

# DEFINING PHYLOGENETIC NETWORKS USING ANCESTRAL PROFILES

ALLAN BAI, PÉTER L. ERDŐS, CHARLES SEMPLE, AND MIKE STEEL

**ABSTRACT.** Rooted phylogenetic networks provide a more complete representation of the ancestral relationship between species than phylogenetic trees when reticulate evolutionary processes are at play. One way to reconstruct a phylogenetic network is to consider its ‘ancestral profile’ (the number of paths from each ancestral vertex to each leaf). In general, this information does not uniquely determine the underlying phylogenetic network. A recent paper considered a new class of phylogenetic networks called ‘orchard networks’ where this uniqueness was claimed to hold. Here we show that an additional restriction on the network, that of being ‘stack-free’, is required in order for the original uniqueness claim to hold. On the other hand, if the additional stack-free restriction is lifted, we establish an alternative result; namely, there is uniqueness within the class of orchard networks up to the resolution of vertices of high in-degree.

## 1. INTRODUCTION

Evolutionary relationships between species are generally represented by phylogenetic trees, where the species at the present appearing as the leaves of the tree, and ancestral species corresponding to interior vertices. Over the last several decades, a wide variety of methods have been developed for reconstructing phylogenetic trees from genomic data [7], and these are now widely used in large-scale studies in systematic biology (e.g., [10, 16]) and associated fields (e.g. in epidemiology to classify strains of viruses such as HIV, influenza, and SARS-Cov2 [19]). However, for certain groups of organisms, the tree model is overly simplistic. This is because of the intricacies of ancestral processes whereby lineages not only split, but sometimes combine together to form new lineages. This latter pattern of evolution is

---

*Date:* December 2, 2020.

*2020 Mathematics Subject Classification.* 05C85, 92D15.

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

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

collectively referred to as ‘reticulation’, and includes the formation of hybrid species, horizontal gene transfer, and endosymbiosis events [5, 8, 11]. Consequently, certain portions of the ‘Tree of Life’ are better described by a phylogenetic network that explicitly exhibits reticulation events. Although there is a well-developed theory for reconstructing phylogenetic trees from various types of data [7, 15], phylogenetic network reconstruction is much more subtle. In particular, for certain types of data it is impossible to distinguish between different (non-isomorphic) phylogenetic networks [13]. One way to address this non-identifiability issue is to work within a subclass of phylogenetic networks that includes phylogenetic trees along with phylogenetic networks that are sufficiently tame. An example is the class of ‘normal’ networks, for which certain reconstructive results have been established [1, 12, 17, 18]. The slightly more general class of ‘tree-child’ networks also allows for unique reconstruction from various types of data [3, 4].

In this paper, we focus on the unique reconstruction of networks from their ‘ancestral profile’, which, roughly speaking, is the number of paths from each ancestral vertex in the network to each extant leaf. It was shown that all binary ‘tree-sibling time-consistent’ and all binary ‘tree-child’ networks are uniquely determined by their ancestral profile [3, 4]. In a recent paper [6], this result was extended to the larger class of binary ‘orchard networks’, which allows for an unbounded number of vertices in the network for a given number of leaves. This contrasts with the classes of binary ‘tree-sibling time-consistent’ and binary ‘tree-child’ networks, for which the size of the network is bounded by the number of leaves. However, the result in [6] omitted an extra condition required for unique reconstruction, namely, the network cannot contain a tower (‘stack’) of reticulations. We show here that this ‘stack-free’ condition is necessary, and that when this extra condition is included the original result claimed in [6] holds. Moreover, this result then generalises (Theorem 3.1) to the class of stack-free orchard networks in which reticulate vertices are allowed to have arbitrarily high in-degree. Note that, the uniqueness is amongst all phylogenetic networks with vertices of arbitrarily high in-degree, that is, the ancestral profile of a stack-free orchard network is always different to the ancestral profile of any other phylogenetic network, even if it is neither orchard nor stack-free. When the stack-free condition is lifted, we describe a second result (Theorem 3.2) which states that, within the class of orchard networks, the ancestral profile of an orchard network uniquely determines the orchard network up to the resolution of vertices of high in-degree.

The structure of the paper is as follows. The next section recalls definitions of phylogenetic networks (which are permitted here to contain vertices of high in-degree), along with the notion of ancestral profile. We also describe the class of orchard networks. This class was introduced and studied independently in [6] (for binary networks) and [9] (for networks that allow

high in-degree). In Section 3, we turn to the question of whether the ancestral profile of an orchard network determines that network (either within the class of orchard networks, or more generally), and state the two main results of the paper. The first main result, Theorem 3.1, states a corrected form of [6, Theorem 2.2] for stack-free networks. The necessary adjustments required for the proof of Theorem 3.1 are given in the Appendix. The second main result, Theorem 3.2, is a reconstructive result that holds when the stack-free condition is removed. Additionally, we discuss the relationships between Theorem 3.1 and the main results in [3, 4]. The proof of Theorem 3.2 is given in Section 4. Some concluding comments are given in Section 5, the last section of the paper.

## 2. PRELIMINARIES

Throughout the paper  $X$  denotes a non-empty finite set and, unless otherwise stated, all paths are directed. For sets  $A$  and  $B$ , we denote the set obtained from  $A$  by removing every element in  $A$  that is also in  $B$  by  $A - B$ . Furthermore, if  $(u, v)$  is an arc of an acyclic directed graph, we say  $u$  is a *parent* of  $v$ .

**Phylogenetic networks.** The following definition of phylogenetic network is slightly more general than in [6]. 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;
- (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 at least two and out-degree one.

We will refer to a phylogenetic network in which every vertex has in-degree at most two as a *binary* phylogenetic network.

We pause to make two technical remarks. First, if  $|X| = 1$ , we additionally allow a single vertex to be a phylogenetic network, in which case, the root is the vertex in  $X$ . Second, suppose that  $\mathcal{N}_1$  and  $\mathcal{N}_2$  are 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  $\varphi : V_1 \rightarrow V_2$  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  $u, v \in V_1$ .

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

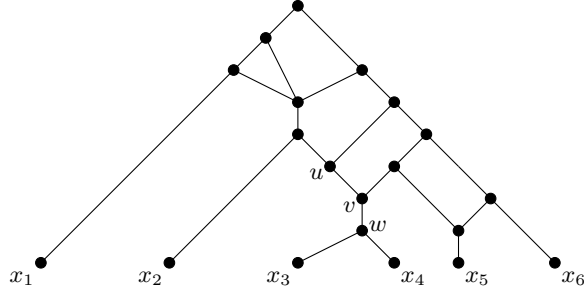


FIGURE 1. A phylogenetic network on  $\{x_1, x_2, \dots, x_6\}$ . The 2-element set  $\{x_3, x_4\}$  is a cherry, while the ordered pair  $\{x_5, x_6\}$  is a reticulated cherry, in which  $x_5$  is the reticulation leaf.

with in-degree one and out-degree two are *tree vertices*, while vertices of in-degree at least 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, a phylogenetic network on  $\{x_1, x_2, \dots, x_6\}$  is shown in Fig. 1. Vertices  $u$  and  $v$  are reticulations, while vertex  $w$  is a tree vertex. Here, as throughout the paper, all arcs are directed down the page.

**Ancestral tuples and ancestral profile.** Let  $\mathcal{N}$  be a phylogenetic network on  $X$  with vertex set  $V$ . Let  $v_1, v_2, \dots, v_t$  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  $\mathcal{N}'$ , 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.

To illustrate these notions consider the two networks  $\mathcal{N}$  and  $\mathcal{N}'$  shown in Fig. 2. Under the labelling of the non-leaf vertices of  $\mathcal{N}$  shown, we have

$$\begin{aligned} \Sigma_{\mathcal{N}} = \{ & (x_1, (1, 1, 0, 0, 0, 0)), (x_2, (1, 0, 1, 1, 0, 0)), \\ & (x_3, (1, 0, 1, 0, 1, 0, 0)), (x_4, (3, 1, 2, 1, 1, 1, 1)) \}. \end{aligned}$$

The other network  $\mathcal{N}'$  in Fig. 2 also realises  $\Sigma_{\mathcal{N}}$ , because under the ordering of the non-leaf vertices of  $\mathcal{N}'$  shown in this figure, we have  $\Sigma_{\mathcal{N}'} = \Sigma_{\mathcal{N}}$ . On the other hand,  $\mathcal{N}$  and  $\mathcal{N}'$  are not isomorphic. To see this observe that the parent of  $x_2$  in  $\mathcal{N}$  has a unique path to  $x_4$  of length 3, while the parent of  $x_2$  in  $\mathcal{N}'$  also has a unique path to  $x_4$  but this path has length 2.

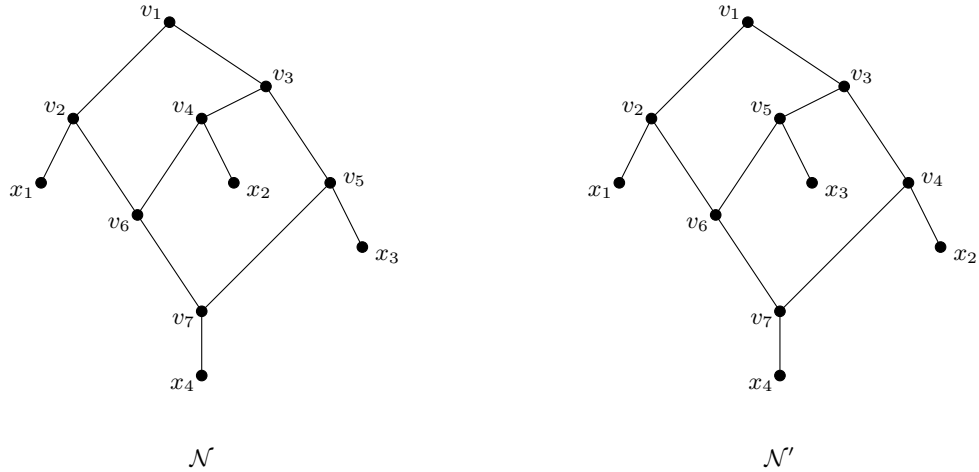


FIGURE 2. Two binary networks  $\mathcal{N}$  and  $\mathcal{N}'$  with the same ancestral profile (for the labelling of the vertices in  $V - X$  shown). However,  $\mathcal{N}$  and  $\mathcal{N}'$  are not isomorphic.

**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. As an example, in Fig. 1,  $\{x_3, x_4\}$  is a cherry, while  $\{x_5, x_6\}$  is a reticulated cherry with  $x_5$  as the reticulation leaf.

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 any resulting vertices of in-degree one and out-degree one. Note that the parent of  $a$  is always suppressed. However, the parent of  $b$  is suppressed only if its in-degree in  $\mathcal{N}$  is exactly two. 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*.

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

$$(1) \quad \mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \dots, \mathcal{N}_k$$

of phylogenetic networks is a *cherry-reduction sequence* of  $\mathcal{N}$  if, for all  $i \in \{1, 2, \dots, k\}$ , 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*. If  $\mathcal{N}$  has a complete cherry-reduction sequence, then  $\mathcal{N}$  is an *orchard network*. It is easily checked that the phylogenetic network shown in Fig. 1 is orchard.

A fundamental property of orchard networks is that if one cherry-reduction sequence leads to a single vertex (in which case  $\mathcal{N}$  is an orchard network), then every maximal cherry-reduction sequence leads to a single vertex (regardless of any choices made during the construction of a cherry-reduction sequence). This result was established for binary orchard networks in [6, Proposition 4.1], and independently shown to hold for general phylogenetic networks in [9, Theorem 1].

**Proposition 2.1.** *Let  $\mathcal{N}$  be an orchard network, and let*

$$\mathcal{N} = \mathcal{N}_0, \mathcal{N}_1, \dots, \mathcal{N}_\ell$$

*be a maximal cherry-reduction sequence. Then this sequence is complete.*

### 3. MAIN RESULTS

In this section we state the two main results of the paper. A *stack* in a phylogenetic network  $\mathcal{N}$  is a pair of reticulations,  $u$  and  $v$ , such that one of the reticulations, say  $u$ , is a parent of the other; that is,  $(u, v)$  is an arc of  $\mathcal{N}$ . We refer to  $(u, v)$  as a *stack arc* of  $\mathcal{N}$ . A phylogenetic network is said to be *stack-free* if it has no stacks.

It was claimed in [6, Theorem 2.2] that, up to isomorphism, every binary orchard network is uniquely determined by its ancestral profile. However, Fig. 2 shows a pair of non-isomorphic phylogenetic networks that are both binary orchard networks, and which have identical ancestral profiles. Notice that both networks in Fig. 2 contain a stack, in particular, reticulations  $v_6$  and  $v_7$ . A corrected version of [6, Theorem 2.2] is Theorem 3.1, the first main result of the paper, which is now extended to allow phylogenetic networks with reticulations of in-degree at least two. The proof follows the same argument as in [6], but some adjustments are required to certain lemmas to allow for the generality beyond binary phylogenetic networks, and (at one point) to impose the stack-free requirement. We describe the required adjustments to the original proof in the Appendix.

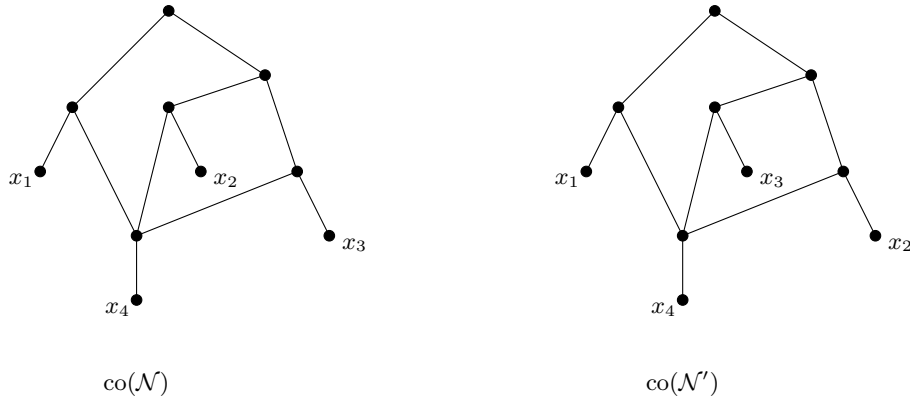


FIGURE 3. The stack identifications  $\text{id}(\mathcal{N})$  and  $\text{id}(\mathcal{N}')$  of the orchard networks  $\mathcal{N}$  and  $\mathcal{N}'$ , respectively, shown in Fig. 2. Observe that both  $\text{id}(\mathcal{N})$  and  $\text{id}(\mathcal{N}')$  are orchard networks, and  $\text{id}(\mathcal{N}) \cong \text{id}(\mathcal{N}')$ .

**Theorem 3.1.** *Let  $\mathcal{N}$  be a stack-free orchard network on  $X$  with vertex set  $V$ . Then, up to isomorphism,  $\mathcal{N}$  is the unique phylogenetic network on  $X$  realising  $\Sigma_{\mathcal{N}}$ .*

We now consider what can be said if the stack-free condition is lifted. Let  $\mathcal{N}$  be a phylogenetic network on  $X$  with vertex set  $V$ . Define a relation  $\sim'$  on  $V - X$  by writing  $u \sim' v$  if  $u$  and  $v$  are reticulations and either  $(u, v)$  or  $(v, u)$  is an arc of  $\mathcal{N}$ . Let  $\sim$  be the transitive closure of  $\sim'$ ; the equivalence classes of vertices under  $\sim$  are called *sinks*. Thus, a phylogenetic network  $\mathcal{N}$  is stack-free if and only if each of its sinks has size 1 (i.e. each reticulation forms its own equivalence class). The *stack identification* of  $\mathcal{N}$ , denoted  $\text{id}(\mathcal{N})$ , is the phylogenetic network obtained from  $\mathcal{N}$  by identifying all the vertices within each sink  $S$  to a single vertex  $v_S$  (and removing any arcs between vertices of the same sink). Observe that  $\text{id}(\mathcal{N})$  can be obtained from  $\mathcal{N}$  by repeatedly deleting each stack arc and identifying its end vertices. Note that  $\text{id}(\mathcal{N})$  is not necessarily a phylogenetic network because it may have arcs in parallel. However, if  $\mathcal{N}$  is orchard, then, as we show in the next section (Lemma 4.1),  $\text{id}(\mathcal{N})$  is also orchard. To illustrate the notion of stack identification, consider the two orchard networks  $\mathcal{N}$  and  $\mathcal{N}'$  shown in Fig. 2. The stack identifications of  $\mathcal{N}$  and  $\mathcal{N}'$  are shown in Fig. 3. Observe that  $\text{id}(\mathcal{N}) \cong \text{id}(\mathcal{N}')$ .

The next theorem is the second main result of the paper.

**Theorem 3.2.** *Let  $\mathcal{N}$  and  $\mathcal{N}'$  be orchard networks on  $X$ . If  $\mathcal{N}$  and  $\mathcal{N}'$  have the same ancestral profile, then  $\text{id}(\mathcal{N}) \cong \text{id}(\mathcal{N}')$ .*

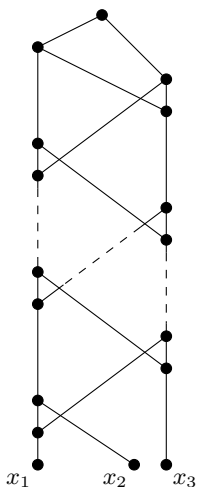


FIGURE 4. An orchard network that is binary and stack-free, for which the total number of vertices is not bounded by the size of its leaf set.

We end this section with a brief discussion concerning Theorem 3.1 and its relationship with the main results in [3, 4]. Let  $\mathcal{N}$  be a phylogenetic network. We say  $\mathcal{N}$  is *tree-child* if every non-leaf vertex of  $\mathcal{N}$  is a parent of a tree vertex or a leaf. Furthermore,  $\mathcal{N}$  is *tree-sibling* if every reticulation has a parent that is also a parent of a tree vertex or a leaf. Also,  $\mathcal{N}$  is *time-consistent* if there is a map  $t$  from the vertex set of  $\mathcal{N}$  to the non-negative integers having the property that if  $(u, v)$  is an arc of  $\mathcal{N}$ , then  $t(u) = t(v)$  if  $(u, v)$  is a reticulation arc; otherwise,  $t(u) < t(v)$ .

The class of stack-free orchard networks includes the class of tree-child networks as a proper subclass. (A proof that a binary tree-child network is orchard is given in [2]. The generalisation to allowing reticulations with in-degree more than two is straightforward.) Moreover, although a tree-child network on a leaf set of size  $n$  can have at most  $n - 1$  reticulations [4], a stack-free orchard network can have arbitrarily many reticulations, as indicated in Fig. 4. The class of stack-free orchard networks also includes tree-sibling time-consistent networks with no stacks. (A proof for the binary case is given in [6]. This proof generalises to allowing reticulations with in-degree more than two.) Like tree-child networks, the number of reticulations of such a network on a leaf set of size  $n$  is linear, in this case at most  $2n - 4$  [3].

Theorem 3.1 in part generalises results in [3, 4]<sup>1</sup>. These papers consider the classes of time-sibling time-consistent and tree-child networks, respectively, in the context of a formation on  $X$  for reconstruction that is equivalent

<sup>1</sup>The results in [3, 4] are slightly stronger than that described here as they allow tree vertices to have out-degree at least two.



to ancestral profile. They establish uniqueness results for tree-child [4, Theorem 1] and for binary tree-sibling time-consistent networks with no stacks [3, Theorem 6]. However, the uniqueness is within the respective classes. Thus, for example, in our terminology, it is shown in [4] that if  $\mathcal{N}$  is a tree-child network on  $X$ , then, up to isomorphism,  $\mathcal{N}$  is the unique tree-child network on  $X$  realising  $\Sigma_{\mathcal{N}}$ .

#### 4. PROOF OF THEOREM 3.2

The proof of Theorem 3.2 makes use of a sequence of lemmas. We begin by showing that the stack identification of an orchard network is orchard.

**Lemma 4.1.** *Let  $\mathcal{N}$  be an orchard network, and let  $e$  be a stack arc of  $\mathcal{N}$ . Suppose that  $\mathcal{N}'$  is obtained from  $\mathcal{N}$  by deleting  $e$  and identifying its end vertices. Then  $\mathcal{N}'$  is an orchard network.*

*Proof.* Let  $e = (u, v)$ . We first show that  $\mathcal{N}'$  has no parallel arcs, that is,  $\mathcal{N}'$  is a phylogenetic network. Assume  $\mathcal{N}'$  has two parallel arcs. Then, by the construction of  $\mathcal{N}'$ , these arcs are directed out of a tree vertex  $t$  and directed into the vertex, say  $v'$ , identifying  $u$  and  $v$ , in which case,  $(t, u)$  and  $(t, v)$  are arcs of  $\mathcal{N}$ . Since  $\mathcal{N}$  is orchard,  $\mathcal{N}$  has a complete cherry-reduction sequence  $\mathcal{S}$ . Applying  $\mathcal{S}$  to  $\mathcal{N}$ , this sequence eventually suppresses  $u$  and  $v$  via cutting a reticulated cherry. Clearly,  $v$  is suppressed before  $u$ . Since  $u$  is a reticulation parent of  $v$ , it follows that prior to  $v$  being suppressed,  $(t, v)$  is cut as part of a cherry reduction of  $\mathcal{S}$ . But this requires  $t$  to have a descendant leaf that is not a descendant of  $v$ . Since  $u$  is the other child of  $t$ , there are no such leaves, and so  $\mathcal{S}$  is not a cherry-reduction sequence of  $\mathcal{N}$ , a contradiction. Thus  $\mathcal{N}'$  has no parallel arcs.

We complete the proof by showing that  $\mathcal{N}'$  is orchard. Let

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

be a complete cherry-reduction sequence for  $\mathcal{N}$ . Since  $u$  and  $v$  are reticulations, and  $u$  is a parent of  $v$ , it follows that, for some  $i \in \{1, 2, \dots, k\}$ , the phylogenetic network  $\mathcal{N}_i$  is obtained from  $\mathcal{N}_{i-1}$  by cutting a reticulated cherry, and then suppressing  $u'$  and  $v$ , where  $u'$  is a parent of  $v$  that is not  $u$  in  $\mathcal{N}$ . A simple induction argument shows that exactly the same sequence of cherry reductions from  $\mathcal{N} = \mathcal{N}_0$  to  $\mathcal{N}_i$  can be applied to  $\mathcal{N}'$  to obtain the cherry-reduction sequence

$$\mathcal{N}' = \mathcal{N}'_0, \mathcal{N}'_1, \mathcal{N}'_2, \dots, \mathcal{N}'_i,$$

where, for all  $j \in \{0, 1, \dots, i-1\}$ , the phylogenetic network is obtained from  $\mathcal{N}_j$  by deleting  $(u, v)$ , and identifying  $u$  and  $v$ , and  $\mathcal{N}'_i \cong \mathcal{N}_i$ . Using the cherry-reduction sequence from  $\mathcal{N}_{i+1}$  to  $\mathcal{N}_k$ , it now follows that there exists a complete cherry-reduction sequence for  $\mathcal{N}'$ , and so  $\mathcal{N}'$  is orchard.  $\square$

The next corollary is an immediate consequence of Lemma 4.1.

**Corollary 4.2.** *Let  $\mathcal{N}$  be an orchard network. Then  $\text{id}(\mathcal{N})$  is an orchard network.*

We next describe two operations on sets of certain ordered pairs. These operations parallel the graph operations of reducing cherries and cutting reticulated cherries. Intuitively, these operations explicitly describe how the ancestral profile of a phylogenetic networks changes if we reduce a cherry or cut a reticulated cherry (see Lemma 4.3).

Let  $X$  be a non-empty set and, for some fixed non-negative integer  $t$ , let

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

be a set of ordered pairs, where  $\sigma(x)$  is a  $t$ -tuple each entry of which is either a non-negative integer or it is a placeholder symbol  $-$ . Note that  $\Sigma$  is an abstraction of  $\Sigma_{\mathcal{N}}$ , where  $\mathcal{N}$  is a phylogenetic network. We now describe two operations on  $\Sigma$  that correspond to the two cherry-reduction operations. Let  $\{a, b\}$  be a 2-element subset of  $X$ . The first operation corresponds to reducing  $b$  when  $\{a, b\}$  is a cherry. Let  $j \in \{1, 2, \dots, t\}$  such that  $\sigma_j(a) = \sigma_j(b) = 1$ , and  $\sigma_j(x) = 0$  for all  $x \in X - \{a, b\}$ . Let  $\Sigma'$  be the set of  $|X - \{b\}|$  ordered pairs obtained from  $\Sigma$  as follows. For all  $x \in X - \{b\}$ , set  $\sigma'(x)$  to be the  $t$ -tuple whose  $i$ -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  $\Sigma'$  has been obtained from  $\Sigma$  by *reducing*  $b$ .

The second operation corresponds to cutting  $\{a, b\}$ , when  $\{a, b\}$  is a reticulated cherry with reticulation leaf  $b$ . Let  $j \in \{1, 2, \dots, t\}$  be such that  $\sigma_j(a) = 1 = \sigma_j(b)$ , and  $\sigma_j(x) = 0$  for all  $x \in X - \{a, b\}$ , and let  $k \in \{1, 2, \dots, t\}$  be such that  $\sigma_k(b) = 1$  and  $\sigma_k(x) = 0$  for all  $x \in X - b$ . The second operation has two types<sup>2</sup>. First, let  $\Sigma'$  be the set of  $|X|$  ordered pairs obtained from  $\Sigma$  as follows. For all  $x \in X - \{b\}$ , set  $\sigma'(x)$  to be the  $t$ -tuple whose  $i$ -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  $\sigma'(b)$  to the  $t$ -tuple whose  $i$ -th entry is

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

---

<sup>2</sup>In the correspondence of cutting a reticulated cherry  $\{a, b\}$ , the two types depend on whether or not the parent of  $b$  is suppressed when cutting  $\{a, b\}$ .

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

Now let  $\Sigma''$  be the set of  $|X|$  ordered pairs obtained from  $\Sigma$  as follows. For all  $x \in X - \{b\}$ , set  $\sigma''(x)$  to be the  $t$ -tuple whose  $i$ -th entry is

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

and set  $\sigma''(b)$  to be the  $t$ -tuple whose  $i$ -th entry is

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

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

The next lemma is established in [6, Lemma 5.1] for binary phylogenetic networks. The extension to phylogenetic networks in which reticulations have in-degree at least two is straightforward and omitted.

**Lemma 4.3.** *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 a phylogenetic network isomorphic to the phylogenetic network obtained from  $\mathcal{N}$  by reducing  $b$ .*
- (ii) *Suppose that  $\{a, b\}$  is a reticulated cherry of  $\mathcal{N}$  with reticulation leaf  $b$ . Then, up to entries with symbol  $-$ , the set of ordered pairs obtained from  $\Sigma_{\mathcal{N}}$  by*
  - (I) *Type-I cutting  $\{a, b\}$  is the ancestral profile of a phylogenetic network isomorphic to the phylogenetic network  $\mathcal{N}'$  obtained from  $\mathcal{N}$  by cutting  $\{a, b\}$  in which the parent of  $b$  is suppressed, and*
  - (II) *Type-II cutting  $\{a, b\}$  is the ancestral profile of a phylogenetic network isomorphic to the phylogenetic network  $\mathcal{N}$  obtained from  $\mathcal{N}$  by cutting  $\{a, b\}$  in which the parent of  $b$  is not suppressed.*

Let  $\mathcal{N}$  be a phylogenetic network on  $X$  with vertex set  $V$ , and let  $v_1, v_2, \dots, v_t$  be a fixed labelling of the vertices in  $V - X$ . For distinct  $i, j \in \{1, 2, \dots, t\}$ , we say  $v_i$  and  $v_j$  are *clones* if, for all  $x \in X$ , we have  $\sigma_i(x) = \sigma_j(x)$ . Characterising which pairs of vertices in an orchard network are clones is crucial to establishing Theorem 3.2. The next lemma gives this characterisation.

**Lemma 4.4.** *Let  $\mathcal{N}$  be an orchard network on  $X$  with vertex set  $V$ . Let  $v_1, v_2, \dots, v_t$  be a fixed labelling of the vertices of  $V - X$ . Then  $v_i$  and  $v_j$  are clones if and only if one of the following holds:*

- (i)  $v_i$  and  $v_j$  belong to the same sink of  $\mathcal{N}$ ; or
- (ii) exactly one of  $v_i$  and  $v_j$  is a reticulation, say  $v_i$ , and there is a reticulation  $v_k$  in the same sink of  $\mathcal{N}$  as  $v_i$  such that  $(v_k, v_j)$  is a (tree) arc of  $\mathcal{N}$ .

Before establishing Lemma 4.4, we give an illustration of the lemma. Consider the orchard network shown in Fig. 1. Every pair of vertices in  $\{u, v, w\}$  are clones. Vertices  $u$  and  $v$  satisfy (i) of Lemma 4.4, while vertices  $u$  and  $w$  (as well as  $v$  and  $w$ ) satisfy (ii) of Lemma 4.4.

*Proof of Lemma 4.4.* It is easily seen that if  $v_i$  and  $v_j$  are vertices for which either (i) or (ii) holds, then  $v_i$  and  $v_j$  are clones. For the converse, suppose that  $v_i$  and  $v_j$  are clones. The proof of the converse is by induction on the sum of the number  $n$  of leaves and the number  $r$  of reticulations of  $\mathcal{N}$ . If  $n + r = 1$ , then  $n = 1$  and  $r = 0$ , and  $\mathcal{N}$  consists of a single vertex, and so the converse holds. If  $n + r = 2$ , then, as  $\mathcal{N}$  is orchard,  $n = 2$  and  $r = 0$ , and  $\mathcal{N}$  consists of two leaves adjoined to the root of  $\mathcal{N}$ . Again, the converse holds.

Now assume that  $n + r \geq 3$ , so  $n \geq 2$  as  $\mathcal{N}$  is orchard, and that the converse holds for all orchard networks in which the sum of the number of leaves and the number of reticulations is at most  $n + r - 1$ . Since  $\mathcal{N}$  is orchard,  $\mathcal{N}$  has a 2-element subset  $\{a, b\}$  of  $X$  such that  $\{a, b\}$  is either a cherry or a reticulated cherry of  $\mathcal{N}$ . First suppose that  $\{a, b\}$  is a cherry of  $\mathcal{N}$ . Let  $p_a$  denote the common parent of  $a$  and  $b$ . Let  $\mathcal{N}'$  denote the phylogenetic network obtained from  $\mathcal{N}$  by reducing  $b$ . By Proposition 2.1,  $\mathcal{N}'$  is orchard. Note that  $\mathcal{N}'$  has the same number of reticulations as  $\mathcal{N}$  but one less leaf. If  $p_a \notin \{v_i, v_j\}$ , then, as  $v_i$  and  $v_j$  are clones of  $\mathcal{N}$ , it follows by Lemma 4.3(i) that  $v_i$  and  $v_j$  are clones of  $\mathcal{N}'$ . Thus, by induction, either (i) or (ii) holds in  $\mathcal{N}'$ . In turn, this implies that either (i) or (ii) holds in  $\mathcal{N}$ . Hence, without loss of generality, we may assume that  $p_a = v_j$ . Let  $g_a$  be the (unique) parent of  $p_a$  in  $\mathcal{N}$ . If  $g_a$  is a tree vertex, then there is a directed path from  $g_a$  to a leaf  $\ell$  such that  $\ell \notin \{a, b\}$ . Since

$$\sigma_i(a) = \sigma_j(a) = 1,$$

that is, there is a path from  $v_i$  to  $a$  in  $\mathcal{N}$ , it follows that there is a path from  $v_i$  to  $g_a$ , and so  $\sigma_i(\ell) \geq 1$ . But  $\sigma_j(\ell) = 0$ , a contradiction as  $v_i$  and  $v_j$  are clones. Hence  $g_a$  is a reticulation.

If  $v_i$  belongs to the sink  $[g_a]$ , then (ii) holds as  $(g_a, v_j)$  is an arc of  $\mathcal{N}$ . So assume that  $v_i$  does not belong to the sink  $[g_a]$ . Since  $\sigma_i(a) = \sigma_j(a) = 1$ ,

there is a path  $P$  in  $\mathcal{N}$  from  $v_i$  to  $g_a$ . Let  $u$  denote the last tree vertex on  $P$ . Since  $v_i \notin [g_a]$ , such a vertex exists and is the parent of a vertex in  $[g_a]$ . Then, as  $u$  is a tree vertex, either there are at least two paths from  $u$  to  $a$  and so  $\sigma_i(a) \geq 2$ , a contradiction as  $\sigma_j(a) = 1$ , or there is a path from  $u$  to a leaf  $\ell \notin \{a, b\}$ . But then  $\sigma_i(\ell) \geq 1$  and  $\sigma_j(\ell) = 0$ , another contradiction as  $v_i$  and  $v_j$  are clones.

Now suppose that  $\{a, b\}$  is a reticulated cherry of  $\mathcal{N}$ . Without loss of generality, we may assume that  $b$  is the reticulation leaf. Let  $p_a$  and  $p_b$  denote the parents of  $a$  and  $b$ , respectively. Let  $\mathcal{N}'$  be the phylogenetic network obtained from  $\mathcal{N}$  by cutting  $\{a, b\}$ . By Proposition 2.1,  $\mathcal{N}'$  is orchard. Note that  $\mathcal{N}'$  has the same number of leaves as  $\mathcal{N}$  but one less reticulation. If  $\{p_a, p_b\} \cap \{v_i, v_j\}$  is empty, then, as  $v_i$  and  $v_j$  are clones of  $\mathcal{N}$ , it follows by Lemma 4.3(ii) that  $v_i$  and  $v_j$  are clones of  $\mathcal{N}'$ . Thus, by induction, either (i) or (ii) holds in  $\mathcal{N}'$ , and therefore in  $\mathcal{N}$ . If  $\{p_a, p_b\} = \{v_i, v_j\}$ , then either  $\sigma_i(a) = 1$  and  $\sigma_j(a) = 0$ , or  $\sigma_j(a) = 1$  and  $\sigma_i(a) = 0$ , a contradiction as  $v_i$  and  $v_j$  are clones. Thus we may assume that

$$|\{p_a, p_b\} \cap \{v_i, v_j\}| = 1.$$

Without loss of generality, suppose that  $v_j \in \{p_a, p_b\}$ . Say  $v_j = p_a$ . Let  $g_a$  be the (unique) parent of  $p_a$  in  $\mathcal{N}$ . Since  $\sigma_i(a) = \sigma_j(a) = 1$ , it follows that there is a path from  $v_i$  to  $g_a$ . If  $g_a$  is a tree vertex, then there is a directed path from  $g_a$  to a leaf  $\ell$  not using  $(g_a, p_a)$ . Observe that  $\ell \neq a$ . If  $\ell = b$ , then  $\sigma_i(b) \geq 2$ , a contradiction as  $\sigma_j(b) = 1$ . On the other hand, if  $\ell \neq b$ , then  $\sigma_i(\ell) \geq 1$  and  $\sigma_j(\ell) = 0$ , another contradiction. Thus we may assume  $v_j = p_b$ .

If  $v_i$  belongs to the same sink  $[p_b]$ , then (i) holds. So assume  $v_i$  does not belong to  $[p_b]$ . Since  $\sigma_j(a) = 0$ , and  $v_i$  and  $v_j$  are clones, it follows that there is no path from  $v_i$  to  $p_a$ . However, as  $\sigma_j(b) = 1$ , there is a path  $P$  from  $v_i$  to  $p_b$ . Let  $u$  denote the last tree vertex on  $P$ . Since  $v_i \notin [p_b]$  such a vertex exists and is the parent of a vertex in  $[p_b]$ . But then, as  $u$  is a tree vertex, either there are at least two paths from  $u$  to  $b$ , in which case  $\sigma_i(b) \geq 2$ , or there is a path from  $u$  to a leaf  $\ell \neq b$ , in which case  $\sigma_i(\ell) \geq 1$ . Both cases contradict that  $v_i$  and  $v_j$  are clones as  $\sigma_j(b) = 1$  and  $\sigma_j(x) = 0$  for all  $x \in X - \{b\}$ . This completes the proof of the lemma.  $\square$

The next two results are consequences of Lemma 4.4. The first result is immediate.

**Corollary 4.5.** *Let  $\mathcal{N}$  be an orchard network on  $X$  with vertex set  $V$ , and suppose that  $\{v_1, v_2, v_3\}$  is a 3-element subset of  $V - X$ . If  $v_i$  and  $v_j$  are clones for all distinct  $i, j \in \{1, 2, 3\}$ , then  $\mathcal{N}$  has a sink of size at least two, in which case at least two of the vertices in  $\{v_1, v_2, v_3\}$  are in the same sink.*

Next let  $\mathcal{N}$  be a phylogenetic network on  $X$  with vertex set  $V$ , and let  $v_i$  and  $v_j$  be distinct vertices of  $V - X$ . We say  $v_i$  and  $v_j$  are a *maximal* pair of clones if  $v_i$  and  $v_j$  are clones, but there is no vertex  $v_k \in V - (X \cup \{v_i, v_j\})$  such that every two elements in  $\{v_i, v_j, v_k\}$  are clones.

**Corollary 4.6.** *Let  $\mathcal{N}$  and  $\mathcal{N}'$  be orchard networks on  $X$ , and suppose  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ . If  $v_i$  and  $v_j$  are a maximal pair of clones of  $\mathcal{N}$ , then  $v_i$  and  $v_j$  are reticulations of  $\mathcal{N}$  if and only if  $v_i$  and  $v_j$  are reticulations of  $\mathcal{N}'$ .*

*Proof.* Since  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ , the vertices  $v_i$  and  $v_j$  are a maximal pair of clones of  $\mathcal{N}'$ . Thus, to prove the lemma, it suffices to show that if  $v_i$  and  $v_j$  are reticulations of  $\mathcal{N}$ , then  $v_i$  and  $v_j$  are reticulations of  $\mathcal{N}'$ . Suppose that  $v_i$  and  $v_j$  are reticulations of  $\mathcal{N}$ . Then, as  $v_i$  and  $v_j$  are a maximal pair of clones, we may assume that  $(v_i, v_j)$  is a stack arc of  $\mathcal{N}$  and  $v_j$  is the parent of a leaf  $\ell$ . In particular,  $\sigma_i(\ell) = \sigma_j(\ell) = 1$  and  $\sigma_i(x) = \sigma_j(x) = 0$  for all  $x \in X - \{\ell\}$ . Now, if  $v_i$  and  $v_j$  are not reticulations of  $\mathcal{N}'$ , then, without loss of generality, we may assume by Lemma 4.4 that  $(v_i, v_j)$  is an arc of  $\mathcal{N}'$ , in which  $v_i$  is a reticulation and  $v_j$  is a tree vertex. But then either there are at least two paths from  $v_i$  to a leaf or there is a path from  $v_i$  to a leaf that is not  $\ell$ . Both possibilities contradict the assumption that  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ . This completes the proof of the corollary.  $\square$

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

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

be a set of ordered pairs, where  $\Sigma(x)$  is a  $t$ -tuple whose entries are either non-negative integers or  $-$  for all  $x \in X$ . We describe a further operation on  $\Sigma$ . This time the operation corresponds to deleting a stack arc and identifying its end vertices. Let  $j, k$ , and  $l$  be distinct element in  $\{1, 2, \dots, t\}$  such that

$$\sigma_j(x) = \sigma_k(x) = \sigma_l(x)$$

for all  $x \in X$ . Let  $\Sigma'$  be the set of  $|X|$  ordered pairs obtained from  $\Sigma$  as follows. For all  $x \in X$ , set  $\sigma'(x)$  to be the  $t$ -tuple whose  $i$ -th entry is

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

Set  $\Sigma' = \{(x, \sigma'(x)) : x \in X\}$ . We say that  $\Sigma'$  has been obtained from  $\Sigma$  by *identifying  $j$* . The proof of the next lemma is routine and omitted.

**Lemma 4.7.** *Let  $\mathcal{N}$  be a phylogenetic network on  $X$  with vertex set  $V$ , and fix an ordering of  $V - X$ . Suppose that  $(v_j, v_k)$  is a stack arc of  $\mathcal{N}$ . Then, up to entries with symbol  $-$ , the set of ordered pairs obtained from  $\Sigma_{\mathcal{N}}$  by identifying  $j$  is the ancestral profile of a phylogenetic network isomorphic to the phylogenetic network obtained from  $\mathcal{N}$  by deleting  $(v_j, v_k)$ , and identifying  $v_j$  and  $v_k$ .*

**Lemma 4.8.** *Let  $\mathcal{N}$  and  $\mathcal{N}'$  be orchard networks on  $X$ . If  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ , then  $\Sigma_{\text{id}(\mathcal{N})} = \Sigma_{\text{id}(\mathcal{N}'})$ .*

*Proof.* Let  $V$  denote the vertex set of  $\mathcal{N}$ , and suppose that  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ . Let  $v_1, v_2, \dots, v_t$  be a fixed labelling of the vertices of  $\mathcal{N}$  in  $V - X$ . Note that, as  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ , the total number of vertices in  $\mathcal{N}$  and  $\mathcal{N}'$  is  $t + |X|$ . The proof is by induction on the number  $s$  of stack arcs of  $\mathcal{N}$ . If  $s = 0$ , then  $\mathcal{N} = \text{id}(\mathcal{N})$  and so, by Lemma 4.4, if  $v_i$  and  $v_j$  are clones of  $\mathcal{N}$ , then exactly one of  $v_i$  and  $v_j$  is a reticulation, say  $v_i$ , and  $(v_i, v_j)$  is a tree arc of  $\mathcal{N}$ . In particular, all sink classes of  $\mathcal{N}$  have size one. We next show that  $\mathcal{N}'$  has no stack arcs.

If  $\mathcal{N}'$  has a stack arc  $e$ , then there exists either a 3-element subset of  $V - X$  such that every pair of elements are clones or the two end vertices of  $e$  form a maximal pair of clones. Since  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ , it follows by Corollaries 4.5 and 4.6 that  $\mathcal{N}$  has a sink of size two, a contradiction. Thus  $\mathcal{N}'$  has no stack arcs, and so  $\mathcal{N}' = \text{id}(\mathcal{N}')$ . Hence  $\Sigma_{\text{id}(\mathcal{N})} = \Sigma_{\text{id}(\mathcal{N}'})$ .

Now assume that  $s \geq 1$  and that the lemma holds for all pairs of orchard networks on the same leaf sets, where one of the networks has at most  $s - 1$  stack arcs. Since  $s \geq 1$ , there exists a stack arc  $(v_i, v_j)$  of  $\mathcal{N}$ , in which case  $v_i$  and  $v_j$  belong to the same sink and are clones of  $\mathcal{N}$ . Since  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ , it follows by Corollaries 4.5 and 4.6 that  $\mathcal{N}'$  has a pair  $v'_i$  and  $v'_j$  of clones, where  $(v'_i, v'_j)$  is a stack arc of  $\mathcal{N}'$  and, for all  $x \in X$ ,

$$\sigma_i(x) = \sigma_j(x) = \sigma_{i'}(x) = \sigma_{j'}(x).$$

Let  $\mathcal{N}_1$  denote the directed graph obtained from  $\mathcal{N}$  by deleting  $(v_i, v_j)$ , and identifying  $v_i$  and  $v_j$ . By Lemma 4.1,  $\mathcal{N}_1$  is orchard. Similarly, let  $\mathcal{N}'_1$  denote the directed graph obtained from  $\mathcal{N}'$  by deleting  $(v'_i, v'_j)$ , and identifying  $v'_i$  and  $v'_j$ . By Lemma 4.1 again,  $\mathcal{N}'_1$  is orchard. Then, as  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ , we deduce by Lemma 4.7 that  $\Sigma_{\mathcal{N}_1} = \Sigma_{\mathcal{N}'_1}$ . Since the number of stack arcs of  $\mathcal{N}_1$  is  $s - 1$ , it follows by the induction assumption that

$$\Sigma_{\text{id}(\mathcal{N}_1)} = \Sigma_{\text{id}(\mathcal{N}'_1)}.$$

But  $\text{id}(\mathcal{N}) \cong \text{id}(\mathcal{N}_1)$  and  $\text{id}(\mathcal{N}') \cong \text{id}(\mathcal{N}'_1)$ , and so  $\Sigma_{\text{id}(\mathcal{N})} = \Sigma_{\text{id}(\mathcal{N}'})$ , thereby completing the proof of the lemma.  $\square$

*Proof of Theorem 3.2.* Suppose that  $\mathcal{N}$  and  $\mathcal{N}'$  are orchard networks on  $X$  with  $\Sigma_{\mathcal{N}} = \Sigma_{\mathcal{N}'}$ . By Corollary 4.2,  $\text{id}(\mathcal{N})$  and  $\text{id}(\mathcal{N}')$  are orchard networks and so, by Lemma 4.8,  $\Sigma_{\text{id}(\mathcal{N})} = \Sigma_{\text{id}(\mathcal{N}'})$ . Thus, by Theorem 3.1,  $\text{id}(\mathcal{N})$  is isomorphic to  $\text{id}(\mathcal{N}')$ .  $\square$

## 5. CONCLUDING COMMENTS

We end by raising two questions concerning orchard networks that may be interesting for future work (even in the case where such networks are assumed to be binary). The first question is whether or not Theorem 3.2 remains true if one removes the requirement that  $\mathcal{N}'$  is an orchard network. Note that Theorem 3.1 requires only that  $\mathcal{N}$  is an orchard network. A second question is whether orchard networks can be characterised succinctly in terms of forbidden subgraphs. For example, a binary phylogenetic network is tree-child if and only if it has no stack reticulations and no (tree) vertex that is the parent of two distinct reticulations [14]. The class of binary ‘tree-based’ networks have also been characterised in a similar way [20]. Such ‘forbidden subgraph’ characterisations have turned out to be particularly helpful in the study of these phylogenetic networks and we expect the same to apply in the study of orchard networks.

## REFERENCES

- [1] 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.
- [2] M. Bordewich, C. Semple, Determining phylogenetic networks from inter-taxa distances, *Journal of Mathematical Biology* 73 (2016) 283–303.
- [3] G. Cardona, M. Llabrés, F. Rosselló, G. Valiente, A distance metric for a class of tree-sibling phylogenetic networks, *Bioinformatics* 24 (2008) 1481–1488.
- [4] G. Cardona, F. Rosselló, G. Valiente, Comparison of tree-child phylogenetic networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 6 (2009) 552–569.
- [5] W.F. Doolittle, Phylogenetic classification and the universal tree, *Science* 284 (1999) 2124–2128.
- [6] P.L. Erdős, C. Semple, M. Steel, (2019). A class of phylogenetic networks reconstructable from ancestral profiles. *Mathematical Biosciences*, 313: 33–40.
- [7] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2004.
- [8] D.H. Huson, R. Rupp, C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge University Press, 2010.
- [9] R. Janssen, Y. Murakami, On cherry-picking and network containment, [arXiv:1812.08065v2](https://arxiv.org/abs/1812.08065v2) (2020).
- [10] 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.
- [11] E.V. Koonin, The turbulent network dynamics of microbial evolution and the statistical tree of life, *Journal of Molecular Evolution* 80 (2015) 244–250.
- [12] S. Linz, C. Semple, Caterpillars on three and four leaves are sufficient to reconstruct binary normal networks, *Journal of Mathematical Biology* 81 (2020) 961–980.
- [13] F. Pardi, C. Scornavacca, Reconstructible phylogenetic networks: Do not distinguish the indistinguishable, *PLoS Computational Biology* 11 (2015) e1004135.
- [14] C. Semple, Phylogenetic networks with every embedded phylogenetic tree a base tree, *Bulletin of Mathematical Biology* 78 (2016) 132–137.
- [15] C. Semple, M. Steel, *Phylogenetics*, Oxford University Press, Oxford, 2003.



- [16] N.S. Upham, J.A. Esselstyn, W. Jetz. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology* 17 (2019) e3000494.
- [17] S.J. Willson, Reconstruction of certain phylogenetic networks from the genomes at their leaves, *Journal of Theoretical Biology* 252 (2008) 338–349.
- [18] S.J. Willson, Properties of normal phylogenetic networks, *Bulletin of Mathematical Biology* 72 (2010) 340–358.
- [19] M. Worobey, J. Pekar, B.B. Larsen, M.I. Nelson, V. Hill, J.B. Joy, A. Rambaut, M.A. Suchard, J.O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America, *Science* 370 (2020) 564–570.
- [20] L. Zhang, On tree-based phylogenetic networks, *Journal of Computational Biology* 23 (2016) 553–565.

#### APPENDIX: ADJUSTMENTS REQUIRED FOR THE PROOF OF THEOREM 3.1

The following lemma replaces [6, Lemma 3.3], in which the requirement that the grandparents of  $b$  (i.e. parents of the parent of  $b$ ), are tree vertices was omitted. Without this extra constraint, the lemma does not hold; an example to illustrate this the phylogenetic network  $\mathcal{N}$  in Fig. 2 by taking  $a = x_3$  and  $b = x_4$ .

**Lemma 5.1.** *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 and all grandparents of  $b$  are tree vertices if and only if*

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

The proof of Lemma 5.1 follows the same argument as the original statement of the lemma.

In addition to Lemma 5.1, the proof of Theorem 3.1 requires two further lemmas. The first replaces [6, Corollary 4.2] (which is correct as stated) and the second is to connect Lemma 5.1 with stack-free orchard networks.

**Lemma 5.2.** *Let  $\mathcal{N}$  be a stack-free 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 a stack-free orchard network.*

**Lemma 5.3.** *Let  $\mathcal{N}$  be a stack-free orchard network. If  $\{a, b\}$  is a reticulated cherry of  $\mathcal{N}$  in which  $b$  is the reticulation leaf, then the grandparents of  $b$  are tree vertices.*

Using these replacement lemmas, together with [6, Proposition 4.1] and [6, Lemma 5.1] replaced by Proposition 2.1 and Lemma 4.3, respectively, the proof of Theorem 3.1 follows the same argument, *mutatis mutandis*, as [6, Theorem 2.2].

SCHOOL OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY,  
CHRISTCHURCH, NEW ZEALAND

*Email address:* `amb337@uclive.ac.nz`

ALFRÉD RÉNYI INSTITUTE OF MATHEMATICS, HUNGARIAN ACADEMY OF SCIENCES,  
BUDAPEST, HUNGARY

*Email address:* `erdos.peter@renyi.hu`

SCHOOL OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY,  
CHRISTCHURCH, NEW ZEALAND

*Email address:* `charles.semple@canterbury.ac.nz`

SCHOOL OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY,  
CHRISTCHURCH, NEW ZEALAND

*Email address:* `mike.steel@canterbury.ac.nz`