Bulletin of Mathematical Biology (2025) 87:168
https://doi.org/10.1007/s11538-025-01549-4

Society for Mathematical Biology

ORIGINAL ARTICLE

Distinguishing Phylogenetic Level-2 Networks with Quartets and Inter-Taxon Quartet Distances

Niels Holtgrefe[1] · Elizabeth S. Allman[2] · Hector Baños[3] · Leo van Iersel[1] · Vincent Moulton[4] · John A. Rhodes[2] · Kristina Wicke[5]

Received: 23 July 2025 / Accepted: 13 October 2025

Abstract

The inference of phylogenetic networks, which model complex evolutionary processes including hybridization and gene flow, remains a central challenge in evolutionary biology. Until now, statistically consistent inference methods have been limited to phylogenetic level-1 networks, which allow no interdependence between reticulate events. In this work, we establish the theoretical foundations for a statistically consistent inference method for a much broader class: semi-directed level-2 networks that are outer-labeled planar and galled. We precisely characterize the features of these networks that are distinguishable from the topologies of their displayed quartet trees. Moreover, we prove that an inter-taxon distance derived from these quartets is circular decomposable, enabling future robust inference of these networks from quartet data, such as concordance factors obtained from gene tree distributions under the Network Multispecies Coalescent model. Our results also have novel identifiability implications across different data types and evolutionary models, applying to any setting in which displayed quartets can be distinguished.

Keywords Phylogenetic network · Semi-directed network · Reticulate evolution · Quartet · Identifiability · Circular split system

1 Introduction

As analysis of genomic datasets has advanced, growing evidence has indicated that hybridization and other reticulate events play a significant role in evolution (see, e.g., Swithers et al. (2012); DeBaun et al. (2023); Sessa et al. (2012)). Hence it is desirable, and increasingly common, to rely on phylogenetic networks instead of phylogenetic trees to depict evolutionary relationships between species. We consider semi-directed (phylogenetic LSA) networks in which the root is suppressed, since the root location cannot be identified under many models of evolution and data types (Solís-Lemus and Ané 2016; Baños 2019; Ané et al. 2024; Gross et al. 2021; Xu and Ané 2022), and such networks might be rooted in practice using an outgroup (Kinene et al. 2016). Limited by both the lack of theoretical identifiability results for complex networks, as well

Extended author information available on the last page of the article

as computational constraints, existing statistically consistent inference methods focus on the class of semi-directed *level-1* networks (Solís-Lemus and Ané 2016; Allman et al. 2019; Kong et al. 2024; Holtgrefe et al. 2025; Allman et al. 2025a). Within this class of networks, reticulate events are assumed to be isolated from each other, forcing the network to consist of only tree-like parts and disjoint cycles. Numerous identifiability results have been obtained for this restricted class of networks, proving that theoretically (most of) a level-1 network can be distinguished from biological data under several evolutionary models (Gross et al. 2021; Baños 2019; Xu and Ané 2022; Allman et al. 2022, 2024b; Englander et al. 2025).

Several results have shown that certain features of higher-level networks (informally, networks with a higher degree of interdependence between reticulate events) can be identified from biological data. Specifically, the *tree-of-blobs* — a tree depicting only the tree-like branching structure of a network — is identifiable from gene tree probabilities under several variants of the Network Multispecies Coalescent model (Allman et al. 2023; Rhodes et al. 2025), as well as from nucleotide sequences under the Jukes-Cantor and Kimura-2-Parameter models (Englander et al. 2025). Similarly, when a *blob* (a 2-edge-connected component of the network) is *outer-labeled planar* (i.e., it can be drawn in the plane without crossing edges and with the taxon labels on the outside), the circular order of its pendant subnetworks is identifiable from the same types of data and under the same models (Rhodes et al. 2025; Englander et al. 2025).

Two recent breakthroughs have shown the identifiability of the full structure of some higher-level networks. On the one hand, (Englander et al. 2025) showed that under the Jukes-Cantor model, semi-directed binary, strongly tree-child level-2 networks can be identified from nucleotide sequences if the networks contain no 3-cycles within blobs. On the other hand, Allman et al. (2025b) showed identifiability of certain semi-directed binary, galled, strongly tree-child level-k networks from gene tree probabilities under variations of the Network Multispecies Coalescent model. Both these results demonstrate that theoretically sound inference methods for networks more general than level-1 are possible, though neither suggest a particular computationally tractable inference procedure.

In this paper, we take a first step in this direction by providing the theoretical foundations for a statistically consistent inference method for a class of level-2 networks. In particular, our main result (Theorem 5.5) revolves around a canonical form that characterizes exactly when semi-directed level-2 networks that are both galled and outer-labeled planar (see Figure 1 for an example of such a network) can and cannot be distinguished by means of displayed quartets. Hence, this canonical form is the most refined network that can theoretically be inferred when using displayed quartets alone, and thus also illuminates the theoretical limits of level-2 network inference methods relying only on quartets. Since our result does not rely on any specific model, immediate identifiability results follow from biological data generated under models of evolution that meet our criteria. In particular, our result shows that the canonical form can be identified under any model in which the topology of displayed quartets can be identified. This includes all aforementioned models (Englander et al. 2025; Rhodes et al. 2025).

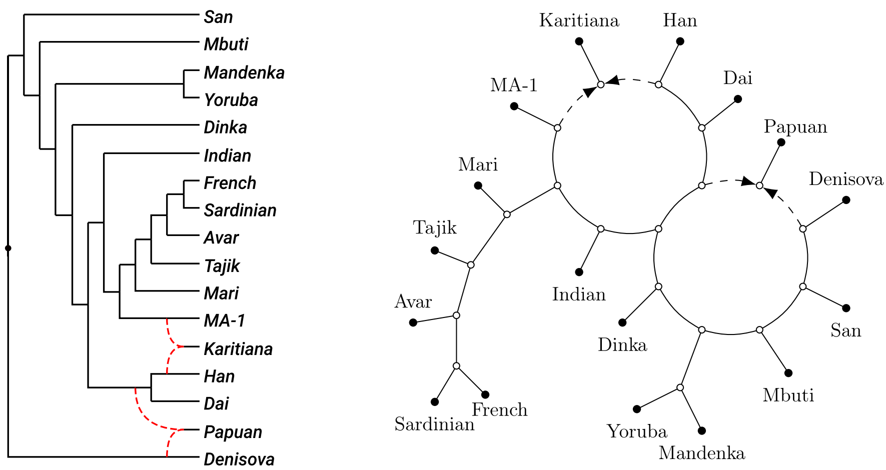

Fig. 1 *Left:* A rooted phylogenetic level-2 network on 17 human populations (Raghavan et al. 2014), where the directions go from left to right and the root is located at the black dot. The network was constructed from 16 complete genomes from modern human worldwide and *MA-1*, a 24,000-year-old anatomically modern human from the Mal'ta site in south-central Siberia. *Right:* The semi-directed phylogenetic network obtained from the rooted network by suppressing its root and only retaining directions of the dashed hybrid edges. The network is level-2, outer-labeled planar, and galled

Our identifiability proof is constructive in nature and at its heart lies a generalization of the inter-taxon quartet distances associated to semi-directed level-1 networks[1] from Allman et al. (2019, 2025a). These distances have the advantage that they can be computed directly from data such as *concordance factors*: summaries of gene tree distributions under the Network Multispecies Coalescent model. We prove that these distances are *circular decomposable* for so-called *bloblets*: networks with a single internal blob (see Theorem 4.7). Essentially, these distances can be decomposed into sums of simpler metrics in a way that respects the circular ordering of the taxa induced by the planar embedding of the network. Interestingly, this approach directly leads to an inference algorithm for computing outer-labeled planar, galled, level-2 networks in the spirit of the existing level-1 inference tools NANUQ and NANUQ$^+$ (Allman et al. 2019, 2025a). An outline of this algorithm is sketched in the discussion; specifics and its implementation will follow in future work.

This paper is structured as follows. Section 2 covers preliminaries on phylogenetic networks, split systems, circular decomposable metrics, and the quartet metric from Rhodes (2019). In Section 3 we prove some auxiliary results used in Section 4 to prove circular decomposability of the inter-taxon quartet distances under consideration. Section 5 contains the main result about our canonical form and the biological identifiability results that follow from it. We end with a discussion in Section 6.

[1] See also (Holtgrefe et al. 2025) for a similar distance.

2 Preliminaries

2.1 Phylogenetic networks

We use standard terminology for phylogenetic networks as in Steel (2016).

Definition 2.1 (Rooted phylogenetic network) A *(binary) rooted (phylogenetic) network* N^+ on a set of at least two taxa X is a rooted directed acyclic graph such that (i) the (unique) root has in-degree zero and out-degree two; (ii) its leaves are of in-degree one and out-degree zero and they are bijectively labeled by elements of X; (iii) all other nodes either have in-degree one and out-degree two (known as *tree nodes*), or in-degree two and out-degree one (known as *hybrid nodes*).

A rooted network admits a natural partial order of its nodes. The *least stable ancestor (LSA)* of a set of leaves Y of a rooted network is the lowest node through which all directed paths from the root to any leaf in Y must pass. Throughout this work, we assume that for all rooted networks on X, the root is the LSA of X (also known as *LSA networks*). The two edges directed towards a hybrid node are called *hybrid edges*. We call a leaf in X a *hybrid (leaf)* if its parent is a hybrid node.

Definition 2.2 (Semi-directed phylogenetic network) A *(binary) semi-directed (phylogenetic) network* N on X is a partially directed graph that can be obtained from a binary rooted phylogenetic LSA network N^+ by undirecting all non-hybrid edges and suppressing the former root.

In the definition above, the rooted network N^+ is called a *rooted partner* of N (see also Figure 2(a) and (b)). We define hybrid nodes and edges in semi-directed networks analogously to those for rooted networks, and note that a semi-directed network may have parallel edges. A semi-directed network without hybrid nodes is a *(binary unrooted) phylogenetic tree*.

Since phylogenetic networks, whether rooted or semi-directed, never have cycles in the (semi-)directed sense, we use the term *cycle* to mean nodes forming a cycle in the fully undirected network. For instance, the level-2 networks in Figure 1 each have three cycles, with the semi-directed network on the right showing two 8-cycles, while the third is a 14-cycle that includes all nodes within the large blob.

A *blob* of a rooted or semi-directed network is a maximal weakly connected subgraph without any cut edges. An *articulation node* of a blob is a node in the blob that is incident to a cut edge of the network. A blob with m articulation nodes is called an *m-blob*. By *contracting* a blob, we mean replacing the blob by a single node, and suppressing it in case this node has degree-2. The *tree-of-blobs* of a semi-directed network is the phylogenetic tree that can be obtained from it by contracting every blob (see Figure 2(c)). In the following definition, we list some additional properties that semi-directed networks may have.

Definition 2.3 Let N be a semi-directed network. Then, we call N

(i) *level-k* for some non-negative integer k if there exist at most k hybrid nodes in each blob of the network;

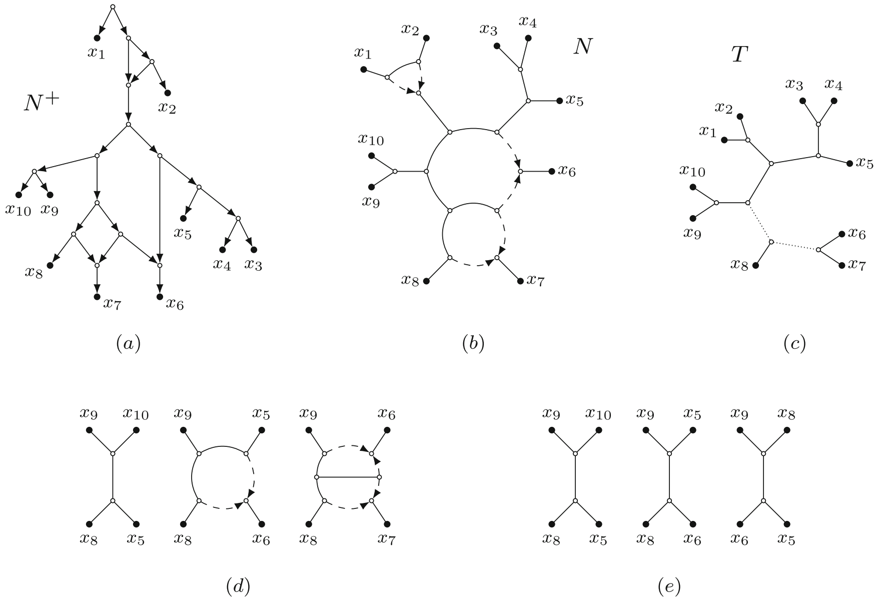

Fig. 2 (a): A rooted phylogenetic network N^+ on $X = \{x_1, \dots, x_{10}\}$ (b): The semi-directed phylogenetic network N on $X = \{x_1, \dots, x_{10}\}$ obtained from N^+. The network N is outer-labeled planar, strictly level-2 and galled (c): A phylogenetic tree T on $X = \{x_1, \dots, x_{10}\}$ that is displayed by N with multiplicity 2. The tree that can be obtained from T by contracting the two dotted edges is the tree-of-blobs of N (d): Three quarnets induced by N (e): Three displayed quartets of N

(ii) *strictly level-k* for some positive integer k if the network is level-k but the network is not level-$(k-1)$;
(iii) a *bloblet (network)*[2] if the network has a single non-leaf blob;
(iv) *k-cycle-free* for some non-negative integer k if the network contains no cycles of length k;
(v) *outer-labeled planar* if the network has a planar embedding with the leaves on the unbounded face;
(vi) *galled* if for every hybrid node h and every pair of partner hybrid edges $e = (v, h)$ and $e' = (v', h)$, there exists a cycle in the network (disregarding edge directions) that contains e and e' and no other hybrid edges of the network.

Note that a galled network can equivalently be defined as one with no hybrid node *ancestral* to another hybrid node in the same blob (i.e., there is no pair of hybrid nodes in the same blob having a path between them in which all directed edges have the same orientation), or in the binary setting of this article, as one where each hybrid node is an articulation node. We also extend the notions of *level* and *strict level* to individual blobs by counting the number of hybrid nodes in that particular blob.

[2] Such networks have also been called *simple*.

The specific classes of networks we consider are given in the next definition. Figure 2(a) and (b) depict a semi-directed network from the class $\mathfrak{N}_2'$ and one of its rooted partners.

Definition 2.4 The (binary) semi-directed phylogenetic networks that are outer-labeled planar, galled, and level-k (for some $k \geq 0$) form the class $\mathfrak{N}_k$. The strictly level-k networks in $\mathfrak{N}_k$ form the subclass $\mathfrak{N}_k'$. The bloblets in $\mathfrak{N}_k$ and $\mathfrak{N}_k'$, respectively, form the subclasses $\mathfrak{B}_k$ and $\mathfrak{B}_k'$.

An *up-down path* between two labeled leaves x_1 and x_2 of a semi-directed network is a path of k edges where the first ℓ edges are directed towards x_1 and the last $k - \ell$ edges are directed towards x_2, where undirected edges are considered bidirected.

Definition 2.5 (Subnetwork) Given a semi-directed network N on X and some $Y \subseteq X$ with $|Y| \geq 2$, the *subnetwork of N induced by Y* is the semi-directed network $N|_Y$ obtained from N by taking the union of all up-down paths between leaves in Y, followed by exhaustively suppressing all degree-2 nodes.

If $|Y| = 4$ with $Y = \{x, y, z, w\}$, we may write $N|_{xyzw}$ to mean $N|_{\{x,y,z,w\}}$. Such a network is called a *quarnet*, or a *quartet* in case N is an unrooted phylogenetic tree (see Figure 2(d)).

Given a semi-directed network N on X and a phylogenetic tree T on X, we say that T is *displayed* by N if it can be obtained from N by deleting exactly one hybrid edge per hybrid node, then recursively deleting all leaves not in X and exhaustively suppressing degree-2 nodes. We use $\mathcal{T}(N)$ to denote the set of phylogenetic trees displayed by N. The *multiplicity* $\mu(T, N)$ of T in $\mathcal{T}(N)$ is the number of distinct choices of hybrid edges in N that when deleted lead to the displayed tree T. If $T \notin \mathcal{T}(N)$, then we set $\mu(T, N) = 0$. We write $\mu(N) = \sum_{T\in\mathcal{T}(N)} \mu(T, N)$, and note that if N has r hybrid nodes, then $|\mathcal{T}(N)| \leq \mu(N) = 2^r$. Lastly, we refer to a quartet tree that is displayed by a quarnet of N as a *displayed quartet of N*. See Figure 2(c) and (e).

2.2 Split systems and circular metrics

This section introduces several known concepts related to splits and metrics, see e.g. Bryant et al. (2007); Allman et al. (2019). Recall that a *pseudo metric $d : X^2 \to \mathbb{R}_{\geq 0}$* on a finite set of elements X is a symmetric non-negative function that (i) satisfies the triangle inequality; and (ii) has the property that $d(x, x) = 0$ for all $x \in X$. In particular, for $x, y \in X$, $d(x, y) = 0$ need not imply $x = y$. All results in this work revolve around pseudo metrics, which we will simply use the word *metrics* for convenience unless we want to emphasize the fact that we are considering pseudo metrics.

A *split* $A|B = B|A$ of a finite set X with at least two elements is a bipartition of X with $A, B \subseteq X$. A split is *empty* if A or B has size 0, and it is *trivial* if A or B has size 1. We say that a cut edge of a semi-directed network N *induces* the split $A|B$ of its leaf set X if the removal of the cut edge disconnects the leaves labeled by elements in A from those labeled by elements in B. We denote the set of splits induced by the cut edges of a semi-directed network N as $\mathrm{Split}(N)$. Similarly, we denote the set of splits induced by the displayed trees of N by $\mathrm{Split}(\mathcal{T}(N))$. To distinguish between the two, we call the former the *induced splits of* N and the latter the *displayed splits of* N. Clearly, $\mathrm{Split}(N) \subseteq \mathrm{Split}(\mathcal{T}(N))$.

The set of all nonempty splits of a finite set X is denoted by $\mathrm{Split}(X)$, and we call a non-empty set $\mathcal{S} \subseteq \mathrm{Split}(X)$ a *split system* on X. A map $\omega : \mathrm{Split}(X) \to \mathbb{R}_{\geq 0}$ is a *weighted split system*. We let the *support* of the weighted split system, denoted by $\mathrm{supp}(\omega)$, be the split system $\mathcal{S} \subseteq \mathrm{Split}(X)$ containing each split that has a positive weight ω. Note that any weighted split system $\omega : \mathrm{Split}(X) \to \mathbb{R}_{\geq 0}$ induces a metric d_ω on X as follows. Given some split $A|B \in \mathrm{Split}(X)$, we define the *split metric* $\delta_{A|B}$ as

$$\delta_{A|B}(x, y) = \begin{cases} 0, & \text{if } x, y \in A \text{ or } x, y \in B; \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

Then, the metric d_ω *induced by* ω is

$$d_\omega(x, y) = \sum_{S \in \mathrm{Split}(X)} \omega(S) \cdot \delta_S(x, y). \tag{2}$$

It is easy to see that the triangle inequality holds for d_ω, so that it is indeed a (pseudo) metric.

Before the next definition, we define a *circular order* $\mathcal{C}$ of a finite set $X = \{x_1, \ldots, x_n\}$ as an ordering of its elements up to reversal and cyclic permutations. We often denote such a circular order as $\mathcal{C} = (x_0, x_1, \ldots, x_n = x_0)$ or simply $\mathcal{C} = (x_1, \ldots, x_n)$.

Definition 2.6 (Circular split system) A split system $\mathcal{S} \subseteq \mathrm{Split}(X)$ is *circular* if there exists a circular order $\mathcal{C} = (x_0, x_1, \ldots, x_n = x_0)$ of the finite set $X = \{x_1, \ldots, x_n\}$ such that each split in $\mathcal{S}$ has the form $S_{ij}(\mathcal{C}) = U_{ij}|V_{ij}$ for some $i \neq j$ with $U_{ij} = \{x_{i+1}, \ldots, x_j\}$ and $V_{ij} = \{x_{j+1}, \ldots, x_i\}$. We say that such a circular order is *congruent* with $\mathcal{S}$.

We sometimes write S_{ij} instead of $S_{ij}(\mathcal{C})$ if $\mathcal{C}$ is clear from the context. The circular split system consisting of all splits $S_{ij}(\mathcal{C})$ for a circular order $\mathcal{C}$ is denoted by $\mathcal{S}(\mathcal{C})$. For example, Figure 3 gives a visualization of a circular split system $\mathcal{S} \subset \mathcal{S}\big((x_1, \ldots, x_n)\big)$. Note that the split system in the figure is also congruent with other circular orders not depicted. We say that a weighted split system is *circular* if its support is a circular split system.

Definition 2.7 (Circular decomposable metric) A metric $d : X^2 \to \mathbb{R}_{\geq 0}$ is *circular decomposable*, if there exists a circular weighted split system $\omega : \mathrm{Split}(X) \to \mathbb{R}_{\geq 0}$ such that $d = d_\omega$.

In other words, a metric d is circular decomposable if it can be decomposed as a sum of weighted split metrics, all corresponding to one circular split system. As shown in Bandelt and Dress (1992) (see also (Bryant et al. 2007)), if d is circular decomposable, this decomposition is the unique way to decompose d into a sum of weighted (weakly

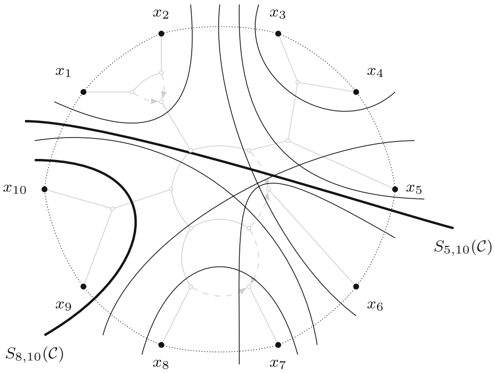

Fig. 3 A visualization of the circular split system $\mathcal{S} \subset \text{Split}(\mathcal{T}(N))$, containing all non-trivial displayed splits (depicted by the black lines) of the outer-labeled planar, semi-directed network N on $X = \{x_1, \ldots, x_{10}\}$ (depicted in gray) from Figure 2(b). $\mathcal{S} \subset \mathcal{S}(\mathcal{C})$ is congruent with the circular order $\mathcal{C} = (x_1, \ldots x_{10})$ (depicted by the dotted lines), one of the induced circular orders of N. Two splits are highlighted in thick black and are labeled using the notation from Definition 2.6

compatible) split metrics, i.e., the circular weighted split system ω is unique. Hence, we may also define the *support* $\text{supp}(d_\omega)$ of a circular decomposable metric d_ω as the unique support of the corresponding circular weighted split system ω.

To prove that a metric is circular decomposable we use the following proposition, which follows from a lemma of Chepoi and Fichet (1998). To avoid confusion, when referring to the split weights defined in the following proposition, we may sometimes write $\alpha_{ij}(d)$ or $\alpha_{ij}(d, \mathcal{C})$ instead of simply α_{ij}.

Proposition 2.8 *Let X be a finite set of elements, $d : X^2 \to \mathbb{R}_{\geq 0}$ a pseudo metric, $\mathcal{C} = (x_0, x_1, \ldots x_n = x_0)$ a circular order of X and $\mathcal{S} \subseteq \mathcal{S}(\mathcal{C})$. For all $S_{ij} \in \mathcal{S}(\mathcal{C})$, let α_{ij} be the* split weight *of S_{ij} associated with d and $\mathcal{C}$, which is defined as*

$$\alpha_{ij} = d(x_i, x_j) + d(x_{i+1}, x_{j+1}) - d(x_i, x_{j+1}) - d(x_{i+1}, x_j).$$

Then, d is circular decomposable with support $\mathcal{S}$ if and only if $\alpha_{ij} \geq 0$ for all $S_{ij} \in \mathcal{S}(\mathcal{C})$ and $\mathcal{S} = \{S_{ij} \in \mathcal{S}(\mathcal{C}) : \alpha_{ij} > 0\}$.

Proof Chepoi and Fichet (1998) proved that any symmetric function $d : X^2 \to \mathbb{R}$ with a zero diagonal can be written as

$$d(x, y) = \frac{1}{2} \sum_{S_{ij} \in \mathcal{S}(\mathcal{C})} \alpha_{ij} \cdot \delta_{S_{ij}}(x, y) \tag{3}$$

for all $x, y \in X$ and for any circular order $\mathcal{C}$ of X. The backward direction then follows. For the forward direction, note that by the main result in Chepoi and Fichet (1998) (see also (Christopher et al. 1996) for an alternative proof), circular decomposable metrics congruent with a circular ordering $\mathcal{C}$ are equivalent to *Kalmanson metrics* (Kalmanson 1975) compatible with $\mathcal{C}$, that is, metrics for which $\alpha_{ij} \geq 0$ for all $S_{ij} \in \mathcal{S}(\mathcal{C})$. So, if d is circular decomposable with support $\mathcal{S}$, all α_{ij} are non-negative. Therefore, 3 gives the circular decomposition of d. Since this decomposition is unique (Bandelt and Dress 1992), it also follows that $\mathcal{S} = \{S_{ij} \in \mathcal{S}(\mathcal{C}) : \alpha_{ij} > 0\}$.

We say that a semi-directed network N on X *induces* a circular order $\mathcal{C}$ of X if there exists an outer-labeled planar representation of N that induces the circular order $\mathcal{C}$ of the leaves in X (see Figure 3). As shown in Rhodes et al. (2025), outer-labeled planar semi-directed networks induce at least one circular order of their leaf sets. Moreover, if the outer-labeled planar network is a bloblet, this circular order is unique. See also (Moulton and Wu 2022) for results related to the planarity of rooted networks. The following result straightforwardly connects the induced circular orders of an outer-labeled planar network to the displayed splits of the network.

Proposition 2.9 *Let N be an outer-labeled planar, semi-directed network on X. Then,* $\mathrm{Split}(\mathcal{T}(N))$ *is a circular split system congruent with any circular order of X induced by N.*

Proof Let $\mathcal{C}$ be an induced circular order of N. Clearly, every tree $T \in \mathcal{T}(N)$ is also outer-labeled planar and induces the order $\mathcal{C}$. Then, $\mathrm{Split}(T)$ is a circular split system congruent with $\mathcal{C}$ for all $T \in \mathcal{T}(N)$ (see, e.g., Gambette et al. (2012)). Hence, $\mathrm{Split}(\mathcal{T}(N)) = \bigcup_{T \in \mathcal{T}(N)} \mathrm{Split}(T)$ is a circular split system congruent with $\mathcal{C}$.

2.3 Quartet metric of a phylogenetic tree

We revisit the metric from Rhodes (2019), which was defined on the leaves of a phylogenetic tree based on the quartets induced by that tree.

Let T be a phylogenetic tree on four leaves $\{x, y, z, w\}$, i.e. T is a *quartet tree*. Then, for any pair of leaves $\{x, y\}$, we let

$$\rho_{xy}(T) = \begin{cases} 0 & \text{if } x \text{ and } y \text{ form a cherry,} \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

Note that $\rho_{xy}(T) = \delta_S(x, y)$, where S is the unique non-trivial split induced by T.

We can use this concept to define a metric d_T for any phylogenetic tree T with at least four leaves. Recall that $T|_{xyzw}$ denotes the quartet tree on leaf set $\{x, y, z, w\}$ induced by the tree T.

Definition 2.10 (Quartet metric) Let T be a phylogenetic tree on leaf set X with $n = |X| \geq 4$. For any pair of leaves $x \neq y$ of X let

$$d_T(x, y) = \sum_{z, w \neq x, y} 2 \cdot \rho_{xy}(T|_{xyzw}) + 2n - 4, \tag{5}$$

with $d_T(x, x) = 0$. Then, d_T is the *quartet metric* of T.

A *metrization* of a tree $T = (V, E)$ is a map $w : E \to \mathbb{R}_{>0}$ that assigns a positive weight to each edge. Next, we define the *quartet metrization* of a phylogenetic tree T

on X, which assigns a weight $w(e)$ to every edge e of T. First note that each internal edge e of T determines a partition of X into 4 non-empty blocks (X_1, X_2, X_3, X_4) where the split associated to e is $X_1 \cup X_2 | X_3 \cup X_4$ and the splits associated to the 4 adjacent edges all lead to one X_i. To obtain the quartet metrization now assign to every internal edge e the weight

$$w(e) = |X_1||X_2| + |X_3||X_4|. \tag{6}$$

Similarly, every pendant edge, incident to a leaf x, induces a tripartition $(\{x\}, X_1, X_2)$ of X. To every such edge e we assign the weight

$$w(e) = |X_1||X_2|. \tag{7}$$

The following inequality will be useful later in this manuscript.

Proposition 2.11 *Let $T = (V, E)$ be a phylogenetic tree on leaf set X with $|X| \geq 4$ and suppose that T is equipped with the quartet metrization $w : E \to \mathbb{R}_{\geq 0}$. Then, for every edge e in T, we have $w(e) \geq |X| - 2$.*

Proof First suppose that e is a pendant edge inducing the tripartition $(\{x\}, X_1, X_2)$ of X. Then, $w(e) = |X_1||X_2|$. The inequality now follows from the fact that $n_1 \cdot n_2 \geq n_1 + n_2 - 1$ if $n_1, n_2 \geq 1$.

Now suppose e is internal inducing the partition (X_1, X_2, X_3, X_4). Then, $w(e) = |X_1||X_2|+|X_3||X_4|$. Again, the inequality follows because $n_1n_2+n_3n_4 \geq \sum_{i=1}^{4} n_i - 2$ whenever all $n_i \geq 1$.

The relationship of the quartet metric defined in Definition 2.10 and the quartet metrization is captured in the following.

Theorem 2.12 *(Rhodes 2019) Let $T = (V, E)$ be a phylogenetic tree on leaf set X with $|X| \geq 4$. Then, for every pair $\{x, y\} \subseteq X$, the quartet metric $d_T(x, y)$ is the distance in T between x and y when T is equipped with the quartet metrization $w : E \to \mathbb{R}_{\geq 0}$.*

Corollary 2.13 *Let T be a phylogenetic tree on at least four leaves. Then, the quartet metric d_T is circular decomposable and its support is* Split(T).

3 Displayed quartet metric of a phylogenetic network

Here we generalize the quartet metric for a tree to semi-directed networks, accounting for the multiplicities of the quartets. Although at first glance this generalization seems primarily of theoretical value since it is unclear whether the quartet multiplicities — and thus the resulting distances — can be computed directly from data, the results here are linked in the next section to a second generalization where the generalized distances can be estimated.

Let N be a semi-directed network on four leaves $\{x, y, z, w\}$, i.e., a quarnet, (possibly a tree). Then, the first step in generalizing the quartet metric is to use the set $\mathcal{T}(N)$ of quartet trees displayed by N, accounting for their multiplicities, to obtain

$$\widetilde{\rho}_{xy}(N) = \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \cdot \rho_{xy}(T). \tag{8}$$

That is, the $\widetilde{\rho}_{xy}$-value for a pair of leaves (x, y) of a quarnet N is simply the average ρ_{xy}-values for all the quartet trees that N displays, weighted by multiplicity. If N is a quartet tree, this reduces to Eq. (4). In case the quarnet is strictly level-1, we obtain that $\mu(N) = 2$ and $\mu(T, N) = 1$, resulting in ρ-values of either $1/2$ or 1. This exactly coincides with the ρ-values used for NANUQ (Allman et al. 2019).

Figure4 illustrates pairwise $\widetilde{\rho}$-values for several outer-labeled planar, galled quarnets. For a quarnet to be galled, every hybrid node must be parental to a leaf, but otherwise the hybrid nodes may be chosen arbitrarily in these examples, as long as their choice results in a valid semi-directed network. Thus, for the quarnet in the second column of Figure 4, exactly one of $\{x, y, z, w\}$ needs to be a hybrid, whereas for the third column exactly one of $\{x, y\}$ and one of $\{z, w\}$ need to be hybrids. As an example, note that — independent of the choice of hybrids — the quarnet in the third column displays the quartet $xy|zw$ with multiplicity 3, the quartet $xz|yw$ with multiplicity 1, and the quartet $xw|yz$ with multiplicity 0, resulting in the shown $\widetilde{\rho}$-values of $1/4$, $3/4$ and 1. Because contracting a 2-blob, or contracting either a 3-blob or a 3-cycle within a blob to a node, has no impact on displayed tree topologies, the values of $\widetilde{\rho}$ hold slightly more generally.

As was the case for phylogenetic trees, this newly defined $\widetilde{\rho}$ gives rise to a metric on the leaf set of an n-leaf semi-directed network. Recall that $N|_{xyzw}$ is the induced quarnet on the leaf set $\{x, y, z, w\}$ of a semi-directed network N.

Definition 3.1 (Displayed quartet metric) Let N be a semi-directed network on leaf set X with $n = |X| \geq 4$. For any pair of leaves $x \neq y$ of X we let

$$\widetilde{d}_N(x, y) = \sum_{z, w \neq x, y} 2 \cdot \widetilde{\rho}_{xy}(N|_{xyzw}) + 2n - 4, \tag{9}$$

with $\widetilde{d}_N(x, x) = 0$. Then, $\widetilde{d}_N$ is the *displayed quartet metric* of N.

Recall that the $\widetilde{\rho}$-value of a quarnet is the average over the $\widetilde{\rho}$-values of each quartet it displays (Eq. (8)). The following now generalizes Lemma 21 from Allman et al. (2019). Informally, it says that the $\widetilde{\rho}$-value of a quarnet of a network N can equivalently be viewed as a weighted average over N's displayed trees of the ρ-values of their induced quartets.

Lemma 3.2 *Let N be a semi-directed network on leaf set X with $|X| \geq 4$. If $\{x, y, z, w\} \subseteq X$, then*

$$\widetilde{\rho}_{xy}(N|_{xyzw}) = \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \cdot \rho_{xy}(T|_{xyzw}).$$

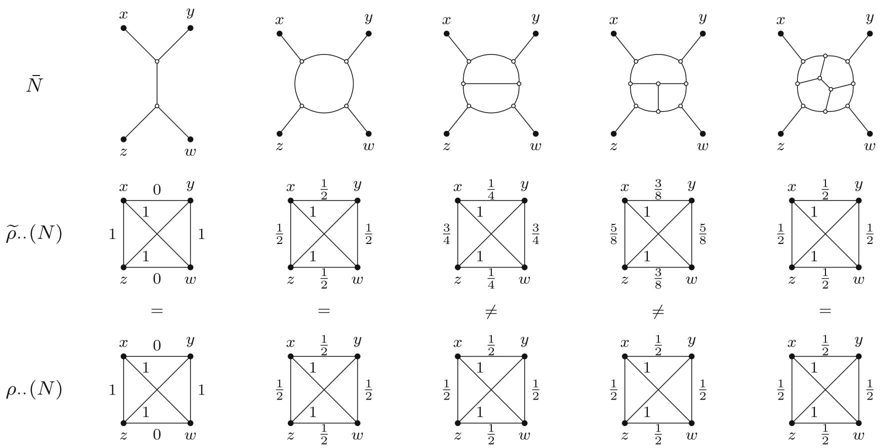

Fig. 4 *Top:* Five undirected graphs $\bar{N}$ that can be obtained by undirecting an outer-labeled planar, galled quarnet N on leaf set $\{x, y, z, w\}$ after contracting 2-blobs, 3-blobs, and 3-cycles within a blob. To make $\bar{N}$ semi-directed again (disregarding the creation of 2-blobs, 3-blobs and 3-cycles), hybrid nodes can be chosen only from the articulation nodes of the undirected blobs since N must be galled, as long as N remains semi-directed. *Middle:* The values $\widetilde{\rho}_{\cdot\cdot}(N)$ (as defined in Eq. (8)) in case N is outer-labeled planar, galled and $\bar{N}$ is as in the top row are shown on edges connecting two taxa. *Bottom:* The values $\rho_{\cdot\cdot}(N)$ (as defined in Eq. (10)) in case N is outer-labeled planar, galled and $\bar{N}$ is as in the top row

Proof Suppose that N has r hybrid nodes and that $N|_{xyzw}$ has $p \leq r$ hybrid nodes. Then, $\mu(N|_{xyzw}) = 2^p$ and $\mu(N) = 2^r$, so that

$$\mu(N|_{xyzw}) = 2^{p-r}\mu(N),$$

and for any quartet tree T' on $\{x, y, z, w\}$,

$$\mu(T', N|_{xyzw}) = 2^{p-r} \sum_{\substack{T \in \mathcal{T}(N): \\ T|_{xyzw} = T'}} \mu(T, N).$$

Thus,

$$\begin{aligned}
\widetilde{\rho}_{xy}(N|_{xyzw}) &= \frac{1}{\mu(N|_{xyzw})} \sum_{T' \in \mathcal{T}(N|_{xyzw})} \mu(T', N|_{xyzw}) \cdot \rho_{xy}(T') \\
&= \frac{1}{2^{p-r}\mu(N)} \sum_{T' \in \mathcal{T}(N|_{xyzw})} 2^{p-r} \sum_{\substack{T \in \mathcal{T}(N): \\ T|_{xyzw} = T'}} \mu(T, N) \cdot \rho_{xy}(T|_{xyzw}) \\
&= \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \cdot \rho_{xy}(T|_{xyzw}).
\end{aligned}$$

□

The next theorem generalizes Theorem 23 of Allman et al. (2019), showing that the displayed quartet metric on a network is simply a weighted sum of the metrics on its displayed trees.

Theorem 3.3 *Let N be a semi-directed network on leaf set X with $|X| \geq 4$ and let $\{x, y\} \subseteq X$. Then,*

$$\widetilde{d}_N(x, y) = \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \cdot d_T(x, y).$$

Proof Let $n = |X|$. Using Definition 2.10 and 3.1, Lemma 3.2, and the equality $\mu(N) = \sum_{T \in \mathcal{T}(N)} \mu(T, N)$, we obtain

$$\begin{aligned}
\widetilde{d}_N(x, y) &\overset{3.1}{=} \sum_{z,w \neq x,y} 2 \cdot \widetilde{\rho}_{xy}(N|_{xyzw}) + 2n - 4 \\
&\overset{3.2}{=} \sum_{z,w \neq x,y} 2 \cdot \left[\frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \cdot \rho_{xy}(T|_{xyzw}) \right] + 2n - 4 \\
&= \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \left[\sum_{z,w \neq x,y} 2\rho_{xy}(T|_{xyzw}) \right] + 2n - 4 \\
&\overset{2.10}{=} \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} \mu(T, N) \cdot d_T(x, y).
\end{aligned}$$

This theorem shows that the metric $\widetilde{d}_N$ is a weighted sum of quartet metrics, with nonnegative weights. Recall that from Corollary 2.13, each of these metrics is circular decomposable and has the induced splits of the corresponding tree as support. Together with the fact that $\mathrm{Split}(\mathcal{T}(N))$ is a circular split system if N is outer-labeled planar (Proposition 2.9) and by the uniqueness of the split decomposition of a circular decomposable metric (Bandelt and Dress 1992), we obtain the following result.

Corollary 3.4 *Let N be a semi-directed, outer-labeled planar network on at least four leaves. Then, the displayed quartet metric $\widetilde{d}_N$ is circular decomposable, with support* $\mathrm{Split}(\mathcal{T}(N))$.

4 NANUQ metric of a phylogenetic network

The displayed quartet metric $\widetilde{d}_N$ of a network N presented in the last section has a straightforward connection to the quartet metrics of displayed trees. However, one of its components, the multiplicity of a displayed quartet, is not likely to be obtainable from data. In this section, we propose a different network generalization of the quartet tree metric that does not account for these multiplicities.

First, we introduce a different generalization of ρ from Section 2.3. Let N be a semi-directed network on four leaves $\{x, y, z, w\}$ (which could be a quartet tree), i.e.,

N is a *quarnet*. Then, in contrast to $\widetilde{\rho}$ of the previous section, we ignore multiplicities and instead define

$$\rho_{xy}(N) = \frac{1}{|\mathcal{T}(N)|} \sum_{T \in \mathcal{T}(N)} \rho_{xy}(T). \tag{10}$$

That is, ρ for a pair of leaves of a quarnet is the *unweighted* average ρ of all the quartet trees N displays. In case N is strictly level-1, $|\mathcal{T}(N)| = 2$, and so this new extension coincides exactly with the $\widetilde{\rho}$-value defined in Eq. (8), and hence also with the ρ that is used for NANUQ (Allman et al. 2019). For other levels, this newly defined ρ and the $\widetilde{\rho}$ from Eq. (8) are generally different (see Figure 4).

As before, ρ induces a corresponding metric on the leaves of a semi-directed network with four or more leaves. Since this metric will be the main generalization of the metric in Allman et al. (2019), we call it the *NANUQ metric*.

Definition 4.1 (NANUQ metric) Let N be a semi-directed network on leaf set X with $n = |X| \geq 4$. For any pair of leaves $x \neq y$ of X, let

$$d_N(x, y) = \sum_{z,w \neq x,y} 2 \cdot \rho_{xy}(N|_{xyzw}) + 2n - 4, \tag{11}$$

with $d_N(x, x) = 0$. Then, d_N is the *NANUQ metric* of N.

An immediate consequence of the definition and Theorem 3.3 is the following, expressing the NANUQ metric as a sum of quartet metrics of the displayed trees of the network, plus an error term.

Lemma 4.2 *Let N be a semi-directed network on leaf set X with $|X| \geq 4$ and let $\{x, y\} \subseteq X$. Then,*

$$d_N(x, y) = \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} d_T(x, y) \cdot \mu(T, N) + \varepsilon(x, y),$$

where $\varepsilon(x, y) = 2 \cdot \sum_{z,w \neq x,y} (\rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw}))$.

Proof Let $n = |X|$. Then,

$$
\begin{aligned}
d_N(x, y) &= \sum_{z,w \neq x,y} 2 \cdot \rho_{xy}(N|_{xyzw}) + 2n - 4 \\
&= \sum_{z,w \neq x,y} 2 \cdot (\widetilde{\rho}_{xy}(N|_{xyzw}) + \rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw})) + 2n - 4 \\
&= \widetilde{d}_N(x, y) + \sum_{z,w \neq x,y} 2 \cdot (\rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw})) \\
&= \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} d_T(x, y) \cdot \mu(T, N) + 2 \sum_{z,w \neq x,y} (\rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw})) \\
&= \frac{1}{\mu(N)} \sum_{T \in \mathcal{T}(N)} d_T(x, y) \cdot \mu(T, N) + \varepsilon(x, y).
\end{aligned}
$$

□

4.1 NANUQ metric is circular decomposable for level-2 bloblet networks

In the rest of this section, we restrict to networks in the class $\mathfrak{B}_2$ with at least four leaves and show that the NANUQ metric is circular decomposable for this class. Recall that this class contains all semi-directed, level-2, outer-labeled planar, galled, bloblet networks. Furthermore, $\mathfrak{B}_2'$ contains all strictly level-2 networks in this class (see Figure 5 for a general example of a network in $\mathfrak{B}_2'$). Note that parallel edges (2-cycles) only appear as a 2-blob for a network in $\mathfrak{N}_2$, and since a k-taxon bloblet in $\mathfrak{B}_2 \subseteq \mathfrak{N}_2$ with $k \geq 3$ cannot have a 2-blob, we do not need to consider parallel edges for such bloblets.

For convenience, we assign a standard partition of the leaves to the networks in $\mathfrak{B}_2'$. As shown in Figure 5, the leaf set X of such a network can be partitioned as $\mathcal{P} = \{A_1, A_2, B_1, B_2, C_1, C_2\}$, uniquely up to the interchange of A with B and 1 with 2. We call $\mathcal{P}$ a *canonical partition* of X. Here, we write $A = A_1 \cup A_2$, $B = B_1 \cup B_2$ and $C = C_1 \cup C_2$, where C contains the two hybrids of N.

We say that a set of leaves S is *on cycle 1* (resp. *on cycle 2*) of N if $S \subseteq A_1 \cup B_1 \cup C_1$ (resp. $S \subseteq A_2 \cup B_2 \cup C_2$), and that the set S is *on the same cycle* if either of these holds.

The following lemma presents an exact expression for the error term ε of Lemma 4.2 for a network in $\mathfrak{B}_2'$. In stating it, indices i are taken modulo 2, so, for example, when $i = 2$ the set A_{i+1} is A_1.

Lemma 4.3 *Let N be a semi-directed network on leaf set X from the class $\mathfrak{B}_2'$ with $|X| \geq 4$ and let $\mathcal{P}$ be a canonical partition of X. Let $\{x, y\}$ be any pair of leaves from X. Then,*

$$
d_N(x, y) = \frac{1}{4} \sum_{T \in \mathcal{T}(N)} d_T(x, y) \cdot \mu(T, N) + \varepsilon(x, y),
$$

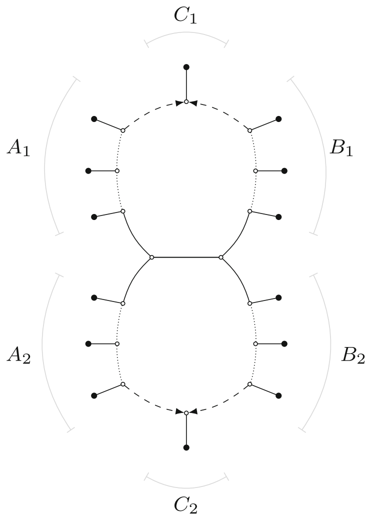

Fig. 5 A general strictly level-2 network from the class $\mathfrak{B}_2'$ with a canonical partition $\mathcal{P} = \{A_1, A_2, B_1, B_2, C_1, C_2\}$ of its (unlabeled) leaves.

where $\varepsilon(x, y) = \varepsilon(y, x)$ *and*

$$\varepsilon(x, y) = \frac{1}{2} \cdot \begin{cases} -|A_1||A_2| - |B_1||B_2| & \textit{if } x \in C_i \textit{ and } y \in C_{i+1}, \\ |A_{i+1}| + |B_{i+1}| & \textit{if } x \in C_i \textit{ and } y \in A_i \cup B_i, \\ -|B_i| & \textit{if } x \in C_i \textit{ and } y \in A_{i+1}, \\ -|A_i| & \textit{if } x \in C_i \textit{ and } y \in B_{i+1}, \\ -1 & \textit{if } x \in A_i \textit{ and } y \in A_{i+1}, \textit{or } x \in B_i \textit{ and } y \in B_{i+1}, \\ 0 & \textit{otherwise}. \end{cases} \tag{12}$$

Proof Since $\mu(N) = 2^2 = 4$, by Lemma 4.2 all that remains is to compute the error term ε, where

$$\varepsilon(x, y) = 2 \sum_{z,w \neq x,y} (\rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw})).$$

Next, note that for any $N|_{xyzw}$ its underlying undirected network $\bar{N} = \overline{N|_{xyzw}}$ with 3-cycles contracted appears — up to relabeling the leaves — in one of the three leftmost columns of the top row of Figure 4. Specifically, if $N|_{xyzw}$ contains no 4-blob, $\bar{N}$ will be the quartet tree in the far-left column, if $N|_{xyzw}$ has a strict level-1 4-blob or a strict level-2 4-blob with a 3-cycle, $\bar{N}$ appears in the second column; otherwise, it appears in the third column. Figure 4 also gives the value of $\rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw})$, with this expression being non-zero only when $N|_{xyzw}$ is strictly level-2 after contracting 3-cycles and x, y are adjacent in the induced circular order. In particular, if these conditions hold, then

$$\rho_{xy}(N|_{xyzw}) - \widetilde{\rho}_{xy}(N|_{xyzw}) = \begin{cases} 1/4 & \text{if } x, y \text{ are on the same cycle,} \\ -1/4 & \text{if } x, y \text{ are on different cycles.} \end{cases}$$

Computing $\varepsilon(x, y)$ is now a matter of checking every quarnet of N on leaves x and y to see if it falls in one of the above two classes contributing a non-zero term. To simplify this, note that the sign of ε is completely determined by whether x and y are on the same cycle or not. Next, note that a quarnet on x and y can only be strictly level-2 after contracting 3-cycles, if (i) the quarnet has the two hybrids in its leaf set (otherwise it would be at most level-1); and (ii) it has two leaves on each cycle (otherwise one of the cycles would be a 3-cycle which would be contracted). Furthermore, recall from above that (iii) the term is non-zero only when x and y are adjacent in the quarnet order.

Thus, to determine $\varepsilon(x, y)$ it suffices to first count the number of choices of the other two leaves z and w such that the induced quarnet satisfies the conditions (i)-(iii). Multiplying this count by $\frac{1}{4} \cdot 2 = \frac{1}{2}$, and then determining the sign as above yields the value.

Case 1: $x \in C_i$ and $y \in C_{i+1}$. Condition (i) is already satisfied. To adhere to condition (iii), we can either pick both leaves z, w in A or both in B. To adhere to condition (ii), z, w must be on different cycles. In total, this gives $|A_1||A_2| + |B_1||B_2|$ choices of z and w.

Case 2: $x \in C_i$ and $y \in A_i \cup B_i$. By condition (i), we must pick the other hybrid in C_{i+1}. Then, to make x and y adjacent (condition (iii)) and to force them to have two leaves on either cycle (condition (ii)), there are exactly $|A_{i+1}| + |B_{i+1}|$ choices.

Case 3: $x \in C_i$ and $y \in A_{i+1}$. Again, we need to pick the other hybrid in C_{i+1} for condition (i). Then, there are $|B_i|$ choices for the fourth leaf to satisfy the other two conditions.

Case 4: $x \in C_i$ and $y \in B_{i+1}$. Analogous to Case 3.

Case 5: $x \in A_i$ and $y \in A_{i+1}$ or $x \in B_i$ and $y \in B_{i+1}$. In this case, we need to pick the two hybrids in C to satisfy all conditions. Hence, there is only one choice.

Case 6: otherwise. In this case, we have that for some $i \in \{1, 2\}$ (1): $\{x, y\} \subseteq A_i$, (2): $\{x, y\} \subseteq B_i$, (3): $x \in A_i$ and $y \in B_i$ (or vice versa), or (4): $x \in A_i$ and $y \in B_{i+1}$ (or vice versa). In all cases, we need to pick the two hybrids. But then, either condition (ii) or condition (iii) is not satisfied. Thus, there are no valid choices. □

Having expressed the NANUQ metric for networks in $\mathfrak{B}_2'$ as a sum of the circular decomposable quartet metric and an explicit error term, we can now show that the NANUQ metric is itself circular decomposable. As a first step we compute bounds on the split weights α_{ij} (see Proposition 2.8).

Given a bloblet N in $\mathfrak{B}_2'$ with leaf set X and canonical partition $\mathcal{P}$, each of its displayed trees can be obtained by removing one hybrid edge for each of the two hybrids. Allowing for multiplicities, this gives rise to four displayed trees (see Figure 6). We denote these trees by $T_{\Sigma_1\Sigma_2}$ with $\Sigma_i \in \{A_i, B_i\}$, where, for example, $\Sigma_i = A_i$ means that the hybrid edge incident to C_i and joined to the subgraph on A_i is kept intact, but its partner hybrid edge (attached to the B_i subgraph) is removed. Hence, when taking

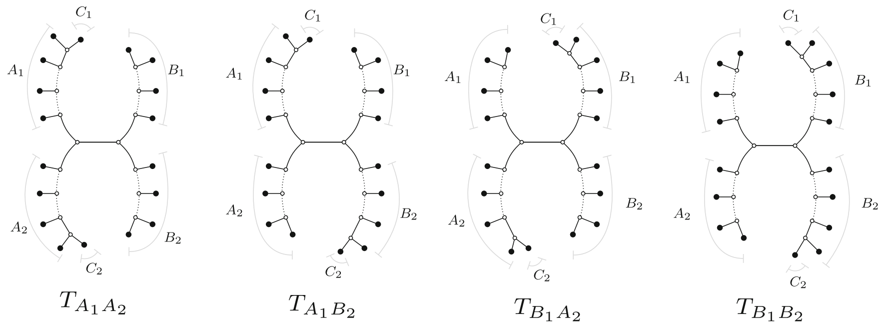

Fig. 6 The four trees displayed by a general semi-directed network from the class $\mathfrak{B}_2'$, using the labeling from Figure 5. If some of the taxon sets A_i, B_i are empty, some of these trees may be isomorphic and have a different topology

multiplicities into account, every network N then has the displayed trees $T_{A_1A_2}$, $T_{A_1B_2}$, $T_{B_1A_2}$ and $T_{B_1B_2}$ (see again Figure 6). Here, multiplicities arise if some of the sets A_i, B_i are empty. For example, the trees $T_{A_1A_2}$ and $T_{B_1A_2}$ are isomorphic if A_1 and B_1 are empty.

We call a non-trivial split of a displayed tree T a *major split* if it separates the leaves in A from those in B. Hence, a major split is of the form $A|B \cup C$, $A \cup C|B$, $A \cup C_1|B \cup C_2$, or $A \cup C_2|B \cup C_1$. As an example, if $|A|, |B| \geq 1$, then $A \cup C_1|B \cup C_2$ is a major split of $T_{A_1B_2}$.

Lemma 4.4 *Let N be a network on leaf set X from the class $\mathfrak{B}_2'$ with $n = |X| \geq 4$, $\mathcal{P}$ a canonical partition of X, and $\mathcal{C} = (x_0, \ldots, x_n = x_0)$ the circular order induced by N. Let $T = T_{\Sigma_1\Sigma_2}$ be a displayed tree of N, d_T the quartet metric on T, and $S_{ij} \in \mathcal{S}(\mathcal{C})$ a split. Then,*

(a) $\alpha_{ij}(d_T) = 0$, *if* $S_{ij} \notin \mathrm{Split}(T)$;
(b) $\alpha_{ij}(d_T) \geq 2|X| - 4 > 0$, *if* $S_{ij} \in \mathrm{Split}(T)$;
(c) $\alpha_{ij}(d_T) > 2|A_1||A_2| + 2|B_1||B_2|$, *if* $S_{ij} \in \mathrm{Split}(T)$ *and* S_{ij} *is a major split of* T.

Proof Note that $\mathcal{C}$ is an induced order for T, so $\mathrm{Split}(T) \subseteq \mathcal{S}(\mathcal{C})$. By Corollary 2.13, d_T is circular decomposable and its support is $\mathrm{Split}(T)$. Statement (a) then follows from Proposition 2.8.

To prove statements (b) and (c), let $S_{ij} \in \mathrm{Split}(T)$ be arbitrary and e the unique edge of T inducing it. Again, by Proposition 2.8 and Corollary 2.13, we already have

$$\alpha_{ij}(d_T) = d_T(x_i, x_j) + d_T(x_{i+1}, x_{j+1}) - d_T(x_i, x_{j+1}) - d_T(x_{i+1}, x_j) > 0.$$

First suppose that $i + 1 \neq j$ and $j + 1 \neq i$. Then, since S_{ij} is the split induced by e, $x_{j+1}x_i|x_{i+1}x_j$ is a quartet induced by T whose internal edge is e. By the 4-point condition for additive metrics on trees (Buneman 1971),

$$\alpha_{ij}(d_T) = d_T(x_i, x_j) + d_T(x_{i+1}, x_{j+1}) - d_T(x_i, x_{j+1}) - d_T(x_{i+1}, x_j) = 2 \cdot w(e),$$

where $w(e)$ is the length of e under the quartet metrization. If $i+1 = j$ or $j+1 = i$, we similarly obtain that $\alpha_{ij} = 2 \cdot w(e)$. From Proposition 2.11, we have that $w(e) \geq |X|-2$ and (b) follows.

Now suppose that S_{ij} is a major split of T. Using Eq. (6), even if some of the sets A_1, A_2, B_1, B_2 are empty, we then always have that $w(e) > |A_1||A_2| + |B_1||B_2|$ and the bound in (c) follows. □

The next lemma shows that if N is a network from the class $\mathfrak{B}_2'$ inducing the circular order $\mathcal{C}$ and $S_{ij} \in \mathcal{S}(\mathcal{C})$, then, if S_{ij} is not induced by any displayed tree of N, its split weight will be zero under d_N.

Lemma 4.5 *Let N be a network on leaf set X from the class $\mathfrak{B}_2'$ with $n = |X| \geq 4$, $\mathcal{C} = (x_0, \ldots x_n = x_0)$ the circular order induced by N, and $S_{ij} \in \mathcal{S}(\mathcal{C})$. If $S_{ij} \notin$ Split$(\mathcal{T}(N))$, then $\alpha_{ij}(d_N) = 0$.*

Proof Recall that $\mu(N) = 4$. For simplicity, we thus show the result for the distances $4 \cdot d_N$. Let $\mathcal{P}$ be a canonical partition of X. Denote by $\widetilde{\mathcal{T}}(N) = \{T_{A_1A_2}, T_{A_1B_2}, T_{B_1A_2}, T_{B_1B_2}\}$ the multiset of trees displayed by N. Then, using the equation from Lemma 4.3 and the fact that $\mu(T, N) = 1$ when summing over all trees T in the multiset $\widetilde{\mathcal{T}}(N)$,

$$4 \cdot d_N(x, y) = \sum_{T \in \widetilde{\mathcal{T}}(N)} d_T(x, y) + 4 \cdot \varepsilon(x, y),$$

with $\varepsilon(x, y)$ as in Lemma 4.3. Let $Y = \{x_i, x_{i+1}, x_j, x_{j+1}\}$ and suppose $S_{ij} \notin$ Split$(\mathcal{T}(N))$. Then S_{ij} is a non-trivial split, and thus $x_{i+1} \neq x_j$ and $x_{j+1} \neq x_i$. Hence, $|Y| = 4$.

As shorthand notation, let $d_{kl} = 4 \cdot d_N(x_k, x_l)$ and $\varepsilon_{kl} = 4 \cdot \varepsilon(x_k, x_l)$. Similarly, for a tree $T \in \widetilde{\mathcal{T}}(N)$, let $d'_{kl}(T) = d_T(x_k, x_l)$, and $d'_{kl} = \sum_{T \in \widetilde{\mathcal{T}}(N)} d'_{kl}(T)$. Then Lemma 4.3 states that $d_{kl} = d'_{kl} + \varepsilon_{kl}$. Lastly, we write $D_1 = d_{i,j} + d_{i+1,j+1}$, $D_2 = d_{i,j+1} + d_{i+1,j}$, so that D_1 and D_2 correspond to the positive and negative terms for α_{ij} from Proposition 2.8. Similarly, we write D'_p, $D'_p(T)$, and E_p with $p \in \{1, 2\}$ for analogs of this expression using d', $d'(T)$, and ε, respectively.

In this notation,

$$\alpha_{ij}(4 \cdot d_N) = D_1 - D_2 = (D'_1 - D'_2) + (E_1 - E_2).$$

Since $S_{ij} \notin$ Split$(\mathcal{T}(N))$ and $\alpha_{ij}(d_T) = D'_1(T) - D'_2(T)$, Lemma 4.4(a) implies that $D'_1(T) - D'_2(T) = 0$ for all $T \in \widetilde{\mathcal{T}}(N)$, and hence $D'_1 - D'_2 = 0$. It is therefore enough to show $E_1 = E_2$. To do so, we consider three cases.

Case $|Y \cap C| = 0$: Since all leaves in Y are non-hybrids and x_i, x_{i+1} are neighbors in $\mathcal{C}$, $\{x_i, x_{i+1}\}$ is a subset of A or of B. Similarly, $\{x_j, x_{j+1}\}$ is a subset of A or of B. If $\{x_i, x_{i+1}\} \subseteq A$ and $\{x_j, x_{j+1}\} \subseteq B$ (or vice versa), then from Eq. (12) we obtain that $E_1 = E_2 = 0$. On the other hand, if $\{x_i, x_{i+1}\} \subseteq A$ and $\{x_j, x_{j+1}\} \subseteq A$ (or both are subsets of B), then by Eq. (12) we get that $E_1 = E_2 \in \{0, -2, -4\}$.

Case $|Y \cap C| = 1$: Without loss of generality, assume $x_i = c_1 \in C_1$ and $x_{i+1} \in B$. Note that either $\{x_j, x_{j+1}\} \subseteq A$ or $\{x_j, x_{j+1}\} \subseteq B$, since x_j, x_{j+1} are non-hybrids adjacent in $\mathcal{C}$. Assume first that $\{x_j, x_{j+1}\}$ is a subset of exactly one of A_1, A_2, B_1, or B_2. Then, since the error terms in Eq. (12) depend only on the specific set in which the leaves are contained, it follows that $\varepsilon_{i,j} = \varepsilon_{i,j+1}$ and $\varepsilon_{i+1,j+1} = \varepsilon_{i+1,j}$ and that $E_1 = E_2$.

Assume now that $x_j \in B_1$ and $x_{j+1} \in B_2$, and hence $x_{i+1} \in B_1$. Then S_{ij} is, for instance, a split of $T_{A_1B_2}$, a contradiction. With the circular order $\mathcal{C}$, it is not possible that $x_j \in B_2$ and $x_{j+1} \in B_1$, since x_{i+1} is not a hybrid. If $x_j \in A_2$ and $x_{j+1} \in A_1$, then the split S_{ij} is a split of $T_{A_1B_2}$ (also $T_{A_1A_2}$), a contradiction. Lastly, since $x_{i+1} \in B$, it is not possible that $x_j \in A_1$ and $x_{j+1} \in A_2$.

Case $|Y \cap C| = 2$: We will show that in this case the split S_{ij} will always be induced by a displayed tree, and hence this case cannot occur. We may again assume $x_i \in C_1$.

If $x_j \in C_2$, we may also assume that $x_{i+1} \in B$ and $x_{j+1} \in A$. Then, S_{ij} is a split in the displayed tree $T_{A_1B_2}$. Similarly, if $x_{j+1} \in C_2$, we may assume that $x_{i+1}, x_j \in B$. Then, S_{ij} is a split in the displayed tree $T_{A_1A_2}$. Finally, if $x_{i+1} \in C_2$, then either A or B is empty. Assuming the latter, then, $x_j, x_{j+1} \in A$ and S_{ij} is a split in $T_{A_1A_2}$. □

The next lemma can be seen as a counterpart to Lemma 4.5. It says that if N is a network from the class $\mathfrak{B}_2'$ inducing the circular order $\mathcal{C}$ and $S_{ij} \in \mathcal{S}(\mathcal{C})$, then, if S_{ij} is induced by some displayed tree of N, its split weight will be strictly positive under d_N.

Lemma 4.6 *Let N be a network on leaf set X from the class $\mathfrak{B}_2'$ with $n = |X| \geq 4$, let $\mathcal{C} = (x_0, \ldots, x_n = x_0)$ be the circular order induced by N, and let $S_{ij} \in \mathcal{S}(\mathcal{C})$. If $S_{ij} \in \mathrm{Split}(\mathcal{T}(N))$, then $\alpha_{ij}(d_N) > 0$.*

Proof Let $\mathcal{P}$ be a canonical partition of X, let $S_{ij} \in \mathrm{Split}(\mathcal{T}(N)) \subseteq \mathcal{S}(\mathcal{C})$ be arbitrary and let $Y = \{x_i, x_{i+1}, x_j, x_{j+1}\}$. Note that we have $|Y| \geq 3$ (as $|Y| \leq 2$ does not correspond to a split S_{ij}). We again show the result for the distances $4 \cdot d_N$, using the same notation as in the proof of Lemma 4.5. We consider four main cases.

Case 1, $|Y \cap C| = 0$ **and** $|Y| = 4$: In this case, all leaves in Y are non-hybrids and thus $|X| \geq 6$. Then, by Eq. (12), $E_1 \geq -4$ and $E_2 \leq 0$. Hence, $E_1 - E_2 \geq -4$. Since $S_{ij} \in \mathrm{Split}(\widetilde{\mathcal{T}}(N))$, there is a tree $T \in \widetilde{\mathcal{T}}(N)$ which induces S_{ij}. Then, relying on the definition of α_{ij} from Proposition 2.8, $\alpha_{ij}(d_T) = D_1'(T) - D_2'(T) \geq 2|X| - 4$ by part (b) of Lemma 4.4, while for all the other trees $T' \in \widetilde{\mathcal{T}}(N)$ by parts (a) and (b) of Lemma 4.4, $\alpha_{ij}(d_{T'}) = D_1'(T') - D_2'(T') \geq 0$. Thus, $D_1' - D_2' \geq 2|X| - 4 \geq 8$. Therefore, $\alpha_{ij}(4 \cdot d_N) = (D_1' - D_2') + (E_1 - E_2) > 0$.

Case 2, $|Y \cap C| = 1$ **and** $|Y| = 4$: In this case, there is exactly one hybrid leaf in Y, and we may assume $x_i \in C_1$, $x_{i+1} \in B$ and that the circular order $\mathcal{C}$ goes clockwise in Figure 5. Then x_j, x_{j+1} are both in B or both in A, which we consider as subcases.

Case 2.1, $x_j, x_{j+1} \in B$: If $x_{i+1} \in B_2$, then because the circular order $\mathcal{C}$ goes clockwise in Figure 5, we must have $x_j, x_{j+1} \in B_2$. Then by Eq. (12), $E_1 = E_2$, so $D_1 - D_2 = D_1' - D_2' > 0$ by Lemma 4.4(b). If $x_{i+1}, x_j, x_{j+1} \in B_1$, or if $x_{i+1} \in B_1$ and $x_j, x_{j+1} \in B_2$, we similarly find $E_1 = E_2$ and $D_1 - D_2 > 0$. Lastly, if $x_{i+1}, x_j \in B_1$ and $x_{j+1} \in B_2$, then $E_1 = 2|A_2| + 2|B_2| - 2$ and $E_2 = -2|A_1|$, so $E_1 - E_2 = 2(|A_2| + |B_2| + |A_1| - 1) > 0$. We again have that $D_1' - D_2' > 0$ by Lemma 4.4(b), so $D_1 - D_2 > 0$.

Case 2.2, $x_j, x_{j+1} \in A$: By Eq. (12), $\varepsilon_{i+1,j} = \varepsilon_{i+1,j+1} = 0$, so $E_1 = \varepsilon_{i,j}$ and $E_2 = \varepsilon_{i,j+1}$. Moreover, we cannot have $x_j, x_{j+1} \in A_2$, because then S_{ij} would not be a split of a tree in $\widetilde{\mathcal{T}}(N)$. If $x_j, x_{j+1} \in A_1$, then $E_1 = E_2$. By Lemma 4.4(b) we know that $D_1' - D_2' > 0$, so we get that $D_1 - D_2 > 0$. If instead $x_j \in A_2$ and $x_{j+1} \in A_1$, then $E_1 = -2|B_1|$ and $E_2 = 2|A_2| + 2|B_2|$. So, $E_1 - E_2 = -2|A_2| - 2|B_1| - 2|B_2| \geq -2|X| + 6$. Note that in this case S_{ij} is a split in both $T_{A_1A_2}$ and $T_{A_1B_2}$. Hence, $D_1' - D_2' \geq 4|X| - 8$ by Lemma 4.4(b). Thus, $D_1 - D_2 > 0$.

Case 3, $|Y \cap C| = 2$ **and** $|Y| = 4$: Without loss of generality, assume that $x_i \in C_1$ and that the circular order $\mathcal{C}$ goes clockwise in Figure 5. We consider three cases depending on the location of the other hybrid in the circular order of the 4 taxa.

Case 3.1, $x_{i+1} \in C_2$: For x_i and x_{i+1} to be neighbors in $\mathcal{C}$, we must have that $|B| = 0$. We now consider two subcases, noting that the case $x_j, x_{j+1} \in A_1$ follows from Case 3.1a by symmetry.

- Case 3.1a, $x_j, x_{j+1} \in A_2$: In this case $\varepsilon_{i,j} = \varepsilon_{i,j+1}$ and $\varepsilon_{i+1,j} = \varepsilon_{i+1,j+1}$, so $E_1 = E_2$. Since $D_1' - D_2' > 0$ by Lemma 4.4(b), $D_1 - D_2 > 0$.
- Case 3.1b, $x_j \in A_2$ and $x_{j+1} \in A_1$: Now, $E_1 = -2|B_1| - 2|B_2| = -2|B| = 0$ and $E_2 = 2|A_1| + 2|B_1| + 2|A_2| + 2|B_2| = 2|A|$ (since $|B| = 0$ in Case 3.1). Hence, $E_1 - E_2 = -2|A| = -2|X| + 4$. Since S_{ij} is a split of $T_{A_1A_2}$, $T_{A_1B_2}$ and $T_{B_1A_2}$, by Lemma 4.4(b), $D_1' - D_2' \geq 6|X| - 12$. Thus, $D_1 - D_2 = (D_1' - D_2') + (E_1 - E_2) \geq 4|X| - 8 > 0$.

Case 3.2, $x_j \in C_2$: We consider three subcases. Note that the case where $x_{i+1} \in B_2$ and $x_{j+1} \in A_2$ is analogous to Case 3.2c. Furthermore, we have that $E_1 = -2|A_1||A_2| - 2|B_1||B_2|$ in all three subcases.

- Case 3.2a, $x_{i+1} \in B_1$ and $x_{j+1} \in A_2$: Thus, $E_2 = -2|B_1| - 2|A_2|$. Next, note that S_{ij} is a major split of $T_{A_1B_2}$. Thus, by Lemma 4.4(c), $D_1' - D_2' \geq D_1'(T_{A_1B_2}) - D_2'(T_{A_1B_2}) > 2|A_1||A_2| + 2|B_1||B_2|$. Since $E_1 - E_2 \geq E_1 \geq -2|A_1||A_2| - 2|B_1||B_2|$, we have $D_1 - D_2 = (D_1' - D_2') + (E_1 - E_2) > 0$.
- Case 3.2b, $x_{i+1} \in B_2$ and $x_{j+1} \in A_1$: Since x_i, x_{i+1} and x_j, x_{j+1} are neighbors in $\mathcal{C}$, $|A_2| = |B_1| = 0$. Thus $E_1 = 0$ and since $\varepsilon_{j,i+1} = 2|A_1| + 2|B_1| = 2|A_1|$ and $\varepsilon_{j+1,i} = 2|A_2| + 2|B_2| = 2|B_2|$, we find $E_2 = 2|A_1| + 2|B_2|$. Thus $E_1 - E_2 = -2|A_1| - 2|B_2| = -2|X| + 4$. Since S_{ij} is a split of T_{A_1,A_2} and T_{A_1,B_2}, by Lemma 4.4(b), $D_1' - D_2' \geq 4|X| - 8$ and it follows that $D_1 - D_2 > 0$.
- Case 3.2c, $x_{i+1} \in B_1$ and $x_{j+1} \in A_1$: We must have $|A_2| = 0$. Moreover, $\varepsilon_{j,i+1} = -2|A_2| = 0$ and $\varepsilon_{j+1,i} = 2|A_2| + 2|B_2| = 2|B_2|$, so $E_2 = 2|B_2|$. Using $|A_2| = 0$ again, $E_1 - E_2 = -2|B_1||B_2| - 2|B_2|$. Then, since S_{ij} is a major split of $T_{A_1A_2}$ and $T_{A_1B_2}$, we have by Lemma 4.4(c) that $D_1'(T_{A_1A_2}) - D_2'(T_{A_1A_2}) > 2|A_1||A_2| + 2|B_1||B_2| = 2|B_1||B_2|$, and similarly $D_1'(T_{A_1B_2}) - D_2'(T_{A_1B_2}) > 2|B_1||B_2|$. Thus, $D_1' - D_2' > 4|B_1||B_2|$ and so $D_1 - D_2 = (D_1' - D_2') + (E_1 - E_2) > 2|B_1||B_2| - 2|B_2| \geq 0$. In particular, $D_1 - D_2 > 0$.

Case 3.3, $x_{j+1} \in C_2$: We consider two subcases. Note that the case with $x_{i+1}, x_j \in B_1$ follows from Case 3.3b by symmetry.

- Case 3.3a, $x_{i+1} \in B_1$ and $x_j \in B_2$: Then, $E_1 = -2|A_1| - 2|A_2| = -2|A|$ and $E_2 = -2|A_1||A_2| - 2|B_1||B_2| - 2$. Hence, $E_1 - E_2 \geq E_1 = -2|A| \geq -2(|X| - 4)$ using that $|B|, |C| \geq 2$ in this case. Then, since S_{ij} is a split in $T_{A_1A_2}$, we have by Lemma 4.4(b) that $D_1' - D_2' \geq 2|X| - 4$. Thus, $D_1 - D_2 > 0$.

- Case 3.3b, $x_{i+1}, x_j \in B_2$: Since x_i and x_{i+1} are neighbors in $\mathcal{C}$, we must have $|B_1| = 0$. We also have $E_1 = -2|A_1| + 2|B_1| + 2|A_1| = 0$ and $E_2 = -2|A_1||A_2| - 2|B_1||B_2| = -2|A_1||A_2|$. Thus, $E_1 - E_2 \geq 0$. Since $D_1' - D_2' > 0$ by Lemma 4.4(b), it follows that $D_1 - D_2 > 0$.

Case 4, $|Y| = 3$: We may assume $x_{i+1} = x_j$, so $E_1 = \varepsilon_{i,j} + \varepsilon_{j,j+1}$ and $E_2 = \varepsilon_{i,j+1}$. Moreover, since S_{ij} is the trivial split $\{x_j\} \mid \big(X \setminus \{x_j\}\big)$ which appears in all four trees in $\widetilde{\mathcal{T}}(N)$, by Lemma 4.4(b), $D_1' - D_2' \geq 8|X| - 16$. If $|Y \cap C| = 0$, the argument for Case 1 applies (with $|X| \geq 5$) , so we may henceforth assume $|Y \cap C| \geq 1$. We are left with the following four subcases, up to symmetry.

Case 4.1, $x_i \in C$ and $x_j, x_{j+1} \notin C$: In this case, we have that $E_1 \geq -2\max\{|A_1|, |A_2|, |B_1|, |B_2|\} - 2 \geq -2(|X| - 2) - 2 = -2|X| + 2$, while $E_2 \leq 2\max\{|A_1| + |B_1|, |A_2| + |B_2|\} \leq 2|X| - 4$. Hence, $E_1 - E_2 \geq -4|X| + 6$. Thus, since $D_1' - D_2' \geq 8|X| - 16$, we obtain $D_1 - D_2 > 0$.

Case 4.2, $x_j \in C$ and $x_i, x_{j+1} \notin C$: In this case we get that $E_1 \geq -4\max\{|A_1|, |A_2|, |B_1|, |B_2|\} \geq -4|X| + 8$, while $E_2 \leq 0$. Hence, $E_1 - E_2 \geq -4|X| + 8$ and again $D_1 - D_2 > 0$ as in Case 4.1.

Case 4.3, $x_i, x_j \in C$ and $x_{j+1} \notin C$: We may assume that $x_i \in C_1, x_j = x_{i+1} \in C_2$. Since x_i and x_{i+1} are neighbors in $\mathcal{C}$, we may also assume $|B| = 0$ and $x_{j+1} \in A_2$. Then, $E_1 = -2|A_1||A_2| - 2|B_1||B_2| + 2|A_1| + 2|B_1| = -2|A_1||A_2| + 2|A_1|$. Since $E_2 = -2|B_1| = 0$, we see $E_1 - E_2 = -2|A_1||A_2| + 2|A_1|$. Now consider the tree $T_{A_1B_2}$ and let e be the edge in this tree inducing the split $S_{ij} = C_2|\,(X \setminus C_2)$. Then, under the quartet metrization w, we have that $w(e) = (|A_1| + 1)|A_2|$ (see Eq. (7)). Hence, by a similar argument as in the proof of Lemma 4.4(c), we obtain that $D_1' - D_2' \geq D_1'(T_{A_1B_2}) - D_2'(T_{A_1B_2}) > 2|A_1||A_2|$. Therefore, $D_1 - D_2 = (D_1' - D_2') + (E_1 - E_2) > 0$.

Case 4.4: $x_i, x_{j+1} \in C$ and $x_j \notin C$. By symmetry, we may assume that $x_i \in C_1$, $x_j \in B_1$ and $x_{j+1} \in C_2$, so $|B_1| = 1$ and $|B_2| = 0$. Then, $E_1 = 2|A_2| + 2|B_2| - 2|A_2| = 2|B_2| = 0$ and $E_2 = -2|A_1||A_2| - 2|B_1||B_2| = -2|A_1||A_2|$. Hence, $E_1 - E_2 \geq 0$, so $D_1 - D_2 > 0$. □

We now show that the NANUQ metric is circular decomposable with support exactly the set of splits induced by the displayed trees of a network. We conjecture that this theorem can be extended to the class $\mathfrak{N}_2$ (for some more details, see the discussion in Section 6).

Theorem 4.7 *Let N be a semi-directed network on at least four leaves from the class $\mathfrak{B}_2$. Then, the NANUQ metric d_N is circular decomposable and its support is* $\mathrm{Split}(\mathcal{T}(N))$.

Proof We may assume that N is in $\mathfrak{B}_2'$, i.e., is strictly level-2, since the result is known for level-1 networks (Allman et al. 2019). By Proposition 2.9, $\mathrm{Split}(\mathcal{T}(N))$ is a circular split system congruent with the circular order induced by N. Then, by Proposition 2.8 and Lemma 4.5 and 4.6, d_N is circular decomposable and has support $\mathrm{Split}(\mathcal{T}(N))$. □

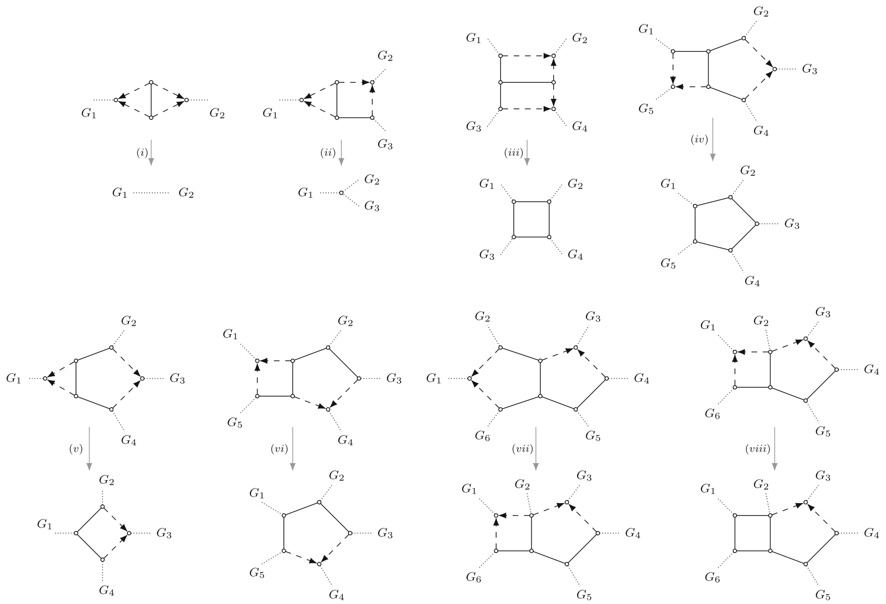

Fig. 7 Illustration of operations (i)-$(viii)$ used in Definition 5.1.

5 Identifiability of a canonical form

In this section, we prove identifiability results for a canonical form of networks in $\mathfrak{B}_2$ and $\mathfrak{N}_2$, which can be obtained by applying a series of operations introduced below. Since the canonical form may not itself be a semi-directed network, its main purpose is not to infer the quartet-related properties from previous sections directly — although quartet information can still be recovered by reversing the operations until a valid semi-directed network is reached. Instead, the canonical form is primarily intended to determine whether a pair of networks is distinguishable. Moreover, it captures the essential features relevant for inference and reflects the kind of output one would expect from an algorithmic method based on our results (see also the discussion in Section 6).

Recall from Section 2 that, given a semi-directed network N on X, the semi-directed network obtained by *contracting* its 2-blobs is the network where every 2-blob in N is replaced by a degree-2 node which is then suppressed (see Figure 7(i)). Similarly, the semi-directed network obtained by *contracting* its 3-blobs is the network where every 3-blob in N is replaced by a degree-3 node (see Figure 7(ii)). By contracting a 3-cycle in a blob B, we mean that the 3-cycle is replaced by a single node (see Figure 7(v)). By *undirecting* a 4-cycle in a semi-directed network, we mean that every directed edge in the 4-cycle is replaced with an undirected edge (see Figure 7(viii)). Since actions such as these result in graphs which are not necessarily semi-directed networks, we refer to these as ‘mixed graphs’.

If N is outer-labeled planar, each k-blob B of N (with $k \geq 4$) induces a unique circular order of the subnetworks around the blob B (Rhodes et al. 2025). The mixed

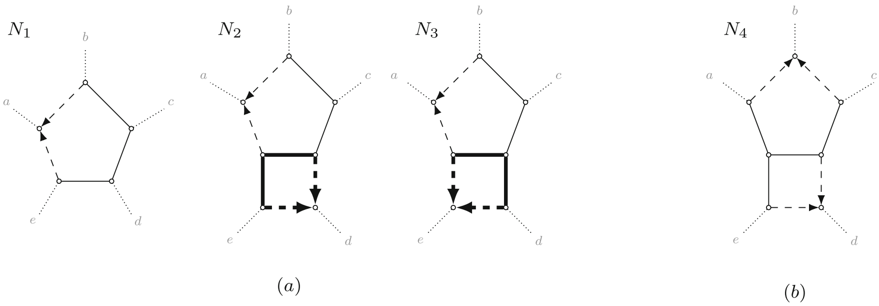

Fig. 8 The four possible 3-cycle-free 5-leaf networks in $\mathfrak{B}_2$, up to relabeling the leaves. Each subfigure contains networks with the same canonical form. The networks N_2 and N_3 both have a dts-undetectable 4-cycle (in thick black), and their canonical form is the network N_1. The network N_4 consists of the split symmetric 5-blob, with its canonical form obtained by replacing the blob by a completely undirected 5-cycle (see also Figure 7(iv)).

graph obtained from N by replacing B by its *representative cycle* is the graph where B is replaced with an undirected k-cycle inducing the same circular order of the subnetworks around B (see e.g. Figure 7(iii) and (iv)).

Leading up to Definition 5.1, we restrict attention to N in $\mathfrak{N}_2$, and B in $\mathfrak{B}_2'$ a k-blob, $k \geq 5$, with hybrid nodes u and v. If B is a 5-blob that is isomorphic to the blob of Figure 8(b), then we say B is *split symmetric*. This naming stems from the fact that the set of splits of the displayed trees in such a blob remains the same when cyclically permuting the leaves.

A 4-cycle in B is a *dts-undetectable 4-cycle (displayed-tree-split undetectable 4-cycle)* if it contains a hybrid node u and a parent p_v of the other hybrid node v. See Figure 8(a) for two dts-undetectable 4-cycles in 5-blobs, and Figure 9(a) for two dts-undetectable 4-cycles in 6-blobs. Every dts-undetectable 4-cycle can be labeled as (p_u, u, p'_u, w), where p_u, p'_u are the two parents of u with p_u, u articulation nodes and one of $\{p'_u, w\}$ a parent of the other hybrid node v. By *suppressing* a dts-undetectable 4-cycle (p_u, u, p'_u, w), we mean deleting the hybrid edge (p_u, u), undirecting the hybrid edge (p'_u, u), and subsequently suppressing the two resulting degree-2 nodes. Suppressing a dts-undetectable 4-cycle turns B into a level-1 blob with the same circular order and with the hybrid v retained as the unique hybrid node in the now-cycle B (see e.g. Figure 7(vi)).

A non-hybrid edge of B incident to an articulation node is called a *frontier edge*. The frontier edges of B are exactly the non-hybrid edges on the 'outside of the blob' in a planar representation, such as in Figures 8 anf 9. Lastly, suppose that B is not a split symmetric 5-blob and does not have a dts-undetectable 4-cycle (note that then $k \geq 6$). A frontier edge $\{s, t\}$ of B is a *dts-undetectable edge (displayed-tree-split undetectable edge)* if (1) both s and t are parents of a hybrid in B, or (2) s is a parent of a hybrid in B and t is part of a 4-cycle in B (or vice versa). See Figure 9(e) for examples. Note that since $k \geq 6$, each blob can have at most one dts-undetectable edge.

The operations and definitions illustrated in Figures 8 to 9, now permit the following.

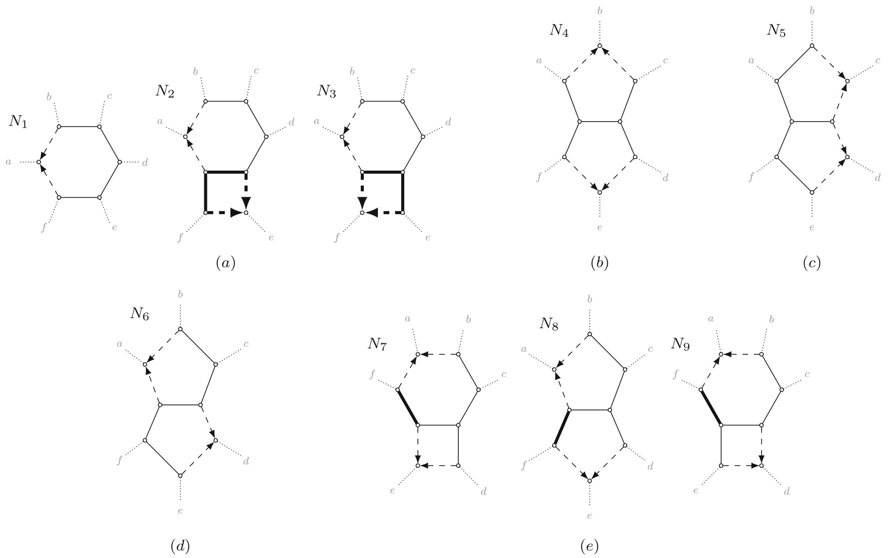

Fig. 9 The nine possible 3-cycle-free 6-leaf networks in $\mathfrak{B}_2$, up to relabeling the leaves. Each subfigure contains networks with the same canonical form. The networks N_2 and N_3 both have a dts-undetectable 4-cycle (in thick black), with canonical form N_1. The networks N_7 and N_8 have a dts-undetectable edge of type 1 (in thick black), and network N_9 has a dts-undetectable edge of type 2 (in thick black). The canonical form for the networks in class (e) is non-binary (i.e, the degree of some non-leaf vertex is strictly greater than 3) and obtained by contracting the thick black edges and undirecting the hybrid edges in the remaining 4-cycle within the blob (see also Figure 7(vii) and (viii))

Definition 5.1 Let N be a semi-directed network on X from the class $\mathfrak{N}_2$. The *canonical form* N^c of N is the mixed graph obtained from N by applying the following operations in order:

(i) contracting every 2-blob;
(ii) contracting every 3-blob;
(iii) replacing every 4-blob with its representative cycle;
(iv) replacing every split symmetric 5-blob with its representative cycle;
(v) contracting 3-cycles in k-blobs ($k \geq 5$);
(vi) suppressing dts-undetectable 4-cycles in k-blobs ($k \geq 5$);
(vii) contracting dts-undetectable edges in k-blobs ($k \geq 6$);
(viii) undirecting the remaining 4-cycles in k-blobs ($k \geq 6$)

Since operations (i)-$(viii)$ only alter the structure of individual blobs, the operations result in the same mixed graph independent of the order in which the blobs are considered. Hence, the canonical form of a network is unique.

The following characterizes those networks whose canonical form matches itself. Note that for 2-, 3- and 4-cycle-free bloblets in $\mathfrak{B}_2'$, the condition that a bloblet is in canonical form is, equivalently, that the two hybrids x_{h_1}, x_{h_2} are not separated by exactly one leaf, $\ldots, x_{h_1}, x_l, x_{h_2}, \ldots$, in the induced circular order.

Observation 5.2 *Let N be a 2-, 3-, and 4-cycle-free semi-directed network from the class $\mathfrak{N}_2$ such that the edge distance between two hybrid nodes in any strictly level-2 blob is not 3. Then, $N \cong N^c$.*

The importance of the canonical form is highlighted in the following.

Lemma 5.3 *Let N_1 and N_2 be two networks from $\mathfrak{N}_2$ on the same leaf set X. Then, $N_1^c \cong N_2^c$ if and only if* $\mathrm{Split}(\mathcal{T}(N_1)) = \mathrm{Split}(\mathcal{T}(N_2))$.

Proof First note that operations (i)-$(viii)$ from Definition 5.1 preserve the tree-of-blobs of a network. Since the displayed splits of a network uniquely determine its displayed quartets, and hence its tree-of-blobs (Allman et al. 2023), it suffices to consider N_1, N_2 with the same tree-of-blobs. Hence, we can treat the blobs separately and may assume $N_1, N_2 \in \mathfrak{B}_2$.

The result is straightforward if $n = |X| \in \{2, 3, 4\}$, so we assume $n \geq 5$ and disregard operations (i)-(iii) in Definition 5.1. It can be checked directly that two networks that are the same after applying operation (iv), which applies to a specific 5-taxon network, have the same displayed splits. We next explain that operations (v) and (vi), which keep networks in $\mathfrak{B}_2$, have no effect on the set of splits of the trees displayed by the network.

Operation (v), contracting a 3-cycle to a node preserves the set of displayed trees, since a 3-cycle displays exactly the same trees as a 3-leaf tree.

If a network has a dts-undetectable 4-cycle, it must have one of two structures. See Figures 8(a) and 9(a) for these in 5- and 6-blobs, and note that for $n \geq 7$ the cycle at the 'top' simply becomes larger (i.e., has more leaves between b and c). The suppression of these 4-cycles by operation (vi) does remove some of the displayed trees, all of which are caterpillars, so that, for instance, in N_2 of Figure 9 the tree $(((f, e), a))...)$ is not displayed anymore. This only removes the split $fe|a....$ As this split is on other displayed trees before and after operation (vi), it has no impact on the collection of splits of the displayed trees.

Next, consider operations (vii) and $(viii)$ of Definition 5.1. We will show that if two networks become isomorphic after applying these operations, they have the same displayed splits. If only operation (vii) (resp. operation $(viii)$) is applied, the pair of networks must be isomorphic to the pair of networks in Figure 11(a) (resp. Figure 11(b)). Lastly, if both operations (vii) and $(viii)$ are applied, the pair of networks comes from the three networks N_7, N_8, and N_9 in Figure 9(e), with possibly more leaves between b and c if $n > 6$. Similar to the previous paragraph, it is straightforward to check that the sets of displayed splits coincide in all three cases.

For the remainder of the proof, we may assume that $N_1 = N_1^c$ and $N_2 = N_2^c$ have the canonical form produced by applying operations (iv)-$(viii)$, so they have no 3-cycles, no dts-undetectable edges, are not symmetric 5-blobs, may be non-binary only as produced by operation (vii), and only have undirected 4-cycles not stemming from dts-undetectable 4-cycles. That $N_1^c = N_2^c$ implies $\mathrm{Split}(\mathcal{T}(N_1)) = \mathrm{Split}(\mathcal{T}(N_2))$ is now obvious. It remains to show the converse.

Suppose then that $N_1 = N_1^c \ncong N_2^c = N_2$. If N_1, N_2 induce different circular orders, then by Rhodes (2025), Thm. 5.3, the displayed quartets of N_1 and N_2 differ, and hence $\mathrm{Split}(\mathcal{T}(N_1)) \neq \mathrm{Split}(\mathcal{T}(N_2))$, as required. Thus we may assume N_1 and N_2 induce the same circular orders henceforth.

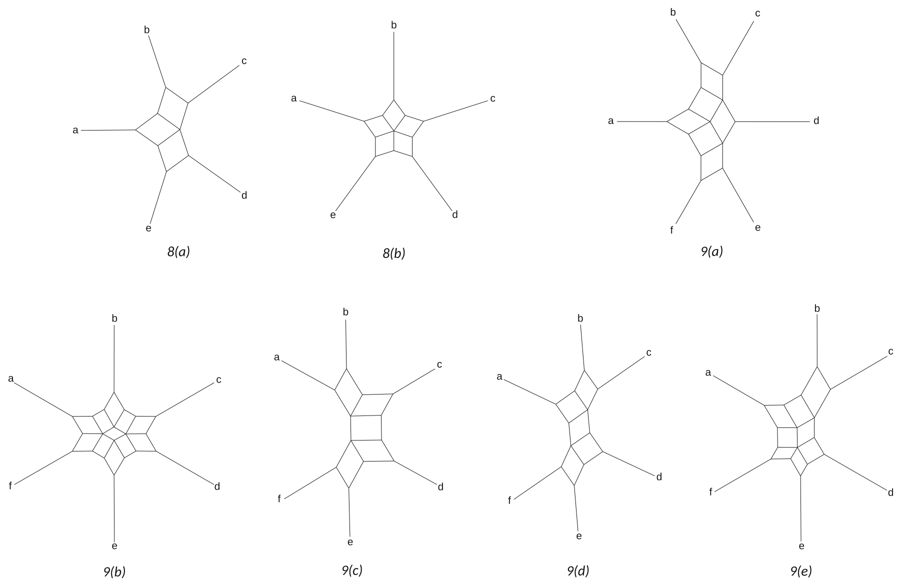

Fig. 10 Splits graphs of the displayed splits of the seven sets of networks with different canonical forms in Figures 8 and 9

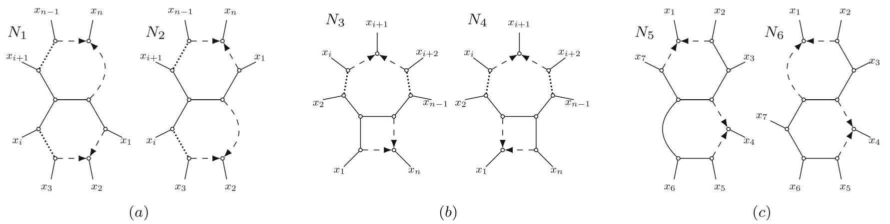

Fig. 11 Three pairs of networks from the class $\mathfrak{B}_2$ used in the proof of Lemma 5.3. Dotted edges indicate locations where additional leaves may be attached. The networks N_1 and N_2 in subfigure (a) have $n \geq 7$ leaves and $4 \leq i \leq n-3$. The networks N_3 and N_4 in subfigure (b) have $n \geq 7$ leaves and $3 \leq i \leq n-4$

We will prove that $\text{Split}(\mathcal{T}(N_1)) \neq \text{Split}(\mathcal{T}(N_2))$ by induction on n, with base cases $n = 5$ and $n = 6$. Since $\text{Split}(\mathcal{T}(N_1))$ and $\text{Split}(\mathcal{T}(N_2))$ are both (unweighted) circular split systems, they can be depicted by a splits graph (see e.g. Huson et al. (2010) for details). Figure 10 shows that the splits graphs of networks from Figures 8 and 9 (i.e., when $n = 5$ or $n = 6$) are distinct for different canonical forms. Hence, the base cases hold.

Now suppose $n \geq 7$ and that the theorem holds for smaller n. We will repeatedly use that if $N_1|_Y{}^c \ncong N_2|_Y{}^c$ for some $Y \subset X$, then $\text{Split}(\mathcal{T}(N_1|_Y)) \neq \text{Split}(\mathcal{T}(N_2|_Y))$ by the induction hypothesis, and hence $\text{Split}(\mathcal{T}(N_1)) \neq \text{Split}(\mathcal{T}(N_2))$. We denote this implication by $(*)$.

We first consider the case where one network, say N_1, is isomorphic to the network N_5 of Figure 11(c). Then, every displayed tree on the network $(N_1|_{X\setminus\{x_1\}})^c$ (resp. $(N_1|_{X\setminus\{x_4\}})^c$) induces the non-trivial split $x_2x_3|x_4x_5x_6x_7$ (resp. $x_5x_6|x_1x_2x_3x_7$).

If $(N_2|_{X\setminus\{x_1\}})^c$ (resp. $(N_2|_{X\setminus\{x_4\}})^c$) does not have this split on every displayed tree, we are done by $(*)$. This implies x_1 and x_4 are hybrids in both networks. Moreover, x_1, x_2, x_3 (resp. x_4, x_5, x_6) must be on the same cycle to get these splits. Since $N_1 \not\cong N_2$ but the networks induce the same circular order, we must then have that N_2 is isomorphic to network N_6 in Figure 11(c). But then $(N_1|_Y)^c \not\cong (N_2|_Y)^c$ for $Y = \{x_1, x_2, x_3, x_4, x_5, x_7\}$ and we are done by $(*)$.

We may now assume that neither N_1 nor N_2 is isomorphic to N_5 in Figure 11(c). Then, we claim there exists a set of leaves $\{x, a, b, c\}$ such that (i) N_1 (and hence N_2) induces the circular order $(x, a, b, c, \ldots)$, (ii) x is a hybrid leaf of N_1, (iii) c is not a hybrid leaf of N_1, (iv) x, a, b are incident to the same k-cycle S of N_1 with $k \geq 6$ (i.e. at least 4 leaves are incident to this cycle). A set satisfying properties (i)-(iii) must exist since $n \geq 7$, and $N_1, N_2 \not\cong N_5$ ensures (iv) is also possible.

Properties (i)-(iv) ensure that by taking the subnetwork of N_1 induced by $X \setminus \{a\}$ (where we write $N_1' = N_1|_{X\setminus\{a\}}$ and $N_2' = N_2|_{X\setminus\{a\}}$), no 3-cycles, dts-undetectable edges, or new 4-cycles are created. Therefore, $(N_1')^c$ can simply be obtained from N_1^c by removing a and suppressing the resulting degree-2 node. Equivalently, N_1^c can be obtained from $(N_1')^c$ by inserting the leaf a between leaves x and b. If $(N_1')^c \cong (N_2')^c$, then since N_1 and N_2 induce the same circular order, N_2^c can also be obtained from $(N_2')^c$ by inserting the leaf a between leaves x and b. This is a contradiction, since then $N_1^c \cong N_2^c$. Hence, we may assume $(N_1')^c \not\cong (N_2')^c$. It then follows from $(*)$ that $\mathrm{Split}(\mathcal{T}(N_1)) \neq \mathrm{Split}(\mathcal{T}(N_2))$. □

In the following proposition, we use the notation $\mathcal{Q}(N) = \mathcal{Q}(\mathcal{T}(N))$ to denote the set of quartet trees induced by the displayed trees of a network N.

Proposition 5.4 *Let N_1 and N_2 be two networks from $\mathfrak{N}_2$ on the same leaf set X. Then, the following are equivalent:*

(i) $N_1^c \cong N_2^c$;
(ii) $\mathcal{Q}(N_1) = \mathcal{Q}(N_2)$;
(iii) $\mathrm{Split}(\mathcal{T}(N_1)) = \mathrm{Split}(\mathcal{T}(N_2))$.

Furthermore, if $N_1, N_2 \in \mathfrak{B}_2 \subseteq \mathfrak{N}_2$ and $|X| \geq 4$, then (i)-(iii) are equivalent to

(iv) $d_{N_1} = d_{N_2}$.

Proof The equivalence between (i) and (iii) was shown as Lemma 5.3. That (iii) implies (ii) follows easily, since one can obtain the set of displayed quartets from the set of splits of the displayed trees. To see that (ii) implies (iii), suppose $\mathcal{Q}(N_1) = \mathcal{Q}(N_2)$. Then, by Allman et al. (2023), N_1 and N_2 have the same trees-of-blobs T, and the splits corresponding to cut edges of N_1 and N_2 are equal. Thus, we can consider each blob of N_1 and N_2 separately, and it suffices to assume that $N_1, N_2 \in \mathfrak{B}_2$. If $|X| \leq 3$, (iii) is trivially true. If instead $|X| \geq 4$, $d_{N_1} = d_{N_2}$, since the distances d_{N_1} and d_{N_2} are obtained from the quartets of N_1 and N_2, and we assumed those sets were equal. By Theorem 4.7 and because the decomposition of a circular decomposable metric is unique, we obtain (iii).

The remaining implications follow immediately since if $N_1, N_2 \in \mathfrak{B}_2$ and $|X| \geq 4$, (ii) implies (iv) and (iv) implies (iii). □

Note that in the previous proposition, (i)-$(iii) \Rightarrow (iv)$ is also true if $N_1, N_2 \in \mathfrak{N}_2$, but for the converse $N_1, N_2 \in \mathfrak{B}_2$ is necessary. In particular, we cannot use a similar argument as before, since we have not proved that $d_{N_1} = d_{N_2}$ implies N_1 and N_2 have the same tree-of-blobs. As mentioned before Theorem 4.7, we conjecture that this is true for networks in the class $\mathfrak{N}_2$. Indeed, proving this extension of Theorem 4.7 to the class $\mathfrak{N}_2$ would suffice to obtain the equivalence (i)-$(iii) \Leftrightarrow (iv)$ in Proposition 5.4 for networks in $\mathfrak{N}_2$, and vice versa.

With a slight reformulation of Proposition 5.4, we obtain the main theorem.

Theorem 5.5 *(a) Let N_1 and N_2 be two semi-directed, outer-labeled planar, level-2, galled networks. Then, N_1 and N_2 are distinguishable from their displayed quartets if and only if N_1 and N_2 have different canonical forms.*

(b) Let N_1 and N_2 be two semi-directed, outer-labeled planar, level-2, galled bloblet networks on at least four leaves. Then, N_1 and N_2 are distinguishable from their pairwise NANUQ distances if and only if N_1 and N_2 have different canonical forms.

As an instance of part (b) of the theorem, the pairwise NANUQ distances for each of the five canonical classes of $\mathfrak{B}_2$ networks on 6 taxa shown in Figure 9 are given in Table 1. These, together with distances obtained by permuting taxon labels, each correspond to a unique labeled canonical form, although only 3 of the forms, 9(b), 9(c), 9(d), determine a unique network in $\mathfrak{B}_2$. Nonetheless, from the splits graphs in Figure 10, the circular order of taxa is immediately seen in all cases, and the splits graphs shapes show that the hybrid node in any k-cycle with $k \geq 5$ and no dts-undetectable edge is determined. Note that hybrid identifiability from this distance does not hold for all 5-blobs, since the symmetry of the splits graph for 8(b) gives no hybrid information.

Looking to the future, for data analysis grounded in the main results of Theorem 5.5, note that when considering outer-labeled-planar networks, the necessary quartet information for applying the theorem is known to be obtainable under several models and data types currently in use for inference. For identifiability from gene trees, these include models of gene tree formation by variations of the Network Multispecies Coalescent (NMSC) process with both independent and common inheritance, and a displayed tree (DT) model in which gene trees are formed on the network without incomplete lineage sorting (Rhodes et al. 2025). Quartet relationships are also identifiable directly from sequence data, assuming a Jukes-Cantor or Kimura-2-Parameter substitution process on the displayed trees of the network (Englander et al. 2025). Although both of these works prove identifiability results beyond that of quartet information, neither suggests a clear path to practical inference. Indeed, although Theorem 5.5 is phrased as an identifiability result, its underlying reliance on the NANUQ distance has the potential to contribute to fast network inference methods.

6 Discussion

In this study, we have focused on the class of outer-labeled planar, galled, level-2, semi-directed networks: a class of phylogenetic networks more general than level-

Table 1 Pairwise NANUQ distances for 6-taxon bloblet networks with the five canonical forms of Figure 9. As proved in Theorem 5.5, these distances distinguish the five canonical forms.

(a) Distances for networks N_1, N_2 and N_3, all with the same canonical form. This class includes the level-1 6-cycle network N_1.

	b	*c*	*d*	*e*	*f*
a	14	17	18	17	14
b		11	16	19	20
c			15	18	19
d				15	16
e					11

(b) Distances for N_4.

	b	*c*	*d*	*e*	*f*
a	14	18	18	17	13
b		14	17	18	17
c			13	17	18
d				14	18
e					14

(c) Distances for N_5.

	b	*c*	*d*	*e*	*f*
a	11	16	18	18	17
b		13	19	19	18
c			14	19	18
d				13	16
e					11

(d) Distances for N_6.

	b	*c*	*d*	*e*	*f*
a	13	16	18	17	16
b		11	17	20	19
c			16	19	18
d				13	16
e					11

(e) Distances for networks N_7, N_8 and N_9, all with the same canonical form.

	b	*c*	*d*	*e*	*f*
a	14	17	18	17	14
b		11	17	19	19
c			16	18	18
d				13	16
e					13

1, which allows some interdependent reticulate events. While this formally means that all blobs are assumed to be outer-labeled planar, galled, and level-2, a careful reading of our proofs shows they in fact permit 2- and 3-blobs not meeting these conditions. Consequently, our results can be applied slightly more generally than the class description initially suggests. We have established that most features of the networks in this class — captured precisely by a canonical form — are identifiable from displayed quartets, implying by results elsewhere (Englander et al. 2025; Rhodes et al. 2025) that identifiability holds from different types of biological data under several models of evolution (see the end of Section 5).

A second contribution of this work is that our proof is constructive and lays the theoretical foundation for a consistent inference algorithm for the canonical form of a class of level-2 networks under the Network Multispecies Coalescent model, in the spirit of NANUQ and NANUQ$^+$ (Allman et al. 2019, 2025a). Notably, this would yield the first algorithm for a subclass of level-2 networks that is known to be statistically consistent for specific models and data types. While both purely combinatorial approaches (e.g., Iersel et al. 2009, 2022,) and computationally cumbersome Bayesian and pseudolikelihood approaches (e.g., Yu and Nakhleh 2015; Zhang et al. 2018; Wen and Nakhleh 2018,) exist, neither addresses potential non-identifiability issues. The specific algorithm, together with its implementation and performance analysis, will follow in future work; however, we sketch an initial outline in the following paragraph.

First, using the existing software tool TINNIK (Allman et al. 2024a), the tree-of-blobs of a network can be consistently constructed from *concordance factors*: summary statistics derived from gene tree probabilities under the Network Multispecies Coalescent model. Once this tree is constructed, it suffices to consider every blob separately, and hence we focus on bloblets for the remainder of this paragraph. Specifically, the concordance factors give information about displayed quartets and they can be used to compute the pairwise NANUQ distances between the leaves of a bloblet (see again Table 1). Then, when input to the consistent method NEIGHBOR- NET for fitting circular decomposable metrics, these distances give rise to a *splits graph* (Bryant and Moulton 2004). As we have shown in Theorem 4.7, the NANUQ distances are circular decomposable and thus, for perfect data, the resulting splits graph depicts the splits of the displayed trees of the network and they correspond to a unique canonical form of the network (Proposition 5.4). Using concepts similar to the *darts* of Allman et al. (2019), this canonical form can be derived directly from the splits graph. While we do not explicitly treat this here, as noted after Theorem 5.5, some of this is intuitive: the circular ordering of the bloblet leaves and their hybrid nodes are easily seen in most cases (see, e.g., Figure 12). Following the NANUQ$^+$ framework (Allman et al. 2025a), with noisy data, a 'best fit' canonical form can be chosen by a more formal criterion.

We end by stating a few open problems related to the theory developed in this work. First, we conjecture that one of our main results (Theorem 4.7) can be extended to more general classes of networks. Specifically, supported by exploratory simulations, we conjecture that this theorem extends to all level-2 outer-labeled planar galled networks (not just bloblets within this class), as is the case for level-1 networks (Allman et al. 2019). Although such a result would not give new identifiability results from biological data (since the tree-of-blobs is already identifiable from quartets (Allman

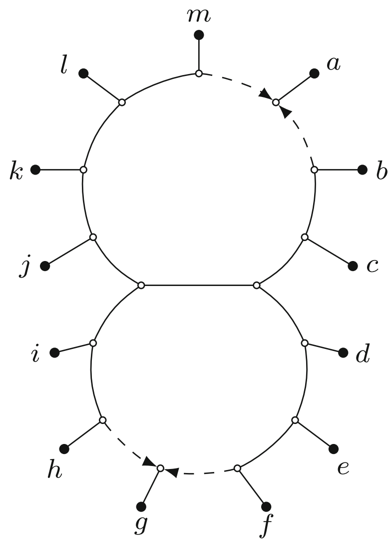

Fig. 12 *Left:* An outer-labeled planar, galled, level-2, semi-directed bloblet network N on leaf set $X = \{a, \ldots, m\}$. *Right:* The splits graph of the pairwise NANUQ distances d_N of the network N, obtained with NEIGHBOR- NET (Bryant and Moulton 2004)

et al. 2023)), it could be useful for inference. In particular, it would allow for obtaining the tree-of-blobs of a network in the class $\mathfrak{N}_2$ directly from the NANUQ distances, instead of relying on TINNIK (Allman et al. 2024a). In addition, we conjecture that the NANUQ metric of outer-labeled planar, galled, level-3, bloblet networks is also circular decomposable. Proving this will either require an even more extensive case analysis than in Lemma 4.6, or a different technique. A final related open problem is whether some of our results can be extended to a parametric family of distances, as introduced in Allman et al. (2025a). Such a result might be used to make an inference algorithm more robust, or for developing heuristics for choosing best-fit blob structures, as in the level-1 case.

Funding This paper is based upon work supported by the National Science Foundation (NSF) under grant DMS-1929284 while all authors were in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the semester program on "Theory, Methods, and Applications of Quantitative Phylogenomics".
NH and LvI were partially supported by the Dutch Research Council (NWO) grant OCENW.M.21.306.
ESA and JAR were partially supported by the National Science Foundation (NSF) grant DMS-205176, and HB by NSF grant DMS-2331660.

Data Availability No data are associated with this article.

Declarations

Competing interests HB serves as a Guest Editor to the Bulletin of Mathematical Biology for the topical collection associated with the ICERM semester program "Theory, Methods and Applications of Quantitative Phylogenomics", but was not involved in the review or editorial decision-making for this manuscript.

References

Allman ES, Baños H, Rhodes JA (2019) NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms Mol Biol 14(1):24. https://doi.org/10.1186/s13015-019-0159-2

Allman ES, Baños H, Rhodes JA (2022) Identifiability of species network topologies from genomic sequences using the logDet distance. J Math Biol 84(5):35. https://doi.org/10.1007/s00285-022-01734-2

Allman ES, Baños H, Mitchell JD, Rhodes JA (2023) The tree of blobs of a species network: identifiability under the coalescent. J Math Biol 86(1):10. https://doi.org/10.1007/s00285-022-01838-9

Allman ES, Baños H, Mitchell JD, Rhodes JA (2024) TINNiK: inference of the tree of blobs of a species network under the coalescent model. Algorithms Mol Biol 19(1):23. https://doi.org/10.1186/s13015-024-00266-2

Allman ES, Baños H, Garrote-Lopez M, Rhodes JA (2024) Identifiability of Level-1 Species Networks from Gene Tree Quartets. Bull Math Biol 86(9):110. https://doi.org/10.1007/s11538-024-01339-4

Allman ES, Baños H, Rhodes JA, Wicke K (2025) NANUQ$^+$: A divide-and-conquer approach to network estimation. Algorithms Mol Biol 20(1):14. https://doi.org/10.1186/s13015-025-00274-w

Allman ES, Ané C, Baños H, Rhodes JA (2025) Beyond level-1: Identifiability of a class of galled tree-child networks. arXiv:2504.21116

Ané C, Fogg J, Allman ES, Baños H, Rhodes JA (2024) Anomalous networks under the multispecies coalescent: theory and prevalence. J Math Biol 88(3):29. https://doi.org/10.1007/s00285-024-02050-7

Baños H (2019) Identifying species network features from gene tree quartets under the coalescent model. Bull Math Biol 81:494–534. https://doi.org/10.1007/s11538-018-0485-4

Bandelt H-J, Dress AW (1992) A canonical decomposition theory for metrics on a finite set. Adv Math 92(1):47–105. https://doi.org/10.1016/0001-8708(92)90061-o

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21(2):255–265. https://doi.org/10.1093/molbev/msh018

Bryant D, Moulton V, Spillner A (2007) Consistency of the neighbor-net algorithm. Algorithms Mol Biol 2:1–11. https://doi.org/10.1186/1748-7188-2-8

Buneman P (1971) "The Recovery of Trees from Measures of Dissimilarity". English. In: Mathematics the the Archeological and Historical Sciences. United Kingdom: Edinburgh University Press, pp.387–395. ISBN: 9780852242131

Chepoi V, Fichet B (1998) A note on circular decomposable metrics. Geometriae Dedicata 69:237–240. https://doi.org/10.1023/A:1004907919611

Christopher G, Farach M, Trick M (1996) "The structure of circular decomposable metrics". In: European Symposium on Algorithms. Springer. pp.486–500. https://doi.org/10.1007/3-540-61680-2_77

DeBaun D, Rabibisoa N, Raselimanana AP, Raxworthy CJ, Burbrink FT (2023) Widespread reticulate evolution in an adaptive radiation. Evolution 77(4):931–945. https://doi.org/10.1093/evolut/qpad011

Englander AK, Frohn M, Gross E, Holtgrefe N, van Iersel L, Jones M, Sullivant S (2025) "Identifiability of Phylogenetic Level-2 Networks under the Jukes-Cantor Model". In: bioRxiv, pp.2025–04. https://doi.org/10.1101/2025.04.18.649493

Gambette P, Berry V, Paul C (2012) Quartets and unrooted phylogenetic networks. J Bioinform Comput Biol 10(04):1250004. https://doi.org/10.1142/S0219720012500047

Gross E, van Iersel L, Janssen R, Jones M, Long C, Murakami Y (2021) Distinguishing level-1 phylogenetic networks on the basis of data generated by Markov processes. J Math Biol 83(3):32. https://doi.org/10.1007/s00285-021-01653-8

Holtgrefe N, Huber KT, van Iersel L, Jones M, Martin S, Moulton V (2025) Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments. Mol Biol Evol 42(4):msaf067. https://doi.org/10.1093/molbev/msaf067

Huson DH, Rupp R, Scornavacca C (2010) Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press, Cambridge
van Iersel L, Keijsper J, Kelk S, Stougie L, Hagen F, Boekhout T (2009) Constructing level-2 phylogenetic networks from triplets. IEEE/ACM Trans Comput Biol Bioinform 6(4):667–681. https://doi.org/10.1109/TCBB.2009.22
van Iersel L, Kole S, Moulton V, Nipius L (2022) An algorithm for reconstructing level-2 phylogenetic networks from trinets. Inf Process Lett 178:106300. https://doi.org/10.1016/j.ipl.2022.106300
Kalmanson K (1975) Edgeconvex circuits and the traveling salesman problem. Can J Math 27(5):1000–1010. https://doi.org/10.4153/cjm-1975-104-6
Kinene T, Wainaina J, Maina S, Boykin L (2016) “Rooting Trees, Methods for”. In: Encyclopedia of Evolutionary Biology. Ed. by R.M. Kliman. Oxford: Academic Press, pp.489–493. ISBN: 978-0-12-800426-5. https://doi.org/10.1016/B978-0-12-800049-6.00215-8
Kong S, Swofford DL, Kubatko LS (2024) Inference of Phylogenetic Networks From Sequence Data Using Composite Likelihood. Syst Biol 74(1):53–69. https://doi.org/10.1093/sysbio/syae054
Moulton V, Wu T (2022) Planar rooted phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform 20(2):1289–1297. https://doi.org/10.1109/TCBB.2022.3185670
Raghavan M et al (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505(7481):87–91. https://doi.org/10.1038/nature12736
Rhodes JA, Baños H, Xu J, Ané C (2025) Identifying circular orders for blobs in phylogenetic networks. Adv Appl Math 163:102804. https://doi.org/10.1016/j.aam.2024.102804
Rhodes JA (2019) Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Trans Comput Biol Bioinform 17(6):2107–2118. https://doi.org/10.1109/tcbb.2019.2917204
Solís-Lemus C, Ané C (2016) Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLOS Genet 12(3):1–21. https://doi.org/10.1371/journal.pgen.1005896
Swithers KS, Soucy SM, Gogarten JP (2012) The Role of Reticulate Evolution in Creating Innovation and Complexity. Int J Evol Biol 2012:1–10. https://doi.org/10.1155/2012/418964
Steel M (2016) Phylogeny. Philadelphia, PA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974485
Sessa EB, Zimmer EA, Givnish TJ (2012) Reticulate evolution on a global scale: A nuclear phylogeny for New World Dryopteris (Dryopteridaceae). Mol Phylogenetics Evol 64(3):563–581. https://doi.org/10.1016/j.ympev.2012.05.009
Wen D, Nakhleh L (2018) Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst Biol 67(3):439–457. https://doi.org/10.1093/sysbio/syx085
Xu J, Ané C (2022) Identifiability of local and global features of phylogenetic networks from average distances. J Math Biol 86(1):12. https://doi.org/10.1007/s00285-022-01847-8
Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genom 16:1–10. https://doi.org/10.1186/1471-2164-16-S10-S10
Zhang C, Ogilvie HA, Drummond AJ, Stadler T (2018) Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol 35(2):504–517. https://doi.org/10.1093/molbev/msx307

Authors and Affiliations

Niels Holtgrefe[1] · Elizabeth S. Allman[2] · Hector Baños[3] · Leo van Iersel[1] · Vincent Moulton[4] · John A. Rhodes[2] · Kristina Wicke[5]

✉ Niels Holtgrefe
n.a.l.holtgrefe@tudelft.nl

Elizabeth S. Allman
e.allman@alaska.edu

Hector Baños
hector.banos@csusb.edu

Leo van Iersel
l.j.j.vaniersel@tudelft.nl

Vincent Moulton
v.moulton@uea.ac.uk

John A. Rhodes
j.rhodes@alaska.edu

Kristina Wicke
kristina.wicke@njit.edu

[1] Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

[2] Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK, USA

[3] Department of Mathematics, California State University San Bernardino, San Bernardino, CA, USA

[4] School of Computing Sciences, University of East Anglia, Norwich, UK

[5] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA