Journal of Classification (2019) 36:200–231
DOI: 10.1007/s00357-018-9279-5

Hierarchies from Lowest Stable Ancestors in Nonbinary Phylogenetic Networks

Katharina T. Huber

University of East Anglia, UK

Vincent Moulton

University of East Anglia, UK

Taoyang Wu

University of East Anglia, UK

Abstract: The reconstruction of the evolutionary history of a set of species is an important problem in classification and phylogenetics. Phylogenetic networks are a generalization of evolutionary trees that are used to represent histories for species that have undergone reticulate evolution, an important evolutionary force for many organisms (e.g. plants or viruses). In this paper, we present a novel approach to understanding the structure of networks that are not necessarily binary. More specifically, we define the concept of a closed set and show that the collection of closed sets of a network forms a hierarchy, and that this hierarchy can be deduced from either the subtrees or subnetworks on all 3-subsets. This allows us to also show that closed sets generalize the concept of the SN-sets of a binary network, sets which have proven very useful in elucidating the structure of binary networks. We also characterize the minimal closed sets (under set inclusion) for a special class of networks (2-terminal networks). Taken together, we anticipate that our results should be useful for the development of new phylogenetic network reconstruction algorithms.

Keywords: Phylogenetic network; Hierarchy; Lower Stable Ancestor; Nonbinary network.

We would like to thank the two anonymous referees for their helpful and constructive comments on a previous version of this paper.

Corresponding Author's Address: T. Wu, School of Computing Sciences, University of East Anglia, UK, email: taoyang.wu@uea.ac.uk.

Published online: 16 November 2018

1. Introduction

Phylogenetic networks are a generalization of evolutionary trees which are used by biologists to represent the evolution of a collection of species X with a reticulate evolutionary history. Essentially, a phylogenetic network N is a rooted, directed acyclic graph (or DAG), with a single root and leaf set labeled by the species in X (see Figure 1) for an example of such a network with leaf-set $X = \{1, 2, \ldots, 8\}$). Internal vertices in N represent ancestors of the species in X, with the root representing the highest common ancestor of all species in X. Those internal vertices which are the child of a single vertex represent a speciation event, and those which are the child of more than one vertex a reticulate event. The latter type of event might, for example, be the hybridization of plant species to form a new hybrid species, or the recombining of viruses to form a new virus.

In recent years, there has been a great deal of work on trying to develop new methods to construct phylogenetic networks from biological data (e.g. from molecular sequences). Some recent reviews concerning phylogenetic networks and approaches to construct them include Gusfield (2014) and Huson, Rupp, and Scornavacca (2010). One approach that has proven helpful in practice is to build up phylogenetic networks from smaller trees or networks. Two specific examples relying on this approach involve building phylogenetic networks from *triplets* or from *trinets*, which are 3-leaved phylogenetic trees and networks, respectively (see next section for formal definitions). They are presented in, for example, Huber et al. (2017), Jansson, Nguyen and Sung (2006), Jansson and Sung (2006), To and Habib (2009), van Iersel and Kelk (2011), and van Iersel and Moulton (2014), and examples of their application to biological data may be found in Huber et al. (2011), Oldman et al. (2016), and van Iersel et al. (2009). These approaches aim to build *binary* phylogenetic networks from *binary* triplets or trinets, that is, networks/triplets/trinets in which the root has two children, the sum of the in-degree and out-degree for every internal vertex is equal to three and each leaf has in-degree one. To do this, they exploit an interesting interplay between certain hierarchies on the leaf-set of the network and the triplets/trinets displayed by a network. However, in practice, the assumption that the networks are binary can be restrictive since, for example, it does not allow for representing uncertainty in the order of divergence or reticulate events, and it may be necessary to allow for triplets/trinets that are not binary (Jetten and van Iersel, in press; Nakhleh, 2011).

In this paper, we consider the problem of extending some of the theory underlying phylogenetic network construction from triplets and trinets to the nonbinary setting. As we shall see, this leads to some new results concerning phylogenetic networks which provide novel insights into their

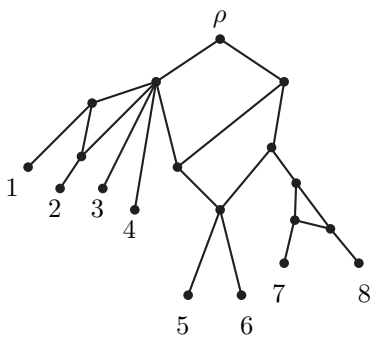

Figure 1. An example of a phylogenetic network N on $X = \{1, \ldots, 8\}$. The arcs in N (and in all subsequent figures) are all directed away from the root ρ.

structure. We expect that these results should prove useful for developing new approaches to constructing phylogenetic networks (see the last section for more details). Note that although nonbinary networks have not been commonly considered in the literature, some work has appeared on constructing (Huber and Moulton, 2012) and comparing (Cardona et al. 2011) certain nonbinary networks. In addition, some structural results have appeared concerning nonbinary tree-based networks (Jetten and van Iersel, in press).

We now summarize the contents of the rest of the paper. In the next section, we present some preliminaries concerning digraphs and phylogenetic networks, including a brief introduction to triplets and trinets. Following on from that section, we then introduce the key new concept of a *closed set*. Recall that, given a non-empty subset Y of the leaf-set X, the *lowest stable ancestor of* Y, $\text{LSA}(Y)$, in a phylogenetic network N is the lowest vertex in N that is a common ancestor of every element in Y and that is contained in every dipath that connects the root of N to some element of Y (Fischer and Huson, 2010). We say that a subset $Y \subseteq X$ is *closed (in* N*)* if $|Y| = 1$, or if $|Y| \geq 2$ and the set of leaves below $\text{LSA}(Y)$ is equal to Y. For example, $\{5, 6\}$ and $\{7, 8\}$ are both closed sets for the phylogenetic network depicted in Figure 1.

After proving some structural results concerning lowest stable ancestors in Section 3, we give a characterization of the closed sets in a phylogenetic network in terms of certain vertices in the network (Theorem 3.6). Using this characterization, in Section 4 we prove that the collection of closed sets in X for a phylogenetic network with leaf-set X is a *hierarchy* (Theorem 4.1), i.e. the collection may be represented as some rooted tree with leaf-set X. We also show that the hierarchy of closed sets is directly related to two further hierarchies on X that can be naturally associated to a

network: namely, hierarchies that arise from the cut arcs and cut vertices of N (vertices and arcs whose removal disconnects N). As we shall see, the closed set, cut arc and cut vertex hierarchies associated to a network are not the same in general, although they are identical for binary networks.

In Section 5, we consider the relationship between the closed sets of a network and the collections of triplets and trinets that it displays. For a binary phylogenetic network N, it is known that the hierarchy corresponding to the cut arcs of N may be retrieved from the collection of triplets displayed by N. One way to show this is to consider so-called SN-sets for arbitrary collections of triplets. The concept of SN-sets was introduced by Jansson and Sung (2006) as part of developing an algorithm to infer binary level-1 networks from triplet systems (see Section 2 for the definition of a level-1 network). Intuitively, each SN-set is a subset which forms the leaf-set of a subnetwork of the network which is produced by their algorithm, and hence the name SN-set ("SubNetwork-set"). These sets turned out to be very useful in elucidating the structure of binary networks in general (see e.g. To and Habib, 2009). Here, we show that the closed sets in a phylogenetic network can be obtained by extending the notion of SN-sets to the nonbinary setting. More specifically, we prove that the collection of closed sets of a phylogenetic network is precisely the collection of (generalized) SN-sets (Corollary 5.7). In addition, we show that the cut arc sets of a phylogenetic network can be obtained from its collection of trinets (Theorem 5.2), which has been proven to hold in the binary setting (van Iersel and Moulton, 2014, Theorem 1).

In Section 6, we consider a certain digraph that can be associated to the collection of trinets in a network, which we call the closure digraph. A simpler version of this digraph was considered in Oldman et al. (2016) for certain binary networks. The closure digraph is of interest since it can be used to help identify certain closed sets in a network. More specifically, using a key result concerning sink sets in the closure digraph (Corollary 6.4), in Section 7 we show that for a special class of phylogenetic networks (2-terminal networks) the sink sets in the closure digraph associated to a phylogenetic network N are precisely the minimal closed sets of N (under set inclusion). We conclude in Section 8, with a discussion of some open problems and possible future directions.

2. Preliminaries

Throughout this paper, X is a finite set with $|X| \geq 3$, unless stated otherwise. A subset $Y \subseteq X$ is called a *singleton* if $|Y| = 1$, and *non-singleton* if $|Y| \geq 2$.

Digraphs. A *directed graph*, or *digraph* for short, $N = (V, E)$, is an ordered pair consisting of a set $V = V(N)$ of *vertices* and a set $E = E(N)$ of *arcs*, that is, ordered pairs (u, v) of distinct vertices $u, v \in V$ (so in particular, there are no loops in N). Suppose N is a digraph and $u, v \in V(N)$. If (u, v) is an arc of N then we say u is a *parent* of v and v a *child* of u. The *in-degree* of u is the number of its parents, and the *out-degree* of u is the number of its children. A *root* of N is a vertex with in-degree 0. A *leaf* of N is a vertex without any children. The set of leaves of N is denoted by $L(N)$. Any vertex in N that is neither a root nor a leaf is referred to as an *interior vertex* of N.

Suppose N is a digraph. Then we call a sequence $P_{u_0,u_k} : u_0, u_1, \ldots, u_k$, $k \geq 1$, of pairwise distinct vertices of N such that (u_{i-1}, u_i) is an arc in N, $1 \leq i \leq k$, a *directed path* (or *dipath* for short) from u_0 to u_k. Moreover, we refer to the vertices u_0 and u_k as the *ends* of P_{u_0,u_k}, and all other vertices of P_{u_0,u_k} as the *interior* vertices of P_{u_0,u_k}. A pair of dipaths in N is called *openly disjoint* if they do not share a vertex other than possibly their ends. A *directed cycle* in N is a dipath P in which the requirement that the ends of P are distinct is replaced by requiring that they coincide. If N does not contain a directed cycle then N is called *acyclic*. Such a graph is sometimes referred to as a *DAG*. A DAG N is called *rooted* if it contains a unique vertex $\rho(N)$ that is the root.

Suppose N is an acyclic digraph and there exists a dipath from u to v for some $u, v \in V(N)$ distinct. Then we write it as $v \prec_N u$, and say that v is *below* u and u is an *ancestor* of v. Note that if the digraph N in question is clear from the context we simply write $v \prec u$ rather than $v \prec_N u$. Furthermore, we write $v \preceq u$ if $u = v$ or $v \prec u$ holds. Given a subset $U \subseteq V(N)$, a vertex $w \in U$ is called *lowest* if no vertex in U is below w. A *common ancestor* of a subset $Y \subseteq V(N)$ is a vertex $w \in V(N)$ that is an ancestor of each vertex in Y. Furthermore, w is called a *lowest common ancestor* of Y if it is lowest among all common ancestors of Y. If u is an interior vertex of N, then we refer to the set of leaves of N below u as the *cluster* $\mathcal{C}(u) = \mathcal{C}_N(u)$ induced by u. In case u is a leaf of N, then we put $\mathcal{C}(u) = \{u\}$.

Suppose N is a digraph. For $v \in V(N)$ we denote by $N - v$ the digraph obtained from N by removing v and all arcs incident with v. We call N *connected* if its underlying undirected graph (i.e., the graph obtained from N by discarding the directions of its arcs) is connected and *disconnected* otherwise. Note that a rooted acyclic digraph is necessarily connected. A vertex v of N is called a *cut vertex* of N if $N - v$ is disconnected. Similarly, a *cut arc* of N is an arc whose removal disconnects N. A directed graph is called *biconnected* if it contains no cut vertices. A *biconnected component* of N, also known as a block of N, is a maximal biconnected subgraph. A bi-

connected component is called *trivial* if it contains precisely one arc (which is necessarily a cut arc), and *non-trivial* otherwise. Finally, we call a vertex $v \in V(N)$ a *terminal vertex* of N if there exists a biconnected component H of N such that v is a lowest vertex in $V(H)$. Note that v could belong to several biconnected components, but at most one of them contains v as a terminal vertex. To illustrate this concept, consider the digraph N depicted in Figure 1. Then the parent vertex of 3 and 4 is not a terminal vertex of N whereas the parent vertex w of 5 and 6 is. Note that w is also contained in the biconnected components of N containing 5 and 6, respectively.

Phylogenetic networks. A *phylogenetic network N (on X)* is a rooted DAG with leaf set X and which does not contain any *degenerate* vertices (i.e., vertices in N that have in-degree and out-degree one). We also denote the leaf set of N by $L(N)$. Note that a phylogenetic network N whose underlying graph is a tree is also called a *phylogenetic tree* (cf. Semple and Steel (2003) for more details concerning phylogenetic trees). To simplify our arguments, we shall assume throughout this paper that all leaves of a phylogenetic network have in-degree one. That is, each leaf v has a unique parent, denoted by $p(v)$. Suppose N is a phylogenetic network. We refer to a vertex of N with in-degree at least two and out-degree one as *a reticulation vertex* of N. For $k \geq 0$ and integer, we call a binary phylogenetic network N *level-k* if each biconnected component of N contains at most k reticulation vertices. Note that a binary phylogenetic network N is a phylogenetic tree if and only if the level of N is zero. Hence, the level of such a phylogenetic network can be regarded as a measure of its deviation from being a phylogenetic tree.

Suppose that N is a phylogenetic network on X and $\emptyset \neq Y \subseteq X$. Extending the notion of a stable ancestor of a subset of X (see Fischer and Huson, 2010 and the Introduction) to subsets $Y \subseteq V(N) - \{\rho_N\}$, we say that a vertex $v \in V(N)$ is a *stable ancestor of Y (in N)* if v is a common ancestor of Y and is contained in every dipath that connects the root of N to some vertex $w \in Y$. Note that if u and v are two stable ancestor of Y then either $u \preceq v$ or $v \preceq u$ must hold. We refer to the unique stable ancestor $w \in V(N)$ that is lowest among all stable ancestors of Y as the *lowest stable ancestor of Y (in N)*, denoted by $\mathrm{LSA}_N(Y)$, or simply $\mathrm{LSA}(Y)$. Note that if $Y = \{y_1, \ldots, y_t\}$ for some $t \geq 1$, then we sometimes write $\mathrm{LSA}_N(y_1, \ldots, y_t)$ rather than $\mathrm{LSA}_N(Y)$. The following two easily proven facts will be useful later on.

Observation 1: Suppose that N is a phylogenetic network on X and $\emptyset \neq Y' \subseteq Y \subseteq V(N) - \{\rho_N\}$. Then $\mathrm{LSA}_N(Y') \prec_N \mathrm{LSA}_N(Y)$.

Observation 2: Suppose that N is a phylogenetic network on X and that $Y \subseteq V(N) - \{\rho_N\}$ contains at least two elements. Then there exists a pair of distinct elements y_1 and y_2 in Y such that $\mathrm{LSA}_N(y_1, y_2) = \mathrm{LSA}_N(Y)$.

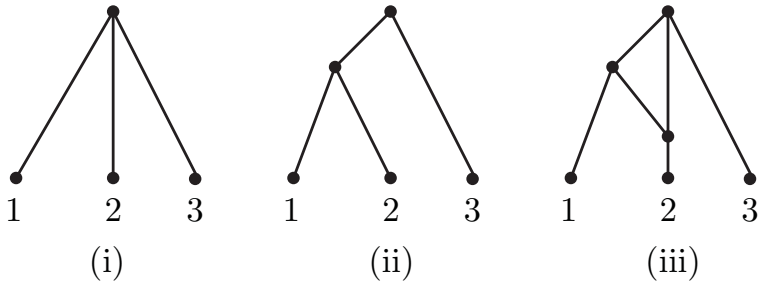

Figure 2. An example of triplets and trinets. (i) A triplet $t_1 = 1|2|3$; (ii) A triplet $t_2 = 12|3$; (iii) A trinet T on $\{1, 2, 3\}$. Here t_1 are t_2 are two triplets displayed by the network N in Figure 1, and T is a trinet induced by N.

We define the *subnet* $N|_Y$ of N on some non-empty $Y \subseteq X$ as the subgraph obtained from N by deleting all vertices that are not on any path from $\text{LSA}(Y)$ to any element in Y and subsequently suppressing all degenerate vertices and all parallel arcs. If the latter results in degenerate vertices then we repeat this whole process until we obtain a digraph containing neither parallel arcs nor degenerate vertices. Note that $N|_X = N$ if and only if $\text{LSA}(X) = \rho(N)$.

Triplets and Trinets. A phylogenetic tree T on a set $Y = \{a, b, c\}$ of size 3 is called a *triplet*. Note that T comes in two possible types. Either (1) T is binary, which implies that T contains two leaves a and b such that $\text{LSA}(a, b) \neq \rho(T)$, in which case we also write $ab|c$ for T, or (2) T is non-binary in which case $\text{LSA}(Z) = \rho(T)$, for all $Z \subseteq Y$ with $|Z| = 2$, and we also write $a|b|c$ for T. An example of these two types of triplets is depicted in Figure 2.

Suppose N is a phylogenetic network on X. Then we say that a triplet $a|b|c$ is *displayed* by N if there exists an interior vertex $r \in V(N)$, such that there exist three pairwise openly disjoint dipaths $P_{r,a}$, $P_{r,b}$, and $P_{r,c}$. Similarly, the triplet $ab|c$ is displayed by N if there exist two distinct interior vertices r and r' in N, such that there exist four pairwise openly disjoint paths $P_{r,r'}, P_{r',a}, P_{r',b}, P_{r,c}$ with r' not contained in $P_{r,c}$. We denote the collection of triplets displayed by N by $\mathcal{R}(N)$.

Let $\mathcal{R}$ be a *triplet system (on X)*, that is, a non-empty set of triplets such that $X = \bigcup_{t \in \mathcal{R}} L(t)$. Then, if Y is a subset of X, we denote by $\mathcal{R}|_Y$ the subsystem of $\mathcal{R}$ consisting of all triplets $t \in \mathcal{R}$ with $L(t) \subseteq Y$. A triplet system $\mathcal{R}$ on X is called *dense* if for each 3-subset $Y \subseteq X$ there exists at least one triplet $t \in \mathcal{R}$ for which $L(t) = Y$.

A phylogenetic network with three leaves is called a *trinet* (see Figure 2 for an example). A *trinet system* $\mathcal{T}$ on X is a non-empty set of trinets such that $\bigcup_{T\in\mathcal{T}} L(T) = X$ and there exist no distinct trinets $T, T' \in \mathcal{T}$ with $L(T) = L(T')$. A trinet system $\mathcal{T}$ on X is called *dense (on X)* if for each subset $Y \subseteq X$ with $|Y| = 3$, there exists precisely one trinet $T \in \mathcal{T}$ such that $L(T) = Y$. Note that the use of the word 'dense' for trinets is slightly different from that for triplets because a phylogenetic network induces precisely one trinet on a subset Y with three leaves but can display more than one triplet on Y; see Figure 2 for an example. For N a phylogenetic network on X, we denote by

$$\mathcal{T}(N) = \{N|_Y \,:\, Y \subseteq X \text{ and } |Y| = 3\},$$

the trinet system on X induced by N. Note that for any phylogenetic network N of X the trinet system $\mathcal{T}(N)$ induced by N is always dense on X.

3. Closed Sets

In this section, we shall give a characterization of the closed sets of a phylogenetic network N on X in terms of terminal vertices. Recall from the Introduction that a subset $A \subseteq X$ with $|A| \geq 2$ is closed (in N) if $\mathcal{C}(\text{LSA}_N(A)) = A$ holds. Note that the set X itself is necessarily closed, and that we use the convention that all singleton subsets of X are also closed.

We begin by proving a useful lemma concerning stable ancestors. To prove this lemma will use the directed point version of Menger's Theorem which we now state for the reader's convenience (for more details see, e.g. Lovász and Plummer, 1986, Theorem 2.4.1).

Theorem 3.1 *[Menger's Theorem] Suppose that D is a digraph with distinguished vertices s and t and that (s,t) is not an arc in D. Then the maximum number of pairwise openly disjoint dipaths from s to t is equal to the minimum size of a vertex set $U \subseteq V(D) - \{s,t\}$ so that each dipath from s to t contains at least one vertex in U.*

Lemma 3.2 *Suppose that N is a phylogenetic network on X, that $w \in V(N)$ is an interior vertex of N on X, and that a and b are two distinct elements in X. Putting $r = \text{LSA}_N(a,b)$, the following assertions hold.*
(i) *If there exist dipaths $P_{w,a}$ and $P_{w,b}$ from w to a and b, respectively, such that the pair $P_{w,a}$ and $P_{w,b}$ is openly disjoint then $w \preceq r$.*
(ii) *There exist dipaths $P_{r,a}$ and $P_{r,b}$ from r to a and b, respectively, such that the pair $P_{r,a}$ and $P_{r,b}$ is openly disjoint.*

Proof. (i): Let $P_{w,a}$ and $P_{w,b}$ denote dipaths from w to a and b, respectively, such that the pair $P_{w,a}$ and $P_{w,b}$ is openly disjoint. Let $P_{\rho,w}$ denote a dipath

from $\rho = \rho(N)$ to w. Then the dipath obtained by concatenating $P_{\rho,w}$ and $P_{w,a}$ is a dipath from ρ to a that contains both w and r in its vertex set. Hence, $w \preceq r$ as otherwise the definition of r implies that r is also a vertex on the dipath $P_{\rho,b}$ from ρ to b obtained by concatenating $P_{\rho,w}$ and $P_{w,b}$. Thus, $r \in (V(P_{w,a}) \cap V(P_{w,b})) - \{w\}$, which is impossible because $P_{w,a}$ and $P_{w,b}$ are openly disjoint.

(ii): Consider the digraph N' obtained from N by adding a new vertex t and two additional arcs (a,t) and (b,t). Then (r,t) is not an arc in N' and the minimum number of vertices in $V(N') - \{r,t\}$ that need to be deleted from N' so that there exists no dipaths from r to t is two. By Menger's Theorem, there exist two dipaths from r to t such that the pair formed by them is openly disjoint. Since the parents of t in N' are a and b, the construction of N' implies that there exist dipaths $P_{r,a}$ and $P_{r,b}$ in N from r to a and b, respectively, such that the pair $P_{r,a}$ and $P_{r,b}$ is openly disjoint.

■

Next, we present a characterization of the terminal vertices of a phylogenetic network. For N a phylogenetic network on X and $v, w \in V(N)$ distinct such that v is a cut vertex of N, we denote by $Z_v(w)$ the connected component of $N - v$ that contains w.

Proposition 3.3 *Suppose that N is a phylogenetic network on X and that $v \in V(N)$ is an interior vertex of N. Then v is a terminal vertex of N if and only if v is a cut vertex of N and there exists no connected component C of $N - v$ and two vertices $u_1, u_2 \in V(C)$ such that $u_1 \prec v \prec u_2$.*

Proof. Assume first that v is a terminal vertex in N. Let H denote the biconnected component of N in which v is a lowest vertex and let $v_1, \ldots, v_t \in V(N)$, $t \geq 1$, denote the children of v. We first show that v is a cut vertex of N. Since v is a lowest vertex in H, it follows that for $1 \leq i \leq t$, we have $v_i \notin V(H)$ and every path from $\rho = \rho(N)$ to v_i must contain v. In other words, there exists no path from ρ to v_i in $N - v$. Hence v is a cut vertex of N.

We next show that there exists no connected component C in $N - v$ for which there exist two vertices $u_1, u_2 \in V(C)$ such that $u_1 \prec v \prec u_2$. Let u_1 and u_2 be two vertices in N with $u_1 \prec v \prec u_2$. It suffices to show that u_2 is contained in $C' = Z_v(\rho)$ and that u_1 is not contained in C'. That u_2 is contained in C' is an immediate consequence of the fact that every parent of v is contained in C'. To see that u_1 is not contained in C' note that since $u_1 \prec v$ there must exist some $1 \leq j \leq t$ such that $u_1 \preceq v_j$. If u_1 were contained in C' then there would exist a path in C' from a parent of v to v_j. By concatenating this path with the dipath from v_j to u_1 in N, whose existence is implied by $u_1 \preceq v_j$, it follows that there exist two distinct paths from v_j to v. Hence v_j must also be contained in H which is impossible.

To see the converse, suppose that v is a cut vertex of N and that there exists no connected component C of $N - v$ and vertices $u_1, u_2 \in V(C)$ with $u_1 \prec v \prec u_2$. Assume for contradiction that v is not a terminal vertex of N. Let u_2 denote a parent of v in N and let H denote the biconnected component of N that contains the arc (u_2, v). Clearly, $v \prec u_2$. Since, by assumption, v cannot be a lowest vertex of H, there must exist a child u_1 of v that is also contained in H. Since $u_1 \prec v$ clearly holds and H is biconnected it follows that there must also exist a path from u_2 to u_1 that does not contain v. But then $u_1 \in V(Z_v(u_2)) - \{v\}$ which is impossible.
■

Corollary 3.4 *Suppose that N is a phylogenetic network on X and that $v \in V(N)$ is a terminal vertex of N. Then a vertex u in $V(N) - \{v\}$ is contained in $Z_v(\rho(N))$ if and only if u is not below v.*

Proof. Put $\rho = \rho(N)$. Suppose first that $u \in V(N) - \{v\}$ such that u is a vertex in $Z_v(\rho)$. Then since v is a terminal vertex in N and so must be an interior vertex of N, Proposition 3.3 implies that u cannot be below v.

Conversely, suppose $u \in V(N) - \{v\}$ is a vertex that is not below v. Then there must exist a dipath from ρ to u that does not contain v. Hence, $u \in V(Z_v(\rho))$.
■

Before stating our characterization of closed sets, we state one more lemma that gives a relationship between a closed set A of a phylogenetic network N and the lowest stable ancestor of A in N. Note that the lemma is trivial if in its statement the word "path" is replace by "dipath". To prove the lemma, we require some further terminology. Suppose that N is a phylogenetic network and $P : v_1, \ldots, v_k$ is an (undirected) path in N. Then we call a vertex v_i, $1 < i < k$, *alternating* if either (i) both (v_i, v_{i+1}) and (v_i, v_{i-1}) are arcs in N or (ii) both (v_{i+1}, v_i) and (v_{i-1}, v_i) are arcs in N. Clearly, P is the underlying undirected path of a dipath in N if and only if P does not contain any alternating vertex. Moreover, if A a closed subset of N and P is a path or a dipath in N, then we call P LSA(A)*-avoiding* if P does not contain LSA(A) in its vertex set.

Lemma 3.5 *Suppose that N is a phylogenetic network on X and that $A \subseteq X$ is a closed set in N. Then* LSA(A) *is a vertex in every (undirected) path that connects $\rho(N)$ to any element in A.*

Proof. Note first that without loss of generality, we may assume that $|A| \geq 2$ as otherwise the lemma clearly holds.

Suppose for contradiction that there is some $x \in A$ such that there exists a path from $\rho = \rho(N)$ to x that does not contain $w :=$ LSA(A). Let

$P^* = P_x^* : v_0 := \rho, \ldots, v_k := x$, $k \geq 1$, denote a w-avoiding path from ρ to x such that the number of alternating vertices in P^* is minimum over all w-avoiding paths between ρ and x. Without loss of generality, we may assume that x is chosen in a way so that the number m of alternating vertices in P^* is minimum over all possible w-avoiding paths between ρ and any element in A.

We show first that $m \geq 2$. Since P^* is w-avoiding and, by the definition of the lowest stable ancestor of a set, every dipath from ρ to x contains w, it follows that P^* is not a dipath in N. Hence, P^* contains at least one alternating vertex. Since neither one of the end vertices of P^* can be alternating, $k \geq 2$ must hold. Hence, (v_0, v_1) and (v_{k-1}, v_k) are two distinct arcs in N and, so, m must be even. Consequently, $m \geq 2$.

We next show that the m-th alternating vertex of P^* is below w. Let $0 < a < b < k$ be such that when starting with v_0 the vertices v_a and v_b are the $(m-1)$-th and m-th alternating vertices in P^*, respectively. Then v_a and v_b are alternating and v_i is not alternating for all $a+1 \leq i \leq k$ and $i \neq b$. Since (v_{k-1}, v_k) is an arc in N, it follows that $v_b, v_{b+1}, \ldots, v_k$ is a dipath from v_b to $v_k = x$. Note that, by the choice of a and b, no vertex on the path $P : v_b, v_{b-1}, \ldots, v_a$ from v_b to v_a can be alternating. Thus, P is a dipath from v_b to v_a. Since N is a rooted DAG, there must exist a dipath $P' : u_1 := \rho, \ldots, u_t := v_b$ from ρ to v_b. Consequently, $u_1, \ldots, u_t, v_{b+1}, \ldots, v_k$ is a dipath from ρ to x. By the definition of a stable ancestor it follows that there exists some $1 \leq i < t$ such that $w = u_i$. Thus, $v_b \prec w$.

Let $y \in X$ denote a leaf of N below v_a and let $w_1 := v_a, w_2, \ldots, w_j := y$ denote an w-avoiding dipath from v_a to y which must exist since $y \prec v_a \prec w$. Hence, $y \in \mathcal{C}(w) = A$, as A is closed. Let P'' denote the path obtained from $v_1, \ldots, v_a, w_2, \ldots, w_j$ by first ignoring directions and then removing all cycles (in case there exist any). Then P'' is a w-avoiding path from ρ to y that contains at most $m-1$ alternating vertices, which contradicts the choice of P^*.

■

We now state our characterization of the closed sets of a phylogenetic network.

Theorem 3.6 *Suppose that N is a phylogenetic network on X and that $A \subseteq X$ is a subset with $2 \leq |A| < |X|$. Then the following statements are equivalent:*
(i) *A is closed in N.*
(ii) *$\mathrm{LSA}(A)$ is a terminal vertex of N and A is closed in N.*
(iii) *there exists a terminal vertex v in N with $A = \mathcal{C}(v)$.*

Proof. (i)⇒(ii): Suppose A is closed in N. It clearly suffices to show that $\mathrm{LSA}(A)$ is a terminal vertex of N. In view of Proposition 3.3, we need to show that (a) $\mathrm{LSA}(A)$ is a cut vertex of N, and (b) there exist no connected component C' of $N - \mathrm{LSA}(A)$ and no two vertices u and u' in C' such that $u \prec \mathrm{LSA}(A) \prec u'$.

To see (a), let $x \in A$. Then, by Lemma 3.5, every path between $\rho := \rho(N)$ and x in N contains $\mathrm{LSA}(A)$. Note that the assumption on $|A|$ implies that $x \neq \mathrm{LSA}(A) \neq \rho$. Hence there is no path in $N - \mathrm{LSA}(A)$ joining ρ and x. Thus, $\mathrm{LSA}(A)$ is a cut vertex of N.

To see (b), put $N' := N - \mathrm{LSA}(A)$. Note that each vertex u in N with $\mathrm{LSA}(A) \prec u$ is contained in $C := Z_{\mathrm{LSA}(A)}(\rho)$ because there exists a dipath from ρ to u that does not contain $\mathrm{LSA}(A)$. We claim that no vertex v in C is below $\mathrm{LSA}(A)$. Indeed, if $v \prec \mathrm{LSA}(A)$ held for some vertex $v \in V(C)$, then there would exist some element $x \in \mathcal{C}(v) \subseteq \mathcal{C}(\mathrm{LSA}(A)) = A$ which is contained in C. Hence, there would exist a path between ρ and x in N' that does not contain $\mathrm{LSA}(A)$, which is a contradiction to Lemma 3.5 since A is closed.

(ii)⇒(iii): This is trivial.

(iii)⇒(i): Assume that v is a terminal vertex of N such that $A = \mathcal{C}(v)$. Then if there exists a dipath from the root ρ to an element $x \in A$ that does not contain v, then ρ and x belong to the same connected component in $N - v$. But this is impossible in view of Proposition 3.3 and the assumption on v. Hence, v must be a stable ancestor of A. Thus, $\mathrm{LSA}(A) \preceq v$ and, so, A is closed in view of $A \subseteq \mathcal{C}(\mathrm{LSA}(A)) \subseteq \mathcal{C}(v) = A$.
■

4. Hierarchies from Networks

A collection $\mathcal{H}$ of subsets of X is called a *hierarchy (on X)* if $A \cap B \in \{\emptyset, A, B\}$ holds for all $A, B \in \mathcal{H}$, and $\mathcal{H}$ contains X and all singletons of X, but not the empty set. In this section, we shall show that the set $\mathcal{H}_{Cl}(N)$ of all closed sets in a network N forms a hierarchy. We shall also show that this hierarchy is closely related to some other hierarchies on X that can be associated to N.

Various ways have been described for associating a hierarchy to a phylogenetic network N on X, two of which we now recall (see, e.g. Dress, Moulton, Steel and Wu (2010) for several examples). The first way concerns the cut arcs of the network. More specifically, define a subset $A \subseteq X$ to be a *cut-arc set (in N)* if either $A = X$ or there exists a cut arc $a = (u, v)$ in N with $u, v \in V(N)$ such that $A = \mathcal{C}(v)$. Clearly, the set $\mathcal{H}_{CA}(N)$ of all cut-arc sets in N is a hierarchy on X, an observation which also follows the result we prove below.

A second way to associate a hierarchy on X to N is via its cut vertices. Call a subset $A \subseteq X$ a *cut-vertex set (of N)* if either $A = X$ or there exists a cut vertex v in N such that A is the leaf set of a connected component of $N - v$ distinct from $Z_v(\rho(N))$. Let $\mathcal{H}_{CV}(N)$ denote the set of cut-vertex sets of N. It is again straight-forward to check that $\mathcal{H}_{CV}(N)$ is a hierarchy on X and that $\mathcal{H}_{CA}(N) \subseteq \mathcal{H}_{CV}(N)$.

Interestingly, even though Theorem 3.6 suggests a close relationship between $\mathcal{H}_{Cl}(N)$ and $\mathcal{H}_{CV}(N)$, this relationship is not in terms of set inclusion since, in general, $\mathcal{H}_{CV}(N)$ is neither a subset nor a superset of $\mathcal{H}_{Cl}(N)$. However, we now introduce a superset $\mathcal{H}^*_{CV}(N)$ of $\mathcal{H}_{CV}(N)$ which we shall show below to be a hierarchy that contains both $\mathcal{H}_{Cl}(N)$ and $\mathcal{H}_{CV}(N)$.

More specifically, we define a subset $A \subseteq X$ to be contained in $\mathcal{H}^*_{CV}(N)$ if either $A \in \mathcal{H}_{CV}(N)$, or there exists a cut vertex v of N such that $A = X - V(Z_v(\rho(N)))$. Since each cut arc of N is incident with a cut vertex of N, it is clear that $\mathcal{H}_{CA}(N) \subseteq \mathcal{H}_{CV}(N) \subseteq \mathcal{H}^*_{CV}(N)$ all hold. To illustrate these concepts, consider the network N on $X = \{1, 2, \ldots, 8\}$ depicted in Figure 1. Let $\mathcal{H}$ be the collection of singletons of X and the set X. Then $\mathcal{H}_{CA}(N) = \mathcal{H} \cup \{\{7, 8\}\}$, $\mathcal{H}_{CV}(N) = \mathcal{H}_{CA}(N) \cup \{\{1, 2\}\}$, $\mathcal{H}_{Cl}(N) = \mathcal{H} \cup \{\{5, 6\}, \{7, 8\}\}$ and $\mathcal{H}^*_{CV}(N) = \mathcal{H}_{CV}(N) \cup \{\{1, 2, 3, 4\}, \{5, 6\}\}$.

Theorem 4.1 *Suppose that N is a phylogenetic network on X. Then $\mathcal{H}_{CA}(N)$, $\mathcal{H}_{Cl}(N)$ and $\mathcal{H}^*_{CV}(N)$ are all hierarchies and*

$$\mathcal{H}_{CA}(N) \subseteq \mathcal{H}_{Cl}(N) \subseteq \mathcal{H}^*_{CV}(N)$$

holds. In addition, if N is binary, then $\mathcal{H}_{CA}(N) = \mathcal{H}_{Cl}(N) = \mathcal{H}^*_{CV}(N)$.

Proof. Clearly $\mathcal{H}_{Cl}(N)$ is a hierarchy on X and, by the above, $\mathcal{H}_{CV}(N)$ is also a hierarchy on X. We break the remainder of the proof into a series of claims. We first claim that $\mathcal{H}^*_{CV}(N)$ is a hierarchy. Let $A_1, A_2 \subseteq X$ denote two distinct elements in $\mathcal{H}^*_{CV}(N)$. We need to show that $A_1 \cap A_2 \in \{\emptyset, A_1, A_2\}$. Without loss of generality, we may assume that $1 < |A_1|, |A_2| < |X|$. For $i = 1, 2$, let v_i be the cut vertex of N associated to A_i as described in the definition of $\mathcal{H}^*_{CV}(N)$ and put $Z_i = Z_{v_i}(\rho)$ where $\rho = \rho(N)$. We consider three cases which reflect the three possible relationships between v_1 and v_2:

Case (1) $v_1 = v_2$: Then $Z_1 = Z_2$. Since $A_1 \neq A_2$ and $\mathcal{H}_{CV}(N)$ is a hierarchy, it suffices to consider the cases that either there exists some $i \in \{1, 2\}$, $i = 1$ say, such that $A_1 \in \mathcal{H}_{CV}(N)$ and $A_2 \notin \mathcal{H}_{CV}(N)$ or $A_1, A_2 \notin \mathcal{H}_{CV}(N)$. In the first case, we have $A_1 = L(C_1)$ for some connected component C_1 of $N - v_1$ distinct from Z_1. Since $A_2 \notin \mathcal{H}_{CV}(N)$ it follows that $A_1 = L(C_1) \subseteq X - L(Z_1) = A_2$ and, so, $A_1 \cap A_2 = A_1$. In the second case, we have $A_i = X - L(Z_1)$ for all $i = 1, 2$. But then $A_1 = A_2$ which is impossible.

Case (2) One of v_1 and v_2 is below the other: Assume without loss of generality that v_2 is below v_1. Then $L(Z_1) \subseteq L(Z_2)$. There are two cases to consider: namely $A_1 \in \mathcal{H}_{CV}(N)$ or $A_1 \notin \mathcal{H}_{CV}(N)$. Suppose first that $A_1 \notin \mathcal{H}_{CV}(N)$. Then $A_1 = X - L(Z_1)$. If $A_2 \in \mathcal{H}_{CV}(N)$ then there exists some connected component C of $N - v_2$ such that $A_2 = L(C)$. Since $L(Z_1) \subseteq L(Z_2)$, we obtain $A_2 = L(C) \subseteq X - L(Z_2) \subseteq X - L(Z_1) = A_1$. If $A_2 \notin \mathcal{H}_{CV}(N)$ then $A_2 = X - L(Z_2)$. Again since $L(Z_1) \subseteq L(Z_2)$ it follows that $A_2 \subseteq A_1$. In either case we obtain $A_1 \cap A_2 = A_2$.

Now, assume $A_1 \in \mathcal{H}_{CV}(N)$. Then swapping the roles of A_1 and A_2 in the previous argument implies $A_1 \subseteq A_2$ in case $A_2 \notin \mathcal{H}_{CV}(N)$. If $A_2 \in \mathcal{H}_{CV}(N)$ then since $\mathcal{H}_{CV}(N)$ is a hierarchy on X it follows that $A_1 \cap A_2 \in \{\emptyset, A_1, A_2\}$.

Case (3) Neither v_1 is below v_2 nor v_2 is below v_1: If $A_1 \in \mathcal{H}_{CV}(N)$ then $A_1 \cap A_2 \in \{\emptyset, A_1, A_2\}$ follows in case $A_2 \in \mathcal{H}_{CV}(N)$ because $\mathcal{H}_{CV}(N)$ is a hierarchy. If $A_2 \notin \mathcal{H}_{CV}(N)$ then $A_2 = X - L(Z_2)$. By assumption on v_1 and v_2, it follows that $A_1 \subseteq X - L(Z_2) = A_2$. Thus, $A_1 \cap A_2 = A_1$. So assume $A_1 \notin \mathcal{H}_{CV}(N)$. Then swapping the roles of A_1 and A_2 in the previous argument implies $A_1 \cap A_2 = A_2$ in case $A_2 \in \mathcal{H}_{CV}(N)$. So assume $A_2 \notin \mathcal{H}_{CV}(N)$. Then $A_2 = X - L(Z_2)$. Since the assumption on v_1 and v_2 implies $L(Z_1) \cup L(Z_2) = X$ it follows that $A_1 \cap A_2 = (X - L(Z_1)) \cap (X - L(Z_2)) = \emptyset$. Thus, $\mathcal{H}^*_{CV}(N)$ is a hierarchy on X, as required.

We next show that the two set-inclusions stated in the theorem hold. We start with establishing that $\mathcal{H}_{CA}(N) \subseteq \mathcal{H}_{Cl}(N)$. Suppose $A \in \mathcal{H}_{CA}(N)$. Without loss of generality, we may assume that A is neither a singleton nor X itself as otherwise the claim clearly follows. Hence, there exist vertices $u, v \in V(N)$ such that (u, v) is a cut-arc and $C(v) = A$. But then v is necessarily a terminal vertex of N. By Theorem 3.6, it follows that A is a closed set in N, as claimed.

It remains to show that $\mathcal{H}_{Cl}(N) \subseteq \mathcal{H}^*_{CV}(N)$. Suppose A in $\mathcal{H}_{Cl}(N)$. Without loss of generality, we may assume that $1 < |A| < |X|$ as otherwise the claim clearly follows again. Since A is closed in N, Theorem 3.6 implies that there exists a terminal vertex v in N such that $A = \mathcal{C}(v)$. Note that, by Proposition 3.3, v is necessarily a cut vertex of N. Let $x \in X$. Then, by Corollary 3.4, x is contained in $A = \mathcal{C}(v)$ if and only if $x \notin V(Z_v(\rho))$. Thus, $A \in \mathcal{H}^*_{CV}(N)$, as claimed.

We conclude the proof of the theorem by showing that the three set inclusions relating $\mathcal{H}_{Cl}(N)$, $\mathcal{H}_{CA}(N)$ and $\mathcal{H}^*_{CV}(N)$ become equalities in case N is binary. To see this, it suffices to show that $\mathcal{H}^*_{CV}(N) \subseteq \mathcal{H}_{CA}(N)$. Suppose N is binary and $A \in \mathcal{H}^*_{CV}(N)$. We need to show that A is a cut-arc set. Without loss of generality, we may assume that $1 < |A| < |X|$. Let v be a cut vertex in N as described in the definition of the elements in $\mathcal{H}^*_{CV}(N)$.

We consider two possible cases, where we put $Z_v := Z_v(\rho)$.

Case (a) $A \in \mathcal{H}_{CV}(N)$: Then there exists some connected component C_1 of $N - v$ distinct from Z_v such that $A = L(C_1)$. Since N is binary, v has at least one but at most two children. If v has one child then let u denote that child. Since v is a cut vertex of N, the arc (v, u) is necessarily a cut arc of N. Since $A = \mathcal{C}(u)$ clearly holds, it follows that A is a cut-arc set.

Suppose v has two children denoted u_1 and u_2. Note that $v \neq \rho$ if u_1 and u_2 are both contained in C_1. Denoting by v' the unique parent of v in this case, it follows that (v', v) is a cut arc of N. Since $A = \mathcal{C}(v)$ clearly holds, A is a cut-arc set. So assume that precisely one of u_1 and u_2, say u_1, is contained in C_1. Then (v, u_1) is a cut arc of N. Since $A = \mathcal{C}(u_1)$ it follows that A is a cut-arc set.

Case (b) $A \notin \mathcal{H}_{CV}(N)$: Then $A = X - L(Z_v)$. Since N is binary, $N - v$ either has two or three connected components. If $N - v$ has two connected components then the same arguments as in Case (a) imply that A is a cut-arc set. So assume that $N - v$ has three connected components. Then v has a unique parent v' and two children, denoted respectively by u_1 and u_2. Note that all of (v', v), (u, u_1) and (u, u_2) must be cut arcs of N. It follows that $A = L(Z_v(u_1)) \cup L(Z_v(u_2))$ as $A = X - L(Z_v)$. Consequently, $A = \mathcal{C}(v)$ and, thus, A must be a cut-arc set.

■

Before concluding this section, we note that closed sets are related to another type of hierarchy that can be related to a phylogenetic network. More specifically, recall that a cluster $C \subseteq X$ is called *tight* in a phylogenetic network N on X if there exists a subset $V_C \subseteq V(N)$ such that (i) for all $v \in V_C$, we have $\mathcal{C}(v) = C$, and (ii) V_C *separates* C from $X - C$, that is, each (undirected) path from C to $X - C$ contains some vertex in V_C. In Dress et al. (2010) it is shown that the tight clusters of a network form a hierarchy. Note that a cut-vertex set of N is not necessarily a tight cluster of N. As a direct corollary of Theorem 3.6 we however obtain:

Corollary 4.2 *Suppose that N is a phylogenetic network on X and that $A \subseteq X$ is a closed set of N. Then A is a tight cluster of N.*

5. Closed Sets from Triplets and Trinets

In this section, we shall see that the closed sets of a phylogenetic network N can be inferred from the triplet or trinet systems induced by N. We start by extending the notion of a closed set to trinet systems on X. Suppose that $\mathcal{T}$ is a trinet system on X and $A \subseteq X$ is a non-empty subset. We say that A is a *closed in* $\mathcal{T}$ if for each trinet $T \in \mathcal{T}$ either $A \cap L(T) = \emptyset$

or $A \cap L(T)$ is a closed set in T. We now show that these concepts agree in case $\mathcal{T}$ is the trinet system displayed by a phylogenetic network.

Theorem 5.1 *Suppose that N is a phylogenetic network on X and that $\emptyset \neq A \subseteq X$. Then A is a closed set in N if and only if A is closed in $\mathcal{T}(N)$.*

Proof. Without loss of generality, we may assume that $1 < |A| < |X|$ as otherwise the theorem clearly holds.

Assume first that A is closed in N. Suppose $T \in \mathcal{T}(N)$ is a trinet such that $A' := A \cap L(T) \neq \emptyset$. Note that if $|A'| = 1$, then A' is closed in T by definition. Moreover, if $|A'| = 3$ then $A' = L(T)$ and so A' is closed in T. So assume $|A'| = 2$. Let $x, y \in X$ and $z \in L(T) - A'$ where $A' := \{x, y\}$. Then $\text{LSA}_N(x, y) \preceq \text{LSA}_N(A) \prec \text{LSA}_N(x, y, z)$, where the $\preceq$ part follows from Observation 1 and the $\prec$ part holds because $\text{LSA}_N(A)$ and $\text{LSA}_N(x, y, z)$ are two common stable ancestors of x and y and hence we have either $\text{LSA}_N(A) \prec \text{LSA}_N(x, y, z)$ or $\text{LSA}_N(x, y, z) \preceq \text{LSA}_N(A)$. However, the latter case implies that $z \preceq \text{LSA}_N(A)$, a contradiction to the fact that A is closed and $z \notin A$. Therefore, $\mathcal{C}_N(\text{LSA}_N(x, y)) = \{x, y\}$ and, so, A' is closed in T.

Conversely, suppose that A is not closed in N. We need to show that there exists a trinet $T \in \mathcal{T}(N)$ such that $A \cap L(T)$ is neither empty nor a closed set in T. By Observation 2, fix $x, y \in A$ such that $\text{LSA}_N(x, y) = \text{LSA}_N(A)$. Since A is not closed in N, there must exist some $z \in X - A$ such that $z \in \mathcal{C}_N(\text{LSA}_N(A))$. Let $T \in \mathcal{T}(N)$ be such that $L(T) = \{x, y, z\}$. Clearly, $A \cap L(T) = \{x, y\} \neq \emptyset$. Moreover, $\mathcal{C}_T(\text{LSA}_T(x, y)) = \{x, y, z\} \neq \{x, y\} = A \cap L(T)$. Thus, $A \cap L(T)$ is not closed in T.
■

Using this result, we now show that the cut-arc sets of a phylogenetic network N can be reconstructed from its trinet system $\mathcal{T}(N)$. This generalizes van Iersel and Moulton (2014, Theorem 1) which considers the binary case.

Theorem 5.2 *Suppose that N is a phylogenetic network on X and that $A \subseteq X$ is a subset such that $2 < |A| < |X|$. Then A is a cut-arc set of N if and only if, for all $x, y \in A$ and $z \notin A$, the set $\{x, y\}$ is a cut-arc set of the trinet induced by N on $\{x, y, z\}$.*

Proof. Suppose that A is a cut-arc set of N. Then there exists a cut arc (u, v) in N with $\mathcal{C}(v) = A$. Let $x, y \in A$ and $z \notin A$ and consider the trinet $T \in \mathcal{R}(N)$ on $\{x, y, z\}$. Then (u, v) induces a cut arc (u', v') in T whose deletion results in two connected components one of which contains $\{x, y\}$ in its vertex set and the other z. Thus, $\{x, y\}$ is a cut-arc set of $N|_{\{x,y,z\}}$.

Conversely, suppose that, for all $x, y \in A$ and $z \notin A$, the set $\{x, y\}$ is a cut-arc set for the trinet on $\{x, y, z\}$ contained in $\mathcal{T}(N)$. Then, for all trinets $T \in \mathcal{T}(N)$, Theorem 4.1 implies that $A \cap L(T)$ is closed in T. By Theorem 5.1 it follows that A is closed in N. Hence, by Theorem 3.6, $w := \text{LSA}(A)$ is a terminal vertex in N and $A = \mathcal{C}(w)$. Let $v \in V(N)$ denote the stable ancestor of A such that $A \subsetneq \mathcal{C}(v)$ while no stable ancestor of A strictly below v has this property. Note that $w \prec v$. We claim that every dipath of N from v to w contains a cut arc of N.

Assume for contradiction that there exists a dipath in N from v to w that contains no cut arc of N. Let u be a vertex in N so that $A \subsetneq \mathcal{C}(u)$ holds and that $A \subsetneq \mathcal{C}(u')$ does not hold for all $u' \in V(N)$ below u. Then $w \prec u \preceq v$ must hold. Choose some $z \in \mathcal{C}(u) - A$, and let $x, y \in A$ such that $\text{LSA}(x, y) = w$. By Lemma 3.2 there exist dipaths from w to x and y, respectively, such that the pair formed by them is openly disjoint. Let $T \in \mathcal{T}(N)$ denote the trinet on $\{x, y, z\}$. We now show that $\{x, y\}$ is not a cut-arc set in T, a contradiction which concludes the proof of the claim. To establish this fact we consider separately the cases $u = v$ and $u \prec v$.

Suppose $u = v$ and fix a dipath P from v to z. Then by the choice of u and v if follows that except for v, none of the vertices in P is an ancestor of x or y. In addition, $v = u$ implies that all dipaths from v to w in N are contained in T, and hence v is also contained in T. Therefore, since there exists no cut arc in N between v and w, there exists no cut arc in T between v and w, and so $\{x, y\}$ is not a cut-arc set in T.

Now suppose $u \prec v$. Fix a dipath $P_{v,u}$ from v to u and a dipath $P_{u,w}$ from u to w. Let w' be the stable ancestor of A contained in $P_{u,w}$ closest to u. Without loss of generality, we may assume that $u \neq w$ as this case can be established in a similar manner. Note that, by the definition of v, we have $w' \neq u$ as u is not a stable ancestor of A. Hence, there exists a dipath $P_{v,w'}$ from v to w' that does not contain u. Starting at v, let $v' \in V(N)$ be the last vertex that is simultaneously contained in $P_{v,w'}$ and $P_{v,u}$. Then v', u, w, and w' must all be vertices of T and each of the dipaths $P_{v',w'}$, $P_{v',u}$, $P_{u,w'}$, and $P_{w',w}$ induce four dipaths in T so that none of them contains a cut arc of T. By the choice of x and y, it follows that $\{x, y\}$ is not a cut-arc set in T. This concludes the proof of the claim.

To show that A is a cut-arc set of N and thus establish the theorem, consider a cut arc (u_1, u_2) in N whose removal disconnects N into two connected components such that the vertex set of one contains w and the vertex set of the other v. Note that such a cut arc must exist by the previous claim. Then u_2 is necessarily a stable ancestor of A. Since $u_2 \prec v$ clearly holds, the choice of v implies that $A = \mathcal{C}(u_2)$. Hence A is a cut-arc set in N.

■

We now turn our attention to triplet systems. We begin by defining the notion of SN-sets for triplet systems which may contain nonbinary triplets. A subset $A \subseteq X$ is called an *SN-set* for a triplet system $\mathcal{R}$ if for $a, b \in A$ distinct and $c \in X - A$, we have $\mathcal{R}|_{\{a,b,c\}} \subseteq \{ab|c\}$, that is, $\mathcal{R}|_{\{a,b,c\}}$ is either $\{ab|c\}$ or $\emptyset$. Note that, by definition, all singletons of X and X itself are SN-sets. We will use the convention that the empty set is not an SN-set for any triplet system. For the triplet system $\mathcal{R}(N)$ displayed by the network N in Figure 1, the subsets $\{5, 6\}$ and $\{7, 8\}$ are SN-sets while $\{5, 6, 7, 8\}$ is not.

The following result is a straightforward generalization of the binary case stated in Jansson and Sung (2006, Lemma 8). Note that the assumption that the triplet system is dense (that is, it contains at least one triplet for each 3-subset) is necessary even for triplet systems that contain only binary triplets.

Lemma 5.3 *Suppose that $\mathcal{R}$ is a dense triplet system on X. Then the set of SN-sets for $\mathcal{R}$ is a hierarchy on X.*

Proof. Assume for contradiction that A and B are two SN-sets for $\mathcal{R}$ such that $A \cap B \notin \{\emptyset, A, B\}$. Then there exists $a \in A$, $b \in B$, and $c \in A \cap B$ such that $a \notin B$ and $b \notin A$. Since $\mathcal{R}$ is dense on X, there exists some $t \in \mathcal{R}$ such that $L(t) = \{a, b, c\}$. Since A and B are SN-sets it follows that $\mathcal{R}|_{\{a,b,c\}} \subseteq \{ac|b\} \cap \{bc|a\} = \emptyset$ which is impossible as $\mathcal{R}$ is dense. ■

We next characterize the SN-sets of a triplet system $\mathcal{R}$ in terms of subsets of X that are closed with respect to a certain closure operation $S_{\mathcal{R}}$ which we now introduce. Suppose that $\mathcal{R}$ is a triplet system on X and $A \subseteq X$. We put $S_{\mathcal{R}}(A) = S_{\mathcal{R}}(A \cup \{c\})$ if there exists $a, b \in A$ and $c \in X - A$ such that $a|bc \in \mathcal{R}$ or $a|b|c \in \mathcal{R}$ holds, and $S_{\mathcal{R}}(A) = A$ otherwise. Note that, by definition, $S_{\mathcal{R}}(\{x\}) = \{x\}$ and $S_{\mathcal{R}}(X) = X$.

Lemma 5.4 *Suppose that $\mathcal{R}$ is a triplet system on X and that $\emptyset \neq A \subseteq X$. Then A is an SN-set for $\mathcal{R}$ if and only if $S_{\mathcal{R}}(A) = A$.*

Proof. Since the lemma clearly holds for $|A| = 1$ and $A = X$, we may assume for the remainder of the proof that $1 < |A| < |X|$.

Suppose first that A is an SN-set for $\mathcal{R}$. Then for all $c \in X - A$ and $a, b \in A$, we have $\mathcal{R}|_{\{a,b,c\}} \subseteq \{ab|c\}$. Thus, the only triplet on $\{a, b, c\}$ contained in $\mathcal{R}$ is $ab|c$. Hence, $S_{\mathcal{R}}(A) = A$.

Conversely, suppose $S_{\mathcal{R}}(A) = A$ and assume for contradiction that A is not an SN-set for $\mathcal{R}$. Then there must exist some $c \in X - A$ such that $\mathcal{R}|_{\{a,b,c\}} \not\subseteq \{ab|c\}$. Swapping the roles of a and b if necessary, we may

assume that $a|bc \in \mathcal{R}$ or $a|b|c \in \mathcal{R}$ (or both). In either case, $S_{\mathcal{R}}(A) \subsetneq S_{\mathcal{R}}(A \cup \{c\}) = S_{\mathcal{R}}(A)$ follows which is impossible.
■

Next we show that the SN-sets associated to a dense triplet system $\mathcal{R}$ can be constructed by applying the $S_{\mathcal{R}}$ closure operations to pairs of elements of X. This generalizes Jansson and Sung (2006, Lemma 7). Note that the density assumption on $\mathcal{R}$ is necessary for Lemma 5.5 to hold.

Lemma 5.5 *Suppose that $\mathcal{R}$ is a dense triplet system on X. If $A \subseteq X$ is an SN-set for $\mathcal{R}$, then $A = S_{\mathcal{R}}(\{x, y\})$ or $A = S_{\mathcal{R}}(\{x\})$, for some $x, y \in A$.*

Proof. Without loss of generality we may assume that $|A| \geq 2$. Choose elements $a, b \in A$ such that $|S_{\mathcal{R}}(\{x, y\})| \leq |S_{\mathcal{R}}(\{a, b\})|$, for all $x, y \in A$. We claim that $A = S_{\mathcal{R}}(\{a, b\})$. Note first that $S_{\mathcal{R}}(\{a, b\}) \subseteq S_{\mathcal{R}}(A) = A$ clearly holds as A is an SN-set. Assume for contradiction that $S_{\mathcal{R}}(\{a, b\}) \neq A$. Then there exists some $c \in A - S_{\mathcal{R}}(\{a, b\})$. The definition of the $S_{\mathcal{R}}$ closure operation combined with the fact that $\mathcal{R}$ is dense implies $\mathcal{R}_{\{a,b,c\}} = \{ab|c\}$. Hence, $b \in S_{\mathcal{R}}(\{a, c\})$. Thus, $S_{\mathcal{R}}(\{a, b\}) \subsetneq S_{\mathcal{R}}(\{a, b, c\}) = S_{\mathcal{R}}(\{a, c\})$, which is impossible.
■

We now relate closed sets for trinet systems with SN-sets for triplet systems. For $\mathcal{T}$ a trinet system on X, we put $\mathcal{R}(\mathcal{T}) := \bigcup_{N \in \mathcal{T}} \mathcal{R}(N)$.

Theorem 5.6 *Suppose that $\mathcal{T}$ is a trinet system on X and $A \subseteq X$. Then A is closed in $\mathcal{T}$ if and only if A is an SN-set for $\mathcal{R}(\mathcal{T})$.*

Proof. Since the theorem holds for $|A| = 1$ and $|A| = |X|$, we may assume for the remainder of the proof that $1 < |A| < |X|$.

Suppose first that A is closed in $\mathcal{T}$. Assume for contradiction that A is not an SN-set of $\mathcal{R}(\mathcal{T})$. Then there exist elements $a, b \in A$ and $c \in X - A$ such that $\mathcal{R}|_{\{a,b,c\}} \not\subseteq \{ab|c\}$. Therefore, there exists a trinet $T \in \mathcal{T}$ on $\{a, b, c\}$ such that $\mathcal{R}(T) \not\subseteq \{ab|c\}$. Swapping the roles of a and b if necessary, we may assume without loss of generality that $ac|b \in \mathcal{R}(T)$ or that $a|b|c \in \mathcal{R}(T)$ holds. In either case, there exists a vertex $r \in V(T)$ such that $c \prec r$ and the pair formed by the dipath from r to a and the dipath from r to b is openly disjoint. By Lemma 3.2(i), it follows that $c \prec r \preceq \mathrm{LSA}_T(a, b)$. Hence, $\mathcal{C}(\mathrm{LSA}_T(a, b)) = \{a, b, c\}$, which contradicts the assumption that A is closed in $\mathcal{T}$.

Conversely, suppose that A is an SN-set of $\mathcal{R}(\mathcal{T})$. Assume for contradiction that A is not closed in $\mathcal{T}$. Then there exists $a, b \in A$, $c \in X - A$, and a trinet $T \in \mathcal{T}$ on $\{a, b, c\}$ such that $\mathcal{C}(\mathrm{LSA}_T(a, b)) = \{a, b, c\}$. Let $r = \mathrm{LSA}_T(a, b)$. Then there exists a dipath $P_{r,c}$ in T from r to c. In addition, by Lemma 3.2(ii), there exist dipaths $P_{r,a}$ and $P_{r,b}$ in T from r to a

and b, respectively, such that the pair formed by them is openly disjoint. We consider two possible cases.

Case (1): $P_{r,c}$ shares no interior vertex with $P_{r,z}$, for all $z \in \{a, b\}$. Then $a|b|c \in \mathcal{R}(T)$. Hence $\mathcal{R}(T) \not\subseteq \{ab|c\}$ and so A cannot be an SN-set for $\mathcal{R}(\mathcal{T})$, which is impossible.

Case (2): There exists some $z \in \{a, b\}$ such that $P_{r,c}$ shares one or more interior vertices with $P_{r,z}$. Let $w \in V(T)$ denote the lowest vertex in $P_{r,c}$ such that the subpath $P_{w,c}$ of $P_{r,c}$ from w to c (i.e., the set of vertices $v \in V(P_{r,c})$ with $c \preceq v \preceq w$) does not share an interior vertex with $P_{r,a}$ and with $P_{r,b}$. Swapping the roles of a and b if necessarily, we may assume without loss of generality that w is a vertex on $P_{r,a}$. Let $P_{r,w}$ denote the subpath of $P_{r,a}$ joining w and r. Considering the vertices r and w and the dipaths $P_{r,w}$, $P_{w,a}$, $P_{w,c}$, and $P_{r,b}$ implies $ac|b \in \mathcal{R}(T)$. Hence, $\mathcal{R}(T) \not\subseteq \{ab|c\}$ which, as observed above, is impossible.

■

Using Theorems 5.1 and 5.6 we immediately obtain:

Corollary 5.7 *Suppose that N is a phylogenetic network on X and that $\emptyset \neq A \subseteq X$. Then A is a closed set in N if and only if A is an SN-set for $\mathcal{R}(N)$.*

6. The Closure Digraph

In Oldman et al. (2016) a certain digraph is associated to trinet systems consisting of level-1 trinets. Using properties of this graph, a method is developed for constructing binary level-1 networks, an important family of binary networks in which no two distinct cycles share a common vertex, from biological datasets. In this section, we shall define and study a generalization of this digraph.

First we introduce some further notation. Suppose $\mathcal{T}$ is a dense trinet system on X. For $x, y \in X$ distinct, let $\kappa_x(y)$ denote the number of elements $z \in X - \{x, y\}$ for which there exists a trinet $T \in \mathcal{T}$ on $\{x, y, z\}$ such that $y \prec \text{LSA}_T(x, z)$. Note that, in general, $\kappa_x(y) \neq \kappa_y(x)$ might hold and that $\kappa_x(y) \leq |X| - 2$ (see Figure 3 for an example).

Now, the *closure digraph* of $\mathcal{T}$, denoted by $D(\mathcal{T})$, is defined as the digraph whose vertex set is X, and any two elements $x, y \in X$ are joined by an arc (x, y) if $\kappa_x(y) = |X| - 2$. An example of a closure digraph for the trinet system of a phylogenetic network is presented in Figure 4. Informally speaking, an arc (x, y) in the closure digraph indicates that every non-singleton set that is closed in $\mathcal{T}$ and contains x must also contain y. More formally:

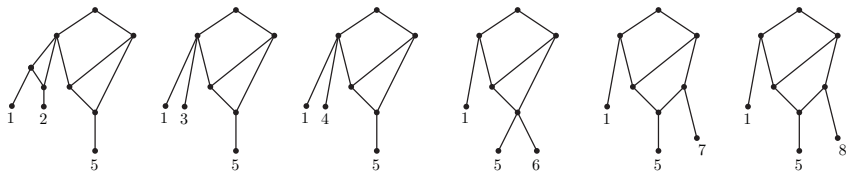

Figure 3. The six trinets that are displayed by the phylogenetic network on $\{1, 2, \cdots, 8\}$ depicted in Figure 1 and contain leaves 1 and 5. This implies $\kappa_1(5) = 6$, while $\kappa_5(1) = 5$.

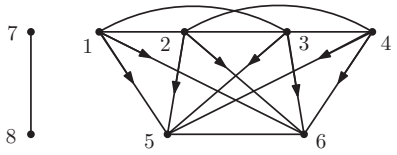

Figure 4. The closure digraph for the trinet system induced by the phylogenetic network depicted in Figure 1. Undirected edges represent bidirected arcs. Figure 4. The arc $(1, 5)$ follows from the example presented in Figure 3.

Lemma 6.1 *Suppose that $\mathcal{T}$ is a dense trinet set on X and that $x, y \in X$ distinct. If (x, y) is an arc in $D(\mathcal{T})$, then each non-singleton set that is closed in $\mathcal{T}$ and contains x must also contain y.*

Proof. Suppose that (x, y) is an arc in $D(\mathcal{T})$ and that A is closed in $\mathcal{T}$ with $x \in A$ and $|A| \geq 2$. Choose some element $a \in A - \{x\}$. Without loss of generality, we may assume $a \neq y$ as otherwise the lemma clearly holds.

Let $T \in \mathcal{T}$ denote the unique trinet on $\{x, y, a\}$. Since $(x, y) \in D(\mathcal{T})$, we have $\kappa_x(y) = |X| - 2$ and, so, $y \prec \text{LSA}_T(x, a)$. Combined with the assumption that A is closed in T, we obtain

$$\{x, y, a\} = \mathcal{C}(\text{LSA}_T(x, a)) \subseteq \mathcal{C}(\text{LSA}_T(A \cap L(T))) = A \cap L(T) \subseteq \{x, y, a\}.$$

Hence, $y \in A$ must hold.
■

Note that even if $\mathcal{T}$ is induced by a binary level-1 network, the converse of Lemma 6.1 need not always hold. For example, suppose N is the network on $X = \{1, 2, 3, 4\}$ depicted in Figure 5. Then $(2, 1)$ is not an arc in the closure digraph $D(\mathcal{T}(N))$. However, each non-singleton set A that is closed in $\mathcal{T}(N)$ must contain 1 if $2 \in A$.

Using Lemma 6.1, we now show that closed sets for dense trinet systems are so-called sink subsets in the closure digraph. Recall that a nonempty subset A of the vertex set of a digraph G is called a *sink subset* in G

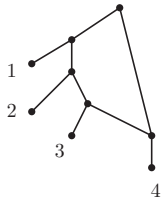

Figure 5. An example illustrating that the converse of Lemma 6.1 does not hold in general—see text for details.

if there exists no arc in G from A to $V - A$, that is, for each arc (x, y) in G with $x \in A$, we have $y \in A$ as well.

Proposition 6.2 *Suppose that $\mathcal{T}$ is a dense trinet set on X and that $A \subseteq X$ is a subset with $|A| > 1$. If A is closed in $\mathcal{T}$, then A is a sink subset in $D(\mathcal{T})$.*

Proof. Assume for contradiction that there exists some $A \subseteq X$ with $|A| \geq 2$ such that A is closed in $\mathcal{T}$ but A is not a sink subset of $D(\mathcal{T})$. Then there exists an arc (x, y) in $D(\mathcal{T})$ with $x \in A$ and $y \in X - A$. Hence, by Lemma 6.1, A cannot be closed in $\mathcal{T}$; a contradiction.
■

Note that the converse of Proposition 6.2 is not true in general. For instance, consider the network N pictured in Figure 1 and its closure digraph $D(\mathcal{T}(N))$ depicted in Figure 4. Then $\{1, 2, 3, 4, 5, 6\}$ is a sink set in $D(\mathcal{T}(N))$, but it is not closed in $\mathcal{T}(N)$. Even so, in the next section we will see that for certain class of networks the converse of Proposition 6.2 does in fact hold.

We now consider properties of the closure digraph of the trinet system induced by a phylogenetic network.

Theorem 6.3 *Suppose that N is a phylogenetic network on X and that $x, y \in X$ distinct. Then (x, y) is an arc of $D(\mathcal{T}(N))$ if either* (i) $y \prec_N p(x)$, *or* (ii) $\mathcal{C}_N(p(x)) = \{x\}$ *and* $y \prec_N \text{LSA}(p(x))$ *hold.*

Proof. Put $\mathcal{T} = \mathcal{T}(N)$. To see that $\kappa_x(y) = |X| - 2$ holds, suppose $z \in X - \{x, y\}$. We claim that $y \prec_T \text{LSA}(x, z)$ where $T \in \mathcal{T}(N)$ is the trinet with leaf set $Y := \{x, y, z\}$.

Assume first that Property (i) holds. Then we have $y \prec_N p(x) \preceq_N \text{LSA}(x, z)$. Hence $y \prec_T p(x) \preceq_T \text{LSA}(x, z)$, from which the claim follows.

Next, assume that Property (ii) holds. Let H denote the digraph obtained from N by removing all vertices that are not on any dipath from $\text{LSA}(Y)$ to some element in Y. Then T is obtained from H by recursively deleting parallel arcs and suppressing degenerate vertices.

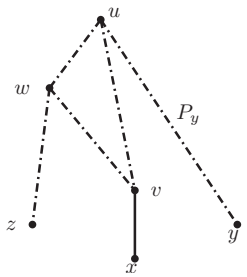

Figure 6. An illustration of Case 2-1 in the proof of Theorem 6.3. Dotted and dashed edges denote dipaths.

Let $v := p(x)$. Together with the assumption that $|\mathcal{C}(v)| = 1 \neq |X|$, it follows that $v \neq \rho := \rho(N)$ and that v must be a reticulation vertex of N. Let $v_1, v_2, \cdots, v_t \in V(N)$ denote the $t \geq 2$ parents of v. Also, let $u := \mathrm{LSA}_N(v)$. Then, by assumption, $y \prec u$. We now consider two possible cases:

Case (1) u is a parent of v: Without loss of generality, we may assume $u = v_1$. Let P_x be the dipath in N consisting of u, v and x. Since $y \prec u$ and $|\mathcal{C}(v)| = 1$, there exists a dipath P_y from u to y in N such that the pair P_x and P_y is openly disjoint in N. Since P_x and P_y also form a pair of openly disjoint dipaths in H, it follows that $u \in V(T)$. Let P'_x and P'_y be the dipaths in T induced by P_x and P_y, respectively. Since P'_x contains either no interior vertex or has v as its only interior vertex, we have $u \preceq \mathrm{LSA}_T(x, z)$ because v is not an ancestor of z. This implies the claim as $y \prec u \preceq \mathrm{LSA}_T(x, z)$ holds in T.

Case (2) u is not a parent of v: We consider two subcases:

Case (2-1) There exists no common ancestor of x and y below u (see Figure 6). This implies that there exists a dipath P_y in N from u to y in which the only ancestor of x is u. Now an argument similar to that used in Case (1) shows that $u \in V(T)$. Note that we may assume that $z \prec_N u$ holds as otherwise u is not an ancestor of z in T, and hence $y \prec u \preceq \mathrm{LSA}_T(x, z)$ holds. In addition, we may further assume that there exists a common ancestor of x and z below u, as otherwise we have $u \preceq \mathrm{LSA}_T(x, z)$.

Let w be a lowest common ancestor of x and z such that w is below u (see Figure 6 for an illustration). Let P_1 be a dipath from u to v in N that contains w, and let P_1^* denote the subpath of P_1 from u to w. Since (u, v) is not an arc in N, Theorem 3.1 combined with $u = \mathrm{LSA}_N(v)$ implies that there exists a dipath P_2 in N from u to v such that the pair P_2 and P_1 is openly disjoint. Let P_x be the dipath from u to x obtained by concatenating P_2 and the arc (v, x). Since w is a lowest common ancestor of x and z, there also exists a dipath P from w to z in which no interior vertex is an ancestor

of x. Let P_z be the dipath from u to z obtained from concatenating P_1^* and P. Then the dipath pair P_x and P_z must be openly disjoint. This implies that w and v are both contained in T, and that $y \prec u \preceq_T \text{LSA}_T(x, z)$ holds.

Case (2-2): There exists a common ancestor of x and y below u: Let w be a lowest common ancestor of x and y below u. Then an argument similar to that in Case (2-1) shows that u, w and v are all vertices in T. This implies $u \preceq_T \text{LSA}_T(x, z)$ and, thus, the claim in this case too.

■

Note that the converse of Theorem 6.3 need not hold in general. For example, consider the arc $(1, 5)$ in the closure digraph depicted in Figure 4. Then neither one of the two conditions in the theorem holds.

We now prove a useful corollary concerning sink subsets in the closure digraph associated to the trinet system of a phylogenetic network. We start with some additional notation. We say that a sink subset A in a digraph G is *minimal* if $|A| > 1$ and every subset $A' \subsetneq A$ with $|A'| > 1$ is not a sink subset in G. Suppose that N is a phylogenetic network on X and that a, b are two vertices in N such that neither one of them is a leaf. We say that a and b are *redundant* if $b \prec a$ and, for each vertex $u \preceq a$, we either have $u \preceq b$ or $b \prec u$. Note that if a and b are redundant then $\mathcal{C}_N(a) = \mathcal{C}_N(b)$.

Corollary 6.4 *Suppose that N is a phylogenetic network on X. Then every sink subset of $D(\mathcal{T}(N))$ has size at least two (or, equivalently, for every $x \in X$, there exists an element $y \in X$ such that (x, y) is an arc in $D(\mathcal{T}(N))$).*

Proof. Put $\mathcal{T} = \mathcal{T}(N)$. Note first that we may assume that N does not contain a redundant pair of vertices as otherwise we may replace N by the phylogenetic network N' obtained from N via the following process. Suppose $a, b \in V(N)$ form a redundant pair of vertices of N. First, delete all vertices $u \in V(N)$ for which $u \prec a$ and $b \prec u$ holds (including their incident arcs). Next, add the arc (a, b) to the resulting graph. Finally, suppress all degenerate vertices of that graph. Clearly, a set is closed in N if and only if it is closed in N'. Furthermore, the closure digraphs for $\mathcal{T}$ and $\mathcal{T}(N')$, respectively, coincide as a pair of elements of X forms an arc in $D(\mathcal{T})$ if and only if it forms an arc in $D(\mathcal{T}(N'))$.

Suppose $x \in X$. We show that there exists an element $y \in X$ such that (x, y) is an arc in $D(\mathcal{T})$. Clearly, if $|\mathcal{C}(p(x))| \geq 2$ then, for any $y \in \mathcal{C}(p(x)) - \{x\}$, we have that (x, y) is an arc of $D(\mathcal{T})$ in view of Theorem 6.3. So assume $|\mathcal{C}(p(x))| = 1$. Note that $p(x)$ is not the root of N as $|X| \geq 3$. Put $u = \text{LSA}(p(x))$. Also note that if $|\mathcal{C}(u)| = 1$ held then u and $p(x)$ would form a redundant pair which is impossible in view of our assumption on N. Hence, $|\mathcal{C}(u)| \geq 2$. Choose some $y \in \mathcal{C}(u) - \{x\}$. Then, by Theorem 6.3, (x, y) must be an arc in $D(\mathcal{T})$.

■

7. 2-Terminal Networks

Suppose that N is a phylogenetic network on X ($\mathcal{T}$ a trinet system) and that $A \subseteq X$ is a closed set in N (in $\mathcal{T}$) of size at least two. Then A is *minimal closed* in N (in $\mathcal{T}$) if each non-singleton subset $A' \subsetneq A$ is not closed in N (in $\mathcal{T}$). In this section, we shall show that for 2-*terminal networks*, that is, networks N for which each biconnected component of N contains at most 2 terminal vertices, the minimal closed sets in N are precisely the minimal sink subsets in the closure digraph $D(\mathcal{T}(N))$.

We begin with a key structural result concerning 2-terminal networks. Note that a similar result is proven in van Iersel et al. (2017, Theorem 3.1) for binary networks, but the binary condition plays an essential part in the proof which necessitates the development of a new approach. Suppose that N is a phylogenetic network and that H is a biconnected component of N. We denote by $r(H)$ the *highest* vertex in H, that is, the necessarily unique vertex in H such that $v \prec r(H)$ holds for all vertices v in H distinct from $r(H)$.

Lemma 7.1 *Suppose that H is a biconnected component in a 2-terminal network N. Then there exists a terminal* vertex u (of N) in H such that $\mathrm{LSA}(u) = r(H)$.

Proof. Note that the lemma clearly holds if H contains only one terminal vertex. Indeed, if u is that vertex then $\mathrm{LSA}(u) \preceq r := r(H)$ holds by definition of $r(H)$. Hence, if $\mathrm{LSA}(u) \neq r$, then $\mathrm{LSA}(u)$ is a cut vertex of H, a contradiction.

So, for the remainder of the proof, assume that H contains precisely two terminal vertices, denoted u_1 and u_2, respectively. For $i = 1, 2$, note that $u_i^* = \mathrm{LSA}(u_i)$ is a vertex of H. Swapping the roles of u_1 and u_2 if necessary, we may assume that u_1^* is not below u_2^*, that is, either u_1^* and u_2^* are not comparable via "$\prec$" or $u_2^* \prec u_1^*$.

To see that $u_1^* = r$, assume for contradiction that $u_1^* \prec r$. Then since H is biconnected and $u_1^* \neq r$ there must exist some $k \geq 3$ and a u_1^*-avoiding path $P : v_1 := r, v_2, \ldots, v_k := u_1$ in H from r to u_1. Since u_1^* is a stable ancestor of u_1, it follows that P contains at least one alternating vertex. Moreover, noting that the arcs (v_1, v_2) and (v_{k-1}, v_k) are distinct as $k \geq 3$, the number m of alternating vertices in P is at least two. Without loss of generality, we may further assume that P is chosen so that every u_1^*-avoiding path in H from r to u_1 contains at least m alternating vertices. Let $1 \leq i < j \leq k$ be such that v_i and v_j are the $(m-1)$-th and m-th alternating vertices of P, respectively. Then the dipath $P_1 : v_j, v_{j+1}, \ldots, v_k$ from v_j to u_1 is a subdipath of P, and hence u_1^*-avoiding (see Figure 7 for an illustration). Since the dipath $P_2 : v_j, v_{j-1}, \ldots, v_i$ is also a subdipath of P we have $v_i \prec v_j$.

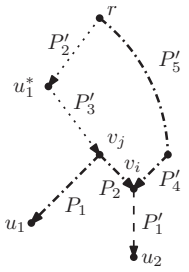

Figure 7. An illustration of the various paths considered in the proof of Lemma 7.1. The concatenation of the dipaths P_2' and P_3' forms the dipath P', and the concatenation of the path P_5' and the dipath P_4' forms the path Q. Finally, the concatenation of P', P_2 and P_1' forms the dipath $\bar{P}$.

Let P' denote a dipath from r to v_j (which exists by the definition of r). Note that the dipath obtained by concatenating P' and P_1 is a dipath in H from r to u_1, and hence contains u_1^* because u_1^* is a stable ancestor of u_1. Since P_1 is u_1^*-avoiding, it follows that u_1^* is a vertex of P'. Hence $v_i \prec v_j \prec u_1^*$. We now prove three claims which will allow us to establish that u_1^* is also a stable ancestor of u_2. This will complete the proof since if u_1^* is a stable ancestor of both u_1 and u_2, then u_1^* must be a cut vertex of H, which is impossible since H is a biconnected component of N and u_1^* is contained in H.

Now, we first claim that $u_1 \preceq v_i$ does not hold. Suppose this is not the case, i.e., $u_1 \preceq v_i$. Then there exists a dipath K from v_i to u_1. Hence, the path R obtained by concatenating the subpath $Q : v_1, \ldots, v_i$ of P with K is a path from r to u_1. Note that since P is u_1^*-avoiding, so is Q. Hence, R is also u_1^*-avoiding. Since, by construction, R has fewer alternating vertices than P this is impossible. Thus, the claim must hold.

Second, we claim that $u_2^* \preceq u_1^*$. To see this, note that since u_1 and u_2 are the only two terminal vertices in H, and, by the previous claim, $u_1 \preceq v_i$ does not hold, we have $u_2 \preceq v_i$ as every non-terminal vertex of N must have a terminal vertex of N below it. Without loss of generality, we may assume that, in fact, $u_2 \prec v_i$ because the case $u_2 = v_i$ can be established in a similar manner. Then there exists a dipath P_1' from v_i to u_2. Hence, the dipath $\bar{P}$ obtained by concatenating P', P_2, and P_1' is a dipath in H from r to u_2. By the definition of a stable ancestor, u_2^* must be a vertex of $\bar{P}$. Since, as was observed above, u_1^* is a vertex of P' and, by assumption, $u_1^* \prec u_2^*$ does not hold, we obtain $u_2^* \preceq u_1^*$, as required for the second claim to hold.

Finally, we claim that u_1^* is also a stable ancestor of u_2. To see this, note first that $u_1 \prec u_2^*$ must hold. Indeed, if $u_1 \prec u_2^*$ did not hold, then u_2^* must be a cut vertex of H since u_2 is the only other terminal vertex contained

in H. But this is impossible as H is a biconnected component of N. Now, assume for contradiction that u_1^* is not a stable ancestor of u_2. Then every u_1^*-avoiding dipath P from r to u_2 (if it exists) must contain u_2^* since u_2^* is a stable ancestor of u_2. But $u_1 \prec u_2^* \preceq u_1^*$ by the previous claims. Hence, the subpath of P from r to u_2^* can be extended to a u_1^*-avoiding dipath from r to u_1. This is impossible as u_1^* is a stable ancestor of u_1.
■

We now show that for 2-terminal networks N, minimal sink subsets in $D(\mathcal{T}(N))$ are closed sets in N.

Proposition 7.2 *Suppose that N is a 2-terminal network on X and that $A \subseteq X$ is a subset with $|A| \geq 2$. If A is a minimal sink subset in $D(\mathcal{T}(N))$, then A is closed in N.*

Proof. Put $\mathcal{T} = \mathcal{T}(N)$ and assume that $A \subseteq X$ is a subset with $|A| \geq 2$ that is also a minimal sink subset in $D(\mathcal{T})$. Using arguments similar to the ones used at the beginning of the proof of Corollary 6.4, we may assume that N does not contain a redundant pair of vertices. The remainder of the proof of the proposition is based on two claims which we establish first. Suppose U is the set of terminal vertices u in N for which, in addition, $\mathcal{C}(u) \cap A \neq \emptyset$ holds. Note that $U \neq \emptyset$ as every element of X is a terminal vertex of N.

Claim 1: For each vertex $u \in U$, either $|\mathcal{C}(u)| = 1$ or $A \subseteq \mathcal{C}(u)$ must hold (but not both). To prove the claim, assume that there exists a vertex $u \in U$ with $|\mathcal{C}(u)| > 1$. We need to show that $A \subseteq \mathcal{C}(u)$. Since u is a terminal vertex of N, Theorem 3.6 implies that $\mathcal{C}(u)$ is closed in N. By Theorem 5.1 and Proposition 6.2, it follows that $\mathcal{C}(u)$ is a sink subset in $D(\mathcal{T})$. Since A is also a sink subset of $D(\mathcal{T})$, the intersection $B = \mathcal{C}(u) \cap A$ is necessarily a sink subset of $D(\mathcal{T})$. By Corollary 6.4, $|B| \geq 2$. Since $B \subseteq A$, the minimality of A implies $B = A$. Thus, $A \subseteq \mathcal{C}(u)$, which completes the proof of Claim 1.

Claim 2: If H is a biconnected component of N that contains a vertex $u \in U$ with $|\mathcal{C}(u)| = 1$, then $\mathcal{C}(r(H)) \subseteq A$. To prove this claim, let u_x denote the unique leaf in $\mathcal{C}(u)$. Note that since N does not contain a redundant pair, u must be the parent of u_x.

Assume first that u is the only terminal vertex of N contained in H. Then $r(H) = \text{LSA}(u)$. Assume for contradiction that there exists some $y \in \mathcal{C}(r(H)) - A$. Then $y \prec r(H) = \text{LSA}(u) = \text{LSA}(p(u_x))$. Hence, by Theorem 6.3, (u_x, y) must be an arc in $D(\mathcal{T})$. Since $u_x \in A$ as $|\mathcal{C}(u)| = 1$, and A is a sink subset of $D(\mathcal{T})$, it follows by Lemma 6.1 that $y \in A$, which is impossible.

Now, suppose that H contains two terminal vertices u_1 and u_2 of N, with $u = u_1$. Put $u_1^* = \text{LSA}(u_1)$, noting that we may assume that

$u_1^* \prec r(H)$ holds since otherwise arguments similar to the ones in the proof of Claim 1 maybe applied. Moreover, $u_2 \prec u_1^*$ as otherwise u_1^* is a cut vertex of H, a contradiction. But then Theorem 6.3 implies for all $y \in \mathcal{C}(u_2)$ that (u_x, y) is an arc in $D(\mathcal{T})$. Since $u_x \in A$ and A is a sink subset of $D(\mathcal{T})$, it follows by Lemma 6.1 that $\mathcal{C}(u_2) \subseteq A$. Since $A \subseteq \mathcal{C}(u_2)$ cannot hold as u_2 is a terminal vertex distinct from u_1, Claim 1 implies that $|\mathcal{C}(u_2)| = 1$. Thus, there exists some $y \in A$ such that $\mathcal{C}(u_2) = \{y\}$. Furthermore, since $u_1^* \prec r(H)$ and H is a biconnected component of a 2-terminal network we have $u_2^* = r(H)$ by Lemma 7.1. Together with $u_2 = p(y)$, Theorem 6.3 implies that (y, z) is an arc in $D(\mathcal{T})$ for all z in $\mathcal{C}(r(H)) - \{y\}$. Combined with the assumption that A is a sink subset of $D(\mathcal{T})$ and $y \in A$, it follows that $\mathcal{C}(r(H)) \subseteq A$. This completes the proof of Claim 2.

Using these claims we now prove that A is closed in N. Suppose $x \in A$. Then, by Theorem 6.3, (x, y) is an arc in $D(\mathcal{T})$, for all $y \in \mathcal{C}(p(x)) - \{x\}$. Hence $\mathcal{C}(p(x)) \subseteq A$. Note that if $p(x)$ is the root $\rho(N)$ of N then $\mathcal{C}(p(x)) = \mathcal{C}(\rho(N)) = X$. Thus, $A = X$ and, so, A is closed in N by definition. Thus, assume for the remainder of the proof that $\rho(N) \neq p(x)$.

Let $p'(x)$ be a parent of $p(x)$ in N and let C denote the biconnected component of N containing the arc $(p'(x), p(x))$. We consider two possible cases:

Case (1) C is a trivial biconnected component of N: Then $(p'(x), p(x))$ is the unique arc of C. Since that arc is clearly a cut arc of N, it follows that $\mathcal{C}(p(x))$ is a cut-arc set for N. Hence, by Theorem 4.1, $\mathcal{C}(p(x))$ is closed in N. Thus, by Theorem 5.1, $\mathcal{C}(p(x))$ is closed in $\mathcal{T}$. Since $(p'(x), p(x))$ is a cut arc of C and N does not contain degenerate vertices, $p(x)$ has at least two children. Hence, $|\mathcal{C}(p(x))| > 1$. By Proposition 6.2, it follows that $\mathcal{C}(p(x))$ must be a sink subset in $D(\mathcal{T})$. Since $\mathcal{C}(p(x)) \subseteq A$, the minimality of A implies $A = \mathcal{C}(p(x))$. Thus, A is closed in N.

Case (2) C is not a trivial biconnected component of N: Let U_C be the set of terminal vertices u in C for which, in addition, $\mathcal{C}(u) \cap A \neq \emptyset$ holds. Note that U_C is not empty as it contains either $p(x)$ or a descendant of $p(x)$. We consider two sub-cases:

Case (2-1) There exists a vertex $u \in U_C$ with $|\mathcal{C}(u)| > 1$: Then, by Claim 1, $A \subseteq \mathcal{C}(u)$. Hence, $x \prec u$, and, therefore, $p(x) \preceq u$. Since u is a terminal vertex of N in C and $(p'(x), p(x))$ is an arc of C, we obtain $p(x) = u$. In view of Theorem 6.3, it follows that for all $y \in \mathcal{C}(u)$, (x, y) is an arc in $D(\mathcal{T})$. Hence, by Lemma 6.1 $\mathcal{C}(u) \subseteq A$. By the minimality of A, we obtain $A = \mathcal{C}(u)$. Thus A is a closed in N.

Case (2-2) $|\mathcal{C}(u)| = 1$, for all $u \in U_C$: We shall construct a sequence of vertices $r_0, r_1, \ldots$ of N which will eventually terminate at a vertex r_k, $k \geq 0$, so that $\mathcal{C}(r_k) = A$ and r_k is either $\rho(N)$ or a terminal vertex of N. Put $r_0 = r(C)$. Then $|\mathcal{C}(r_0)| > 1$ because C is non-trivial and N

does not contain any redundant pair of vertices. By Claim 2, $\mathcal{C}(r_0) \subseteq A$. Hence, if $r_0 = \rho(N)$, then $X = \mathcal{C}(\rho(N)) = \mathcal{C}(r_0) \subseteq A \subseteq X$, which implies $\mathcal{C}(r_0) = X = A$. Hence, A is closed in N in this case. So suppose $r_0 \neq \rho(N)$. If r_0 is a terminal vertex of N, then Theorem 3.6 implies that $\mathcal{C}(r_0)$ is closed in N. By Theorem 5.1 and Proposition 6.2, $\mathcal{C}(r_0)$ is a sink subset in $D(\mathcal{T})$, and by minimality of A, $\mathcal{C}(r_0) = A$.

So, assume $r_0 \neq \rho(N)$ and that r_0 is not a terminal vertex of N. Then there exists some biconnected component C_1 of N that contains r_0 so that $r_0 \prec r_1 := r(C_1)$ holds. Furthermore, let $u^1 \in V(C_1)$ denote a terminal vertex of N for which $u^1 \prec r_0$ holds. Then $\mathcal{C}(u^1) \subseteq \mathcal{C}(r_0) \subseteq A$ and so $u^1 \in U$. Note that since $(p'(x), p(x))$ is an arc in C, we have $x \notin \mathcal{C}(u^1)$. Hence, $A \not\subseteq \mathcal{C}(u^1)$. By Claim 1, $|\mathcal{C}(u^1)| = 1$, and so by Claim 2, $\mathcal{C}(r_1) \subseteq A$. With r_1 playing the role of r_0 in the argument used in the last paragraph, if $r_1 = \rho(N)$ or r_1 is a terminal vertex of N, then $\mathcal{C}(r_1) = A$. Therefore, r_1 must be contained in a biconnected component C_2 of N which contains a terminal vertex $u^2 \in C_2$ with $u^2 \prec r_1 \prec r_2 := r(C_2)$ and $\mathcal{C}(r_2) \subseteq A$.

Since N is finite, this process of constructing vertices r_i, $i \geq 0$ must terminate at some stage $k \geq 0$ resulting in a vertex r_k such that $\mathcal{C}(r_k) = A$ and r_k is either $\rho(N)$ or a terminal vertex of N.
■

We now characterize sets that are minimal closed in 2-terminal networks.

Theorem 7.3 *Suppose that N is a 2-terminal network on X and $A \subseteq X$ with $|A| \geq 2$. Then the following assertions are equivalent.*
(i) *A is minimal closed in N.*
(ii) *A is minimal closed in $\mathcal{T}(N)$.*
(iii) *A is a minimal sink subset in the closure digraph $D(\mathcal{T}(N))$.*

Proof. $(i) \iff (ii)$: This is a direct consequence of Theorem 5.1.

$(ii) \Rightarrow (iii)$: Suppose that A is a minimal closed set in $\mathcal{T} := \mathcal{T}(N)$. Then, by Proposition 6.2, A is a sink subset in $D(\mathcal{T})$. Assume for contradiction that A is not a minimal sink subset in $D(\mathcal{T})$. Then there exists a minimal sink subset $B \subseteq X$ in $D(\mathcal{T})$ with $B \subsetneq A$. By Proposition 7.2, B must be closed in N. Hence, by Theorem 5.1, B is also a closed in $\mathcal{T}$. Thus, $B = A$ by the minimality of A which is impossible.

$(iii) \Rightarrow (ii)$: Put $\mathcal{T} := \mathcal{T}(N)$ and suppose that A is a minimal sink subset in $D(\mathcal{T})$. Then, by Proposition 7.2, A is closed in N. Assume for contradiction that A is not minimal closed in N. Then there exists some $B \subsetneq A$ that is minimal closed in N. By the equivalence of Assertions (i) and (ii) in Theorem 7.3, B must be a minimal closed set in $\mathcal{T}$. Hence, $|B| \geq 2$ by the definition of a minimal closed set of N. By Proposition 6.2,

B is a sink subset in $D(\mathcal{T})$. Thus, $A = B$ by the minimality of A which is impossible.
■

To illustrate the last theorem, consider the network N on $X = \{1, 2, \ldots, 8\}$ in Figure 1. Then $A := \{7, 8\}$ is minimal closed in N, and A is also a minimal sink subset in the closure digraph $D(\mathcal{T}(N))$ (see Figure 4). On the other hand, $A' := \{1, 2, 3, 4, 5, 6\}$ is a sink subset in the closure digraph $D(\mathcal{T}(N))$ but it is not a minimal sink subset because $\{5, 6\}$ is a sink subset. Hence Theorem 7.3 implies that A' is not minimal closed in N. Indeed, A' is not even a closed set in N.

Since a level-2 (and hence also a level-1) network is necessarily a 2-terminal network, Theorem 7.3 can be viewed as a significant generalization of a result presented in Oldman, Wu, van Iersel and Moulton (2016, Theorem 1 in the Appendix), which characterizes minimal sink subsets in the closure digraph induced by level-1 networks using minimal cut-arc sets.

8. Conclusions and Future Directions

In this paper we have introduced the concept of a closed set in a phylogenetic network. We have seen that these sets provide a natural way to extend the notion of SN-sets for binary networks to general networks, and that the closed sets of a network are closely related to the triplets and trinets that it displays.

In Theorem 7.3, we showed that we can characterize the closed sets of a 2-terminal network in terms of minimal sink subsets of the closure digraph associated to the triplets displayed by the network. It would be interesting to know whether or not this result also holds for networks in general, although this appears to be quite difficult to decide. In addition, it could also be of interest to better understand properties of 2-terminal networks (or more generally, k-terminal networks, $k \geq 1$, which can be defined in the obvious way). For example, are 2-terminal networks defined by their trinets? Note that level-2 networks enjoy this property (van Iersel and Moulton 2014).

In general, a phylogenetic network is not determined by its trinets (even if it is binary) (Huber, van Iersel, Moulton and Wu 2015). However, by Theorem 5.2 it follows that the cut arc hierarchy $\mathcal{H}_{CA}(N)$ can be constructed from the trinets of a phylogenetic network N. It would be interesting to know whether or not the cut vertex hierarchy $\mathcal{H}_{CV}(N)$ or the related hierarchy $\mathcal{H}^*_{CV}(N)$ can also be reconstructed from trinets. More generally, it could be useful to understand which other features of networks are determined by their trinets.

In this paper we have concentrated on theoretical properties of closed sets. However, there are associated algorithmic questions that are also of

interest. For example, note that combined with an algorithm similar to the one presented in Jansson and Sung (2004, Figure 4), Lemma 5.5 can be used to compute, for any dense triplet system $\mathcal{R}$ on a set X, the associated family of SN-sets for $\mathcal{R}$ in $O(|X|^5)$ time. However, it would be interesting to know whether there may be a more efficient algorithm for computing closed sets along the lines of the one presented in Jansson et al. (2006) for computing SN-sets. This might also use results presented in Fischer and Huson (2010) for computing lowest stable ancestors.

Solutions to these sorts of problems should eventually lead to new algorithms for computing phylogenetic networks. One possible approach to develop such an algorithm could be to use Theorem 7.3 as a basis for computing level-2 networks (or more generally 2-terminal networks). This might follow the approach that was used in Oldman et al. (2016) to construct level-1 networks in a bottom up fashion from level-1 trinets. In particular, first a dense set of level-2 trinets would be computed from biological data and then, using the closure digraph of this set, a minimal sink subset would be found. For this subset a simple level-2 network could then be derived, and the subset replaced by a single element in such a way that this whole process could be repeated. However, various problems would need to be overcome to make this approach work. For example, new methods need to be developed to associate level-2 trinets to biological data, and robust ways need to be found for combining level-2 trinets into level-2 networks.

References

CARDONA, G., LLABRES, M., ROSSELLO, F., and VALIENTE, G. (2011), "Comparison of Galled Trees," *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8*, 410–427.

DRESS, A., MOULTON, V., STEEL, M., and WU, T. (2010), "Species, Clusters and the 'Tree of life': A Graph-Theoretic Perspective," *Journal of Theoretical Biology, 265*, 535–542.

FISCHER, J., and HUSON, D. (2010), "New Common Ancestor Problems in Trees and Directed Acyclic Graphs," *Information Processing Letters, 110*, 331–335.

GUSFIELD, D. (2014), *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*, MIT Press.

HUBER, K.T., and MOULTON, V. (2012), "Encoding and Constructing 1-Nested Phylogenetic Networks with Trinets," *Algorithmica, 616*, 714–738.

HUBER, K.T., VAN IERSEL, L., MOULTON, V., SCORNAVACCA, C., and WU, T. (2017), "Reconstructing Phylogenetic Level-1 Networks from Nondense Binet and Trinet Sets," *Algorithmica, 77*, 173–200.

HUBER, K.T., VAN IERSEL, L., MOULTON, V., and WU, T. (2015), "How Much Information is Needed to Infer Reticulate Evolutionary Histories," *Systematic Biology, 64*, 102–111.

HUBER, K.T., VAN IERSEL, L., KELK, S., and SUCHECKI, R. (2011), "A Practical Algorithm for Reconstructing Level-1 Phylogenetic Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8*, 635–649.

HUSON, D.H., RUPP, R., and SCORNAVACCA, C. (2010), *Phylogenetic Networks: Concepts, Algorithms and Applications*, Cambridge University Press.

JANSSON, J., NGUYEN, N., and SUNG, W.-K. (2006), "Algorithms for Combining Rooted Triplets into a Galled Phylogenetic Network," *SIAM Journal of Computing, 35*, 1098–1121.

JANSSON, J., and SUNG, W.-K. (2006), "Inferring a Level-1 Phylogenetic Network from a Dense Set of Rooted Triplets," *Theoretical Computer Science, 363*, 60–68.

JETTEN, L., and VAN IERSEL, L. (2016), "Nonbinary Tree-Based Phylogenetic Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press.

LOVÁSZ, L., and PLUMMER, M.D. (1986), *Matching Theory (Vol. 121, North-Holland Mathematics Studies)*, Elsevier Science Ltd.

NAKHLEH, L. (2011), "Evolutionary Phylogenetic Networks: Models and Issues," in *Problem Solving Handbook in Computational Biology and Bioinformatics*, Springer, pp. 125–158.

OLDMAN, J., WU, T., VAN IERSEL, L., and MOULTON, V. (2016), "Trilonet: Piecing Together Small Networks to Reconstruct Reticulate Evolutionary Histories," *Molecular Biology and Evolution, 33*, 2151–2162.

SEMPLE, C., and STEEL, M. (2003), *Phylogenetics,* Oxford University Press.

TO, T.-H., and HABIB, M. (2009), "Level-k Phylogenetic Networks are Constructable from a Dense Triplet Set in Polynomial Time", in *Annual Symposium on Combinatorial Pattern Matching*, Springer, pp. 275–288.

VAN IERSEL, L., KEIJSPER, J., KELK, S., STOUGIE, L., HAGEN, F., and BOEKHOUT, T. (2009), "Constructing Level-2 Phylogenetic Networks from Triplets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6*, 667–681.

VAN IERSEL, L., and KELK, S. (2011), "Constructing the Simplest Possible Phylogenetic Network from Triplets," *Algorithmica, 60*, 207–235.

VAN IERSEL, L., and MOULTON, V. (2014), "Trinets Encode Tree-Child and Level-2 Phylogenetic Networks," *Journal of Mathematical Biology, 68*, 1707–1729.

VAN IERSEL, L., MOULTON, V., DE SWART, E., and WU, T. (2017), "Binets: Fundamental Building Blocks for Phylogenetic Networks," *Bulletin of Mathematical Biology, 79*, 1135–1154.