with $\sigma_{i'-1}$, where $y_{i'} \in \{x_i, y_i\}$, and $\mathcal{N}_{i'}$ has $\{x_i, y_i\}$ as a cherry but $\mathcal{N}_{i'-1}$ does not. If $x_{i'} = \ell_\rho$ or $x_{i'} \in \{\ell_1, \ell_2, \ldots, \ell_r\}$, we contradict the construction by the choice of $\{x_i, y_i\}$. Also, if $x_{i'} \notin \{\ell_\rho, \ell_1, \ell_2, \ldots, \ell_r\}$, then we again contradict the construction. Hence, we may assume for the remainder of the proof that exactly one of v_j and $v_{j'}$ is a reticulation in $\mathcal{N}_{i-1}$.

Let

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_{i-1}, y_{i-1}), (x_i, -)$$

be the sequence returned by the construction. We prove by induction on i that σ is a cherry-picking sequence for $\mathcal{P}$ whose weight is at most $h(\mathcal{N})$.

If $i = 1$, then $\mathcal{N}$ consists of the single vertex in X and the construction correctly returns such a sequence.

Now suppose that $i \geq 2$, and consider the first iteration of the construction. Either $\{x_1, y_1\}$ is a cherry or a reticulated cherry of $\mathcal{N}_0$. If $\{x_1, y_1\}$ is a cherry, then $\{x_1, y_1\}$ is a cherry of each tree in $\mathcal{P}$. In this instance, let $\mathcal{P}'$ denote the set of phylogenetic X'-trees obtained from $\mathcal{P}$ by picking x_1, where $X' = X - \{x_1\}$. Observe that $\mathcal{N}_1$ is a tree-child network on X' that displays $\mathcal{P}'$.

Now assume that $\{x_1, y_1\}$ is a reticulated cherry with x_1 as the reticulation leaf. Let $\mathcal{P}_1$ be the subset of trees in $\mathcal{P}$ not displayed by $\mathcal{N}_1$ and let $\mathcal{P}_2 = \mathcal{P} - \mathcal{P}_1$. Note that $\{x_1, y_1\}$ is a cherry of each tree in $\mathcal{P}_1$. For each tree in $\mathcal{P}_1$, delete the edge incident with x_1, suppress any resulting degree-two vertex, and reattach x_1 to the rest of the tree containing y_1 by subdividing an edge with a new vertex and adding an edge joining this vertex and x_1 so that the resulting phylogenetic X-tree is displayed by $\mathcal{N}_1$. It is easily seen that this is always possible. Let $\mathcal{P}_1'$ denote the resulting collection of trees obtained from $\mathcal{P}_1$. For this instance, let $\mathcal{P}' = \mathcal{P}_1' \cup \mathcal{P}_2$ and observe that $\mathcal{N}_1$ displays $\mathcal{P}'$.

To complete the induction it suffices to show that if

$$\sigma' = (x_2, y_2), (x_3, y_3), \ldots, (x_{i-1}, y_{i-1}), (x_i, -)$$

is a cherry-picking sequence for $\mathcal{P}'$ whose weight $w(\sigma')$ is at most $h(\mathcal{N}_1)$, then σ is a cherry-picking sequence for $\mathcal{P}$ whose weight $w(\sigma)$ is at most $h(\mathcal{N}_0)$, that is, at most $h(\mathcal{N})$. First assume that $\{x_1, y_1\}$ is a cherry of $\mathcal{N}_0$. Then, as σ' satisfies (P) and $x_1 \notin \mathcal{L}(\mathcal{N}_1)$, it follows that σ also satisfies (P) and so σ is a cherry-picking sequence for $\mathcal{P}$. Since x_1 only appears once as the first coordinate of an ordered pair in σ, we have

$$w(\sigma) = w(\sigma') \leq h(\mathcal{N}_1) = h(\mathcal{N}_0).$$

Now assume that $\{x_1, y_1\}$ is a reticulated cherry of $\mathcal{N}_0$ with x_1 as the reticulation leaf. Without loss of generality, let v_1 denote the associated reticulation, so that $x_1 = \ell_1$. We next show that σ satisfies (P). If the in-degree of v_1 is at least three in $\mathcal{N}_0$, then v_1 exists in $\mathcal{N}_1$ and so, by construction, x_1 does not appear as the second coordinate of an ordered pair in σ' as well as in σ. Therefore, if the in-degree of v_1 is at least three in $\mathcal{N}_0$, then σ satisfies (P).

Now suppose that the in-degree of v_1 is two in $\mathcal{N}_0$. To establish that σ satisfies (P), assume to the contrary that x_1 appears as the second coordinate of an ordered pair in σ'. Let (z, x_1) denote the first such ordered pair. Then, at some iteration j, either $\{z, x_1\}$ is a cherry or a reticulated cherry of $\mathcal{N}_{j-1}$. If $\{z, x_1\}$ is a cherry of $\mathcal{N}_{j-1}$, then, since $x_1 = \ell_1$, we are in Step 3(a) in iteration j of the construction and so the ordered pair should be (x_1, z); a contradiction. On the other hand, if $\{z, x_1\}$ is a reticulated cherry of $\mathcal{N}_{j-1}$, then z is the reticulation leaf of $\{z, x_1\}$ and, by construction of σ, one of the parents of v_1 in $\mathcal{N}_0$ is the parent of two reticulations in $\mathcal{N}_0$, namely v_1 and the reticulation for which, by construction, there is a tree path starting at this reticulation and ending at z; a contradiction as $\mathcal{N}_0$ is tree-child. Hence σ satisfies (P). Thus, since each tree in $\mathcal{P}_1$ has $\{x_1, y_1\}$ as a cherry and σ' is a cherry-picking sequence for $\mathcal{P}'$, it follows that σ is a cherry-picking sequence for $\mathcal{P}$. Furthermore, as $w(\sigma) = w(\sigma') + 1$ and $h(\mathcal{N}_0) = h(\mathcal{N}_1) + 1$,

$$w(\sigma) = w(\sigma') + 1 \leq h(\mathcal{N}_1) + 1 = h(\mathcal{N}_0).$$

This completes the proof of the lemma. □

Proof of Theorem 2.1. Combining Corollary 3.2 and Lemma 3.4 establishes the theorem. □

We next establish Theorem 2.2.

Proof of Theorem 2.2. We first show that $h(\mathcal{P}) \leq s_+(\mathcal{P})$. Let $\mathcal{P}'$ be a set of phylogenetic trees obtained from $\mathcal{P}$ by attaching a set Z such that $X \cap Z$ is empty and $s_+(\mathcal{P}) = s(\mathcal{P}')$. It follows by Theorem 2.1 that there is a tree-child network $\mathcal{N}'$ on $X \cup Z$ that displays $\mathcal{P}'$ with $h(\mathcal{N}') = s(\mathcal{P}')$. Observe that $\mathcal{N}'$ displays $\mathcal{P}$. Let $\mathcal{N}$ be the phylogenetic network on X obtained from $\mathcal{N}'$ by deleting every vertex that is not on a directed path from the root to a leaf in X, and suppressing any resulting non-root vertex of degree two. Noting that no deleted vertex is used to display a phylogenetic tree in $\mathcal{P}$, it is easily checked that, up to the root having out-degree one, $\mathcal{N}$ displays $\mathcal{P}$. Furthermore, $h(\mathcal{N}) \leq h(\mathcal{N}')$. Therefore, by Theorem 2.1,

$$h(\mathcal{P}) \leq h(\mathcal{N}) \leq h(\mathcal{N}') = s(\mathcal{P}') = s_+(\mathcal{P}).$$

In particular, $h(\mathcal{P}) \leq s_+(\mathcal{P})$.

To prove the converse, $h(\mathcal{P}) \geq s_+(\mathcal{P})$, let $\mathcal{N}$ be a phylogenetic network on X that displays $\mathcal{P}$ and $h(\mathcal{N}) = h(\mathcal{P})$. Let $\mathcal{N}'$ be the phylogenetic network

obtained by attaching a new leaf to each reticulation edge in $\mathcal{N}$, i.e. for each reticulation edge e, subdivide e with a new vertex u and add a new edge (u, z_e), where $z_e \notin X$. It is easily checked that $\mathcal{N}'$ is tree-child and $h(\mathcal{N}') = h(\mathcal{N})$. Let Z denote the set of new leaves attached to $\mathcal{N}$. For each tree $\mathcal{T}$ in $\mathcal{P}$, let $\mathcal{T}_r$ denote a binary refinement of $\mathcal{T}$ that is displayed by $\mathcal{N}$, and let $\mathcal{T}_r'$ be a binary phylogenetic tree with leaf set $X \cup Z$ that is displayed by $\mathcal{N}'$ and obtained from $\mathcal{T}_r$ by attaching Z. Note that $\mathcal{T}_r'$ is a binary refinement of a tree that can be obtained from $\mathcal{T}$ by attaching Z. Set

$$\mathcal{P}_r' = \{\mathcal{T}_r' : \mathcal{T} \in \mathcal{P}\},$$

and note that $h_{\rm tc}(\mathcal{P}_r') \leq h(\mathcal{N}')$ as $\mathcal{N}'$ is tree-child and displays $\mathcal{P}_r'$. Since each tree in $\mathcal{P}_r'$ is a binary refinement of a tree that can be obtained from a tree in $\mathcal{P}$ by attaching Z, we have $s_+(\mathcal{P}) \leq s(\mathcal{P}_r')$. Thus, by Theorem 2.1,

$$s_+(\mathcal{P}) \leq s(\mathcal{P}_r') = h_{\rm tc}(\mathcal{P}_r') \leq h(\mathcal{N}') = h(\mathcal{N}) = h(\mathcal{P}),$$

and so $s_+(\mathcal{P}) \leq h(\mathcal{P})$. □

We end this section with the pseudocode of an algorithm—called Construct Tree-Child Network—that constructs a tree-child network from a cherry-picking sequence. Specifically, given a cherry-picking sequence σ for a set $\mathcal{P}$ of phylogenetic X-trees, Construct Tree-Child Network returns a tree-child network $\mathcal{N}$ on X that displays $\mathcal{P}$ and $h(\mathcal{N}) \leq w(\sigma)$. This is the same construction as that used to prove Lemma 3.1 and so the proof of its correctness is not given.

Algorithm. Construct Tree-Child Network
Input. A set $\mathcal{P}$ of phylogenetic X-trees, and a cherry-picking sequence

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s), (x_{s+1}, -)$$

for $\mathcal{P}$.
Output. A tree-child network $\mathcal{N}$ on X that displays $\mathcal{P}$ and $h(\mathcal{N}) \leq w(\sigma)$.

Step 1. If $|X| = 1$, set $\mathcal{N}_{s+1}$ to be the phylogenetic network consisting of the single vertex x_{s+1} in X and return $\mathcal{N}_{s+1}$. Otherwise, set $\mathcal{N}_{s+1}$ to be the phylogenetic network consisting of the single edge (ρ, x_{s+1}) and set $i = s$.

Step 2. Depending on which holds, do exactly one of the following three steps.

- **(a)** If $x_i \in \mathcal{L}(\mathcal{N}_{i+1})$ and the parent p_i of x_i is a reticulation in $\mathcal{N}_{i+1}$, then obtain $\mathcal{N}_i$ from $\mathcal{N}_{i+1}$ by subdividing the edge directed into y_i with a new vertex u and adding a new edge (u, p_i).
- **(b)** If $x_i \in \mathcal{L}(\mathcal{N}_{i+1})$ and the parent p_i of x_i is not a reticulation in $\mathcal{N}_{i+1}$, then obtain $\mathcal{N}_i$ from $\mathcal{N}_{i+1}$ by subdividing the edge directed into y_i with a new vertex u, subdividing the edge (p_i, x_i) with a new vertex v, and adding a new edge (u, v).
- **(c)** Else $x_i \notin \mathcal{L}(\mathcal{N}_{i+1})$, and obtain $\mathcal{N}_i$ from $\mathcal{N}_{i+1}$ by subdividing the edge directed into y_i with a new vertex u and adding a new edge (u, x_i).

Step 3. If $i = 1$, then set $\mathcal{N}$ to be the network obtained from $\mathcal{N}_i$ by deleting the unique edge incident with ρ and return $\mathcal{N}$. Otherwise, decrement i by one and go to Step 2.

Now, let σ be a leaf-added cherry-picking sequence for a set $\mathcal{P}$ of phylogenetic X-trees. Then there exists a set $\mathcal{P}'$ of phylogenetic trees on $X \cup Z$ obtained from $\mathcal{P}$ by attaching Z such that σ is a cherry-picking sequence for $\mathcal{P}'$. It is straightforward to check that the network $\mathcal{N}$ on X resulting from calling Construct Tree-Child Network for $\mathcal{P}'$ and σ and, subsequently, restricting to vertices and edges on a path from the root to leaves in X as described in the first direction of the proof of Theorem 2.2 displays $\mathcal{P}$ and $h(\mathcal{N}) \leq w(\sigma)$.

4. Bounding the maximum number of auxiliary leaves

In light of Theorem 2.2, a natural question to ask is how many auxiliary leaves need to be attached to a given set $\mathcal{P}$ of phylogenetic X-trees in order to calculate $h(\mathcal{P})$. Attaching auxiliary leaves to $\mathcal{P}$ is necessary whenever $h(\mathcal{P}) < h_{\mathrm{tc}}(\mathcal{P})$. Here, we provide an upper bound on the number of auxiliary leaves in terms of $h_{\mathrm{tc}}(\mathcal{P})$. We start by introducing two operations that, repeatedly applied, transform any phylogenetic network $\mathcal{N}$ that displays $\mathcal{P}$ into a tree-child network without increasing $h(\mathcal{N})$ and that displays a set of phylogenetic trees obtained from $\mathcal{P}$ by attaching auxiliary leaves.

Let $\mathcal{N}$ be a phylogenetic network on X, and let (u, v) be an edge in $\mathcal{N}$ such that u and v are reticulations. Obtain a phylogenetic network $\mathcal{N}'$ from $\mathcal{N}$ by contracting (u, v) and, for each resulting pair of parallel edges, repeatedly deleting one of the two edges in parallel and suppressing the resulting degree-two vertex. We say that $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by a *contraction*.

Lemma 4.1. *Let $\mathcal{P}$ be a collection of phylogenetic X-trees, and let $\mathcal{N}$ be a phylogenetic network on X that displays $\mathcal{P}$. Let $\mathcal{N}'$ be a phylogenetic network obtained from $\mathcal{N}$ by a contraction. Then $\mathcal{N}'$ displays $\mathcal{P}$ and $h(\mathcal{N}') \leq h(\mathcal{N})$.*

Proof. Let (u, v) be the edge in $\mathcal{N}$ that is incident with two reticulations and contracted in the process of obtaining $\mathcal{N}'$ from $\mathcal{N}$. Furthermore, let w be the vertex in $\mathcal{N}'$ that results from identifying u and v. We have $d^-(w) \leq d^-(u) + d^-(v) - 1$ while all other reticulations w' in $\mathcal{N}'$ correspond to a reticulation u' in $\mathcal{N}$ and $d^-(w') = d^-(u')$. It now follows that $h(\mathcal{N}') \leq h(\mathcal{N})$.

Now, let $\mathcal{T}$ be a tree in $\mathcal{P}$. If v is not used to display $\mathcal{T}$ in $\mathcal{N}$, then u is also not used to display $\mathcal{T}$ in $\mathcal{N}$, and it is easily seen that $\mathcal{N}'$ displays $\mathcal{T}$ without using w. On the other hand, if v is used to display $\mathcal{T}$ in $\mathcal{N}$, then exactly one parent, t say, of v is used to display $\mathcal{T}$ in $\mathcal{N}$. If $t \neq u$, it is clear that $\mathcal{N}'$ displays $\mathcal{P}$. Furthermore, if $t = u$, then exactly one parent of u, say s, is used to display $\mathcal{T}$ in $\mathcal{N}$. Now, regardless of whether or not s is also a parent of v in which case s is suppressed in obtaining $\mathcal{N}'$ from $\mathcal{N}$, it again follows that $\mathcal{N}'$ displays $\mathcal{T}$. Hence $\mathcal{N}'$ displays each tree in $\mathcal{P}$ and the lemma follows. □

We call a phylogenetic network with no edges whose end vertices are both reticulations *stack free*. It follows from repeated applications of Lemma 4.1 that if $\mathcal{P}$ is a collection of phylogenetic X-trees, then there is a stack-free network $\mathcal{N}$ on X that displays $\mathcal{P}$ such that $h(\mathcal{N}) = h(\mathcal{P})$.

For the second operation, let $\mathcal{N}$ be a phylogenetic network on X, and let u be a tree vertex in $\mathcal{N}$ whose two children are both reticulations. Furthermore, let (u, v) be a reticulation edge, and let $z \notin X$. Obtain a phylogenetic network $\mathcal{N}'$ on $X \cup \{z\}$ from $\mathcal{N}$ by subdividing the edge (u, v) with a new vertex w and adding a new edge (w, z). We say that $\mathcal{N}'$ has been obtained from $\mathcal{N}$ by a *leaf-attaching* operation. Figure 3 illustrates (a) a contraction and (b) a leaf-attaching operation.

We are now in a position to establish the main result of this section.

Theorem 4.2. *Let $\mathcal{P}$ be a collection of phylogenetic X-trees. There exists a set Z of auxiliary leaves with the following two properties.*

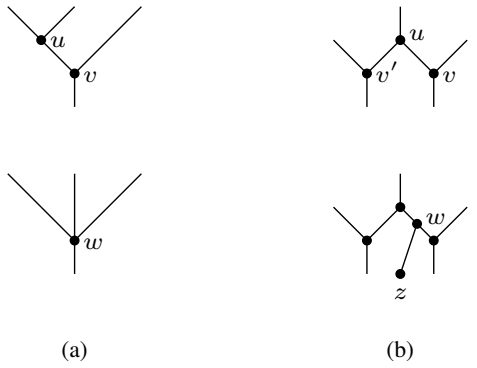

Figure 3: The phylogenetic networks at the bottom are obtained from their respective networks at the top by (a) a contraction and (b) a leaf-attaching operation.

(i) $|Z| \leq h_{\mathrm{tc}}(\mathcal{P})$, *and*

(ii) *there is a collection* $\mathcal{P}_Z$ *of phylogenetic trees with leaf set* $X \cup Z$ *obtained from* $\mathcal{P}$ *by attaching* Z *such that* $h(\mathcal{P}) = h_{\mathrm{tc}}(\mathcal{P}_Z)$.

Proof. Let $\mathcal{N}$ be a stack-free network on X that displays $\mathcal{P}$ with $h(\mathcal{N}) = h(\mathcal{P})$. By Lemma 4.1, $\mathcal{N}$ exists. Now obtain a phylogenetic network $\mathcal{N}_Z$ from $\mathcal{N}$ by a minimum number of repeated applications of the leaf-attaching operation until each tree vertex in the resulting network has at least one child that is a tree vertex or a leaf. Clearly, $h(\mathcal{N}) = h(\mathcal{N}_Z)$. Moreover, since no leaf-attaching operation results in a new edge in $\mathcal{N}_Z$ that is incident with two reticulations, $\mathcal{N}_Z$ is stack free. It now follows that $\mathcal{N}_Z$ is tree-child. Let $Z = \mathcal{L}(\mathcal{N}_Z) - X$. Then, by construction, the size of Z is equal to the number of tree vertices in $\mathcal{N}$ whose two children are both reticulations. Let $\mathcal{P}_Z$ be a set of phylogenetic trees obtained from $\mathcal{P}$ by attaching Z to $\mathcal{P}$ such that $\mathcal{N}_Z$ displays $\mathcal{P}_Z$. Since $\mathcal{N}$ displays $\mathcal{P}$, such a set $\mathcal{P}_Z$ always exists. By construction, $h(\mathcal{P}) \geq h_{\mathrm{tc}}(\mathcal{P}_Z)$. Moreover, as each tree in $\mathcal{P}$ is a restriction of a tree in $\mathcal{P}_Z$, it follows that $h(\mathcal{P}) \leq h_{\mathrm{tc}}(\mathcal{P}_Z)$; thereby establishing part (ii) of the theorem.

Using the construction of the previous paragraph, we now establish part (i) of the theorem. Let E_r be the set of reticulation edges in $\mathcal{N}$, and let V_t be the set of tree vertices of $\mathcal{N}$ whose children are both reticulations. Recall that $|V_t| = |Z|$. We next make two observations. First, each vertex in V_t is incident with two edges in E_r. Second, each edge in E_r is incident with at most one vertex in V_t. In summary, this implies that $|Z| \leq \frac{1}{2}|E_r|$.

Furthermore, we have $|E_r| = h(\mathcal{P}) + |V_r|$, where V_r is the set of reticulations in $\mathcal{N}$. Therefore, as $|V_r| \leq h(\mathcal{P})$, we have $|E_r| \leq 2h(\mathcal{P})$. As $h(\mathcal{P}) \leq h_{\mathrm{tc}}(\mathcal{P})$, it now follows that

$$|Z| \leq \tfrac{1}{2}|E_r| \leq \tfrac{1}{2} \cdot 2h(\mathcal{P}) \leq \tfrac{1}{2} \cdot 2h_{\mathrm{tc}}(\mathcal{P}) = h_{\mathrm{tc}}(\mathcal{P}).$$

This establishes part (i) of the theorem. □

5. Scoring an optimum forest

For a collection $\mathcal{P}$ of binary phylogenetic X-trees, acyclic-agreement forests characterise $h(\mathcal{P})$ for when $\mathcal{P}$ consists of exactly two trees. Indeed, many algorithms and theoretical results that deal with Minimum Hybridisation for two trees are deeply-anchored in the notion of acyclic-agreement forests [1, 2, 4, 16]. In this section, we establish a particular hardness result that contributes to an explanation of why acyclic-agreement forests appear, however, to be of little use to solve Minimum Hybridisation for more than two trees. This result is a particular instance of a conjecture in [13, page 1626].

For the purpose of the upcoming definitions, we regard the root of a binary phylogenetic X-tree $\mathcal{T}$ as a vertex labelled ρ at the end of a pendant edge adjoined to the original root. Furthermore, we view ρ as an element of the leaf set of $\mathcal{T}$; thus $\mathcal{L}(\mathcal{T}) = X \cup \{\rho\}$. Let $\mathcal{T}$ and $\mathcal{T}'$ be two binary phylogenetic X-trees. An *agreement forest* $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ for $\mathcal{T}$ and $\mathcal{T}'$ is a partition of $X \cup \{\rho\}$ such that $\rho \in \mathcal{L}_\rho$ and the following conditions are satisfied:

(i) For all $i \in \{\rho, 1, 2, \ldots, k\}$, we have $\mathcal{T}|\mathcal{L}_i \cong \mathcal{T}'|\mathcal{L}_i$.

(ii) The trees in
$$\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$$
and
$$\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$$
are vertex-disjoint subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

Now, let $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Let $G_\mathcal{F}$ be the directed graph that has vertex set $\mathcal{F}$ and an arc from $\mathcal{L}_i$ to $\mathcal{L}_j$ precisely if $i \neq j$ and

(iii) the root of $\mathcal{T}(\mathcal{L}_i)$ is an ancestor of the root of $\mathcal{T}(\mathcal{L}_j)$ in $\mathcal{T}$, or the root of $\mathcal{T}'(\mathcal{L}_i)$ is an ancestor of the root of $\mathcal{T}'(\mathcal{L}_j)$ in $\mathcal{T}'$.

We call $\mathcal{F}$ an *acyclic-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ if $G_{\mathcal{F}}$ has no directed cycle. Moreover, if $\mathcal{F}$ contains the smallest number of elements over all acyclic-agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, we say that $\mathcal{F}$ is a *maximum acyclic-agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$, in which case, we denote this number minus one by $m_a(\mathcal{T}, \mathcal{T}')$.

Baroni et al. [2] established the following characterisation for when a collection of binary phylogenetic X-trees contains exactly two trees.

Theorem 5.1. *Let $\mathcal{P} = \{\mathcal{T}, \mathcal{T}'\}$ be a collection of two binary phylogenetic X-trees. Then $h(\mathcal{P}) = m_a(\mathcal{T}, \mathcal{T}')$.*

Let $\mathcal{N}$ be a phylogenetic network on X with root ρ that displays a set $\mathcal{P}$ of binary phylogenetic X-trees. As above, we regard ρ as a vertex at the end of a pendant edge adjoined to the original root. We obtain a forest from $\mathcal{N}$ by deleting all reticulation edges, repeatedly contracting edges where one end-vertex has degree one and is not in $X \cup \{\rho\}$, deleting isolated vertices not in $X \cup \{\rho\}$ and, lastly, suppressing all vertices with in-degree one and out-degree one. Let $\{\mathcal{S}_\rho, \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k\}$ be the forest obtained from $\mathcal{N}$ in this way. We say that $\mathcal{F} = \{\mathcal{L}(\mathcal{S}_\rho), \mathcal{L}(\mathcal{S}_1), \mathcal{L}(\mathcal{S}_2), \ldots, \mathcal{L}(\mathcal{S}_k)\}$ is the forest *induced* by $\mathcal{N}$. Moreover, $\mathcal{F}$ is said to be *optimum* if $\mathcal{N}$ is a tree-child network with $h_{\mathrm{tc}}(\mathcal{P}) = h(\mathcal{N})$. For example, up to regarding ρ as a new vertex that is adjoined to the original root of the two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ shown in Figure 1,

$$\mathcal{F} = \{\{\rho, 1, 2, 6\}, \{3\}, \{4\}, \{5\}\} \text{ and } \mathcal{F}' = \{\{\rho, 1, 3, 5, 6\}, \{2\}, \{4\}\}$$

is the induced forest of $\mathcal{N}$ and $\mathcal{N}'$, respectively.

In [13], the authors investigate Minimum Hybridisation for three trees and conjecture that, given a set $\mathcal{P}$ of three binary phylogenetic X-trees and the induced forest of a phylogenetic network $\mathcal{N}$ that displays $\mathcal{P}$ and $h(\mathcal{P}) = h(\mathcal{N})$, it is NP-hard to determine $\mathcal{N}$. For $|\mathcal{P}| \geq 3$, we affirmatively answer their conjecture in the context of tree-child networks. More precisely, using cherry-picking sequences, we show that the following decision problem is NP-complete.

Scoring Optimum Forest
Instance. A non-negative integer k, a collection $\mathcal{P}$ of binary phylogenetic X-trees, an optimum forest $\mathcal{F}$ induced by a tree-child network $\mathcal{N}$ on X that displays $\mathcal{P}$.
Question. Is $h_{\mathrm{tc}}(\mathcal{P}) \leq k$?

If $|\mathcal{P}| = 2$, then, by Observation 1.1, $h_{\mathrm{tc}}(\mathcal{P}) = h(\mathcal{P})$ and $\mathcal{F}$ is a maximum acyclic-agreement forest with $h_{\mathrm{tc}}(\mathcal{P}) = |\mathcal{F}| - 1$. Hence, Scoring Optimum Forest is polynomial time when $|\mathcal{P}| = 2$. However, the general problem is NP-complete.

Theorem 5.2. *The problem* Scoring Optimum Forest *is* NP*-complete.*

The remainder of this section consists of the proof of Theorem 5.2. To establish the result, we use a reduction from a particular instance of the NP-complete problem Shortest Common Supersequence. Let Σ be a finite alphabet, and let W be a finite subset of words in Σ^*. A word $z \in \Sigma^*$ is a *common supersequence of* W if each word in W is a subsequence of z.

Shortest Common Supersequence (SCS)
Instance. A non-negative integer k, a finite alphabet Σ, and a finite subset W of words in Σ^*.
Question. Is there a supersequence of the words in W with at most k letters?

Timkovskii [23, Theorem 2] established the next theorem. The *orbit* of a letter in Σ is the set of its occurrences in the words in W. Note that if a word in W uses a letter, b say, twice, then that word contributes two occurrences to the orbit of b.

Theorem 5.3. *The decision problem* SCS *is* NP*-complete even if each word in W has* 3 *letters and the size of all orbits is* 2*.*

A consequence of Theorem 5.3 is the next corollary.

Corollary 5.4. *The decision problem* SCS *is* NP*-complete even if each word in W has* 3 *letters, the size of all orbits is at most* 2*, and no word in W contains a letter twice.*

PROOF. Let k, Σ, and W be an instance of SCS, where each word W has 3 letters and all orbits have size 2. Let Y be the subset of W that consists of those words in W in which no letter occurs twice. Observe that if $b \in \Sigma$ and b occurs twice in a word in W, then no word in Y contains b. Furthermore, with regards to Y, each word has 3 letters, the size of all orbits is at most 2, and no word contains a letter twice. Let t denote the number of distinct letters that occur in two distinct words in $W - Y$. Note that the construction of Y and the computation of t can both be done in time polynomial in $|W|$. The corollary will follow from Theorem 5.3 by showing that SCS with parameters Σ and W has a supersequence of length at most

$$2|W - Y| + t + k$$

if and only if SCS with parameters Σ and Y has a supersequence of length at most k.

Suppose that SCS with parameters Σ and Y has a supersequence z of length at most k. Now iteratively extend z to a sequence z' as follows. Let $w \in W - Y$. Then w contains two occurrences of a letter, b say, in Σ. Let d denote the third letter in w. Note that b occurs in no other word in Σ and d occurs in exactly one other word in Σ. First assume that d occurs in a word in Y. Depending on whether d is the first, second, or third letter in w, extend z by adding bb to the end of z, adding b at the beginning and b at the end of z, or adding bb at the beginning of z, respectively. The resulting sequence is a supersequence for $Y \cup \{w\}$. Second assume that d does not occur as a word in Y. Then d occurs in a word w' in $W - Y$. Let c denote the letter occurring twice in w'. Extend z by adding w to the beginning of z and then, to the resulting sequence, add two occurrences of c after w, add one occurrence of c before w and one occurrence of c after w, or two occurrences of c before w depending on whether d is the first, second, or third letter of w', respectively. The resulting sequence is a supersequence for $Y \cup \{w, w'\}$. Taking the resulting sequence and repeating this process for each remaining word in $W - (Y \cup \{w\})$ or $W - (Y \cup \{w, w'\})$, respectively, we eventually obtain a supersequence z' for W. Moreover, z' has length

$$2|W - Y| + t + k.$$

For the converse, suppose that there is a supersequence z of W of length

$$2|W - Y| + t + k.$$

Let z' be the sequence obtained from z by deleting each occurrence of a letter that occurs twice in a word in W and deleting exactly one occurrence of a letter that occurs in two distinct words in $W - Y$ and, hence, does not occur in a word in Y. It is easily checked that z' is a supersequence of Y. Furthermore, since there are $2|W - Y|$ deletions of the first type and t deletions of the second type, it follows that z' has length k. This completes the proof of the corollary. □

The decision problem described in the statement of Corollary 5.4 is the one we will use for the reduction in proving Theorem 5.2. Let k, Σ, and W be an instance of SCS such that each word in W has 3 letters, the size of all orbits is at most 2, and no word in W contains a letter twice. Without loss of generality, we may assume that, for each $\ell \in \Sigma$, there is a word in W containing ℓ, and that $|W| \geq 3$, so no letter is contained in each word. Let

$$W = \{w_1, w_2, \ldots, w_q\}$$

and, for each $i \in \{1, 2, \ldots, q\}$, let $w_i = w_{i1}\, w_{i2}\, w_{i3}$. Also, let

$$o(\Sigma) = \ell_1, \ell_2, \ldots, \ell_{|\Sigma|}$$

denote a fixed ordering of the letters in Σ. For each w_i in W, we denote the sequence obtained from $o(\Sigma)$ by removing each of the three letters in w_i by $o(\Sigma) - w_i$.

We now construct an instance of Scoring Optimum Forest. A *rooted caterpillar* is a binary phylogenetic tree $\mathcal{T}$ whose leaf set can be ordered, say $x_1, x_2, \ldots, x_n$, so that $\{x_1, x_2\}$ is a cherry and if p_i denotes the parent of x_i, then, for all $i \in \{3, 4, \ldots, n\}$, we have (p_i, p_{i-1}) as an edge in $\mathcal{T}$. Here, we denote the rooted caterpillar by $(x_1, x_2, \ldots, x_n)$.

Now, for each $i \in \{1, 2, \ldots, q\}$, let $\mathcal{T}_i$ denote the rooted caterpillar

$$\mathcal{T}_i = (\alpha, w_{i1}, w_{i2}, w_{i3}, \beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}, o(\Sigma) - w_i),$$

and let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_q\}$. Note that each of the trees in $\mathcal{P}$ has leaf set

$$X = \{\alpha, \beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}\} \cup \Sigma$$

and $\mathcal{P}$ can be constructed in time polynomial in the size of Σ and W.

We next establish a lemma that reveals a relationship between the weight of a cherry-picking sequence for $\mathcal{P}$ and the length of a supersequence for W. Let

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s), (x_{s+1}, -)$$

be a cherry-picking sequence for a set $\mathcal{P}$ of binary phylogenetic X-trees. For each $i \in \{1, 2, \ldots, s\}$, we say that (x_i, y_i) *corresponds* to n trees in $\mathcal{P}$ if $\{x_i, y_i\}$ is a cherry in exactly n trees obtained from $\mathcal{P}$ by picking $x_1, x_2, \ldots, x_{i-1}$, where $1 \leq n \leq |\mathcal{P}|$.

Lemma 5.5. *Let $k \leq 2|\Sigma|$ be a positive integer. Then there is a cherry-picking sequence of $\mathcal{P}$ of weight k if and only if there is a supersequence of W of length k.*

Proof. First suppose there is a common supersequence z of W of length k. Let

$$z = m_1\, m_2\, \cdots\, m_k,$$

and let σ denote the sequence

$$(m_1, \alpha), \ldots, (m_k, \alpha), (\beta_1, \alpha), \ldots, (\beta_{q|\Sigma|}, \alpha), (\ell_1, \alpha), \ldots, (\ell_{|\Sigma|}, \alpha), (\alpha, -).$$

Since z is a supersequence of W, it is easily seen that σ is a cherry-picking sequence of $\mathcal{P}$. Moreover,

$$w(\sigma) = (k + |X|) - |X| = k.$$

Now suppose that there is a cherry-picking sequence

$$\sigma = (x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s), (x_{s+1}, -)$$

of $\mathcal{P}$ of weight k. Without loss of generality, we may assume that each ordered pair in σ is essential. We first show that there is a positive integer i' such that $\mathcal{P}_{i'}$ is obtained from $\mathcal{P}$ by picking $x_1, x_2, \ldots, x_{i'}$ and each tree in $\mathcal{P}_{i'}$ has a cherry consisting of two elements in $\{\alpha, \beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}\}$. If not, then there is a word w_i in W such that either $(w_{ij}, \beta_{q|\Sigma|})$ or $(\beta_{q|\Sigma|}, w_{ij})$ is an ordered pair in σ, where $j \in \{1, 2, 3\}$. Now, by considering a word not containing w_{ij} and its associated tree in $\mathcal{P}$, it is easily seen that $w(\sigma) \geq q|\Sigma| - 1$ as each of the elements in $\{\beta_1, \beta_2, \ldots, \beta_{q|\Sigma|-1}\}$ appears at least twice as the first element of an ordered pair in σ. But then

$$q|\Sigma| - 1 > 2|\Sigma|$$

as $q \geq 3$ and $|\Sigma| \geq 3$, contradicting the assumption $k \leq 2|\Sigma|$.

Consider the first i' ordered pairs in σ. For each tree $\mathcal{T}_i$ in $\mathcal{P}$, there are exactly three ordered pairs whose first and second elements are in

$$S = \{\alpha, w_{i1}, w_{i2}, w_{i3}\}$$

and picking $x_1, x_2, \ldots, x_{i'}$ from $\mathcal{T}_i$ picks three elements of S. Since the size of all orbits is at most 2, such an ordered pair corresponds to at most two trees in $\mathcal{P}$. We next construct a sequence σ' of ordered pairs obtained from σ. We start by modifying the first i' ordered pairs of σ as follows:

(a) Amongst the first i' ordered pairs, replace each ordered pair of the form (α, ℓ) with (ℓ, α), where $\ell \in \Sigma$.

(b) With the sequence obtained after (a) is completed, sequentially move along the sequence to the i'-th ordered pair replacing each ordered pair of the form (ℓ, ℓ'), where $\ell, \ell' \in \Sigma$ in one of the following ways:

   (i) If (ℓ, ℓ') corresponds to exactly one tree in $\mathcal{P}$, then replace it with (ℓ, α) or (ℓ', α) depending on whether (ℓ', α) or (ℓ, α), respectively, is an earlier ordered pair.

   (ii) If (ℓ, ℓ') corresponds to two trees, $\mathcal{T}_i$ and $\mathcal{T}_j$ say, in $\mathcal{P}$ and the order of the letters ℓ and ℓ' is the same in w_i and w_j, then replace it with (ℓ, α) or (ℓ', α) depending on whether (ℓ', α) or (ℓ, α), respectively, is an earlier ordered pair.

   (iii) If (ℓ, ℓ') corresponds to two trees, $\mathcal{T}_i$ and $\mathcal{T}_j$ say, in $\mathcal{P}$ and the order of the letters ℓ and ℓ' in w_i is not the same as that in w_j, then replace it with (ℓ, α) if (ℓ, α) occurs as an ordered pair before (ℓ', α) earlier in the sequence; otherwise, (ℓ', α) occurs as an ordered pair before (ℓ, α) earlier in the sequence and so replace it with (ℓ', α).

With this modification of σ after (b) is completed, let σ'_1 denote the subsequence of the first i' ordered pairs whose coordinates are in $\Sigma \cup \{\alpha\}$, and let σ'_2 denote the subsequence $\sigma - \sigma'_1$. Let σ' denote the concatenation of σ'_1 and σ'_2.

Now, consider each tree $\mathcal{T}_i$ in $\mathcal{P}$ together with its corresponding ordered pairs in σ and the associated ones in σ'. Let $\mathcal{P}_{|\sigma'_1|}$ be the set of trees obtained from $\mathcal{P}$ by picking $x'_1, x'_2, \ldots, x'_{|\sigma'_1|}$, where x'_j is the first coordinate of the

j-th ordered pair in σ' for each $j \in \{1, 2, \ldots, |\sigma'_1|\}$. A routine check shows that this picking sets the tree corresponding to $\mathcal{T}_i$ in $\mathcal{P}_{|\sigma'_1|}$ to be the rooted caterpillar

$$(\alpha, \beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}, o(\Sigma) - w_i).$$

In particular, $(w_{i1}, \alpha), (w_{i2}, \alpha), (w_{i3}, \alpha)$ is subsequence of σ'_1.

We next extend σ'_1 to a cherry-picking sequence for $\mathcal{P}$ of weight at most k. Consider σ and σ'. If σ_1 denotes the subsequence of ordered pairs in σ corresponding to σ'_1, then $|\sigma_1| = |\sigma'_1|$ and

$$|\sigma| - |\sigma_1| \geq q|\Sigma| + |\Sigma| + 1$$

as each of the elements in $\{\beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}\} \cup \Sigma$ as well as at least one element in $\{\beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}\} \cup \Sigma \cup \{\alpha\}$ appears as the first coordinate of an ordered pair in $\sigma - \sigma_1$. Here, an element in $\{\beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}\} \cup \Sigma$ may be counted twice as it appears as the first coordinate of two ordered pairs in $\sigma - \sigma_1$. It follows that the sequence of ordered pairs that is the concatenation of σ'_1 and

$$(\beta_1, \alpha), (\beta_2, \alpha), \ldots, (\beta_{q|\Sigma|}, \alpha), (\ell_1, \alpha), (\ell_2, \alpha), \ldots, (\ell_{|\Sigma|}, \alpha), (\alpha, -)$$

is a cherry-picking sequence of $\mathcal{P}$ whose weight is at most k.

Let σ'_1 be the sequence

$$(m_1, \alpha), (m_2, \alpha), \ldots, (m_{k'}, \alpha).$$

Since $(w_{i1}, \alpha), (w_{i2}, \alpha), (w_{i3}, \alpha)$ is a subsequence of σ'_1 for each tree $\mathcal{T}_i$,

$$m_1\, m_2 \cdots m_{k'}$$

is a common supersequence of W. Moreover, as $w(\sigma) = k$, we have $k' \leq k$. It follows that there is a supersequence of W of length k. □

To complete the proof of Theorem 5.2, let

$$\mathcal{F} = \{\{\rho, \alpha, \beta_1, \beta_2, \ldots, \beta_{q|\Sigma|}\}\} \cup \{\{\ell\} : \ell \in \Sigma\}$$

be a partition of $X \cup \{\rho\}$. We next show that $\mathcal{F}$ is an optimum forest induced by a tree-child network on X with root ρ and that displays $\mathcal{P}$. Let $z = z_1 z_2 \cdots z_k$ be a common supersequence of W of minimum length, and suppose this length is k. Since all orbits have size at most 2 and z is of

minimum length, each letter in Σ appears at most twice in z, and so $k \leq 2|\Sigma|$. Let $\mathcal{T}$ be the 'multi-labelled' rooted caterpillar

$$\mathcal{T} = (\alpha,\, z_1, z_2, \ldots, z_k,\, \beta_1,\, \beta_2,\, \ldots,\, \beta_{q|\Sigma|},\, o(\Sigma))$$

and let $\mathcal{N}$ be the tree-child network with root ρ obtained from $\mathcal{T}$ as follows. For each $\ell \in \Sigma$, identify the leaves labelled ℓ and adjoin a new pendant edge to the identified vertex with the leaf-end labelled ℓ. Since z is a common supersequence of W, it is easily checked that $\mathcal{N}$ displays $\mathcal{P}$. Furthermore, $h(\mathcal{N}) = k$. By Theorem 2.1 and Lemma 5.5, $h_{\mathrm{tc}}(\mathcal{P}) = k$, and so, as $\mathcal{F}$ is induced by $\mathcal{N}$, it follows that $\mathcal{F}$ is an optimum forest for $\mathcal{P}$.

Now, given an arbitrary phylogenetic network, it can be verified in polynomial time whether it is tree-child, it displays $\mathcal{P}$ [21], its hybridisation number is at most k, and it induces $\mathcal{F}$. Hence, Scoring Optimum Forest is in NP.

Theorem 5.2 now follows by combining Corollary 5.4 with Theorem 2.1 and Lemma 5.5.

6. Concluding remarks

In this paper, we have generalised the concept of cherry-picking sequences as introduced in [12] and shown how this generalisation can be used to characterise the minimum number of reticulation events that is needed to explain any set of phylogenetic X-trees in the space of tree-child networks as well as in the space of all phylogenetic networks. To see that these two minima can be different for a fixed set of phylogenetic trees, consider the set $\mathcal{P}$ of trees presented in Figure 2. It was shown in [13, 15] that $h(\mathcal{P}) = 3$ and that there are six phylogenetic networks each of which displays $\mathcal{P}$ and has a hybridisation number of three. However, none of these six phylogenetic networks is tree-child. Moreover, using cherry-picking sequences a straightforward check shows that $h_{\mathrm{tc}}(\mathcal{P}) = 4$. Furthermore, we have shown that Scoring Optimum Forest is NP-complete. Hence, given an optimum forest, it is computationally hard to compute $h_{\mathrm{tc}}(\mathcal{P})$ for when $\mathcal{P}$ is a set of binary phylogenetic X-trees, where $|\mathcal{P}| \geq 3$. This contrasts with the two-tree case for which Scoring Optimum Forest is polynomial-time solvable and further hints at that agreement forests are of limited use beyond the two-tree case.

Of course, restricting to collections of binary phylogenetic trees, one could generalise the definition of an acyclic-agreement forest for two binary phylogenetic trees to more than two trees in the most obvious way. That is, one requires Conditions (i), (ii), and (iii) in the definition of an acyclic-agreement forest to hold for each tree in an arbitrarily large collection of binary phylogenetic X-trees. With this generalisation in mind and observing that the number of components in a forest that is induced by a tree-child network is equal to its number of reticulations plus one, one might conjecture that, given a set $\mathcal{P}$ of binary phylogenetic X-trees, the number of components in a maximum acyclic-agreement forest for $\mathcal{P}$ is the same as the minimum number of components in an optimum forest for $\mathcal{P}$. To see that this is not true, we refer back to Figure 1. Let $\mathcal{F}$ be the forest induced by $\mathcal{N}$, and let $\mathcal{F}'$ be the forest induced by $\mathcal{N}'$. Since $|\mathcal{F}'| = 3$, a maximum acyclic-agreement forest for $\mathcal{P}$ has at most three elements. Moreover, since $h(\mathcal{N})$=3 and $\mathcal{N}$ is tree-child, we have $h_{\rm tc}(\mathcal{P}) \leq 3$. Indeed, it can be checked that $h_{\rm tc}(\mathcal{P}) = 3$. Moreover, there is no tree-child network that displays $\mathcal{P}$ and induces an optimum forest that is also a maximum acyclic-agreement forest for $\mathcal{P}$. Consequently, an approach that exploits maximum acyclic-agreement forests for a set $\mathcal{P}$ of binary phylogenetic trees to compute $h_{\rm tc}(\mathcal{P})$, such as computing a maximum acyclic-agreement forest $\mathcal{F}$ for $\mathcal{P}$ and, subsequently, scoring $\mathcal{F}$ in a way that reflects the number of edges that are directed into each reticulation vertex in a network that induces $\mathcal{F}$, is unlikely to give the desired result.

Lastly, from a computational viewpoint, the introduction of acyclic-agreement forests [2] has triggered significant progress towards the development of ever faster algorithms to solve Minimum Hybridisation for when the input contains exactly two phylogenetic trees (e.g. see [1, 4, 8, 9, 20, 24]). We look forward to seeing a similar development now for solving Minimum Hybridisation for arbitrarily many phylogenetic trees by using cherry-picking sequences. In turn, this is likely to be of benefit to biologists who often wish to infer evolutionary histories that are not entirely tree-like and for data sets that usually consists of more than two phylogenetic trees.

Acknowledgements. We thank the New Zealand Marsden Fund for their financial support.

References.

[1] B. Albrecht, C. Scornavacca, A. Cenci, and D. H. Huson (2012). Fast

computation of minimum hybridization networks. *Bioinformatics*, **28**, 191–197.

[2] M. Baroni, S. Grünewald, V. Moulton, and C. Semple, (2005), Bounding the number of hybridization events for a consistent evolutionary history, *Journal of Mathematical Biology*, **51**, 171–182.

[3] M. Bordewich and C. Semple (2007). Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics*, **155**, 914–928.

[4] M. Bordewich and C. Semple (2007). Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4**, 458–466.

[5] M. Bordewich and C. Semple (2016). Determining phylogenetic networks from inter-taxa distances. *Journal of Mathematical Biology,* **73**, 283–303.

[6] G. Cardona, F. Rosselló, and G. Valiente (2009). Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **6**, 552–569.

[7] Z. Z. Chen and L. Wang (2012). Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**, 372–384.

[8] Z. Z. Chen and L. Wang (2013). An ultrafast tool for minimum reticulate networks. *Journal of Computational Biology*, **20**, 38–41.

[9] J. Collins, S. Linz, and C. Semple (2011). Quantifying hybridization in realistic time. *Journal of Computational Biology*, **18**, 1305–1318.

[10] J.-M. Drezen, J. Gauthier, T. Josse , A. Bézier, E. Herniou, E. Huguet (2016). Foreign DNA acquisition by invertebrate genomes. *Journal of Invertebrate Pathology*, doi: 10.1016/j.jip.2016.09.004.

[11] P. J. Humphries, S. Linz, and C. Semple (2013). On the complexity of computing the temporal hybridization number for two phylogenies. *Discrete Applied Mathematics*, **161**, 871–880.

[12] P. J. Humphries, S. Linz, and C. Semple (2013). Cherry picking: a characterization of the temporal hybridization number for a set of phylogenies. *Bulletin of Mathematical Biology*, **75**, 1879–1890.

[13] L. van Iersel, S. Kelk, N. Lekić, C. Whidden, and N. Zeh, (2016). Hybridization number on three rooted binary trees is EPT. *SIAM Journal on Discrete Mathematics*, **30**, 1607–1631.

[14] L. van Iersel, C. Semple, and M. Steel (2010). Locating a tree in a phylogenetic network. *Information Processing Letters*, **110**, 1037–1043.

[15] S. Kelk (2012). Personal communication.

[16] S. Kelk, L. van Iersel, N. Lekić, S. Linz, C. Scornavacca, and L. Stougie (2012). Cycle killer ... qu'est-ce que c'est? On the comparative approximability of hybridization number and directed feedback vertex set. *SIAM Journal on Discrete Mathematics*, **26**,1635–1656.

[17] J. Mallet, N. Besansky, and M. W. Hahn (2016). How reticulated are species? *BioEssays*, **38**, 140–149.

[18] T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, International Wheat Genome Sequencing Consortium, K. S. Jakobsen, B. B. Wulff, B. Steuernagel, K. F. Mayer, and O. A. Olsen (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.

[19] B. M. E. Moret, L. Nakhleh, T. Warnow, C. R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme (2004). Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**, 13–23.

[20] T. Piovesan and S. Kelk (2012). A simple fixed parameter tractable algorithm for computing the hybridization number of two (not necessarily binary) trees, *IEEE Transactions on Computational Biology and Bioinformatics*, **10**, 18–25.

[21] J. Simpson. Tree structure in phylogenetic networks. PhD thesis, University of Canterbury, *in preparation*.

[22] S. M. Soucy, J. Huang, and J. P. Gogarten (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, **16**, 472–482.

[23] V. G. Timkovskii (1989). Complexity of common subsequence and supersequence problems and related problems. *Cybernetics*, **25**, 565–580.

[24] Y. Wu and J. Wang (2010). Fast computation of the exact hybridization number of two phylogenetic trees. In: International Symposium on Bioinformatics Research and Applications, Springer, pp. 203–214.

[25] Y. Wu (2010). Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics*, **26**, i140–i148.