HYPERCUBES AND HAMILTON CYCLES OF DISPLAY SETS OF ROOTED PHYLOGENETIC NETWORKS

JANOSCH DÖCKER, SIMONE LINZ, AND CHARLES SEMPLE

Abstract. In the context of reconstructing phylogenetic networks from a collection of phylogenetic trees, several characterisations and subsequently algorithms have been established to reconstruct a phylogenetic network that collectively embeds all trees in the input in some minimum way. For many instances however, the resulting network also embeds additional phylogenetic trees that are not part of the input. However, little is known about these inferred trees. In this paper, we explore the relationships among all phylogenetic trees that are embedded in a given phylogenetic network. First, we investigate some combinatorial properties of the collection $\mathcal{P}$ of all rooted binary phylogenetic trees that are embedded in a rooted binary phylogenetic network $\mathcal{N}$. To this end, we associated a particular graph G, which we call rSPR graph, with the elements in $\mathcal{P}$ and show that, if $|\mathcal{P}| = 2^k$, where k is the number of vertices with in-degree two in $\mathcal{N}$, then G has a Hamilton cycle. Second, by exploiting rSPR graphs and properties of hypercubes, we turn to the well-studied class of rooted binary level-1 networks and give necessary and sufficient conditions for when a set of rooted binary phylogenetic trees can be embedded in a level-1 network without inferring any additional trees. Lastly, we show how these conditions translate into a polynomial-time algorithm to reconstruct such a network if it exists.

1. Introduction

Phylogenetic networks, which are used to represent treelike and non-treelike ancestral relationships between a set of present-day species, generalise phylogenetic trees by allowing for cycles in the underlying graph. In the case of rooted phylogenetic networks, vertices with in-degree at least two, called *reticulations*, represent non-treelike events such as hybridisation or lateral gene transfer that cannot be represented by a single rooted phylogenetic tree, whereas vertices with in-degree one represent treelike speciation events. Software to reconstruct phylogenetic networks frequently uses molecular sequence data or a collection of conflicting phylogenetic trees as input [2, 5, 18, 21, 31]. If a phylogenetic network $\mathcal{N}$ is reconstructed from a set of phylogenetic trees such that $\mathcal{N}$ embeds each tree of the input, then it may be necessary for $\mathcal{N}$ to not only embed the input, but also a number of additional phylogenetic trees. For example, referring to the two rooted phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ that are shown in Figure 1, every rooted phylogenetic network that embeds

Date: August 1, 2023.

Key words and phrases. Display set, Gray code, hypercube, Hamilton cycle, phylogenetic network, subtree prune and regraft.

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

$\mathcal{T}_1$ and $\mathcal{T}_2$ has at least two reticulations and embeds at least one tree that is distinct from $\mathcal{T}_1$ and $\mathcal{T}_2$. In Figure 1 and, in fact, in all figures of this paper, arcs of rooted phylogenetic networks are directed down the page and arrowheads are omitted. Clearly, if the input consists of all rooted binary phylogenetic trees on a fixed leaf set, then no rooted phylogenetic network that embeds each tree in the input infers any additional tree. Such networks are called universal networks and exist, for example, for the class of tree-based networks [11, 35]. However, for more structurally restricted network classes such as level-1 or tree-child networks whose number of reticulations is bounded linearly in the number of leaves [25], no universal network exist. Moreover, in practice, one is often interested in a subset of all rooted binary phylogenetic trees on a fixed leaf set. It is consequently more realistic to ask which collections of phylogenetic trees can be embedded in a network without inferring any additional tree? In this paper, we approach this question from two angles.

First, we investigate the relationships among all rooted binary phylogenetic trees that are embedded in a given rooted binary phylogenetic network $\mathcal{N}$. We refer to the set of all such trees as the *display set* (formally defined in the next section) of $\mathcal{N}$. It is well-known that the size of the display set of a rooted binary phylogenetic network with exactly k reticulations is at most 2^k and that this bound is sharp, such as for normal networks [20, 33]. However, not all rooted phylogenetic networks with k reticulations have a display set of size 2^k. An example is shown in Figure 1. To explore the relationships among the elements of a display set $\mathcal{P}$ of a rooted phylogenetic network, we associate an undirected graph with $\mathcal{P}$. Referring to this graph as the *rSPR graph* of $\mathcal{P}$, its vertex set is $\mathcal{P}$ and two vertices are connected by an edge precisely if they are one rooted subtree prune and regraft (rSPR) operation [17] apart. The rSPR operation induces a metric on the space of all rooted binary phylogenetic trees with a fixed leaf set. It is used to compare pairs of phylogenetic trees and to search for an optimum tree in tree space [1, 7, 12, 28]. We show that the rSPR graph of a display set of a rooted binary phylogenetic network is always connected. In turn this implies that, if the rSPR graph of an arbitrary collection of rooted binary phylogenetic trees is not connected, then any rooted phylogenetic network that embeds each tree in the collection infers additional phylogenetic trees. Moreover, if $\mathcal{P}$ is the display set of a rooted binary phylogenetic network with k reticulations and has size 2^k, then its rSPR graph G has a Hamilton cycle. Hence, in the spirit of [13], it is possible to systematically traverse G, thereby visiting each element in $\mathcal{P}$ exactly once.

Second, we turn to level-1 networks that are phylogenetic networks whose underlying cycles do not intersect. We characterise when a set $\mathcal{P}$ of rooted binary phylogenetic trees is the display set of a rooted binary level-1 network, in which case it is possible to reconstruct such a network that embeds each tree in $\mathcal{P}$ and does not infer any additional tree. Our characterisation again employs rSPR graphs and establishes necessary and sufficient conditions for $\mathcal{P}$ to be the display set of a rooted binary level-1 network. To provide a flavour of the characterisation, a necessary condition is that the rSPR graph of $\mathcal{P}$ is isomorphic to the k-dimensional hypercube for some non-negative integer k. Although hypercubes have previously been used in research on phylogenetic trees (e.g. in the context of Buneman graphs and maximum parsimony [29, Section 5.5] as well as in developing a lower bound on the minimum number of reticulations needed to explain a collection of conflicting

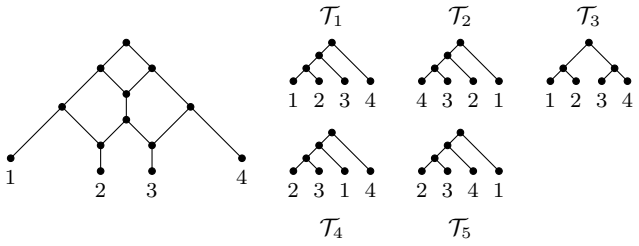

FIGURE 1. A phylogenetic network with three reticulations whose display set consists of the five phylogenetic trees shown on the right.

phylogenetic trees [34]), their application to studying the display set of a phylogenetic network is new to this paper. Subsequent to the characterisation, we present a polynomial-time algorithm to decide whether or not $\mathcal{P}$ is the display set of a rooted binary level-1 network and, if so, to reconstruct such a network. This algorithm can also be easily modified to enumerate all rooted binary level-1 network whose display set is $\mathcal{P}$. Relatedly, there exists earlier work [19, 30] on reconstructing a (single) rooted binary level-1 network whose number of reticulations is minimised over all rooted binary level-1 networks whose display set is a superset of $\mathcal{P}$, where the focus in [19] is on the case $|\mathcal{P}| = 2$. Although these earlier algorithms can potentially be exploited further to also decide if there exists a rooted binary level-1 network whose display set is $\mathcal{P}$, the purpose of the present paper is to demonstrate the applicability of rSPR graphs to studying display sets of rooted binary phylogenetic networks. Indeed, we expect that rSPR graphs will be used in the future to investigate related questions that go beyond level-1 networks.

The remainder of the paper is organised as follows. Section 2 provides definitions and terminology that is used in the subsequent sections. We then establish basic properties of rSPR graphs in Section 3 and, in particular, hamiltonicity of any rSPR graph with 2^k vertices whose underlying collection of rooted binary phylogenetic trees is the display set of a rooted phylogenetic network with k reticulations in Section 4. In Section 5, we use rSPR graphs to characterise when a set $\mathcal{P}$ of rooted binary phylogenetic trees is the display set of a rooted level-1 network. Lastly, in Section 6, we show that it takes polynomial time to decide if the necessary and sufficient conditions established in Section 5 are satisfied and, if so, to reconstruct a level-1 network whose display set is $\mathcal{P}$.

2. Preliminaries

This section gives definitions and terminology on phylogenetic trees and networks as well as on hypercubes that is used in the following sections. Throughout this paper, X denotes a non-empty finite set.

Phylogenetic networks. A *rooted binary phylogenetic network* $\mathcal{N}$ *on* X is a rooted acyclic directed graph with no parallel arcs or loops that satisfies the following three properties:

(i) the (unique) root has in-degree zero and out-degree two;
(ii) a vertex of 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 $|X| = 1$, then we additionally allow $\mathcal{N}$ to consist of the single vertex in X. Let v be a vertex of $\mathcal{N}$. If v has out-degree zero, then v is called a *leaf*, and X is referred to as the *leaf set* of $\mathcal{N}$. Furthermore, if v has in-degree one and out-degree two, v is referred to as a *tree vertex*, and if it has in-degree two and out-degree one, v is referred to as a *reticulation*. For an arc (u, v) of $\mathcal{N}$, we say that u is a *parent* of v and, equivalently, that v is a *child* of u. Also, if v is a reticulation, then (u, v) is called a *reticulation arc*. Lastly, if u is a vertex of $\mathcal{N}$, then $C_{\mathcal{N}}(u)$ denotes the subset of X whose elements x have the property that there is a directed path from u to x in $\mathcal{N}$. Such a subset of X is referred to as a *cluster* of $\mathcal{N}$ and we denote the set of clusters of $\mathcal{N}$ by $C(\mathcal{N})$. Note that each element in X is a cluster of $\mathcal{N}$. If there is no ambiguity, we sometimes refer to $C_{\mathcal{N}}(u)$ as $C(u)$.

We next consider different classes of phylogenetic networks that are well known in the literature. For an excellent overview on the different classes, we refer the interested reader to Kong et al. [22] (and references therein). Let $\mathcal{N}$ be a rooted binary phylogenetic network on X, and let $e = (u, v)$ be a reticulation arc of $\mathcal{N}$. Then e is called a *shortcut* of $\mathcal{N}$ if there exists a directed path from u to v that avoids e. Now, if each non-leaf vertex of $\mathcal{N}$ has a child that is a tree vertex or leaf, then $\mathcal{N}$ is a *tree-child* network. Moreover, if $\mathcal{N}$ is tree-child and does not contain a shortcut, then $\mathcal{N}$ is a *normal* network. With a view towards the underlying cycles of $\mathcal{N}$, we say that $\mathcal{N}$ is a *level*-1 network if no two underlying cycles of $\mathcal{N}$ have a common vertex. Lastly, a *rooted binary phylogenetic X-tree* is a rooted binary phylogenetic network on X with no reticulations. In what follows, we will refer to a rooted binary phylogenetic network and a rooted binary phylogenetic tree as a *phylogenetic network* and a *phylogenetic tree*, respectively, since all such networks and trees in this paper are rooted and binary.

We next define three types of subtrees of a phylogenetic X-tree $\mathcal{T}$. Let V be a subset of the vertices of $\mathcal{T}$. First, we write $\mathcal{T}(V)$ to denote the minimal rooted subtree of $\mathcal{T}$ that connects all elements in V. Second, the *restriction of $\mathcal{T}$ to V*, denoted by $\mathcal{T}|V$, is the rooted phylogenetic tree obtained from $\mathcal{T}(V)$ by suppressing each vertex with in-degree one and out-degree one. In what follows, V is typically a subset of X. Third, a subtree of $\mathcal{T}$ is called *pendant* if it can be detached from $\mathcal{T}$ by deleting a single arc.

Now, let $\mathcal{N}$ be a phylogenetic network on X with k reticulations, and let $\mathcal{T}$ be a phylogenetic X-tree. We say that $\mathcal{T}$ is *displayed* by $\mathcal{N}$ if there exists a subgraph of $\mathcal{N}$ that is a subdivision of $\mathcal{T}$. For a set $\mathcal{P}$ of phylogenetic X-trees, we say that $\mathcal{N}$ *displays* $\mathcal{P}$ if each element in $\mathcal{P}$ is displayed by $\mathcal{N}$. Moreover the set of all phylogenetic X-trees that are displayed by $\mathcal{N}$ is called the *display set* of $\mathcal{N}$ and denoted by $T(\mathcal{N})$. Since the in-degree of each reticulation is two, it immediately follows that $|T(\mathcal{N})| \leq 2^k$. Furthermore, we say that $T(\mathcal{N})$ is *maximum* if $|T(\mathcal{N})| = 2^k$. The class of phylogenetic networks whose display set is maximum strictly

contains the class of normal networks [20, 33]. However, not every phylogenetic network $\mathcal{N}$ with k reticulations has a display set of size 2^k. An example is shown in Figure 1. Moreover, a sufficient but not necessary condition for a phylogenetic network with k reticulations to display strictly less than 2^k phylogenetic trees is the existence of an arc that is incident with two reticulations.

Let $\mathcal{N}$ and $\mathcal{N}'$ be two phylogenetic networks on X with vertex and arc sets V and E, and V' and E', respectively. Then $\mathcal{N}$ and $\mathcal{N}'$ are *isomorphic* if there is a bijection $\psi : V \to V'$ such that $\psi(x) = x$ for all $x \in X$, and $(u, v) \in E$ if and only if $(\psi(u), \psi(v)) \in E'$ for all $u, v \in V$. If $\mathcal{N}$ and $\mathcal{N}'$ are isomorphic, we write $\mathcal{N} \cong \mathcal{N}'$ and, otherwise, we write $\mathcal{N} \ncong \mathcal{N}'$.

rSPR and agreement forests. Let $\mathcal{T}$ be a phylogenetic X-tree. For the purposes of the upcoming definitions, we view the root of $\mathcal{T}$ as a vertex ρ adjoined to the original root by a pendant arc. Furthermore, we regard ρ as part of the label set of $\mathcal{T}$, that is, $\mathcal{L}(\mathcal{T}) = X \cup \{\rho\}$. Let $e = (u, v)$ be an arc of $\mathcal{T}$ that is not incident with ρ. Let $\mathcal{T}'$ be a phylogenetic X-tree obtained from $\mathcal{T}$ by deleting e and reattaching the resulting rooted subtree that contains v via a new arc f in the following way: Subdivide an arc of the component that contains ρ with a new vertex u', join u' and v with f, and suppress u. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *rooted subtree prune and regraft* (rSPR) operation. The *rSPR distance* between any two phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, denoted by $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$, is the minimum number of rSPR operations that transform $\mathcal{T}$ into $\mathcal{T}'$. It is well known that $\mathcal{T}'$ can always be obtained from $\mathcal{T}$ by a sequence of single rSPR operations and, so, the distance is well defined.

Now, let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees. An *agreement forest* $\mathcal{F} = \{\mathcal{L}_\rho, \mathcal{L}_1, \ldots, \mathcal{L}_k\}$ for $\mathcal{T}$ and $\mathcal{T}'$ is a partition of $X \cup \{\rho\}$ such that $\rho \in \mathcal{L}_\rho$ and the following two properties are satisfied.

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

An agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is a *maximum agreement forest* if it has the smallest number of elements amongst all agreement forests for $\mathcal{T}$ and $\mathcal{T}'$. The following theorem links the rSPR distance between two phylogenetic trees and the size of a maximum agreement forest for the same trees.

Theorem 2.1. [7] *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted phylogenetic X-trees, and let $\mathcal{F}$ be a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = |\mathcal{F}| - 1$.*

Let $\mathcal{P}$ be a set of phylogenetic X-trees. The *rSPR graph of $\mathcal{P}$* is the graph $G = (V, E)$ with $V = \mathcal{P}$ and for which $\{\mathcal{T}, \mathcal{T}'\}$ is an edge in E precisely if $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$. Let $\mathcal{N}$ be a phylogenetic network. For ease of reading, we often refer to the rSPR graph of $T(\mathcal{N})$ as the *rSPR graph of $\mathcal{N}$*. To illustrate, Figure 2 shows two phylogenetic networks with their rSPR graphs.

Gray codes and hypercubes. Let k be a non-negative integer, and let $n = 2^k$. We refer to a string s as a *k-bit string* if s has length k and each bit of s is either

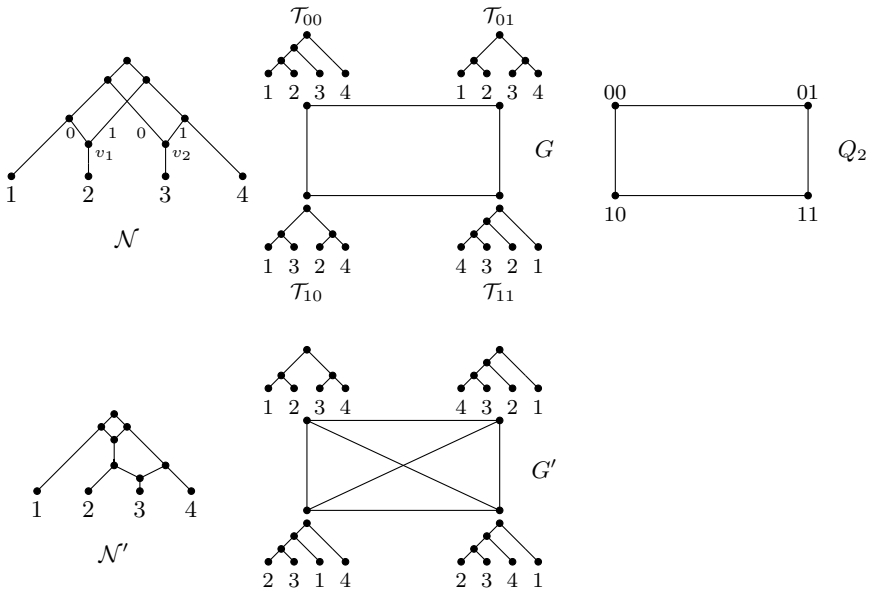

FIGURE 2. Two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ with their rSPR graphs G and G', respectively. For each 2-bit string s, $\mathcal{T}_s$ denotes the phylogenetic tree in $T(\mathcal{N})$ that is encoded by s under the ordered binary assignment for $\mathcal{N}$ as indicated by v_1, v_2, and the assignment of 0 or 1 to each reticulation arc of $\mathcal{N}$.

0 or 1. For two k-bit strings s and s', we denote the Hamming distance between s and s' by $d(s, s')$. Furthermore, an ordering $(s_1, s_2, \ldots, s_n)$ on all k-bit strings is called a *(cyclic) Gray code* if $d(s_1, s_n) = 1$ and, for each $j \in \{1, 2, \ldots, n-1\}$, $d(s_j, s_{j+1}) = 1$ [14]. It is well known that such an ordering on $(s_1, s_2, \ldots, s_n)$ exists (see, for example, [15, 26]).

Let k be a non-negative integer, and let B be the set of all k-bit strings. If $k = 0$, then the only element in B is the empty string. The *k-dimensional hypercube* Q_k is the undirected graph whose vertex set is B and for which $\{s, s'\}$ is an edge in Q_k precisely if $d(s, s') = 1$. Observe that the number of edges in Q_k is $2^{k-1}k$. Moreover, the edge set of Q_k can naturally be partitioned into k sets $E_1, E_2, \ldots, E_k$ such that, for each $i \in \{1, 2, \ldots, k\}$, we have $|E_i| = 2^{k-1}$ and each edge $\{s, s'\}$ of Q_k is an element of E_i if and only if the i-th bit of s and the i-th bit of s' are not the same. We refer to E_i as the i-th *bit edge subset* of Q_k. By way of example, Q_2 is shown in Figure 2, where E_1 contains the two vertical edges and E_2 contains the two horizontal edges.

We end this section with a well-known theorem whose proof is straightforward [26], and that establishes an equivalence between finding a Gray code for all k-bit strings and finding a Hamilton cycle of Q_k.

Theorem 2.2. *Let k be an integer with $k \geq 2$, and let $n = 2^k$. Furthermore, let $C = (s_1, s_2, \ldots, s_n)$ be an ordering on all k-bit strings. Then C is a Gray code if*

and only if

$$\{\{s_1, s_2\}, \{s_2, s_3\}, \ldots, \{s_{n-1}, s_n\}, \{s_n, s_1\}\}$$

is the edge set of a Hamilton cycle of Q_k.

3. Properties of rSPR Graphs

Let $\mathcal{N}$ be a phylogenetic network, and let $R = \{v_1, v_2, \ldots, v_k\}$ be the set of reticulations in $\mathcal{N}$. For each $i \in \{1, 2, \ldots, k\}$, let u_i and u_i' be the two parents of v_i in $\mathcal{N}$. Furthermore, let

$$\phi : \{(u_i, v_i), (u_i', v_i) : i \in \{1, 2, \ldots, k\}\} \to \{0, 1\}$$

be a map that assigns either 0 or 1 to each reticulation arc such that

$$\{\phi((u_i, v_i)), \phi((u_i', v_i))\} = \{0, 1\}$$

for each $v_i \in R$. We refer to ϕ as a *binary assignment* for $\mathcal{N}$. Moreover, (u_i, v_i) is called the 1-*arc of* v_i *under* ϕ if $\phi(u_i) = 1$ and, otherwise, (u_i, v_i) is called the 0-*arc of* v_i *under* ϕ. This definition extends in the obvious way to (u_i', v_i).

Now, let $\mathcal{N}$ be a phylogenetic network on X, and let ϕ be a binary assignment for $\mathcal{N}$. Let R be the set of reticulations in $\mathcal{N}$, and let $|R| = k$. Fix an ordering on the elements in R, say $(v_1, v_2, \ldots, v_k)$. Let s be a k-bit string, and let S be the directed spanning tree of $\mathcal{N}$ such that, for each $i \in \{1, 2, \ldots, k\}$, S uses the 1-arc of v_i under ϕ if the i-th bit of s is 1 and S uses the 0-arc of v_i under ϕ if the i-th bit of s is 0. Furthermore, let $\mathcal{T}_s$ denote the phylogenetic X-tree that is obtained from S by repeatedly suppressing vertices of in-degree one and out-degree one, deleting vertices with out-degree zero that are not in X, and deleting vertices with in-degree zero and out-degree one. Note that the last operation of deleting a vertex with in-degree zero and out-degree one is, for example, necessary for $\mathcal{T}_s$ to be a phylogenetic tree if $\mathcal{N}$ has an underlying 3-cycle that contains the root and S contains the unique reticulation arc of the 3-cycle that is not incident with the root. We say that s *encodes* $\mathcal{T}_s$ *under* ϕ. By construction, $\mathcal{T}_s \in T(\mathcal{N})$. Each k-bit string encodes a unique element in $T(\mathcal{N})$ under ϕ. Moreover two distinct k-bit strings may encode the same element in $\mathcal{T}(\mathcal{N})$ under ϕ. To illustrate, Figure 2 shows, for each 2-bit string s, the phylogenetic tree $\mathcal{T}_s$ that is encoded under the binary assignment as indicated for the phylogenetic network $\mathcal{N}$ shown in the same figure.

Notational remark. Let $\mathcal{N}$ be a phylogenetic network. Throughout this section and the next, we denote a binary assignment ϕ of $\mathcal{N}$ and a fixed ordering $(v_1, v_2, \ldots, v_k)$ on the reticulations of $\mathcal{N}$ with $k \geq 0$ by

$$(\mathcal{N}, \phi, (v_1, v_2, \ldots, v_k)).$$

Furthermore, we refer to $(\mathcal{N}, \phi, (v_1, v_2, \ldots, v_k))$ as an *ordered binary assignment* of $\mathcal{N}$.

Lemma 3.1. *Let* $\mathcal{N}$ *be a phylogenetic network on* X*, and let* s *and* s' *be two* k*-bit strings such that* $d(s, s') = 1$*. Furthermore, let* $\mathcal{T}_s$ *and* $\mathcal{T}_{s'}$ *be the two phylogenetic* X*-trees that are encoded by* s *and* s' *under* ϕ*, respectively. Then* $d_{\mathrm{rSPR}}(\mathcal{T}_s, \mathcal{T}_{s'}) \leq 1$.

Proof. Let k be the number of reticulations in $\mathcal{N}$. Throughout this proof, let $(\mathcal{N}, \phi, (v_1, v_2, \ldots, v_k))$ be an ordered binary assignment of $\mathcal{N}$. If $\mathcal{T}_s \cong \mathcal{T}_{s'}$, then the result clearly follows. We may therefore assume that $\mathcal{T}_s \ncong \mathcal{T}_{s'}$. Let S (resp. S') be the directed spanning tree of $\mathcal{N}$ such that, for each $i \in \{1, 2, \ldots, k\}$, S (resp. S') uses the 1-arc of v_i under ϕ if the i-th bit of s (resp. s') is 1 and, otherwise, S (resp. S') uses the 0-arc of v_i under ϕ. As $d(s, s') = 1$, there exists exactly one reticulation v_j in $\mathcal{N}$ such that one of S and S' uses the 1-arc of v_j under ϕ while the other uses the 0-arc of v_j under ϕ. Now, obtain a directed acyclic graph $\mathcal{N}'$ from $\mathcal{N}$ by deleting each arc that is directed into a reticulation and not used by S or S', and subsequently, applying any of the following three operations until no further operation is possible.

(i) Suppress a vertex of in-degree one and out-degree one.
(ii) Delete a vertex with out-degree zero that is not in X.
(iii) Delete a vertex of in-degree zero and out-degree one.

By construction, v_j is the only vertex of $\mathcal{N}'$ with in-degree 2. Moreover, since $\mathcal{T}_s \ncong \mathcal{T}_{s'}$, the two arcs that are directed into v_j are not in parallel. Hence $\mathcal{N}'$ is a phylogenetic network. Furthermore, as $\mathcal{T}_s, \mathcal{T}_{s'} \in T(\mathcal{N})$, we also have $\mathcal{T}_s, \mathcal{T}_{s'} \in T(\mathcal{N}')$. Since $\mathcal{T}_s \ncong \mathcal{T}_{s'}$ and, consequently, each phylogenetic network that displays $\mathcal{T}_s$ and $\mathcal{T}_{s'}$ has at least one reticulation, it now follows from [4, Proposition 2] that $d_{\mathrm{rSPR}}(\mathcal{T}_s, \mathcal{T}_{s'}) = 1$. This completes the proof of the lemma. □

The next lemma shows that the rSPR graph of a phylogenetic network is always connected.

Lemma 3.2. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. If there exists a phylogenetic network $\mathcal{N}$ with $T(\mathcal{N}) = \mathcal{P}$, then the rSPR graph of $\mathcal{P}$ is connected.*

Proof. Suppose that $\mathcal{P}$ is the display set of a phylogenetic network $\mathcal{N}$ on X that has k reticulations. Let $(\mathcal{N}, \phi, (v_1, v_2, \ldots, v_k))$ be an ordered binary assignment of $\mathcal{N}$. Furthermore, let $n = 2^k$, and let B be the set of all k-bit strings. Consider an ordering, say $(s_1, s_2, \ldots, s_n)$, on the elements in B that is a Gray code. Now, for each $j \in \{1, 2, \ldots, n\}$, let $\mathcal{T}_j$ be the phylogenetic X-tree that is encoded by s_j under ϕ, and let S_j be the directed spanning tree of $\mathcal{N}$ that, for each $i \in \{1, 2, \ldots, k\}$, uses the 1-arc of v_i under ϕ if the i-th bit of s_j is 1 and, otherwise, S_j uses the 0-arc of v_i under ϕ. Note that we may have $\mathcal{T}_j = \mathcal{T}_{j'}$ for two distinct elements $j, j' \in \{1, 2, \ldots, n\}$. Let G be the undirected graph without loops whose vertex set is $\{S_1, S_2, \ldots, S_n\}$ and for which $\{S_j, S_{j'}\}$ is an edge precisely if $d_{\mathrm{rSPR}}(\mathcal{T}_j, \mathcal{T}_{j'}) \leq 1$. By Lemma 3.1, it follows that each of

$$\{S_1, S_2\}, \{S_2, S_3\}, \ldots, \{S_{n-1}, S_n\}$$

is an edge in G and, hence, G is connected. If $|T(\mathcal{N})| = n$, then, as the elements $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n$ are pairwise distinct, it is straightforward to check that G is isomorphic to the rSPR graph of $\mathcal{N}$. Assume that $\mathcal{T}_j \cong \mathcal{T}_{j'}$ for two distinct elements $j, j' \in \{1, 2, \ldots, n\}$. Since $d_{\mathrm{rSPR}}(\mathcal{T}_j, \mathcal{T}_{j'}) = 0$, the edge $\{S_j, S_{j'}\}$ exists in G. Moreover, $\{S_j, S_l\}$ is an edge in G if and only if $\{S_{j'}, S_l\}$ is an edge in G with $l \in \{1, 2, \ldots, n\}$. It now follows that the undirected graph obtained from G by deleting $S_{j'}$ is connected. By construction, repeating this vertex deletion operation

until there exists no further pair of vertices S_j and $S_{j'}$ with $\mathcal{T}_j \cong \mathcal{T}_{j'}$ results in a connected graph that is isomorphic to the rSPR graph of $\mathcal{N}$. □

Although it is well-known that each set $\mathcal{P}$ of phylogenetic trees can be displayed by some phylogenetic network $\mathcal{N}$ such that $\mathcal{P} \subseteq T(\mathcal{N})$, it is not always possible to find a phylogenetic network with display set $\mathcal{P}$. In this case, each phylogenetic network that displays $\mathcal{P}$ infers additional phylogenetic trees that are not contained in $\mathcal{P}$. It is therefore of interest to characterise sets of phylogenetic trees that are equal to the display set of some phylogenetic network. The next corollary, which follows from the contrapositive of Lemma 3.2, makes a first step in this direction and gives a sufficient condition for when there exists no phylogenetic network whose display set is equal to a given set of phylogenetic trees.

Corollary 3.3. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. If the rSPR graph for $\mathcal{P}$ is not connected, then there exists no phylogenetic network $\mathcal{N}$ on X such that $T(\mathcal{N}) = \mathcal{P}$.*

In addition to the last corollary that does not impose any restrictions on the phylogenetic networks under consideration, Section 5 establishes necessary and sufficient conditions for when a set of phylogenetic trees is the display set of a level-1 network.

Now, let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees. Accounting for ρ in the definition of an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, there are $2|X| - 1$ forests of size two that can be obtained from $\mathcal{T}$ by deleting a single arc. Hence, it can be checked in polynomial time if $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$ because, in this case, one of the $2|X| - 1$ forests is guaranteed to be a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. In turn, for an arbitrary-sized set $\mathcal{P}$ of phylogenetic trees, it can be checked in time that is polynomial in $|\mathcal{P}|$ and $|X|$ if the rSPR graph of $\mathcal{P}$ is connected. Note that the converse of Corollary 3.3 is not true. For example, it is straightforward to check that each phylogenetic network that displays the three phylogenetic trees $\mathcal{T}_{01}$, $\mathcal{T}_{10}$, and $\mathcal{T}_{00}$ that are shown in Figure 2 has at least two reticulations and displays strictly more than three phylogenetic trees.

4. Phylogenetic Networks with a Maximum Display Set

In this section, we consider phylogenetic networks $\mathcal{N}$ that have the property $|T(\mathcal{N})| = 2^k$, where k is the number of reticulations of $\mathcal{N}$. As noted earlier the well-studied class of normal networks has this property [20, 33]. Moreover, as we will make more precise in the next section, if we ignore the trivial reticulations of a level-1 network $\mathcal{N}$, then $\mathcal{N}$ also has this property.

Theorem 4.1. *Let $\mathcal{N}$ be a phylogenetic network on X with k reticulations, and let G be the rSPR graph of $\mathcal{N}$. If $|T(\mathcal{N})| = 2^k$, then the k-dimensional hypercube Q_k is a spanning subgraph of G. In particular, G has a Hamilton cycle if $k \geq 2$.*

Proof. Let $(\mathcal{N}, \phi, (v_1, v_2, \ldots, v_k))$ of $\mathcal{N}$ be an ordered binary assignment of $\mathcal{N}$, and let $n = 2^k$. Let B be the set of all k-bit strings, and let $(s_1, s_2, \ldots, s_n)$ be an ordering on the elements in B that is a Gray code. Now consider Q_k. By Theorem 2.2,

$\{\{s_1, s_2\}, \{s_2, s_3\}, \ldots, \{s_{n-1}, s_n\}, \{s_n, s_1\}\}$ is the edge set of a Hamilton cycle in Q_k.

We complete the proof by showing that Q_k is a spanning subgraph of G. Let

$$\psi : \{s_1, s_2, \ldots, s_n\} \to T(\mathcal{N})$$

be a map that assigns each s_j with $j \in \{1, 2, \ldots, n\}$, to the phylogenetic tree that is encoded by s_j under ϕ. As $T(\mathcal{N})$ is maximum, ψ is a bijection. Consider an edge $\{s_j, s_{j'}\}$ in Q_k. As $d(s_j, s_{j'}) = 1$, it follows from Lemma 3.1 that $d_{\mathrm{rSPR}}(\psi(s_j), \psi(s_{j'})) \leq 1$. In fact, again as $T(\mathcal{N})$ is maximum, we have

$$d_{\mathrm{rSPR}}(\psi(s_j), \psi(s_{j'})) = 1.$$

Thus, if $\{s_j, s_{j'}\}$ is an edge in Q_k, then $\{\psi(s_j), \psi(s_{j'})\}$ is an edge in G. It now follows that, as Q_k has a Hamilton cycle, so does G, thereby establishing the theorem. □

Referring back to Figure 2, observe that each of the two phylogenetic networks $\mathcal{N}$ and $\mathcal{N}'$ that is shown in this figure is normal and has two reticulations. Furthermore the rSPR graph G of $\mathcal{N}$ is Q_2, and the rSPR graph of $\mathcal{N}'$ is the complete graph on four vertices, which can be obtained from Q_2 by adding two additional edges.

5. Characterising Display Sets of Level-1 Networks

In this section, we characterise when a set $\mathcal{P}$ of phylogenetic trees is the display set of a level-1 network. This characterisation is phrased in terms of the rSPR graph of $\mathcal{P}$. Unlike the previous sections that gave explicit binary assignments as well as a mapping from the set of all bit strings of a given length to a collection of phylogenetic trees, for ease of reading, bijections between vertices of a hypercube and vertices of an rSPR graph are implicit in this and the next section. We start by giving some further definitions.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees each with root ρ, and suppose that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$. We refer to a subset X' of X as a *moving subtree* for $\mathcal{T}$ and $\mathcal{T}'$ if the bipartition

$$\{(X \cup \{\rho\}) - X', X'\}$$

of $X \cup \{\rho\}$ is a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Intuitively, if X' is a moving subtree for $\mathcal{T}$ and $\mathcal{T}'$, then $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by pruning and regrafting the pendant subtree $\mathcal{T}|X'$. Note that $\mathcal{T}$ and $\mathcal{T}'$ may not have a unique moving subtree. If X' is a moving subtree for $\mathcal{T}$ and $\mathcal{T}'$, we associate this move with the ordered pair (X', Y'), where Y' is the minimal cluster in $C(\mathcal{T}) \cap C(\mathcal{T}')$ properly containing X'.

Let $\mathcal{P}$ be a set of phylogenetic X-trees such that $|\mathcal{P}| = 2^k$ for some non-negative integer k, and let G be the rSPR graph of $\mathcal{P}$. Suppose that there is a (graph) isomorphism from G to Q_k. Under this isomorphism, for all $i \in \{1, 2, \ldots, k\}$, let E_i denote the subset of edges of G corresponding to the i-th bit edge subset of Q_k throughout the remainder of the paper. Now label each edge e of G with the ordered pair that is associated with a moving subtree for the end vertices of e, and

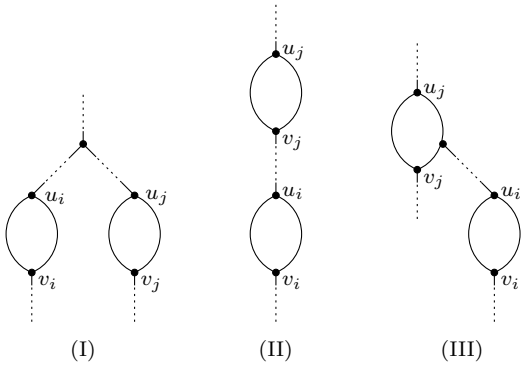

FIGURE 3. The three ways in which the clusters of two reticulations v_i and v_j and the clusters of their two respective source vertices u_i and u_j of a level-1 network can interact. Note that the ordered pairs $(C(v_i), C(u_i)) = (X_i, Y_i)$ and $(C(v_j), C(u_j)) = (X_j, Y_j)$ satisfy (I) in the definition of the nested subtree property for the level-1 network on the left-hand side, and (II) and (III) of the same definition for the level-1 network in the middle and right-hand side, respectively. Each solid arc of an underlying cycle indicates a directed path of arbitrary length.

suppose that, for all $i \in \{1, 2, \ldots, k\}$, this labelling can be done so that each edge in E_i has the same label, (X_i, Y_i) say. Then G is said to have the *nested subtree property* if, for all distinct $i, j \in \{1, 2, \ldots, k\}$, the ordered pairs (X_i, Y_i) and (X_j, Y_j) satisfy one of the following:

(I) $Y_i \cap Y_j = \emptyset$;
(II) $Y_i \subseteq X_j$; and
(III) $Y_i \subset Y_j$ and $X_j \cap Y_i = \emptyset$.

It is easily checked that, for all distinct ordered pairs (X_i, Y_i) and (X_j, Y_j), at most one of (I)–(III) holds, and if (X_i, Y_i) and (X_j, Y_j) are distinct and satisfy one of (I)–(III), then $Y_i \neq Y_j$. In this section, we will always view G as having each of its edges labelled with the ordered pair associated with a corresponding moving subtree represented by the edge.

Properties (I)–(III) of the nested subtree property capture the way certain clusters of a level-1 network interact. In particular, let $\mathcal{N}$ be a level-1 network, and let u be a vertex of an underlying cycle $\mathcal{C}$ of $\mathcal{N}$, and let v be the (unique) reticulation of $\mathcal{C}$. If u is the root of $\mathcal{N}$ or no arc of $\mathcal{N}$ that is directed into u lies on $\mathcal{C}$, then u is called the *source vertex* of v. Since no two underlying cycles of $\mathcal{N}$ intersect, it is easily seen that this notion is well defined. Now, if v_i and v_j are distinct reticulations of $\mathcal{N}$, and u_i and u_j are the source vertices of v_i and v_j, respectively, then it turns out that the ordered pairs $(C(v_i), C(u_i))$ and $(C(v_j), C(u_j))$ satisfy one of (I)–(III) as illustrated in Figure 3.

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

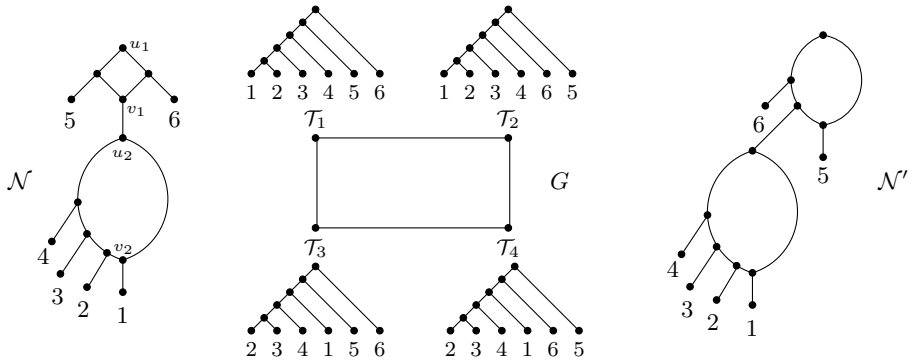

FIGURE 4. The rSPR graph G of a level-1 network $\mathcal{N}$ with $T(\mathcal{N}) = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4\}$, and a level-1 network $\mathcal{N}'$ such that $T(\mathcal{N}) = T(\mathcal{N}')$.

Theorem 5.1. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Then $\mathcal{P}$ is the display set of a level-1 network on X if and only if, for some non-negative integer k, the rSPR graph of $\mathcal{P}$ is isomorphic to Q_k and has the nested subtree property.*

Note that, in the statement of Theorem 5.1, if $\mathcal{P}$ is the display set of a level-1 network, then $|\mathcal{P}| = 2^k$ for some non-negative integer k.

Example. Let $X = \{1, 2, 3, 4, 5, 6\}$, and consider the level-1 network $\mathcal{N}$ on X shown in Figure 4. The set $T(\mathcal{N})$ consists of the four phylogenetic trees $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, and $\mathcal{T}_4$ which are illustrated as the vertices of the rSPR graph G of this set in the same figure. Now, $\{1, 2, 3, 4\}$ is a moving subtree for $\mathcal{T}_1$ and $\mathcal{T}_2$ as well as for $\mathcal{T}_3$ and $\mathcal{T}_4$. In both instances, the ordered pair corresponding to this moving subtree is

$$(\{1, 2, 3, 4\}, X).$$

Similarly, $\{1\}$ is a moving subtree for $\mathcal{T}_1$ and $\mathcal{T}_3$ and for $\mathcal{T}_2$ and $\mathcal{T}_4$. The corresponding ordered pair in both instances is

$$(\{1\}, \{1, 2, 3, 4\}).$$

Since $(\{1, 2, 3, 4\}, X)$ and $(\{1\}, \{1, 2, 3, 4\})$ satisfy (II), and there are only two ordered pairs to compare, it follows that G has the nested subtree property. Note that $(C(v_1), C(u_1)) = (\{1, 2, 3, 4\}, X)$ and $(C(v_2), C(u_2)) = (\{1\}, \{1, 2, 3, 4\})$.

Now observe that we could instead have labelled the edges $\{\mathcal{T}_1, \mathcal{T}_2\}$ and $\{\mathcal{T}_3, \mathcal{T}_4\}$ of G with the ordered pair $(\{5\}, X)$, in which case, $(\{5\}, X)$ and $(\{1\}, \{1, 2, 3, 4\})$ satisfy (III). Thus the choice of ordered pair for the edges $\{\mathcal{T}_1, \mathcal{T}_2\}$ and $\{\mathcal{T}_3, \mathcal{T}_4\}$ is not unique. Indeed, in the proof of Theorem 5.1 this choice leads to the construction of another level-1 network $\mathcal{N}'$ whose display set is also $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4\}$. By way of comparison, $\mathcal{N}'$ is shown in Figure 4.

The remainder of this section establishes Theorem 5.1. We first provide some additional terminology and preliminary results. Subsequently, we establish separately the two directions of Theorem 5.1 as Lemmas 5.5 and 5.6.

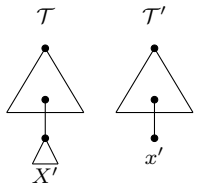

FIGURE 5. A generic example of the subtree reduction that reduces a phylogenetic X-tree $\mathcal{T}$ to a phylogenetic tree $\mathcal{T}'$ on leaf set $(X - X') \cup \{x'\}$. Triangles indicate subtrees.

Let v be a reticulation of a level-1 network $\mathcal{N}$. We say that v is *non-trivial* if the unique underlying cycle of $\mathcal{N}$ that contains v has at least four vertices and, otherwise, we say that v is *trivial*. Let v be a trivial reticulation of $\mathcal{N}$. Obtain a phylogenetic network $\mathcal{N}'$ from $\mathcal{N}$ by deleting one of the two arcs directed into v and suppressing the two resulting degree-2 vertices. (If one of the two arcs is incident with the root of $\mathcal{N}$, choose the other arc to delete.) As $\mathcal{N}$ is level-1, so is $\mathcal{N}'$. Repeating this step for each remaining trivial reticulation in $\mathcal{N}'$ results in a level-1 network, say $\mathcal{N}^*$, with no trivial reticulation. We refer to $\mathcal{N}^*$ as the *essential level*-1 *network* with respect to $\mathcal{N}$. Since no two underlying cycles of $\mathcal{N}$ have a common vertex, $\mathcal{N}^*$ is unique. Moreover, we have the following observation.

Observation 5.2. [24, Theorem 3.1] *Let $\mathcal{N}$ be a level-1 network and let $\mathcal{N}^*$ be the essential level-1 network with respect to $\mathcal{N}$. Then $T(\mathcal{N}) = T(\mathcal{N}^*)$. Moreover, if $\mathcal{N}^*$ has k reticulations, then $|T(\mathcal{N}^*)| = 2^k$.*

The next corollary is an immediate consequence of Theorem 4.1 and Observation 5.2 as well as an immediate consequence of Theorem 5.1.

Corollary 5.3. *Let G be the rSPR graph of a level-1 network with at least two non-trivial reticulations. Then G has a Hamilton cycle.*

Let $\mathcal{T}$ be a phylogenetic X-tree, and suppose that X' is a subset of X such that X' is a cluster of $\mathcal{T}$. Let $\mathcal{T}'$ be the phylogenetic tree obtained from $\mathcal{T}$ by replacing the pendant subtree whose leaf set is X' with a new leaf, x' say. That is, $\mathcal{T}'$ is obtained from $\mathcal{T}$ by deleting all vertices v (and their incident arcs) such that $C(v)$ is a proper subset of X' and label the vertex u whose cluster is X' with x'. Note that the leaf set of $\mathcal{T}'$ is $(X - X') \cup \{x'\}$. We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *subtree reduction on* X'. The leaf x' is referred to as the *replacement leaf*. A generic example of a subtree reduction is shown in Figure 5.

Lemma 5.4. [7, Proposition 3.2] *Let $\mathcal{T}$ and $\mathcal{T}'$ be two distinct phylogenetic X-trees, and suppose that $Y \subseteq X$ such that $\mathcal{T}|Y \cong \mathcal{T}'|Y$. Let $\mathcal{T}_1$ and $\mathcal{T}_1'$ be the phylogenetic trees obtained from $\mathcal{T}$ and $\mathcal{T}'$, respectively, by applying a subtree reduction on Y with new replacement leaf y. Then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$ if and only if $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_1') = 1$. In particular, $\{(X \cup \{\rho\}) - X', X'\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ if and only if either*

(i) *$Y \subseteq X'$ and $\{(X \cup \{\rho\}) - X', (X' - Y) \cup \{y\}\}$ is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$, or*

(ii) $Y \subseteq (X \cup \{\rho\}) - X'$ *and* $\{((X \cup \{\rho\}) - (X' \cup Y)) \cup \{y\}, X'\}$ *is an agreement forest for* $\mathcal{T}_1$ *and* $\mathcal{T}_1'$.

We are now ready to prove the two directions of Theorem 5.1 beginning with the necessary direction.

Lemma 5.5. *Let* $\mathcal{N}$ *be a level-*1 *network on* X*, and let* k *be the number of non-trivial reticulations of* $\mathcal{N}$*. Then the rSPR graph of* $\mathcal{N}$ *is isomorphic to* Q_k *and has the nested subtree property.*

Proof. By Observation 5.2 and the paragraph prior to it, we may assume that $\mathcal{N}$ has no trivial reticulations. Let $\{v_1, v_2, \ldots, v_k\}$ denote the set of reticulations of $\mathcal{N}$ and, for all $i \in \{1, 2, \ldots, k\}$, let u_i denote the source vertex of v_i. Furthermore, let X_i and Y_i denote the clusters $C(v_i)$ and $C(u_i)$, respectively, of $\mathcal{N}$. For the proof of the lemma, we will prove a stronger statement. In particular, we will additionally show that there is an isomorphism that maps the rSPR graph of $\mathcal{N}$ to Q_k such that, for all $i \in \{1, 2, \ldots, k\}$, the subset of edges of G corresponding to the i-th bit edge subset of Q_k can each be labelled (X_i, Y_i), and that this choice of labelling verifies that G has the nested subtree property. The proof is by induction on k. If $k = 0$, then $\mathcal{N}$ is a phylogenetic X-tree and the rSPR graph of $\mathcal{N}$ is isomorphic to Q_0, and the stronger statement, and thus the lemma, immediately follows. If $k = 1$, then the rSPR graph of $\mathcal{N}$ is isomorphic to Q_1, and the stronger statement, and therefore the lemma, immediately follows again. Now suppose that $k \geq 2$ and the stronger statement holds for all level-1 networks on X with at most $k - 1$ reticulations.

For each $i \in \{1, 2, \ldots, k\}$, recall that u_i is the source vertex of v_i. Without loss of generality, we may assume that u_k is a source vertex at maximum distance from the root of $\mathcal{N}$. Let p_1 and p_2 denote the parents of v_k. Since $k \geq 2$, it follows by the maximality of u_k that neither p_1 nor p_2 is the root of $\mathcal{N}$. If either (p_1, v_k) or (p_2, v_k) is a shortcut, we may assume that (p_1, v_k) is the shortcut. Note that at most one of (p_1, v_k) and (p_2, v_k) is a shortcut; otherwise, (p_1, v_k) and (p_2, v_k) are parallel arcs. Let $\mathcal{N}_1$ be the level-1 network on X obtained from $\mathcal{N}$ by deleting the arc (p_2, v_k) and suppressing the two resulting degree-two vertices. Since $\mathcal{N}$ has k reticulations, $\mathcal{N}_1$ has $k - 1$ reticulations and it follows by the induction assumption that there is an isomorphism φ_1 that maps the rSPR graph G_1 of $\mathcal{N}_1$ to Q_{k-1} such that, for all $i \in \{1, 2, \ldots, k-1\}$, if E_i^1 denotes the subset of edges of G_1 corresponding to the i-th bit edge subset of Q_{k-1}, then we can label each edge in E_i^1 with (X_i, Y_i) and this labelling verifies that G_1 has the required nested subtree property of the stronger statement. Since $\mathcal{N}$ is level-1, for each $i \in \{1, 2, \ldots, k-1\}$, the clusters of u_i and v_i are Y_i and X_i, respectively, in $\mathcal{N}_1$. This setup is illustrated in Figure 6 for when $k = 3$, and neither (p_1, v_3) nor (p_2, v_3) is a shortcut in $\mathcal{N}$. Furthermore, the same figure illustrates the rest of the inductive proof.

Now let $\mathcal{N}_2$ be the level-1 network on X obtained from $\mathcal{N}$ by deleting the arc (p_1, v_k) and suppressing the two resulting degree-two vertices. We next construct from G_1 the rSPR graph of $\mathcal{N}_2$. By the maximality of u_k, if $\mathcal{T}$ and $\mathcal{T}'$ are vertices of G_1, then $\mathcal{T}|Y_k \cong \mathcal{T}'|Y_k$. Let Z_k denote the cluster $C(p_2)$ of $\mathcal{N}$. Note that, as (p_2, v_k) is not a shortcut, $X_k \subset Z_k \subseteq Y_k$. Let G_2 be the graph obtained from G_1 by replacing each vertex $\mathcal{T}$ of G_1 with the phylogenetic X-tree $\mathcal{S}$ obtained from $\mathcal{T}$ by

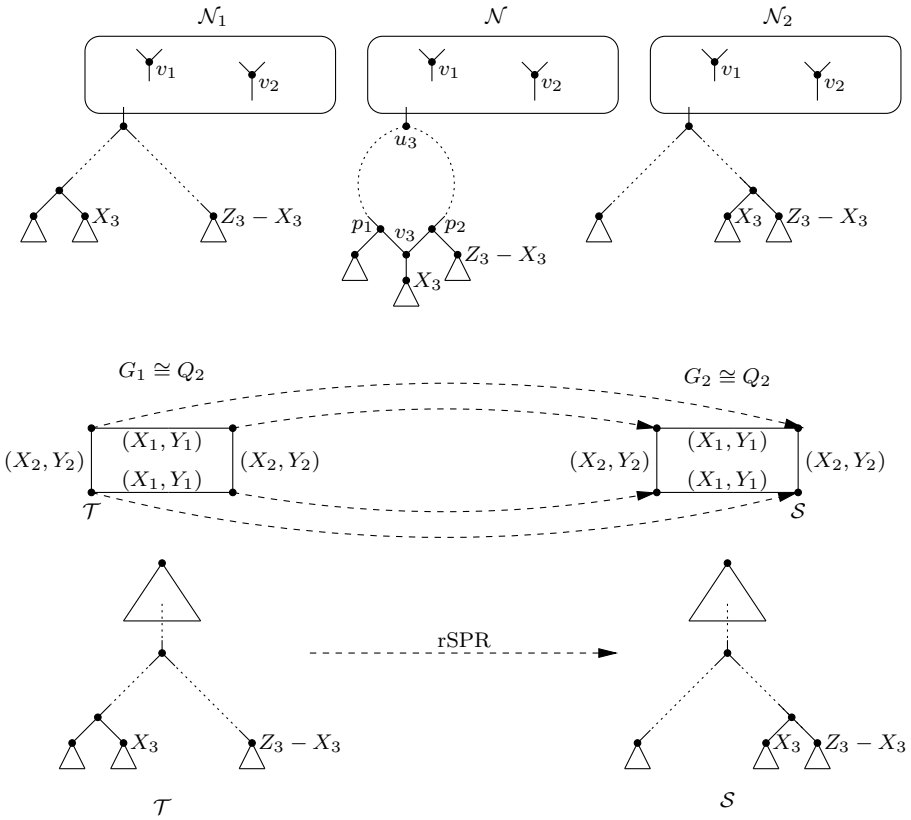

FIGURE 6. Setup as described in the proof of Lemma 5.5 for when $k = 3$, and neither (p_1, v_3) nor (p_2, v_3) is a shortcut in $\mathcal{N}$. Triangles indicate subtrees of their respective level-1 network or phylogenetic tree. Furthermore, vertices whose clusters are X_3 or $Z_3 - X_3$ are labelled accordingly.

a single rSPR operation that prunes the pendant subtree whose leaf set is X_k and regrafts it to the arc directed into the vertex whose cluster is $Z_k - X_k$. Again, this is illustrated in Figure 6. For ease of reading, we say that $\mathcal{S}$ is the vertex in G_2 whose *partner* is $\mathcal{T}$. Evidently, G_2 is isomorphic to Q_{k-1} under the isomorphism φ_2 that is obtained from φ_1 by replacing every vertex $\mathcal{T}$ in G_1 with its partner $\mathcal{S}$ in G_2. Furthermore, by construction, if $\mathcal{S}$ and $\mathcal{S}'$ are vertices of G_2, then $\mathcal{S}|Y_k \cong \mathcal{S}'|Y_k$.

We next show that G_2 is the rSPR graph of $\mathcal{N}_2$ and the labelling of its edges verifies the stronger statement. Since the vertex set of G_1 is the display set of $\mathcal{N}_1$, it follows by the maximality of u_k and construction that the vertex set of G_2 is the display set of $\mathcal{N}_2$. Let $\mathcal{T}$ and $\mathcal{T}'$ be vertices of G_1, and let $\mathcal{S}$ and $\mathcal{S}'$ be the partners of $\mathcal{T}$ and $\mathcal{T}'$, respectively, in G_2. Let $\mathcal{T}_1$, $\mathcal{T}_1'$, $\mathcal{S}_1$, and $\mathcal{S}_1'$ denote the phylogenetic trees obtained from $\mathcal{T}$, $\mathcal{T}'$, $\mathcal{S}$, and $\mathcal{S}'$, respectively, by applying a subtree reduction on Y_k with replacement leaf y_k. First assume that $\mathcal{T}$ and $\mathcal{T}'$ are adjacent in G_1, and let (X', Y') denote the ordered pair labelling the edge joining $\mathcal{T}$ and $\mathcal{T}'$. Then

$\{(X \cup \{\rho\}) - X', X'\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and Y' is the minimal cluster in $C(\mathcal{T}) \cap C(\mathcal{T}')$ properly containing X'. Since $\mathcal{T}|Y_k \cong \mathcal{T}'|Y_k$, it follows by Lemma 5.4 that either (i) $Y_k \subseteq X'$ and

$$\{(X \cup \{\rho\}) - X', (X' - Y_k) \cup \{y_k\}\}$$

is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$, or (ii) $Y_k \subseteq (X \cup \{\rho\}) - X'$ and

$$\{((X \cup \{\rho\}) - (X' \cup Y_k)) \cup \{y_k\}, X'\}$$

is an agreement forest for $\mathcal{T}_1$ and $\mathcal{T}_1'$. If (i) holds, then $(Y' - Y_k) \cup \{y_k\}$ is the minimal cluster in $C(\mathcal{T}_1) \cap C(\mathcal{T}_1')$ properly containing $(X' - Y_k) \cup \{y_k\}$. Furthermore, if (ii) holds, then, by the maximality of u_k, either $Y_k \subset Y'$, in which case, $(Y' - Y_k) \cup \{y_k\}$ is the minimal cluster in $C(\mathcal{T}_1) \cap C(\mathcal{T}_1')$ properly containing X', or $Y_k \cap Y' = \emptyset$, in which case, Y' is the minimal cluster in $C(\mathcal{T}_1) \cap C(\mathcal{T}_1')$ properly containing X'.

Now, by the single rSPR operation in which $\mathcal{S}$ is obtained from $\mathcal{T}$ and $\mathcal{S}'$ is obtained from $\mathcal{T}'$, it follows that $\mathcal{T}_1 \cong \mathcal{S}_1$ and $\mathcal{T}_1' \cong \mathcal{S}_1'$. Thus, as $\mathcal{S}|Y_k \cong \mathcal{S}'|Y_k$, it follows by Lemma 5.4 that

$$\{(X \cup \{\rho\}) - X', X'\}$$

is an agreement forest for $\mathcal{S}$ and $\mathcal{S}'$, and Y' is the minimal cluster in $C(\mathcal{S}) \cap C(\mathcal{S}')$ properly containing X'. Hence $\mathcal{S}$ and $\mathcal{S}'$ are correctly joined by an edge labelled (X', Y') in G_2. Moreover, if $\mathcal{T}$ and $\mathcal{T}'$ are not adjacent in G_1, then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') > 1$ and so, by Lemma 5.4, $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_1') > 1$. Thus, as $\mathcal{T}_1 \cong \mathcal{S}_1$ and $\mathcal{T}_1' \cong \mathcal{S}_1'$, we have $d_{\mathrm{rSPR}}(\mathcal{S}_1, \mathcal{S}_1') > 1$. Therefore, by Lemma 5.4, $d_{\mathrm{rSPR}}(\mathcal{S}, \mathcal{S}') > 1$. We deduce that G_2 is the rSPR graph of $\mathcal{N}_2$ and, as the labelling of the edges of G_1 verifies that G_1 has the nested subtree property, the labelling of the edges of G_2 has the required nested subtree property of the stronger statement.

We now construct the rSPR graph of $\mathcal{N}$ from G_1 and G_2 as follows. Take G_1 and G_2 and, for each vertex $\mathcal{T}$ in G_1, join $\mathcal{T}$ to its partner $\mathcal{S}$ in G_2 and label the edge (X_k, Y_k). Call the resulting graph G. By construction, there is an isomorphism that maps G to Q_k such that, for all $i \in \{1, 2, \ldots, k\}$ each edge in the subset of edges of G corresponding to the i-th bit subset of Q_k is labelled (X_i, Y_i). Furthermore, the end vertices of an edge joining a vertex $\mathcal{T}$ in G_1 with a vertex $\mathcal{S}$ in G_2 have an agreement forest $\{(X \cup \{\rho\}) - X_k, X_k\}$, that is, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{S}) = 1$, and Y_k is the minimal cluster in $C(\mathcal{T}) \cap C(\mathcal{S})$ containing X'.

We next show that if $\mathcal{T}$ is a vertex in G_1 and $\mathcal{S}$ is a vertex in G_2, but $\mathcal{S}$ is not the partner of $\mathcal{T}$, then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{S}) > 1$. Say $\mathcal{T}$ and $\mathcal{S}$ are such vertices, but $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{S}) = 1$. Then $\mathcal{T}$ and $\mathcal{S}$ have an agreement forest $\{(X \cup \{\rho\}) - Z, Z\}$, where Z is a cluster of $\mathcal{T}$ and $\mathcal{S}$, and $\mathcal{T}|Z \cong \mathcal{S}|Z$. Since $\mathcal{T}$ and $\mathcal{S}$ have pendant subtrees with leaf set Y_k, but $\mathcal{T}|Y_k \ncong \mathcal{S}|Y_k$, it follows that $Z \subset Y_k$. As $\mathcal{T}$ is a vertex of G_1 and $\mathcal{S}$ is a vertex of G_2, this implies that $Z = X_k$, in which case, $\mathcal{S}$ is the partner of $\mathcal{T}$, a contradiction. Thus, as G_1 and G_2 are the rSPR graphs of $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively, and a phylogenetic tree is displayed by $\mathcal{N}$ if and only if it is either displayed by $\mathcal{N}_1$ or displayed by $\mathcal{N}_2$, it follows that G is the rSPR graph of $\mathcal{N}$.

Lastly, let i and j be distinct elements of $\{1, 2, \ldots, k\}$. If $k \notin \{i, j\}$, then, by construction, (X_i, Y_i) and (X_j, Y_j) satisfy one of (I)–(III) in the definition of the nested subtree property. So assume that $i = k$. Now u_j is the source vertex of v_j,

and $C(u_j) = Y_j$ and $C(v_j) = X_j$ in $\mathcal{N}$. Say $Y_j \cap Y_k \neq \emptyset$. Then, by the choice of u_k, we have that $Y_k \subset Y_j$. Since $\mathcal{N}$ is level-1, it is easily seen that either $Y_k \subseteq X_j$ or $X_j \cap Y_k = \emptyset$, and so (X_j, Y_j) and (X_k, Y_k) satisfy either (II) or (III). Thus, G is the rSPR graph of $\mathcal{N}$ and the labelling of the edges of G has the required nested subtree property of the stronger statement. This completes the proof of the lemma. □

The next lemma is the converse of Lemma 5.5, and thereby completes the proof of Theorem 5.1.

Lemma 5.6. *Let $\mathcal{P}$ be a set of phylogenetic X-trees and let G be the rSPR graph of $\mathcal{P}$. If G is isomorphic to Q_k for some non-negative integer k, and G has the nested subtree property, then there is a level-1 network $\mathcal{N}$ on X whose display set is $\mathcal{P}$.*

Proof. Suppose that there is an isomorphism from G to Q_k and that G has the nested subtree property. Under this isomorphism, we may assume that, for all $i \in \{1, 2, \ldots, k\}$, each edge in E_i, the subset of edges of G corresponding to the i-th bit edge subset of Q_k, has the label (X_i, Y_i) and, for all distinct $i, j \in \{1, 2, \ldots, k\}$ the ordered pairs (X_i, Y_i) and (X_j, Y_j) satisfy one of (I)–(III). Note that, as G is isomorphic to Q_k, we have $|\mathcal{P}| = 2^k$ for some non-negative integer k. The proof is by induction on k. If $k = 0$, then $|\mathcal{P}| = 1$, and the lemma trivially holds by choosing $\mathcal{N}$ to be the phylogenetic X-tree in $\mathcal{P}$. Now suppose that $k \geq 1$ and that the lemma holds for all sets of phylogenetic trees of size 2^{k-1} with the same leaf set.

Amongst the ordered pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k)$, choose (X_i, Y_i) to be an ordered pair with the property that there is no (X_j, Y_j) for which $Y_j \subset Y_i$. Such an ordered pair exists, otherwise $Y_i = Y_j$ for some $i \neq j$ contradicting the assumption that the ordered pairs satisfy the nested subtree property. Furthermore, if there is an ordered pair (X_j, Y_j) such that $X_j \cap Y_i \neq \emptyset$, then, as G satisfies the nested subtree property, $Y_i \subseteq X_j$.

Consider the graph obtained from G by deleting the edges in E_i. The resulting graph has exactly two components, G_1 and G_2 say, and each component is isomorphic to Q_{k-1}. Observe that, as G has the nested subtree property, G_1 and G_2 also have this property. We next show that if $\mathcal{T}$ and $\mathcal{T}'$ are vertices of G_1, then $\mathcal{T}|Y_i \cong \mathcal{T}'|Y_i$. Certainly, $Y_i \in C(\mathcal{T})$ and $Y_i \in C(\mathcal{T}')$ as all vertices of G_1 are incident with an edge labelled (X_i, Y_i) in G. Assume that $\mathcal{T}$ and $\mathcal{T}'$ are adjacent in G_1. Let (X_j, Y_j) denote the ordered pair labelling the edge joining $\mathcal{T}$ and $\mathcal{T}'$ in G_1, where $i \neq j$. Then

$$\{(X \cup \{\rho\}) - X_j, X_j\}$$

is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. By the choice of (X_i, Y_i), one of (I) $Y_i \cap Y_j = \emptyset$, (II) $Y_i \subseteq X_j$, and (III) $Y_i \subset Y_j$ and $X_j \cap Y_i = \emptyset$ holds. Thus either $Y_i \subseteq (X \cup \{\rho\}) - X_j$ or $Y_i \subseteq X_j$. Therefore, as $\{(X \cup \{\rho\}) - X_j, X_j\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, it follows that in all cases $\mathcal{T}|Y_i \cong \mathcal{T}'|Y_i$. Since Q_{k-1} is connected, we can repeatedly apply this argument to eventually show that $\mathcal{T}|Y_i$ and $\mathcal{T}'|Y_i$ are isomorphic for all vertices $\mathcal{T}$ and $\mathcal{T}'$ in G_1. Similarly, if $\mathcal{S}$ and $\mathcal{S}'$ are vertices of G_2,

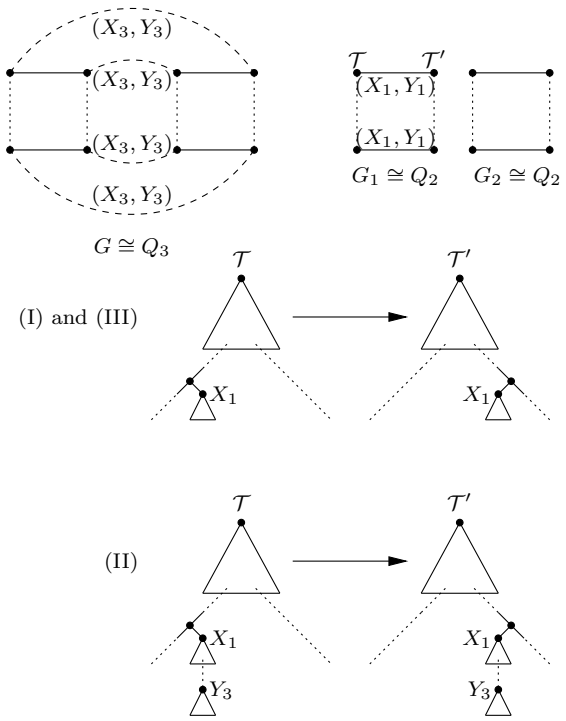

FIGURE 7. Setup as described in the proof of Lemma 5.6 for when $k = i = 3$, $j = 1$, and E_1 contains the solid edges of G, E_2 contains the dotted edges of G, and E_3 contains the dashed edges of G. Vertices whose clusters are X_1 or Y_3 are labelled accordingly. If (I) or (III) in the definition of the nested subtree property applies to (X_1, Y_1) and (X_3, Y_3), then $X_1 \cap Y_3 = \emptyset$. Otherwise, if (II) applies, then $Y_3 \subseteq X_1$. As $\mathcal{T}$ and $\mathcal{T}'$ are joined by an edge in G_1 that is labelled (X_1, Y_1), it follows that $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by a single rSPR operation that prunes and regrafts the pendant subtree $\mathcal{T}|X_1$. Hence, regardless of which of (I)–(III) applies, as $\{(X \cup \{\rho\}) - X_1, X_1\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, where X is the leaf set of $\mathcal{T}$ and $\mathcal{T}'$, we have $\mathcal{T}|Y_3 \cong \mathcal{T}'|Y_3$.

then $\mathcal{S}|Y_i \cong \mathcal{S}'|Y_i$. For $k = i = 3$ and $j = 1$, the preceding argument is illustrated in Figure 7.

Let $\mathcal{P}_1$ denote the set of vertices of G_1, and let $\mathcal{P}_1'$ denote the collection of phylogenetic trees on $(X - Y_i) \cup \{y_i\}$ obtained by replacing each phylogenetic X-tree in $\mathcal{P}_1$ with the phylogenetic tree on $(X - Y_i) \cup \{y_i\}$ resulting from a subtree reduction on Y_i, where the replacement leaf is y_i. By Lemma 5.4, the rSPR graph G_1' of $\mathcal{P}_1'$ can be obtained from G_1 by replacing each vertex with the phylogenetic tree in $\mathcal{P}_1'$ resulting from this subtree reduction, and replacing those ordered pairs (X_j, Y_j) in which $Y_i \subseteq X_j$ with

$$((X_j - Y_i) \cup \{y_i\}, (Y_j - Y_i) \cup \{y_i\})$$

and in which $Y_i \subset Y_j$ and $X_j \cap Y_i = \emptyset$ with

$$(X_j, (Y_j - Y_i) \cup \{y_i\}).$$

Using the fact that G_1 has the nested subtree property, a routine check shows that G_1' also has the nested subtree property and so, by the induction assumption, there is a level-1 network $\mathcal{N}_1'$ on $(X - Y_i) \cup \{y_i\}$ whose display set is $\mathcal{P}_1'$. Now let $\mathcal{N}_1$ be the level-1 network on X obtained from $\mathcal{N}_1'$ by replacing y_i with the (pendant) subtree $\mathcal{T}|Y_i$, where $\mathcal{T} \in \mathcal{P}_1$. That is, $\mathcal{N}_1$ is obtained from $\mathcal{N}_1'$ by identifying the root of a phylogenetic tree isomorphic to $\mathcal{T}|Y_i$ with the vertex y_i. Since the display set of $\mathcal{N}_1'$ is $\mathcal{P}_1'$, it follows that the display set of $\mathcal{N}_1$ is $\mathcal{P}_1$.

Let $\mathcal{S}$ be a vertex of G_2, and recall that if $\mathcal{S}'$ is also a vertex of G_2, then $\mathcal{S}|Y_i \cong \mathcal{S}'|Y_i$. Let $\mathcal{T}$ be the unique vertex of G_1 such that $\{\mathcal{T}, \mathcal{S}\}$ is an edge in G. Furthermore, let u (resp. u') be the vertex of $\mathcal{T}$ (resp. $\mathcal{S}$) such that $X_i \subset C(u)$ (resp. $X_i \subset C(u')$) and u (resp. u') has no child whose cluster is a proper superset of X_i. Note that $C(u) \subseteq Y_i$ and $C(u') \subseteq Y_i$ as (X_i, Y_i) labels the edge $\{\mathcal{T}, \mathcal{S}\}$. Let $\mathcal{N}$ be the phylogenetic network obtained from $\mathcal{N}_1$ in one of the following two ways:

(i) If $(C(u) - X_i) \cap (C(u') - X_i) = \emptyset$ or $C(u') \subseteq C(u)$, subdivide the arc directed into the (unique) vertex whose cluster is $C(u') - X_i$, subdivide the arc directed into the (unique) vertex whose cluster is X_i, and adjoin an arc from the first to the second of these subdivisions.
(ii) If $C(u) \subseteq C(u')$, subdivide the arc directed into the (unique) vertex whose cluster is $C(u')$, subdivide the arc directed into the (unique) vertex whose cluster is X_i, and adjoin an arc from the first to the second of these subdivisions.

The arcs that get subdivided in (i) and (ii) exist as X_i is a cluster of $\mathcal{T}$ and $\mathcal{S}$, and $\mathcal{T}|((X \cup \{\rho\}) - X_i) \cong \mathcal{S}|((X \cup \{\rho\}) - X_i)$. By the choice of (X_i, Y_i), the network $\mathcal{N}$ is level-1. If $\mathcal{T}$ is the unique vertex of G_1 in G adjacent to $\mathcal{S}$, then $\mathcal{T}|(X - X_i) \cong \mathcal{S}|(X - X_i)$ and $\mathcal{T}|X_i \cong \mathcal{S}|X_i$. Therefore, by construction, $\mathcal{N}$ displays $\mathcal{S}$. Since the choice of $\mathcal{S}$ is arbitrary in the sense that $\mathcal{S}|Y_i \cong S'|Y_i$ for all vertices $\mathcal{S}'$ of G_2, the lemma now follows. □

6. Reconstructing Level-1 Networks for a Given Display Set

Using the characterisation established in the last section, this section describes a polynomial-time algorithm, called Construct Level-1 Network, that, given a collection $\mathcal{P}$ of phylogenetic trees, reconstructs a level-1 network whose display set is $\mathcal{P}$ or returns that no such network exists. As a corollary, we derive that in fact all such networks can be reconstructed. Before we present the algorithm, we establish several results on the properties of rSPR graphs and hypercubes that are needed to show that Construct Level-1 Network runs in polynomial time.

Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees, and suppose that $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by a single rSPR operation. In particular, $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by deleting an arc (u, v) and reattaching the resulting rooted subtree that contains v with a new arc (u', v). Ignoring the suppressing of u, if the underlying path joining

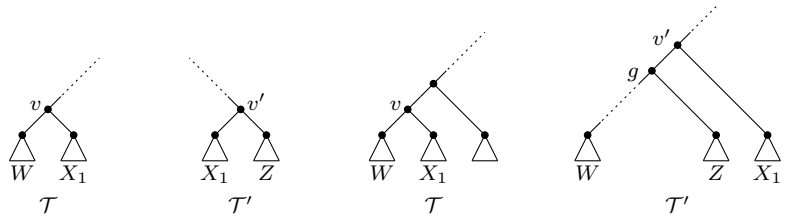

FIGURE 8. Setup of $\mathcal{T}$ and $\mathcal{T}'$ as used in the proof of Proposition 6.1 depending on whether X_1 is a moving subtree of Type I (left) or Type II (right). Triangles indicate pendant subtrees with at least one leaf.

u and u' consists of at most two arcs, then we say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *rooted nearest neighbour interchange* (rNNI) operation. Note that if this path consists of one arc, then $\mathcal{T} \cong \mathcal{T}'$. The *rNNI distance* between any two phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ is the minimum number of rNNI operations that transforms $\mathcal{T}$ into $\mathcal{T}'$, and is denoted by $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}')$. Like the rSPR operation, rNNI is reversible and $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}')$ is well defined as there is always a sequence of rNNI operations that transforms $\mathcal{T}$ into $\mathcal{T}'$ [7, 27].

Again, let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees with $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$. We denote the set of moving subtrees for $\mathcal{T}$ and $\mathcal{T}'$ by $M(\mathcal{T}, \mathcal{T}')$. Furthermore, for a cluster X' of $\mathcal{T}$, we call the vertex, v say, of $\mathcal{T}$, such that $C(v) = X'$, the *parent* of X' and the vertex, u say, of $\mathcal{T}$, such that (u, v) is the arc of $\mathcal{T}$ directed into v, the *grandparent* of X'. Lastly, if $X' \subseteq X$ and $|X'| = 3$, then $\mathcal{T}|X'$ is called a *rooted triple* of $\mathcal{T}$. If $X' = \{a, b, c\}$ and the underlying paths joining a and b, and joining the root of $\mathcal{T}$ and c are disjoint, then we denote the rooted triple $\mathcal{T}|\{a, b, c\}$ by $ab|c$ or, equivalently, $ba|c$. The set of rooted triples of $\mathcal{T}$ is denoted by $\mathcal{R}(\mathcal{T})$. It is well known that $\mathcal{R}(\mathcal{T}) = \mathcal{R}(\mathcal{T}')$ if and only if $\mathcal{T} \cong \mathcal{T}'$ (see, for example, [29]).

The next proposition bounds the number of moving subtree for two phylogenetic trees. More specifically, it shows that, if two phylogenetic trees have rSPR distance one, then there exists a unique moving subtree unless the two trees also have rNNI distance one in which case there are exactly three moving subtrees.

Proposition 6.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees, and suppose that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$. Then $|M(\mathcal{T}, \mathcal{T}')| \in \{1, 3\}$. Moreover, $|M(\mathcal{T}, \mathcal{T}')| = 3$ if and only if $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$.*

Proof. Let ρ denote the root of $\mathcal{T}$ and $\mathcal{T}'$. Since $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$, it follows that $|M(\mathcal{T}, \mathcal{T}')| \geq 1$. Let $X_1 \in M(\mathcal{T}, \mathcal{T}')$, and let v and v' be the grandparents of X_1 in $\mathcal{T}$ and $\mathcal{T}'$, respectively. Furthermore, let u and u' be the vertices of $\mathcal{T}$ and $\mathcal{T}'$, respectively, such that $C_{\mathcal{T}}(u)$ and $C_{\mathcal{T}'}(u')$ is the minimal cluster in $C(\mathcal{T}) \cap C(\mathcal{T}')$ properly containing X_1. Let $W = C_{\mathcal{T}}(v) - X_1$. Since rNNI is reversible, we may view the rSPR operations corresponding to the moving subtrees in $M(\mathcal{T}, \mathcal{T}')$ as transforming $\mathcal{T}$ into $\mathcal{T}'$. Furthermore, we may also assume that either $W \cap C_{\mathcal{T}'}(v') = \emptyset$, in which case, we say X_1 is a *Type I* moving subtree, or $W \subseteq C_{\mathcal{T}'}(v')$, in which case, we say X_1 is a *Type II* moving subtree.

Now, if X_1 is a Type I moving subtree, let $Z = C_{\mathcal{T}'}(v') - X_1$, and if X_1 is a Type II moving subtree, let Z be the cluster of $\mathcal{T}'$ such that $Z \cap (X_1 \cup W) = \emptyset$, and the grandparent, g say, of Z in $\mathcal{T}'$ is on the path from v' to the parent of W and (v', g) is an arc in $\mathcal{T}'$. The setup of $\mathcal{T}$ and $\mathcal{T}'$ depending on whether X_1 is a moving subtree of Type I or Type II is illustrated in Figure 8. Note that g exists; otherwise, $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 0$. Observe that W and Z are both clusters of $\mathcal{T}$ and $\mathcal{T}'$, and X_1, W, and Z are (pairwise) disjoint subsets of X. Moreover, if $x_1 \in X_1$, $w \in W$, and $z \in Z$, then $x_1w|z \in \mathcal{R}(\mathcal{T})$, and either $x_1z|w \in \mathcal{R}(\mathcal{T}')$ if X_1 is a Type I moving subtree or $wz|x_1 \in \mathcal{R}(\mathcal{T}')$ if X_1 is a Type II moving subtree.

Now suppose that $X_2 \in M(\mathcal{T}, \mathcal{T}') - \{X_1\}$. It follows by the observations at the end of the last paragraph that $(X_1 \cup W \cup Z) \cap X_2 \neq \emptyset$. Otherwise, $\{(X \cup \{\rho\}) - X_2, X_2\}$ is not an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ as $x_1w|z$ is a rooted triple of $\mathcal{T}|((X \cup \{\rho\}) - X_2)$ but $x_1w|z$ is not a rooted triple of $\mathcal{T}'|((X \cup \{\rho\}) - X_2)$. Furthermore, using the same observations, a similar analysis establishes that either $X_2 = W$, or $Z \subseteq X_2$ and $(X_1 \cup W) \cap X_2 = \emptyset$.

Assume that $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') \neq 1$. Then $C_{\mathcal{T}}(u) - (X_1 \cup W \cup Z)$ is non-empty. Let $\ell \in C_{\mathcal{T}}(u) - (X_1 \cup W \cup Z)$, and note that $x_1w|\ell \in \mathcal{R}(\mathcal{T})$, but $x_1w|\ell \notin \mathcal{R}(\mathcal{T}')$. Also, if X_1 is a Type I moving subtree, then either $x_1\ell|z \in \mathcal{R}(\mathcal{T})$ and $x_1z|\ell \in \mathcal{R}(\mathcal{T}')$, or $z\ell|x_1 \in \mathcal{R}(\mathcal{T})$ and $x_1z|\ell \in \mathcal{R}(\mathcal{T}')$, while if X_1 is a Type II moving subtree, then $x_1\ell|z \in \mathcal{R}(\mathcal{T})$ and $\ell z|x_1 \in \mathcal{R}(\mathcal{T}')$. Thus $X_2 \neq W$ as $\{(X \cup \{\rho\}) - W, W\}$ is not an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$, and so $Z \subseteq X_2$ and $(X_1 \cup W) \cap X_2 = \emptyset$. Therefore, as $x_1w|\ell \in \mathcal{R}(\mathcal{T})$ but $x_1w|\ell \notin \mathcal{R}(\mathcal{T}')$, it follows that $Z \cup \{\ell\} \subseteq X_2$. But then, as $(X_1 \cup W) \cap X_2 = \emptyset$, we have $z\ell|x_1 \in \mathcal{R}(\mathcal{T})$ and $z\ell|x_1 \in \mathcal{R}(\mathcal{T}')$, a contradiction. Thus if $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') \neq 1$, then $|M(\mathcal{T}, \mathcal{T}')| = 1$.

Now assume that $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$. Then

$$C_{\mathcal{T}}(u) = C_{\mathcal{T}'}(u') = X_1 \cup W \cup Z,$$

and it is easily checked that each of X_1, W, and Z is a moving subtree for $\mathcal{T}$ and $\mathcal{T}'$. Moreover, these are the only moving subtrees for $\mathcal{T}$ and $\mathcal{T}'$ as we argued earlier that if $X_2 \in M(\mathcal{T}, \mathcal{T}') - \{X_1\}$, then either $X_2 = W$, or $Z \subseteq X_2$ and $(X_1 \cup W) \cap X_2 = \emptyset$. Hence if $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$, then $|M(\mathcal{T}, \mathcal{T}')| = 3$. This completes the proof of the proposition. □

The next two lemmas and Proposition 6.4 establish properties of hypercubes and graphs that are isomorphic to a hypercube. Let H be a graph, and suppose that H is isomorphic to Q_k for some non-negative integer k. Then there is an isomorphism φ from the set of vertices of H to the set B of all k-bit strings such that if u and v are adjacent vertices in H, then $\varphi(u)$ and $\varphi(v)$ differ in exactly one position. Under φ, the i-th bit edge subset of H is the subset of edges whose end vertices differ precisely in the i-th position. However, φ is not the only such isomorphism. We next show that, regardless of the isomorphism, the collection of subsets of the edge set of H corresponding to the bit edge subsets of Q_k is always the same. The following result can be found, for example, in [23].

Lemma 6.2. *Let k be a non-negative integer, and let e and f be adjacent edges of Q_k. Then Q_k has a unique 4-cycle containing e and f.*

Lemma 6.3. *Let H be a graph, and suppose that φ is a (graph) isomorphism between H and Q_k for some non-negative integer $k \geq 2$. If v_1, v_2, v_3, v_4, v_1 is a 4-cycle of H, then the edges $\{\varphi(v_1), \varphi(v_2)\}$ and $\{\varphi(v_3), \varphi(v_4)\}$ are in the same bit edge subset of Q_k.*

Proof. Let v_1, v_2, v_3, v_4, v_1 be a 4-cycle of H. For all $i \in \{1, 2, 3, 4\}$, we have $\varphi(v_i)$ is a k-bit string. Since v_1 and v_2, and v_1 and v_4 are adjacent, it follows that $\varphi(v_1)$ and $\varphi(v_2)$ differ in precisely one bit, say the i-th bit, and $\varphi(v_1)$ and $\varphi(v_4)$ differ in precisely one bit, say the j-th bit. Furthermore, as $\varphi(v_2) \neq \varphi(v_4)$, we have $i \neq j$. Also, as H is isomorphic to Q_k, the vertices v_1 and v_3 are not adjacent in H, and so $\varphi(v_1)$ and $\varphi(v_3)$ differ in precisely two bits. But v_3 is adjacent to v_2 and v_4, so these two bits are the i-th and j-th bits. Hence, without loss of generality, we may assume that the i-th and j-th bits of $\varphi(v_1)$, $\varphi(v_2)$, $\varphi(v_3)$, and $\varphi(v_4)$ are 0 0, 1 0, 1 1, and 0 1, respectively. It now follows that $\{\varphi(v_1), \varphi(v_2)\}$ and $\{\varphi(v_3), \varphi(v_4)\}$ are in the same bit edge subset of Q_k. □

Proposition 6.4. *Let H be a graph, and suppose that φ_1 and φ_2 are both (graph) isomorphisms between H and Q_k for some non-negative integer k. Then*

$$\{\varphi_1^{-1}(E_1), \varphi_1^{-1}(E_2), \ldots, \varphi_1^{-1}(E_k)\} = \{\varphi_2^{-1}(E_1), \varphi_2^{-1}(E_2), \ldots, \varphi_2^{-1}(E_k)\}$$

where, for all $i \in \{1, 2, \ldots, k\}$, the sets $\varphi_1^{-1}(E_i)$ and $\varphi_2^{-1}(E_i)$ are the subsets of $E(H)$ mapped to E_i under φ_1 and φ_2, respectively.

Proof. Suppose that H is isomorphic to Q_k for some non-negative integer k. If $k \in \{0, 1\}$, then the proposition trivially holds, so assume that $k \geq 2$. Let v_1 be a vertex of H. Then, under an isomorphism between H and Q_k, each of the k edges incident with v_1 are assigned to distinct bit edge subsets of Q_k. Let f_1 be an edge incident with v_1 in H. We next show that, regardless of the choice of isomorphism between H and Q_k, the bit edge subset of Q_k containing the image of f_1 is always the same subset of edges of Q_k.

Since H is isomorphic to Q_k, it follows that H has a Hamilton cycle

$$v_1,\, e_1,\, v_2,\, e_2,\, v_3,\, \ldots,\, v_{2^k},\, e_{2^k},\, v_1$$

starting at v_1. Consider Algorithm 1 which traverses this Hamilton cycle. We next show that the image of F is a bit edge subset of Q_k. By Lemma 6.2, Line 4 is well defined, that is, there is a unique edge of H incident with v_{i+1} that is in the unique 4-cycle containing f_i and e_i. Thus, by Lemma 6.3, regardless of the choice of isomorphism between H and Q_k, the bit edge subset of Q_k that contains the image of f_1 also contains the image of f_2 and, more generally, the bit edge subset of Q_k that contains the image of f_i also contains the image of f_{i+1}. Hence, the bit edge subset of Q_k that contains the image of f_1, contains the images of each of the edges in F.

Now, by construction, it is easily checked that every vertex of H is incident with exactly one edge in F, that is, F is a perfect matching of H, and so $|F| = 2^{k-1}$. Since a bit edge subset of Q_k has size 2^{k-1}, it follows that the image of F is a bit edge subset of Q_k. Repeating this process for each of the remaining edges incident with v_1 establishes the proposition. □

Algorithm 1: COMPUTE BIT EDGE SUBSET

Input: A graph H that is isomorphic to Q_k for some integer $k \geq 2$, a Hamilton cycle v_1, e_1, v_2, e_2, v_3, $\ldots$, v_{2^k}, e_{2^k}, v_1 of H and an edge f_1 incident with v_1 in H.

Output: A subset F of the edges of H.

```
F ← {f_1}
for i ← 1 to 2^k − 1 do
    if f_i ≠ e_i then
        Set f_{i+1} to be the edge of H incident with v_{i+1} that is in the unique
          4-cycle containing f_i and e_i
    else
        f_{i+1} ← f_i
    F ← F ∪ {f_{i+1}}
return F
```

Let G be the rSPR graph of a collection $\mathcal{P}$ of phylogenetic X-trees such that $G \cong Q_k$. Furthermore, let $e = \{\mathcal{T}, \mathcal{T}'\}$ be an edge in the i-th bit edge subset E_i of G for some $i \in \{1, 2, \ldots, k\}$. If (X', Y') is an ordered pair such that $X' \in M(\mathcal{T}, \mathcal{T}')$ and Y' is the minimal cluster in $C(\mathcal{T}) \cap C(\mathcal{T}')$ that properly contains X', we say that (X', Y') is an *ordered pair for* e. Moreover, (X', Y') is said to *verify* E_i if (X', Y') is an ordered pair for each edge in E_i. Now suppose that

$$O = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k))$$

is a sequence of distinct ordered pairs such that each (X_i, Y_i) with $i \in \{1, 2, \ldots, k\}$ verifies E_i. If, for each pair $i, j \in \{1, 2, \ldots, k\}$ and $i \neq j$, the two ordered pairs (X_i, Y_i) and (X_j, Y_j) satisfy one of (I)–(III) in the definition of the nested subtree property, then we say that O *verifies* the nested subtree property of G.

The next two lemmas establish properties of ordered pairs for edges of an rSPR graph. These properties are then used in Proposition 6.7 to show that, provided an rSPR graph G has the nested subtree property, any ordered pair that labels an edge $\{\mathcal{T}, \mathcal{T}'\}$ with $d_{\text{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$ and verifies its associated bit edge subset can also be used to verify the nested subtree property of G.

Lemma 6.5. *Let $\{\mathcal{T}, \mathcal{T}'\}$ be an edge of an rSPR graph G. If (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) are ordered pairs for e, then $Y_1 = Y_2 = Y_3$, $X_1 \cup X_2 \cup X_3 = Y_1$, and the three sets X_1, X_2, and X_3 are pairwise disjoint.*

Proof. By Proposition 6.1, we have $M(\mathcal{T}, \mathcal{T}') = \{X_1, X_2, X_3\}$ and $d_{\text{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$. The lemma now follows from the definition of an rNNI move. □

Lemma 6.6. *Let G be the rSPR graph for a collection $\mathcal{P}$ of phylogenetic X-trees such that $G \cong Q_k$ for some non-negative integer k. Let (X', Y') be an ordered pair that verifies E_i for some $i \in \{1, 2, \ldots, k\}$. Then X' and Y' are clusters of each element in $\mathcal{P}$.*

Proof. Let $\mathcal{T}$ be a vertex of G. Since $G \cong Q_k$, so E_i is a perfect matching of G, it follows that $\mathcal{T}$ is incident with an edge $e \in E_i$. Hence, X' and Y' are both elements in $C(\mathcal{T})$. □

Proposition 6.7. *Let G be the rSPR graph for a collection $\mathcal{P}$ of phylogenetic X-trees such that $G \cong Q_k$ for some non-negative integer k. Let $e = \{\mathcal{T}, \mathcal{T}'\}$ be an edge of E_i with $i \in \{1, 2, \ldots, k\}$ such that (X_i^1, Y_i^1), (X_i^2, Y_i^2), and (X_i^3, Y_i^3) are ordered pairs for e. Suppose that the sequence of ordered pairs*

$$((X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k))$$

verifies the nested subtree property of G and that $(X_i, Y_i) = (X_i^1, Y_i^1)$. If (X_i^ℓ, Y_i^ℓ) with $\ell \in \{2, 3\}$ verifies E_i, then the sequence

$$((X_1, Y_1), (X_2, Y_2), \ldots, (X_{i-1}, Y_{i-1}), (X_i^\ell, Y_i^\ell), (X_{i+1}, Y_{i+1}), \ldots, (X_k, Y_k))$$

also verifies the nested subtree property of G.

Proof. By Proposition 6.1, we first observe that $M(\mathcal{T}, \mathcal{T}') = \{X_i^1, X_i^2, X_i^3\}$ and $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$. Now, suppose that (X_i^ℓ, Y_i^ℓ) with $\ell \in \{2, 3\}$ verifies E_i. Without loss of generality, we may assume that $\ell = 2$. Let f be an edge of G such that $f \in E_j$ with $i \neq j$. We consider three cases depending on which of the three properties in the definition of the nested subtree property the two ordered pairs (X_i, Y_i) and (X_j, Y_j) satisfy. To this end, recall Lemma 6.5, which we will freely use throughout all three cases.

First, suppose that (X_i, Y_i) and (X_j, Y_j) satisfy (I). As $Y_i \cap Y_j = \emptyset$ it follows that $Y_i^2 \cap Y_j = \emptyset$. Thus (X_i^2, Y_i^2) and (X_j, Y_j) also satisfy (I).

Second, suppose that (X_i, Y_i) and (X_j, Y_j) satisfy (II). Clearly, if $Y_i \subseteq X_j$, then it immediately follows that $Y_i^2 \subseteq X_j$. Hence (X_i^2, Y_i^2) and (X_j, Y_j) also satisfy (II). We may therefore assume that $Y_j \subseteq X_i$. As $Y_j \subseteq X_i \subset Y_i$, we have $Y_j \subset Y_i^2$. Moreover, because $Y_j \subseteq X_i$ and $X_i \cap X_i^2 = \emptyset$, it follows that $X_i^2 \cap Y_j = \emptyset$. Thus (X_i^2, Y_i^2) and (X_j, Y_j) satisfy (III).

Third, suppose that (X_i, Y_i) and (X_j, Y_j) satisfy (III). Similar to the previous case, if $Y_i \subset Y_j$ and $X_j \cap Y_i = \emptyset$, then $Y_i^2 \subset Y_j$ and $X_j \cap Y_i^2 = \emptyset$ and, so, (X_i^2, Y_i^2) and (X_j, Y_j) also satisfy (III). Therefore, assume that $Y_j \subset Y_i$ and $X_i \cap Y_j = \emptyset$. If $X_i^2 \cap Y_j = \emptyset$, then, as $Y_j \subset Y_i^2$, it follows that (X_i^2, Y_i^2) and (X_j, Y_j) again satisfy (III). On the other hand, if $X_i^2 \cap Y_j \neq \emptyset$, we consider X_i^3 to complete the argument. Assume that $X_i^3 \cap Y_j \neq \emptyset$. Then, as $X_i^1 \cup X_i^2 \cup X_i^3 = Y_i$, we have $Y_j \subseteq X_i^2 \cup X_i^3$. As $d_{\mathrm{rNNI}}(\mathcal{T}, \mathcal{T}') = 1$, each element in $M(\mathcal{T}, \mathcal{T}') = \{X_i^1, X_i^2, X_i^3\}$ is a cluster of $\mathcal{T}$ and $\mathcal{T}'$, and $X_i^2 \cup X_i^3$ is a cluster of at most one of $\mathcal{T}$ and $\mathcal{T}'$. But by Lemma 6.6, Y_j is a cluster of $\mathcal{T}$ and $\mathcal{T}'$; a contradiction. Hence $X_i^3 \cap Y_j = \emptyset$, and so $Y_j \subset Y_i$. Thus (X_i^2, Y_i^2) and (X_j, Y_j) satisfy (II).

For all three cases, it now follows that

$$((X_1, Y_1), (X_2, Y_2), \ldots, (X_{i-1}, Y_{i-1}), (X_i^2, Y_i^2), (X_{i+1}, Y_{i+1}), \ldots, (X_k, Y_k))$$

verifies the nested subtree property for G, thereby establishing the proposition. □

We are now in a position to present the algorithm Construct Level-1 Network that constructs a level-1 network whose display set is a given set of phylogenetic trees if such a network exists.

Algorithm 2: Construct Level-1 Network (Part 1)

Input: A collection $\mathcal{P}$ of phylogenetic X-trees.
Output: A level-1 network $\mathcal{N}$ on X with $T(\mathcal{N}) = \mathcal{P}$ if such a network exists or, otherwise, a statement saying that no such network exists.

$k \leftarrow \log_2 |\mathcal{P}|$
if k is not a non-negative integer **then**
  return "There is no level-1 network on X whose display set is $\mathcal{P}$."
if $k = 0$ **then**
  return the unique element in $\mathcal{P}$.
Construct the rSPR graph G of $\mathcal{P}$ and, for each pair $\mathcal{T}, \mathcal{T}' \in \mathcal{P}$ with $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$, compute $M(\mathcal{T}, \mathcal{T}')$.
if G is not isomorphic to Q_k **then**
  return "There is no level-1 network on X whose display set is $\mathcal{P}$."
for $i \leftarrow 1$ **to** k **do**
  if there exists an ordered pair (X_i, Y_i) that verifies E_i and, for each $j \in \{1, 2, \ldots, i-1\}$, the pairs (X_i, Y_i) and (X_j, Y_j) satisfy one of (I)–(III) in the definition of the nested subtree property **then**
    set (X_i, Y_i) to be such an ordered pair
  else
    return "There is no level-1 network on X whose display set is $\mathcal{P}$."
Set $((X_1^0, Y_1^0), \ldots, (X_k^0, Y_k^0))$ to be a permutation of $((X_1, Y_1), \ldots, (X_k, Y_k))$ such that, for each pair $i, j \in \{1, 2, \ldots, k\}$ with $i < j$ either $Y_i^0 \cap Y_j^0 = \emptyset$ or $Y_i^0 \subset Y_j^0$.
$\mathcal{P}_0 \leftarrow \mathcal{P}$ and $X_0 \leftarrow X$
for $i \leftarrow 1$ **to** k **do**
  $X_i \leftarrow (X_{i-1} - Y_i^{i-1}) \cup \{y_i\}$
  Obtain $\mathcal{P}_i$ from $\mathcal{P}_{i-1}$ by replacing each $\mathcal{T} \in \mathcal{P}_{i-1}$ with the phylogenetic tree on X_i resulting from a subtree reduction on Y_i^{i-1} with replacement leaf y_i.
  for $j \leftarrow 1$ **to** k **do**
    if $Y_i^{i-1} \subseteq X_j^{i-1}$ **then**
      $(X_j^i, Y_j^i) \leftarrow ((X_j^{i-1} - Y_i^{i-1}) \cup \{y_i\}, (Y_j^{i-1} - Y_i^{i-1}) \cup \{y_i\})$
    else if $Y_i^{i-1} \subset Y_j^{i-1}$ and $X_j^{i-1} \cap Y_i^{i-1} = \emptyset$ **then**
      $(X_j^i, Y_j^i) \leftarrow (X_j^{i-1}, (Y_j^{i-1} - Y_i^{i-1}) \cup \{y_i\})$
    else
      $(X_j^i, Y_j^i) \leftarrow (X_j^{i-1}, Y_j^{i-1})$

Algorithm 2: CONSTRUCT LEVEL-1 NETWORK (Part 2)

26 $\mathcal{T} \leftarrow$ the unique phylogenetic X_k-tree in $\mathcal{P}_k$
27 $\mathcal{N}_k \leftarrow \mathcal{T}$
28 $i \leftarrow k$
29 **repeat**
30 $\quad \mathcal{T} \leftarrow$ an element in $\mathcal{P}_{i-1}$
31 $\quad$ Obtain $\mathcal{N}_i'$ from $\mathcal{N}_i$ by replacing the leaf labelled y_i with $\mathcal{T}|Y_i^{i-1}$.
32 $\quad$ Set $\mathcal{S}$ to be an element in $\mathcal{P}_{i-1}$ such that $\mathcal{T}|Y_i^{i-1} \not\cong \mathcal{S}|Y_i^{i-1}$.
33 $\quad$ Set u to be the vertex of $\mathcal{T}$ such that $X_i^{i-1} \subset C_{\mathcal{T}}(u)$ and no child w of u in $\mathcal{T}$ satisfies $X_i^{i-1} \subset C_{\mathcal{T}}(w)$.
34 $\quad$ Set u' to be the vertex of $\mathcal{S}$ such that $X_i^{i-1} \subset C_{\mathcal{S}}(u')$ and no child w of u' in $\mathcal{S}$ satisfies $X_i^{i-1} \subset C_{\mathcal{S}}(w)$.
35 $\quad$ **if** $(C_{\mathcal{T}}(u) - X_i^{i-1}) \cap (C_{\mathcal{S}}(u') - X_i^{i-1}) = \emptyset$ or $C_{\mathcal{S}}(u') \subseteq C_{\mathcal{T}}(u)$ **then**
36 $\quad\quad$ set $\mathcal{N}_{i-1}$ to be the network obtained from $\mathcal{N}_i'$ by subdividing the arc directed into the (unique) vertex whose cluster is $C_{\mathcal{S}}(u') - X_i^{i-1}$ with a new vertex v, subdividing the arc directed into the (unique) vertex whose cluster is X_i^{i-1} with a new vertex v', and adding the new arc (v, v')
37 $\quad$ **else if** $C_{\mathcal{T}}(u) \subseteq C_{\mathcal{S}}(u')$ **then**
38 $\quad\quad$ set $\mathcal{N}_{i-1}$ to be the network obtained from $\mathcal{N}_i'$ by subdividing the arc directed into the (unique) vertex whose cluster is $C_{\mathcal{S}}(u')$ with a new vertex v, subdividing the arc directed into the (unique) vertex whose cluster is X_i^{i-1} with a new vertex v', and adding the new arc (v, v')
39 $\quad i \leftarrow i - 1$
40 **until** $i < 1$
41 $\mathcal{N} \leftarrow \mathcal{N}_0$
42 **return** $\mathcal{N}$

Theorem 6.8. *Let $\mathcal{P}$ be a set of phylogenetic X-trees. Then* CONSTRUCT LEVEL-1 NETWORK *correctly decides if $\mathcal{P}$ is the display set of a level-1 network on X and, if so, reconstructs such a network. Moreover the running time of the algorithm is $O(|\mathcal{P}|^2|X|^2)$.*

Proof. Let G be the rSPR graph of $\mathcal{P}$. Suppose that $G \cong Q_k$ for some non-negative integer k. Let $\{E_1, E_2, \ldots, E_k\}$ be the partition of the edge set of G such that each E_i with $i \in \{1, 2, \ldots, k\}$ corresponds to the i-th bit edge subset of Q_k. By Proposition 6.4, this partition is well defined. Hence, it follows from Theorem 5.1 that CONSTRUCT LEVEL-1 NETWORK correctly decides whether or not $\mathcal{P}$ is the display set of a level-1 network on X. That is, the algorithm completes Lines 1–13 without returning "There is no level-1 network on X whose display set is $\mathcal{P}$" if and only if $\mathcal{P}$ is the display set of a level-1 network on X. Now suppose that the algorithm completes Lines 1–13 without returning "There is no level-1 network on X whose display set is $\mathcal{P}$". It then follows from the construction given in the inductive proof of Lemma 5.6 that Lines 14–42 of CONSTRUCT LEVEL-1 NETWORK correctly reconstruct a level-1 network whose display set is $\mathcal{P}$.

In preparation for the running time analysis, we discuss implementation details and suitable data structures next. For all directed and undirected graphs, we use adjacency lists to store arcs (resp. edges), and red-black trees to perform binary set operations and comparisons in time at most $O(|X|\log|X|)$. For details about these data structures, we refer the interested reader to [9]. Several steps in Construct Level-1 Network involve finding clusters both in phylogenetic trees and phylogenetic networks. Let $\mathcal{N}$ be a level-1 network on X. Asano et al. [3, Theorem 4] show how to compute a certain cluster representation of $\mathcal{N}$ in time $O(|X|)$. Roughly speaking, the authors describe a clever way to number the leaves of $\mathcal{N}$ such that each cluster can be described by a discrete interval. A similar approach was previously taken by Day [10] to obtain efficient representations for clusters in phylogenetic trees. Using this interval-based representation, we can perform each of the following operations in time at most $O(|X|)$, which we will freely use throughout the remainder of the proof.

(i) Find a cluster $C_{\mathcal{N}}(u)$ of a vertex u in $\mathcal{N}$.
(ii) For a given subset $X' \subseteq X$, decide if there is a vertex u in $\mathcal{N}$ with $C_{\mathcal{N}}(u) = X'$ and in the case of existence also find all vertices with this property.
(iii) For a given subset $X' \subseteq X$, find a vertex u in $\mathcal{N}$ with $X' \subset C_{\mathcal{N}}(u)$ and no child w of u satisfies $X' \subset C_{\mathcal{N}}(w)$.

Note that (i)–(iii) also apply to trees as every phylogenetic tree is a level-1 network.

It remains to show that the running time of Construct Level-1 Network is $O(|\mathcal{P}|^2|X|^2)$. We first bound the number of arcs of a level-1 network by a function that only depends linearly on $|X|$. Since any level-1 network on X is also tree child, it follows from [8] that such a network has at most $|X|-1$ reticulations. Hence, as each level-1 network $\mathcal{N}_i$ (resp. $\mathcal{N}_i'$) with $i \in \{0,1,2,\ldots,k\}$ that is reconstructed in Lines 14–42 of the algorithm has at most k reticulations, it follows from [25, Lemma 2.1] that $\mathcal{N}_i$ (resp. $\mathcal{N}_i'$) has at most

$$3k + 2|X| - 2 \leq 3(|X|-1) + 2|X| - 2 = 5|X| - 5$$

arcs. Now, noting that each of Lines 1–5, 8, 15, 26–30 and 39–42 takes time $O(1)$, we next detail the running time of the remaining steps.

Line 6. Let $\mathcal{T}$ and $\mathcal{T}'$ be two phylogenetic X-trees. As mentioned at the end of Section 3, it can be checked in polynomial time if $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = 1$. A straightforward way to implement this check is as follows. For each arc (u,v) in $\mathcal{T}$, let $X_v = C_{\mathcal{T}}(v)$. Then check if there is a vertex v' in $\mathcal{T}'$ with $C_{\mathcal{T}'}(v') = X_v$, $\mathcal{T}|X_v \cong \mathcal{T}'|X_v$ and $\mathcal{T}|(X - X_v) \cong \mathcal{T}'|(X - X_v)$. Since deciding if two phylogenetic trees are isomorphic can be done in $O(|X|)$, e.g. with the algorithm given by Gusfield [16], the above can be implemented such that the check if $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = 1$ takes time $O(|X|^2)$ in total. Hence, it takes the same time to compute $M(\mathcal{T},\mathcal{T}')$ and time $O(|\mathcal{P}|^2|X|^2)$ to reconstruct the rSPR graph G of $\mathcal{P}$ together with the set of moving subtrees for each edge.

Line 7. Checking if G is isomorphic to Q_k takes time $O(|\mathcal{P}|\log_2|\mathcal{P}|)$ [6].

Lines 9–13. Given a moving subtree X', it takes time $O(|X|)$ to compute the ordered pair (X', Y') and testing two such pairs for equality takes time $O(|X| \log |X|)$. Now, recall that $|E_i| = 2^{k-1}$ for each $i \in \{1, 2, \ldots, k\}$. Then, as the number of moving subtrees for two phylogenetic trees is at most three (see Proposition 6.1), it takes time $O(2^{k-1}|X| \log |X|)$ to compute all ordered pairs that verify E_i for each $i \in \{1, 2, \ldots, k\}$. Hence, by Proposition 6.7 it takes time $O(2^{k-1}k|X| \log |X|)$ to compute a sequence

$$((X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k))$$

of ordered pairs that pairwise satisfy one of (I)–(III) in the definition of the nested subtree property if G has the nested subtree property. With $k = \log_2 |\mathcal{P}|$, it follows that Lines 9–13. take time $O((|X| \log |X|)(|\mathcal{P}| \log |\mathcal{P}|))$.

Line 14. We can obtain the permutation in Line 14 by $O(k^2)$ comparisons that each involve checking if $Y_i \cap Y_j = \emptyset$ or $Y_i \subset Y_j$ in time $O(|X| \log |X|)$. If both of these checks fail, we swap the positions of (X_i, Y_i) and (X_j, Y_j). Hence, Line 14 takes time $O(k^2|X| \log |X|)$ which, with $k = \log_2 |\mathcal{P}|$, is $O((\log |\mathcal{P}|)^2|X| \log |X|)$.

Lines 16–25. Line 17 takes time $O(|X| \log |X|)$, Line 18 takes time $O(|\mathcal{P}||X|)$, and Lines 19–25 take time $O(k|X| \log |X|)$. The outer loop is executed k times and, so, Lines 16–25 take time $O(k|X| \log |X| + k|\mathcal{P}||X| + k^2|X| \log |X|)$. Since $k = \log_2 |\mathcal{P}|$ that is $O((|X| \log |X|)(|\mathcal{P}| \log |\mathcal{P}|))$.

Lines 29–40. Each of Lines 31, 33, 34, 36 and 38 takes time $O(|X|)$, each of Lines 35 and 37 takes time $O(|X| \log |X|)$, and Line 32 takes time $O(|\mathcal{P}||X|)$ since Y_i^{i-1} is a cluster of $\mathcal{S}$ and $\mathcal{T}$ by construction and non-isomorphism between two phylogenetic trees can be checked in time $O(|X|)$ [16]. The loop is executed k times and, so, Lines 29–40 take time $O(k|\mathcal{P}||X| \log |X|)$, that is again $O((|X| \log |X|)(|\mathcal{P}| \log |\mathcal{P}|))$.

It now follows that Line 6 is the most time-consuming step and Construct Level-1 Network takes time $O(|\mathcal{P}|^2|X|^2)$ as claimed. This completes the proof of the theorem. □

We remark that Whidden and Matsen [32] have shown that the rSPR graph for a collection $\mathcal{P}$ of phylogenetic X-trees can be computed in time $O(|\mathcal{P}||X|^2)$. If their result can be extended to not only deciding if $d_{\text{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$ for any pair $\mathcal{T}, \mathcal{T}' \in \mathcal{P}$ but, additionally, to compute the set $M(\mathcal{T}, \mathcal{T}')$ in the same time, then the running time of Construct Level-1 Network can be improved further since Line 6 is the current bottleneck in the running time analysis (see the proof of Theorem 6.8).

The next corollary shows that we cannot only reconstruct a level-1 network $\mathcal{N}$ for a collection $\mathcal{P}$ of phylogenetic trees such that $T(\mathcal{N}) = \mathcal{P}$ if such a network exists but, in fact, reconstruct all level-1 networks that have this property. Suppose that the rSPR graph G of $\mathcal{P}$ is isomorphic to Q_k for some non-negative integer k. Then by iterating over all sequences of ordered pairs $((X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k))$

that verify the nested subtree property of G, the next corollary is a consequence of Lemma 5.6 and Theorem 6.8.

Corollary 6.9. *Let G be the rSPR graph of a collection of phylogenetic X-trees such that $G \cong Q_k$ for some non-negative integer k. For each E_i with $i \in \{1, 2 \ldots, k\}$, let n_i the number of ordered pairs that verify E_i. If G has the nested subtree property, then there are $\Pi_{i=1}^{k} n_i$ level-1 networks on X with no trivial reticulation whose display set is $\mathcal{P}$. Moreover, each such network can be reconstructed in polynomial time.*

Acknowledgements. We thank the referee for their constructive comments.

References

[1] Allen, B. L., Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics, 5:1–15.

[2] Allman, E., Baños, H., Rhodes, J. A. (2019). NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms for Molecular Biology, 14:24.

[3] Asano, T., Jansson, J., Sadakane, K., Uehara, R., Valiente, G. (2012). Faster computation of the Robinson–Foulds distance between phylogenetic networks. Information Sciences, 197:77–90.

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

[5] Barrat-Charlaix, P., Vaughan, T., Neher, R. A. (2022). TreeKnit: Inferring ancestral reassortment graphs of influenza viruses. PLoS Computational Biology, 18:e1010394.

[6] Bhat, K. V. S. (1980). On the complexity of testing a graph for n-cube. Information Processing Letters, 11:16–19.

[7] Bordewich, M., Semple, C. (2005). On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combinatorics, 8:409–423.

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

[9] Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C. (2009). Introduction to Algorithms, 3rd edition. The MIT Press.

[10] Day, W. H. (1985). Optimal algorithms for comparing trees with labeled leaves. Journal of Classification, 2:7–28.

[11] Francis, A., Steel, M. (2015). Which phylogenetic networks are merely trees with additional arcs? Systematic Biology, 64:768–777.

[12] Goloboff, P. A. (2008). Calculating SPR distance between trees. Cladistics, 24:591–597.

[13] Gordon, K., Ford, E., St. John, K. (2013). Hamiltonian walks of phylogenetic treespaces. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10:1076–1079.

[14] Gray, F. (1953). Pulse code communication. U.S. Patent 2,632,058.

[15] R. P. Grimaldi (2003). Discrete and Combinatorial Mathematics: An Applied Introduction, 5th edition. Pearson, Addison Wesley.

[16] Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. Networks, 21:19–28.

[17] Hein, J., Jiang, T., Wang, L., Zhang, K. (1996). On the complexity of comparing evolutionary trees. Discrete Applied Mathematics, 71:153–169.

[18] Huson, D., Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic Biology, 61:1061–1067.

[19] T. N. D. Huynh, J. Jansson, N. B. Nguyen, W.-K. Sung. Constructing a smallest refining galled phylogenetic network. In: RECOMB 2005, Lecture Notes in Bioinformatics 3500, pp. 265–280.

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

[21] van Iersel, L., Janssen, R., Jones, M., Murakami, Y., Zeh, N. (2022). A practical fixed-parameter algorithm for constructing tree-child networks from multiple binary trees, Algorithmica, 84:917–960.
[22] Kong, S., Pons, J. C., Kubatko, L., Wicke, K. (2022). Classes of explicit phylogenetic networks and their biological and mathematical significance. Journal of Mathematical Biology, 84:47.
[23] Laborde, J.-M., Hebbare, S. P. R. (1982). Another characterization of hypercubes. Discrete Mathematics, 39:161–166
[24] Linz, S., Semple, C. (2022). Non-essential arcs in phylogenetic networks. Journal of Computer and System Sciences, 128:1–17.
[25] McDiarmid, C., Semple, C., Welsh D. (2015). Counting phylogenetic networks. Annals of Combinatorics 19:205–224.
[26] Mütze, T. (2022). Combinatorial Gray codes—an updated survey, arXiv:2202.01280.
[27] Robinson, D. F. (1971). Comparison of leaf labeled trees with valency three. Journal of Combinatorial Theory, 11:105–119.
[28] St. John, K. (2017). The shape of phylogenetic treespace. *Systematic Biology*, 66:e83–e94.
[29] Semple, C., Steel, M. (2003). Phylogenetics. Oxford University Press.
[30] Simpson, J. R. (2019). Tree structure in phylogenetic networks. PhD thesis. University of Canterbury.
[31] Solís-Lemus, C., Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genetics, 12:e1005896.
[32] Whidden, C., Matsen, F. A. (2018). Efficiently inferring pairwise subtree prune-and-regraft adjacencies between phylogenetic trees. In: Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics, pp. 77–91.
[33] Willson, S. J. (2012). Tree-average distances on certain phylogenetic networks have their weights uniquely determined. Algorithms for Molecular Biology, 7:13.
[34] Wu, Y. (2010). Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. Bioinformatics, 26:i140–i148
[35] L. Zhang (2016). On tree-based phylogenetic networks (2016). Journal of Computational Biology, 23:553–565.

Department of Computer Science, University of Tübingen, Germany

Current address: School of Computer Science, University of Auckland, Auckland, New Zealand

Email address: janosch.doecker@auckland.ac.nz

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

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

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

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