A CLASS OF PHYLOGENETIC NETWORKS RECONSTRUCTABLE FROM ANCESTRAL PROFILES

PÉTER L. ERDŐS, CHARLES SEMPLE, AND MIKE STEEL

Abstract. Rooted phylogenetic networks provide an explicit representation of the evolutionary history of a set X of sampled species. In contrast to phylogenetic trees which show only speciation events, networks can also accommodate reticulate processes (for example, hybrid evolution, endosymbiosis, and lateral gene transfer). A major goal in systematic biology is to infer evolutionary relationships, and while phylogenetic trees can be uniquely determined from various simple combinatorial data on X, for networks the reconstruction question is much more subtle. Here we ask when can a network be uniquely reconstructed from its 'ancestral profile' (the number of paths from each ancestral vertex to each element in X). We show that reconstruction holds (even within the class of all networks) for a class of networks we call 'orchard networks', and we provide a polynomial-time algorithm for reconstructing any orchard network from its ancestral profile. Our approach relies on establishing a structural theorem for orchard networks, which also provides for a fast (polynomial-time) algorithm to test if any given network is of orchard type. Since the class of orchard networks includes tree-sibling tree-consistent networks and tree-child networks, our result generalise reconstruction results from 2008 and 2009. Orchard networks allow for an unbounded number k of reticulation vertices, in contrast to tree-sibling tree-consistent networks and tree-child networks for which k is at most $2|X| - 4$ and $|X| - 1$, respectively.

1. Introduction

Phylogenetic trees and networks have become a ubiquitous tool for representing evolutionary relationships in systematics biology [7] and other areas of classification (for example, language evolution and epidemiology). From early sketches by Charles Darwin and Ernst Haeckel in the 19$^{\text{th}}$ century,

Date: May 1, 2019.

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

Key words and phrases. Tree-child networks, orchard networks, accumulation phylogenies, ancestral profiles, path-tuples.

The first author was supported in part by the National Research, Development and Innovation Office (NKFIH grants K 116769 and KH 126853). The second and third authors were supported by the New Zealand Marsden Fund (UOC1709).

more complex and detailed trees are now revealing the finer details of portions of the 'tree of life'. Today, biologists routinely build phylogenetic trees on hundreds of species, such as the recent tree of (nearly) all ∼10,000 species of birds [14]. Phylogenetic trees have a leaf set X that consists of the sampled organisms (typically, a group of present-day species); the root of the tree represents the most recent common ancestor of the species in X. Current methods for inferring phylogenetic trees trees generally use genomic data from the species in X, and apply one of several possible reconstruction methods. While many of these methods are statistically based, they are ultimately founded on underlying combinatorial uniqueness results concerning trees [7, 17].

Although phylogenetic trees have proved a convenient representation for many groups of species including, for example, mammals and birds, in other domains of life evolution is not always described as a simple vertical process of speciation (where lineages split in two as new species form) and extinction. Instead, various reticulate processes allow for a 'horizontal' component. Two main examples include the formation of hybrid species (such as in certain plant or fish species), and the exchange of genes between species in a process called lateral gene transfer (such as in bacteria). An additional reticulate process relevant to early life on earth is endosymbiosis in which organelles are incorporated into cells.

For these reasons, phylogenetic networks (acyclic directed graphs with a single root vertex and leaves forming the set X) have been proposed as a more flexible and accurate representation of evolutionary history [6, 15]. Accordingly, there has been considerable recent interest in extending the mathematical foundation of phylogenetic tree reconstruction to networks [11]. This extension faces a number of mathematical obstacles. In particular, while trees can be encoded and reconstructed in several ways (for example, based on their associated system of clusters, path distances between pairs of leaves, and induced 3-leaf subtrees), none of these approaches extends to networks, except for in very special cases [9, 12, 19]. This has led to various approaches being proposed, which usually involve one or more of the following:

(i) not distinguishing between phylogenetic networks that are similar in a certain way [16];
(ii) considering reconstruction only within a limited subclass of phylogenetic networks [2]; and
(iii) allowing types of information for X beyond what is normally used for tree reconstruction [1].

Approach (ii) has received the most attention so far, with some positive results (for example, for reconstructing the subclass of normal networks

from their induced trees [20]). In this paper, we focus more on approach (iii), and, although we restrict to a class of subnetworks (which we call 'orchard networks'), our reconstruction result has the additional strength that it can distinguish between any two networks from information on X provided at least one of them is an orchard network. To provide some intuition, informally, a phylogenetic network is an orchard network if it can be reduced to a single vertex by recursively finding a pair of leaves that form either a cherry or a reticulated cherry, and then applying a cherry reduction to that pair of leaves.

The type of information on X we consider is the following. View the interior (non-leaf) vertices of a phylogenetic network $\mathcal{N}$ as being labelled. In the biological setting, this label could correspond, for example, to the genome of the ancestral species at this vertex (or some sub-genome that is sufficiently detailed to distinguish this ancestral vertex from others). For each species x in the leaf set X, suppose we can count the number of directed paths in the network from each ancestral genome (i.e. interior vertex) to x. This 'ancestral profile' is thus an ordered tuple of numbers, one tuple for each leaf in X (note that current technology does not yet provide this information, so our approach is in the spirit of earlier mathematical results in phylogenetics that preceded the data required for their application). It turns out that such information is not enough to distinguish between an arbitrary pair of networks (we provide an example). However, if the underlying network $\mathcal{N}$ is an orchard network, our main result shows that no other network (orchard or not) can have the same ancestral profile. Moreover, we present and justify a polynomial-time algorithm for reconstructing any orchard network from its ancestral profile. Our arguments rely on a structural property of orchard networks which also implies that there is a polynomial-time algorithm for testing whether or not an arbitrary network is an orchard network.

Our results generalise earlier work in [4, 5] which considered the more restricted classes of 'tree-sibling time-consistent' networks and 'tree-child' networks, respectively. These authors use equivalent information on X for reconstruction, however, their reconstruction result faces two limitations that are lifted here. First, the uniqueness results of [4, 5] hold only within the class of tree-sibling time-consistent networks and tree-child networks, whereas we show that ancestral profiles can distinguish an orchard network from any other network. Second, neither tree-sibling time-consistent networks nor tree-child networks can have too many reticulate vertices (at most and $2n-4$ and $n-1$, respectively, where $n = |X|$), whereas orchard networks can have arbitrarily many reticulate vertices (independent of n).

Our results are also related to (and partly motivated by) earlier work by [1] and [18] on 'accumulation phylogenies'. This involved a different subclass of networks (called 'regular' in these papers, and 'cluster networks' in [11]),

which neither contains, nor is contained in the subclass of orchard networks. A limitation of this subclass is that (unlike orchard networks) they do not allow 'redundant arcs' (an arc (u, v) for which there is another path in the network from u to v). Allowing redundant arcs has a strong biological motivation since even if each reticulation events happens instantaneously between two contemporaneous species, redundant arcs can still appear in the resulting network if not all species at the present are sampled. The results in [1, 18] also assume any two networks being considered are within this same subclass. In summary, our results are not directly related to this earlier work on accumulation phylogenies, apart from using a related type of information.

The paper is organised as follows. The next section contains some necessary definitions along with the statement of the main result (Theorem 2.2) and deduces, as a consequence, the main result (Theorem 1) in [5]. This section also provides examples to justify various claims. Section 3 describes some preliminary lemmas, which apply more generally than for ancestral profiles, and in Section 4 we state and prove the structural property of orchard networks that allows for an easy test as to whether or not an arbitrary network is of orchard type. The proof of Theorem 2.2 is established in Section 5. We end the paper with a brief discussion in Section 6.

Lastly, just as we completed the write-up of this paper, a manuscript [13] was posted on arXiv that also considers the class of orchard networks (referred to as "cherry-picking networks" in [13]). The focus of that manuscript is quite different to that of this paper; nevertheless, it contains an independent and different proof of the structural property of orchard networks which is needed as a lemma for Theorem 2.2 in this paper.

2. Main Result

Throughout the paper X denotes a non-empty finite set and, unless otherwise stated, all paths are directed. For vertices u and v of a directed graph D, we say v is *reachable* from u if there is a path in D from u to v. Furthermore, for sets A and B, we denote the set obtained from A by removing every element in A that is also in B by $A - B$. If $|B| = 1$, say $B = \{b\}$, we denote this by $A - b$.

Phylogenetic networks. A *phylogenetic network on* X is a rooted acyclic directed graph with no arcs in parallel and satisfying the following properties:

(i) the (unique) root has in-degree zero and out-degree two;

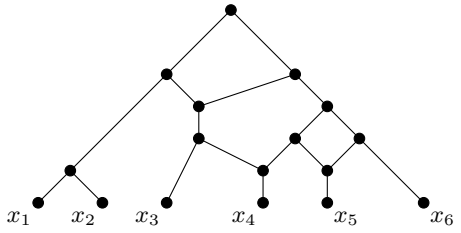


FIGURE 1. A phylogenetic network $\mathcal{N}$ on $\{x_1, x_2, \ldots, x_6\}$. Here, $\{x_1, x_2\}$ is a cherry and $\{x_3, x_4\}$ is a reticulated cherry with x_4 the reticulation leaf.

(ii) a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is X; and
(iii) all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one.

For technical reasons, if $|X| = 1$, we additionally allow a single vertex to be a phylogenetic network, in which case, the root is the vertex in X. Phylogenetic networks as defined here are also referred to as 'binary phylogenetic networks' in the literature.

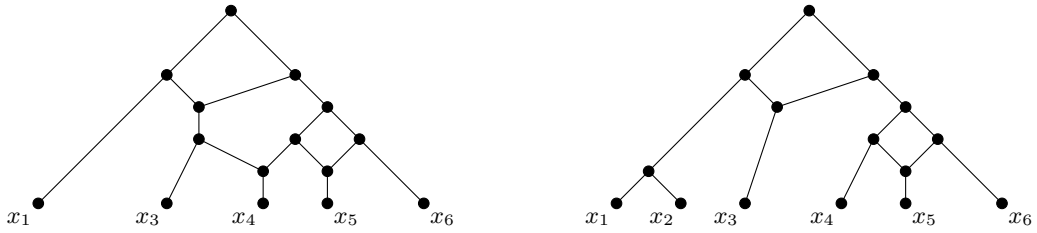
Let $\mathcal{N}$ be a phylogenetic network on X. The vertices with out-degree zero are the *leaves* of $\mathcal{N}$, and so X is called the *leaf set* of $\mathcal{N}$. Furthermore, vertices with in-degree one and out-degree two are *tree vertices*, while vertices of in-degree two and out-degree one are *reticulations*. The arcs directed into a reticulation are called *reticulation arcs*, all other arcs are *tree arcs*. To illustrate, an example of a phylogenetic network with leaf set $\{x_1, x_2, \ldots, x_6\}$ and three reticulations is shown in Fig. 1.

Lastly, let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two phylogenetic networks on X with vertex and arc sets V_1 and E_1, and V_2 and E_2, respectively. We say $\mathcal{N}_1$ is *isomorphic* to $\mathcal{N}_2$ if there exists a bijection $\varphi : V_1 \to V_2$ such that $\varphi(x) = x$ for all $x \in X$, and $(u, v) \in E_1$ if and only if $(\varphi(u), \varphi(v)) \in E_2$ for all $u, v \in V_1$.

Ancestral tuples and ancestral profile. Let $\mathcal{N}$ be a phylogenetic network on X with vertex set V. Let $v_1, v_2, \ldots, v_t$ be a fixed (arbitrary) labelling of the vertices in $V - X$. For all $x \in X$, the *ancestral tuple* of x, denoted $\sigma(x)$, is the t-tuple whose i-th entry is the number of paths in $\mathcal{N}$ from v_i to x. Denoted by $\Sigma_{\mathcal{N}}$, we call the set

$$\Sigma_{\mathcal{N}} = \{(x, \sigma(x)) : x \in X\},$$

of ordered pairs the *ancestral profile* of $\mathcal{N}$. Furthermore, if $\mathcal{N}'$ is a phylogenetic network on X and, up to an ordering of the non-leaf vertices of $\mathcal{N}'$,



(i) $\mathcal{N}_1$ (ii) $\mathcal{N}_2$

FIGURE 2. $\mathcal{N}_1$ has been obtained from $\mathcal{N}$ in Fig. 1 by reducing x_2, while $\mathcal{N}_2$ has been obtained from $\mathcal{N}$ by cutting $\{x_3, x_4\}$.

we have $\Sigma_{\mathcal{N}'} = \Sigma_{\mathcal{N}}$, we say $\mathcal{N}'$ *realises* $\Sigma_{\mathcal{N}}$. Lastly, although $\Sigma_{\mathcal{N}}$ depends on the ordering of the vertices in $V - X$, the ordering is fixed and so the labelling can be effectively ignored.

Cherries and reticulated cherries. Let $\mathcal{N}$ be a phylogenetic network on X, and let $\{a, b\}$ be a 2-element subset of X. Let p_a and p_b denote the parents of a and b, respectively. We say $\{a, b\}$ is a *cherry* of $\mathcal{N}$ if $p_a = p_b$. Furthermore, if one of the parents, say p_b, is a reticulation and (p_a, p_b) is an arc in $\mathcal{N}$, then $\{a, b\}$ is a *reticulated cherry* of $\mathcal{N}$, in which case, b is the *reticulation leaf* of the reticulated cherry. Observe that p_a is necessarily a tree vertex. For the phylogenetic network shown in Fig. 1, $\{x_1, x_2\}$ is a cherry, while $\{x_3, x_4\}$ is a reticulated cherry in which x_4 is the reticulation leaf. Furthermore, in Fig. 1, $\{x_4, x_5\}$ is neither a cherry nor a reticulated cherry.

We next describe two operations associated with cherries and reticulated cherries that are central to this paper. Let $\mathcal{N}$ be a phylogenetic network. First suppose that $\{a, b\}$ is a cherry of $\mathcal{N}$. Then *reducing* b is the operation of deleting b and suppressing the resulting vertex of in-degree one and out-degree one. If the parent of a and of b is the root of $\mathcal{N}$, then reducing b is the operation of deleting b as well as deleting the root of $\mathcal{N}$, thus leaving only the isolated vertex a. Now suppose that $\{a, b\}$ is a reticulated cherry of $\mathcal{N}$ in which b is the reticulation leaf. Then *cutting* $\{a, b\}$ is the operation of deleting the reticulation arc joining the parents of a and b, and suppressing the two resulting vertices of in-degree one and out-degree one. It is easily seen that the operations of reducing a cherry and cutting a reticulated cherry both result in a phylogenetic network. Collectively, we refer to these two operations as *cherry reductions*. To illustrate, the phylogenetic network shown in Fig. 2(i) (resp. Fig. 2(ii)) has been obtained from the phylogenetic network in Fig. 1 by reducing x_2 (resp. cutting $\{x_3, x_4\}$).

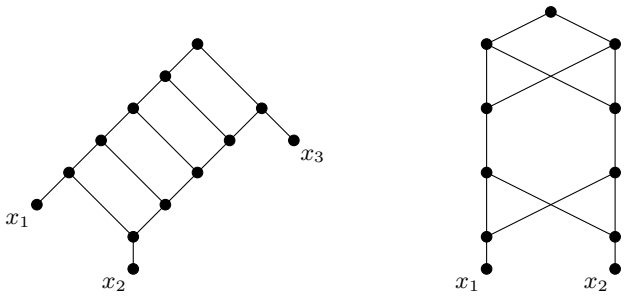


FIGURE 3. (i) An orchard network and (ii) a non-orchard network.

Orchard networks. For a phylogenetic network $\mathcal{N}$, the sequence

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k \tag{1}$$

of phylogenetic networks is a *cherry-reduction sequence of* $\mathcal{N}$ if, for all $i \in \{1, 2, \ldots, k\}$, the phylogenetic network $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by a (single) cherry reduction. The sequence is *maximal* if $\mathcal{N}_k$ has no cherries or reticulated cherries. If $\mathcal{N}_k$ consists of a single vertex, the sequence is *complete*, in which case, $\mathcal{N}$ is called an *orchard network*. Observe that if (1) is complete, then the leaf set of $\mathcal{N}_{k-1}$ has size two and the parent of each leaf is the root of $\mathcal{N}_{k-1}$. It is easily checked that the phylogenetic network shown in Fig. 1 is an orchard network. In Section 4, we show that if $\mathcal{N}$ is an orchard network, then every maximal sequence of cherry reductions of an orchard network $\mathcal{N}$ is complete. Thus if we want to construct a complete cherry-reduction sequence for an orchard network, the order in which the reductions are applied does not matter. In turn, this provides an easy test to decide whether or not an arbitrary network is orchard.

One of the most well-studied classes of phylogenetic networks is the class of tree-child networks. Introduced in [5], a phylogenetic network is *tree-child* if every non-leaf vertex is the parent of a tree vertex or a leaf. Tree-child networks are examples of orchard networks [3], but there exist orchard networks that are not tree-child. Indeed, while the size of the leaf set bounds the total number of vertices of a tree-child network [5], the total number of vertices in an orchard network is not necessarily bounded by the size of its leaf set. For example, the phylogenetic network shown in Fig. 3(i) is an orchard network with exactly three leaves but, by extending it in the obvious way, we can produce an orchard network with an arbitrarily large odd number of vertices and still with exactly three leaves. Furthermore, not all phylogenetic networks are orchard networks as Fig. 3(ii) illustrates.

For this paper, a second relevant class of phylogenetic networks is the class of tree-sibling time-consistent networks. Let $\mathcal{N}$ be a phylogenetic network. We say $\mathcal{N}$ is *tree-sibling* if every reticulation has a parent that is also the parent of a tree vertex or a leaf. Furthermore, $\mathcal{N}$ is *time-consistent* if there is a map t from the vertex set of $\mathcal{N}$ to the non-negative integers such that if (u, v) is a reticulation arc of $\mathcal{N}$, then $t(u) = t(v)$; otherwise, $t(u) < t(v)$. We refer to such a mapping as a *temporal labelling*. In the literature, time-consistent networks are also referred to as *temporal* networks. Like tree-child networks, the class of tree-sibling time-consistent networks is a proper subclass of orchard networks. For completeness, we include a proof of containment. To see that it is proper, it is shown in [4] that, unlike orchard networks, the number of reticulations of a tree-sibling time-consistent network is bounded by the size of its leaf set.

Lemma 2.1. *Let $\mathcal{N}$ be a tree-sibling time-consistent network. Then $\mathcal{N}$ is an orchard network.*

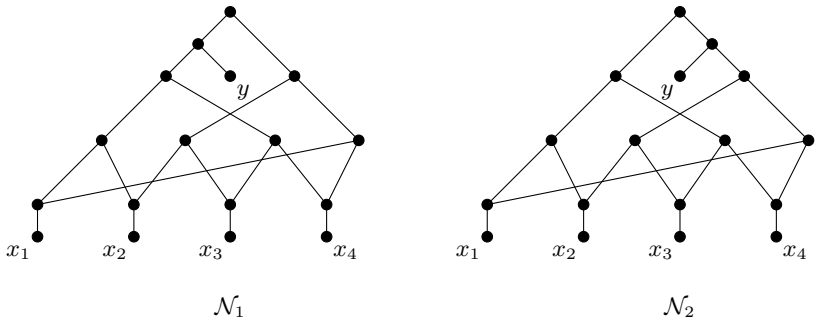
Proof. Clearly, the lemma holds if $\mathcal{N}$ has no reticulations. Therefore we may assume that $\mathcal{N}$ has at least one reticulation. We first show that $\mathcal{N}$ has either a cherry or a reticulated cherry. Let t be a temporal labelling of the vertices of $\mathcal{N}$, and let v be a reticulation with the property that $t(v) \geq t(v')$ for all reticulations v' of $\mathcal{N}$. Since $\mathcal{N}$ is tree-sibling, v has a parent, u say, that is the parent of a vertex w which is either a tree vertex or a leaf. By maximality, no reticulations are reachable from v or w. Therefore, if two leaves are reachable from either v or w, then $\mathcal{N}$ has a cherry. If this does not occur, then w is a leaf and that the (unique) child, x say, of v is also a leaf. In particular, $\{w, x\}$ is a reticulated cherry of $\mathcal{N}$.

To complete the proof, let $\mathcal{N}'$ be obtained from $\mathcal{N}$ by a cherry reduction. Clearly, $\mathcal{N}'$ is also tree-sibling. Furthermore, it is easily checked that the mapping t' from the vertex set of $\mathcal{N}'$ to the non-negative integers given by $t'(u) = t(u)$ is a temporal labelling of $\mathcal{N}'$. Thus $\mathcal{N}'$ is tree-sibling time-consistent. The lemma now follows. □

Main result. The following theorem is the main result of the paper.

Theorem 2.2. *Let $\mathcal{N}$ be an orchard network on X with vertex set V. Then, up to isomorphism, $\mathcal{N}$ is the unique phylogenetic network on X realising $\Sigma_{\mathcal{N}}$. Furthermore, up to isomorphism, $\mathcal{N}$ can be reconstructed from $\Sigma_{\mathcal{N}}$ in time $O(|X|^3|V|^3)$.*

It is worth emphasising that the uniqueness of $\mathcal{N}$ in the statement of Theorem 2.2 is amongst all phylogenetic networks on X, not just within



$\mathcal{N}_1$ $\mathcal{N}_2$

FIGURE 4. Two non-isomorphic phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$, but $\Sigma_{\mathcal{N}_1} = \Sigma_{\mathcal{N}_2}$.

the class of orchard networks on X. Furthermore, if $\mathcal{N}$ is not an orchard network, then the outcome of Theorem 2.2 does not necessarily hold. In particular, consider the two phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ in Fig. 4. It is easily checked that by fixing an ordering of the non-leaf vertices of each of $\mathcal{N}_1$ and $\mathcal{N}_2$ so that the parent of y is in the same position in both orderings, we have $\Sigma_{\mathcal{N}_1} = \Sigma_{\mathcal{N}_2}$. But $\mathcal{N}_1$ is not isomorphic to $\mathcal{N}_2$.

Theorem 2.2 generalises results of Cardona et al. [4] and Cardona et al. [5]. Let $\mathcal{N}$ be a phylogenetic network on X with vertex set V and let $x_1, x_2, \ldots, x_n$ be a fixed ordering of the leaves in X. For all $v \in V - X$, the *path tuple* of v, denoted $\pi(v)$, is the n-tuple whose i-th entry is the number of paths in $\mathcal{N}$ from v to x_i. Let $\Pi_{\mathcal{N}}$ denote the multiset

$$\{\pi(v) : v \in V - X\}$$

of path tuples of $\mathcal{N}$. If $\mathcal{N}'$ is a phylogenetic network on X and, up to an ordering of X, we have $\Pi_{\mathcal{N}'} = \Pi_{\mathcal{N}}$, we say $\mathcal{N}'$ *realises* $\Pi_{\mathcal{N}}$. The next theorem was established in [4] and [5].

Theorem 2.3. *Let $\mathcal{N}$ be a phylogenetic network on X.*

(i) *If $\mathcal{N}$ is tree-sibling time-consistent, then, up to isomorphism, $\mathcal{N}$ is the unique tree-sibling time-consistent network on X realising $\Pi_{\mathcal{N}}$.*
(ii) *If $\mathcal{N}$ is tree-child, then, up to isomorphism, $\mathcal{N}$ is the unique tree-child network on X realising $\Pi_{\mathcal{N}}$.*

Furthermore, for both instances, up to isomorphism, $\mathcal{N}$ can be constructed from $\Pi_{\mathcal{N}}$ in time polynomial in the size of X.

Let $\mathcal{N}$ be a phylogenetic network on X with vertex set V. The set $\Sigma_{\mathcal{N}}$ and multiset $\Pi_{\mathcal{N}}$ are equivalent in the amount of information they provide. To see this, let $x_1, x_2, \ldots, x_n$ and $v_1, v_2, \ldots, v_t$ be fixed orderings of the vertices in X and $V - X$, respectively. Then, for all $i \in \{1, 2, \ldots, t\}$, the

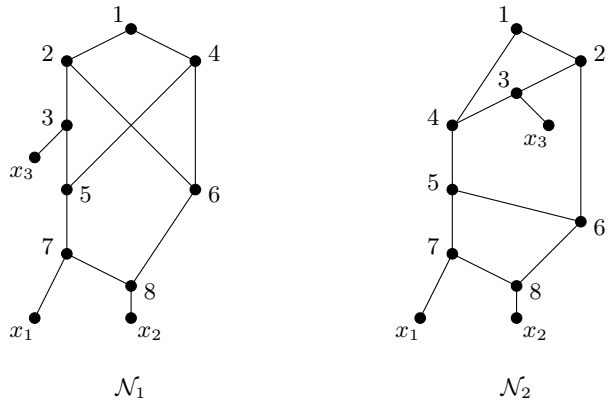


FIGURE 5. Two orchard networks $\mathcal{N}_1$ and $\mathcal{N}_2$ with $\Gamma_{\mathcal{N}_1} = \Gamma_{\mathcal{N}_2}$, but $\Sigma_{\mathcal{N}_1} \neq \Sigma_{\mathcal{N}_2}$.

n-tuple $\pi(v_i)$ is the tuple whose j-th entry is the i-th entry of $\sigma(x_j)$ for all $j \in \{1, 2, \ldots, n\}$. Similarly, each ordered pair in $\Sigma_{\mathcal{N}}$ can be obtained from $\Pi_{\mathcal{N}}$. Thus Theorem 2.2 generalises Theorem 2.3 in two ways. First, it shows that the latter holds for the more general class of orchard networks and, second, the uniqueness is not confined to the class of networks being constructed.

We end the section with three remarks. Firstly, Theorem 2.2 is not the first reconstruction result concerning the class of orchard networks. Although this class was not named, it is shown in [3] that orchard networks are reconstructible from their so-called multiset distance matrices. See [3, Theorem 3.4]. We have no doubt that, over time, the class of orchard networks will be realised to be reconstructible in other ways as well.

The second remark concerns a related, but weaker, notion to that of ancestral tuples called ancestral sets. Let $\mathcal{N}$ be a phylogenetic network on X with vertex set V. For all $x \in X$, the *ancestral set* of x is

$$\gamma(x) = \{v \in V - X : \ x \text{ is reachable from } v\}.$$

Thus $\gamma(x)$ is the set of non-leaf vertices v in $\mathcal{N}$ for which there is a directed path from v to x. Observe that, for all $x \in X$, the root of $\mathcal{N}$ is always an element of $\gamma(x)$ and so $\gamma(x)$ is non-empty. Let $\Gamma_{\mathcal{N}}$ denote the set

$$\{(x, \gamma(x)) : x \in X\}$$

of ordered pairs. Given $\Sigma_{\mathcal{N}}$, it is clear that we can construct $\Gamma_{\mathcal{N}}$ in time $O(|V|)$.

To see that ancestral sets is a weaker notion than ancestral tuples, consider the two orchard networks $\mathcal{N}_1$ and $\mathcal{N}_2$ shown in Fig. 5, where the non-leaf vertices have been labelled $1, 2, \ldots, 8$. For each $i \in \{1, 2\}$, the ancestral sets

of x_1, x_2, and x_3 are $\{1,2,3,4,5,7\}$, $\{1,2,\ldots,8\}$, and $\{1,2,3\}$, respectively. But $\mathcal{N}_1$ is not isomorphic to $\mathcal{N}_2$. Note that, for a fixed ordering of $1,2,\ldots,8$, the ancestral tuple of x_2 differs in $\mathcal{N}_1$ and $\mathcal{N}_2$ even though the ancestral tuples of x_1 and x_3 are the same for $\mathcal{N}_1$ and $\mathcal{N}_2$. Nevertheless, despite this example, the ancestral sets of a phylogenetic network $\mathcal{N}$ do provide some information regarding the structure of $\mathcal{N}$. As this is of possible independent interest, we highlight this in the next section where the preliminary lemmas are established in terms of ancestral sets.

The third remark concerns the relationship between orchard networks and the increasingly prominent class of tree-based networks [8]. A phylogenetic network $\mathcal{N}$ on X with root ρ and vertex set V is *tree-based* if it has, as a subgraph, a rooted subtree with root ρ, vertex set V, and leaf set X. Note that ρ in the subtree may have out-degree one. It is shown in [10] that the class of orchard networks is a proper subclass of tree-based networks. To see that it is proper, observe that the non-orchard networks $\mathcal{N}_1$ and $\mathcal{N}_2$ in Fig. 4 are both tree-based. Thus, the networks in this figure also show that Theorem 2.2 does not extend to tree-based networks.

3. Preliminary Lemmas

In this section, we establish several results that will be used in the proof of Theorem 2.2. These results show that the ancestral sets, and thus the ancestral tuples, of an arbitrary phylogenetic network recognise and distinguish cherries and reticulated cherries.

Lemma 3.1. *Let $\mathcal{N}$ be a phylogenetic network on X, and let a and b be distinct elements in X. Then $\gamma(a) \subseteq \gamma(b)$ if and only if the parent of b is reachable from the parent of a.*

Proof. Let p_a and p_b denote the parents of a and b, respectively. If p_b is reachable from p_a, then it is clear that $\gamma(a) \subseteq \gamma(b)$. To prove the converse, suppose that $\gamma(a) \subseteq \gamma(b)$. Then $p_a \in \gamma(b)$ and so, by definition, b is reachable from p_a. In turn, this implies that p_b is reachable from p_a. □

The next corollary immediately follows from Lemma 3.1 and the fact that phylogenetic networks are acyclic.

Corollary 3.2. *Let $\mathcal{N}$ be a phylogenetic network on X, and let $\{a,b\}$ be a 2-element subset of X. Then $\{a,b\}$ is a cherry in $\mathcal{N}$ if and only if $\gamma(a) = \gamma(b)$.*

Lemma 3.3. *Let $\mathcal{N}$ be a phylogenetic network on X, and let $\{a,b\}$ be a 2-element subset of X. Then $\{a,b\}$ is a reticulated cherry of $\mathcal{N}$ in which b is the reticulation leaf if and only if*

(i) $\gamma(a) \subsetneq \gamma(b)$,
(ii) *there is no* $x \in X - b$ *such that* $\gamma(a) \subset \gamma(x)$, *and*
(iii) $\left|\gamma(b) - \bigcup_{x \in X-b} \gamma(x)\right| = 1$.

Proof. Let p_a and p_b denote the parents of a and b, respectively. It is easily checked that if $\{a, b\}$ is a reticulated cherry in which b is the reticulation leaf, then (i)–(iii) hold. So suppose that (i)–(iii) hold. Since (i) holds, it follows by Lemma 3.1 that there is a directed path P in $\mathcal{N}$ from p_a to p_b. If p_b is a tree vertex, then $\mathcal{N}$ has a leaf, c say, reachable from p_b such that $c \neq b$. This implies that $\gamma(a) \subset \gamma(c)$, contradicting (ii). Therefore p_b is a reticulation. Lastly, assume (p_a, p_b) is not an arc in $\mathcal{N}$. Let u denote the vertex on P immediately prior to p_b. If u is a tree vertex, then $\mathcal{N}$ has a leaf $c' \neq b$ reachable from u with $\gamma(a) \subset \gamma(c')$, contradicting (ii). On the other hand, if u is a reticulation, then

$$\left|\gamma(b) - \bigcup_{x \in X-b} \gamma(x)\right| \geq 2,$$

contradicting (iii). Thus (p_a, p_b) is an arc and so $\{a, b\}$ is a reticulated cherry in which b is the reticulation leaf. □

4. Order Does Not Matter

Let $\mathcal{N}$ be an orchard network. Then, by definition, there exists a complete cherry-reduction sequence for $\mathcal{N}$. But, how do we find such a sequence and does the order in which we apply the cherry reductions matter? The next proposition says that if we take $\mathcal{N}$ and repeatedly apply cherry reductions until no more is possible, we always construct a complete cherry-reduction sequence. A vertex on a directed path is *non-terminal* if it is neither the first nor last vertex on the path.

Proposition 4.1. *Let* $\mathcal{N}$ *be an orchard network, and let*

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_\ell \tag{2}$$

be a maximal sequence of cherry reductions. Then this sequence is complete.

Proof. Let X denote the leaf set of $\mathcal{N}$, and suppose (2) is not complete. Paralleling (2), we begin by constructing a sequence

$$\mathcal{N} = \mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_\ell$$

of rooted acyclic directed graphs as follows. If $\mathcal{N}_1$ is obtained from $\mathcal{N}_0$ by reducing a leaf of a cherry, then $\mathcal{M}_1$ is obtained from $\mathcal{M}_0$ by deleting the same leaf but not suppressing the resulting vertex of in-degree one and out-degree one. Similarly, if $\mathcal{N}_1$ is obtained from $\mathcal{N}_0$ by cutting a reticulated

cherry, then $\mathcal{M}_1$ is obtained from $\mathcal{M}_0$ by deleting the same reticulation arc but not suppressing the two resulting vertices of in-degree one and out-degree one. More generally, if $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by reducing a leaf of a cherry, that is, deleting a leaf b say and suppressing its parent p_b, then $\mathcal{M}_i$ is obtained from $\mathcal{M}_{i-1}$ by deleting b as well as deleting every non-terminal vertex on the (unique) path from p_b to b in $\mathcal{M}_{i-1}$. Note that each of these non-terminal vertices has in-degree one and out-degree one in $\mathcal{M}_{i-1}$. On the other hand, if $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by cutting a reticulated cherry, that is, deleting a reticulation arc (p_a, p_b) and suppressing p_a and p_b, then $\mathcal{M}_i$ is obtained from $\mathcal{M}_{i-1}$ by deleting (p_a, p_b). Observe that, for all i, if we suppress every vertex in $\mathcal{M}_i$ of in-degree one and out-degree one, we obtain $\mathcal{N}_i$. Thus $\mathcal{M}_i$ is a subdivision of $\mathcal{N}_i$ for all i, that is, $\mathcal{N}_i$ can be obtained from $\mathcal{M}_i$ by suppressing all vertices of in-degree one and out-degree one for all i. Furthermore, as (2) is not complete, the root ρ of $\mathcal{N}$ is never deleted and so, for all i, the root of $\mathcal{M}_i$ is also ρ and has out-degree two in $\mathcal{M}_i$.

We now analyse $\mathcal{M}_\ell$. Since (2) is maximal and not complete, $\mathcal{N}_\ell$ has at least one reticulation. This implies that $\mathcal{M}_\ell$ has at least one vertex of in-degree two and out-degree one. We next show that every non-terminal vertex in $\mathcal{M}_\ell$ on a path from ρ to a vertex of in-degree two and out-degree one has degree three.

4.1.1. *Let v be a vertex of in-degree two and out-degree one in $\mathcal{M}_\ell$. If u is a non-terminal vertex of $\mathcal{M}_\ell$ on a path in $\mathcal{M}_\ell$ from ρ to v, then u has degree three in $\mathcal{M}_\ell$.*

Proof. Suppose u is a vertex of in-degree one and out-degree one on a path from ρ to v in $\mathcal{M}_\ell$. In $\mathcal{N}$, the vertex u has degree three. Therefore, for some $i \in \{1, 2, \ldots, \ell\}$, we have that $\mathcal{N}_i$ is obtained from $\mathcal{N}_{i-1}$ by a cherry reduction in which an arc incident with u is deleted. Now, as v is a vertex of in-degree two and out-degree one in $\mathcal{M}_\ell$, it follows that v is a reticulation in $\mathcal{N}_\ell$, and therefore a reticulation in $\mathcal{N}_i$. Thus there is a path P in $\mathcal{N}_i$ from u to v. It is now easily checked that no cherry reduction applied to $\mathcal{N}_{i-1}$ in which an arc incident with u and not lying on P is deleted is possible. Hence u has degree-three. □

We now complete the proof of the proposition. Since $\mathcal{N}$ is orchard, there is a sequence

$$\mathcal{N} = \mathcal{N}'_0, \mathcal{N}'_1, \mathcal{N}'_2, \ldots, \mathcal{N}'_k$$

of cherry reductions such that $\mathcal{N}'_k$ consists of a single vertex. Let i be the smallest index such that $\mathcal{N}'_i$ is obtained from $\mathcal{N}'_{i-1}$ by cutting a reticulated cherry in which the deleted reticulation arc, (u, v) say, has the property that v is in $\mathcal{M}_\ell$ and it has in-degree two and out-degree one in $\mathcal{M}_\ell$. Observe that, by the choice of i, no vertex of in-degree two and out-degree one is reachable

from v in $\mathcal{M}_\ell$ except v itself. As (2) is maximal, this implies that there is a unique vertex, ℓ_v say, in X that is reachable from v in $\mathcal{M}_\ell$.

Now, u is a tree vertex in $\mathcal{N}'_{i-1}$ whose other child, in addition to v, is a leaf. By (4.1.1), u has degree-three in $\mathcal{M}_\ell$. Furthermore, as u is a tree vertex in $\mathcal{N}'_{i-1}$, it follows that u has in-degree one and out-degree two in $\mathcal{M}_\ell$. Let w denote the child of u in $\mathcal{M}_\ell$ that is not v. At least one vertex in X is reachable from w in $\mathcal{M}_\ell$ and this vertex is not ℓ_v. If, in $\mathcal{M}_\ell$, there is no vertex reachable from w with in-degree two and out-degree one, then (2) is not maximal. Therefore, in $\mathcal{M}_\ell$ there is such a vertex w' reachable from w. In $\mathcal{N}$, the vertex w' is a reticulation, and so there is a $j \in \{1, 2, \ldots, k\}$ such that $\mathcal{N}'_j$ is obtained from $\mathcal{N}'_{j-1}$ by cutting a reticulated cherry in which a reticulation arc directed into w' is deleted. Since (u, v) is the reticulation arc directed into v that is deleted, it follows $j < i$. But, by the choice of i, we have $i < j$; a contradiction. We conclude that (2) is complete. □

The following corollary is an immediate consequence of Proposition 4.1.

Corollary 4.2. *Let $\mathcal{N}$ be an orchard network, and let $\{a, b\}$ be a cherry or a reticulated cherry of $\mathcal{N}$. If $\mathcal{N}'$ is obtained from $\mathcal{N}$ by reducing b if $\{a, b\}$ is a cherry or cutting $\{a, b\}$ if $\{a, b\}$ is a reticulated cherry, then $\mathcal{N}'$ is an orchard network.*

Since deciding if a given pair of leaves of a phylogenetic network is either a cherry or a reticulated cherry takes constant time and a cherry reduction also takes constant time, the last corollary gives a polynomial-time algorithm for deciding if an arbitrary phylogenetic network $\mathcal{N}$ is orchard. In particular, repeatedly find a cherry or a reticulated cherry, and apply the appropriate cherry reduction until this process is no longer possible. This takes at most $O(|V|)$ iterations, where V is the vertex of $\mathcal{N}$. If at the completion of this process, we have a phylogenetic network consisting of a single vertex, then $\mathcal{N}$ is orchard; otherwise, $\mathcal{N}$ is not orchard. Observe that if $\mathcal{N}$ is orchard with n leaves and k reticulations, then this process consists of $n + k - 1$ cherry reductions.

5. Proof of Theorem 2.2

In this section, we prove Theorem 2.2. For a phylogenetic network $\mathcal{N}$, Corollary 3.2 and Lemma 3.3 show that it is straightforward to recognise cherries and reticulated cherries of $\mathcal{N}$ using only the ancestral sets, and thus the ancestral tuples, of $\mathcal{N}$. This fact is freely used throughout this section. We next describe two operations on tuples that parallel the operations of reducing a cherry and cutting a reticulated cherry.

Let X be a non-empty finite set and, for some fixed t, let

$$\Sigma = \{(x, \sigma(x)) : x \in X\}$$

be a set of ordered pairs, where, for all $x \in X$, we have that $\sigma(x)$ is a t-tuple whose entries are either non-negative integers or $-$. Note that the symbol $-$ is going to be used as a placeholder. Let $\{a, b\}$ be a 2-element subset of X. The first operation will be used only in association with reducing b when $\{a, b\}$ is a cherry. Let $j \in \{1, 2, \ldots, t\}$ such that $\sigma_j(a) = \sigma_j(b) = 1$, but $\sigma_j(x) = 0$ for all $x \in X - \{a, b\}$. Let Σ' be the set of $|X - b|$ ordered pairs obtained from Σ as follows. For all $x \in X - b$, set $\sigma'(x)$ so that the i-th entry is

$$\sigma'_i(x) = \begin{cases} \sigma_i(x), & \text{if } i \neq j; \\ -, & \text{if } i = j. \end{cases}$$

Set $\Sigma' = \{(x, \sigma'(x)) : x \in X - b\}$. We say that Σ' has been obtained from Σ by *reducing* b.

The second operation will be used only in association with cutting $\{a, b\}$ when $\{a, b\}$ is a reticulated cherry in which b is the reticulation leaf. Let $j \in \{1, 2, \ldots, t\}$ such that $\sigma_j(a) = 1 = \sigma_j(b)$ but $\sigma_j(x) = 0$ for all $x \in X - \{a, b\}$, and let $k \in \{1, 2, \ldots, t\}$ such that $\sigma_k(b) = 1$ but $\sigma_k(x) = 0$ for all $x \in X - b$. Let Σ' be the set of $|X|$ ordered pairs obtained from Σ as follows. For all $x \in X - b$, set $\sigma'(x)$ so that the i-th entry is

$$\sigma'_i(x) = \begin{cases} \sigma_i(x), & \text{if } i \notin \{j, k\}; \\ -, & \text{if } i \in \{j, k\}; \end{cases}$$

and set $\sigma'(b)$ so that the i-th entry is

$$\sigma'_i(b) = \begin{cases} \sigma_i(b) - \sigma_i(a), & \text{if } i \notin \{j, k\}; \\ -, & \text{if } i \in \{j, k\}. \end{cases}$$

Set $\Sigma' = \{(x, \sigma'(x) : x \in X\}$. We say that Σ' has been obtained from Σ by *cutting* $\{a, b\}$.

Lemma 5.1. *Let $\mathcal{N}$ be a phylogenetic network on X with vertex set V and $|X| \geq 2$, and fix an ordering of $V - X$. Let $\{a, b\}$ be a 2-element subset of X.*

(i) *If $\{a, b\}$ is a cherry of $\mathcal{N}$, then, up to entries with symbol $-$, the set of ordered pairs obtained from $\Sigma_{\mathcal{N}}$ by reducing b is the ancestral profile of the phylogenetic network $\mathcal{N}'$ obtained from $\mathcal{N}$ by reducing b.*

(ii) *If $\{a, b\}$ is a reticulated cherry of $\mathcal{N}$ in which b is the reticulation leaf, then, up to entries with symbol $-$, the set of ordered pairs obtained from $\Sigma_{\mathcal{N}}$ by cutting $\{a, b\}$ is the ancestral profile of the phylogenetic network $\mathcal{N}'$ obtained from $\mathcal{N}$ by cutting $\{a, b\}$.*

Proof. We prove the lemma for (ii). The proof of the lemma for (i) is similar, but easier, and omitted. Suppose $\{a, b\}$ is a reticulated cherry of $\mathcal{N}$ in which b is the reticulation leaf, and $\mathcal{N}'$ is obtained from $\mathcal{N}$ by cutting $\{a, b\}$. Let Σ' be the set of ordered pairs obtained from $\Sigma_{\mathcal{N}}$ by cutting $\{a, b\}$. We will show that Σ' is the ancestral profile of a phylogenetic network isomorphic to $\mathcal{N}'$.

Let V denote the vertex set of $\mathcal{N}$, and fix an ordering $v_1, v_2, \ldots, v_t$ of the vertices in $V - X$. Let p_a and p_b denote the parents of a and b, respectively, in $\mathcal{N}$. Set

$$U_a = \{v_j \in V - X : \ \sigma_j(a) = 1 = \sigma_j(b),\ \sigma_j(x) = 0 \text{ for all } x \in X - \{a, b\}\}$$

and

$$U_b = \{v_k \in V - X : \ \sigma_k(b) = 1,\ \sigma_k(x) = 0 \text{ for all } x \in X - b\}.$$

Observe that U_a and U_b are both non-empty as $p_a \in U_a$ and $p_b \in U_b$, but $U_a \cap U_b$ is empty.

Now consider Σ'. To obtain Σ' from $\Sigma_{\mathcal{N}}$, we chose (i) an entry in $\sigma(a)$, say j, such that $\sigma_j(a) = 1 = \sigma_j(b)$ but $\sigma_j(x) = 0$ for all $x \in X - \{a, b\}$, and (ii) an entry in $\sigma(b)$, say k, such that $\sigma_k(b) = 1$ but $\sigma_k(x) = 0$ for all $x \in X - b$. In particular, these chosen entries correspond to vertices, v_j and v_k say, in U_a and U_b, respectively.

Let $\mathcal{N}_1$ denote the phylogenetic network obtained from $\mathcal{N}$ by bijectively relabelling the vertices in U_a with the vertices in U_a so that p_a is relabelled v_j, and bijectively relabelling the vertices in U_b with the vertices in U_b so that p_b is relabelled v_k. Clearly, $\mathcal{N}_1$ is isomorphic to $\mathcal{N}$ and $\Sigma_{\mathcal{N}}$ is the ancestral profile of $\mathcal{N}_1$. Furthermore, it is easily checked that, up to isomorphism, Σ' is the ancestral profile of the phylogenetic network $\mathcal{N}_1'$ obtained from $\mathcal{N}_1$ by cutting $\{a, b\}$. But $\mathcal{N}_1'$ is isomorphic to $\mathcal{N}'$, thereby completing the proof of the lemma. □

With Lemma 5.1 in hand, we next prove the uniqueness part of Theorem 2.2

Proof of the uniqueness part of Theorem 2.2. The proof is by induction on the sum of the number n of leaves and the number k of reticulations in $\mathcal{N}$. If $n + k = 1$, then $n = 1$ and $k = 0$, and $\mathcal{N}$ consists of the single vertex in X, and so uniqueness holds. If $n + k = 2$, then, as $\mathcal{N}$ is orchard, $n = 2$ and $k = 0$, in which case, $\mathcal{N}$ consists of two leaves attached to the root. Again, uniqueness holds. Now suppose that $n + k \geq 3$ and the uniqueness holds for all orchard networks for which the sum of the number of leaves and the number of reticulations is at most $n + k - 1$. Note that, as $\mathcal{N}$ is orchard, $n \geq 2$.

Since $\mathcal{N}$ is orchard, it has either a cherry or a reticulated cherry. Thus, by Corollary 3.2 and Lemma 3.3, it is possible to find a 2-element subset $\{a, b\}$ of X using only $\Sigma_{\mathcal{N}}$ such that $\{a, b\}$ is either a cherry or a reticulated cherry of $\mathcal{N}$. If the latter, we can also determine from $\Sigma_{\mathcal{N}}$ which of a and b is the reticulation. Without loss of generality, we may assume b is the reticulation leaf. Depending on whether $\{a, b\}$ is a cherry or a reticulated cherry, let $\mathcal{N}'$ be obtained from $\mathcal{N}$ by reducing b or cutting $\{a, b\}$, respectively, and let Σ' be the set of ordered pairs obtained from $\Sigma_{\mathcal{N}}$ by reducing b or cutting $\{a, b\}$, respectively. Regardless of the way $\mathcal{N}'$ and Σ' are obtained, it follows by Corollary 4.2 and Lemma 5.1 that $\mathcal{N}'$ is an orchard network and, up to isomorphism, Σ' is the ancestral profile of $\mathcal{N}'$. Furthermore, $\mathcal{N}'$ has either $n-1$ leaves and k reticulations if $\{a, b\}$ is a cherry, or n leaves and $k-1$ reticulations if $\{a, b\}$ is a reticulated cherry. Therefore, by the induction assumption, up to isomorphism, $\mathcal{N}'$ is the unique phylogenetic network whose ancestral profile is Σ'.

Now let $\mathcal{N}_1$ be a phylogenetic network on X such that $\Sigma_{\mathcal{N}}$ is the ancestral profile of $\mathcal{N}_1$. Note that $\mathcal{N}_1$ has the same number of non-leaf vertices as $\mathcal{N}$, but not necessarily the same number of reticulations. First assume $\{a, b\}$ is a cherry of $\mathcal{N}$. Then, by Corollary 3.2, $\{a, b\}$ is a cherry of $\mathcal{N}_1$. Let $\mathcal{N}_1'$ denote the phylogenetic network obtained from $\mathcal{N}_1$ by reducing b. By Lemma 5.1(i), up to isomorphism, Σ' is the ancestral profile of $\mathcal{N}_1'$. Thus, by the induction assumption, $\mathcal{N}_1'$ is isomorphic to $\mathcal{N}'$. Since $\{a, b\}$ is a cherry of $\mathcal{N}_1$ and $\mathcal{N}$, it follows that $\mathcal{N}_1$ is isomorphic to $\mathcal{N}$.

Lastly, assume $\{a, b\}$ is a reticulated cherry of $\mathcal{N}$. Then, by Lemma 3.3, $\{a, b\}$ is a reticulated cherry of $\mathcal{N}_1$ in which b is the reticulation leaf. Let $\mathcal{N}_1'$ be the phylogenetic network obtained from $\mathcal{N}_1$ by cutting $\{a, b\}$. By Lemma 5.1(ii), up to isomorphism, Σ' is the ancestral profile of $\mathcal{N}_1'$. Hence, by the induction assumption, $\mathcal{N}_1'$ is isomorphic to $\mathcal{N}'$. As $\{a, b\}$ is a reticulated cherry of $\mathcal{N}$ and $\mathcal{N}_1$ in which b is the reticulation leaf, we have that $\mathcal{N}_1$ is isomorphic to $\mathcal{N}$. This completes the proof of the uniqueness part of Theorem 2.2. □

5.1. **The algorithm.** Let $\mathcal{N}$ be an orchard network on X, and let Σ denote the ancestral profile of $\mathcal{N}$. Called Orchard Tuple, we next describe an algorithm which takes as its input X and Σ, and returns a phylogenetic network $\mathcal{N}_1$ on X that is isomorphic to $\mathcal{N}$. The proof that the algorithm works correctly is essentially the same as that used to prove the uniqueness part of Theorem 2.2, and so it is omitted. The running time of the algorithm follows its description.

1. If $|X| = 1$, then return the phylogenetic network consisting of the single vertex in X.

2. Else, find a 2-element subset, $\{a, b\}$ say, of X such that either (I) $\gamma(a) = \gamma(b)$ or (II) $\gamma(a) \subset \gamma(b)$, there is no $x \in X - b$ with $\gamma(a) \subseteq \gamma(x)$, and

$$\left| \gamma(b) - \bigcup_{x \in X - b} \gamma(x) \right| = 1.$$

   (a) If $\{a, b\}$ satisfies (I) (in which case $\{a, b\}$ is a cherry), then
      (i) Reduce b in Σ to give the set Σ' of $|X - b|$ ordered pairs.
      (ii) Apply Orchard Tuple to input $X' = X - b$ and Σ'. Construct $\mathcal{N}_1$ from the returned phylogenetic network $\mathcal{N}_1'$ on X' by subdividing the arc incident to a with a new vertex p_a, and adjoining a new leaf b via the new arc (p_a, b). If $|X'| = 1$, then set $\mathcal{N}_1$ to be the phylogenetic network consisting of the leaves a and b adjoined to the root. Return $\mathcal{N}_1$.
   (b) Else, $\{a, b\}$ satisfies (II) (in which case $\{a, b\}$ is a reticulated cherry and b is the reticulation leaf).
      (i) Cut $\{a, b\}$ in Σ to give the set Σ' of $|X|$ ordered pairs.
      (ii) Apply Orchard Tuple to X and Σ'. Construct $\mathcal{N}_1$ from the returned phylogenetic network $\mathcal{N}_1'$ on X by subdividing the arcs incident to a and b with new vertices p_a and p_b, respectively, and adding the new arc (p_a, p_b). Return $\mathcal{N}_1$.

We now consider the running time of Orchard Tuple. The input to the algorithm is a set X and the ancestral profile of an orchard network $\mathcal{N}$ on X whose entries are either a non-negative integer or the symbol $-$. Let V denote the vertex set of $\mathcal{N}$. As noted earlier, the set $\Gamma_{\mathcal{N}} = \{(x, \gamma(x)) : x \in X\}$ can be determined from Σ in $O(|V|)$ time. This is a preprocessing step and it will have no effect on the theoretical running time. Except for when $|X| \in \{1, 2\}$, in which case, Orchard Tuple runs in constant time, each iteration begins by finding a 2-element subset of X satisfying either (I) or (II). This takes $O(|X|^2|V|)$ time as there are $O(|X|^2)$ two-element subsets of X and each subset takes $O(|V|)$ time to decide if is satisfies either (I) or (II). Once such a 2-element is found, we construct Σ'. Regardless of the way Σ' is constructed, this takes $O(|X||V|)$ time. When $\mathcal{N}_1'$ is returned, we augment to $\mathcal{N}_1$ in constant time, and so each iteration takes $O(|X|^3|V|^2)$ time.

When we recurse, Σ' is the ancestral profile of an orchard network with either one less leaf or one less reticulation than an orchard network for which Σ is the ancestral profile. Thus the total number of iterations is $O(|V|)$. We conclude that Orchard Tuple completes in $O(|X|^3|V|^3)$ time. This completes the proof of Theorem 2.2.

6. Conclusion

The main result of this paper, Theorem 2.2, shows that the ancestral profile of an orchard network $\mathcal{N}$ on X uniquely determines $\mathcal{N}$ amongst all phylogenetic networks on X. This generalises results in both [4] and [5], which considered tree-sibling time-consistent networks and tree-child networks (subclasses of orchard networks whose number of reticulations is at most linear in the number of leaves). Curiously, these later results have a different motivation compared to what motivated Theorem 2.2. There the motivation is to construct a distance measure (metric) on the classes of tree-sibling time-consistent networks and tree-child networks which is computable in polynomial time. Recalling that they considered the equivalent notion of path-tuples, for two tree-sibling time-consistent (resp. tree-child) networks $\mathcal{N}_1$ and $\mathcal{N}_2$, the distance between $\mathcal{N}_1$ and $\mathcal{N}_2$ is the value

$$|\Pi_{\mathcal{N}_1} \triangle \Pi_{\mathcal{N}_2}|\,,$$

where the symmetric difference and the cardinality operator refer to multisets. It is easily checked that this same measure extends to the class of orchard networks.

As noted in the introduction, our result does not relate to specific biological data that is readily available at present. However, a type of data that might provide ancestral profile information would be genomic fragments that follow lineage splitting and reticulation events, so that when a reticulation occurs, a trace of each fragment in the incoming lineage is preserved in (different regions of) the reticulate genome.

Lastly, we end with a question asked by one of the referees. For a given orchard network $\mathcal{N}$, is it possible to count the number of complete cherry-reduction sequences of $\mathcal{N}$?

Acknowledgements

We thank the three anonymous referees for their careful reading of the paper and constructive comments.

References

[1] M. Baroni, M. Steel, Accumulation phylogenies, Annals of Combinatorics 10 (2006) 19–30.

[2] M. Bordewich, K.T. Huber, V. Moulton, C. Semple, Recovering normal networks from shortest inter-taxa distance information, Journal of Mathematical Biology 77 (2018) 571–594.

[3] M. Bordewich, C. Semple, Determining phylogenetic networks from inter-taxa distances, Journal of Mathematical Biology 73 (2016) 283–303.
[4] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, A distance metric for a class of tree-sibling phylogenetic networks 24 (2008) 1481–1488.
[5] G. Cardona, F. Rosselló, G. Valiente, Comparison of tree-child phylogenetic networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics 6 (2009) 552–569.
[6] W.F. Doolittle, Phylogenetic classification and the universal tree, Science 284 (1999) 2124–2128.
[7] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Sunderland, MA, 2004.
[8] A. R. Francis, M. Steel, Which phylogenetic networks are merely trees with additional arcs?, Systematic Biology 64 (2015) 768–777.
[9] P. Gambette, K.T. Huber, On encodings of phylogenetic networks of bounded level, Journal of Mathematical Biology 65 (2012) 157–180.
[10] K.T. Huber, L. van Iersel, R. Janssen, M. Jones, V. Moulton, Y. Murakami, C. Semple, Rooting for phylogenetic networks, in preparation.
[11] D.H. Huson, R. Rupp, C. Scornavacca, Phylogenetic Networks: Concepts, Algorithms and Applications, Cambridge University Press, 2010.
[12] L. van Iersel, V. Moulton, Trinets encode tree-child and level-2 phylogenetic networks, Journal of Mathematical Biology 68 (2014) 1707–1729.
[13] R. Janssen, Y. Murakami, Solving phylogenetic network containment problem using cherry-picking sequences, arXiv:1812.08065 (2018).
[14] W. Jetz, G.H. Thomas, J.B. Joy, K. Hartmann, A.O. Mooers, The global diversity of birds in space and time, Nature 491 (2012) 444–448.
[15] E.V. Koonin, The turbulent network dynamics of microbial evolution and the statistical tree of life, Journal of Molecular Evolution 80 (2015) 244–250.
[16] F. Pardi, C. Scornavacca, Reconstructible phylogenetic networks: Do not distinguish the indistinguishable, PLoS Computational Biology 11 (2015) e1004135.
[17] C. Semple, M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.
[18] S.J. Willson, Reconstruction of certain phylogenetic networks from the genomes at their leaves, Journal of Theoretical Biology 252 (2008) 338–349.
[19] S.J. Willson, Properties of normal phylogenetic networks, Bulletin of Mathematical Biology 72 (2010) 340–358.
[20] S.J. Willson, Regular networks can be uniquely constructed from their trees, IEEE/ACM Transactions on Computational Biology and Bioinformatics 8 (2011) 785–796.

Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary

E-mail address: erdos.peter@renyi.mta.hu

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

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

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

E-mail address: mike.steel@canterbury.ac.nz