On Encodings of Phylogenetic Networks of Bounded Level

Philippe Gambette, Katharina Huber

Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

On Encodings of Phylogenetic Networks of Bounded Level

Philippe Gambette · Katharina T. Huber

Received: date / Accepted: date

Abstract Phylogenetic networks have now joined phylogenetic trees in the center of phylogenetics research. Like phylogenetic trees, such networks canonically induce collections of phylogenetic trees, clusters, and triplets, respectively. Thus it is not surprising that many network approaches aim to reconstruct a phylogenetic network from such collections. Related to the well-studied perfect phylogeny problem, the following question is of fundamental importance in this context: When does one of the above collections encode (i.e. uniquely describe) the network that induces it? For the large class of level-1 (phylogenetic) networks we characterize those level-1 networks for which an encoding in terms of one (or equivalently all) of the above collections exists. In addition, we show that three known distance measures for comparing phylogenetic networks are in fact metrics on the resulting subclass and give the diameter for two of them. Finally, we investigate the related concept of indistinguishability and also show that many properties enjoyed by level-1 networks are not satisfied by networks of higher level.

This work was supported by the French ANR projects ANR-06-BLAN-0148-01 (GRAAL) and ANR-08-EMER-011-01 (PhylARIANE). The authors would like to thank the organizers of the MIEP 2008 workshop where this work was initiated and also the two referees for their helpful comments and suggestions. Finally, they would also like to thank the London Mathematical Society for supporting part of this work.

Philippe Gambette
I.M.L.,
C.N.R.S., Université Marseille 2, France.
E-mail: philippe.gambette@gmail.com

Katharina T. Huber
School of Computing Sciences,
University of East Anglia,
Norwich, NR4 7TJ, UK.
E-mail: Katharina.Huber@cmp.uea.ac.uk

The final publication is available at www.springerlink.com
DOI: 10.1007/s00285-011-0456-y

Keywords Phylogenetic networks · triplets · clusters · supernetwork · level-1 network · level-k network · weak hierarchy · consistency · metric · indistinguishable

Mathematics Subject Classification (2000) 92B10

1 Introduction

An improved understanding of the complex processes that drive evolution has lent support to the idea that reticulate evolutionary events, such as lateral gene transfer or hybridization, are more common than originally thought, rendering a phylogenetic tree (essentially a rooted leaf labelled graph-theoretical tree) too simplistic a model to fully understand the complex processes that drive evolution. Reflecting this, phylogenetic networks have now joined phylogenetic trees in the center of phylogenetics research. Influenced by the diversity of questions posed by evolutionary biologists that can be addressed with a phylogenetic network, various alternative definitions of these types of networks have been developed over the years (see e. g. (Huson et al, 2011) for a recent overview). These include split networks (Bryant and Moulton, 2004; Bandelt et al, 1995; Holland et al, 2004) as well as ancestral recombination graphs (Song and Hein, 2005), TOM networks (Willson, 2006), level-k networks[1] with k a non-negative integer that in a some sense captures how complex the network structure is, networks for studying the evolution of polyploid organisms (Moulton and Huber, 2006), tree-child and tree-sibling networks (Cardona et al, 2008), to name just a few.

Apart from split networks which aim to give an implicit model of evolution and are not the focus of this paper, all other phylogenetic networks mentioned above aim to provide an explicit model of evolution. Although slightly different in detail, they are all based on the concept of a leaf-labelled rooted connected directed acyclic graph (see the next section for a definition). For the convenience of the reader, we depict an example of a phylogenetic network in the form of a level-1 network in Fig. 1(a). Concerning these types of phylogenetic networks, it should be noted that they are closely related to *galled trees* (Wang et al, 2001; Gusfield et al, 2003) and that, in addition to constituting the first layer of the hierarchy of level-k networks, they also give rise to a large subclass of the class of tree-sibling networks (Arenas et al, 2008).

Due to the rich combinatorial structure of phylogenetic networks, different combinatorial objects have been used to reconstruct them from biological data.

[1] Note that these networks were originally introduced in Choy et al (2004), but the definition commonly used now is slightly different with the main difference being that every vertex of the network with indegree 2 must have outdegree 1 (see e.g. Jansson et al (2006)).

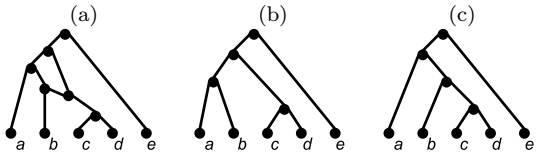


Fig. 1 (a) A level-1 network N. (b) and (c) The phylogenetic trees that form the tree system $\mathcal{T}(N)$.

For a set X of taxa (e.g. species or organisms), these include *cluster systems* of X, that is, collections of non-empty subsets of X (Bandelt and Dress, 1989; Huson and Rupp, 2008), *triplet systems* on X, that is, collections of phylogenetic trees with just three leaves which are generally called *(rooted) triplets* (Jansson and Sung, 2006; To and Habib, 2009), and *tree systems*, that is, collections of phylogenetic trees which all have leaf set X (Semple, 2007). The underlying rationale being that any phylogenetic network N induces a *hardwired* cluster system $\mathcal{C}(N)$, a triplet system $\mathcal{R}(N)$ and a tree system $\mathcal{T}(N)$. Again we defer the precise definitions to later sections of this paper, and remark that for the level-1 network N with leaf set $X = \{a, b \ldots, e\}$ depicted in Fig. 1(a), the cluster system $\mathcal{C}(N)$ is the set $\bigcup_{x \in X}\{\{x\}\} \cup \{\{a, b\}, \{c, d\}, \{b, c, d\}, Y, X\}$, where $Y = X - \{e\}$, and the tree system $\mathcal{T}(N)$ consists of the phylogenetic trees depicted in Fig. 1(b) and (c), respectively. Denoting a phylogenetic tree t on x, y, z such that the root of t is the parent vertex of z and the parent vertex of x and y by $z|xy$ (or equivalently by $xy|z$) then the triplet system $\mathcal{R}(N)$ consists of all triplets of the form $e|xy$ with $x, y \in Y$ distinct, plus the triplets $a|cd$, $b|cd$, $c|ab$, $a|bc$, $d|ab$ and $a|bd$.

Although undoubtedly highly relevant for phylogenetic network reconstruction, the following fundamental question has however remained largely unanswered so far (the main exception being the case where N is in fact a phylogenetic tree in which case this question is closely related to the well-studied *perfect phylogeny problem* – see e.g. Grünewald and Huber (2007) for a recent overview): When do the systems $\mathcal{C}(N)$, $\mathcal{R}(N)$, or $\mathcal{T}(N)$ induced by a phylogenetic network N *encode* N, that is, there is no other phylogenetic network N' for which the corresponding systems for N and N' coincide?

Complementing the insights for when N is a phylogenetic tree alluded to above, answers were recently provided for $\mathcal{R}(N)$ in case N is a very special type of level-k network, $k \geq 2$, (van Iersel et al, 2009b) and for $\mathcal{T}(N)$ for the special case that N is a regular network (Willson, 2010). Undoubtedly these are important first results. However, there are many types of phylogenetic networks which are encoded by the tree system they induce but which are not regular. Similarly, there are many types of phylogenetic networks which are encoded by the triplet system they induce but they do not belong to that special class of level-k networks considered in van Iersel et al (2009b). An example for both cases is the level-1 network depicted in Fig. 1(a) modified by

subdividing the incoming arc of the parent of b by a new vertex v and then adding an arc from v to a new labelled leaf. Although one might be tempted to speculate that all level-1 networks enjoy this property, this is not the case since the level-1 networks depicted in Fig. 1(a) and Fig. 3(b), respectively, induce the same tree system and the same triplet system. The main result of this paper shows that these observations are not a coincidence. More precisely, in Theorem 1 we establish that a level-1 network N is encoded by the triplet system $\mathcal{R}(N)$ (or equivalently by the tree system $\mathcal{T}(N)$ or equivalently the *softwired* cluster system $\mathcal{S}(N) = \mathcal{S}(\mathcal{T}(N)) := \bigcup_{T \in \mathcal{T}(N)} \mathcal{C}(T)$ which arises in the context of the *softwired interpretation* of N (Huson and Rupp, 2008) and contains $\mathcal{C}(N)$) if and only if, when ignoring directions, N does not contain a cycle of length 4. Consequently the number of non-isomorphic (see below) phylogenetic networks which all induce the same tree system as N (or equivalently the same triplet system or the same cluster system $\mathcal{S}(N)$) grows exponentially in the number of cycles of N of length 4. Furthermore, Theorem 1 implies that three known distance measures for phylogenetic networks are in fact metrics on the resulting subclass of level-1 networks and for two of them we establish their diameter on that class. It is of course highly tempting to speculate that a similar characterization might hold for higher values of k. However as our analysis of level-2 networks shows this is not the case. Moreover, we show that it is possible for a level-2 network to be encoded by some of the above systems without being encoded by the others.

The paper is organized as follows. In the next section, we present the definition of a level-1 network plus surrounding and relevant terminology. In Section 3, we present the definitions of the cluster system $\mathcal{C}(N)$ and the tree system $\mathcal{T}(N)$ induced by a phylogenetic network N. This also completes the definition of the cluster system $\mathcal{S}(N)$ given in the introduction. Subsequent to this, we show that for any level-1 network N, the cluster systems $\mathcal{S}(N)$ and $\mathcal{C}(N)$ are weak hierarchies (Proposition 1) which are well-known objects in cluster analysis. In addition, we show that, in general, this property is not enjoyed by level-k networks with $k \geq 2$. In Section 4, we present the definition of the triplet system $\mathcal{R}(N)$ induced by a phylogenetic network N. Subsequent to this, we first investigate the system $\mathcal{R}(N)$ in case N is a structurally very simple level-1 network and then establish that the triplet system induced by a phylogenetic tree T is contained in the triplet system of a level-1 network N if and only if $T \in \mathcal{T}(N)$ holds (Proposition 2). In Section 5, we prove our main result (Theorem 1) and, in Section 6, we study the restriction of three known distance measures to the resulting subclass of level-1 networks. In Section 7, we turn our attention to higher level networks and show that an encoding of such a network in terms of one of the above systems does not imply that it is also encoded by the other systems (Proposition 5). In Section 8, we conclude with some general remarks concerning the accurate reconstruction of phylogenetic networks from triplets and phylogenetic trees in general and level-k networks in particular.

To ease the presentation of our results, in all figures the (unique) root of a network is the top vertex and all arcs are directed downwards, away from the root.

2 Basic terminology and results concerning level-1 networks

In this section we present the definitions of a phylogenetic network and of a level-k network, $k \geq 0$. In addition we also provide the basic and relevant terminology surrounding these structures.

Suppose X is a non-empty finite set. For any directed graph G, we denote the vertex set of G by $V(G)$, the set of *leaves* of G (i. e. the vertices of G with indegree 1 and outdegree 0) by $L(G)$ and the set of arcs of G by $A(G)$. Furthermore, we put $V^-(G) := V(G) - L(G)$. The arcs in $A(G)$ whose removal disconnect G in the sense that for any two vertices in the resulting graph there does not exist a (possibly undirected) path between them are called the *cut-arcs* of G. A cut-arc of G that is incident with a leaf of G is called trivial *trivial*.

A *phylogenetic network* N on X is a rooted directed acyclic graph (DAG) that satisfies the following additional properties:

(P1) $L(N) = X$.
(P2) Exactly one vertex of N, called the *root* and denoted by ρ_N, has indegree 0 and outdegree 2.
(P3) All vertices of N that are not contained in $L(N) \cup \{\rho_N\}$ are either *split vertices*, that is, have indegree 1 and outdegree 2 or *reticulation vertices*, that is, have indegree 2 and outdegree 1.

The set of reticulation vertices of N is denoted by $r(N)$. A phylogenetic network N with $r(N) = \emptyset$ is called a *(rooted) phylogenetic tree (on X)*. Two phylogenetic networks N and N' which both have leaf set X are said to be *isomorphic* if there exists a bijection from $V(N)$ to $V(N')$ which is the identity on X and induces a graph isomorphism between N and N'.

To present the definition of a level-k network, we need to introduce some terminology concerning rooted DAGs first. Suppose G is a rooted connected DAG with at least 2 vertices. Then we denote the graph obtained from G by ignoring the directions on G by $U(G)$. If H is a graph with at least 2 vertices then we call H *biconnected* if H does not contain a vertex whose removal disconnects it. A *biconnected component* of H is a maximal subgraph of H that is biconnected. If G is a phylogenetic network and B is a rooted sub-DAG such that $U(B)$ is a biconnected component of $U(G)$ then we call B a *block*.

Following Choy et al (2004), we call a phylogenetic network N a *level-k (phylogenetic) network* for some non-negative integer k if each block of N contains at most k reticulation vertices. The quantity k is sometimes referred to as the *level* of N. Note that some authors define a level-1 network N to be a phylogenetic network without the above outdegree requirement on the elements in $r(N)$ (see e.g. Jansson and Sung (2006); Choy et al (2005)) or the above indegree requirement on the elements in $r(N)$ (see e. g. Rosselló and Valiente (2009)). Also the requirement that each block contains at most k reticulation vertices is sometimes replaced for level-1 networks by the requirement that the cycles in $U(N)$ are node disjoint (see e.g. Jansson et al (2006)). Although in spirit capturing the same idea, the difference between these definitions is that, according to Choy et al (2005), the structure depicted in Fig. 2 is a level-1 network whereas according to van Iersel et al (2009a) it is not. See also Rosselló and Valiente (2009) for more on this.

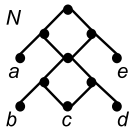


Fig. 2 Using the definition in Choy et al (2005) or in Jansson and Sung (2006), N is a level-1 network. However, using the definition in Jansson et al (2006) or in van Iersel et al (2009a), N is not a level-1 network. In fact, it is not even a phylogenetic network.

Regarding the definition of a level-k network sensu van Iersel et al (2009a), it should be noted that the phylogenetic network depicted in Fig. 3(a) is a level-1 network. However the alternative level-1 network N depicted in Fig. 3(b) is a less parsimonious representation of the same biological information (expressed in terms of the systems $\mathcal{T}(N)$, $\mathcal{R}(N)$, $\mathcal{C}(N)$, and $\mathcal{S}(N)$) than the former in the sense that the arcs in gray are redundant for displaying that information. To avoid these types of level-1 networks which cannot be encoded by any of the 4 systems of interest in this paper, we follow van Iersel et al (2009b) and require that every block in a level-1 network which is not a cut-arc contains at least 4 vertices.

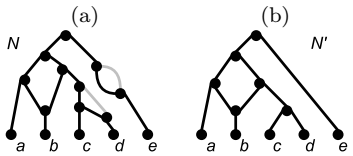


Fig. 3 The level-1 network N depicted in (a) induces and thus represents the same triplet system $\mathcal{R}(N)$, cluster systems $\mathcal{C}(N)$ and $\mathcal{S}(N)$, and tree system $\mathcal{T}(N)$ as the level-1 network N' presented in (b). However, N' is a less parsimonious representation of those 4 systems.

For $k = 1, 2$, it was shown in van Iersel et al (2009a) (see also Jansson and Sung (2006) for the case $k = 1$) that level-k networks can be built up by chaining together structurally very simple level-k networks called *simple level-k networks* (see also (Gambette et al, 2009) for more on this). More precisely, a level-k network N, $k \geq 0$, is called *simple* if every cut-arc of N is a trivial cut-arc. For example, the network obtained by first contracting the outgoing arcs of the root of the level-1 network N' depicted in Fig. 3(b) and then contracting the incoming arcs of c and d (retaining the label c), is a simple level-1 network on $\{a, b, c\}$.

From now on and unless stated otherwise, X is a non-empty finite set and all phylogenetic networks have leaf set X.

3 The systems $\mathcal{C}(N)$, $\mathcal{T}(N)$, and $\mathcal{S}(N)$

In this section, we introduce for a phylogenetic network N the associated systems $\mathcal{C}(N)$, $\mathcal{T}(N)$, and $\mathcal{S}(N)$ already mentioned in the introduction. In addition, we prove that in case N is a level-1 network the associated systems $\mathcal{C}(N)$ and $\mathcal{S}(N)$ are weak hierarchies. We conclude with presenting an example that shows that higher level networks do not enjoy this property in general. We start with some definitions.

Suppose N is a phylogenetic network. Then we say that a vertex $a \in V(N)$ is *below* a vertex $b \in V(N)$, denoted by $a \preceq_N b$, if there exists a path P_{ba} (possibly of length 0) from b to a. In this case, we also say that b is *above* a. Every vertex $v \in V(N)$ therefore induces a non-empty subset $C(v) = C_N(v)$ of X which comprises of all leaves of N below v (see e.g. Semple and Steel (2003, page 51)). We collect the subsets $C(v)$ induced by the vertices v of N this way in the set $\mathcal{C}(N)$, i.e. we put $\mathcal{C}(N) = \bigcup_{v \in V(N)} \{C(v)\}$. For convenience, we refer to any collection $\mathcal{C}$ of non-empty subsets of X as a *cluster system (on X)* and to the elements of $\mathcal{C}$ as *clusters* of X. It should be noted that in case N is a phylogenetic tree, the cluster system $\mathcal{C}(N)$ is a *hierarchy (on X)*, that is, for any two clusters $C_1, C_2 \in \mathcal{C}(N)$ we have that $C_1 \cap C_2 \in \{\emptyset, C_1, C_2\}$. Hierarchies are sometimes also called *laminar families*, and it is well-known that the cluster systems $\mathcal{C}(T)$ induced by a phylogenetic tree T uniquely determines that tree (see e.g. Semple and Steel (2003, page 51)).

In the context of phylogenetic network construction, the concept of a *weak hierarchy (on X)* was introduced in Bandelt and Dress (1989). These objects are defined as follows. Suppose $\mathcal{C}$ is a cluster system on X. Then $\mathcal{C}$ is called a *weak hierarchy (on X)* if

$$C_1 \cap C_2 \cap C_3 \in \{C_1 \cap C_2, C_2 \cap C_3, C_1 \cap C_3\} \tag{1}$$

holds for any three elements $C_1, C_2, C_3 \in \mathcal{C}$. Note that any hierarchy is in particular a weak hierarchy and that any subset of a weak hierarchy is again a weak hierarchy. Also note that weak hierarchies are well-known objects in classical hypergraph and abstract convexity theory (Bandelt and Dress, 1989) (see also the reference therein and Barthélémy et al (2004)), and that they where originally introduced into cluster analysis as *medinclus* in Batbedat (1988).

We will establish the main result of this section (Proposition 1) by showing that the cluster system $\mathcal{S}(N)$ associated to a level-1 network N is a weak hierarchy. To do this, we first need to complete the definition of the softwired cluster system $\mathcal{S}(N)$ given in the introduction, which relies on the definition of the system $\mathcal{T}(N)$. We will do this next.

Suppose N is a phylogenetic network. Then we say that a phylogenetic tree T is *displayed* by N if the leaf set of T is X and T can be obtained from N via a series of vertex deletions, arc deletions, and vertex suppressions (see also (Huson et al, 2011)). For a vertex v the latter operation is defined as deleting v plus its incoming and outgoing arcs a_1 and a_2, respectively, from N and adding an arc from the tail of a_1 to the head of a_2. The set $\mathcal{T}(N)$ is then the collection of all phylogenetic trees that are displayed by N. Note that in addition to the cluster $C_N(v)$, a cluster system

$$\mathcal{S}_N(v) = \{C_T(v) : T \in \mathcal{T}(N)\}$$

can be associated to every vertex $v \in V(N)$. Clearly, $C_N(v) \in \mathcal{S}_N(v)$ holds for every $v \in V(N)$ and $\mathcal{S}(N) = \bigcup_{v \in V(N)} \mathcal{S}_N(v)$.

To link clusters of X with level-1 networks on X, we say that a cluster C on X is *level-1-consistent* if there exists a level-1 network N such that $C \in \mathcal{S}(N)$. More generally, we say that a cluster system $\mathcal{C}$ is *level-1-consistent* if there exists a level-1 network N such that $\mathcal{C} \subseteq \mathcal{S}(N)$ holds. Thus, the cluster system $\mathcal{S}(N)$ associated to any level-1 network N is level-1-consistent.

Proposition 1 *A level-1-consistent cluster system is a weak hierarchy. In particular, the systems $\mathcal{S}(N)$ and $\mathcal{C}(N)$ associated to a level-1 network N are weak hierarchies.*

Proof: Suppose $\mathcal{C}$ is a level-1-consistent cluster system. Let N denote a level-1 network such that $\mathcal{C} \subseteq \mathcal{S}(N)$. Since, as was remarked above, a subset of a weak hierarchy is again a weak hierarchy, it clearly suffices to show that $\mathcal{S}(N)$ is a weak hierarchy. To observe this, consider the phylogenetic tree T_1 on X obtained from N by randomly deleting for each reticulation vertex of N one of its incoming arcs and suppressing any resulting degree 2 vertex. If this renders the root ρ_N of N to have degree 1, identify ρ_N with its unique child. Construct a phylogenetic tree T_2 on X in a similar way but this time deleting for each

reticulation vertex the other incoming arc. Clearly, $\mathcal{S}(N) = \mathcal{C}(T_1) \cup \mathcal{C}(T_2)$. Since $\mathcal{C}(T_i)$, $i = 1, 2$, is a hierarchy and the union of two hierarchies is a weak hierarchy (Bandelt and Dress, 1989, page 149), the proposition follows. ∎

In general the number of elements in a weak hierarchy on X is at most $\binom{|X|+1}{2}$ (Bandelt and Dress, 1989). However Proposition 1 combined with (Kanj et al, 2008, Lemma 6.8) trivially implies that for the special case of level-1-consistent cluster systems this general bound for weak hierarchies can be improved to a bound that is linear in $|X|$.

We remark in passing that a similarity measure $D_{\mathcal{C}} : X \times X \to \mathbb{R}$ can be associated to any cluster system $\mathcal{C}$ of X by putting $D_{\mathcal{C}}(a, b) = |\{C \in \mathcal{C} : a, b \in C\}|$, $a, b \in X$. Proposition 1 combined with the main result from Bandelt and Dress (1989) implies that any level-1-consistent cluster system $\mathcal{C}$ can be uniquely reconstructed from its associated similarity measure $D_{\mathcal{C}}$. Using the well-known *Farris transform* (see e. g. Semple and Steel (2003, page 149), and Dress et al (2007) for a recent overview) a similarity measure can be canonically transformed into a distance measure $D^{\mathcal{C}}$ on X. For a set Y such a measure is defined as a map from $Y \times Y$ into the non-negative reals that is symmetric, satisfies the triangle inequality, and vanishes on the main diagonal. Distance measures were investigated in Chan et al (2006) from an algorithmical point of view in the context of representing them in terms of an *ultrametric* level-1 network. These networks are generalizations of ultrametric phylogenetic trees in the sense that every path from the root of the network to any of its leaves is of the same length.

We conclude this section with remarking that as the example of the level-2 network N presented in Fig. 4 shows, the result analogous to Proposition 1 does not hold for level-2 networks in general since $\{\{a, b, c\}, \{a, b, d\}, \{b, c, d\}\} \subseteq \mathcal{S}(N)$ and $\{a, b, c\} \cap \{a, b, d\} \cap \{b, c, d\} = \{b\}$ but the intersection of any 2 of the participating 3-sets is of size 2.

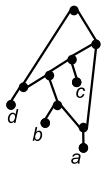
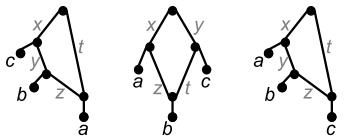


Fig. 4 A level-2 network N for which $\mathcal{S}(N)$ is not a weak hierarchy.

4 Triplet systems induced by simple level-1 networks and by cluster systems

In this section, we turn our attention to triplet system induced by simple level-1 networks and also by cluster systems. In particular, we establish a property of these networks (Lemma 1) which, with regards to the encoding problem for level-1 networks, has turned out to be fundamental. Furthermore we show that a triplet system induced by a phylogenetic tree T is contained in the triplet system induced by a level-1 network N if and only if T is displayed by N (Proposition 2). Along the way, we establish various relationships between the triplet system, the softwired cluster system and the tree system induced by a level-1 network. We start with presenting the definition of the triplet system $\mathcal{R}(N)$ induced by a phylogenetic network N.

For a phylogenetic network N, we say that a triplet $x|yz$ is *consistent* with N if $x, y, z \in X$ and there exist two vertices $u, v \in V(N)$ and pairwise internally vertex-disjoint paths in N from u to y, u to z, v to u and v to x. Note that a triplet system $\mathcal{R}$ is called *consistent* with a phylogenetic network N if every triplet in $\mathcal{R}$ is consistent with N. For convenience, we will sometimes say that a phylogenetic network N is consistent with a triplet t (or a triplet system $\mathcal{R}$) if t (or $\mathcal{R}$) is consistent with N. The set of all triplets consistent with a phylogenetic network N is denoted by $\mathcal{R}(N)$, and we say that N *reflects* a triplet system $\mathcal{R}$ if $\mathcal{R} = \mathcal{R}(N)$. For example and ignoring the arc labels for the moment, each of the three simple level-1 networks $\mathcal{L}^i(\mathcal{R})$ on $X = \{a, b, c\}$, $i \in X$, depicted in Fig. 5 reflects the triplet system $\mathcal{R} = \{a|bc, c|ab\}$.



$\mathcal{L}^a(\mathcal{R})$ $\mathcal{L}^b(\mathcal{R})$ $\mathcal{L}^c(\mathcal{R})$

Fig. 5 Ignoring the arc labels for the moment, the three non-isomorphic simple level-1 networks on $\{a, b, c\}$ that all reflect the triplet system $\mathcal{R} = \{a|bc, c|ab\}$.

Now suppose that N is one of the simple level-1 networks $\mathcal{L}^i(\mathcal{R})$, $i \in X = \{a, b, c\}$, on X depicted in Fig. 5. Assume that $d \notin X$ and let $e = uv \in A(N)$ denote a non-cut arc of N, that is, e is not a cut arc of N. Then we denote by $N_e \oplus d$ the level-1 network obtained from N by adding a new vertex w to $V(N) \cup \{d\}$ and replacing e by the arcs uw, wv, and wd. If the knowledge of e is of no relevance, then we will write $N \oplus d$ rather than $N_e \oplus d$.

The next result is fundamental to the proof of our main result (Theorem 1) as it assures us that although all three simple level-1 networks depicted in Fig. 5 reflect the same triplet system, this property is lost when attaching an additional leaf to a non-cut arc of each of them.

Lemma 1 *Suppose* $X = \{a, b, c, d\}$ *and* $\mathcal{R} = \{a|bc, c|ab\}$. *Then, the triplet systems*

$$\mathcal{R}(\mathcal{L}^a(\mathcal{R}) \oplus d), \;\; \mathcal{R}(\mathcal{L}^b(\mathcal{R}) \oplus d), \;\; \textit{and } \mathcal{R}(\mathcal{L}^c(\mathcal{R}) \oplus d).$$

are all distinct.

Proof: For each of the simple level-1 networks $\mathcal{L}^x(\mathcal{R})$, $x \in \{a, b, c\}$, consider the arc labeling indicated in Fig. 5. With $r := a|bc$ and $r' := c|ab$, we detail the triplet systems induced by the simple level-1 networks obtained from them by attaching a new leaf d in Table 1. Since no two of those systems are the same, the lemma follows. ∎

simple level-1 network	replaced arc	induced triplet system
$\mathcal{L}^a(\mathcal{R})$	x	$\{r, r', a\|bd, a\|cd, d\|ab, d\|ac, d\|bc\}$
	y	$\{r, r', a\|bd, a\|cd, c\|ad, c\|bd, d\|ab\}$
	z	$\{r, r', a\|bd, a\|cd, b\|ad, c\|ad, c\|bd\}$
	t	$\{r, r', b\|ad, c\|ad, d\|ab, d\|ac, d\|bc\}$
$\mathcal{L}^b(\mathcal{R})$	x	$\{r, r', b\|ad, c\|ad, c\|bd, d\|ab, d\|bc\}$
	y	$\{r, r', a\|bd, a\|cd, b\|cd, d\|ab, d\|bc\}$
	z	$\{r, r', a\|bd, b\|ad, c\|ad, c\|bd, d\|bc\}$
	t	$\{r, r', a\|bd, a\|cd, b\|cd, c\|bd, d\|ab\}$
$\mathcal{L}^c(\mathcal{R})$	x	$\{r, r', c\|ad, c\|bd, d\|ab, d\|ac, d\|bc\}$
	y	$\{r, r', a\|bd, a\|cd, c\|ad, c\|bd, d\|bc\}$
	z	$\{r, r', a\|bd, a\|cd, b\|cd, c\|ad, c\|bd\}$
	t	$\{r, r', a\|cd, b\|cd, d\|ab, d\|ac, d\|bc\}$

Table 1 The triplet system induced by attaching a new leaf to one of the non-cut arcs of the simple level-1 networks depicted in Fig. 5.

Turning our attention to cluster systems of X, note that any cluster $C \subsetneq X$ induces a triplet system

$$\mathcal{R}(C) = \{c_1c_2|x : c_1, c_2 \in C \text{ distinct and } x \in X - C\}$$

on X. Thus, any non-empty cluster system $\mathcal{C}$ on X induces the triplet system $\mathcal{R}(\mathcal{C}) := \bigcup_{C \in \mathcal{C} - \{X\}} \mathcal{R}(C)$ on X. Note that $\mathcal{R}(\mathcal{C})$ is dense on X, where a triplet system $\mathcal{R}$ on X is called *dense* if for any three distinct elements $a, b, c \in X$ there exists a triplet $t \in \mathcal{R}$ such that $L(t) = \{a, b, c\}$.

The next result is central for establishing the main result of this section (Proposition 2).

Lemma 2 *Suppose N is a level-k network, $k \geq 0$, with at least 3 leaves. Then the following holds*

(i) If $t \in \mathcal{R}(N)$ then there exists a phylogenetic tree $T \in \mathcal{T}(N)$ with $t \in \mathcal{R}(T)$.
(ii) $\mathcal{R}(N) = \bigcup_{T \in \mathcal{T}(N)} \mathcal{R}(T) = \bigcup_{C \in \mathcal{S}(N)} \mathcal{R}(C)$.

Proof: (i) Suppose $t \in \mathcal{R}(N)$ and assume that $x_1, x_2, x_3 \in X$ with $t = x_1x_2|x_3$. Then there exist distinct vertices $u, v \in V(N)$ such that among the paths from u to x_1, u to x_2, v to u and v to x_3 no two of them share an interior vertex. Then the phylogenetic tree obtained from N by deleting, for any vertex $v \in r(N)$, one of the incoming arcs as specified below and suppressing the resulting degree two vertex is clearly contained in $\mathcal{T}(N)$ and $t \in \mathcal{R}(T)$ holds. If v is a vertex on one of the four paths above, call it P, then delete that incoming arc of v that is not also an arc on P. Otherwise, delete one of the two incoming arcs of v making sure that the resulting graph is a phylogenetic tree on X.

(ii) Clearly $\bigcup_{T \in \mathcal{T}(N)} \mathcal{R}(T) = \bigcup_{T \in \mathcal{T}(N)} \bigcup_{C \in \mathcal{S}(T)} \mathcal{R}(C) = \bigcup_{C \in \mathcal{S}(N)} \mathcal{R}(C)$ holds and it is straightforward to see that $\bigcup_{T \in \mathcal{T}(N)} \mathcal{R}(T) \subseteq \mathcal{R}(N)$. The converse set inclusion follows from (i). ∎

To establish Proposition 2, we require some more terminology. Suppose $\mathcal{R}$ is a triplet system on X and $\mathcal{C}$ is a cluster system on X. Then we associate to $\mathcal{R}$ the cluster system

$$\mathcal{S}(\mathcal{R}) = \{C \subsetneq X : x_1, x_2 \in C \text{ distinct and } x_3 \in X - C \text{ implies } x_1x_2|x_3 \in \mathcal{R}\}$$

on X and to $\mathcal{C}$ the tree system

$$\mathcal{T}(\mathcal{C}) = \{T \in \mathcal{T}(X) : \mathcal{C}(T) \subseteq \mathcal{C}\}$$

on X where $\mathcal{T}(X)$ denotes the space of all phylogenetic trees on X.

Next, suppose that N is a phylogenetic network and that $a, b, c \in V(N)$ are vertices of N with $a \preceq_N b$ and $c \preceq_N b$. Then we call b a *common ancestor* of a and c. A *lowest common ancestor* $lca_N(a, c)$ of a and c is a common ancestor of a and c and no other vertex below $lca_N(a, c)$ is a common ancestor of a and c. Note that in a level-0 or level-1 network N, the lowest common ancestor between any two distinct leaves of N is always unique whereas this need not be the case for level-k networks with larger k. More generally, suppose C is cluster of X. Then a vertex $v \in V(N)$ is called a *lowest common ancestor* of C if v is a common ancestor of every pair of elements in C and no other vertex in N below v satisfies this property. In view of the fact that this vertex is again unique in a level-1 network we will denote it by $lca_N(C)$.

Proposition 2 *Suppose N is a level-k network, $k \geq 0$, with at least 3 leaves. Then $\mathcal{S}(N) \subseteq \mathcal{S}(\mathcal{R}(N))$ and $\mathcal{T}(N) \subseteq \mathcal{T}(\mathcal{S}(N))$. Moreover, if $k \leq 1$ we have*

(i) $\mathcal{S}(N) = \mathcal{S}(\mathcal{R}(N))$.
(ii) $\mathcal{T}(N) = \mathcal{T}(\mathcal{S}(N))$.
(iii) *A phylogenetic tree* T *is displayed by a level-k network* N *if and only if* $\mathcal{R}(T) \subseteq \mathcal{R}(N)$.

Proof: Suppose $C \in \mathcal{S}(N)$. Then $x_1x_2|x_3 \in \mathcal{R}(C)$ holds for any two distinct elements x_1, $x_2 \in C$ and any $x_3 \in X - C$. By Lemma 2, $x_1x_2|x_3 \in \bigcup_{C' \in \mathcal{S}(N)} \mathcal{R}(C') \subseteq \mathcal{R}(N)$ and so $C \in \mathcal{S}(\mathcal{R}(N))$, as required. That $\mathcal{T}(N) \subseteq \mathcal{T}(\mathcal{S}(N))$ holds is trivial.

Assume for the remainder of the proof that $k \leq 1$. Clearly (i) – (iii) hold in case $k = 0$. So assume $k = 1$.

(i) Suppose $C \in \mathcal{S}(\mathcal{R}(N))$. Then $C \neq X$ and there exist distinct elements $a, b \in C$ such that $lca_N(C) = lca_N(a, b)$. Let $x \in X - C$. Then since $C \in \mathcal{S}(\mathcal{R}(N))$, we have $ab|x \in \mathcal{R}(N)$. By Lemma 2(i) there exists a phylogenetic tree $T \in \mathcal{T}(N)$ with $ab|x \in \mathcal{R}(N)$. Since $lca_N(C) = lca_N(a, b) \in V(T)$ it follows that $C \in \mathcal{C}(T) \subseteq \mathcal{S}(N)$.

(ii) Suppose $T \in \mathcal{T}(\mathcal{S}(N))$. For all $v \in V(N)$ put $E_N(v) = \{w \in V(N) : C_N(v) = C_N(w)\}$. Note that $|E_N(v)| = 2$ if and only if v is a reticulation vertex of N or the unique child of such a vertex of N and that $|E_N(v)| = 1$ holds otherwise. Consider the map $\phi : V(T) \to V(N)$ that maps every vertex $v \in V(T)$ to a vertex $w_v := \phi(v)$ in $V(N)$ such that $C(v) = C(w_v)$ and every element $w \in E_N(w_v)$ distinct from w_v lies on a path from the root ρ_N of N to w_v. By the definition of $\mathcal{T}(\mathcal{S}(N))$ it follows that ϕ is well-defined. Moreover, ϕ is clearly injective. In fact, since w_v is never a reticulation vertex of N, the map ϕ induces a bijection between the vertices of T and the vertex set $V(N) - r(N)$.

Let $x, y \in V(T)$ such that xy is an arc in T and $\phi(x)$ and $\phi(y)$ are joined in N by a path P_{xy} of length 2. Then the interior vertex r_{xy} of P_{xy} must be a reticulation vertex of N and every reticulation vertex of N must be the interior vertex of such a path. Let $a_{r_{xy}}$ denote that incoming arc of r_{xy} whose tail is x. Then deleting for all reticulation vertices r of N the incoming arc distinct from a_r (suppressing any resulting degree 2 vertices) results in a tree $T' \in \mathcal{T}(N)$ that is isomorphic with T (in fact T' is T).

(iii) That $\mathcal{R}(T) \subseteq \mathcal{R}(N)$ holds whenever a phylogenetic tree T is displayed by a level-1 network N follows from Lemma 2(ii). Conversely, suppose T is a phylogenetic tree with $\mathcal{R}(T) \subseteq \mathcal{R}(N)$. By (i) and (ii) it suffices to show that $T \in \mathcal{T}(\mathcal{S}(\mathcal{R}(N)))$. Thus, we need to show that $\mathcal{C}(T) \subseteq \mathcal{S}(\mathcal{R}(N))$. So suppose $C \in \mathcal{C}(T)$ and let $x_1, x_2 \in C$ distinct, and $x_3 \in X - C$. Then $x_1x_2|x_3 \in \mathcal{R}(T) \subseteq \mathcal{R}(N)$. Hence, $C \in \mathcal{S}(\mathcal{R}(N))$. ∎

Note that as the example of the level-2 network depicted in Fig. 6 shows, the analogous relationships (i. e. Proposition 2(i) and (iii)) do not hold in general for the corresponding systems.

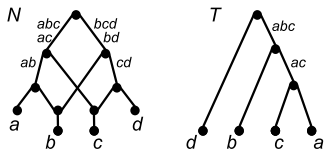


Fig. 6 A level-2 network N with $bc|d, bc|a \in \mathcal{R}(N)$ but $\{b, c\} \notin \mathcal{S}(N)$ and so $\mathcal{S}(\mathcal{R}(N)) \neq \mathcal{S}(N)$. A phylogenetic tree T with $\mathcal{C}(T) \subseteq \mathcal{S}(N)$ and $\mathcal{R}(T) \subseteq \mathcal{R}(N)$ but $T \notin \mathcal{T}(N)$. For ease of verification, we have labelled an arc by the softwired cluster(s) its head induces and also write xyz for a cluster $\{x, y, z\}$.

5 Encodings of level-1 networks

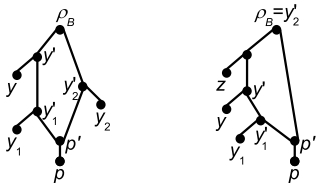
In this section, we characterize those level-1 networks N that are encoded by the triplet system $\mathcal{R}(N)$, or equivalently the tree system $\mathcal{T}(N)$, or equivalently the cluster system $\mathcal{S}(N)$ they induce (Theorem 1). Note in this context that the cluster system $\mathcal{C}(N)$ induced by a level-1 network N is in general not an encoding of N. To see this, consider for example a simple level-1 network N with $n \geq 3$ leaves such that the parent of each leaf lies on the same path from the root ρ_N of N to its reticulation vertex h_N. Then $\rho_N h_N$ is an arc of N and $\mathcal{C}(N) = \mathcal{C}(T)$ holds for the phylogenetic tree $T \in \mathcal{T}(N)$ obtained via deleting the arc $\rho_N h_N$ from N and identifying ρ_N with its unique child.

Bearing in mind that there exist triplet systems which can be reflected by more than one level-1 network, we denote the collection of all level-1 networks that reflect a triplet system $\mathcal{R}$ by $\mathfrak{L}_1(\mathcal{R})$. Clearly, if $\mathcal{R}$ is reflected by a level-1 network N then $N \in \mathfrak{L}_1(\mathcal{R}(N))$ and so $|\mathfrak{L}_1(\mathcal{R}(N))| \geq 1$. Similarly, we denote for a tree system $\mathcal{T}$ the collection of all level-1 networks N for which $\mathcal{T} = \mathcal{T}(N)$ holds by $\mathfrak{L}_1(\mathcal{T})$, and for a cluster system $\mathcal{C}$ the collection of all level-1 networks N for which $\mathcal{C} = \mathcal{S}(N)$ holds by $\mathfrak{L}_1(\mathcal{C})$. As in the case of triplet systems, there exist tree systems $\mathcal{T}$ and cluster systems $\mathcal{C}$ with $|\mathfrak{L}_1(\mathcal{T})| > 1$ and $|\mathfrak{L}_1(\mathcal{C})| > 1$, respectively.

The next lemma serves as a stepping stone in the proof of Theorem 1 and is concerned with analyzing level-1 networks N that have a large enough *leaved* block B, that is, B is a block of N and every arc that starts at a vertex in B ends either in a vertex of B or a leaf of N. Calling those leaves of N also the leaves of B, we remark that a leaved block is a simple level-1 network on its set of leaves.

To present the proof of the lemma, we require some more notation and definitions. Suppose N is a phylogenetic network with at least 3 leaves. Then we call a subset $\{x, y\} \subseteq X$ a *cherry* of N if there exists a vertex $v \in V(N)$ such that $vx, vy \in A(N)$. Now suppose N is a level-1 network and x is either (i) a leaf of a leaved block B of N that is adjacent with a parent of the reticulation vertex of B or (ii) a leaf in a cherry of N. Then we denote by $N - x$ the level-1 network obtained from N by removing x and, in case of (i), its parent y and all their incident arcs and, in case of (ii), its incident arc, both times suppressing resulting degree 2 vertices and, if the roof of N has been rendered a degree 1 vertex, identifying it with its unique child. Also, we say that N is a *strict* level-1 network if N is not a phylogenetic tree and associate to a triplet system $\mathcal{R}$ and some $x \in \bigcup_{t \in \mathcal{R}} L(t)$, the triplet system $\mathcal{R}_x := \{t \in \mathcal{R} : x \notin L(t)\}$. Recall that for a directed graph G, we put $V^-(G) = V(G) - L(G)$.

Obviously, a level-1 network N that has a leaved block B with at least 5 non-leaf vertices must have at least 4 leaves. Also one of the two distinct parent vertices $y_1', y_2' \in V(B)$ of the reticulation vertex p' of B could be the root $\rho = \rho_B$ of B. Note however that y_i' is adjacent with a leaf of B whenever $y_i' \neq \rho$, $i = 1, 2$. We denote that leaf by y_i. Also note that since $|V^-(B)| \geq 5$, at least one of the paths $P_{\rho y_i'}$ from ρ to y_i', $i = 1, 2$, must contain a vertex y' distinct from ρ and y_i. Let y denote the leaf of B adjacent with y'. Note that without loss of generality we may assume that $i = 1$ and that y' is the predecessor of y_1' on the path $P_{\rho y_1'}$. Let p denote the leaf of B adjacent with p'. We depict the two possible configurations for B for the case $|V^-(B)| = 5$ in Fig. 7.



(a) (b)

Fig. 7 The two possible configurations for B in case $|V^-(B)| = 5$ (see text for details).

Lemma 3 *Suppose N is a level-1 network with at least 3 leaves such that, in addition to every block having at least 5 vertices, N also has a leaved block. Then $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$.*

Proof: We prove the lemma by induction on the number n of leaves of N. Suppose that B is a leaved block of N and assume that the notations and assumptions made above for a leaved block of a level-1 network apply to B. To see the base case of the induction assume that $n = 4$. Then B must equal N

and that $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ holds is a straightforward consequence of Lemma 1. To establish the induction step, assume that the induction hypothesis holds for all level-1 networks on $n-1$ leaves, as specified in the statement of the lemma. We distinguish the cases that $|V^-(B)| > 5$ and that $|V^-(B)| = 5$.

Suppose first that $|V^-(B)| = 5$. Then either $\rho = y_2'$ and so B has, in addition to the leaves y_1, y, p, precisely one more leaf z, or $\rho \neq y_2'$ and the leaves of B are y_1, y_2, y and p (cf Fig. 7). We only consider the case $\rho \neq y_2'$ since the arguments for the case $\rho = y_2'$ are similar. Then $B - y_1$ is a phylogenetic tree on the leaves y, p, y_2, i.e. the triplet $t := y|py_2$. Put $t' = y_2|py \in \mathcal{R}(N)$ and $\mathcal{R}^{t'} := \mathcal{R}_{y_1} - t'$. Since $N - y_1$ is either a phylogenetic tree or a strict level-1 network such that each of its blocks has at least 5 vertices, the induction hypothesis implies $|\mathfrak{L}_1(\mathcal{R}(N - y_1))| = 1$. Thus, $N - y_1$ is the unique level-1 network that reflects $\mathcal{R}^{t'}$. Note that the only way to turn $N - y_1$ into a level-1 network that, in addition to reflecting $\mathcal{R}^{t'}$, is also consistent with t' is to replace t by one of the 3 level-1 networks depicted in Fig. 5 with y playing the role of a, p playing the role of b and y_2 playing the role of c. If with $\mathcal{R} = \{t, t'\}$ we had that that simple level-1 network B' were the network $\mathcal{L}^c(\mathcal{R})$ then the triplet $p|y_1y$ would not be contained in $\mathcal{R}(N)$ which is impossible. Also since $\{p|yy_1, y_1|py_2, y|py_1\} \subseteq \mathcal{R}(N)$ it is impossible for B' to be the network $\mathcal{L}^a(\mathcal{R})$. Consequently B' must be the network $\mathcal{L}^b(\mathcal{R})$. Since $y_1|py$, $y|py_1$, and $y_2|y_1p$ are triplets in $\mathcal{R}(N)$, Lemma 1 implies that the only way to transform $N - y_1$ into a level-1 network that in addition to including y_1 in its leaf set also reflects $\mathcal{R}(N)$ is to subdivide the arc y_2p of B' by a vertex v and adding the arc vy_1 to the arc set of $N - y_1$. But that network is N and so $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ must hold.

Now assume that $|V^-(B)| > 5$ holds. Let $N \ominus y_1$ denote the level-1 network obtained from N by adding an arc from the tail of the incoming arc a of y_1' to p' and deleting y_1, y_1' and the arcs $y_1'y_1$, $y_1'p'$ and a. Then since every block in the level-1 network $N \ominus y_1$ clearly has at least 5 vertices, the induction hypothesis implies $|\mathfrak{L}_1(\mathcal{R}(N \ominus y_1))| = 1$. Combined with the fact that $y|y_1p$, $y_1p|y_2$ and $y_1|py_2$ are triplets in $\mathcal{R}(N)$ it follows that $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ in case $\rho \neq y_2'$. If $\rho = y_2'$ then $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ follows from the fact that $y|y_1p, p|yy_1 \in \mathcal{R}(N)$. ∎

We are now ready to prove our main result.

Theorem 1 *Suppose N is a level-1 network with at least 3 leaves. Then the following statements are equivalent*

(i) *N has a block with four vertices.*
(ii) $|\mathfrak{L}_1(\mathcal{R}(N))| > 1$.
(iii) $|\mathfrak{L}_1(\mathcal{S}(N))| > 1$.
(iv) $|\mathfrak{L}_1(\mathcal{T}(N))| > 1$.

Proof: (i) $\Rightarrow$ (iv): This is an immediate consequence of the fact that all simple level-1 networks depicted in Fig. 5 induce the same set of phylogenetic trees.

(iv) $\Rightarrow$ (iii): Suppose that N is a level-1 network with $|\mathfrak{L}_1(\mathcal{T}(N))| > 1$. Then there exists a level-1 network $N' \in \mathfrak{L}_1(\mathcal{T}(N))$ distinct from N with $\mathcal{T}(N) = \mathcal{T}(N')$. But then $\mathcal{S}(N) = \bigcup_{T\in\mathcal{T}(N)} \mathcal{C}(T) = \bigcup_{T\in\mathcal{T}(N')} \mathcal{C}(T) = \mathcal{S}(N')$ and so $N' \in \mathfrak{L}_1(\mathcal{S}(N))$. Thus, $|\mathfrak{L}_1(\mathcal{S}(N))| > 1$.

(iii) $\Rightarrow$ (ii): Suppose that N is a level-1 network with $|\mathfrak{L}_1(\mathcal{S}(N))| > 1$. Then there exists a level-1 network $N' \in \mathfrak{L}_1(\mathcal{S}(N))$ distinct from N such that $\mathcal{S}(N) = \mathcal{S}(N')$. But then Lemma 2(ii) implies

$$\mathcal{R}(N) = \bigcup_{C\in\mathcal{S}(N)} \mathcal{R}(C) = \bigcup_{C\in\mathcal{S}(N')} \mathcal{R}(C) = \mathcal{R}(N')$$

and so $N' \in \mathfrak{L}_1(\mathcal{R}(N))$. Hence, $|\mathfrak{L}_1(\mathcal{R}(N))| > 1$.

(ii) $\Rightarrow$ (i) We will show by induction on the number n of leaves of N that if every block in N has at least 5 vertices then $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$. Suppose N is a level-1 network with n leaves such that every block of N has at least 5 vertices. Note that we may assume that N has at least one such block since otherwise N is a phylogenetic tree and so $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ clearly holds. But then $n \geq 4$. If $n = 4$ then $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ is a straightforward consequence of Lemma 1.

Suppose $n > 4$. Assume for every level-1 network N_0 with $n_0 < n$ leaves that $|\mathfrak{L}_1(\mathcal{R}(N_0))| = 1$ holds whenever N_0 is a phylogenetic tree or every block in N_0 has at least 5 vertices. We distinguish the cases that N has a cherry and that it does not. Clearly, if N does not have a cherry then it must have a leaved block and so $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ follows by Lemma 3.

Now suppose that N has a cherry $\{x, y\} \subseteq X$ and assume for contradiction that there exists a level-1 network N' in $\mathfrak{L}_1(\mathcal{R}(N))$ distinct from N. Without loss of generality, we may assume that this cherry is as far away from the root of N as possible. Then since N is a strict level-1 network all of whose blocks have at least 5 vertices, $N - x$ must enjoy the same property with regards to its blocks. But then, by induction hypothesis, $|\mathfrak{L}_1(\mathcal{R}(N - x))| = 1$ and so $N - x$ is the unique level-1 network that reflects $\mathcal{R}(N - x) = \mathcal{R}_x$. Since by the choice of x, for every leaf z in N distinct from x and y, only the triplet $z|xy$ out of the 3 possible triplets on $\{x, y, z\}$ is contained in $\mathcal{R}(N) = \mathcal{R}(N')$, it follows that $\{x, y\}$ must also be a cherry in N'. But then $N = N'$ which is impossible. Thus, $|\mathfrak{L}_1(\mathcal{R}(N))| = 1$ must hold in this case too which completes the proof of the theorem. ∎

Note that the implication $(ii) \Rightarrow (i)$ in the proof of Theorem 1 can also be obtained using an alternative 2-phase strategy that is based on so called SN-sets which were originally introduced in Jansson and Sung (2006). For any

triplet system $\mathcal{R}$ on X and any subset $Y \subseteq X$, the SN-set associated to Y is recursively defined as $SN(Y) = SN(Y \cup \{c\})$ if there exists some $y, y' \in Y$ distinct and some $c \in X - Y$ such that $yc|y' \in \mathcal{R}$, and $SN(Y) = Y$ otherwise. In Jansson and Sung (2006), it was shown that in case $\mathcal{R}$ is dense, a rooted leaf-labelled tree can be associated to the set $\Sigma_{\mathcal{R}}$ of strongly connected components of a certain graph $G_{\mathcal{R}}$ that can be associated to the SN-sets of X of the form $SN(\{a, b\})$ with $a, b \in X$ distinct. This tree is sometimes called the *SN-tree* $\mathcal{T}_{\mathcal{R}}$ associated to $\mathcal{R}$ and is used in (Jansson and Sung, 2006) to compute a level-1 network from such a triplet system in $O(|X|^3)$ time. Then the first step of the alternative strategy consists of establishing that the edge set of $\mathcal{T}_{\mathcal{R}(N)}$ is in bijective correspondence with the set of cut arcs of any level-1 network N' with $\mathcal{R}(N') = \mathcal{R}(N)$. Ignoring the blocks of N for the moment by viewing each one of them as being collapsed into a single vertex this implies that the structure of N is uniquely determined by $\mathcal{R}(N)$. The second and final step is concerned with establishing the structure of the blocks of N. But this is a consequence of Lemma 1 since by collapsing for each vertex v of a block of N the set of vertices of N that are reachable from v by crossing the cut arc of N incident with v into a single vertex, and making the corresponding adjustments to the triplet system $\mathcal{R}(N)$, the block becomes a simple level-1 network and the triplet system a triplet system on the leaf set of that block.

Theorem 1 immediately implies the following corollary about the number of networks that reflect $\mathcal{R}(N)$.

Corollary 1 *Let N be a level-1 network with at least 3 leaves. Then the number of non-isomorphic level-1 networks N' that reflect $\mathcal{R}(N)$ (or equivalently for which $\mathcal{T}(N) = \mathcal{T}(N')$ or equivalently $\mathcal{S}(N) = \mathcal{S}(N')$ holds) is 3^b, where b is the number of blocks of N that have 4 vertices.*

Returning to the problem of encodings of level-1 networks, we remark that phylogenetic trees on X can also be viewed as trees together with a bijective labelling map between X and the leaf set of such trees. Taking this point of view, phylogenetic trees were generalized in Moulton and Huber (2006) to *MUL-trees* by allowing two or more leaves of such a tree to have the same label. Note that in Fellows et al (2003) such trees are called *rl-trees*. For example, the tree obtained from the phylogenetic tree depicted in Fig. 1(c) by replacing the leaf labelled a by the cherry labelled $\{a, b\}$ is such a tree. In fact, this MUL-tree is the MUL-tree induced by the level-1 network N depicted in Fig. 3(b) i. e. it contains all paths from the root of N to all leaves of N. For a level-1 network N it is easily seen that the MUL-tree $\mathcal{M}(N)$ induced by N this way is in fact an encoding of N in the sense that N is the unique level-1 network that can give rise to $\mathcal{M}(N)$.

6 Metrics

The problem of comparing phylogenetic networks has recently received a considerable amount of attention in the literature resulting in e.g. the definition of metrics for so called *time consistent tree child* and *time consistent tree sibling networks* see e.g. (Cardona et al, 2009a,b) and also (Huson et al, 2011) for a recent overview). Denoting the class of all phylogenetic networks on X by $\mathcal{N}$, and the subclass of all level-1 network on X which do not contain a block with 4 vertices by $\mathcal{C}_1^-$, we carry this theme further by establishing in this section that 3 distance measures that were originally introduced in (Huson et al, 2011) are in fact metrics on $\mathcal{C}_1^-$. In addition we present the diameter of two of them on that subclass. Before we start, we remark that since level-1 networks contained in $\mathcal{C}_1^-$ are allowed to contain blocks B that contain an arc from the root of B to the reticulation vertex of B, the class C_1^- is different from the class of time consistent tree-child or time-consistent tree-sibling networks.

To start, recall that a distance measure D on $\mathcal{N}$ is called a *metric* if the following property is satisfied for all networks $N_1, N_2 \in \mathcal{N}$:

(2) $D(N_1, N_2) = 0$ if and only if N_1 and N_2 are isomorphic.

Note that Property 2 is called the *separation property* in (Cardona et al, 2009a,b). Also note that a distance measure is sometimes called a metric and a distance measure that satisfies Property 2 a proper metric.

Two types of distance measures for phylogenetic networks that have their origin in the problem of comparing phylogenetics trees are the *triplet distance* D_{tri} and the *Robinson-Foulds distance* D_{RF}. For T_1 and T_2 two phylogenetic trees (i.e. level-0 networks) the former is defined as

$$D_{tri}(T_1, T_2) = |\mathcal{R}(T_1) \Delta \mathcal{R}(T_2)|/2$$

and the latter as

$$D_{RF}(T_1, T_2) = |\mathcal{S}(T_1) \Delta \mathcal{S}(T_2)|/2.$$

where for any two set A and B the symmetric difference between A and B is denoted by $A \Delta B$. Note however that both distance measures are different from the Robinson Foulds distance and the triplet distance introduced in (Cardona et al, 2009a) and (Cardona et al, 2009b), respectively. Also note that both can be canonically extended to obtain distance measures D_{tri} and D_{RF} on $\mathcal{N}$ by replacing T_1 and T_2 by phylogenetic networks N_1 and N_2, respectively.

Complementing the above two distance measures, a further distance measure D_{tree} on $\mathcal{N}$ was introduced in Huson et al (2011). For two phylogenetic networks N_1 and N_2 in $\mathcal{N}$, this distance measure is defined as

$$D_{tree}(N_1, N_2) = |\mathcal{T}(N_1) \Delta \mathcal{T}(N_2)|/2.$$

In view of Theorem 1, we immediately obtain

Corollary 2 *The distance measures D_{tri}, D_{RF} and D_{tree} are metrics on $\mathcal{C}_1^-$.*

To better understand the range of values a metric D on $\mathcal{N}$ (or a subclass $\mathcal{C}$ of $\mathcal{N}$) can attain, the *diameter* $diam(D, \mathcal{C})$ of D is sometimes used which is defined as

$$diam(D, \mathcal{C}) := \max\{D(N_1, N_2) : N_1, N_2 \in \mathcal{C}\}.$$

To establish our main result of this section (Theorem 2), we first introduce a leaf-labelled rooted DAG that is central for establishing $diam(D_{RF}, \mathcal{C}_1^-)$. Let $\mathcal{G}(n) = (G(n), \phi)$ denote the leaf-labelled graph on $n := |X| \geq 3$ leaves with $\phi : X \to L(G(n))$ a bijective map and $G(n)$ a rooted DAG that satisfies the following property: In addition to enjoying Properties (P2) and (P3) (and thus also (P1) once we have identified X with $L(G(n))$), $G(n)$ consists of precisely one leaved block on n leaves that has an arc joining its root and its reticulation vertex. Note that once the aforementioned identification has been carried out, $\mathcal{G}(n)$ is clearly a level-1 network. Also note that the networks $\mathcal{L}^a(\mathcal{R})$ and $\mathcal{L}^c(\mathcal{R})$ on $X = \{a, b, c\}$ depicted in Fig. 5 are examples of $\mathcal{G}(3)$. For clarity of our arguments and in case the knowledge of the leaf-labelling map ϕ is of no relevance, we omit it from our discussion. In this case we simply view $\mathcal{G}(n)$ as a level-1 network.

Concerning $\mathcal{S}(\mathcal{G}(n))$, we remark that a phylogenetic tree T with $n \geq 3$ leaves is known to have $2n-1$ vertices and thus induces $2n-1$ distinct clusters. Since $|\mathcal{T}(\mathcal{G}(n))| = 2$ and the two trees contained in $\mathcal{T}(\mathcal{G}(n))$ only have $n+1$ clusters in common (i.e. X and its singletons) it follows that $|\mathcal{S}(\mathcal{G}(n))| = 3n-3$. Moreover, the following result holds.

Proposition 3 *Suppose N is a level-1 network with $n \geq 3$ leaves. Then*

$$|\mathcal{S}(N)| \leq |\mathcal{S}(\mathcal{G}(n))| = 3n - 3$$

Proof: We prove the proposition by induction on n. Clearly the stated inequality holds for $n = 3$. So assume that it holds up to and including some $n \geq 3$ and let N be a level-1 network on $n+1$ leaves. We distinguish the cases that N has a cherry (i) and that N does not have a cherry (ii).

(i): Suppose $\{x, y\}$ is a cherry of N. Then deleting x plus its incident arc (suppressing the resulting degree 2 vertex) results in a level-1 network N^- on n leaves. Clearly $|\mathcal{S}(N)| = |\mathcal{S}(N^-)| + 2$ and so, by induction hypothesis,

$$|\mathcal{S}(N)| = |\mathcal{S}(N^-)|+2 \leq |\mathcal{S}(\mathcal{G}(n))|+2 = 3n-3+2 \leq 3(n+1)-3 = |\mathcal{S}(\mathcal{G}(n+1))|$$

(ii): Since N does not have a cherry it must contain a leaved block B with $|V(B)| \geq 4$. Consider first the case that $|V(B)| \geq 5$. Let x be a leaf

of N that is adjacent to a vertex of B that is not the reticulation vertex h_B of B. Then deletion of x plus its incident arc (suppressing the resulting degree 2 vertex in B) results in a level-1 network N^- on n leaves. Clearly $|\mathcal{S}(N)| = |\mathcal{S}(N^-)| + 3$. Arguments similar to the ones used in Case (i) imply that $|\mathcal{S}(N)| \leq |\mathcal{S}(\mathcal{G}(n+1))|$.

Finally assume that $|V(B)| = 4$. Let x be the leaf of N that is adjacent with h_B. Then deleting x and h_B plus their incident arcs (suppressing resulting degree 2 vertices in B) results again in a level-1 network N^- on n leaves. Note that $|\mathcal{S}(N)| = |\mathcal{S}(N^-)| + 3$ again holds. Proceeding as in the case that $|V(B)| \geq 5$ yields $|\mathcal{S}(N)| \leq |\mathcal{S}(\mathcal{G}(n+1))|$. ∎

We next introduce a leaf-labelled DAG that has turned out to be central for establishing the diameter of D_{tree}. Let $\mathcal{G}'(n) = (G'(n), \phi')$ denote the leaf-labelled graph on $n := |X| \geq 4$ leaves with $\phi' : X \to L(G'(n))$ a bijective map and $G'(n)$ a rooted DAG that satisfies the following property: In addition to enjoying Properties (P2) and (P3) (and thus also (P1) once we have identified X with $L(G'(n))$), $G'(n)$ consists of $l := \lfloor \frac{n-1}{3} \rfloor$ blocks B_i, $1 \leq i \leq l$, with $|V(B_i)| = 5$ and an arc joining the root ρ_i of B_i with its reticulation vertex h_i and $h_i\rho_{i+1} \in A(G'(n))$, $1 \leq i \leq l-1$. In addition, $G'(n)$ has $((n-1) \mod 3)$ extra leaves one of which is a child of the root $\rho_{G'(n)}$ of $G'(n)$ and the distance of the other leaf to $\rho_{G'(n)}$ is 3. Note that once the aforementioned identification has been carried out, $\mathcal{G}'(n)$ is clearly a level-1 network. We present two examples of $\mathcal{G}'(12)$ with $X = \{x_1, \ldots, x_{12}\}$ in Fig. 8.

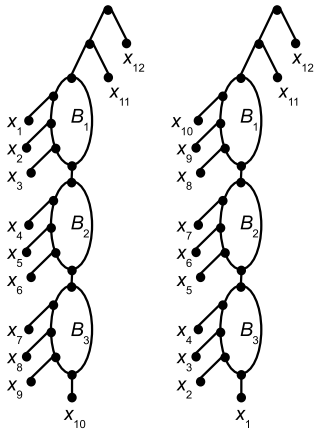


Fig. 8 Two examples of $\mathcal{G}'(12)$ on $X = \{x_1, \ldots, x_{12}\}$ – see the proof of Theorem 2 for details.

We are now ready to establish the aforementioned diameter results.

Theorem 2 *Suppose $n \geq 4$. Then*

(i) $diam(D_{tree}, \mathcal{C}_1^-) = 2^{\lfloor \frac{n-1}{3} \rfloor}$.
(ii) $diam(D_{RF}, \mathcal{C}_1^-) = 2n - 4$.

Proof: (i) We first establish that $diam(D_{tree}, \mathcal{C}_1^-) \leq 2^{\lfloor \frac{n-1}{3} \rfloor}$ must hold. To this end, note first that $diam(D_{tree}, \mathcal{C}_1^-) \leq \max\{|\mathcal{T}(N)| : N \in \mathcal{C}_1^-\}$. Next note that for any network $N \in \mathcal{C}_1^-$ we have $|\mathcal{T}(N)| \leq 2^b$ where b is the number of blocks in N. Also note that every network $N \in \mathcal{C}_1^-$ can be transformed into a (multifurcating) phylogenetic tree T_N by collapsing every block B of N into a vertex v_B with $|V(B)| - 1$ outgoing arcs. Since a multifurcating phylogenetic tree on n leaves can have at most $\lfloor \frac{n-1}{3} \rfloor$ non-leaf vertices of outdegree 4 the stated bound follows.

To see that this bound is sharp, put $X = \{x_1, \ldots, x_n\}$, $n \geq 4$, and assume that $G'(n)$ is embedded in the plane as indicated in either one of the two networks depicted in Fig. 8 (ignoring the leaf labelling for the moment). Moreover and starting from the unique leaf of $G'(n)$ below the reticulation vertex furthest away from the root $\rho_{G'(n)}$ of $G'(n)$, proceed in clockwise fashion to enumerate the leaves of $G'(n)$ by $1, 2, \ldots, n$ with the leaf that is a child of $\rho_{G'(n)}$ receiving n (if it exists). As above, put $l := \lfloor \frac{n-1}{3} \rfloor$.

Let N_1 denote the graph $\mathcal{G}'(n) = (G'(n), \phi_1')$ with leaf labelling map $\phi_1' : X \to L(G'(n))$ defined as $\phi_1'(x_i) := i$, $1 \leq i \leq n$. Let N_2 denote the graph $\mathcal{G}'(n) = (G'(n), \phi_2')$ with leaf labelling map $\phi_2' : X \to L(G'(n))$ defined as $\phi_2'(x_i) = 3l + 2 - i$, $1 \leq i \leq 3l + 1$, and $\phi_2'(x_i) = \phi_1'(x_i)$ otherwise. Note that the latter case only applies if $n \neq 3l + 1$. (In Fig. 8, we depict N_1 and N_2 for the case $X = \{x_1, \ldots, x_{12}\}$). Then it is easy to see that $\mathcal{T}(N_1) \cap \mathcal{T}(N_2) = \emptyset$. Since $|\mathcal{T}(N_i)| = 2^l$, $i = 1, 2$, we have $D_{tree}(N_1, N_2) = 2^l$.

(ii) That $2n-4$ is an upper bound for $diam(D_{RF}, \mathcal{C}_1^-)$ follows from Proposition 3 as each level-1 network contained in $\mathcal{C}_1^-$ induces at most $3n-3$ softwired clusters (including X and its singletons). Hence, for any two level-1 networks $N_1, N_2 \in \mathcal{C}_1^-$ we have $|\mathcal{N}_1 \Delta \mathcal{N}_2| \leq 2(3n - 3) - 2 - 2n = 4n = 4n - 8$ which immediately implies the stated upper bound.

To see that the bound is sharp put $X = \{x_1, \ldots, x_n\}$, $n \geq 4$, and assume that $G(n)$ is embedded in the plane such that for every interior vertex v on the path from the root $\rho_{G(n)}$ of $G(n)$ to the reticulation vertex $h_{G(n)}$ of $G(n)$ the left child of v always a leaf. Starting from the leaf below $h_{G(n)}$ proceed in clockwise fashion to enumerate the leaves of $G(n)$ by $1, 2, \ldots, n$.

Let N_1 denote the graph $\mathcal{G}(n) = (G(n), \phi_1)$ with leaf-labelling map $\phi_1 : X \to L(G(n))$ defined as $\phi_1(x_i) = i$, $1 \leq i \leq n$. Let N_2 denote the graph $\mathcal{G}(n) = (G(n), \phi_2)$ with leaf-labelling map $\phi : X \to G(n)$ defined as $\phi(x_{\sigma(i)}) = i$, $1 \leq i \leq n$, where $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is the permutation given by

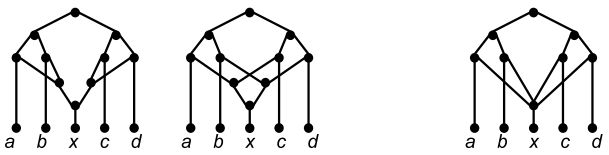
$\sigma(i) = i/2$ if i is even and $\lfloor \frac{n}{2} \rfloor + \frac{i+1}{2}$ otherwise. Then it is not difficult to see that X and its singletons are the only clusters common to $\mathcal{S}(N_1)$ and $\mathcal{S}(N_2)$. By Proposition 3, $D_{RF}(N_1, N_2) = (2(3n-3) - 2n) - 2)/2 = 2n - 4$ follows. ∎

We remark that the bound for D_{RF} is also attained for 2 rooted caterpillar phylogenetic trees on n leaves (i. e. a rooted phylogenetic tree on n leaves with a unique cherry) where the leaf ordering of one is the reverse of the leaf ordering of the other.

7 Encodings and indistinguishability

In this section, we turn our attention to encodings of level-k networks with $k \geq 2$. Our main result is summarized in Proposition 5 where we present examples of level-2 networks that are encoded by some of the induced systems of interest in this paper but not by others.

To start, consider the two simple level-3 networks depicted in Fig. 9(i) and (ii), respectively, which originally appeared in (Huson et al, 2011) in slightly different form (see also (Moret et al, 2004)). As can be quickly checked, both networks induce the same hardwired cluster system and also the same tree system (and thus also the same softwired cluster system and the same triplet system). Consequently, neither of them encodes those networks. By canonically extending the definitions of the above four systems plus their surrounding terminology to also apply to the leaf labelled rooted DAG depicted in Fig. 9(iii) it is easy to check that the four systems induced by that graph coincide with the corresponding systems induced by the networks in Fig. 9(i) and (ii).



(i) (ii) (ii)

Fig. 9 Two indistinguishable simple level-3 networks (i,ii), and a multicombination network resulting from a series of multicombination contractions applied to either one of them (iii).

To also include such graphs in our discussion, we now extend our definition of a phylogenetic network N (and thus also the definition of the set $r(N)$) to allow reticulation vertices to have indegree 2 or more. However and in case of ambiguity, we follow Huson et al (2011) and refer to a reticulation vertex with indegree strictly greater than 2 as a *multicombination*, and to a

phylogenetic network containing such reticulation vertices as a *multicombining network*. In fact, any phylogenetic network N that contains arcs of the form r_2r_1 with $r_1, r_2 \in r(N)$ can be transformed into a multicombining network using a series of *multicombination contractions*, that is, contractions of the arcs r_2r_1. Extending the definition of a level of a phylogenetic network to a multicombining network N by defining the level of N as the maximum sum, among all blocks B of N, of the indegrees, minus one, of all vertices in $B \cap r(N)$, it should be noted that this operation is level preserving. Within this more general framework, we formalize the notion of indistinguishability which was already indicated in (Huson et al, 2011) as follows: We say that two phylogenetic networks N and N' are *indistinguishable* if $\mathcal{C}(N) = \mathcal{C}(N')$ and $\mathcal{T}(N) = \mathcal{T}(N')$ hold and *distinguishable* otherwise. Note that $\mathcal{S}(N) = \mathcal{S}(N')$ and $\mathcal{R}(N) = \mathcal{R}(N')$ must hold whenever two phylogenetic networks N and N' are indistinguishable.

It is easy to see, that the phylogenetic networks presented in Figs. 9(i) and (ii) give rise to the multicombining network depicted in Fig. 9(iii) using a series of multicombination contractions. That the observed indistinguishability property holds is not a coincidence as the following result shows.

Proposition 4 *Suppose N is a level-k network (possibly containing multicombination vertices) with $k \geq 2$ and N' is a multicombining network that can be obtained from N by a series of multicombination contractions. Then N and N' are indistinguishable.*

Proof: Since N' is a multicombining network that can be obtained from N by a series of multicombination contractions, it suffices to show that at each step the hardwired cluster system and the tree system induced by N is preserved. Suppose $r_1, r_2 \in r(N)$ with $r_2r_1 \in A(N)$. Assume without loss of generality that N' is the network resulting from a multicombination contraction of r_2r_1 and denote the generated vertex by r. Let $y_1, \ldots, y_t \in V(N)$, $t \geq 2$, denote the parents of r_2, and let $r_2, z_1, \ldots, z_{t'} \in V(N)$, $t' \geq 2$, denote the parents of r_1. Furthermore, let x denote the (unique) child of r_1. We illustrate these configurations for the case $t = 3$ and $t' = 2$ in Fig. 10(a) and (b), respectively. For a positive integer m, put $[m] := \{1, \ldots, m\}$.

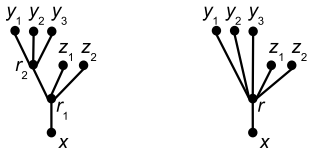


Fig. 10 The situation for the networks N and N' – see the proof of Proposition 4 for details.

To see that $\mathcal{C}(N) = \mathcal{C}(N')$ holds, note that in N, the set of leaves simultaneously below y_i, $i \in [t]$, z_i, $i \in [t']$, r_1, and r_2 equals the set of leaves below x. In N', the set of leaves simultaneously below y_i, $i \in [t]$, z_i, $i \in [t']$, and r equals the set of leaves below x. Moreover, the set of leaves below x in N and N' is the same. Hence $\mathcal{C}(N) = \mathcal{C}(N')$.

We next show that $\mathcal{T}(N) = \mathcal{T}(N')$ holds. Suppose $T \in \mathcal{T}(N)$ and assume that $y_j r_2 \in A(N)$ is the unique arc in N that was not deleted to obtain N'. For the arcs in N with head r_1, we have the following two cases:

(i) If $r_2 r_1$ is the arc that was not deleted in N then T is displayed by N' as all arcs $y_i r \in A(N')$ with $i \in [t] - j$ and all arcs $z_i r \in A(N')$ with $i \in [t']$ may be deleted from N'.

(ii) If $z_{j'} r_1$ with $j' \in [t']$ is the arc which was not deleted in N then T is displayed by N', as all arcs $y_i r \in A(N')$ with $i \in [t]$ and all arcs $z_i r \in A(N')$ with $i \in [t'] - j'$ may be deleted from N'.

Thus, $T \in \mathcal{T}(N')$ follows in both cases. Conversely, for any tree $T \in \mathcal{T}(N')$, we need to delete all but one arc in $\bigcup_{i\in[t]}\{y_i r\} \cup \bigcup_{i\in[t']}\{z_i r\}$ from N'. For any of the $t + t'$ possibilities and by reversing our arguments in the above analysis, we can determine which arcs to delete from N (i. e. all but one with head r_2 and all but one with head r_1) to see that $T \in \mathcal{T}(N)$ holds. Hence, $\mathcal{T}(N) = \mathcal{T}(N')$, as required. Thus, N and N' are indistinguishable. ∎

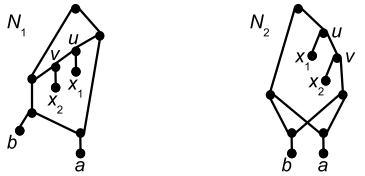
In consequence, two level-k, $k \geq 2$, networks N_1 and N_2 are indistinguishable if the multicombining networks N_i' obtained from N_i, $i = 1, 2$, at the point when the respective multicombination contraction series stabilize are isomorphic.

We next present 2 examples of distinguishable level-2 networks. For the networks considered in each example the induced triplet system is the same. However in one case the induced softwired cluster systems coincide whereas in the other they do not (Proposition 5). The example presented in Proposition 5(ii) is of particular interest in the context of Theorem 1 as it implies that our arguments for establishing that theorem do not readily translate into arguments for level-k networks with $k \geq 2$. More precisely, they show that adding additional leaves to both networks that make up that example by subdividing the arc from x_1's parent to x_2's, and attaching additional leaves, results in two distinct level-2 networks that still reflect the same triplet system.

Proposition 5 *(i) There exist two non-isomorphic distinguishable simple level-2 networks N_1 and N_2 that have the same number of edges, vertices, and reticulation vertices, and $\mathcal{R}(N_1) = \mathcal{R}(N_2)$ and $\mathcal{S}(N_1) \neq \mathcal{S}(N_2)$ hold.*

(ii) There exist two non-isomorphic distinguishable simple level-2 networks N_1 and N_2 that have the same number of edges, vertices, reticulation vertices, and $\mathcal{R}(N_1) = \mathcal{R}(N_2)$ and $\mathcal{S}(N_1) = \mathcal{S}(N_2)$ hold.

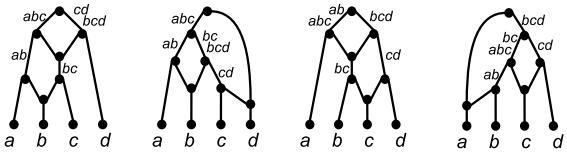
Proof: (i) The networks N_1 and N_2 depicted in Fig. 11 induce the same triplet system $\{a|x_1b,\ b|x_1a,\ x_1|ab,\ a|x_2b,\ b|x_2a,\ x_2|ab,\ x_1|x_2a, a|x_1x_2,\ x_1|x_2b,\ b|x_1x_2\}$ but different softwired cluster systems. More precisely, $\mathcal{S}(N_1) = \{\{a\}, \{b\}, \{x_1\}, \{x_2\}, \{a, x_2\}, \{b, x_2\}, \{x_1, x_2\}, \{a, b, x_2\}, \{a, x_1, x_2\}, \{b, x_1, x_2\}, \{a, b, x_1, x_2\}\}$ and $\mathcal{S}(N_2) = \mathcal{S}(N_1) \cup \{\{a, x_2\}\}$



(i) (ii)

Fig. 11 Two non-isomorphic distinguishable simple level-2 networks that induce the same triplet systems but different softwired cluster systems.

(ii) The four networks presented in Fig. 12 induce the same triplet system $\{a|bc,\ a|bd,\ a|cd,\ b|cd,\ c|ab,\ d|ab,\ d|ac,\ d|bc\}$ and the same softwired cluster system $\{\{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{b, c\}, \{c, d\}, \{a, b, c\}, \{b, c, d\}, \{a, b, c, d\}\}$. ∎



(i) (ii) (iii) (iv)

Fig. 12 Four non-isomorphic distinguishable simple level-2 networks that induce the same triplet and softwired cluster systems. Again, the conventions from Fig. 6 apply.

8 Conclusion

In this paper we have investigated under what circumstances a level-1 network N is encoded by the triplet system, softwired cluster system, or tree system it induces. In particular we have shown that this is the case if and only if N

does not have a block with 4 vertices. For the resulting subclass $\mathcal{C}_1^-$ of level-1 networks this implies that three known distance measures for phylogenetic networks are in fact metrics on $\mathcal{C}_1^-$ and for two of them we have established their diameter. Along the way we have shown that a softwired cluster system induced by a level-1 network is a weak hierarchy, that a phylogenetic tree T is displayed by a level-1 network N if and only if $\mathcal{R}(T) \subseteq \mathcal{R}(N)$ holds and that a level-2 network may be encoded by some of the systems of interest in this paper without being encoded by the other ones. In fact our results show that many of the properties that we establish for level-1 networks are not enjoyed by level-2 networks and thus networks of higher level.

Regarding the accurate reconstruction of phylogenetic networks from triplets or phylogenetic trees, our results have profound consequences as they imply that any such network N that contains a block of size 4 cannot be encoded by any one of the three systems of interest in this paper. Furthermore if N contains one of the networks discussed in the context of Proposition 5 then N is not encoded by $\mathcal{R}(N)$ and possibly $\mathcal{S}(N)$.

References

Arenas M, Valiente G, Posada D (2008) Characterization of reticulate networks based on the coalescent. Molecular Biology and Evolution 25:2517–2520

Bandelt HJ, Dress AWM (1989) Weak hierarchies associated with similarity measures: an additive clustering technique. Bulletin of Mathematical Biology 51:113–166

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human population using median networks. Genetics 141:743–753

Barthélémy JP, Brucker F, Osswald C (2004) Combinatorial optimization and hierarchical classifications. 4OR: A Quaterly Journal of Operations Research 2(3):179–219

Batbedat A (1988) Les isomorphismes HTE et HTS, après la bijection de Benzécri-Johnson. Metron 46:47–59

Bryant D, Moulton V (2004) NeighborNet: An agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution 21(2):255–265

Cardona G, Llabrés M, Rosselló F, Valiente G (2008) A distance metric for a class of tree-sibling phylogenetic networks. Bioinformatics 24(13):1481–1488

Cardona G, Llabres M, Rossello F, Valiente G (2009a) Metrics for phylogenetic networks I: Generalization of the robinson-foulds metric. IEEE/ACM Transactions in Computational Biology and Bioinformatics 6(1):46 – 61

Cardona G, Llabres M, Rossello F, Valiente G (2009b) Metrics for phylogenetic networks II: Nodal and triplets metrics. IEEE/ACM Transactions in Computational Biology and Bioinformatics 6(3):454 – 469

Chan HL, Jansson J, Lam TW, Yiu SM (2006) Reconstructing an ultrametric galled phylogenetic network from a distance matrix. Journal of Bioinformatics and Computational Biology 4(4):807–832

Choy C, Jansson J, Sadakane K, Sung WK (2004) Computing the maximum agreement of phylogenetic networks. In: Proceedings of Computing: the Tenth Australasian Theory Symposium (CATS'04), Electronic Notes in Theoretical Computer Science, vol 91, pp 134–147

Choy C, Jansson J, Sadakane K, Sung WK (2005) Computing the maximum agreement of phylogenetic networks. Theoretical Computer Science 335(1):93–107

Dress AWM, Huber KT, Moulton V (2007) Some uses of the Farris transform in mathematics and phylogenetics - a review. Annals of Combinatorics 11(1):1–37

Fellows M, Hallet M, Stege U (2003) Analogs & duals of the mast problem for sequences & trees. Journal of Algorithms 49(1):192–216

Gambette P, Berry V, Paul C (2009) The structure of level-k phylogenetic networks. In: Proceedings of the 20^{th} Annual Symposium on Combinatorial Pattern Matching (CPM'09)

Grünewald S, Huber KT (2007) Identifying and defining trees. In: Gascuel O, Steel M (eds) Reconstructing Evolution, New Mathematical and Computational Advances, Oxford University Press, pp 217–246

Gusfield D, Eddhu S, Langley C (2003) Efficient reconstruction of phylogenetic networks with constrained recombination. In: Proceedings of the 2003 IEEE Computational Systems Bioinformatics Conference (CSB2003), pp 363–374

Holland BR, Huber KT, Moulton V, Lockhart PJ (2004) Using consensus networks to visualize contradictory evidence for species phylogeny. Molecular Biology and Evolution 21(7):1459–1461

Huson DH, Rupp R (2008) Summarizing multiple gene trees using cluster networks. In: Proceedings of the eighth Workshop on Algorithms in Bioinformatics (WABI'08), Springer Verlag, Lecture Notes in Computer Science, vol 5251, pp 296–305

Huson DH, Rupp R, Scornavacca C (2011) Phylogenetic Networks. Cambridge University Press

van Iersel L, Keijsper J, Kelk S, Stougie L, Hagen F, Boekhout T (2009a) Constructing level-2 phylogenetic networks from triplets. IEEE/ACM Transactions in Computational Biology and Bioinformatics 6(4):667–681

van Iersel L, Kelk S, Mnich M (2009b) Uniqueness, intractability and exact algorithms: reflections on level-k phylogenetic networks. Journal of Bioinformatics and Computational Biology In press

Jansson J, Sung WK (2006) Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theoretical Computer Science 363(1):60–68

Jansson J, Nguyen NB, Sung WK (2006) Algorithms for combining rooted triplets into a galled phylogenetic network. SIAM Journal on Computing 35(5):1098–1121

Kanj IA, Nakhleh L, Than C, Xia G (2008) Seeing the trees and their branches in the network is hard. Theoretical Computer Science 401:153–164

Moret BME, Nakhleh L, Warnow T, Linder CR, Tholse A, Padolina A, Sun J, Timme R (2004) Phylogenetic networks: Modeling, reconstructibility, and accuracy. IEEE/ACM Transactions in Computational Biology and Bioinformatics 1(1):13–23

Moulton V, Huber KT (2006) Phylogenetic networks from multi-labeled trees. Journal of Mathematical Biology 52(5):613–632

Rosselló F, Valiente G (2009) All that glisters is not galled. Mathematical Biosciences 221(1):54–59

Semple C (2007) Hybridization networks. In: Gascuel O, Steel M (eds) Reconstructing Evolution, New Mathematical and Computational Advances, Oxford University Press, pp 277–314

Semple C, Steel M (2003) Phylogenetics. Oxford University Press

Song YS, Hein J (2005) Constructing minimal ancestral recombination graphs. Journal of Computational Biology 12(2):147–169

To TH, Habib M (2009) Level-k phylogenetic network can be constructed from a dense triplet set in polynomial time. In: Proceedings of the 20^{th} Annual Symposium on Combinatorial Pattern Matching (CPM'09)

Wang L, Zhang K, Zhang L (2001) Perfect phylogenetic networks with recombination. In: Proceedings of the 16th ACM Symposium on Applied Computing (SAC'01), pp 46–50

Willson S (2010) Regular networks are determined by their trees. IEEE/ACM Transactions in Computational Biology and Bioinformatics To appear

Willson SJ (2006) Unique solvability of certain hybrid networks from their distances. Annals of Combinatorics 10(1):165–178