CHERRY PICKING IN FORESTS: A NEW CHARACTERIZATION FOR THE UNROOTED HYBRID NUMBER OF TWO PHYLOGENETIC TREES

KATHARINA T. HUBER, SIMONE LINZ, AND VINCENT MOULTON

Abstract. Phylogenetic networks are a special type of graph which generalize phylogenetic trees and that are used to model non-treelike evolutionary processes such as recombination and hybridization. In this paper, we consider *unrooted* phylogenetic networks, i.e. simple, connected graphs $\mathcal{N} = (V, E)$ with leaf set X, for X some set of species, in which every internal vertex in $\mathcal{N}$ has degree three. One approach used to construct such phylogenetic networks is to take as input a collection $\mathcal{P}$ of phylogenetic trees and to look for a network $\mathcal{N}$ that contains each tree in $\mathcal{P}$ and that minimizes the quantity $r(\mathcal{N}) = |E| - (|V| - 1)$ over all such networks. Such a network always exists, and the quantity $r(\mathcal{N})$ for an optimal network $\mathcal{N}$ is called the *hybrid number of* $\mathcal{P}$. In this paper, we give a new characterization for the hybrid number in case $\mathcal{P}$ consists of two trees. This characterization is given in terms of a *cherry picking sequence* for the two trees, although to prove that our characterization holds we need to define the sequence more generally for two forests. Cherry picking sequences have been intensively studied for collections of *rooted* phylogenetic trees, but our new sequences are the first variant of this concept that can be applied in the unrooted setting. Since the hybrid number of two trees is equal to the well-known *tree bisection and reconnection distance* between the two trees, our new characterization also provides an alternative way to understand this important tree distance.

1. Introduction

Phylogenetic networks are a special type of graph which generalize phylogenetic trees and that are used to model non-treelike evolutionary processes such as recombination and hybridization [12]. There are several classes of phylogenetic networks (see e.g. [21] for a recent review), but in this paper, we shall restrict our attention to *unrooted* phylogenetic networks (see e.g. [14]). Such a network is a simple, connected graph $\mathcal{N} = (V, E)$ with leaf set X, where X is some set of species, and every internal vertex in $\mathcal{N}$ has degree three. The *reticulation number* $r(\mathcal{N})$ of $\mathcal{N}$ is defined as $|E| - (|V| - 1)$. If $r(\mathcal{N}) = 0$, then $\mathcal{N}$ has no cycles, in which case it is called an *(unrooted binary) phylogenetic tree* on X. An example of a phylogenetic network $\mathcal{N}$ with leaf set $X = \{1, \ldots, 6\}$ and $r(\mathcal{N}) = 1$ is given in the left of Figure 1

Date: May 12, 2025.

2020 *Mathematics Subject Classification.* 05C05; 92D15.

Key words and phrases. phylogenetic network; forest; TBR distance; cherry picking sequence; hybrid number.

The second author thanks the New Zealand Marsden Fund for their financial support.

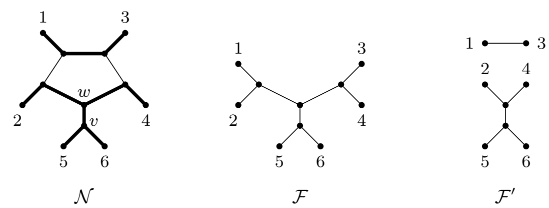

FIGURE 1. Left: A phylogenetic network $\mathcal{N}$ with leaf set $X = \{1, \ldots, 6\}$. Middle: A forest $\mathcal{F}$ that contains a single phylogenetic tree. Right: A forest $\mathcal{F}'$ comprising two components. Both $\mathcal{F}$ and $\mathcal{F}'$ are displayed by $\mathcal{N}$; the way in which $\mathcal{F}'$ is displayed by $\mathcal{N}$ is indicated in bold.

(the tree in the centre of this figure is a phylogenetic tree); formal definitions are given in the next section.

One approach commonly used to construct an unrooted phylogenetic network is to take as input a set $\mathcal{P}$ of phylogenetic trees and to infer a network that embeds or 'displays' [14, 17] each tree in $\mathcal{P}$ (see Figure 1 for an example of a phylogenetic tree displayed in a network). Such a network always exists since, for example one can be constructed using the so-called display graph [5] for all trees in $\mathcal{P}$ and subsequently refining the vertices of that graph whose degree is greater than three by introducing new edges. From a biological viewpoint, it is of interest to not reconstruct an arbitrary unrooted phylogenetic network that displays each element in $\mathcal{P}$, but one whose reticulation number is minimized over all such networks. This is known as the *Unrooted Hybridization Number* problem [14]. If $\mathcal{N}$ is an unrooted phylogenetic network that displays each tree in $\mathcal{P}$ and whose reticulation number is minimized over all such networks, then $r(\mathcal{N})$ is denoted by $h(\mathcal{P})$ and is called the *hybrid number of* $\mathcal{P}$.

Intriguingly, in [14, Theorem 3] it is shown that, if $\mathcal{P}$ contains only two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on the same leaf set, then $h(\mathcal{P})$ is in fact equal to the *tree bisection and reconnection (TBR) distance* between $\mathcal{T}$ and $\mathcal{T}'$ (see e.g. [1] and also [26]). Informally, this distance is defined as the minimum number of so-called tree bisection and reconnection moves required to convert $\mathcal{T}$ into $\mathcal{T}'$, where such a move on a phylogenetic tree is essentially the process of cutting an edge in that tree and then reconnecting the resulting two trees by introducing a new edge between two edges, one from each of the two trees. The TBR distance has been studied for several years, and is still a topic of current research [6, 18, 19, 25, 27]. Furthermore, computing this distance is an NP-hard optimization problem [1]. Hence, gaining a better understanding of the hybrid number of two phylogenetic trees is not only of interest for constructing phylogenetic networks but also for shedding new light on computing the TBR distance between them.

In this paper, we establish a new characterization for the hybrid number of two phylogenetic trees (Theorem 4.1) and, therefore the TBR distance (Theorem 4.2). Our approach is motivated by the concept of a *cherry picking sequence* that was introduced in [11] in the context of two *rooted* (binary) phylogenetic trees. A *cherry* in such a tree is a pair of leaves that are adjacent to the same vertex. Roughly speaking, for a pair of rooted phylogenetic trees $\mathcal{R}$ and $\mathcal{R}'$ on the same leaf set X, a cherry picking sequence for $\mathcal{R}$ and $\mathcal{R}'$ is an ordering or sequence σ of the elements in X, say $\sigma = (x_1, x_2, \ldots, x_n)$, $n = |X|$, such that each x_i, $i \in \{1, 2, \ldots, n\}$, is a leaf in a cherry in the phylogenetic trees obtained from $\mathcal{R}$ and $\mathcal{R}'$ by pruning off $x_1, x_2, \ldots, x_{i-1}$. In [11] it is shown that a cherry picking sequence for two rooted phylogenetic trees $\mathcal{R}$ and $\mathcal{R}'$ exists precisely if $\mathcal{R}$ and $\mathcal{R}'$ can be embedded in a *rooted* binary phylogenetic network $\mathcal{N}$ on X that satisfies certain structural properties (namely it is 'time consistent' and 'tree-child'). The number of elements x_i in σ with $i \in \{1, 2, \ldots, n\}$ for which the two trees obtained from $\mathcal{R}$ and $\mathcal{R}'$ by pruning off $x_1, x_2, \ldots, x_{i-1}$ have cherries $\{x_i, y_i\}$ and $\{x_i, y_i'\}$ such that $y_i \neq y_i'$ is called the *weight* of σ. In addition, it is shown in the same paper that the minimum weight over all cherry picking sequences for $\mathcal{R}$ and $\mathcal{R}'$ equates to the minimum number of reticulations (vertices with in-degree two and out-degree one) that are required to display $\mathcal{R}$ and $\mathcal{R}'$ by a rooted phylogenetic network that is time consistent and tree-child. Since the publication of [11], cherry picking sequences have been extensively studied and generalized to larger classes of rooted phylogenetic networks and arbitrarily large collections of rooted phylogenetic trees (see e.g. [16, 24]). This theoretical work has, in turn, resulted in the development of practical algorithms to reconstruct rooted phylogenetic networks from a set of rooted phylogenetic trees [4, 13]. In a related line of research, cherry picking operations have recently been used to define and compute distances between phylogenetic networks [22, 23].

In this paper, we shall introduce an analog of a cherry picking sequence for two unrooted phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on the same leaf set. In particular, the main result of this paper (Theorem 4.1) proves that the hybrid number and, consequently, the TBR distance of $\mathcal{T}$ and $\mathcal{T}'$ can be given in terms of the minimum weight of a cherry picking sequence using an alternative weight to the rooted setting, and the minimum is again taken over all such sequences for $\mathcal{T}$ and $\mathcal{T}'$. Interestingly, to obtain this result we found that it is necessary to define a cherry picking sequence for two *forests* on the same leaf set (a forest is a collection of phylogenetic trees, see e.g. the right-hand side of Figure 1), and then apply such a sequence to the special case where the two forests are both phylogenetic trees. Intuitively, this is the case because, instead of only pruning leaves of cherries as in the rooted case that is described in the previous paragraph, a cherry picking sequence whose weight characterizes the hybrid number of $\mathcal{T}$ and $\mathcal{T}'$ allows for operations that split a phylogenetic tree into two smaller phylogenetic trees by deleting edges that are not necessarily pendant. It is also interesting to note that our definition of a cherry picking sequence has some similarity with certain data reduction rules that have been recently introduced for establishing parameterized algorithms to compute the TBR distance between two phylogenetic trees [19].

We now summarize the content of the rest of this paper. In Section 2, we present some preliminaries. Subsequently, in Section 3, we introduce the main concept of a cherry picking sequence for two forests, and show that such a sequence always exists for any pair of forests. In Section 4, we show that if σ is a cherry picking sequence for two forests $\mathcal{F}$ and $\mathcal{F}'$ on X, then there is an unrooted phylogenetic network $\mathcal{N}$ on X that displays $\mathcal{F}$ and $\mathcal{F}'$ such that $r(\mathcal{N})$ is at most the weight of σ (Theorem 4.5). In Section 5, we prove four technical lemmas that are concerned with how forests are displayed in networks. These lemmas allow us to complete the proof of our main result (Theorem 4.1) in Section 6, where we prove that if $\mathcal{N}$ is a phylogenetic network that displays two forests $\mathcal{F}$ and $\mathcal{F}'$ on X, then there is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ whose weight is at most $r(\mathcal{N})$ (Theorem 6.1). We conclude in Section 7 with a brief discussion of some future directions of research.

2. Preliminaries

In this section we present some basic notation and definitions.

Graphs. The graphs that we shall consider in this paper will be multi-graphs. We denote such a graph G by an ordered pair $G = (V, E)$, where $V = V(G)$ are the vertices of G and $E = E(G)$ are the edges of G. No graph considered will contain a loop, but it may contain a *multi-edge*, that is, a pair of vertices which are joined by more than one edge. In case there is a single edge in G joining two vertices u and v we call it an *edge* and denote it by $\{u, v\}$. Note that if $E(G)$ contains only edges, then $E(G)$ is a set. An edge e of a graph G is called a *cut-edge* of G if the deletion of e disconnects G. If e is a cut-edge that contains a vertex of G of degree one then we call e a *trivial cut-edge*. A *subdivision* of G is a graph obtained from G by replacing edges in G by paths containing at least one edge. If G has two edges $\{u, v\}$, $\{v, w\}$, $u, v, w \in V(G)$ and v has degree two and is not incident with two edges $\{u, v\}$ or two edges $\{v, w\}$, the process of replacing these two edges by a single edge $\{u, w\}$ and removing v is called *suppressing* the vertex v. If G has a multi-edge between two vertices u and v in $V(G)$, the process of replacing the edges by a single edge $\{u, v\}$ is called *suppressing* the multi-edge. A *leaf* of G is a vertex in $V(G)$ of degree at most one. The set of leaves of G is denoted by $L(G)$. Finally, we define $r(G) = |E| - (|V| - 1)$ and refer to $r(G)$ as the *reticulation number of* G which is also known as the cyclomatic number of G [3].

Phylogenetic trees and networks. Throughout the paper, X denotes a finite set with $n = |X| \geq 1$, where we think of X as a set of species.

A *phylogenetic network* $\mathcal{N}$ on X is a simple, connected graph (V, E), with leaf set $X \subseteq V$ and such that all non-leaf vertices have degree three [14]. Note that this type of graph is sometimes called a *binary* phylogenetic network. Also note that if $X = \{x\}$, then we shall regard the graph consisting of the single vertex x as being a phylogenetic network, and will denote it by x. If $\mathcal{N}$ is a tree, that is, it contains no cycles, then we call $\mathcal{N}$ a *phylogenetic tree (on X)*. Suppose that $\mathcal{T}$ is a phylogenetic tree on X and $Y \subseteq X$ is a non-empty subset. We denote by $\mathcal{T}(Y)$ the minimal subtree of $\mathcal{T}$ that connects the elements in Y. Note that $\mathcal{T}(Y)$ need not be a phylogenetic tree on Y because it might contain vertices that have degree two. We therefore denote the phylogenetic tree obtained by suppressing all vertices in $\mathcal{T}(Y)$ of degree two by $\mathcal{T}|Y$ and refer to this as being the *restriction* of $\mathcal{T}$ to Y.

Suppose that $\mathcal{T}$ is a phylogenetic tree and that $\mathcal{N}$ is a phylogenetic network. We call $\mathcal{T}$ a *pendant subtree* of $\mathcal{N}$ if either $\mathcal{N}$ equals $\mathcal{T}$ or there is a cut-edge e in $\mathcal{N}$ so that $\mathcal{T}$ is equal to one of the two connected components obtained from $\mathcal{N}$ by deleting e and suppressing any resulting degree-two vertices. In the latter case, we shall also refer to $\mathcal{T}$ as being a *proper* pendant subtree of $\mathcal{N}$, and we denote the cut-edge e that gives rise to $\mathcal{T}$ by $e_{\mathcal{T},\mathcal{N}}$ or just $e_{\mathcal{T}}$ if $\mathcal{N}$ is clear from the context. If $\mathcal{T}'$ is a phylogenetic tree, and $\mathcal{T}$ is a proper pendant subtree of $\mathcal{T}'$, we let $\mathcal{T}' - \mathcal{T}$ denote the phylogenetic tree on $X \setminus L(\mathcal{T})$ obtained by deleting $\mathcal{T}$ and $e_{\mathcal{T}}$ from $\mathcal{T}'$ and suppressing the resulting degree-two vertices (in case there are any).

An edge of $\mathcal{N}$ that contains a leaf x of $\mathcal{N}$ is called a *pendant edge* in $\mathcal{N}$ (so that, in particular, x is a proper pendant subtree of $\mathcal{N}$ and e_x is the pendant edge of x). In case $\mathcal{N}$ has at least two leaves, a *cherry* of $\mathcal{N}$ consists of two distinct leaves x and y of $\mathcal{N}$ such that either $\mathcal{N}$ is equal to the edge $\{x, y\}$, or x and y are adjacent to a common vertex (so that, in particular, in the latter case the phylogenetic tree $\{x, y\}$ is a proper pendant subtree of $\mathcal{N}$). We shall denote a cherry consisting of leaves x and y by (x, y), where the order of x and y is unimportant. We call $\mathcal{N}$ *pendantless* if it contains no proper pendant subtree with at least two leaves (e.g. for the phylogenetic network $\mathcal{N}$ in Figure 1, since the subtree with leaf set $\{5, 6\}$ is a cherry of $\mathcal{N}$, it follows that $\mathcal{N}$ is not pendantless). We also need notation for two further special cases. Suppose that $\mathcal{T}$ is a pendant subtree of $\mathcal{N}$. If $\mathcal{T}$ has leaf set $\{x, y, z\} \subseteq X$ and $\mathcal{T}$ is a proper pendant subtree of $\mathcal{N}$ such that (x, y) is a cherry of $\mathcal{N}$, then we denote $\mathcal{T}$ by $((x, y), z)$ (for brevity, if $\mathcal{T} = \mathcal{N}$, then we shall also denote $\mathcal{T}$ by $((x, y), z)$ although the choice of the cherry (x, y) is unimportant). In addition, if $\mathcal{T}$ has leaf set $\{x, y, z, w\} \subseteq X$ and (x, y), (z, w) are both cherries in $\mathcal{N}$, then we denote $\mathcal{T}$ by $((x, y), (z, w))$.

Forests. A *forest* $\mathcal{F}$ on X is a set of phylogenetic trees whose collective leaf set equals X and no two leaves have the same label. Abusing notation, we will consider a phylogenetic tree $\mathcal{T}$ as also being a forest (i.e. the singleton set that consists of $\mathcal{T}$). A pair (x, y) with $x, y \in X$ distinct is called a *cherry* of a forest $\mathcal{F}$ if it is a cherry in one of the trees in $\mathcal{F}$.

For an edge e in a phylogenetic tree $\mathcal{T}$ on X we denote by $\mathcal{T} - e$ the forest obtained from $\mathcal{T}$ by deleting e and suppressing degree-two vertices (if any) in the resulting trees. If $\mathcal{F}$ is a forest on X, $|X| \geq 2$, and $x \in X$, then we let $\mathcal{T}_x$ denote the tree in $\mathcal{F}$ that contains x in its leaf set. We denote by $\mathcal{F} - x$ the forest obtained from $\mathcal{F}$ by removing $\mathcal{T}_x$ from $\mathcal{F}$ in case $|L(\mathcal{T}_x)| = 1$ and replacing $\mathcal{T}_x$ by $\mathcal{T}_x - x$ otherwise. Also, for e an edge in $\mathcal{F}$, we let $\mathcal{T}_e$ denote the tree in $\mathcal{F}$ with e in its edge set, and denote by $\mathcal{F} - e$ the forest obtained by replacing $\mathcal{T}_e$ by $\mathcal{T}_e - e$. For example, referring to Figure 1 for $x = 4$, the tree $\mathcal{T}_x$ in $\mathcal{F}'$ is $((2, 4), (5, 6))$ and $\mathcal{F}' - 4$ is the forest comprising the cherry with leaf set $\{1, 3\}$ and the three-leaf tree $((5, 6), 2)$. Similarly, for e being the pendant edge incident with 4, the tree $\mathcal{T}_e$ in $\mathcal{F}'$ is again $((2, 4), (5, 6))$ and $\mathcal{F} - e$ is the forest comprising the cherry with leaf set $\{1, 3\}$, the isolated vertex 4, and the three-leaf tree $((5, 6), 2)$.

Displaying phylogenetic trees and forests. Suppose that $\mathcal{N}$ is a phylogenetic network on X, and that $\mathcal{T}$ is a phylogenetic tree on some subset $Y \subseteq X$. Then we say that $\mathcal{T}$ is *displayed* by $\mathcal{N}$ if $\mathcal{T}$ can be obtained from a subtree $\mathcal{N}[\mathcal{T}]$ of $\mathcal{N}$ by suppressing all vertices in $\mathcal{N}[\mathcal{T}]$ with degree two. We shall refer to $\mathcal{N}[\mathcal{T}]$ as an *image* of $\mathcal{T}$ in $\mathcal{N}$. For example in Figure 1, for $\mathcal{T} = ((2, 4), (5, 6))$, a subgraph $\mathcal{N}[\mathcal{T}]$ of $\mathcal{N}$ with leaf set $\{2, 4, 5, 6\}$ is indicated in bold. Note that an image of $\mathcal{T}$ in $\mathcal{N}$ is isomorphic to a subdivision of $\mathcal{T}$ and that $\mathcal{T}$ could have several images in $\mathcal{N}$.

Usually the image of $\mathcal{T}$ in $\mathcal{N}$ that we are considering is clear from the context, but in case we want to make this clearer we shall refer to it explicitly.

Suppose that $\mathcal{F}$ is a forest on X. Then we say that $\mathcal{N}$ *displays* $\mathcal{F}$ if every tree in $\mathcal{F}$ is displayed by $\mathcal{N}$, and we can choose an image $\mathcal{N}[\mathcal{T}]$ for each tree $\mathcal{T}$ in $\mathcal{F}$ such that for any distinct trees $\mathcal{T}, \mathcal{T}' \in \mathcal{F}$ the images $\mathcal{N}[\mathcal{T}]$ and $\mathcal{N}[\mathcal{T}']$ do not share a vertex. If $\mathcal{N}$ displays $\mathcal{F}$, then we refer to the set $\mathcal{N}[\mathcal{F}] = \{\mathcal{N}[\mathcal{T}] : \mathcal{T} \in \mathcal{F}\}$ as an *image* of $\mathcal{F}$ in $\mathcal{N}$. For example, considering again the forest $\mathcal{F}'$ and the phylogenetic network $\mathcal{N}$ pictured in Figure 1, the subgraph $\mathcal{N}[\mathcal{F}']$ of $\mathcal{N}$ is indicated in bold.

3. Cherry picking sequences

In this section, we define the concept of a cherry picking sequence for two forests having the same leaf set, and show that such a sequence always exists for any pair of forests. We begin by making a simple but important observation, whose proof is straight-forward and omitted.

Observation 3.1. *If we are given two forests $\mathcal{F}$ and $\mathcal{F}'$ on X and two distinct elements $x, y \in X$ such that (x, y) is a cherry in $\mathcal{F}$ then precisely one of the following must hold: (i) (x, y) is a cherry in $\mathcal{F}'$; (ii) (x, z) or (y, z) is a cherry in $\mathcal{F}'$ for some $z \in X \setminus \{x, y\}$, (iii) x and y are in the same tree in $\mathcal{F}'$, but neither x nor y is in a cherry of $\mathcal{F}'$; or (iv) x and y are in different trees in $\mathcal{F}'$, but neither x nor y is in a cherry of $\mathcal{F}'$.*

Exploiting this observation, we shall now define a cherry picking sequence for two forests $\mathcal{F}$ and $\mathcal{F}'$ on the same leaf set X as follows: We call a sequence $\sigma = (x_1, x_2, \ldots, x_m)$, $m \geq |X| \geq 1$, of elements in X a *cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ (of length m)* if there are forests $\mathcal{F}[i]$ and $\mathcal{F}'[i]$, $1 \leq i \leq m$, such that $\mathcal{F}[1] = \mathcal{F}$, $\mathcal{F}'[1] = \mathcal{F}'$, $\mathcal{F}[m] = \mathcal{F}'[m] = \{x_m\}$ and, for each $1 \leq i \leq m-1$, precisely one of the following cases holds. The different cases are illustrated in Figure 2 and an example of a cherry picking sequence is presented in Figure 3.

- (C1) (x_i, y) is a cherry in both $\mathcal{F}[i]$ and $\mathcal{F}'[i]$, and $\mathcal{F}[i+1] = \mathcal{F}[i] - x_i$ and $\mathcal{F}'[i+1] = \mathcal{F}'[i] - x_i$;
- (C2) (x_i, y) is a cherry in $\mathcal{F}[i]$ and one of the following three cases holds:
  - (a) (x_i, z) is a cherry in $\mathcal{F}'[i]$ with $y \neq z$, $\mathcal{F}'[i] = \mathcal{F}'[i+1]$ and either
    - (i) $\mathcal{F}[i+1]$ equals $\mathcal{F}[i] - e_{x_i}$ or $\mathcal{F}[i] - e_y$; or
    - (ii) $((x_i, y), p)$ is a proper pendant subtree of a tree in $\mathcal{F}[i]$, $p \in X$, and $\mathcal{F}[i+1] = \mathcal{F}[i] - e_{\mathcal{S}}$ for some $\mathcal{S} \in \{p, ((x_i, y), p)\}$; or
    - (iii) $((x_i, y), (p, q))$ is a proper pendant subtree of a tree in $\mathcal{F}[i]$, $p \neq q \in X$, and $\mathcal{F}[i+1] = \mathcal{F}[i] - e_{\mathcal{S}}$ for some $\mathcal{S} \in \{p, q, (p, q)\}$;
  - (b) x_i and y are both leaves in some tree $\mathcal{T}' \in \mathcal{F}'[i]$ and neither x_i nor y is contained in a cherry in $\mathcal{T}'$, and either
    - (i) $\mathcal{F}'[i+1] = \mathcal{F}'[i]$ and $\mathcal{F}[i+1] = \mathcal{F}[i] - e_{x_i}$ or $\mathcal{F}[i+1] = \mathcal{F}[i] - e_y$; or
    - (ii) for e the edge in $\mathcal{T}'$ that contains the vertex adjacent to x_i and is not on the path between x_i and y, $\mathcal{F}'[i+1] = \mathcal{F}'[i] - e$ and $\mathcal{F}[i+1] = \mathcal{F}[i]$;

(c) x_i and y are contained in different trees of $\mathcal{F}'[i]$, neither of them is an isolated vertex or contained in a cherry of $\mathcal{F}'[i]$, $\mathcal{F}'[i] = \mathcal{F}'[i+1]$, and $\mathcal{F}[i+1] = \mathcal{F}[i] - e_{x_i}$ or $\mathcal{F}[i+1] = \mathcal{F}[i] - e_y$;

or the same holds with the roles of $\mathcal{F}$ and $\mathcal{F}'$ reversed.

(C3) x_i is a component of $\mathcal{F}[i]$, $\mathcal{F}[i+1] = \mathcal{F}[i] - x_i$, and $\mathcal{F}'[i+1] = \mathcal{F}'[i] - x_i$, or the same holds with the roles of $\mathcal{F}$ and $\mathcal{F}'$ reversed.

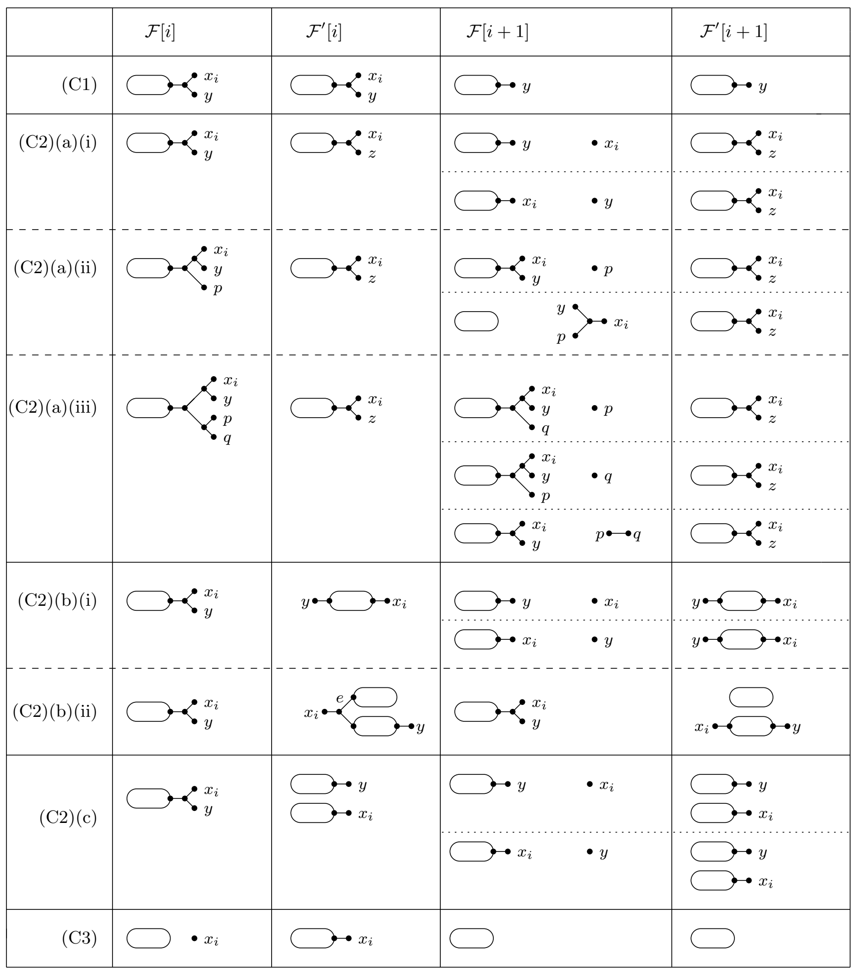

FIGURE 2. The different cases as described in the definition of a cherry picking sequence. Ovals indicate subtrees. The roles of $\mathcal{F}$ and $\mathcal{F}'$ in (C2) and (C3) could also be reversed.

Note that in cases (C1) and (C3), $\mathcal{F}[i+1]$ and $\mathcal{F}'[i+1]$ both have one leaf less than $\mathcal{F}[i]$ and $\mathcal{F}'[i]$, respectively, whereas in all sub-cases of (C2) one of $\mathcal{F}[i+1]$ and $\mathcal{F}'[i+1]$ is obtained by removing an edge from $\mathcal{F}[i]$ or $\mathcal{F}'[i]$. For simplicity, we shall call the removal of a leaf or an edge as described in each of the cases in (C1)–(C3) a *reduction*, or, if we want to be more specific, we may, for example, say that we use a (C2)(a)(ii)-reduction (applied to x_i).

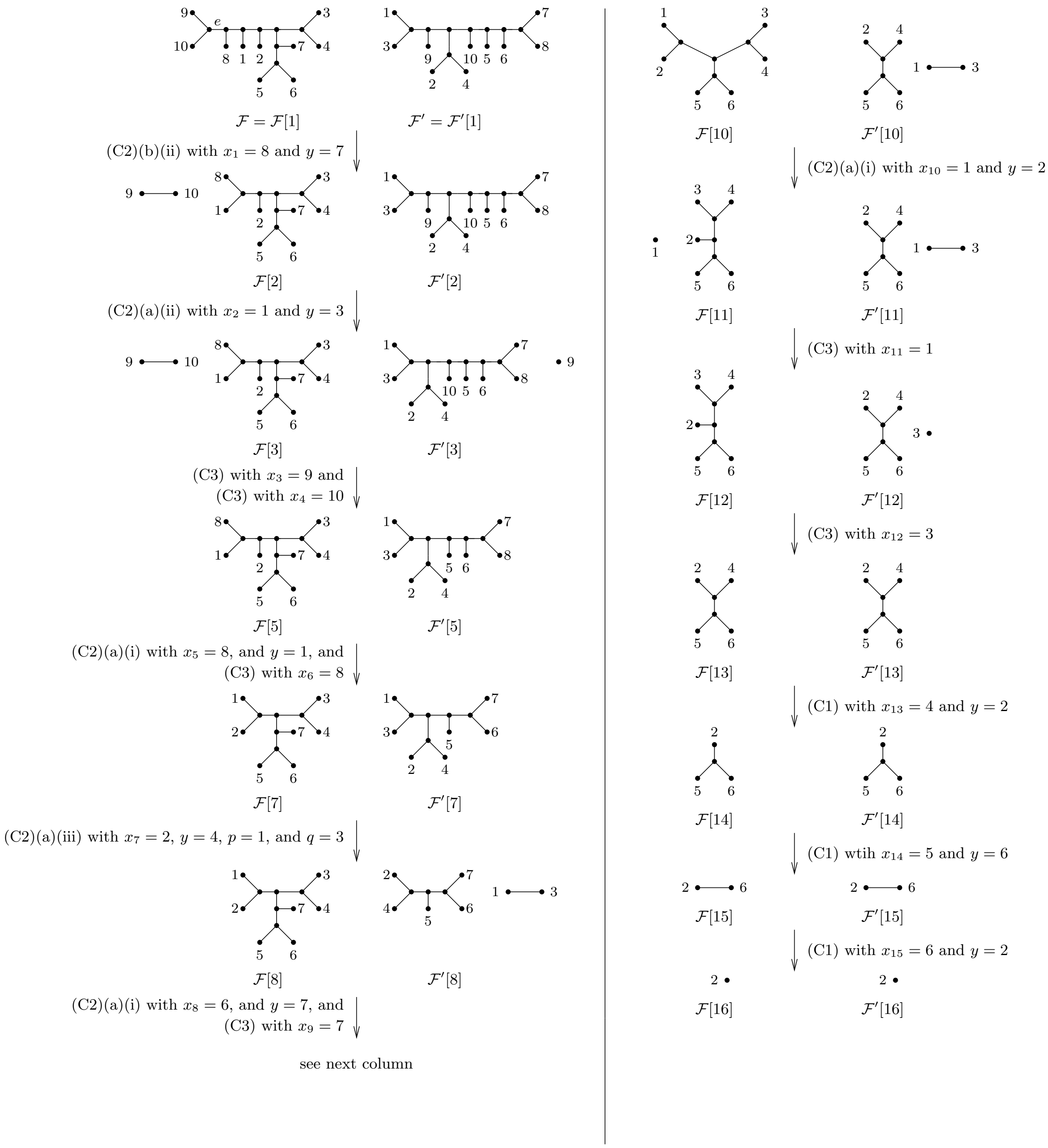

FIGURE 3. For the two depicted forests $\mathcal{F}$ and $\mathcal{F}'$ the sequence $(8, 1, 9, 10, 8, 8, 2, 6, 7, 1, 1, 3, 4, 5, 6, 2)$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$. The reductions applied within the sequence are indicated next to the arrows.

Let $\sigma = (x_1, \ldots, x_m)$ be a cherry picking sequence for two forests $\mathcal{F}$ and $\mathcal{F}'$ with $m \geq 2$. For each $1 \leq i \leq m-1$, we set $c(x_i) = 0$ if one of the reductions in (C1) and (C3) is applied to x_i and, otherwise, $c(x_i) = 1$. In addition, we define the *weight* of σ by $w(\sigma) = \sum_{i=1}^{m-1} c(x_i)$ if $m \geq 2$ and 0 if $m = 1$. To illustrate these definitions, let $X = \{1, \ldots, 6\}$ and consider the two forests $\mathcal{F}$ and $\mathcal{F}'$ on X in Figure 1. Then $(8, 1, 9, 10, 8, 8, 2, 6, 7, 1, 1, 3, 4, 5, 6, 2)$ gives a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with weight 6, as can be verified by looking at the reductions depicted in Figure 3.

We now show that a cherry picking sequence always exists for any pair of forests with the same leaf set.

Proposition 3.2. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are forests on X. Then there exists a cherry picking sequence σ for $\mathcal{F}$ and $\mathcal{F}'$ of length m for some $m \geq |X|$.*

Proof. Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are two forests on X, $n = |X| \geq 1$. We use induction on n. If $n = 1$, then we can assume $X = \{x\}$, and so $\sigma = (x)$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ (since $\mathcal{F} = \mathcal{F}' = \{x\}$).

So suppose that $n > 1$, and that for any two forests $\mathcal{F}_1$ and $\mathcal{F}_2$ on a set Y with $1 \leq |Y| \leq n-1$, there exists a cherry picking sequence for $\mathcal{F}_1$ and $\mathcal{F}_2$ of length at least $|Y|$.

First, suppose that there is no cherry in either $\mathcal{F}$ or $\mathcal{F}'$. Then the edge sets of $\mathcal{F}$ and $\mathcal{F}'$ must both be empty. Pick any $x \in X$ and apply a (C3)-reduction to obtain forests $\mathcal{F}_1 = \mathcal{F} - x$ and $\mathcal{F}_2 = \mathcal{F}' - x$ on $X \setminus \{x\}$. By induction, there is a cherry picking sequence $(x_1, \ldots, x_p)$ for $\mathcal{F}_1$ and $\mathcal{F}_2$ with $p \geq n-1$. Hence $(x, x_1, \ldots, x_p)$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with length at least n.

Now suppose that there exist $x, y \in X$ such that (x, y) is a cherry in one of the forests, say $\mathcal{F}$. If x or y is a component of $\mathcal{F}'$ then, as in the previous case, applying a (C3)-reduction and our induction hypothesis yields a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$. So assume that neither x nor y is a component of $\mathcal{F}'$. Then, by Observation 3.1, setting $\mathcal{F}[i] = \mathcal{F}$ and $\mathcal{F}'[i] = \mathcal{F}'$, without loss of generality we must be in the situation given in (C1) or one of the cases given in (C2) in the definition of a cherry picking sequence with $x = x_i$.

In case (C1), we can apply the reduction $\mathcal{F}_1 = \mathcal{F}[i] - x$ and $\mathcal{F}_2 = \mathcal{F}'[i] - x$ to $\mathcal{F}[i]$ and $\mathcal{F}'[i]$. By induction, there is a cherry picking sequence $(x_1, \ldots, x_p)$ for $\mathcal{F}_1$ and $\mathcal{F}_2$ with $p \geq n-1$. Hence $(x, x_1, \ldots, x_p)$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with length $p+1$ which is at least n.

For case (C2), since one of the cases (a), (b) or (c) must hold, we can apply the reduction $\mathcal{F}_1 = \mathcal{F}[i] - e_x$ and $\mathcal{F}_2 = \mathcal{F}'[i]$ (which occurs in cases (a)(i), (b)(i) and (c)). This implies that x is a component in $\mathcal{F}_1$. So, we can then apply a (C3)-reduction to $\mathcal{F}_1$ and $\mathcal{F}_2$ with $x_i = x$ which results in two forests $\mathcal{F}_1''$ and $\mathcal{F}_2''$ on $X \setminus \{x\}$. By induction, there is a cherry picking sequence $(x_1, \ldots, x_p)$ for $\mathcal{F}_1''$ and $\mathcal{F}_2''$ with $p \geq n-1$. It therefore follows that $(x, x, x_1, \ldots, x_p)$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with length $p+2 > n$. $\square$

Note that in the proof of Theorem 3.2, we did not need to apply either a (C2)(a)(ii), a (C2)(a)(iii), or a (C2)(b)(ii)-reduction. However, as we shall see below, we require these additional reductions to ensure that we obtain a cherry picking sequence with the desired weight in the proof of Theorem 6.1.

4. Bounding the hybrid number from above using cherry picking sequences

The *hybrid number* $h(\mathcal{F},\mathcal{F}')$ of two forests $\mathcal{F}$ and $\mathcal{F}'$ on X is defined to be

$$h(\mathcal{F},\mathcal{F}') = \min\{r(\mathcal{N}) : \mathcal{N} \text{ displays } \mathcal{F} \text{ and } \mathcal{F}'\},$$

where $r(\mathcal{N})$ denotes the reticulation number of $\mathcal{N}$ defined in Section 2. By Proposition 3.2 and Theorem 4.5 (below), note that there always exists a phylogenetic network on X that displays $\mathcal{F}$ and $\mathcal{F}'$. In this and the next two sections, we shall focus on proving the main result of this paper:

Theorem 4.1. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are forests on X. Then*

$$h(\mathcal{F},\mathcal{F}') = \min\{w(\sigma) : \sigma \textit{ is a cherry picking sequence for } \mathcal{F} \textit{ and } \mathcal{F}'\}.$$

To prove Theorem 4.1, first note that the right-hand side of the equality in Theorem 4.1 exists by Proposition 3.2. In this section, we shall prove a result (Theorem 4.5 below) from which it directly follows that the hybrid number for two forests can be no larger than the weight of an optimal cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$. Then, after proving some supporting lemmas in the next section, in Section 6 we shall prove a result (Theorem 6.1), from which it directly follows that the weight of an optimal cherry picking sequence for two forests is no larger than the hybrid number for the forests. The proof of Theorem 4.1 then follows immediately.

Before proceeding, denoting by $d_{TBR}(\mathcal{T},\mathcal{T}')$ the TBR distance between two phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ on X [1, 26], we note that as an immediate corollary of Theorem 4.1 and [14, Theorem 3] we obtain:

Theorem 4.2. *Suppose that $\mathcal{T}$ and $\mathcal{T}'$ are phylogenetic trees in X. Then*

$$d_{TBR}(\mathcal{T},\mathcal{T}') = \min\{w(\sigma) : \sigma \textit{ is a cherry picking sequence for } \mathcal{T} \textit{ and } \mathcal{T}'\}.$$

To prove Theorem 4.5 we will use the following two lemmas.

Lemma 4.3. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are forests on X, $|X| \geq 2$, such that either (i) both $\mathcal{F}$ and $\mathcal{F}'$ contain a cherry (x,y), $x,y \in X$, or (ii) $\mathcal{F}$ contains a component x, $x \in X$. If $\mathcal{N}'$ is a phylogenetic network on $X\backslash\{x\}$ that displays $\mathcal{F}-x$ and $\mathcal{F}'-x$, then there is a phylogenetic network $\mathcal{N}$ on X with $r(\mathcal{N}) = r(\mathcal{N}')$ that displays $\mathcal{F}$ and $\mathcal{F}'$.*

Proof. We obtain a phylogenetic network $\mathcal{N}$ from $\mathcal{N}'$ as follows. For (i) we insert a new pendant edge containing x into the pendant edge in $\mathcal{N}'$ containing y to obtain $\mathcal{N}$.

For (ii), suppose that x is a leaf in some tree $\mathcal{T}'$ in $\mathcal{F}'$. If $\mathcal{T}'$ is equal to x, then we insert a new pendant edge containing x into any edge of $\mathcal{N}'$ to obtain $\mathcal{N}$, and if $\mathcal{T}'$ is the cherry (x, y), $y \in X$, then we insert a new pendant edge containing x into the pendant edge in $\mathcal{N}'$ containing y to obtain $\mathcal{N}$. Now, suppose $|L(\mathcal{T}')| \geq 3$. Then there exists a vertex u in $\mathcal{T}'$ that is adjacent with x and two further such edges in $\mathcal{T}'$. Let e be the edge in $\mathcal{T}' - x$ that results from removing the pendant edge containing x from $\mathcal{T}'$ and suppressing u. We then insert a pendant edge with leaf x anywhere into an edge contained in the image of e in $\mathcal{N}'$ to obtain $\mathcal{N}$.

In all of these cases, clearly $\mathcal{N}$ displays $\mathcal{F}$ and $\mathcal{F}'$, and $r(\mathcal{N}) = r(\mathcal{N}')$. □

Lemma 4.4. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are forests on X, $|X| \geq 2$, and that e is an edge in $\mathcal{F}'$. If $\mathcal{N}'$ is a phylogenetic network on X that displays $\mathcal{F}$ and $\mathcal{F}' - e$, then there is a phylogenetic network $\mathcal{N}$ on X that displays $\mathcal{F}$ and $\mathcal{F}'$ with $r(\mathcal{N}) = r(\mathcal{N}') + 1$.*

Proof. Suppose that $\mathcal{N}'$ displays $\mathcal{F}$ and $\mathcal{F}' - e$. If e is an isolated cherry $\{x, y\}$ in $\mathcal{F}'$, $x, y \in X$, then insert a new edge into $\mathcal{N}'$ whose end vertices subdivide the two pendant edges in $\mathcal{N}'$ that contain x and y. If $e = \{u, x\}$ contains a leaf x, $x \in X$, and a vertex u with degree 3 in $\mathcal{F}'$, then let f be the edge in $\mathcal{F}' - e$ that results from suppressing u after removing e from $\mathcal{F}'$, and insert a new edge into $\mathcal{N}'$ whose end vertices subdivide the pendant edge containing x in $\mathcal{N}'$ and any edge in the image of the edge f in $\mathcal{N}'$. And, if $e = \{u, v\}$, with u and v both having degree 3 in $\mathcal{F}$, then let f and g be the two edges in $\mathcal{F}' - e$ that result from suppressing the vertices u and v in $\mathcal{F} - e$, and insert a new edge into $\mathcal{N}'$ whose end vertices subdivide any edge in the image of f in $\mathcal{N}'$ and any edge in the image of g in $\mathcal{N}'$ (noting that the images of f and g must be disjoint since $\mathcal{N}'$ displays $\mathcal{F} - e$, and f and g are contained in different components of $\mathcal{F} - e$).

In all three cases it is straight-forward to see that $\mathcal{N}$ is a phylogenetic network that displays $\mathcal{F}$ and $\mathcal{F}'$, and that $r(\mathcal{N}) = r(\mathcal{N}') + 1$. □

Theorem 4.5. *Suppose that σ is a cherry picking sequence for two forests $\mathcal{F}$ and $\mathcal{F}'$ on X. Then there is a phylogenetic network $\mathcal{N}$ on X that displays $\mathcal{F}$ and $\mathcal{F}'$ with $w(\sigma) \geq r(\mathcal{N})$.*

Proof. Clearly, we can assume $|X| \geq 2$. Let $\sigma = (x_1, x_2, \ldots, x_m)$, $m \geq 2$, so that $\mathcal{F}[1] = \mathcal{F}$ and $\mathcal{F}'[1] = \mathcal{F}'$ in the definition of a cherry picking sequence. We use induction on the weight $w(\sigma)$ of σ.

Base Case: Suppose $w(\sigma) = 0$. Then only the reductions in (C1) or (C3) are applied at stage i in σ, for all $1 \leq i \leq m - 1$. Now, clearly the two forests $\mathcal{F}[m-1]$ and $\mathcal{F}'[m-1]$ are displayed by the phylogenetic tree $\{x_{m-1}, x_m\}$. So, by applying Lemma 4.3 $(m-2)$-times it follows that $\mathcal{F}$ and $\mathcal{F}'$ are displayed by some phylogenetic tree $\mathcal{T}$ on X. Since $w(\sigma) = r(\mathcal{T}) = 0$, the base case holds.

Now suppose that the theorem holds for all cherry picking sequences σ' for two forests with $0 \leq w(\sigma') \leq w(\sigma) - 1$, and that $w(\sigma) \geq 1$. Let $1 \leq i \leq m - 1$ be the smallest i such that $c(x_i) = 1$, i.e. i is the first time that we apply a (C2)-reduction in σ.

Suppose $i > 1$. If we can find a phylogenetic network $\mathcal{N}'$ that displays $\mathcal{F}[i]$ and $\mathcal{F}'[i]$ so that, for the cherry picking sequence $\sigma' = (x_i, \ldots, x_m)$ for $\mathcal{F}[i]$ and $\mathcal{F}'[i]$, we have $w(\sigma') \geq r(\mathcal{N}')$, then by applying Lemma 4.3 $(i-1)$ times to the sequence $(x_1, \ldots, x_{i-1})$, it follows that there is a phylogenetic network $\mathcal{N}$ that displays $\mathcal{F}$ and $\mathcal{F}'$ with $w(\sigma) = w(\sigma') \geq r(\mathcal{N}') = r(\mathcal{N})$. Thus we may assume without loss of generality that $i = 1$.

So, let $\sigma = (x_1, \ldots, x_m)$, $m \geq 2$, be a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with $w(\sigma) > 0$ and such that a (C2)-reduction is applied to x_1. Then by induction, there must be a phylogenetic network $\mathcal{N}'$ on $\{x_1, \ldots, x_m\}$ that displays $\mathcal{F}[2]$ and $\mathcal{F}'[2]$ with $w((x_1, \ldots, x_m)) \geq r(\mathcal{N}')$. But since a (C2)-reduction corresponds to removing an edge from one of $\mathcal{F}$ or $\mathcal{F}'$, by Lemma 4.4 it follows that there is a phylogenetic network $\mathcal{N}$ on X that displays $\mathcal{F}$ and $\mathcal{F}'$ with

$$r(\mathcal{N}) = r(\mathcal{N}') + 1 \geq w(\sigma') + 1 = w(\sigma).$$

This completes the proof of the theorem. □

We end this section by noting that the proof of Theorem 4.5 in combination with the constructive proofs of Lemmas 4.3 and 4.4 lend themselves to a polynomial-time algorithm that reconstructs a phylogenetic network $\mathcal{N}$ from a given cherry picking sequence σ such that $w(\sigma) \geq r(\mathcal{N})$.

5. Four technical lemmas

In this section, we shall prove four technical lemmas (Lemmas 5.1, 5.3, 5.4 and 5.5) that are concerned with how we can reduce the reticulation number of a phylogenetic network displaying two forests whilst retaining the displaying property. We shall use these lemmas to prove Theorem 6.1 in the next section.

We start by introducing some further definitions. A *pseudo-network* $\mathcal{N}$ *on* X is a connected multi-graph with leaf set $X \subseteq V(\mathcal{N})$ such that all non-leaf vertices have degree three. Note that any multi-edge in a pseudo-network can contain only two edges, and that any pseudo-network that does not contain a multi-edge is a phylogenetic network. Note also that the suppression of a multi-edge and the subsequent suppression of the resulting degree-two vertices may create a new multi-edge. In addition, we extend the notions introduced for phylogenetic networks to pseudo-networks in the natural way.

Now, let $\mathcal{N}$ be a pseudo-network on X. A maximal 2-connected[1] subgraph of $\mathcal{N}$ that is not an edge is called a *blob* of $\mathcal{N}$. In particular, note that a multi-edge in $\mathcal{N}$ could be a blob. Let $\mathcal{B}$ be a blob of $\mathcal{N}$, and let $e = \{u, v\}$ be an edge of $\mathcal{N}$. We say that $\mathcal{B}$ is *incident* with e (or that e is *incident* with $\mathcal{B}$) if precisely one of u and v is contained in $V(\mathcal{B})$. Furthermore, if e is incident with $\mathcal{B}$ and contains a leaf $x \in X$, then we say that $\mathcal{B}$ is *incident* with x. We denote the subset of leaves of $\mathcal{N}$

[1] A graph G is *2-connected* if the graph obtained by removing any vertex from G is connected.

that are incident with $\mathcal{B}$ by $L(\mathcal{B})$. Now let E be the subset of edges of $\mathcal{N}$ that are incident with $\mathcal{B}$. We refer to $\mathcal{B}$ as a *pendant blob* of $\mathcal{N}$ if at most one edge in E is a non-trivial cut-edge of $\mathcal{N}$ and all other edges in E are incident with a leaf.

In what follows, for a pseudo-network $\mathcal{N}$ on X that displays two forests $\mathcal{F}$ and $\mathcal{F}'$, we shall be interested in understanding how $r(\mathcal{N})$ changes if we transform $\mathcal{N}$ into a phylogenetic network that displays $\mathcal{F}$ and $\mathcal{F}'$ by removing certain leaves or edges. To this end, suppose that $|X| \geq 2$, and $\mathcal{N}$ contains precisely one multi-edge. We shall call a sequence $\mathcal{N}_1, \ldots, \mathcal{N}_k$, $k \geq 2$, of distinct pseudo-networks on X a *simplification sequence for* $\mathcal{N}$ if $\mathcal{N} = \mathcal{N}_1$, each $\mathcal{N}_{i+1}$ is obtained from $\mathcal{N}_i$ by suppressing a single multi-edge in $\mathcal{N}_i$ and suppressing the resulting vertices of degree two for all $1 \leq i \leq k-1$, and $\mathcal{N}_k$ is a phylogenetic network on X.

Lemma 5.1. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are two forests on X with $|X| \geq 2$ and that $\mathcal{N}$ is a pseudo-network on X with precisely one multi-edge such that $\mathcal{N}$ displays $\mathcal{F}$ and $\mathcal{F}'$. In addition, suppose that the blob in $\mathcal{N}$ that contains the multi-edge is incident with at least two cut-edges of $\mathcal{N}$. Then there exists a simplification sequence $\mathcal{N}_1, \ldots, \mathcal{N}_k$ for $\mathcal{N}$, for some $k \geq 2$. Moreover, $\mathcal{N}_k$ displays $\mathcal{F}$ and $\mathcal{F}'$ and $r(\mathcal{N}) > r(\mathcal{N}_k)$.*

Proof. Let $\mathcal{B}$ be the blob in $\mathcal{N}$ that contains the unique multi-edge in $\mathcal{N}$, and let u and v be the two vertices contained in the multi-edge. In addition let p and q be the vertices in $\mathcal{N}$ such that $\{u,p\}$ and $\{v,q\}$ are in $E(\mathcal{N})$. Then since $\mathcal{B}$ is incident with at least two cut-edges it follows that $p \neq q$. Thus, the pseudo-network obtained by suppressing the multi-edge and any resulting degree-two vertex is either a phylogenetic network on X, or contains a blob that is incident with at least two cut-edges and contains a unique multi-edge. This immediately implies that there is a unique simplification sequence $\mathcal{N}_1, \ldots, \mathcal{N}_k$ for $\mathcal{N}$, for some $k \geq 2$.

We now show that $\mathcal{N}_i$ displays the forests $\mathcal{F}$ and $\mathcal{F}'$, for all $1 \leq i \leq k$. For some $i \in \{1, \ldots, k-1\}$, suppose that $\mathcal{N}_i$ displays the forests $\mathcal{F}$ and $\mathcal{F}'$, that $e = \{u,v\}$ and $e' = \{u,v\}$ are edges of $\mathcal{N}_i$ and that, e is suppressed to obtain $\mathcal{N}_{i+1}$. Choose images for $\mathcal{F}$ and $\mathcal{F}'$ in $\mathcal{N}$, respectively. It is straight-forward to check that if e is not contained in the image of $\mathcal{F}$ or $\mathcal{F}'$, then the pseudo-network $\mathcal{N}_{i+1}$ displays $\mathcal{F}$ and $\mathcal{F}'$. Moreover, if one of e and e', say e, is contained in the image of $\mathcal{F}$ and e' is contained in the image of $\mathcal{F}'$, then clearly we can alter the image of $\mathcal{F}$ so that only e' is contained in the image of $\mathcal{F}$ and $\mathcal{F}'$. Since $\mathcal{N}$ displays $\mathcal{F}$ and $\mathcal{F}'$, it follows that $\mathcal{N}_{i+1}$ also does for all $1 \leq i \leq k-1$.

Finally, we note that $r(\mathcal{N}) > r(\mathcal{N}_k)$ since clearly $r(\mathcal{N}_i) = r(\mathcal{N}_{i+1}) + 1$, for all $1 \leq i \leq k-1$. □

We now consider the effect of removing certain leaves from a phylogenetic network that displays two forests. To this end, we define for a phylogenetic network $\mathcal{N}$ on X and $x \in X$, the graph $\mathcal{N} - x$ to be the pseudo-network on $X \setminus \{x\}$ obtained from $\mathcal{N}$ by removing x, the pendant edge $e_x \in E(\mathcal{N})$, and suppressing the resulting degree-two vertex. We first make a simple observation whose proof is straight-forward.

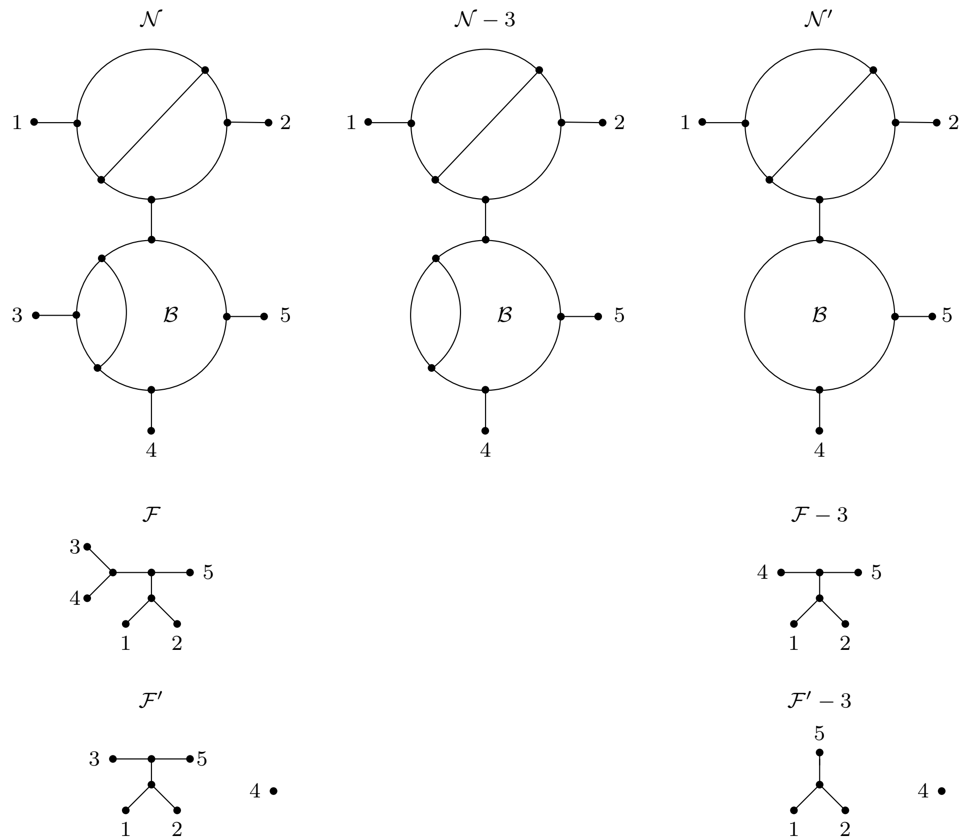

FIGURE 4. An explicit example of Lemma 5.3 applied to $x = 3$ of a phylogenetic network $\mathcal{N}$ on $\{1, 2, \ldots, 5\}$ with a pendant blob $\mathcal{B}$ that displays two forests $\mathcal{F}$ and $\mathcal{F}'$ on $\{1, 2 \ldots, 5\}$. Then $\mathcal{N} - 3$ is not a phylogenetic network, and there exists a phylogenetic network $\mathcal{N}'$ on $\{1, 2, 4, 5\}$ with $r(\mathcal{N}) > r(\mathcal{N}')$ that displays $\mathcal{F} - 3$ and $\mathcal{F}' - 3$.

Observation 5.2. *Suppose that $\mathcal{N} = (V, E)$ is a phylogenetic network on X with $|X| \geq 2$, $x \in X$, and $\{u, x\}$, $\{u, v\}, \{u, w\} \in E$ with $u, v, w \in V \setminus \{x\}$. Then $\mathcal{N} - x$ is a pseudo-network, and $\mathcal{N} - x$ is a phylogenetic network on $X \setminus \{x\}$ if and only if $\{v, w\} \notin E$.*

Using Observation 5.2, we obtain the following lemma, which is illustrated in Figure 4.

Lemma 5.3. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are two forests on X with $|X| \geq 2$ and that $\mathcal{N}$ is a phylogenetic network on X that displays $\mathcal{F}$ and $\mathcal{F}'$. In addition, suppose that there is a pendant blob $\mathcal{B}$ of $\mathcal{N}$ with $|L(\mathcal{B})| \geq 2$ and some $x \in L(\mathcal{B})$ such that the pseudo-network $\mathcal{N} - x$ on $X \setminus \{x\}$* is not *a phylogenetic network. Then there exists a phylogenetic network $\mathcal{N}'$ on $X \setminus \{x\}$ such that $r(\mathcal{N}) > r(\mathcal{N}')$ and $\mathcal{N}'$ displays $\mathcal{F} - x$ and $\mathcal{F}' - x$.*

Proof. Suppose first that $X = \{x, y\}$. Then we can take $\mathcal{N}'$ to be the vertex y to complete the proof of the lemma.

Now, suppose that $|X| \geq 3$. Note that $\mathcal{N} - x$ clearly displays $\mathcal{F} - x$ and $\mathcal{F}' - x$. Let v and w be the two vertices adjacent to the non-leaf vertex of e_x. Since $\mathcal{N}$ is a phylogenetic network, we have $v \neq w$. Moreover, by Observation 5.2, $\{v, w\}$ is an edge of $\mathcal{N}$ and $\mathcal{N} - x$ contains the unique multi-edge between v and w. This multi-edge is contained in a blob $\mathcal{B}'$ of $\mathcal{N} - x$. Since $\mathcal{B}$ is a pendant blob in $\mathcal{N}$, it follows that $\mathcal{B}'$ is a pendant blob in $\mathcal{N} - x$. Furthermore, since $|L(\mathcal{B})| \geq 2$ and $|X| \geq 3$, either $|L(\mathcal{B}')| \geq 2$, or $|L(\mathcal{B}')| = 1$ and $\mathcal{B}'$ is incident with at least one non-trivial cut-edge. The proof of the lemma now follows immediately by applying Lemma 5.1 to $\mathcal{N} - x$ and $\mathcal{B}'$. □

We now consider what happens when we remove certain edges from a phylogenetic network that displays two forests.

Lemma 5.4. *Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are two forests on X with $|X| \geq 2$ and that $\mathcal{N}$ is a phylogenetic network on X that displays $\mathcal{F}$ and $\mathcal{F}'$. Let $\mathcal{N}[\mathcal{F}]$ and $\mathcal{N}[\mathcal{F}']$ denote images of $\mathcal{F}$ and $\mathcal{F}'$ in $\mathcal{N}$, respectively. If e is an edge in a pendant blob $\mathcal{B}$ of $\mathcal{N}$ with $|L(\mathcal{B})| \geq 2$ and*

- (a) *e is neither contained in $\mathcal{N}[\mathcal{F}]$ nor in $\mathcal{N}[\mathcal{F}']$, then there is a phylogenetic network $\mathcal{N}'$ on X that displays $\mathcal{F}$ and $\mathcal{F}'$ with $r(\mathcal{N}) > r(\mathcal{N}')$.*
- (b) *e is not in $\mathcal{N}[\mathcal{F}']$ but there exists an edge f in $\mathcal{F}$ such that e is in the image of f in $\mathcal{N}[\mathcal{F}]$, then there is a phylogenetic network $\mathcal{N}'$ on X that displays $\mathcal{F} - f$ and $\mathcal{F}'$ and $r(\mathcal{N}) > r(\mathcal{N}')$.*

Proof. Let e be as in the statement of the lemma. Remove e from $\mathcal{B}$ and suppress the two resulting degree-two vertices. Then it is straight-forward to check that the resulting multi-graph $\mathcal{N}'$ is a pseudo-network on X with $r(\mathcal{N}) > r(\mathcal{N}')$ that displays $\mathcal{F}$ and $\mathcal{F}'$ in case (a) and $\mathcal{F} - f$ and $\mathcal{F}'$ in case (b).

If $\mathcal{N}'$ is a phylogenetic network, this completes the proof of the lemma. Otherwise, since by assumption $\mathcal{B}$ must be incident with at least two cut-edges, it follows that we can apply Lemma 5.1 to complete the proof. □

Our final lemma in this section can be used to remove two special types of pendant blobs from a network. Before stating the lemma, we first need a new definition. Suppose that $\mathcal{N}$ is a phylogenetic network on X, $|X| \geq 1$, and that $\mathcal{B}$ is a pendant blob of $\mathcal{N}$ with $|L(\mathcal{B})| \leq 1$. If $\mathcal{N} \neq \mathcal{B}$, then let $\{u, v\}$ be the cut-edge that disconnects $\mathcal{B}$ from $\mathcal{N}$ with $v \in V(\mathcal{B})$. We define the phylogenetic network $\mathcal{N} - \mathcal{B}$ as follows:

- Suppose that $|L(\mathcal{B})| = 0$, so that $\mathcal{N} \neq \mathcal{B}$ as $\mathcal{N}$ has a non-empty leaf set. Then first delete all vertices of $\mathcal{B}$ and any edges containing them to obtain a graph $\mathcal{G}$. If the pseudo-network obtained by then suppressing u in $\mathcal{G}$ is a phylogenetic network, then this is defined to be $\mathcal{N} - \mathcal{B}$. Otherwise, suppressing u in $\mathcal{G}$ must result in a multi-edge that contains the two vertices of $\mathcal{G}$ adjacent with u, say w, w'. To obtain $\mathcal{N} - \mathcal{B}$ we subdivide the edge $\{w, w'\}$ in the graph $\mathcal{G}$ by adding in a new vertex u' and then add the edge $\{u, u'\}$.

- Suppose $|L(\mathcal{B})| = 1$, so that $L(\mathcal{B}) = \{x\}$, for some $x \in X$. If $\mathcal{N} = \mathcal{B}$, then define $\mathcal{N} - \mathcal{B} = \{x\}$. Otherwise, we let $\mathcal{N} - \mathcal{B}$ be the phylogenetic network obtained by first removing all vertices in $\mathcal{B}$ and edges that contain them from $\mathcal{N}$ and then adding the edge $\{x, u\}$ (i.e. the graph obtained by replacing $\mathcal{B}$ with leaf x).

Lemma 5.5. *Suppose that $\mathcal{N}$ displays two forests $\mathcal{F}$ and $\mathcal{F}'$ and that $\mathcal{N}$ is a phylogenetic network that contains a pendant blob $\mathcal{B}$ with $|L(\mathcal{B})| \leq 1$. Then $\mathcal{N} - \mathcal{B}$ displays $\mathcal{F}$ and $\mathcal{F}'$ and $r(\mathcal{N}) > r(\mathcal{N} - \mathcal{B})$.*

Proof. It is straight-forward to see that $\mathcal{N} - \mathcal{B}$ displays $\mathcal{F}$ and $\mathcal{F}'$.

Now consider the blob $\mathcal{B}$. We first show that $r(\mathcal{B}) \geq 2$. Towards a contradiction, assume that $r(\mathcal{B}) < 2$. Then $r(\mathcal{B}) = 1$, since $r(\mathcal{B}) \neq 0$ as otherwise $\mathcal{B}$ would be a tree which is impossible as it is 2-connected. Consider the graph $\mathcal{B}'$ that is obtained from $\mathcal{B}$ by suppressing all vertices in $\mathcal{B}$ with degree 2. Then $\mathcal{B}'$ is 2-connected and every vertex in $\mathcal{B}'$ must have degree 3. Moreover, since $r(\mathcal{B}') = r(\mathcal{B}) = 1$ there must be an edge whose removal from $\mathcal{B}'$ gives a tree, since a connected graph G with $r(G) = 1$ contains a unique cycle [3, Corollary 2, p.29]. But, if we remove an edge from $\mathcal{B}'$ then every vertex in the resulting graph must have degree at least 2, and in a tree that is not a single vertex there must be at least two vertices with degree 1, a contradiction. So $r(\mathcal{B}) \geq 2$.

Next, suppose $|L(\mathcal{B})| = 0$. Then if the pseudo-network obtained by suppressing u in $\mathcal{G}$ in the definition of $\mathcal{N} - \mathcal{B}$ is a phylogenetic network, then it is straight-forward to check that $r(\mathcal{N} - \mathcal{B}) = r(\mathcal{N}) - r(\mathcal{B})$, and so $r(\mathcal{N}) > r(\mathcal{N} - \mathcal{B})$. Otherwise, it is straight-forward to check that $r(\mathcal{N} - \mathcal{B}) = r(\mathcal{N}) - r(\mathcal{B}) + 1$. Hence, as $r(\mathcal{B}) \geq 2$, it follows that $r(\mathcal{N}) > r(\mathcal{N} - \mathcal{B})$.

And, finally, it is straight-forward to check that if $|L(\mathcal{B})| = 1$, then $r(\mathcal{N} - \mathcal{B}) = r(\mathcal{N}) - r(\mathcal{B})$ and so $r(\mathcal{N}) > r(\mathcal{N} - \mathcal{B})$. □

6. Bounding the weight of a cherry picking sequence from above by the hybrid number

In this section, we complete the proof of Theorem 4.1 by establishing the following result.

Theorem 6.1. *Suppose that $\mathcal{N}$ is a phylogenetic network on X that displays two forests $\mathcal{F}$ and $\mathcal{F}'$ on X. Then there is a cherry picking sequence σ for $\mathcal{F}$ and $\mathcal{F}'$ with $r(\mathcal{N}) \geq w(\sigma)$.*

Before proving this result, we first state a useful lemma that will help us to deal with (C1) and (C3) reductions. Its proof is straight-forward and left to the reader.

Lemma 6.2. *Suppose that $\mathcal{N}$ is a phylogenetic network on X, $|X| \geq 2$, and that $\mathcal{F}$ and $\mathcal{F}'$ are forests on X that are displayed by $\mathcal{N}$. If $x \in X$ is such that $\mathcal{N} - x$ is a phylogenetic network then $r(\mathcal{N} - x) = r(\mathcal{N})$, and $\mathcal{N} - x$ displays both $\mathcal{F} - x$ and $\mathcal{F}' - x$. Moreover, if $\mathcal{N}$ has a cherry, then the cherry contains a leaf x such that $\mathcal{N} - x$ is a phylogenetic network and either (a) there is a cherry (x, y) contained in both $\mathcal{F}$ and $\mathcal{F}'$, $y \in X \setminus \{x\}$, or (b) x is a component in one of $\mathcal{F}$ or $\mathcal{F}'$.*

Before proving Theorem 6.1, as the proof is quite technical, we give a brief road-map of its main points. The proof works using induction on $r(\mathcal{N})$. In the base case we have $r(\mathcal{N}) = 0$, and so $\mathcal{N}$ is a phylogenetic tree. In this case we can continually pick off leaves that are contained in cherries to obtain a cherry picking sequence σ with $w(\sigma) = r(\mathcal{N}) = 0$.

Now, in case $r(\mathcal{N}) > 0$, we essentially repeatedly look for non cut-edges e in $\mathcal{N}$ that are contained in the image of one of the forests displayed by $\mathcal{N}$ but not in the other. If we can find such an edge e, by carefully considering how the images of the forests are embedded in $\mathcal{N}$, we can then complete the proof by removing e from $\mathcal{N}$, applying Lemmas 5.3 and 5.4 and some (C2)-reduction, and using the fact that $r(\mathcal{N}) > r(\mathcal{N} - e)$. We use this tool within a series of claims to show that either we can prove the theorem by induction, or that we can further restrict the way that the forests are embedded in $\mathcal{N}$. This eventually allows us to finally apply our induction hypothesis one final time and thus complete the proof.

More specifically, we first show that we can assume that $\mathcal{N}$ satisfies a certain property (P). Then, after noting that $\mathcal{N}$ must be pendantless (Claim 1), we show that we can also assume that $\mathcal{N}$ must have a pendant blob $\mathcal{B}$ with at least two leaves such that no leaf of $\mathcal{B}$ is a component of $\mathcal{F}$ or $\mathcal{F}'$ and every edge in $\mathcal{B}$ is in $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$ (Claims 2, 3 and 4). In Claims 5 and 6 we then show that either $\mathcal{F}$ or $\mathcal{F}'$ must contain a cherry (x, y) that, out of the four possibilities (i)–(iv) in Observation 3.1, only satisfies (ii). Then in Claims 7–11 we show that we can assume that there is a component in $\mathcal{F}$ that contains a cherry (x, y) and a component in $\mathcal{F}'$ that contains a cherry (x, z) with $x, y, z \in L(\mathcal{B})$. This allows us to assume that $((x, y), p)$ or $((x, y), (p, q))$ is a pendant subtree of one of the forests for some $x, y, p, q \in L(\mathcal{B})$ distinct. This permits us to carry out a case analysis in Claims 12–14 in order to complete the proof.

Proof of Theorem 6.1: In what follows, for two sequences $\sigma = (x_1, \ldots, x_m)$, $m \geq 1$, and $\sigma' = (y_1, \ldots, y_l)$, $l \geq 1$, we denote by $\sigma \circ \sigma'$ the sequence $(x_1, \ldots, x_m, y_1, \ldots, y_l)$.

We prove the theorem using induction on $r(\mathcal{N}) \geq 0$.

Base Case: Suppose $r(\mathcal{N}) = 0$. Then $\mathcal{N}$ is a phylogenetic tree on X. If $|X| = 1$, then $X = \{x\}$, for some x, and (x) is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with $w((x)) = 0$. Now, suppose $m = |X| > 1$. Order the elements in X as $x_1, x_2, \ldots, x_m$ so that, for all $1 \leq i \leq m - 1$, x_i is contained in a cherry of $\mathcal{N}|\{x_i, \ldots, x_m\}$ which is possible since any phylogenetic tree with two or more leaves always contains a cherry. Then, by applying Lemma 6.2 and using either a (C1) or (C3) reduction for each x_i, $1 \leq i \leq m - 1$, depending on whether (a) or (b) holds in Lemma 6.2

for x_i, respectively, it follows that $(x_1, \ldots, x_m)$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ with $r(\mathcal{N}) = w((x_1, \ldots, x_m)) = 0$. Hence the base case holds.

Now, suppose that q is a positive integer and that the theorem holds for all phylogenetic networks with reticulation number strictly less than q. Suppose that $\mathcal{N}$ is a phylogenetic network with $r(\mathcal{N}) = q$ that displays $\mathcal{F}$ and $\mathcal{F}'$. Then to complete the proof of the theorem we need to show that:

(CP) There is a cherry picking sequence σ for $\mathcal{F}$ and $\mathcal{F}'$ with $r(\mathcal{N}) \geq w(\sigma)$.

Note first that we will assume from now on that $|X| \geq 3$ since for $|X| \in \{1, 2\}$, we can clearly find a cherry picking sequence σ such that $w(\sigma) = 0$ and so $r(\mathcal{N}) > 0 = w(\sigma)$ and hence (CP) holds.

Starting with $\mathcal{N}$, repeatedly look for elements $x \in X$ that satisfy (a) there is a cherry (x, y) contained in both $\mathcal{F}$ and $\mathcal{F}'$, $y \in X \setminus \{x\}$, or (b) x is a component in either $\mathcal{F}$ or $\mathcal{F}'$, and such that removing x and e_x and suppressing the resulting degree-2 vertex results in a phylogenetic network until this is no longer possible. Then we obtain a sequence $(x_1, \ldots, x_m)$, $x_i \in X$, $m \geq 1$, where, for $\mathcal{N}_1 = \mathcal{N}$, $\mathcal{N}_{i+1} = \mathcal{N}_i - x_i$, $\mathcal{F}[1] = \mathcal{F}$, $\mathcal{F}[i+1] = \mathcal{F}[i] - x_i$, $\mathcal{F}'[1] = \mathcal{F}'$ and $\mathcal{F}'[i+1] = \mathcal{F}'[i] - x_i$, $1 \leq i \leq m$, we have for all $1 \leq i \leq m$ that x_i satisfies (a) or (b) and $\mathcal{N}_{i+1} = \mathcal{N}_i - x_i$ is a phylogenetic network on $X \setminus \{x_1, \ldots, x_i\}$, $\mathcal{N}_i$ displays the forests $\mathcal{F}[i]$, $\mathcal{F}'[i]$ and, by Lemma 6.2, $r(\mathcal{N}_i) = r(\mathcal{N})$.

In particular, $\mathcal{N}_{m+1}$ is a phylogenetic network on $X \setminus \{x_1, \ldots, x_m\}$ with $r(\mathcal{N}_{m+1}) = q$ that displays $\mathcal{F}[m+1]$ and $\mathcal{F}'[m+1]$, and there is no $x \in X \setminus \{x_1, \ldots, x_m\}$ such that $\mathcal{N}_{m+1} - x$ is a phylogenetic network and either (a) or (b) holds.

Now suppose that we could show that (CP) holds for $\mathcal{N}_{m+1}$ along with the forests $\mathcal{F}[m+1]$ and $\mathcal{F}'[m+1]$ (which we shall do below). Then there must exist a cherry picking sequence σ' for $\mathcal{F}[m+1]$ and $\mathcal{F}'[m+1]$ with $r(\mathcal{N}_{m+1}) \geq w(\sigma')$. Hence, $\sigma = (x_1, \ldots, x_m) \circ \sigma'$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$. Since we have applied a (C1) or (C3) reduction to each x_i depending on whether (a) or (b) holds for x_i, respectively, we obtain $w(\sigma') = w(\sigma)$. Thus $r(\mathcal{N}) = r(\mathcal{N}_{m+1}) \geq w(\sigma') = w(\sigma)$ and so (CP) holds for $\mathcal{N}$.

Thus, to show that (CP) holds for $\mathcal{N}$, we assume from now on that $\mathcal{N}$ satisfies the following property:

(P): There is no $x \in X$ such that $\mathcal{N} - x$ is a phylogenetic network on $X \setminus \{x\}$ and either (a) there is a cherry (x, y) contained in both $\mathcal{F}$ and $\mathcal{F}'$, $y \in X \setminus \{x\}$, or (b) x is a component in either $\mathcal{F}$ or $\mathcal{F}'$.

Claim 1: $\mathcal{N}$ is pendantless.

Proof of Claim 1: Suppose $\mathcal{N}$ is *not* pendantless. Then $\mathcal{N}$ contains a cherry. Thus by Lemma 6.2 there must be some $x \in X$ such that $\mathcal{N} - x$ is a phylogenetic network and (a) or (b) is satisfied for x. But this contradicts the fact that $\mathcal{N}$ satisfies property (P). ■

By Claim 1 there is some pendant blob in $\mathcal{N}$. Let $\mathcal{B}$ be such a blob.

Claim 2: $|L(\mathcal{B})| \geq 2$.

Proof of Claim 2: If $|L(\mathcal{B})| \leq 1$, then by applying Lemma 5.5 to $\mathcal{N}$ and using induction it follows that (CP) holds. ■

Claim 3: No leaf in $L(\mathcal{B})$ is a component in $\mathcal{F}$ or $\mathcal{F}'$.

Proof of Claim 3: Suppose that some $x \in L(\mathcal{B})$ is a component of $\mathcal{F}$. Then by property (P), the pseudo-network $\mathcal{N} - x$ is not a phylogenetic network. So we can apply a (C3)-reduction to x and, since $|L(\mathcal{B})| \geq 2$ by Claim 2, we can apply Lemma 5.3 to $\mathcal{F}$, $\mathcal{F}'$, $\mathcal{N}$ and x to see that (CP) must hold for $\mathcal{N}$ by induction. ■

In what follows, we choose images $\mathcal{N}[\mathcal{F}]$ and $\mathcal{N}[\mathcal{F}']$ of $\mathcal{F}$ and $\mathcal{F}'$ in $\mathcal{N}$, respectively, as defined in Section 2.

Claim 4: Every pendant edge incident to $\mathcal{B}$ and every edge in $\mathcal{B}$ is contained in $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$ (or both).

Proof of Claim 4: If a pendant edge were not in $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$, then either $\mathcal{F}$ or $\mathcal{F}'$ would contain an element in $L(\mathcal{B})$ that is a component in $\mathcal{F}$ or $\mathcal{F}'$, which contradicts our assumption in Claim 3. And, if there is some $e \in E(\mathcal{B})$ such that e is neither contained in $\mathcal{N}[\mathcal{F}]$ nor in $\mathcal{N}[\mathcal{F}']$, then since $|L(\mathcal{B})| \geq 2$ by Claim 2, Lemma 5.4(a) implies that there exists a phylogenetic network $\mathcal{N}'$ on X that displays $\mathcal{F}$ and $\mathcal{F}'$ and $r(\mathcal{N}) > r(\mathcal{N}')$. We can then apply induction to see that (CP) holds for $\mathcal{N}$. ■

Claim 5: Suppose that (x, y) is a cherry in a component $\mathcal{T}$ of $\mathcal{F}$ or $\mathcal{F}'$ with $x, y \in L(\mathcal{B})$, and γ is the path in $\mathcal{T}$ between x and y. If u, v are the vertices in $\mathcal{B}$ adjacent to x, y, respectively, then $\{x, u\}$ and $\{y, v\}$ are pendant edges of $\mathcal{B}$, the image $\mathcal{N}[\gamma]$ of γ contains both $\{x, u\}$ and $\{y, v\}$ and has length at least three, and every edge in $\mathcal{N}[\gamma]$ not equal to $\{x, u\}$ and $\{y, v\}$ is contained in $E(\mathcal{B})$.

Proof of Claim 5: Since $\mathcal{N}$ is pendantless by Claim 1, $u \neq v$ and $\{x, u\}$ and $\{y, v\}$ are pendant edges of $\mathcal{B}$. Moreover, u and v are clearly contained in $\mathcal{N}[\gamma]$, and $\mathcal{N}[\gamma]$ must contain at least three edges since $u \neq v$. And every edge in $\mathcal{N}[\gamma]$ not equal to $\{x, u\}$ and $\{y, v\}$ is contained in $E(\mathcal{B})$ since $\mathcal{B}$ is a pendant blob in $\mathcal{N}$. ■

Claim 6: If $\mathcal{T}$ is any tree in $\mathcal{F}$ or in $\mathcal{F}'$ that contains a cherry (x, y) with $x, y \in L(\mathcal{B})$, then out of the four cases (i)–(iv) in Observation 3.1 only case (ii) holds for x and y.

Proof of Claim 6: Assume without loss of generality that $\mathcal{T}$ is a component in $\mathcal{F}$ and that (x, y) is a cherry in $\mathcal{T}$. Let u, v, γ and $\mathcal{N}[\gamma]$ be as in Claim 5.

Suppose Observation 3.1(iv) holds, i.e. x and y are contained in different components in $\mathcal{F}'$ and neither x nor y is in a cherry of $\mathcal{F}'$. Then as x and y are in different components of $\mathcal{F}'$, there must exist an edge e' in $\mathcal{N}[\gamma]$ that is not in $\mathcal{N}[\mathcal{F}']$. So, by Claim 2 and applying Lemma 5.4(b) to e', there exists a phylogenetic network $\mathcal{N}'$ on X with $r(\mathcal{N}) > r(\mathcal{N}')$ that displays $\mathcal{F}'$ and either $\mathcal{F} - e_x$ or $\mathcal{F} - e_y$, say

$\mathcal{F} - e_x$. Since $x, y \in L(\mathcal{B})$ neither x nor y is an isolated vertex in $\mathcal{F}'$ by Claim 3, and so we have applied a (C2)(c)-reduction to $\mathcal{F}$ and $\mathcal{F}'$ to obtain $\mathcal{F} - e_x$. Moreover, by induction, there is a cherry picking sequence σ' for $\mathcal{F} - e_x$ and $\mathcal{F}'$ with $r(\mathcal{N}') \geq w(\sigma')$. Hence, $\sigma = (x) \circ \sigma'$ is a cherry picking sequence for $\mathcal{F}$ and $\mathcal{F}'$ such that $r(\mathcal{N}) \geq r(\mathcal{N}') + 1 \geq w(\sigma') + 1 = w(\sigma)$, and so (CP) must hold.

Now, suppose Observation 3.1(iii) holds, i.e, x and y are leaves in the same tree $\mathcal{T}'$ in $\mathcal{F}'$, but neither x nor y is in a cherry of $\mathcal{T}'$. If there exists an edge in $\mathcal{N}[\gamma]$ that is not contained in $\mathcal{N}[\mathcal{F}']$ then we can apply a similar argument to the one used in the previous paragraph to see that (CP) must hold for $\mathcal{N}$ using a (C2)(b)(i)-reduction applied to x or y and induction. So, suppose every edge in $\mathcal{N}[\gamma]$ is contained in $\mathcal{N}[\mathcal{F}']$. Let e and e' be the edges in $E(\mathcal{B})$ incident to u and v, respectively, but not in $\mathcal{N}[\gamma]$ (so, in particular, neither e nor e' is a cut-edge). Note that $e \neq e'$ since if $e = e'$, then as every edge in $\mathcal{B}$ is contained in $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$ by Claim 4, this would imply that either $\mathcal{F}$ or $\mathcal{F}'$ contains a cycle, a contradiction. Moreover, since every edge in $\mathcal{B}$ is contained in $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$ by Claim 4 and (x, y) is a cherry in $\mathcal{F}$, it follows that one of the edges e, e', say e', is in $\mathcal{N}[\mathcal{F}']$ but not in $\mathcal{N}[\mathcal{F}]$, and that e' is in the image $\mathcal{N}[f]$ of the edge f in $\mathcal{T}'$ that contains the vertex in $\mathcal{T}'$ that is adjacent to y but that is not in the path in $\mathcal{T}'$ between x and y. So, by Lemma 5.4(b) applied to e' with the roles of $\mathcal{F}$ and $\mathcal{F}'$ reversed, there exists a phylogenetic network $\mathcal{N}'$ that displays $\mathcal{F}$ and $\mathcal{F}' - f$ with $r(\mathcal{N}) > r(\mathcal{N}')$. Since this corresponds to applying a (C2)(b)(ii)-reduction to f, we can apply induction again to $\mathcal{N}'$ to see that (CP) must hold.

Last, suppose Observation 3.1(i) holds, i.e., that (x, y) is a cherry in both $\mathcal{F}$ and $\mathcal{F}'$. Note that by property (P), neither the pseudo-network $\mathcal{N} - x$ nor the pseudo-network $\mathcal{N} - y$ is a phylogenetic network. Moreover, every edge in $\mathcal{N}[\gamma]$ must be contained in both $\mathcal{N}[\mathcal{F}]$ and $\mathcal{N}[\mathcal{F}']$, otherwise we could use Lemma 5.4(b) and a (C2)(b)(i)-reduction and induction to see that (CP) must hold. Let e and e' be the edges in $E(\mathcal{B})$ incident to u and v, respectively, but not in $\mathcal{N}[\gamma]$, so that $e \neq e'$ and neither e nor e' is a cut-edge of $\mathcal{N}$. Note that since every edge in $\mathcal{B}$ is contained in $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$ by Claim 4, and (x, y) is a cherry in both $\mathcal{F}$ and $\mathcal{F}'$, it follows without loss of generality, that e is in $\mathcal{N}[\mathcal{F}]$ but not in $\mathcal{N}[\mathcal{F}']$, and that e' is in $\mathcal{N}[\mathcal{F}']$ but not in $\mathcal{N}[\mathcal{F}]$.

We now show that the length of the path $\mathcal{N}[\gamma]$ must be three (noting that we are assuming that it has length at least three). Suppose not. First suppose that the length of $\mathcal{N}[\gamma]$ is four. Let w be the vertex in $\mathcal{N}[\gamma]$ that is distance two from both x and y. Then, by Claim 4, e and e' must be in either $\mathcal{N}[\mathcal{F}]$ or $\mathcal{N}[\mathcal{F}']$. Since (x, y) is a cherry in both $\mathcal{F}$ and $\mathcal{F}'$, it follows that the edge incident to $\mathcal{B}$ and not in $\mathcal{N}[\gamma]$ is neither in $\mathcal{N}[\mathcal{F}]$ nor $\mathcal{N}[\mathcal{F}']$. Hence, w must be contained in the cut-edge of $\mathcal{N}$ that is incident to $\mathcal{B}$. Now, if e and e' have no vertex in common then, since $\mathcal{N} - x$ is a pseudo-network that is not a phylogenetic network, when we remove the edge $\{u, x\}$ from $\mathcal{N}$ and suppress the vertex u, we must obtain a multi-edge that contains the two vertices in $\mathcal{B}$ adjacent to u that are not equal to x. But this is impossible since one of these two vertices must be equal to w, and the degree of w in $\mathcal{N}$ is 3. And, if e and e' have a vertex in common, say w', then $\mathcal{B}$ must be a 4-cycle, with leaves x, y and another leaf z, $z \in X$, such that $\{w', z\}$ is a pendant edge of $\mathcal{B}$. But this is impossible by property (P) since (z, y) is then a cherry in both $\mathcal{F}$ and $\mathcal{F}'$ and $\mathcal{N} - z$ is a phylogenetic network. Finally, note that since (x, y) is cherry in both $\mathcal{F}$ and $\mathcal{F}'$ and $\mathcal{B}$ is a pendant blob, it follows that the length of $\mathcal{N}[\gamma]$ cannot be greater than four by Claim 4.

Now, assume that the length of the path $\mathcal{N}[\gamma]$ is three. If e and e' have no vertex in common, then since $\mathcal{N} - x$ is a pseudo-network that is not a phylogenetic network, when we remove the edge $\{u, x\}$ from $\mathcal{N}$ and suppress the vertex u, we must obtain a multi-edge that contains the two vertices in $\mathcal{B}$ adjacent to u that are not equal to x. But this is impossible since one of these two vertices must be equal to v as $\mathcal{N}[\gamma]$ has length three, and e and e' have no vertex in common. And, finally for the case where $\mathcal{N}[\gamma]$ has length three, if e and e' have a vertex in common, say c, then $\mathcal{B}$ is the 3-cycle u, v, c and there is a vertex $c' \neq u, v$ with $c' \notin X$ such that $\{c, c'\}$ is a non-pendant cut-edge of $\mathcal{N}$. So we can apply a (C1)-reduction to x, make a new network $\mathcal{N}'$ on $X \setminus \{x\}$ with $r(\mathcal{N}) > r(\mathcal{N}')$ by removing x and all of the vertices in $\mathcal{B}$ and the edges in $\mathcal{N}$ containing them and adding the edge $\{c', y\}$, and then apply induction to see that (CP) holds. ■

Claim 7: Neither $\mathcal{F}$ nor $\mathcal{F}'$ contains a cherry in the form of an isolated edge $\{x, y\}$ with $x, y \in L(\mathcal{B})$.

Proof of Claim 7: Suppose that $\mathcal{T}$ is a cherry in $\mathcal{F}$ in the form of the isolated edge $\{x, y\}$ with $x, y \in L(\mathcal{B})$. Let u, v, γ and $\mathcal{N}[\gamma]$ be as in Claim 5. By Claim 6, Observation 3.1(ii) holds and so we may assume that there is a cherry in $\mathcal{F}'$, say (x, z), with $z \in X \setminus \{y\}$. Now, all of the edges in $\mathcal{N}[\gamma]$ must be contained in $\mathcal{N}[\mathcal{F}']$, otherwise we can apply Lemma 5.4(b) to any edge on $\mathcal{N}[\gamma]$ not contained in $\mathcal{N}[\mathcal{F}']$ and a (C2)(a)(i)-reduction to x_i plus induction to see that (CP) holds. But this implies that the edge e in $\mathcal{B}$ that contains the vertex adjacent to x and that is not contained in $\mathcal{N}[\gamma]$ must be contained in the image of the path in $\mathcal{F}'$ between x and z. Indeed, if this were not the case then, as e must be contained in $\mathcal{N}[\mathcal{F}']$ by Claim 4, this would imply that (x, z) is not a cherry in $\mathcal{F}'$. Thus, as $\{x, y\}$ is a component in $\mathcal{F}$ and e is not a cut-edge for $\mathcal{N}$, we can apply Lemma 5.4(b) to e and the roles of $\mathcal{F}$ and $\mathcal{F}'$ reversed to see that (CP) holds by applying a (C2)(a)(i)-reduction to x and induction. ■

Claim 8: There exists some tree $\mathcal{T}$ in $\mathcal{F}$ that contains a cherry (x, y) with $x, y \in L(\mathcal{B})$, and such that $|L(\mathcal{B}) \cap L(\mathcal{T})| \geq 3$.

Proof of Claim 8: First note that there must be some cherry (x, y) in $\mathcal{F}$ with $x, y \in L(\mathcal{B})$. Indeed, suppose first that $\mathcal{N} = \mathcal{B}$. Then, $L(\mathcal{B}) = X$, and so, by Claim 3, $\mathcal{F}$ must contain a tree $\mathcal{T}$ with $|L(\mathcal{T})| > 1$ and $L(\mathcal{T}) \subseteq L(\mathcal{B})$ and so such a cherry must exist. So suppose $\mathcal{N} \neq \mathcal{B}$. Note that if $\mathcal{F}$ contains a tree $\mathcal{T}$ with $L(\mathcal{B}) \subseteq L(\mathcal{T})$, then since $|L(\mathcal{T})| \geq |L(\mathcal{B})| \geq 2$, there must be a cherry (x, y) in $\mathcal{T}$ with $x, y \in L(\mathcal{B})$, otherwise $\mathcal{N}$ would not display $\mathcal{T}$. And, if $\mathcal{F}$ contains no tree $\mathcal{T}$ with $L(\mathcal{B}) \subseteq L(\mathcal{T})$, then $\mathcal{F}$ must contain a tree $\mathcal{T}'$ with $L(\mathcal{T}')$ a proper subset of $L(\mathcal{B})$, otherwise $\mathcal{N}$ would again not display $\mathcal{F}$. But then, $\mathcal{T}'$ contains a cherry (x, y) with $x, y \in L(\mathcal{B})$.

Now, let $\mathcal{T}$ be the tree in $\mathcal{F}$ that contains a cherry (x, y) with $x, y \in L(\mathcal{B})$. First note that, by Claim 7, $|L(\mathcal{T})| \geq 3$. We now show that we must also have

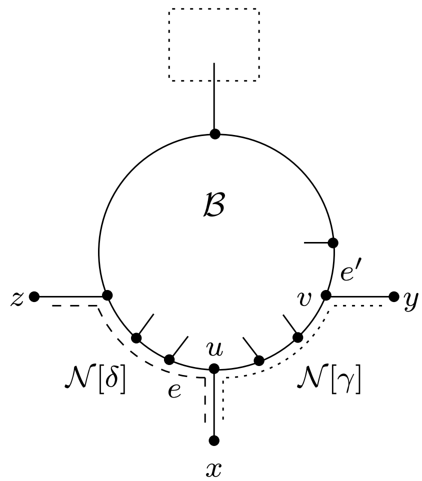

FIGURE 5. A schematic representation illustrating the images $\mathcal{N}[\gamma]$ and $\mathcal{N}[\delta]$ of the cherries (x, y) and (x, z) in the blob $\mathcal{B}$ as well as the associated edges e and e' used in the proof of Theorem 6.1. Note that we show that $z \in L(\mathcal{B})$ in Claim 11.

$|L(\mathcal{B})| \geq 3$. Suppose for contradiction that this is not the case. Then, by Claim 2, we must have $L(\mathcal{B}) = \{x, y\}$, and since $|L(\mathcal{T})| \geq 3$, we must also have $\mathcal{N} \neq \mathcal{B}$. Moreover, by Claim 3, neither x nor y is a component of $\mathcal{F}'$. Hence, (x, y) must be a cherry in $\mathcal{F}'$ since otherwise $\mathcal{N}$ would not display $\mathcal{F}'$. But this is a contradiction since it implies that (x, y) is a cherry in both $\mathcal{F}$ and $\mathcal{F}'$ so that Observation 3.1(i) holds, which is not possible by Claim 6. Thus, $|L(\mathcal{B})| \geq 3$.

Now, if $\mathcal{N} = \mathcal{B}$, then since $|L(\mathcal{T})| \geq 3$ and $|L(\mathcal{B})| \geq 3$, it immediately follows that $|L(\mathcal{B}) \cap L(\mathcal{T})| \geq 3$. So, suppose $\mathcal{N} \neq \mathcal{B}$. Then as $|L(\mathcal{B})| \geq 3$ it follows that there is some $s \in L(\mathcal{B}) \setminus \{x, y\}$. If $s \in L(\mathcal{T})$, then $|L(\mathcal{B}) \cap L(\mathcal{T})| \geq 3$ and Claim 8 follows. Otherwise $s \notin L(\mathcal{T})$ and so there must be some tree $\mathcal{T}'$ in $\mathcal{F}$ with $s \in L(\mathcal{T}')$. Now, since $|L(\mathcal{T})| \geq 3$, it follows that $\mathcal{N}[\mathcal{T}]$ must contain the non-pendant cut-edge incident to $\mathcal{B}$, since otherwise $|L(\mathcal{B}) \cap L(\mathcal{T})| \geq 3$. Thus, as $\mathcal{N}[\mathcal{T}']$ is disjoint from $\mathcal{N}[\mathcal{T}]$ by the definition of the image of a forest, we must have $L(\mathcal{T}') \subseteq L(\mathcal{B})$. But $\mathcal{T}'$ is not a single leaf by Claim 3, and it is not a cherry in the form of an isolated edge by Claim 7. Hence $|L(\mathcal{T}')| \geq 3$, and so $\mathcal{T}'$ contains a cherry (r, t) with $r, t \in L(\mathcal{B})$ and $|L(\mathcal{T}') \cap L(\mathcal{B})| \geq 3$. ■

Now, by Claim 8, we shall assume that we have a tree $\mathcal{T}$ in $\mathcal{F}$ that has a cherry (x, y) with $x, y \in L(\mathcal{B})$ and such that $|L(\mathcal{B}) \cap L(\mathcal{T})| \geq 3$. Moreover, by Claim 6, Observation 3.1(ii) holds, and so, as well as the cherry (x, y) in $\mathcal{T}$, there is a cherry in $\mathcal{F}'$, say (x, z), with $z \in X \setminus \{y\}$. Let $\mathcal{T}'$ denote the tree in $\mathcal{F}'$ that contains the cherry (x, z). For the cherry (x, y) in $\mathcal{T}$ and γ the path in $\mathcal{T}$ between x and y, let u, v and $\mathcal{N}[\gamma]$ be as in Claim 5. In addition, let δ be the path in $\mathcal{F}'$ between x and z and let $\mathcal{N}[\delta]$ be its image in $\mathcal{N}[\mathcal{F}']$, noting that $\mathcal{N}[\delta]$ must have length at least three by Claim 5. We picture this configuration in Figure 5.

Claim 9: Every edge in $\mathcal{N}[\gamma]$ or in $\mathcal{N}[\delta]$ that is in $E(\mathcal{B})$ must be contained in $\mathcal{N}[\mathcal{F}']$ or $\mathcal{N}[\mathcal{F}]$, respectively. In particular, since $x, y \in L(\mathcal{B})$, $\mathcal{N}[\gamma]$ is contained in $\mathcal{N}[\mathcal{F}']$.

Proof of Claim 9: We consider $\mathcal{N}[\gamma]$; the proof for $\mathcal{N}[\delta]$ is similar. Suppose g is an edge that is in $\mathcal{N}[\gamma]$ and $E(\mathcal{B})$, but g is not in $\mathcal{N}[\mathcal{F}']$. Then we can apply a (C2)(a)(i)-reduction to x or y, remove the edge g and apply Lemma 5.4(b) to the removed edge plus induction to see that (CP) must hold. Thus, $\mathcal{N}[\gamma]$ must be contained in $\mathcal{N}[\mathcal{F}']$, by Claim 3. ■

Claim 10: Let e and e' be the edges in $\mathcal{N}$ that are not in the path $\mathcal{N}[\gamma]$ and that contain u and v, respectively (see Figure 5). Then e is in $\mathcal{N}[\mathcal{T}]$, e' is not in $\mathcal{N}[\mathcal{T}]$, and both e and e' are in $\mathcal{N}[\mathcal{F}']$.

Proof of Claim 10: Note that neither e nor e' is a cut-edge. Moreover, we can assume that e must be contained in $\mathcal{N}[\mathcal{T}]$. Indeed, if not then e must be contained in $\mathcal{N}[\mathcal{F}']$, by Claim 4. Thus, since $\mathcal{N}[\gamma]$ is contained in $\mathcal{N}[\mathcal{F}']$ by Claim 9, e is an edge in $\mathcal{N}[\delta]$. Thus, we can apply a (C2)(a)(i) reduction to z, remove e, and apply Lemma 5.4(b) to e plus induction to see that (CP) holds. Hence, in summary, since (x, y) is a cherry in $\mathcal{T}$, we must have that e is an edge in $\mathcal{N}[\mathcal{T}]$ and that e' is not an edge in $\mathcal{N}[\mathcal{T}]$.

Now, since e' is not an edge in $\mathcal{N}[\mathcal{T}]$, by Claim 4 it follows that e' is in $\mathcal{N}[\mathcal{F}']$. Moreover e is also in $\mathcal{N}[\mathcal{F}']$, since otherwise as (x, z) is a cherry in $\mathcal{T}'$ it follows that $\mathcal{F}'$ contains the component $((x, y), z)$ and so we could apply a (C2)(a)(i)-reduction to z, remove e and apply Lemma 5.4(b) to e plus induction to see that (CP) must hold. ■

Claim 11: $z \in L(\mathcal{B})$, every edge in $\mathcal{N}[\delta]$ that does not contain x or z is in $E(\mathcal{B})$, and $\mathcal{N}[\delta]$ is contained in $\mathcal{N}[\mathcal{T}]$.

Proof of Claim 11: Suppose that z is not contained in $L(\mathcal{B})$. Then we must have $\mathcal{N} \neq \mathcal{B}$, and $\mathcal{N}[\delta]$ must contain a non-pendant cut-edge, such that one of the vertices in this cut-edge, w say, is contained in $V(\mathcal{B})$.

Consider the edge e' (see Figure 5). Then since e' is in $\mathcal{N}[\mathcal{F}']$ by Claim 10 and it is not a cut-edge, there must be a cherry in $\mathcal{T}'$, (r, s) say, with $r, s \in L(\mathcal{B}) \setminus \{x, y\}$ (since otherwise $\mathcal{T}'$ would only have four leaves and we could apply a (C2)(a)(i)-reduction and Lemma 5.4(b) to e'). But then by Claim 6, without loss of generality $\mathcal{T}$ must contain a cherry (r, a), with $a \in X \setminus \{x, y, s\}$. Moreover , since w is a vertex in $\mathcal{N}[\mathcal{T}]$ we must have $a \in L(\mathcal{B})$ as every edge in $E(\mathcal{B})$ and $\mathcal{N}[\delta]$ is contained in $\mathcal{N}[\mathcal{T}]$ by Claim 9. Hence, instead of considering the cherries (x, y) and (x, z) in $\mathcal{T}$ and $\mathcal{T}'$, we can instead consider the cherries (r, s) and (r, a) in $\mathcal{T}'$ and $\mathcal{T}$, respectively, and, reversing the roles of $\mathcal{T}$ and $\mathcal{T}'$, apply our arguments to these cherries instead. In particular, it follows that we may assume $z \in L(\mathcal{B})$.

The last statements in the claim now follow immediately by Claims 3 and 9. ■

Now, since $\mathcal{T}$ contains the cherry (x, y) and $|L(\mathcal{B}) \cap L(\mathcal{T})| \geq 3$, we will assume that either $((x, y), p)$ or $((x, y), (p, q))$ is a pendant subtree of $\mathcal{T}$ with $x, y, p, q \in L(\mathcal{B})$ distinct. Note that by Claim 5, the length of the path $\mathcal{N}[\delta]$ is at least three.

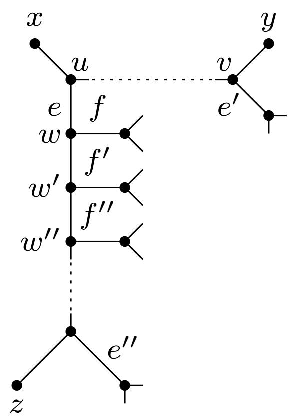

FIGURE 6. The configuration in $\mathcal{N}$ that underpins the proofs of Claims 12–14 of Theorem 6.1.

We next show that if the length of $\mathcal{N}[\delta]$ is three, then (CP) holds, which completes the proof of the theorem. Indeed, suppose that $\mathcal{N}[\delta]$ has length three. Then without loss of generality $z = p$. Moreover, if $\mathcal{T}$ contains $((x, y), p)$, then setting $\mathcal{S} = ((x, y), p)$, we can apply a (C2)(a)(ii)-reduction to $\mathcal{S}$, remove the edge e'', and apply Lemma 5.4(b) plus induction to see that (CP) must hold. And if $\mathcal{T}$ contains $((x, y), (p, q))$ then we must have $\mathcal{T} = ((x, y), (p, q))$, in which case we can apply a (C2)(a)(iii)-reduction to q, remove the edge e'', and apply Lemma 5.4(b) plus induction to see that (CP) must hold. Now, suppose that the length of the path $\mathcal{N}[\delta]$ is greater than three. Let w, w' and w'' be the vertices in $\mathcal{N}[\delta]$ that are distance two, three and four from x, respectively. Let f, f' and f'' be the edges in $\mathcal{N}$ that are not in $\mathcal{N}[\delta]$ and that contain the vertices w, w' and w'', respectively (see Figure 6).

Claim 12: f is in $\mathcal{N}[\mathcal{T}]$.

Proof of Claim 12: Suppose not. Then, since f is not in $\mathcal{N}[\mathcal{F}']$ because (x, z) is a cherry in $\mathcal{F}'$, by Claim 4 f must be a cut-edge of $\mathcal{N}$. Moreover, f is not a pendant edge otherwise $\mathcal{F}'$ would contain a component that is a vertex which contradicts Claim 3. Thus, f' is not a cut-edge of $\mathcal{N}$ as $\mathcal{B}$ is a pendant blob, and so f' is contained in $\mathcal{N}[\mathcal{T}]$ by Claim 4.

Now, if $z = w''$, then $\mathcal{T}$ must contain $((x, y), z)$ as a proper pendant subtree and we can apply a (C2)(a)(ii)-reduction to $((x, y), z)$, remove the edge f', and apply Lemma 5.4(b) plus induction to see that (CP) must hold. And, if $z \neq w''$, then as f'' is not a cut-edge of $\mathcal{N}$ (since f is) it follows that f'' is in $\mathcal{N}[\mathcal{T}]$ by Claim 4. But then if $\mathcal{T}$ contains $((x, y), p))$ as a pendant subtree then it must be proper because two edges on $\mathcal{N}[\delta]$ must be incident with w''. Hence, we can apply a (C2)(a)(ii)-reduction to p, remove the edge f', and apply Lemma 5.4(b) plus induction to see that (CP) must hold. And, if $\mathcal{T}$ contains $((x, y), (p, q))$ as a pendant subtree, then either, without loss of generality, $z = p$, or f' is in the image of the path from x to p (or q). Furthermore, $((x, y), (p, q))$ must be proper because two of the edges that share a vertex with f' or f'' are also contained in $\mathcal{N}[\delta]$. Hence, we can either apply a (C2)(a)(ii)-reduction to q and remove f'' or a (C2)(a)(iii)-reduction to (p, q) and remove f' to see that (CP) holds using Lemma 5.4(b) and induction. ■

Claim 13: If $\mathcal{T}$ contains $((x, y), z)$ as a pendant subtree, then f is in the image of the pendant edge in $\mathcal{T}$ that contains p and $((x, y), z)$ is a proper pendant subtree of $\mathcal{T}$.

Proof of Claim 13: Suppose that the claim is not true. Then since e is an edge in $\mathcal{N}[\mathcal{T}]$ and f is an edge in $\mathcal{N}[\mathcal{T}]$ by Claim 12, $\{w, w'\}$ is in the image of the pendant edge in $\mathcal{T}$ that contains p, and $((x, y), p)$ is a proper pendant subtree of $\mathcal{T}$. Now since every edge in $\mathcal{N}[\delta]$ is contained in $\mathcal{N}[\mathcal{T}]$ by Claim 11, the edge e'' must be in the image of the pendant edge in $\mathcal{T}$ that contains p. Moreover, e'' is not a cut-edge of $\mathcal{N}$ since $p \in L(\mathcal{B})$. so we can apply a (C2)(a)(iii)-reduction to p, remove e'' and apply Lemma 5.4(b) plus induction to see that (CP) holds. ■

Claim 14: If $\mathcal{T}$ contains $((x, y), (p, q))$ as a pendant subtree, then f is in the image of the pendant edge in $\mathcal{T}$ that shares a vertex with the cherry (p, q) and $((x, y), (p, q))$ is a proper pendant subtree of $\mathcal{T}$.

Proof of Claim 14: Suppose that the claim is not true. Then using similar arguments to those used in the proof of Claim 13, it follows that without loss of generality, $z = p$, and either $w'' = z$ and f' is in the image of the pendant edge in $\mathcal{F}'$ that contains q, or f' is a cut-edge that is not a pendant edge, and f'' is in the image of the pendant edge in $\mathcal{F}'$ that contains q. In either case, $((x, y), (p, q))$ is a proper pendant subtree of $\mathcal{T}$. Furthermore, in either case by removing f' or f'' respectively and applying a (C2)(a)(iii)-reduction to q we can use Lemma 5.4(b) and induction to see that (CP) holds. ■

Finally, to complete the proof of the theorem, if $\mathcal{T}$ contains $((x, y), z)$ as a pendant subtree, then, by Claim 13, we can apply a (C2)(a)(ii)-reduction to p, remove f and apply Lemma 5.4(b) plus induction to see that (CP) must hold. And, if $\mathcal{T}$ contains $((x, y), (p, q))$ as a pendant subtree, then by Claim 14, we can apply a (C2)(a)(iii)-reduction to (p, q), remove f and apply Lemma 5.4(b) plus induction to see that (CP) must hold. □

7. Discussion

In this paper, we have proven that the hybrid number of two binary phylogenetic trees can be given in terms of cherry picking sequences. There are several interesting future directions of research. For example, Humphries et al. [11] have considered a certain type of cherry picking sequence to characterize arbitrary size collections of rooted phylogenetic trees that can be displayed by a rooted phylogenetic network that is time consistent and tree-child. It would be interesting to investigate extensions of our results to computing the hybrid number of an arbitrary collection of not necessarily binary phylogenetic trees (or forests) that can be displayed by an unrooted phylogenetic network of a given class such as tree-child or orchard (see [8] for the definition of such networks).

As mentioned in the introduction, the hybrid number of two unrooted binary phylogenetic trees is equal to the TBR distance between the two trees. Recently, there has been some interest in improving algorithms to compute the TBR distance (see e.g. [19, 20]), and some of the data reductions rules introduced in these papers are similar to the rules that we used to define cherry picking sequences (for example, the reduction described in (C2)(a)(ii) is similar to the (*,3,*)-reduction in [19, Section 3.1]). It would be interesting to see if there is a deeper connection between these two concepts, and to investigate whether, for example, the results presented here could be used to improve depth-bounded search tree algorithms to compute the TBR distance. To investigate this connection, a first step might be to understand whether or not all of the reductions in the definition of a cherry picking sequence are needed, or whether a somewhat simpler list of reductions might be found to define a cherry picking sequence so that Theorem 4.1 still holds.

Finally, we have considered the hybrid number $h(\mathcal{F}, \mathcal{F}')$ for $\mathcal{F}$ and $\mathcal{F}'$ being an arbitrary pair of binary forests on X. In case we restrict h to pairs of phylogenetic trees on X, h is the TBR distance on the set of phylogenetic trees on X. It would be interesting to understand properties of h for pairs of forests in general (e. g. is h related to some kind of TBR distance on the set of forests on X?).

Acknowledgment The authors would like to thank the anonymous referees for numerous comments and suggestions to improve the exposition of the paper.

References

[1] Allen, B., Steel, M. (2001) Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics, 5:1–15.
[2] Baroni, B., Grünewald, S., Moulton, V., Semple, S. (2005) Bounding the number of hybridization events for a consistent evolutionary history. Journal of Mathematical Biology. 51:171–182.
[3] Berge C. (2001) The Theory of Graphs. Dover Publications Inc.
[4] Bernardini, G., van Iersel, L., Julien, E., Stougie L. (2023) Reconstructing phylogenetic networks via cherry picking and machine learning. Algorithms for Molecular Biology. 18:13.
[5] Bryant, D., Lagergren, J. (2006) Compatibility of unrooted phylogenetic trees is FPT. Theoretical Computer Science. 351:296–302.
[6] Chen, J., Fan, J.-H., Sze, S.-H. (2015) Parameterized and approximation algorithms for maximum agreement forest in multifurcating trees. Theoretical Computer Science. 562:496–512.
[7] Döcker, J., van Iersel, L., Kelk, S., Linz, S. (2019) Deciding the existence of a cherry-picking sequence is hard on two trees. Theoretical Computer Science. 260:131–143.
[8] Döcker, J., Linz, S. (2024). Characterising rooted and unrooted tree-child networks. arXiv preprint, arXiv:2408.17105.
[9] Huber, K. T., Moulton, V., Steel, M., and Wu, T. (2016) Folding and unfolding phylogenetic trees and networks. Journal of Mathematical Biology. 73:1761–1780.
[10] Humphries, P. J, Linz, S., Semple, C. (2013) On the complexity of computing the temporal hybridization number for two phylogenies. Discrete Applied Mathematics 161:871–880.
[11] Humphries, P. J., Linz, S., Semple, C. (2013) Cherry picking: a characterization of the temporal hybridization number for a set of phylogenies. Bulletin of Mathematical Biology. 75:1879–1890.
[12] Huson, D. H., Rupp, R. and Scornavacca, C. (2010 Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press.

[13] van Iersel, L., Janssen, R., Jones, M., Murakami, Y., Zeh, N. (2022) A practical fixed-parameter algorithm for constructing tree-child networks from multiple binary trees. Algorithmica. 84:917–960.
[14] van Iersel, L., Kelk, S., Stamoulis, G., Stougie, L., Boes, O. (2018) On unrooted and root-uncertain variants of several well-known phylogenetic network problems. Algorithmica. 80:2993–3022.
[15] van Iersel, L., Semple, C., Steel, M. (2010) Locating a tree in a phylogenetic network. Information Processing Letters. 110:1037–1043.
[16] Janssen, R., Murakami, Y. (2021) On cherry-picking and network containment. Theoretical Computer Science. 856:121–150.
[17] Kanj, I. A., Nakhleh, L., Than, C., Xia, G. (2008) Seeing the trees and their branches in the network is hard. Theoretical Computer Science. 401:153–164.
[18] Kelk, S., Linz, S. (2019) A tight kernel for computing the tree bisection and reconnection distance between two phylogenetic trees. SIAM Journal on Discrete Mathematics. 33:1556–1574.
[19] Kelk, S., Linz, S. (2020) New reduction rules for the tree bisection and reconnection distance. Annals of Combinatorics. 24:475–502.
[20] Kelk, S., Linz S., Meuwese R. (2024) Deep kernelization for the tree bisection and reconnection (TBR) distance in phylogenetics. Journal of Computer and System Sciences. 142, Article no. 103519.
[21] Kong, S., Pons, J.C., Kubatko, L., Wicke, K. Classes of explicit phylogenetic networks and their biological and mathematical significance. Journal of Mathematical Biology. 2022 May;84(6):1–44.
[22] Landry, K., Teodocio, A., Lafond, M. Tremblay-Savard, O. (2023) Defining phylogenetic network distances using cherry operations. Transactions on Computational Biology and Bioinformatics. 20:1654–1666.
[23] Landry, K., Tremblay-Savard, O., Lafond, M. A fixed-parameter tractable algorithm for finding agreement cherry-reduced subnetworks in level-1 orchard networks. Journal of Computational Biology, in press.
[24] Linz, S., Semple, C. (2019) Attaching leaves and picking cherries to characterise the hybridisation number for a set of phylogenies. Advances in Applied Mathematics. 105:102–129.
[25] St John, K. (2017) The shape of phylogenetic treespace. Systematic Biology, 66:e83.
[26] Semple, C., Steel, M. (2003) Phylogenetics. Oxford University Press.
[27] Whidden, C., Beiko, R. G., Zeh, N. (2013) Fixed-parameter algorithms for maximum agreement forests. SIAM Journal on Computing. 42:1431–1466.
[28] Zhang, L., Ng, Y .K., Wu, T., Zheng, Y. (2011) Network model and efficient method for detecting relative duplications or horizontal gene transfers. In: IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). pp. 214–219.

8. Statements and declarations

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
The authors have no relevant financial or non-financial interests to disclose.
All authors contributed to the study conception and design. All authors read and approved the final manuscript.
Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

School of Computing Sciences, University of East Anglia, Norwich, UK

Email address: k.huber@uea.ac.uk

School of Computer Science, University of Auckland, Auckland, New Zealand

Email address: s.linz@auckland.ac.nz

School of Computing Sciences, University of East Anglia, Norwich, UK

Email address: v.moulton@uea.ac.uk