TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Frohn, M., Holtgrefe, N., van Iersel, L., Jones, M., & Kelk, S. (2026). Bounds on the Sequence Length Sufficient to Reconstruct Binary Level-1 Phylogenetic Networks Under the CFN Model. *Annals of Combinatorics*. https://doi.org/10.1007/s00026-026-00830-0

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Ann. Comb.
© 2026 The Author(s)
https://doi.org/10.1007/s00026-026-00830-0

Annals of Combinatorics

Bounds on the Sequence Length Sufficient to Reconstruct Binary Level-1 Phylogenetic Networks Under the CFN Model

Martin Frohn, Niels Holtgrefe, Leo van Iersel, Mark Jones and Steven Kelk

Abstract. Phylogenetic trees and networks are graphs used to model evolutionary relationships, with trees representing strictly branching histories and networks allowing for events in which lineages merge, called *reticulation* events. While the question of data sufficiency has been studied extensively in the context of trees, it remains largely unexplored for networks. In this work, we take a first step in this direction by establishing bounds on the amount of genomic data required to reconstruct binary level-1 semi-directed phylogenetic networks, which are binary networks in which reticulation events are indicated by directed edges, all other edges are undirected, and cycles are vertex disjoint. For this class, methods have been developed recently that are statistically consistent. Roughly speaking, such methods are guaranteed to reconstruct the correct network assuming infinitely long genomic sequences. Here we consider the question whether networks from this class can be uniquely and correctly reconstructed from finite sequences. Specifically, we present an inference algorithm that takes as input genetic sequence data, and demonstrate that the sequence length sufficient to reconstruct the correct network with high probability, under the CFN model of evolution, scales logarithmically, polynomially, or polylogarithmically with the number of taxa, depending on the parameter regime. As part of our contribution, we also present novel inference rules for quartet data in the semi-directed phylogenetic network setting.

Keywords. Phylogenetic network, Quartets, Inference rules, Distance method, Sequence alignment, Sequence length.

Published online: 21 May 2026

1. Introduction

Phylogenetic networks are graphs that model the evolutionary history of a set of taxa, for example different species, which are represented by the leaves of the network. Evolutionary events in which lineages merge, such as hybridization or recombination, are represented by so-called *reticulation vertices* that have more than one parent. Rooted phylogenetic networks are directed acyclic graphs with a single root, see Fig. 1a. However, since the root location is not identifiable under most models [7,17], we consider semi-directed phylogenetic networks here, in which the incoming edges of reticulation vertices are directed, all other edges are undirected and the root is suppressed, see Fig. 1b. For formal definitions, see Sect. 2.

Recently, several methods have been developed for reconstructing semi-directed level-1 phylogenetic networks[1], which are networks in which cycles of the underlying undirected graph are vertex disjoint, from biological data consisting of gene trees or aligned DNA sequences [3,4,21,28,38]. Such networks are relevant in practice because they allow for both speciation (divergence) and reticulation (convergence) events to explain the evolutionary relationships between taxa while assuming that reticulation events are (in some sense) isolated [8,27]. However, although some of the methods mentioned above are statistically consistent, i.e., guaranteed to reconstruct the correct network given perfect data, these papers do not discuss how much data is needed. In particular, statistically consistent methods for reconstructing networks from DNA sequences [21,30] may only be successful when provided sequences that are infinitely long.

For phylogenetic trees, data sufficiency (also called *sample complexity*) has been studied for various substitution models, starting with the seminal work in Refs. [14,15], its improvements [26,31,32], variations [11,12], and generalizations [40,43]. Recently, lower bounds on the sample complexity have also been established under a multi-species coalescent model [20]. Typically, such reconstruction methods are *fast converging* [23], i.e., given lower and upper bounds on edge lengths and a DNA sequence alignment of polynomial size (in the number of taxa), the method recovers the true tree topology with high probability. A potential drawback of fast converging methods is that they may rely on known edge length bounds and absence of such bounds might even lead to statistical inconsistencies [11,12,23]. Methods that are fast converging without requiring any knowledge of the bounds on edge lengths are called *absolute fast converging* [40].

In a recent paper, a first result on data sufficiency for phylogenetic network inference was presented [41] focusing on reconstructing binary semi-directed level-1 phylogenetic networks from SNPs, basically a binary sequence alignment, in which each site (column) is assumed to have evolved down a tree displayed by the network in such a way that its state changed exactly once.

[1] In the directed setting, level-1 networks were originally called 'galled trees' [18].

TABLE 1. The growth rate of the number of sites sufficient to correctly reconstruct a binary semi-directed level-1 network N under the CFN model with high probability as a function of the number of taxa n (see Theorem 5.5 for a precise statement)

$\max\{\text{depth}(N), lc(N)\}$	Range of mutation probabilities $p(e)$, $e \in E(N)$	
	$[c_1, c_2]$ for constants $c_1, c_2 \geq 0$	$\left[\frac{1}{\log n}, \frac{\log\log n}{\log n}\right]$
$\Theta(1)$	Logarithmic	Polylog
$\Theta(\log\log n)$	Polylog	Polylog
$\Theta(\log n)$	Polynomial	Polylog

This growth rate depends on the maximum of the network depth and the maximum cycle length $lc(N)$ (first column, see Sect. 2 for definitions) and on the bounds on the mutation probabilities. In the second column, these bounds are constants, while in the third column, they are functions of n. Inheritance probabilities are bounded by a function of the mutation probabilities (see Proposition 5.4)

The latter assumption, also called *homoplasy-free* or *infinite sites*, is a severely restricting condition on the data that we do not assume in this work.

In this article, we introduce a new method to reconstruct binary semi-directed level-1 phylogenetic networks from biological data under the Cavender–Farris–Neyman (CFN) model [34]: the two-state analog of the well-known Jukes–Cantor model. We assume a substitution process along each network edge, without incomplete lineage sorting, where at reticulation nodes, each site is independently inherited from one parent according to inheritance probabilities. This model also accommodates homoplastic sites, allowing for convergent or parallel mutations. Our method reconstructs binary semi-directed level-1 networks up to the suppression of 2-cycles, 3-cycles, and reticulation vertices in 4-cycles, as do other methods [3,4,41], since this is best possible for displayed-quartet-based methods [7].

Our main result is that the introduced method reconstructs the correct network with high probability, under the CFN model of evolution, for input sequences whose required length scales logarithmically, polynomially, or polylogarithmically with the number of taxa, depending on the parameter regime, see Table 1.

A central concept in our approach is the notion of a *quartet profile*: the set of quartet trees displayed by a network for a given set of four taxa, see Fig. 1. Such a quartet profile may either contain one quartet tree (representing tree-like evolution), or multiple quartet trees (representing a cycle, and thus network-like evolution). Our method differentiates between the two by applying the four-point method [35]—a key characteristic in all absolute fast

converging methods for tree inference [40]—to estimated pairwise distances. By precisely characterizing a strategically chosen subset of these quartet profiles that capture the structure of the network, we can significantly reduce the number of quartet profiles that must be inferred directly from the biological data. Then we apply newly developed inference rules to deduce the remaining set of quartet profiles, which enables us to leverage an existing reconstruction algorithm [16] to reconstruct the full network. Note also that quartet profiles have been shown to be identifiable under several models, with proofs either relying on algebraic invariants [13] or on concordance factors (a type of summary statistic) [36].

The inference rules we develop are also interesting in their own right, as they extend and generalize classical results from tree-based phylogenetics. While inference rules for induced quartet trees of phylogenetic trees have been extensively studied, tracing back to Ref. [10] (see also Ref. [37] and references therein), their use in a network setting has been very limited so far. The TiNNiK software tool [1,2] for inferring the *tree-of-blobs*—a tree capturing the global branching structure of a semi-directed network—relies on a rule to propagate quartet classifications by marking additional quartets as belonging to reticulate regions. However, this rule is not an inference rule in the sense used here. In addition, Ref. [22] introduced inference rules for *quarnets* (4-leaf subnetworks) of undirected level-1 networks (see also Ref. [25]), but these do not directly translate to our semi-directed setting. In our work, we define and study the closure of a set of quartet profiles, building on well-established ideas from the tree literature. This also leads to a novel quarnet encoding result, showing that any semi-directed level-1 network is encoded by some linear-sized set of its induced quarnets (see Corollary 3.10). Although our focus is on level-1 networks, the validity of most of our inference rules is shown for networks of any level, opening up broader avenues for future research.

The structure of the paper is as follows. In Sect. 2, we cover preliminaries on phylogenetic networks and the CFN model. In Sect. 3, we introduce novel quartet profile inference rules and use these to define the closure operation. These techniques are used in Sect. 4, where we present our main distance-based inference algorithm, which is analyzed under the CFN model in Sect. 5. We end with a discussion in Sect. 6.

2. Preliminaries

2.1. Phylogenetic Networks

For the sake of generality, we allow phylogenetic networks to contain parallel edges, which can result from a variety of biological processes [24]; see also, e.g., Refs. [1,7,36]. Formally, we call a graph that may contain parallel edges a *multi-graph*; it is *directed* if all edges are directed, and *mixed* if it may contain both directed and undirected edges.

Definition 2.1. (*Rooted network*) A *(binary) rooted (phylogenetic) network* N^+ on a set of at least two taxa $\mathcal{X}$ is a directed acyclic multi-graph such that (i)

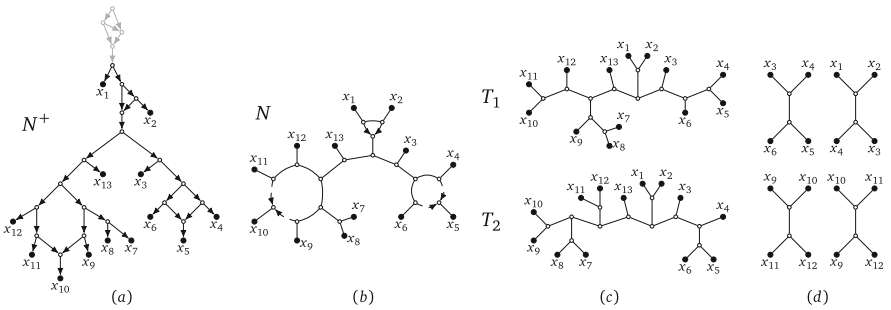

FIGURE 1. **a** A rooted phylogenetic network N^+ on $\mathcal{X} = \{x_1, \ldots, x_{13}\}$. The gray vertices are above the LSA of the network. Hence, removing the gray vertices and edges from N^+ gives an LSA network. **b** The semi-directed phylogenetic level-1 network N on $\mathcal{X}$ obtained from this LSA network. **c** Two unrooted phylogenetic trees T_1 and T_2 on $\mathcal{X}$ displayed by N. **d** Four quartets displayed by N. The quartet $x_3x_4|x_5x_6$ does not form a quartet profile of N, since N also displays another quartet ($x_3x_6|x_4x_5$, not depicted in the figure) on the same four leaves. The quartet $x_1x_2|x_3x_4$ forms a trivial quartet profile, and the two remaining quartets together form the quartet profile $(x_9x_{10}x_{11}x_{12})$

there is a unique vertex with indegree zero (the *root*), which has out-degree two; (ii) each vertex with out-degree zero (a *leaf*) has in-degree one and the set of such vertices is bijectively labeled by $\mathcal{X}$; (iii) all other vertices either have in-degree one and out-degree two (*tree vertices*), or in-degree two and out-degree one (*reticulation vertices*).

The *lowest stable ancestor (LSA)* of a rooted network on $\mathcal{X}$ is the lowest vertex through which all directed paths from the root to any leaf in $\mathcal{X}$ pass. Throughout this work, we assume that all rooted networks on $\mathcal{X}$ are *LSA networks*: networks in which the root is the LSA (see Fig. 1a). This is a common assumption, since within standard modeling frameworks, the structure above the LSA cannot be resolved from quartet data [7]. The two edges directed towards a reticulation vertex are called *reticulation edges*.

Definition 2.2. (*Semi-directed network*) A *(binary) semi-directed (phylogenetic) network* N on at least two taxa $\mathcal{X}$ is a mixed multi-graph that can be obtained from a binary rooted phylogenetic LSA network N^+ on $\mathcal{X}$ by undirecting all non-reticulation edges and suppressing the former root.

The reticulation edges of a semi-directed network are precisely its directed edges, and the reticulation vertices are the vertices with in-degree two. A leaf x is *below* a reticulation vertex r if there is a path from r to x consisting of no edges oriented towards r. A *cut-edge* is an edge whose removal separates the graph into two connected components and the cut-edge is *nontrivial* if the

two components contain at least two vertices each. A *blob* of any multi-graph (including directed and mixed multi-graphs) is a maximal subgraph without any cut-edges. A blob is an *m-blob* if it has m cut-edges incident to it and it is *internal* if it does not contain a leaf. A (rooted or semi-directed) network is *level-k*, for some integer $k \geq 0$, if there are at most k reticulation vertices in each blob of the network. In this work, we mostly focus on level-1 networks, which consist of only tree-like parts and vertex disjoint cycles since we assume our networks to be binary. See Fig. 1b for an example of a semi-directed level-1 network. We denote the maximum topological length of a cycle in N by $lc(N)$. A semi-directed level-1 network with a single internal blob and which contains exactly one reticulation vertex is called a *sunlet (network)*.

A level-0 network is commonly known as a phylogenetic tree. We distinguish between a *rooted phylogenetic tree* (a level-0 rooted network) and an *unrooted phylogenetic tree* (a level-0 semi-directed network). An unrooted phylogenetic tree on at least two leaves $\mathcal{X}$ is a *displayed tree* of a semi-directed network N on $\mathcal{X}$ if the tree can be obtained from the network by deleting one reticulation edge per reticulation vertex, undirecting the remaining reticulation edges, exhaustively deleting leaves not in $\mathcal{X}$, and suppressing degree-2 vertices. See Fig. 1c for an example.

In Definition 2.3, we formalize the notion of a *subnetwork* (or *subtree* for trees). To this end, we let an *up-down path* between two leaves x_1 and x_2 of a semi-directed network be a path of k edges where the first ℓ edges are undirected or directed towards x_1 and the last $k - \ell$ edges are undirected or directed towards x_2.

Definition 2.3. (*Subnetwork*) Given a semi-directed network N on $\mathcal{X}$ and some $\mathcal{Y} \subseteq \mathcal{X}$ with $|\mathcal{Y}| \geq 2$, the *subnetwork* of N induced by $\mathcal{Y}$ is the semi-directed network $N|_{\mathcal{Y}}$ obtained from N by taking the union of all up-down paths between leaves in $\mathcal{Y}$, followed by exhaustively suppressing degree-2 vertices.

Given a semi-directed network N on $\mathcal{X}$, a 4-leaf subtree of a displayed tree of N is called a *displayed quartet*.[2] A displayed quartet on $\{a, b, c, d\}$ has four trivial cut-edges: one cutting off each leaf. Next to that, the quartet has exactly one nontrivial cut-edge inducing a nontrivial partition of its leaves ($ab|cd$, $ac|bd$, or $ad|bc$). In general, given a semi-directed network N on $\mathcal{X}$ and a partition $A|B$ of $\mathcal{X}$ (with A and B both non-empty), we say that $A|B$ is a *split* of N if there exists a cut-edge of N whose removal disconnects the leaves in A from those in B. The split and the cut-edge are *nontrivial* if the corresponding partition is nontrivial, that is, if $|A|, |B| \geq 2$. As already used above, for splits with few leaves, we sometimes omit the set notation, e.g., we use $ab|cd$ instead of $\{a, b\}|\{c, d\}$.

The main data type used in this work is the following. See Fig. 1d for an example.

[2] In Proposition A.1 of Appendix A, we prove that taking displayed trees and inducing subnetworks are commutative. Hence, equivalently, a displayed quartet can also be defined as a displayed tree of a 4-leaf subnetwork.

Definition 2.4. (*Quartet profile*) A *quartet profile* on four leaves $\{a, b, c, d\} \subseteq \mathcal{X}$ is a set of quartet trees on $\{a, b, c, d\}$. Given a semi-directed network N on $\mathcal{X}$, the quartet profile of N on $\{a, b, c, d\} \subseteq \mathcal{X}$ is the set of displayed quartets of N on $\{a, b, c, d\}$.

If the quartet profile on $\{a, b, c, d\}$ contains a single quartet tree, we denote it by the split induced by this tree, say $ab|cd$, and call it a *trivial* quartet profile. Note that this implies that the quartet $ab|cd$ is an induced 4-leaf subtree of every displayed tree of the network. If instead the quartet profile contains two quartet trees, we denote the quartet profile by the circular ordering that is congruent with the nontrivial splits of the two trees, i.e., in which each split corresponds to opposite pairs in the ordering. For instance, if the quartet profile on $\{a, b, c, d\}$ contains exactly the trees with splits $ab|cd$ and $ad|bc$, then the quartet profile will be denoted by $(abcd)$ (or equivalently, e.g., $(bcda)$ or $(dcba)$). See also Fig. 1d, where the quartet trees $x_9x_{10}|x_{11}x_{12}$ and $x_9x_{12}|x_{10}x_{11}$ form a quartet profile and both are congruent with the circular ordering $(x_9x_{10}x_{11}x_{12})$, which also matches the ordering of the leaves around the left-most cycle of the network N in Fig. 1b. The leaf set of a quartet profile q is denoted by $\mathcal{L}(q)$ and the set of all quartet profiles of a network N is denoted by $\mathcal{Q}(N)$. In general, a *set of quartet profiles* $\mathcal{Q}$ *on* $\mathcal{X}$ is a set such that $\mathcal{X} = \bigcup_{q \in \mathcal{Q}} \mathcal{L}(q)$ and there is at most one quartet profile per 4-leaf subset of $\mathcal{X}$.

It follows from Ref. [36] that the quartet profiles (referred to there as '4-taxon circular order information') of any *outer-labeled planar* network (which include the level-1 networks we consider [33]) never contain all three possible quartet trees. We emphasize that knowing whether a quartet profile is displayed by a network amounts to knowing *all* displayed quartets for the considered set of four leaves. Hence, a network has only trivial quartet profiles if and only if all its displayed trees induce the same set of quartets. Since quartets uniquely encode a tree [37], this is equivalent to the network having a unique displayed tree.

We end with a useful result specifying exactly which features of a network are uniquely determined and can be reconstructed from quartet profiles. Throughout this manuscript, we will use the notation $\tilde{N}$ introduced in this proposition, which represents the main object we aim to construct. For convenience, we may occasionally refer to $\tilde{N}$ as a "network", although it is not one in the strict sense of the term. By *suppressing* a k-cycle in a semi-directed network, we mean replacing the cycle by a single vertex and suppressing resulting degree-2 vertices. For convenience, we refer to the operation of undirecting reticulation edges as *suppressing reticulation vertices*.

Proposition 2.5. [7,16] *Let N be a semi-directed level-1 network on n leaves and denote by $\tilde{N}$ the mixed graph obtained from N by suppressing 2-cycles, 3-cycles, and reticulation vertices in 4-cycles. Then $\tilde{N}$ is uniquely characterized by $\mathcal{Q}(N)$ and can be reconstructed from $\mathcal{Q}(N)$ in $\mathcal{O}(n^2)$ time.*

2.2. The CFN Model and Distances in Phylogenetic Networks

Given a rooted phylogenetic network $N = (V, E)$, for each $v \in V$, we associate a random variable $Y(v)$ with a binary state space, corresponding to purine or pyrimidine bases in DNA sequences. For each edge $e = (u, v) \in E$ that is not a reticulation edge, we define $p(e)$ as the probability that the states of $Y(u)$ and $Y(v)$ differ and call it the *mutation probability* of e. We make the standard exclusion $p(e) \neq 0$ and $p(e) \neq 1/2$ because such a probability would mean that either there is no change on e or variables $Y(v)$ and $Y(u)$ are random with respect to each other [14]. Moreover, biologically, nucleotide states are more likely to remain unchanged than to mutate. Therefore, $p(e) \in (0, 1/2)$. Observe that probabilities $p(e)$ yield a Markov model on the displayed trees of N, called the CFN model [34], which is an instance of the general Markov model for transition matrices

$$M^e = \begin{pmatrix} 1 - p(e) & p(e) \\ p(e) & 1 - p(e) \end{pmatrix}$$

and a uniform root distribution. For reticulation edges $e = (u, v) \in E$, we also define the *inheritance probability* $\gamma(e) \in (0, 1)$ as the proportion of sites in the alignment of v contributed by u. The following result is well-known in the literature (e.g., page 13 in Ref. [19]).

Lemma 2.6. *For the path in a phylogenetic tree constituted by edges $e_1, \ldots, e_m$ and with endpoints v_1 and v_2, we have*

$$\mathbb{P}[Y(v_1) \neq Y(v_2)] = \frac{1}{2}\left(1 - \prod_{i=1}^{m}(1 - 2p(e_i))\right)$$

under the CFN model.

Observe that Lemma 2.6 can be applied to every displayed tree T of N. Hence, we can encode the magnitude of the mutation probabilities in the CFN model as edge lengths of N as follows: for each displayed tree T of N, we define a $n \times n$ matrix $D(T)$ of pairwise distances between leaves $\mathcal{X} = \{x_1, \ldots, x_n\}$ such that $D(T)$ are additive distances on T. Specifically, for the unique path in T constituted by edges $e_1, \ldots, e_m$ and with endpoints $x_i, x_j \in \mathcal{X}$, we define

$$D(T)_{ij} = -\frac{1}{2}\log\left(1 - 2 \cdot \mathbb{P}[Y(x_i) \neq Y(x_j)]\right) = -\sum_{l=1}^{m}\frac{1}{2}\log\left(1 - 2p(e_l)\right)$$

to adjust the distance matrix to our choice of mutation probabilities under the CFN model. This means, $D(T)_{ij}$ is the sum of positive edge lengths $\ell(e_j)$, $j \in \{1, \ldots, m\}$, defined in general by

$$\ell(e) = -\frac{1}{2}\log\theta(e) \qquad \forall\, e \in E,\ \theta(e) = 1 - 2p(e).$$

In other words, $D(T)$ defines additive distances on T. Notice that this definition of $D(T)$ coincides with the paralinear distance [29] and the LogDet distance [39] for T under the CFN model.

Now, let $\mathcal{S}$ be a sequence alignment on $\mathcal{X}$ of length k generated under the CFN model. Hence, $\mathcal{S}$ is a collection of binary vectors $s \in \{0, 1\}^k$. We

let $d_H(i,j)$ denote the *Hamming distance* between the sequences $s^i, s^j \in \mathcal{S}$ to define the *dissimilarity score* of sequences s^i and s^j as $h^{ij} = d_H(i,j)/k$. Then the expectation $\mathbb{E}[h^{ij}]$ can be seen as the probability $\mathbb{P}[Y(x_i) \neq Y(x_j)]$ between sequences s^i and s^j. Therefore, we define the *empirical distance* of s^i and s^j by

$$d_{ij} = -\frac{1}{2}\log\left(1 - 2h^{ij}\right)$$

to ensure that d_{ij} converges to $D(T)_{ij}$ in probability as k grows for any fixed displayed tree T of N [39]. This means, for fixed inheritance probabilities $\gamma(e)$, any reticulation vertex v in a rooted binary phylogenetic network N and displayed trees T_1 and T_2 of N which differ only in distinct reticulation edges e_1 and e_2 incident to v, respectively, we know that d_{ij} converges to $\gamma(e_1)D(T_1)_{ij} + \gamma(e_2)D(T_2)_{ij}$ in probability as k grows.

Using the empirical distances as input, Ref. [14] provided an absolute fast converging method to reconstruct the tree T with probability $1 - o(1)$ if

$$k > \frac{c\log n}{\left(1 - \max\limits_{e\in E}\theta(e)\right)^2 \left(\min\limits_{e\in E}\theta(e)\right)^{\mathrm{depth}(T)}} \tag{1}$$

for a constant c and

$$\mathrm{depth}(T) = \max_{e=(v_1,v_2)\in E(T)} \max_{i\in\{1,2\}} \{\text{shortest topological distance from } v_i \text{ to a leaf after removing } e\}.$$

As a shorthand notation, we denote $\mathrm{depth}(N)$ as the largest chosen number $\mathrm{depth}(T)$ among all displayed trees T of N. In other words, $\mathrm{depth}(N)$ is an upper bound across all internal vertices v and displayed trees T of N on the shortest topological distance between v and a leaf of the subtree of T induced by v. Table 2 shows different values for the bound (1) on the number of sites k. Notice that $\mathrm{depth}(T) = \Theta(1)$ and $\mathrm{depth}(T) = \Theta(\log n)$ occur for the completely unbalanced and completely balanced binary tree T, respectively. Additionally, $\mathrm{depth}(T) = \mathcal{O}(\log\log n)$ typically occurs for Yule trees [14]. Furthermore, if values $\theta(e)$ are restricted to be either large or small, then the bound on k becomes logarithmic or polynomial, respectively [32].

TABLE 2. Asymptotic behavior of the righthand side of bound (1) for different growth rates for $\mathrm{depth}(T)$ and different ranges for mutation probabilities $p(e)$, $e \in E(T)$, with n the number of taxa

Depth(T)	Range of mutation probabilities $p(e)$, $e \in E(T)$	
	$[c_1, c_2]$ for constants $c_1, c_2 \geq 0$	$\left[\frac{1}{\log n}, \frac{\log\log n}{\log n}\right]$
$\Theta(1)$	Logarithmic	Polylog
$\Theta(\log\log n)$	Polylog	Polylog
$\Theta(\log n)$	Polynomial	Polylog

We call the length of a shortest path between two vertices v_i, v_j in a tree T the *graphical distance* between v_i and v_j and denote it by D_{ij}. The matrix of pairwise graphical distances between vertices in T is denoted by D. Here, the length of a path is either the number of edges of a path if T is unweighted or the sum of the edge lengths of the path if T is weighted. Then we will make use of the following characterization of additive distances [9]:

Lemma 2.7. (Four-point condition) *A graph is a tree if and only if it is connected, 3-cycle-free and has graphical distance D satisfying the four-point condition:*

$$D_{ij} + D_{kl} \leq \max\{D_{ik} + D_{jl}, D_{il} + D_{jk}\}$$

for all vertices v_i, v_j, v_k, v_l in the graph. Equivalently, the four-point condition requires that exactly one of the following properties holds:

$$\begin{aligned} D_{ij} + D_{kl} &< D_{ik} + D_{jl} = D_{il} + D_{jk}, \\ D_{ik} + D_{jl} &< D_{ij} + D_{kl} = D_{il} + D_{jk}, \\ D_{il} + D_{jk} &< D_{ik} + D_{jl} = D_{ij} + D_{kl} \end{aligned}$$

for all vertices v_i, v_j, v_k, v_l in the graph.

3. Inference Rules and Dyadic Closure

In this section, we develop inference rules for quartet profiles and utilize them to characterize a small set of quartet profiles that suffice to reconstruct the whole network. Although our primary focus is on level-1 networks, we prove that several of the inference rules hold more generally for networks of any level, which could open further directions for research.

3.1. Inference Rules

Inference rules for quartets of unrooted phylogenetic trees are rules of the type $ab|cd + ab|ce \mapsto ab|de$, meaning that if $ab|cd$ and $ab|ce$ are quartets induced by a tree, this implies that $ab|de$ is a quartet induced by the same tree. Such inference rules have been extensively studied for phylogenetic trees [e.g., 6,14], since they allow to extract more quartets from only limited quartet information available. More recently, inference rules for *quarnets* (4-leaf subnetworks) of fully undirected networks have also been studied [22]. The method presented in this paper partially relies on new inference rules for quartet profiles. We note that since we study semi-directed networks and hence the subnetwork induced by fewer leaves might force the loss of information on reticulation events, these rules are inherently different from the undirected quarnet inference rules presented in Ref. [22].

Formally, an *inference rule* for quartet profiles is a rule of the type

$$q_1 + \cdots + q_k \mapsto q_1' + \cdots + q_l'$$

that can be applied to a set of distinct quartet profiles $\mathcal{Q} = \{q_1, \ldots, q_k\}$ and yields a set of new quartet profiles $\mathcal{Q}' = \{q_1, \ldots, q_l\}$. Recall that each q_i forms a set of quartet trees: if $|q_i| = 1$, it is identified with the split of its unique

quartet tree, e.g., $ab|cd$, and if $|q_i| = 2$, it is represented using circular ordering notation, e.g. $(abcd)$. We say that an inference rule is *valid* if the following implication holds for all semi-directed networks N (possibly restricted to a certain subclass of semi-directed networks)

$$\mathcal{Q} \subseteq \mathcal{Q}(N) \;\Rightarrow\; \mathcal{Q}' \subseteq \mathcal{Q}(N).$$

The *order* of an inference rule is its input size k. A second order inference rule is called *dyadic*, while a third order rule is *triadic*.

We first introduce two well-known dyadic inference rules for quartets of unrooted phylogenetic trees (see e.g. Refs. [6,14]):

$$ab|cd + ac|de \mapsto ab|de + bc|de; \tag{R1}$$

$$ab|cd + ab|ce \mapsto ab|de. \tag{R2}$$

In the following proposition, we prove that any rule that is valid for quartets of an unrooted phylogenetic tree—which thus includes inference rules R1 and R2—is valid for semi-directed networks, when considering (necessarily trivial) quartet profiles. We note in passing that this also means that these are valid inference rules for quarnet-splits (as defined in Ref. [16]).

Proposition 3.1. *Any valid quartet inference rule for unrooted phylogenetic trees is a valid quartet profile inference rule for semi-directed networks.*

Proof. Denote by R a quartet inference rule $q_1 + \cdots + q_k \mapsto q'_1 + \cdots + q'_l$ (with $k, l \geq 1$) that is valid for any unrooted phylogenetic tree. Let N be a semi-directed network and assume that all q_i are also trivial quartet profiles of N (i.e., there is only one displayed quartet tree for each of the corresponding four-leaf subsets). Next, let T be a displayed tree of N on the same leaf set. Then since all quartet profiles q_i are trivial, all of them are also induced quartets of the tree T. Hence, by the valid rule R, all the q'_j are induced quartets of T. Since T was an arbitrary displayed tree of N on the same leaf set, all q'_j are quartets of any of the displayed trees of N, and hence they are trivial quartet profiles of N. □

Building upon the two inference rules R1 and R2, we introduce two novel inference rules (the first one being dyadic, and the second one triadic) that also consider nontrivial quartet profiles:

$$(abcd) + ac|de \mapsto ab|de + bc|de; \tag{R3}$$

$$(abcd) + ab|de + bc|de \mapsto (abce). \tag{R4*}$$

In the following lemma, we show that these inference rules are valid for any semi-directed network.

Lemma 3.2. *The quartet profile inference rules R3 and R4* are valid for any semi-directed network.*

Proof. We first consider rule R3 and assume that $(abcd)$ and $ac|de$ are quartet profiles of N. Then we can partition the displayed trees of N into the set $\mathcal{T}_1$ with quartets $ab|cd$, and the set $\mathcal{T}_2$ with quartets $ad|bc$. Since $ac|de$ is a trivial quartet profile of N, it is a quartet of all of these displayed trees. We can now

apply the valid quartet inference rule R1 to the quartets in $\mathcal{T}_1$, and obtain that $ab|de$ and $bc|de$ are quartets of those trees. After switching the roles of a and c, we can rewrite R1 to $bc|ad + ac|de \mapsto ab|de + bc|de$, and we obtain that $ab|de$ and $bc|de$ are also quartets of the trees in $\mathcal{T}_2$. Since this covers all displayed trees, these two quartets are (trivial) quartet profiles of N.

For rule R4*, suppose that $(abcd)$, $ab|de$, and $bc|de$ are quartet profiles of N. We again have the two sets of displayed trees, $\mathcal{T}_1$ with quartets $ab|cd$, and $\mathcal{T}_2$ with quartets $ad|bc$. Furthermore, in any of these displayed trees, $ab|de$ and $bc|de$ are quartets. Now, note that we can rewrite rule R2 to $ab|cd + ab|de \mapsto ab|ce$ by switching the roles of c and d. Applying this rule to the trees in $\mathcal{T}_1$ gives us that $ab|ce$ are quartets in those trees. Similarly, we can also rewrite rule R2 to $bc|da + bc|de \mapsto bc|ae$. Applying this rule to the trees in $\mathcal{T}_2$ gives us that $bc|ae$ are quartets in those trees. Since we have considered all displayed trees, this means that $(abce)$ is a quartet profile of N. □

In case the semi-directed network N is level-1, inference rule R4* can be simplified to a dyadic inference rule:

$$(abcd) + bc|de \mapsto (abce). \tag{R4}$$

This is useful, since exhaustively applying dyadic inference rules to a set of quartets requires less time than applying higher order inference rules. The following lemma shows validity.

Lemma 3.3. *The quartet profile inference rule R4 is valid for any semi-directed level-1 network.*

Proof. Suppose that $(abcd)$ and $bc|de$ are quartet profiles of a semi-directed level-1 network N. It suffices to consider the case where N has exactly five leaves, otherwise we can consider the subnetwork induced by $\{a, b, c, d, e\}$. Moreover, we can disregard 2- and 3-cycles, since they do not change the displayed quartets. We now consider two cases, depending on whether N is a sunlet or not.

Case 1: *N is a sunlet.* Since $bc|de$ is a trivial quartet profile, the subnetwork of N induced by $\{b, c, d, e\}$ also induces the split $bc|de$. Now note that a 4-leaf subnetwork of a sunlet network induces a nontrivial split if its leaf set excludes the leaf below the reticulation, and it does not induce a nontrivial split otherwise. Hence, a must be the leaf below the reticulation of the sunlet. Furthermore, the quartet profile $(abcd)$ indicates that the leaves of N are either ordered as (a, b, c, d, e) or (a, b, c, e, d) around the sunlet (see Fig. 2a). In both cases, $(abce)$ is a quartet profile of N.

Case 2: *N is not a sunlet.* The existence of the quartet profile $(abcd)$ shows that N *does* have a 4-cycle. Consequently, N consists of one four-cycle with an incident nontrivial cut-edge. Since $bc|de$ is a quartet profile, this cut-edge must induce the split $abc|de$. Hence, N is one of the four networks in Fig. 2b, and it follows that $(abce)$ is a quartet profile. □

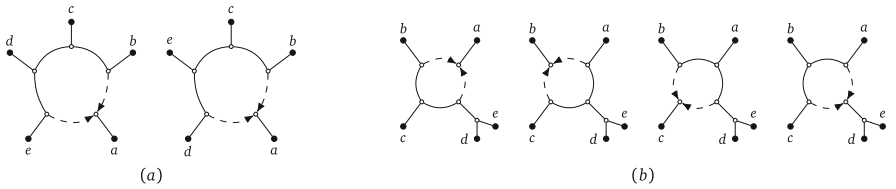

FIGURE 2. All semi-directed level-1 networks on $\{a, b, c, d, e\}$ with quartet profiles $(abcd)$ and $bc|de$ (disregarding possible 2- and 3-cycles). The networks in subfigure **a** are sunlets that have a 5-cycle and the networks in subfigure **b** are not sunlets, having a 4-cycle and a node forming its own internal blob

3.2. Representative Quartet Profiles and Dyadic Closure

In this section, we present a small set of carefully selected quartet profiles that uniquely determine a semi-directed level-1 network. To do so, we first introduce some additional technical definitions.

Let N be a semi-directed level-1 network on $\mathcal{X}$ and B an internal k-blob. Since N is level-1, such blobs are either cycles or single degree-3 vertices. Thus, they naturally induce a *circular order* $\mathcal{C} = (X_1, \ldots, X_k)$ of subsets of $\mathcal{X}$ (with $X_1| \ldots |X_k$ a k-partition of $\mathcal{X}$), defined up to reversal and cyclic permutations. See Fig. 3a for an example. For each of the subsets X_i, we call the unique leaf $x_i \in X_i$ that is topologically the closest to B (in terms of the number of edges) its *representative leaf*. To ensure uniqueness, ties are resolved by choosing the leaf with the smallest index. The set of representative leaves x_i corresponding to the circular order $\mathcal{C}$ is denoted by $\mathcal{R}(\mathcal{C})$.

Let uv be a nontrivial cut-edge of a semi-directed network N. Then u (resp. v) is contained in a $(k+1)$-blob B_X (resp. $(l+1)$-blob B_Y) for some $k+1 \geq 3$ (resp. $l+1 \geq 3$). Recall that uv induces a nontrivial split $X|Y$ of the leaves of N. Moreover, B_X (resp. B_Y) induces a circular order $\mathcal{C}_1 = (X_1, \ldots, X_k, Y)$ (resp. $\mathcal{C}_2 = (X, Y_1, \ldots, Y_l)$), where $X = \bigcup_i X_i$ (resp. $Y = \bigcup_j Y_j$). Then, we let $\mathcal{D} = (X_1, \ldots, X_k)|(Y_1, \ldots, Y_l)$ be the *2-circular order* induced by uv, capturing both the induced split and the two induced circular orders. Note that X_1 and X_k (resp. Y_1 and Y_l) are the two neighbors of Y (resp. X) in $\mathcal{C}_1$ (resp. $\mathcal{C}_2$). In the 2-circular order $\mathcal{D}$, these are fixed as the outermost elements of the two sequences $(X_1, \ldots, X_k)$ and $(Y_1, \ldots, Y_l)$. Hence, unlike circular orders (which are defined up to cyclic permutation and reversal), $\mathcal{D}$ is defined only up to reversal of the two sequences, as it also records where the cut-edge attaches, namely between X_1 and X_k and between Y_1 and Y_l. See Fig. 3b and c for two examples. The set of representative leaves $\{x_1, \ldots, x_k, y_1, \ldots, y_l\}$ corresponding to $\mathcal{D}$ is again denoted by $\mathcal{R}(\mathcal{D})$, where $x_i \in \mathcal{R}(\mathcal{C}_1)$ and $y_j \in \mathcal{R}(\mathcal{C}_2)$.

We now extend the concept of a representative quartet from Ref. [14]—originally developed for rooted trees—to the context of quartet profiles and semi-directed level-1 networks.

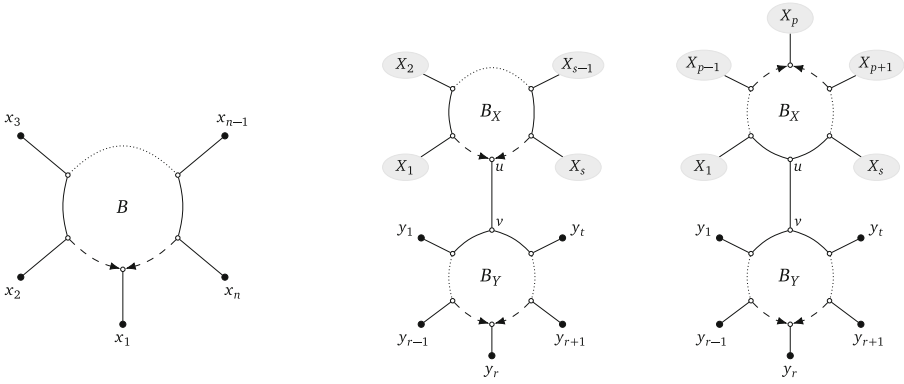

FIGURE 3. **a** A sunlet network on $\{x_1, \ldots, x_n\}$. The blob B induces the circular order $(\{x_1\}, \ldots, \{x_n\})$. **b**, **c** Two types of semi-directed level-1 networks on $X_1 \cup \ldots \cup X_s \cup \{y_1\} \cup \ldots \cup \{y_t\}$, with $s, t \geq 2$, $1 \leq r \leq t$ and $1 \leq p \leq s$ (for subfigure **c**). The y_j represents leaves, and the X_i represents subsets of leaves. The networks both have a nontrivial cut-edge uv (with its endpoints in the blobs B_X and B_Y) inducing the 2-circular order $(X_1, \ldots, X_s)|(\{y_1\}, \ldots, \{y_t\})$

Definition 3.4. (*Representative quartet profile*) Let N be a semi-directed level-1 network with leaf set $\mathcal{X}$. Then the set $\mathcal{Q}_R(N) \subseteq \mathcal{Q}(N)$ of *representative quartet profiles* is constructed as follows. For each cut-edge with induced 2-circular order $\mathcal{D} = (X_1, \ldots, X_k)|(Y_1, \ldots, Y_l)$ of $\mathcal{X}$ and $\mathcal{R}(\mathcal{D}) = \{x_1, \ldots, x_k, y_1, \ldots, y_l\}$,

(i) $x_1x_k|y_1y_l \in \mathcal{Q}_R(N)$.

For each k-blob B ($k \geq 4$) with induced circular order $\mathcal{C} = (X_1, \ldots, X_k)$ of $\mathcal{X}$ such that X_1 is the set of leaves below the reticulation in B and $\mathcal{R}(\mathcal{C}) = \{x_1, \ldots, x_k\}$,

(ii) $x_ix_{i+1}|x_{i+2}x_{i+3} \in \mathcal{Q}_R(N)$ for all $i \in \{2, \ldots, k-3\}$;
(iii) $(x_1x_ix_{i+1}x_{i+2}) \in \mathcal{Q}_R(N)$ for all $i \in \{2, \ldots, k-2\}$.

Note that the representative quartet profiles are uniquely determined and do not depend on the specific representation of the circular or 2-circular order (i.e., up to reversal and cyclic permutation, and reversal, respectively), as all such representations yield the same set. As an example, consider the network in Fig. 4. Then the quartet profiles $x_1x_7|x_9x_{11}$, $x_1x_2|x_5x_7$, and $(x_4x_5x_7x_1)$ are representative quartet profiles of types (i), (ii), and (iii), respectively. Whereas, for example, $x_1x_2|x_5x_{11}$, $x_1x_5|x_2x_3$, and $(x_4x_5x_7x_2)$ are not representative.

In general, there is exactly one representative quartet profile of type (i) for every nontrivial split of a network, and there are $(k-4)+(k-3) = 2k-7$ representative quartet profiles of types (ii) and (iii) for every k-blob with $k \geq 4$. Hence, among all level-1 networks with a fixed number of leaves, the size

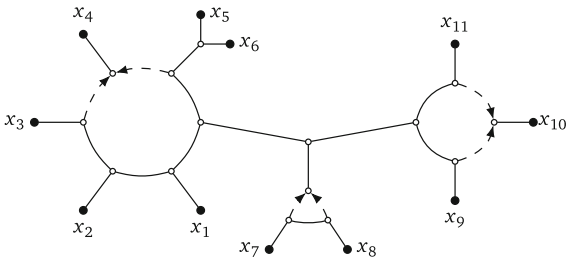

FIGURE 4. A semi-directed network on $\{x_1, \ldots, x_{11}\}$ used to illustrate the concept of a representative quartet profile

of $\mathcal{Q}_R(N)$ is largest for a sunlet network—which has one large blob and no nontrivial splits. The following observation follows.

Observation 3.5. *If N is a semi-directed level-1 network on n leaves, then $|\mathcal{Q}_R(N)| \leq 2n - 7$.*

In the remainder of this section, we prove that the representative quartet profiles are sufficient to build up the complete set of quartet profiles, solely using dyadic inference rules. We formalize this by means of the following definition and stress that the triadic rule R4* is not being applied.

Definition 3.6. (*Dyadic closure*) Let $\mathcal{Q}$ be a set of quartet profiles. The set $\mathrm{cl}_2(\mathcal{Q})$ of quartet profiles obtained from $\mathcal{Q}$ by exhaustively applying inference rules R1–R4 is the *dyadic closure* of $\mathcal{Q}$.

Observe that the dyadic closure operation cl_2 applied to a set of quartet profiles $\mathcal{Q}$ indeed satisfies the standard closure axioms: *extensivity* ($\mathcal{Q} \subseteq \mathrm{cl}_2(\mathcal{Q})$), *monotonicity* ($\mathcal{Q}_1 \subseteq \mathcal{Q}_2 \implies \mathrm{cl}_2(\mathcal{Q}_1) \subseteq \mathrm{cl}_2(\mathcal{Q}_2)$), and *idempotency* ($\mathrm{cl}_2(\mathrm{cl}_2(\mathcal{Q})) = \mathrm{cl}_2(\mathcal{Q})$).

Lemma 3.7. *Let N be a sunlet network. If $\mathcal{Q}$ is a set of quartet profiles such that $\mathcal{Q}_R(N) \subseteq \mathcal{Q} \subseteq \mathcal{Q}(N)$, then $\mathcal{Q}(N) = \mathrm{cl}_2(\mathcal{Q})$.*

Proof. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the leaves of N such that x_1 is the leaf below the reticulation and such that N induces the circular ordering $(x_1, \ldots, x_n)$ of its leaves. See also Fig. 3a.

By monotonicity of the dyadic closure (i.e., $\mathcal{Q}_1 \subseteq \mathcal{Q}_2 \implies \mathrm{cl}_2(\mathcal{Q}_1) \subseteq \mathrm{cl}_2(\mathcal{Q}_2)$; see also the remark after Definition 3.6), it suffices to consider the minimal case $\mathcal{Q} = \mathcal{Q}_R(N)$. Throughout the proof, we denote by $\mathcal{Q}_t(N)$ (resp. $\mathcal{Q}_c(N)$) the trivial (resp. nontrivial) quartet profiles in $\mathcal{Q}(N)$. We have already shown that the inference rules R1–R4 are valid for N (Proposition 3.1 and Lemmas 3.2 and 3.3), so $\mathrm{cl}_2(\mathcal{Q}) \subseteq \mathcal{Q}_t(N) \cup \mathcal{Q}_c(N)$. Thus, it suffices to show that $\mathrm{cl}_2(\mathcal{Q}) \supseteq \mathcal{Q}_t(N) \cup \mathcal{Q}_c(N)$. We prove the lemma by induction on n. The base cases with $n \leq 4$ follow trivially.

Let u_i denote the non-leaf neighbor of x_i, $1 \leq i \leq n$. Note that u_1 is the reticulation. Let $T = N|_{\{x_2,\ldots,x_n\}}$ and $N' = N|_{\{x_1,\ldots,x_{n-1}\}}$. All vertices except

x_1 and u_1 lie on up-down paths between leaves of T, and all vertices except x_n lie on such paths in N'. Hence, by Definition 2.3, T is obtained by removing x_1 and u_1 and suppressing u_2 and u_n, while N' is obtained by removing x_n and suppressing u_n. Thus, T is a tree and N' is a sunlet.

Since T is a tree, each quartet profile in $\mathcal{Q}_R(T)$ is a trivial quartet profile. Moreover, each quartet profile of N not involving x_1 is also trivial. Combining this with Definition 3.4—which implies that the leaf set of a representative quartet profile in T corresponds to the leaf set of a representative quartet profile in N—we obtain $\mathcal{Q}_R(T) \subseteq \mathcal{Q}_R(N)$. It follows from Ref. [14, Lem. 2] that exhaustively applying inference rules R1 and R2 to $\mathcal{Q}_R(T)$ results in $\mathcal{Q}(T)$: the set of all quartet profiles induced by the tree T. Since $\mathcal{Q}(T) = \mathcal{Q}_t(N)$, we get that $\mathrm{cl}_2(\mathcal{Q}) \supseteq \mathcal{Q}_t(N)$.

Since N' is a sunlet and it induces the same circular ordering as N except for the removal of the leaf x_n, each quartet profile of N' is a quartet profile of N. As before, in light of Definition 3.4, the leaf set of a representative quartet profile in N' corresponds to the leaf set of a representative quartet profile in N. So, $\mathcal{Q}_R(N') \subseteq \mathcal{Q}_R(N) = \mathcal{Q}$. In particular, $\mathcal{Q}_R(N) = \mathcal{Q}_R(N') \cup \{x_{n-3}x_{n-2}|x_{n-1}x_n, (x_1x_{n-2}x_{n-1}x_n)\}$. Thus, by the induction hypothesis, $\mathrm{cl}_2(\mathcal{Q}) \supseteq \mathcal{Q}(N')$. Since we also already have that $\mathrm{cl}_2(\mathcal{Q}) \supseteq \mathcal{Q}_t(N)$, it remains to show that $q = (x_1x_ix_jx_n) \in \mathrm{cl}_2(\mathcal{Q})$ for all $1 < i < j < n$. We consider two cases. (1): $j < n-1$. Then both $(x_1x_ix_jx_{n-1})$ (since it is in $\mathcal{Q}(N')$) and $x_ix_j|x_{n-1}x_n$ (since it is in $\mathcal{Q}_t(N)$) are in $\mathrm{cl}_2(\mathcal{Q})$. Applying rule R4 to these two quartet profiles shows that $q \in \mathrm{cl}_2(\mathcal{Q})$. (2): $j = n-1$. Then we may also assume that $i < n-2$, otherwise q is in $\mathcal{Q}$ and thus trivially in $\mathrm{cl}_2(N)$. Both $(x_1x_nx_{n-1}x_{n-2})$ (since it is in $\mathcal{Q}_R(N)$) and $x_nx_{n-1}|x_{n-2}x_i$ (since it is in $\mathcal{Q}_t(N)$) are in $\mathrm{cl}_2(\mathcal{Q})$, so applying rule R4 shows that $q \in \mathrm{cl}_2(\mathcal{Q})$. □

To prove the main theorem in this section (Theorem 3.9), the following lemma will first show a slightly weaker result about the dyadic closure. In particular, given a semi-directed level-1 network N, we call a quartet profile $q \in \mathcal{Q}(N)$ a *spanning quartet profile* if there exists a circular order $\mathcal{C}$ or 2-circular order $\mathcal{D}$ induced by an edge or a blob of N such that $\mathcal{L}(q) \subseteq \mathcal{R}(\mathcal{C})$ or $\mathcal{L}(q) \subseteq \mathcal{R}(\mathcal{D})$, respectively. That is, q is a quartet profile containing only representative leaves of some induced (2-)circular order of N. We denote the set of all spanning quartet profiles of N by $\mathcal{Q}_S(N)$ and note that $\mathcal{Q}_R(N) \subseteq \mathcal{Q}_S(N)$.

Lemma 3.8. *Let N be a semi-directed level-1 network. If $\mathcal{Q}$ is a set of quartet profiles such that $\mathcal{Q}_R(N) \subseteq \mathcal{Q} \subseteq \mathcal{Q}(N)$, then $\mathcal{Q}_S(N) \subseteq \mathrm{cl}_2(\mathcal{Q})$.*

Proof. As in the previous lemma, it suffices to consider the minimal case $\mathcal{Q} = \mathcal{Q}_R(N)$ due to the monotonicity of the dyadic closure. We may assume that the number of leaves n of N is at least 4, otherwise the result follows trivially. Since the presence of 2 cycles does not influence the quartet profiles, we will also assume that N has no 2 cycles. We prove the theorem by induction on the number of nontrivial splits.

The base case, where N has no nontrivial splits, implies that N has a single internal blob to which all leaves are attached. Further, this internal blob

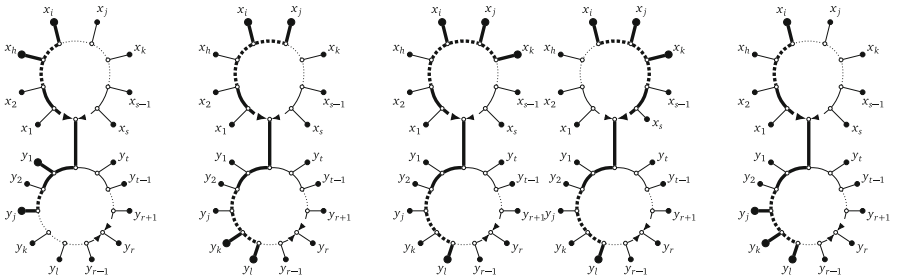

FIGURE 5. Illustration supporting the four claims in the proof of Lemma 3.8. Each subfigure depicts one option of how the considered quartet profile is displayed in the shown network. The quartet profiles are (from left to right): $x_h x_i | y_1 y_j$, $x_i x_j | y_k y_l$, $(x_i x_j x_k y_l)$, and $x_i y_j | y_k y_l$. The subfigure corresponding to Claim 3 shows both quartets $x_i x_j | x_k y_l$ and $x_j x_k | x_i y_l$ contributing to the nontrivial quartet profile $(x_i x_j x_k y_l)$

cannot be a single vertex, since N is binary and $n \geq 4$. Thus, N is a sunlet and the result then follows from the previous lemma.

Now assume that N has at least one nontrivial split. Let B_Y be a *lowest* internal blob of N, i.e., B_Y has only one incident nontrivial cut-edge uv (with v in B_Y) such that v is not the reticulation of B_Y. Let B_X be the blob that contains u. See Fig. 3b and c for an illustration. For simplicity, we also assume that both B_X and B_Y are nontrivial, i.e., if they are 3-blobs, they are not single vertices. Note that this transformation does not change any of the quartet profiles. Let $\mathcal{D} = (X_1, \ldots, X_s)|(\{y_1\}, \ldots, \{y_t\})$ be the induced two-circular order of uv, where $Y = \{y_1, \ldots, y_t\}$, $X = X_1 \cup \ldots \cup X_s$ and $\mathcal{R}(\mathcal{D}) = \{x_1, \ldots, x_s, y_1, \ldots, y_t\}$ contains the representative leaves. Here, y_r with $1 \leq r \leq t$ is the leaf below the reticulation of B_Y. Note that N can be of two different types (see again Fig. 3), depending on whether the leaves in Y are below the reticulation of B_X (type 1, shown in Fig. 3b), or whether some set X_p (with $1 \leq p \leq s$) is below the reticulation of B_X (type 2, shown in Fig. 3c).

Let $N_X = N|_{X \cup \{y_1\}}$ and let $N_Y = N|_{Y \cup \{x_h\}}$, where $x_h \in \{x_1, \ldots, x_s\}$, $1 \leq h \leq s$, is the leaf topologically closest to the blob B_Y, with ties resolved by choosing the leaf with the smallest index. Then $\mathcal{Q}_R(N) = \mathcal{Q}_R(N_X) \cup \mathcal{Q}_R(N_Y) \cup \{x_1 x_s | y_1 y_t\}$. Also observe that N_Y is a sunlet inducing a unique circular order $\mathcal{C}$ of all singleton sets of leaves in $Y \cup \{x_h\}$. Then each leaf of N_Y is a representative leaf of the singleton set it is contained in. So, each quartet profile of N_Y is a spanning quartet profile, i.e., $\mathcal{Q}_S(N_Y) = \mathcal{Q}(N_Y)$. Combining this with the induction hypothesis, we then get that $\mathcal{Q}_S(N_X) \cup \mathcal{Q}(N_Y) \cup \{x_1 x_s | y_1 y_t\} \subseteq \mathrm{cl}_2(\mathcal{Q})$. We now prove four claims, showing that certain quartet profiles in $\mathcal{Q}_S(N)$ are in $\mathrm{cl}_2(\mathcal{Q})$. Together with the fact that $\mathcal{Q}_S(N_X) \subseteq \mathrm{cl}_2(\mathcal{Q})$ and $\mathcal{Q}(N_Y) \subseteq \mathrm{cl}_2(\mathcal{Q})$,

Claims 2–4 then show that $\mathcal{Q}_S(N) \subseteq \mathrm{cl}_2(\mathcal{Q})$. For each claim, Fig. 5 illustrates how the corresponding quartet profile is displayed in the network, in the case where the network is of type 1 and each set X_i is a singleton $\{x_i\}$. For the figure, we assume $h < i$ and $r - 1 > l$, such that x_h appears to the left of the leaves x_i, x_j, x_k, and y_r appears to the right of the leaves y_j, y_k, y_l in the circular ordering induced by the shown planar representation.

Claim 1: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x_h, x_i, y_1, y_j\}$ with $1 \leq i \leq s$ such that $i \neq h$ and $1 < j \leq t$.* Without loss of generality, we assume that $1 \leq h < i \leq s$. Note that every such q has the form $x_h x_i | y_1 y_j$ (see Fig. 5). Observe that $q_1 = x_1 x_s | y_1 y_t$ is in $\mathrm{cl}_2(\mathcal{Q})$. If $p \in \{1, i, s\}$, then $q_2 = (x_1 x_i x_s y_1)$ is in $\mathrm{cl}_2(\mathcal{Q})$ (Case 1(i)). Otherwise, without loss of generality, we may assume that $q_2 = x_1 x_i | x_s y_1$ is in $\mathrm{cl}_2(\mathcal{Q})$ (Case 1(ii)). In Case 1(i) (resp. Case 1(ii)), we can apply rule R3 (resp. rule R1) to q_1 and q_2, to obtain that $q_3 = x_1 x_i | y_1 y_t$ is in $\mathrm{cl}_2(\mathcal{Q})$. (Note that this follows trivially if $i = s$, since then $q_3 = q_1$). If $h = 1$, the proof is done, so we assume that $h > 1$. An analogous proof then shows that $q_4 = x_1 x_h | y_1 y_t$ is in $\mathrm{cl}_2(\mathcal{Q})$. Applying rule R2 to q_3 and q_4 then shows that q is in $\mathrm{cl}_2(\mathcal{Q})$.

Claim 2: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x_i, x_j, y_k, y_l\}$ with $1 \leq i < j \leq s$ and $1 \leq k < l \leq t$.* Again, every such q has the form $x_i x_j | y_k y_l$ (see Fig. 5). First assume that $k > 1$, the proof being analogous for $k = 1$ by changing y_1 to y_2 throughout. By the previous claim, we know that $q_1 = x_h x_i | y_1 y_k$ and $q_2 = x_h x_j | y_1 y_k$ are in $\mathrm{cl}_2(\mathcal{Q})$. Applying rule R2 shows $q_3 = x_i x_j | y_1 y_k \in \mathrm{cl}_2(\mathcal{Q})$. (This follows directly if $h \in \{i, j\}$.) Similarly, it follows that $q_4 = x_i x_j | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$. Then again applying rule R2 to q_3 and q_4 gives that $q = x_i x_j | y_k y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$.

Claim 3: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x_i, x_j, x_k, y_l\}$ with $1 \leq i < j < k \leq s$ and $1 \leq l \leq t$.* If N is of type 1 or of type 2 with $p \in \{i, j, k\}$, then q has the form $(x_i x_j x_k y_l)$ (Case 3(i), see Fig. 5). If N is of type 2 with $p \notin \{i, j, k\}$, then we may assume without loss of generality that q has the form $x_i x_j | x_k y_l$ (Case 3(ii)). Next, observe that in Case 3(i) (resp. Case 3(ii)), $q_1 = (x_i x_j x_k y_1)$ (resp. $q_1 = x_i x_j | x_k y_1$) is in $\mathcal{Q}(N_X) \subseteq \mathrm{cl}_2(\mathcal{Q})$ and that $q_2 = x_j x_k | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$ by Claim 2. Then applying rule R4 (resp. rule R2) to q_1 and q_2 proves that $q \in \mathrm{cl}_2(\mathcal{Q})$.

Claim 4: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x_i, y_j, y_k, y_l\}$ with $1 \leq i \leq s$ and $1 \leq j < k < l \leq t$.* If $r \in \{j, k, l\}$, then q has the form $(x_i y_j y_k y_l)$ (Case 4(i)). Otherwise, we may assume without loss of generality that q has the form $x_i y_j | y_k y_l$ (Case 4(ii), see Fig. 5). In Case 4(i) (resp. Case 4(ii)), $q_1 = (x_h y_j y_k y_l)$ (resp. $q_1 = x_h y_j | y_k y_l$) is in $\mathcal{Q}(N_Y) \subseteq \mathrm{cl}_2(\mathcal{Q})$, so we may assume that $i \neq h$. Then $q_2 = x_h x_i | y_k y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$ by Claim 2. Applying rule R4 (resp. rule R4) to q_1 and q_2 proves that $q \in \mathrm{cl}_2(\mathcal{Q})$. □

Theorem 3.9. *Let N be a semi-directed level-1 network. If $\mathcal{Q}$ is a set of quartet profiles such that $\mathcal{Q}_R(N) \subseteq \mathcal{Q} \subseteq \mathcal{Q}(N)$, then $\mathcal{Q}(N) = \mathrm{cl}_2(\mathcal{Q})$.*

Proof. Since the inference rules R1–R4 are valid for N (Proposition 3.1and Lemmas 3.2 and 3.3), we have that $\mathcal{Q}(N) \supseteq \mathrm{cl}_2(\mathcal{Q})$. It remains to show that $\mathcal{Q}(N) \subseteq \mathrm{cl}_2(\mathcal{Q})$.

By Lemma 3.8, we have $\mathcal{Q}_S(N) \subseteq \mathrm{cl}_2(\mathcal{Q})$. Using monotonicity and idempotency of the dyadic closure (see the paragraph after Definition 3.6), this implies $\mathrm{cl}_2(\mathcal{Q}_S(N)) \subseteq \mathrm{cl}_2(\mathrm{cl}_2(\mathcal{Q})) = \mathrm{cl}_2(\mathcal{Q})$. Therefore, if we show $\mathcal{Q}(N) \subseteq \mathrm{cl}_2(\mathcal{Q}_S(N))$, the inclusion $\mathcal{Q}(N) \subseteq \mathrm{cl}_2(\mathcal{Q})$ will follow. So, it suffices to prove $\mathcal{Q}(N) \subseteq \mathrm{cl}_2(\mathcal{Q})$ for $\mathcal{Q} = \mathcal{Q}_S(N)$. We prove the theorem by induction on the number of nontrivial splits, with the base case again following from Lemma 3.7.

Consider the exact same setup as in the proof of Lemma 3.8. Then by the induction hypothesis, we have that $\mathcal{Q}(N_X) \cup \mathcal{Q}(N_Y) \subseteq \mathrm{cl}_2(\mathcal{Q})$. We slightly extend the notation by using x'_i to indicate an arbitrary leaf in the set X_i, as opposed to the notation x_i that was used for the representative leaf of X_i. With slight abuse of notation, when considering leaves x'_i and x'_j, we allow $i = j$, in which case the leaves are meant to be two distinct leaves in $X_i = X_j$. We now prove three claims for quartet profiles $q \in \mathcal{Q}(N)$.

Claim 5: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x'_i, x'_j, y_k, y_l\}$ with $1 \leq i, j \leq s$ and $1 \leq k < l \leq t$.* Observe that the subnetwork $N|_{\mathcal{L}(q)}$ always induces the split $x'_i x'_j | y_k y_l$ (see also Fig. 3b and c). Hence, every displayed quartet of N induces the split $x'_i x'_j | y_k y_l$, implying that the quartet profile q has the form $x'_i x'_j | y_k y_l$. Since $x_i x_j | y_k y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$, we can assume without loss of generality that $x'_i \neq x_i$. We consider two cases.

Case (i): $i \neq j$. Observe that $x_i x'_i | x_j y_1$ is in $\mathcal{Q}(N_X) \subseteq \mathrm{cl}_2(\mathcal{Q})$ and $x_i x_j | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$. Then by rule R1, $q_1 = x_i x'_i | y_1 y_l$ and $q_2 = x'_i x_j | y_1 y_l$ are in $\mathrm{cl}_2(\mathcal{Q})$. By symmetry, $q_3 = x'_j x_i | y_1 y_l$ is then also in $\mathrm{cl}_2(\mathcal{Q})$. (Note that this follows trivially if $x_j = x'_j$.) Applying rule R2 to q_1 and q_3 gives that $q_4 = x'_i x'_j | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$. If $k = 1$, we are done. Otherwise, again using symmetry, we get that $q_5 = x'_i x'_j | y_1 y_k$ is in $\mathrm{cl}_2(\mathcal{Q})$. Applying rule R2 to q_4 and q_5 gives that q is in $\mathrm{cl}_2(\mathcal{Q})$.

Case (ii): $i = j$. Then x'_i and x'_j are two distinct leaves in $X_i = X_j$. Let $m \neq i, j$ be arbitrary, and note that x_m always exists since $i = j$. Observe that $x_i x'_i | x_m y_1$ is in $\mathcal{Q}(N_X) \subseteq \mathrm{cl}_2(\mathcal{Q})$ and $x_i x_m | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$. Then by rule R1, $q_1 = x_i x'_i | y_1 y_l$ and $q_2 = x'_i x_m | y_1 y_l$ are in $\mathrm{cl}_2(\mathcal{Q})$. Similarly, by swapping x'_i for x'_j, $q_3 = x'_j x_m | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$ (which is trivial if $x'_j = x_i$). Applying rule R2 to q_2 and q_3 gives that $q_4 = x'_i x'_j | y_1 y_l$ is in $\mathrm{cl}_2(\mathcal{Q})$. The rest of the proof is analogous to Case (i).

Claim 6: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x'_i, x'_j, x'_k, y_l\}$ with $1 \leq i, j, k \leq s$ and $1 \leq l \leq t$.* The proof of this claim is analogous to the proof of Claim 3 in the proof of Lemma 3.8, but instead relying on Claim 5 instead of Claim 2.

Claim 7: *$q \in \mathrm{cl}_2(\mathcal{Q})$ if $\mathcal{L}(q) = \{x'_i, y_j, y_k, y_l\}$ with $1 \leq i \leq s$ and $1 \leq j < k < l \leq t$.* Again, the proof of this claim is analogous to the proof of Claim 4 in the proof of Lemma 3.8, but instead relying on Claim 5 instead of Claim 2.

To conclude the proof of the theorem, observe that any quartet profile in $\mathcal{Q}(N) \setminus (\mathcal{Q}(N_X) \cup \mathcal{Q}(N_Y) \cup \mathcal{Q}_S(N))$ is shown to be in $\mathrm{cl}_2(\mathcal{Q})$ by one of the three claims. □

Recall that Proposition 2.5 stated that given a semi-directed level-1 network N, the mixed graph $\tilde{N}$ obtained from N by suppressing 2-cycles, 3-cycles

and reticulation vertices in 4-cycles, is uniquely characterized by quartet profiles. Combining the previous theorem with Proposition 2.5 and Observation 3.5, we obtain the following corollary, where we use the term *quarnet* for a 4-leaf subnetwork of a network.,[3]

Corollary 3.10. *Let N be a semi-directed level-1 network on $n \geq 4$ leaves. Then*

(a) *there exists a set of $\mathcal{O}(n)$ quartet profiles of N that uniquely characterizes $\tilde{N}$;*
(b) *there exists a set of $\mathcal{O}(n)$ quarnets of N that uniquely characterizes N.*

Proof. By Observation 3.5, there are $\mathcal{O}(n)$ representative quartet profiles. By the previous theorem, the full set of quartet profiles can be obtained from these, and by Proposition 2.5, this set uniquely characterizes $\tilde{N}$. This yields part (a).

For part (b), we first note that by Proposition A.1, a quartet profile on leaf set $\{a, b, c, d\}$—that is, the set of 4-leaf subtrees induced by $\{a, b, c, d\}$ in all displayed trees of N—can equivalently be obtained as the set of displayed trees of the 4-leaf subnetwork/quarnet $N|_{\{a,b,c,d\}}$. Hence, by part (a), there exists a set of $\mathcal{O}(n)$ quarnets that uniquely charachterizes $\tilde{N}$.

To recover the full network structure N (i.e., to determine the resolution of 2-cycles, 3-cycles, and reticulation in 4-cycles), it suffices to consider specific quarnets. For each 4-cycle, a single quarnet with one leaf on each of the four sides of the 4-cycle determines the placement of the reticulation. For each 3-blob in $\tilde{N}$ inducing a circular order (X_1, X_2, X_3), one quarnet with one leaf from each of two parts and two leaves from the remaining part suffices to determine its structure.

Finally, 2-cycles can only be placed on cut-edges of N, as otherwise the resulting network would not be level-1. To detect 2-cycles on a trivial cut-edge, it suffices to consider any quarnet containing the corresponding leaf. For a nontrivial cut-edge inducing a 2-circular order $(X_1, \dots, X_k) \,|\, (Y_1, \dots, Y_\ell)$, it suffices to consider a quarnet containing two leaves from (distinct parts of) the X-side and two leaves from (distinct parts of) the Y-side. The corresponding quarnet then contains a nontrivial cut-edge, and the presence or absence of a 2-cycle on this cut-edge in the quarnet directly reflects its presence or absence in the original network. Consequently, this determines whether a 2-cycle is present on the given cut-edge of N.

Altogether, since there are $\mathcal{O}(n)$ blobs and cut-edges in N, this yields a linear-sized set of quarnets that uniquely characterizes N. □

[3] In Ref. [16], it was shown that any algorithm that sequentially adds leaves to reconstruct a semi-directed level-1 network from quarnets must rely on $\Omega(n \log n)$ quarnets in the worst case. Although part (b) of Corollary 3.10 may appear to contradict this bound, it does not violate the result presented there. In particular, the bound from Ref. [16] is a lower bound on the *query complexity* meaning that for any algorithm relying on queries to an oracle (which returns quarnets of the network to be reconstructed), there exists an adversarial strategy that forces the algorithm to make $\Omega(n \log n)$ queries. In contrast, Corollary 3.10b instead says that once a network is known, a linear-sized set of quarnets exists that completely characterizes it.

Algorithm 1: Dyadic Closure Network Construction (DCNC) method

Input: a set of quartet profiles $\mathcal{Q}$ on $\mathcal{X}$
Output: a semi-directed level-1 network N' on $\mathcal{X}$ with 2-cycles, 3-cycles and reticulation vertices in 4-cycles suppressed, Insufficient, or Inconsistent

compute the dyadic closure $\mathrm{cl}_2(\mathcal{Q})$ analogous to [14]
if $\mathrm{cl}_2(\mathcal{Q})$ *contains more than one quartet profile for some set of four leaves* **then**
    return Inconsistent
else if $|\mathrm{cl}_2(\mathcal{Q})| < \binom{n}{4}$ **then**
    return Insufficient
else
    $N' \leftarrow$ semi-directed level-1 network with 2-cycles, 3-cycles and reticulation vertices in 4-cycles suppressed, obtained by applying the algorithm from [16] to $\mathcal{Q}$
    return N'

4. The Dyadic Closure Method

In this section, we present an algorithm that reconstructs (most of) a semi-directed level-1 network from a sequence alignment on $\mathcal{X}$. We follow a strategy similar as in Ref. [14] and as in the analysis of the 'short quartet support method' in Ref. [40].

As a first ingredient, we present the DCNC (Dyadic Closure Network Construction) method (see Algorithm 1), which takes as input a set of quartet profiles, and outputs a level-1 network, inconsistent (if two quartet profiles are incompatible), or insufficient (if the quartet profiles do not provide enough information to construct a network). From Theorem 3.9, we can draw an immediate conclusion for the DCNC method, formulated in Corollary 4.1. Recall from Proposition 2.5 that we use the notation $\tilde{N}$ to denote the mixed graph obtained from a semi-directed level-1 network N by suppressing 2-cycles, 3-cycles, and reticulation vertices in 4-cycles.

Corollary 4.1. *Let N be a semi-directed level-1 network on $\mathcal{X}$ and let $\mathcal{Q}$ be a set of quartet profiles such that $\mathcal{Q}_R(N) \subseteq \mathcal{Q} \subseteq \mathcal{Q}(N)$. Then the network N' returned by the DCNC method satisfies $N' = \tilde{N}$.*

Definition 4.2. (*width of a quartet*) For a sequence alignment $\mathcal{S}$ on $\mathcal{X}$, the *width* of any quartet tree or quartet profile on four leaves $Q \subseteq \mathcal{X}$ is defined as the maximum dissimilarity score h^{ij} of sequences s^i and s^j among all $x_i, x_j \in Q$.

Our second ingredient, the L1QPC (Level-1 Quartet Profile Construction) method (see Algorithm 2), infers a set of quartet profiles from a sequence alignment on $\mathcal{X}$, where the quartets considered have bounded width.

Algorithm 2: Level-1 Quartet Profile Construction (L1QPC) method

Input: a real number $w \geq 0$, an empirical distance matrix d from a sequence alignment on $\mathcal{X}$

Output: a set of quartet profiles $\mathcal{Q}_w$ of width at most w

for $\{x_i, x_j, x_k, x_l\} \subseteq \mathcal{X}$ *with width at most* w *and* $d_{ij} + d_{kl} \leq d_{il} + d_{jk} \leq d_{ik} + d_{jl}$ **do**
  $\Delta_1 \leftarrow (d_{il} + d_{jk}) - (d_{ij} + d_{kl})$;
  $\Delta_2 \leftarrow (d_{ik} + d_{j\ell}) - (d_{il} + d_{jk})$;
  if $\Delta_1 = \Delta_2$ **then**
    Add trivial quartet profiles $x_ix_j|x_kx_l$, $x_ix_k|x_jx_l$ and $x_ix_l|x_jx_k$ to $\mathcal{Q}_w$.
  else if $\Delta_1 > \Delta_2$ **then**
    Add trivial quartet profile $x_ix_j|x_kx_l$ to $\mathcal{Q}_w$.
  else
    Add quartet profile $(x_ix_jx_kx_l)$ to $\mathcal{Q}_w$.
return Q_w

Intuitively, increasing the input parameter w for the L1QPC method increases the number of quartet profiles that are returned. This decreases the confidence that all inferred quartet profiles are correct, as there is a higher probability that at least one profile is incorrect. On the other hand, a small value of w increases this confidence, but forces L1QPC to return only a small set of short-width quartet profiles which might be insufficient to guarantee the inference of the dyadic closure (see Theorem 3.9).

Now, we are ready to state the main algorithm (see Algorithm 3). The DC method does a binary search over the space of parameters w to find a large enough set of quartet profiles such that the representative quartet profiles are also returned, but small enough such that the dyadic closure can be inferred correctly. The algorithm iteratively applies the previous algorithms L1QPC and DCNC, making it a polynomial-time algorithm:

Proposition 4.3. *The* DC *method can be implemented to run in* $\mathcal{O}(n^5 \cdot \log n)$ *time.*

Proof. Since the DC method performs a binary search over n^2 distances, both L1QPC and DCNC are called at most $\mathcal{O}(\log n)$ times. Clearly, L1QPC takes $\mathcal{O}(n^4)$ time, whereas DCNC can be implemented to run in $\mathcal{O}(n^5)$ time. Specifically, the implementation outlined in Ref. [14, Thm. 5] shows that the dyadic closure (using only rules R1 and R2) can be computed in $\mathcal{O}(n^5)$ time. Since our additional rules R3 and R4 are also dyadic, this time complexity remains unchanged. Furthermore, the algorithm in Ref. [16] runs in $\mathcal{O}(n^2)$ time (Proposition 2.5), proving that DCNC indeed takes $\mathcal{O}(n^5)$ time. □

In the next section, we show that the DC method works as intended. This means, we will show that there exists a set of quartet profiles $\mathcal{Q}_w$ and a semi-directed level-1 network $\tilde{N}$ (with 2-cycles, 3-cycles and reticulation vertices in 4-cycles suppressed) returned by Algorithm 3 such that

1. $\mathcal{Q}_R(N) \subseteq \mathcal{Q}_w$,
2. $\mathcal{Q}_w \subseteq \mathcal{Q}(N)$,
3. the L1QPC method finds Q_w during the binary search with high probability.

The following technical result establishes condition (1). We will extend this construction to show that conditions (2) and (3) hold with high probability given a sufficient number of sites k in the sequence alignment $\mathcal{S}$ on $\mathcal{X}$.

Lemma 4.4. *Let $N = (V, E)$ be a semi-directed level-1 network on $\mathcal{X}$, $\mathcal{S}$ be a sequence alignment on $\mathcal{X}$ and let*

$$w > \frac{1}{2}\left(1 - \min_{e \in E} \theta(e)^{depth(N)+lc(N)}\right).$$

Assume for all $x_i, x_j \in \mathcal{X}$ that if $\mathbb{P}[Y(x_i) \neq Y(x_j)] < w$ in each displayed tree of N, then $h^{ij} < w$, where h^{ij} is the dissimilarity score between the sequences

Algorithm 3: Dyadic Closure (DC) method

Input: a sequence alignment S on $\mathcal{X}$
Output: a semi-directed level-1 network N' on $\mathcal{X}$ with 2-cycles, 3-cycles and reticulation vertices in 4-cycles suppressed, or Fail

for $x_i, x_j \in \mathcal{X}$ **do**
  Compute dissimilarity score h^{ij} from $\mathcal{S}$;
  $d_{ij} \leftarrow -\frac{1}{2}\log(1 - 2h^{ij})$;
$N' \leftarrow$ None
for *dissimilarity score w derived from $\mathcal{S}$* **do** `// perform a binary search`
  $\mathcal{Q}_w \leftarrow \text{L1QPC}(w, d)$
  if $\text{DCNC}(\mathcal{Q}_w) =$ Insufficient **then**
    increase w
  else if $\text{DCNC}(\mathcal{Q}_w) =$ Inconsistent **then**
    decrease w
  else if $\text{DCNC}(\mathcal{Q}_w) = \tilde{N}_w$ *for some semi-directed level-1 network N_w* **then**
    $N' \leftarrow \tilde{N}_w$
    increase w
if $N' =$ None **then**
  return Fail
return N'

in $\mathcal{S}$ associated to x_i and x_j. Then there exists $\epsilon > 0$ such that

$$\mathcal{Q}_R(N) \subseteq \mathcal{Q}_{w-\epsilon}$$

for the set of quartet profiles $\mathcal{Q}_{w-\epsilon}$ of width at most $w - \epsilon$.

Proof. We consider quartet profiles of types (i), (ii), and (iii). First, let $x_1x_k|y_1y_l \in \mathcal{Q}_R(N)$ and let T be any displayed tree of N. Then the path $e_1e_2 \dots e_m$ connecting taxa $z_i, z_j \in \{x_1, x_k, y_1, y_l\}$ in T yields

$$\begin{aligned}\mathbb{P}\left[Y(z_i) \neq Y(z_j)\right] &= \frac{1}{2}\left(1 - \prod_{i=1}^{m}(1-2p(e_i))\right) \leq \frac{1}{2}\left(1 - \min_{e\in E}\theta(e)^{\mathrm{depth}(T)+lc(N)}\right) \\ &\leq \frac{1}{2}\left(1 - \min_{e\in E}\theta(e)^{\mathrm{depth}(N)+lc(N)}\right) < w.\end{aligned}$$

We assumed that $h^{ij} < w$ for $\mathbb{P}\left[Y(x_i) \neq Y(x_j)\right] < w$. Hence, there exists $\epsilon > 0$ such that

$$x_1x_k|y_1y_l \in \mathcal{Q}_{w-\epsilon}.$$

Next, let $x_ix_{i+1}|x_{i+2}x_{i+3} \in \mathcal{Q}_R(N)$. Then similar to our first case

$$\mathbb{P}\left[Y(z_i) \neq Y(z_j)\right] \leq \frac{1}{2}\left(1 - \min_{e\in E}\theta(e)^{\mathrm{depth}(N)}\right) < w,$$

leading to the same conclusion. Finally, $(x_1x_ix_{i+1}x_{i+2}) \in \mathcal{Q}_R(N)$ is analogous to our first case. □

5. Performance of the DC Method Under the CFN Model

First, we provide sufficient conditions for the L1QPC method to return the correct set of quartet profiles $\mathcal{Q}_w$. To this end, for a semi-directed level-1 network N under the CFN model, we define a function s which maps the reticulation edges e of N to the positive real numbers by equations

$$\begin{aligned}&\frac{\gamma(e)}{1-\gamma(e)} \cdot \frac{\log \min_{e'\in E}\theta(e')}{\log \max_{e'\in E}\theta(e')} + \epsilon \\ &= \begin{cases} \frac{2s(e)^2-7s(e)}{2s(e)^2+s(e)-1} & \text{if } \frac{\gamma(e)}{1-\gamma(e)} \cdot \frac{\log \min_{e'\in E}\theta(e')}{\log \max_{e'\in E}\theta(e')} < 1 \\ \frac{4s(e)^3+4s(e)^2-s(e)-1}{12s(e)^2+4s(e)+1} & \text{otherwise} \end{cases}\end{aligned} \tag{2}$$

for a sufficiently small $\epsilon > 0$ such that $s(e) > 1/2$. Observe that function s is well-defined because the left hand side of Eq. (2) is a positive real number defined by our model parameters and the right hand side either converges to 1 or diverges for increasing $s(e)$.

Lemma 5.1. *Let N be a semi-directed level-1 network on leafset $\mathcal{Y} = \{x_i, x_j, x_k, x_l\}$ with $(x_ix_jx_kx_l) \in \mathcal{Q}_R(N)$, containing a cycle of length four and displayed trees T_1 and T_2. Let e be the reticulation edge from N present in T_1 and let d be the matrix of dissimilarity scores on $\mathcal{Y}$ defined by a sequence alignment on $\mathcal{Y}$. Then for the width w of $\mathcal{Y}$, the L1QPC method infers $(x_ix_jx_kx_l)$ from*

d if the inheritance probability $\gamma(e)$ *satisfies* $D = \gamma(e)D(T_1) + (1-\gamma(e))D(T_2)$ *such that*

$$\max\{|d_r - D_r| : r \in \{ij, kl, ik, jl\}\} < -\frac{\gamma(e)}{4s(e)} \log \max_{e' \in E} \theta(e'), \tag{3}$$

$$\max\{|d_r - D_r| : r \in \{il, jk, ik, jl\}\} < \frac{\gamma(e)-1}{4s(e)} \log \max_{e' \in E} \theta(e'). \tag{4}$$

The proof of Lemma 5.1 is deferred to Appendix B as it is rather technical. Throughout this section, we will make repeated use of the following Azuma–Hoeffding concentration inequalities [5]:

Lemma 5.2. *Let* k *be a positive integer, let* S *be any set and let* $Y = (Y_1, \ldots, Y_k)$ *be independent random variables taking values in* S*. For some* $t \in \mathbb{R}$*, let* $L : S^k \to \mathbb{R}$ *be a function such that* $|L(a) - L(b)| \leq t$ *whenever* a *and* b *differ in only one coordinate. Then for* $\lambda > 0$*,*

$$\mathbb{P}\left[L(Y) - \mathbb{E}[L(Y)] \geq \lambda\right] \leq \exp\left(-\frac{\lambda^2}{2t^2k}\right) \text{ and } \mathbb{P}\left[L(Y)\right.$$
$$\left. -\mathbb{E}[L(Y)] \leq -\lambda\right] \leq \exp\left(-\frac{\lambda^2}{2t^2k}\right).$$

To simplify the continued use of Lemma 5.2, we consider k to be the number of sites of a sequence alignment $\mathcal{S}$ on $\mathcal{X}$ under the CFN model. Let $Y = (Y_1, \ldots, Y_k)$ be independent random variables taking values $Y_{ij} = s_i^j$ for $s^j \in \mathcal{S}$, $i \in \{1, \ldots, k\}$ and, for all $i, j \in \{1, \ldots, n\}$, define the function $L : \{0,1\}^{n \times k} \to \mathbb{R}$ by

$$M \mapsto \frac{1}{k} d_H\left((M_{hi})_{h \in \{1,\ldots,n\}}, (M_{hj})_{h \in \{1,\ldots,n\}}\right),$$

i.e., $|L(M) - L(N)| \leq 1/k$ for all $M, N \in \{0,1\}^{n \times k}$ which differ in only one coordinate. Then we arrive at following special case of Lemma 5.2:

Corollary 5.3. *For* $x_i, x_j \in \mathcal{X}$*, the number of sites* k *and* $\lambda > 0$*,*

$$\mathbb{P}\left[h^{ij} - \mathbb{E}[h^{ij}] \geq \lambda\right] \leq \exp\left(-\frac{\lambda^2 k}{2}\right) \text{ and } \mathbb{P}\left[h^{ij} - \mathbb{E}[h^{ij}] \leq -\lambda\right] \leq \exp\left(-\frac{\lambda^2 k}{2}\right).$$

Now, by bounding the compound changing probabilities $\mathbb{P}[Y(v_1) \neq Y(v_2)]$ (see Lemma 2.6) and fixing the inheritance probabilities $\gamma(e)$, we can derive a lower bound on the probability that the L1QPC method returns the correct quartet profile.

Proposition 5.4. *Let* $N = (V, E)$ *be a semi-directed level-1 network and let* $E_{\max}$ *be an upper bound on the compound changing probability over all paths in* N *between taxa* $\{x_i, x_j, x_k, x_l\}$*. Suppose* k *sites evolve under the CFN model on* N *and assume* $N_{|(x_i,x_j,x_k,x_l)}$ *contains a reticulation edge* e*. Then the probability that* L1QPC *fails to return the correct quartet profile on taxa* $\{x_i, x_j, x_k, x_l\}$ *for a fixed inheritance probability* $\gamma(e) \in (0,1)$ *is at most*

$$12 \exp\left(-\frac{k}{8}\left(1 - \max_{e' \in E} \theta(e')^{\min\{\gamma(e), 1-\gamma(e)\}/(2s(e))}\right)^2 (1 - 2E_{\max})^2\right).$$

Equivalently, L1QPC *returns the correct quartet profile with probability* $1 - o(1)$ *if*

$$\min\{\gamma(e), 1-\gamma(e)\} \geq 2s(e) \cdot \frac{\log\left(1-\sqrt{M/k}\right)}{\log \max_{e' \in E} \theta(e')} \tag{5}$$

where M *is a sufficiently large positive number increasing in* $E_{\max}$ *with* $k > M$.

Proof. The quartet profile $(x_i x_j x_k x_l) \in \mathcal{Q}_R(N_{|\{x_i,x_j,x_k,x_l\}})$ is not trivial because $N_{|(x_i,x_j,x_k,x_l)}$ contains a reticulation edge e. Let γ and s denote $\gamma(e)$ and $s(e)$, respectively. Then we know from Proposition 5.1 that L1QPC returns $(x_i x_j x_k x_l)$ for $\gamma \in (0,1)$ such that for displayed trees T_1 and T_2 of $N_{|\{x_i,x_j,x_k,x_l\}}$ and $D = \gamma D(T_1) + (1-\gamma) D(T_2)$,

$$\max\{|d_r - D_r| : r \in \{ij, kl, ik, jl\}\} < -\frac{\gamma}{4s} \log \max_{e \in E} \theta(e),$$
$$\max\{|d_r - D_r| : r \in \{il, jk, ik, jl\}\} < \frac{\gamma - 1}{4s} \log \max_{e \in E} \theta(e).$$

Then the probability that L1QPC fails to return the correct quartet profile on taxa $\{x_i, x_j, x_k, x_l\}$ for a fixed inheritance probability $\gamma \in (0,1)$ and $m = \max_{e \in E} \theta(e)$ is upper bounded by

$$\begin{aligned}
&\sum_{r \in \{ij,kl\}} \mathbb{P}\left[|d_r - D_r| \geq -\frac{\gamma}{4s} \log m\right] \\
&\quad + \sum_{r \in \{il,jk\}} \mathbb{P}\left[|d_r - D_r| \geq \frac{\gamma - 1}{4s} \log m\right] \\
&\quad + \sum_{r \in \{ik,jl\}} \mathbb{P}\left[|d_r - D_r| \geq -\frac{\max\{\gamma, 1-\gamma\}}{4s} \log m\right].
\end{aligned}$$

Recall that $d_r = -\log(1-2h^r)/2$, $D(T_i)_r = -\log(1-2\mathbb{E}_i[h^r])/2$, $i \in \{1,2\}$, where $\mathbb{E}_i[h^r]$ denotes the expectation $\mathbb{E}[h^r]$ in T_i. Then for $r \in \{ij, kl, il, jk, ik, jl\}$,

$$\begin{aligned}
|d_r - D_r| &= \frac{1}{2} \left|\log(1-2h^r) - \gamma \log(1-2\mathbb{E}_1[h^r]) - (1-\gamma)\log(1-2\mathbb{E}_2[h^r])\right| \\
&= \frac{1}{2} \left|\log \frac{1-2h^r}{(1-2\mathbb{E}_1[h^r])^\gamma (1-2\mathbb{E}_2[h^r])^{1-\gamma}}\right|.
\end{aligned}$$

Case 1: $d_r \leq D_r$. Without loss of generality $r \in \{ij, kl\}$. Then we want to calculate

$$\mathbb{P}\left[\frac{1}{2} \log \frac{1-2h^r}{(1-2\mathbb{E}_1[h^r])^\gamma (1-2\mathbb{E}_2[h^r])^{1-\gamma}} \geq -\frac{\gamma}{4s} \log m\right] \tag{6}$$

which is equivalent to

$$\mathbb{P}\left[1 - 2h^r \geq (1-2\mathbb{E}_1[h^r])^\gamma (1-2\mathbb{E}_2[h^r])^{1-\gamma} m^{-\gamma/(2s)}\right]. \tag{7}$$

Let $\mathbb{E}[h^r] = \max\{\mathbb{E}_1[h^r], \mathbb{E}_2[h^r]\}$. Then probability (7) is upper bounded by

$$\mathbb{P}\left[1 - 2h^r \geq (1 - 2\mathbb{E}[h^r])\, m^{-\gamma/(2s)}\right].$$

After shifting the inequality by $(2\mathbb{E}[h^r] - 1)$, we arrive at

$$\mathbb{P}\left[h^r - \mathbb{E}[h^r] \leq -\frac{1}{2}\left(m^{-\gamma/(2s)} - 1\right)(1 - 2\mathbb{E}[h^r])\right].$$

Since $m^{\gamma/(2s)} \in (0, 1)$, we can apply Corollary 5.3 to upper bound (6) by

$$\exp\left(-\frac{k}{8}\left(m^{-\gamma/(2s)} - 1\right)^2 (1 - 2\mathbb{E}[h^r])^2\right).$$

Case 2: $d_r > D_r$. Without loss of generality $r \in \{ij, kl\}$. Then we want to calculate

$$\mathbb{P}\left[\frac{1}{2}\log\frac{(1 - 2\mathbb{E}_1[h^r])^\gamma (1 - 2\mathbb{E}_2[h^r])^{1-\gamma}}{1 - 2h^r} \geq -\frac{\gamma}{4s}\log m\right]. \tag{8}$$

Analogous to Case 1, for $\mathbb{E}[h^r] = \max\{\mathbb{E}_1[h^r], \mathbb{E}_2[h^r]\}$, we can upper bound (8) by

$$\mathbb{P}\left[h^r - \mathbb{E}[h^r] \geq \frac{1}{2}\left(m^{-\gamma/(2s)} - 1\right)(1 - 2h^r)\right]. \tag{9}$$

Let $\epsilon = \left(1 - m^{\gamma/(2s)}\right)/2 \in (0, 1)$.

Case 2.1: $h^r - \mathbb{E}[h^r] \leq \epsilon(1 - 2\mathbb{E}[h^r])$. Then

$$(1 - 2h^r) \geq (1 - 2\mathbb{E}[h^r]) - 2\epsilon(1 - 2\mathbb{E}[h^r]) = (1 - 2\mathbb{E}[h^r])(1 - 2\epsilon).$$

Hence, (9) is upper bounded by

$$\mathbb{P}\left[h^r - \mathbb{E}[h^r] \geq \frac{1}{2}\left(m^{-\gamma/(2s)} - 1\right)(1 - 2\epsilon)(1 - 2\mathbb{E}[h^r])\right].$$

Thus, Corollary 5.3 yields the following upper bound for (8):

$$\exp\left(-\frac{k}{8}\left(m^{-\gamma/(2s)} - 1\right)^2 (1 - 2\epsilon)^2 (1 - 2\mathbb{E}[h^r])^2\right)$$
$$= \exp\left(-\frac{k}{8}\left(m^{-\gamma/(2s)} - 1\right)^2 m^{\gamma/s} (1 - 2\mathbb{E}[h^r])^2\right).$$

Case 2.2: $h^{ij} - \mathbb{E}[h^r] > \epsilon(1 - 2\mathbb{E}[h^r])$. Then, we apply Corollary 5.3 directly to obtain

$$\mathbb{P}\left[h^r - \mathbb{E}[h^r] \geq \epsilon(1 - 2\mathbb{E}[h^r])\right] \leq \exp\left(-\frac{k}{2}\epsilon^2 (1 - 2\mathbb{E}[h^r])^2\right).$$

Hence, in total, (8) is upper bounded by

$$2\exp\left(-\frac{k}{8}\left(1 - m^{\gamma/(2s)}\right)^2 (1 - 2\mathbb{E}[h^r])^2\right).$$

We conclude that the probability of failure for any fixed choice of γ is upper bounded by

$$\begin{aligned}
&\sum_{r\in\{ij,kl\}} 2\exp\left(-\frac{k}{8}\left(1-m^{\gamma/(2s)}\right)^2(1-2\mathbb{E}[h^r])^2\right)\\
&\quad+\sum_{r\in\{il,jk\}} 2\exp\left(-\frac{k}{8}\left(1-m^{(1-\gamma)/(2s)}\right)^2(1-2\mathbb{E}[h^r])^2\right)\\
&\quad+\sum_{r\in\{il,jk\}} 2\exp\left(-\frac{k}{8}\left(1-m^{\max\{\gamma,1-\gamma\}/(2s)}\right)^2(1-2\mathbb{E}[h^r])^2\right)\\
&\quad\leq 12\exp\left(-\frac{k}{8}\left(1-m^{\min\{\gamma,1-\gamma\}/(2s)}\right)^2(1-2E_{\max})^2\right).
\end{aligned}$$

If we upper bound this probability by a small $\epsilon > 0$, then we obtain bounds on γ:

$$\begin{aligned}
&12\exp\left(-\frac{k}{8}\left(1-m^{\min\{\gamma,1-\gamma\}/(2s)}\right)^2(1-2E_{\max})^2\right)\leq\epsilon\\
&\quad\Leftrightarrow\ m^{\min\{\gamma,1-\gamma\}/(2s)}\leq 1-\sqrt{\frac{8}{k}\ln\left(\frac{12}{\epsilon}\right)(1-2E_{\max})^{-2}}\\
&\quad\text{or } m^{\min\{\gamma,1-\gamma\}/(2s)}\geq 1+\sqrt{\frac{8}{k}\ln\left(\frac{12}{\epsilon}\right)(1-2E_{\max})^{-2}}.
\end{aligned}$$

Then for the constant $M := 8\ln(12/\epsilon)(1-2\mathbb{E}_{\max})^{-2} > 1$ and $k > M$, we use the fact that $\log m < 0$ to arrive at

$$\begin{aligned}
&\min\{\gamma,1-\gamma\}\geq\frac{2s}{\log m}\log\left(1-\sqrt{M/k}\right)\ \text{ or }\ \min\{\gamma,1-\gamma\}\\
&\quad\leq\frac{2s}{\log m}\log\left(1+\sqrt{M/k}\right)<0.
\end{aligned}$$

Thus, our claim follows. □

Finally, we arrive at our main result. Here, for our choice of function s at the beginning of the section, we denote $s_{\max}$ as the maximum value in the image of s. This value increases the closer inheritance probabilities to binary states. Hence, bounds on the inheritance probabilities yield bounds on $s_{\max}$, too.

Theorem 5.5. *Suppose k sites evolve under the CFN model on a phylogenetic level-1 network N, then the* DC *method returns $\tilde{N}$ with probability $1-o(1)$, if*

$$k > \frac{c\cdot\log n}{\left(1-\max\limits_{e\in E}\theta(e)^{\min\{\gamma_{\min},1-\gamma_{\max}\}/s_{\max}}\right)^2\left(\min\limits_{e\in E}\theta(e)\right)^{2(depth(N)+lc(N))}} \tag{10}$$

where c is a fixed constant, $\gamma_{\min}$ and $\gamma_{\max}$ are the minimum and maximum attained inheritance probability, respectively. Equivalently, the DC *method returns $\tilde{N}$ with probability $1 - o(1)$, if inequality (5) holds for all inheritance probabilities.*

Proof. We show that there exists a set of quartet profiles $\mathcal{Q}_w$ of width w such that $\mathcal{Q}_R(N) \subseteq \mathcal{Q}_w \subseteq \mathcal{Q}(N)$ and the L1QPC method returns $\mathcal{Q}_w$ with high probability. To this end, let

$$\tau \in \left(0, \frac{1}{6} \min_{e \in E} \theta(e)^{\mathrm{depth}(N)+lc(N)}\right),$$

$$S_\tau = \left\{\{x_i, x_j\} \,:\, h^{ij} < \frac{1}{2} - \tau\right\},$$

$$Z_\tau = \left\{\{x_i, x_j, x_k, x_l\} \subseteq \mathcal{X} \,:\, h^r < \frac{1}{2} - 2\tau \;\forall\, r \in \{ij, ik, il, jk, jl, kl\}\right\}.$$

Observe that $\{x_i, x_j\} \in S_{2\tau}$ for $x_i, x_j \in Q$, $Q \in Z_\tau$, and

$$\mathbb{P}\left[\exists\, w \in \{h^{ij} \,:\, 1 \le i, j \le n\} \,:\, \mathcal{Q}_R(N) \subseteq \mathcal{Q}_w \subseteq \mathcal{Q}(N)\right]$$
$$\ge \mathbb{P}\left[\mathcal{Q}_R(N) \subseteq Z_\tau \text{ and L1QPC returns the correct quartet profile for } Q \in Z_\tau\right].$$

Consider $x_i, x_j \in \mathcal{X}$.

Case 1: $\mathbb{E}[h^{ij}] \ge 1/2 - \tau$. Then from Corollary 5.3, it follows

$$\mathbb{P}\left[\{x_i, x_j\} \in S_{2\tau}\right] = \mathbb{P}\left[h^{ij} < \frac{1}{2} - 2\tau\right] \le \mathbb{P}\left[h^{ij} - \mathbb{E}[h^{ij}] \le -\tau\right] \le \exp\left(-\tau^2 k/2\right).$$

Case 2: $\mathbb{E}[h^{ij}] < 1/2 - 3\tau$. Then from Corollary 5.3, it follows

$$\mathbb{P}\left[\{x_i, x_j\} \notin S_{2\tau}\right] = \mathbb{P}\left[h^{ij} \ge \frac{1}{2} - 2\tau\right] \le \mathbb{P}\left[h^{ij} - \mathbb{E}[h^{ij}] \ge \tau\right] \le \exp\left(-\tau^2 k/2\right).$$

Hence, we conclude

$$\mathbb{P}\left[\{x_i, x_j\} \in S_{2\tau} \;\forall\, x_i, x_j \in \mathcal{X},\, \mathbb{E}[h^{ij}] < \frac{1}{2} - 3\tau\right]$$
$$+ \mathbb{P}\left[\{x_i, x_j\} \notin S_{2\tau} \;\forall\, x_i, x_j \in \mathcal{X},\, \mathbb{E}[h^{ij}] \ge \frac{1}{2} - \tau\right]$$
$$\ge 1 - 2 \cdot \binom{n}{2} \exp\left(-\tau^2 k/2\right). \tag{11}$$

As shorthand notation, we write $\mathbb{P}[A]$ for the sum of probabilities in (11). Assume that $\{x_i, x_j\} \in S_{2\tau}$ for all $x_i, x_j \in \mathcal{X}$ with $E[h^{ij}] < 1/2 - 3\tau$ in some displayed tree of N. This assumption is well founded due to bound (11) and implies

$$h^{ij}, E^{ij}_{\min} < \frac{1}{2}\left(1 - \min_{e \in E} \theta(e)^{\mathrm{depth}(N)+lc(N)}\right) \qquad \forall\, x_i, x_j \in \mathcal{X}.$$

Then $\mathcal{Q}_R(N) \subseteq Z_\tau$ by Lemma 4.4. Therefore,

$$\mathbb{P}\left[\mathcal{Q}_R(N) \subseteq Z_\tau \text{ and L1QPC returns the correct quartet profile for } Q \in Z_\tau\right]$$
$$\ge \mathbb{P}\left[\mathcal{Q}_R(N) \subseteq Z_\tau,\ \text{L1QPC returns the correct quartet profile for } Q \in Z_\tau\right.$$

and event A occurs]

$\geq \mathbb{P}$ [L1QPC returns the correct quartet profile for $Q \in Z_\tau$ and event A occurs] .

Now, let $\mathcal{Q}_{\max}$ denote the set of quartets $Q \subseteq \mathcal{X}$ such that $\max\{E[h^{ij}] : x_i, x_j \in Q\} < 1/2 - \tau$. Without loss of generality, L1QPC returns a nontrivial quartet profile $(x_i x_j x_k x_l)$ for Q. For a reticulation edge e in $N_{|\{x_i,x_j,x_k,x_l\}}$, $(x_i x_j x_k x_l)$ is the correct quartet profile with probability at least

$$1 - 12\exp\left(-\frac{k}{8}\left(1 - \max_{e' \in E} \theta(e')^{\min\{\gamma(e), 1-\gamma(e)\}/(2s(e))}\right)^2 \tau^2\right) \tag{12}$$

by Proposition 5.4. Otherwise, $(x_i x_j x_k x_l)$ is trivial and Theorem 8 in Ref. [14] applies which equates to setting $\gamma(e) = s(e) = 1$ in formula (12). Therefore, we can apply bound (11) to infer that $Z_r \subseteq \mathcal{Q}_{\max}$ with high probability. Hence,

$$\begin{aligned}
&\mathbb{P}\,[\text{L1QPC returns the correct quartet profile for } Q \in Z_\tau \text{ and event } A \text{ occurs}] \\
&\quad \geq \mathbb{P}\,[\text{L1QPC returns the correct quartet profile for } Q \in Q_{\max} \text{ and event } A \text{ occurs}] \\
&\quad \geq 1 - 12\exp\left(-\frac{k}{8}\left(1 - \max_{e \in E} \theta(e)^{\min\{\gamma_{\min}, 1-\gamma_{\max}\}/s_{\max}}\right)^2 \tau^2\right) - (n^2 - n)\exp\left(-\frac{k}{2}\tau^2\right)
\end{aligned}$$

where the second inequality follows from the Bonferonni inequality. Thus, there exists a constant c such that

$$\begin{aligned}
&\mathbb{P}\left[\exists\, w \in \{h^{ij} : 1 \leq i, j \leq n\} : Q_R(N) \subseteq Q_w \subseteq \mathcal{Q}(N)\right] \\
&\quad \geq 1 - n^c \exp\left(-\frac{k}{8}\left(1 - \max_{e \in E} \theta(e)^{\min\{\gamma_{\min}, 1-\gamma_{\max}\}/s_{\max}}\right)^2 \tau^2\right).
\end{aligned}$$

□

Corollary 5.6. *Suppose k sites evolve under the CFN model on a phylogenetic level-1 network N with fixed inheritance probabilities. Then for depth$(N) = \mathcal{O}(\log\log n)$, a constant upper bound on the number of edges in each cycle in N and constant bounds on the mutation probabilities, the* DC *method returns $\tilde{N}$ with probability $1 - o(1)$, if $k =$ polylog(n).*

Proof. Theorem 5.5 tells us that, for a fixed constant c,

$$\frac{c \cdot \log n}{\left(1 - \max_{e \in E} \theta(e)^{\min\{\gamma_{\min}, 1-\gamma_{\max}\}/s_{\max}}\right)^2 \left(\min_{e \in E} \theta(e)\right)^{2(\text{depth}(N) + lc(N))}} \tag{13}$$

is an upper bound on the given number of sites k to determine $\tilde{\mathcal{N}}$ with high probability. Since we assume that inheritance probabilities are fixed, $\gamma_{\min}$ and $\gamma_{\max}$ satisfy inequality (5). From Proposition 5.4, we know that the inheritance probabilities adhering to inequality (5) are fully characterized by parameters k, $E_{\max}$ and $\theta(e) = 1 - 2p(e)$, $e \in E(N)$. The latter two are bounded by constants because we assume that the mutation probabilities $p(e)$ admit constant bounds. Recall from the beginning of this section that $s_{\max}$ is a function of probabilities $p(e)$ and $\gamma(e)$, $e \in E(N)$, implying that $s_{\max}$ is constantly

bounded, too. Finally, we assume $\text{depth}(N) = \mathcal{O}(\log \log n)$ and that $\text{lc}(N)$ is constant. This means, for suitable constants a and b, bound (13) simplifies to

$$\frac{c \cdot \log n}{a^2 \cdot b^{\mathcal{O}(\log \log n)}} = \frac{c \cdot \log n}{a^2 \cdot (\log n)^{\mathcal{O}(1)}} = \mathcal{O}\left((\log n)^{c'}\right)$$

for some constant c'. □

Table 1 in the introduction broadens Corollary 5.6 to further conditions to derive lower bounds on the number of sites of a sequence alignment from Theorem 5.5. For example, if we assume that our measures $\text{depth}(N)$ and $\text{lc}(N)$ of the network topology are independent of the number of taxa, i.e., are bounded by constants, then Corollary 5.6 can be strengthened to only require a logarithmic sequence length. In other words, the number of sites sufficient to reconstruct the correct binary level-1 phylogenetic network is logarithmic in the number of taxa of the network. The same argument can be extended in the opposite direction: if $\text{depth}(N)$ and $\text{lc}(N)$ grow logarithmically in the number of taxa, then Corollary 5.6 can be relaxed to conclude that the sufficient sequence length is a polynomial in the number of taxa. Table 1 details further bounds in case we vary the bounds on mutation probabilities, too. Observe that the bounds on inheritance probabilities are solely defined by the bounds on mutation probabilities. Hence, to derive conclusions on the sufficient sequence length, we do not consider any further assumptions on the inheritance probabilities.

6. Discussion

In this manuscript, we have provided a first analysis of the amount of genomic data required to accurately reconstruct phylogenetic networks, without a homoplasy-free restriction. While our results are based on a fairly simple model of evolution and depend on bounds on several parameters, they offer an optimistic outlook, showing that the required sequence length scales logarithmically, polynomially, or polylogarithmically with the number of taxa. In particular, these results suggest that accurate inference of evolutionary histories in the presence of reticulate events is possible in certain cases, even with a restricted amount of genomic data available. Moreover, we hope to have established a necessary bridge between convergence questions in the extensively studied context of tree inference and their mostly unstudied counterparts in network inference. This connection opens the door to rigorous analyses of broader classes of networks and more complex evolutionary models. It may also lead to new insights into the behavior and reliability of existing software tools for network reconstruction.

Although the analysis of our DC method assumes a CFN model of evolution, the method itself can be applied much more generally, relying only on a distance matrix between taxa as input. Hence, as a first step, one might extend our analysis to related models of evolution, such as the Jukes-Cantor, Kimura-2-Parameter and Kimura-3-Parameter models, or possibly models allowing for

ILS (incomplete lineage sorting). Another direction could be to instead consider different types of distances that can be obtained from genomic data (see e.g. Refs. [3,4,42]).

Extending our results to networks more general than level-1 will require additional work on the algorithm itself. Although most of our proposed inference rules are valid for networks of any level, we have not established closure properties—a key component of our method—when going beyond level-1. In particular, we do not expect our current set of inference rules to suffice, since networks beyond level-1 are not all outer-labeled planar and some quartet profiles could contain all three possible quartet trees on four leaves, a case that is not covered by our current set of rules. A natural next step may be to first study the tree-of-blobs of a network—the tree obtained by contracting each blob to a single vertex—which could serve as a bridge to more general network classes. It must be noted that one cannot simply rely on existing theory for trees however, since a trivial quartet profile of a network is not necessarily induced by its tree-of-blobs (see also the distinction between *T-quartets* and *B-quartets* from Ref.[1]).

We end by noting that our inference rules naturally extend to semi-directed quarnets (4-leaf subnetworks). Closure properties for this case may be an interesting direction for future research (as studied for undirected quarnet inference rules in Ref. [22]).

Funding This work was partially supported by grant OCENW.M.21.306 of the Dutch Research Council (NWO).

Data Availability No datasets were generated or analyzed for this study. All information needed to evaluate the findings are contained within the manuscript.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A. Displaying Trees and Inducing Subnetworks are Commutative

In the following proposition, we prove that taking displayed trees and taking subnetworks are commutative operations. That is, the set of displayed trees of a subnetwork is the same as the set of induced subtrees of the displayed trees of the full network.

Proposition A.1. *Let N be a semi-directed network on $\mathcal{X}$ and let T be an unrooted phylogenetic tree on $\mathcal{Y} \subseteq \mathcal{X}$ with $|\mathcal{Y}| \geq 2$. Then T is a displayed tree of $N|_{\mathcal{Y}}$ if and only if T is a subtree induced by $\mathcal{Y}$ of some displayed tree on $\mathcal{X}$ of N.*

Proof. First note that each reticulation vertex r of $N|_{\mathcal{Y}}$ corresponds to a unique reticulation vertex r' in N. We denote the set of these reticulation vertices in N by R. Let P be the set of reticulation vertices of N not in R. Since we allow parallel edges, a reticulation vertex r of N is in R if and only if it is on an up-down path in N between two leaves in $\mathcal{Y}$.

Suppose that T is a displayed tree of $N|_{\mathcal{Y}}$. Without loss of generality, suppose we label the edges such that T can be obtained from $N|_{\mathcal{Y}}$ by deleting the reticulation edge e_i^1 for each pair of reticulation edges (e_i^1, e_i^2) (with reticulation vertex $r_i \in R$) in $N|_{\mathcal{Y}}$. Let T' be the displayed tree of N obtained by removing the edge e_i^1 for each $r_i' \in R$, and arbitrarily removing one reticulation edge for each reticulation vertex in P. Then T is an induced subtree $T'|_{\mathcal{Y}}$ of T', proving the forward implication.

For the backward implication, suppose that T is an induced subtree of some displayed tree T' of N. Without loss of generality, assume that we label the edges such that T' can be obtained from N by deleting the reticulation edge e_i^1 for each pair of reticulation edges (e_i^1, e_i^2) (with reticulation vertex $r_i \in R \cup P$) in N. Similar to before, T can then be obtained from $N|_{\mathcal{Y}}$ by removing the reticulation edge e_i^1 for each $r_i \in R$. This proves the other implication. □

B. Proof of Lemma 5.1

Recall, for a semi-directed level-1 network N under the CFN model, we have a well-defined function s which maps the reticulation edges e of N to the positive real numbers by equations

$$\frac{\gamma(e)}{1-\gamma(e)} \cdot \frac{\log \min_{e' \in E} \theta(e')}{\log \max_{e' \in E} \theta(e')} + \epsilon = \begin{cases} \frac{2s(e)^2 - 7s(e)}{2s(e)^2 + s(e) - 1} & \text{if } \frac{\gamma(e)}{1-\gamma(e)} \cdot \frac{\log \min_{e' \in E} \theta(e')}{\log \max_{e' \in E} \theta(e')} < 1 \\ \frac{4s(e)^3 + 4s(e)^2 - s(e) - 1}{12s(e)^2 + 4s(e) + 1} & \text{otherwise} \end{cases}$$

for a sufficiently small $\epsilon > 0$ such that $s(e) > 1/2$.

Lemma B.1. *Let N be a semi-directed level-1 network on leafset $\mathcal{Y} = \{x_i, x_j, x_k, x_l\}$ with $(x_i x_j x_k x_l) \in \mathcal{Q}_R(N)$, containing a cycle of length four and displayed trees T_1 and T_2. Let e be the reticulation edge from N present in T_1 and let d be the matrix of dissimilarity scores on $\mathcal{Y}$ defined by a sequence alignment on $\mathcal{Y}$. Then for the width w of $\mathcal{Y}$, the L1QPC method infers $(x_i x_j x_k x_l)$ from d if the inheritance probability $\gamma(e)$ satisfies $D = \gamma(e)D(T_1) + (1-\gamma(e))D(T_2)$ such that*

$$\max\{|d_r - D_r| : r \in \{ij, kl, ik, jl\}\} < -\frac{\gamma(e)}{4s(e)} \log \max_{e' \in E} \theta(e'), \tag{3}$$

$$\max\{|d_r - D_r| : r \in \{il, jk, ik, jl\}\} < \frac{\gamma(e)-1}{4s(e)} \log \max_{e' \in E} \theta(e'). \tag{4}$$

Proof. Let γ and s denote the inheritance probability and function $s(e)$ for the reticulation edge e from N present in T_1. Let $\ell_1, \ell_2, \ell_3, \ell_4$ be the edge lengths of the internal edges e_1, e_2, e_3, e_4 in N on paths from x_i to x_j, from x_j to x_k, from x_k to x_l and from x_l to x_i, respectively. Without loss of generality $e_3 \in E(T_1) \backslash E(T_2)$ and $e_4 \in E(T_2) \backslash E(T_1)$. Assume $\gamma \in (0,1)$ such that $D = \gamma D(T_1) + (1-\gamma)D(T_2)$ and conditions (3) and (4). Then $(x_i x_j x_k x_l) \in \mathcal{Q}_R(N)$ only if

$$\begin{aligned} D_{ik} + D_{jl} - (D_{ij} + D_{kl}) &= \gamma(\ell_1 + 2\ell_2 + \ell_3) + (1-\gamma)(2\ell_1 + \ell_2 + \ell_4) \\ &\quad - [\gamma(\ell_1 + \ell_3) + (1-\gamma)(2\ell_1 + \ell_2 + \ell_4)] \\ &= (2-\gamma)\ell_1 + (\gamma+1)\ell_2 + \gamma\ell_3 + (1-\gamma)\ell_4 \\ &\quad - [(2-\gamma)\ell_1 + (1-\gamma)\ell_2 + \gamma\ell_3 + (1-\gamma)\ell_4] = 2\gamma\ell_2, \end{aligned} \tag{14}$$

$$\begin{aligned} D_{ik} + D_{jl} - (D_{il} + D_{jk}) &= \gamma(\ell_1 + 2\ell_2 + \ell_3) + (1-\gamma)(2\ell_1 + \ell_2 + \ell_4) \\ &\quad - [\gamma(\ell_1 + 2\ell_2 + \ell_3) + (1-\gamma)(\ell_2 + \ell_4)] \\ &= (2-\gamma)\ell_1 + (\gamma+1)\ell_2 + \gamma\ell_3 + (1-\gamma)\ell_4 \\ &\quad - [\gamma\ell_1 + (\gamma+1)\ell_2 + \gamma\ell_3 + (1-\gamma)\ell_4] = 2(1-\gamma)\ell_1, \end{aligned} \tag{15}$$

$$\begin{aligned} D_{il} + D_{jk} - (D_{ij} + D_{kl}) &= \gamma(\ell_1 + 2\ell_2 + \ell_3) + (1-\gamma)(\ell_2 + \ell_4) \\ &\quad - [\gamma(\ell_1 + \ell_3) + (1-\gamma)(2\ell_1 + \ell_2 + \ell_4)] \\ &= 2\gamma\ell_2 - 2(1-\gamma)\ell_1. \end{aligned} \tag{16}$$

Without loss of generality, $d_{ij} + d_{kl} \leq d_{il} + d_{jk}$. Then our claim follows if we show that

$$d_{il} + d_{jk} - d_{ij} - d_{kl} < d_{ik} + d_{jl} - d_{il} - d_{jk}$$

Equivalently, we show that $2(d_{il} + d_{jk}) < d_{ik} + d_{jl} + d_{ij} + d_{kl}$. To this end, for $f = -\frac{1}{2}\log \max_{e \in E} \theta(e) \leq -\frac{1}{2}\log(1-2p(e))$, $e \in E$, we get

$$\begin{aligned} d_{ij} + d_{kl} &\overset{(3)}{>} D_{ij} + D_{kl} - \frac{\gamma}{s} f \\ &\overset{(16)}{=} D_{il} + D_{jk} - 2\gamma\ell_2 + 2(1-\gamma)\ell_1 - \frac{\gamma}{s} f \end{aligned}$$

$$\overset{(4)}{>} d_{il} + d_{jk} + \left(2 - \frac{1}{s}\right)(1-\gamma)\ell_1 - \left(2 + \frac{1}{s}\right)\gamma\ell_2.$$

Since $d_{ij} + d_{kl} \leq d_{il} + d_{jk}$, it follows that

$$\left(2 - \frac{1}{s}\right)(1-\gamma)\ell_1 < \left(2 + \frac{1}{s}\right)\gamma\ell_2. \tag{17}$$

Moreover, we have

$$\begin{aligned}
&d_{ik} + d_{jl} + d_{ij} + d_{kl} \\
&\overset{(3)}{>} D_{ik} + D_{jl} - \frac{\gamma}{s}f + D_{ij} + D_{kl} - \frac{\gamma}{s}f \\
&\overset{(15),(16)}{=} D_{il} + D_{jk} + 2(1-\gamma)\ell_1 + D_{il} + D_{jk} - 2\gamma\ell_2 + 2(1-\gamma)\ell_1 - \frac{2\gamma}{s}f \\
&= 2\left(D_{il} + D_{jk} + 2(1-\gamma)\ell_1 - \gamma\ell_2 - \frac{\gamma}{s}f\right) \\
&\geq 2\left(D_{il} + D_{jk} + \frac{1-\gamma}{s}f + \left(2 - \frac{1}{s}\right)(1-\gamma)\ell_1 - \left(1 + \frac{1}{s}\right)\gamma\ell_2\right) \\
&\overset{(4)}{>} 2\left(d_{il} + d_{jk} + \left(2 - \frac{1}{s}\right)(1-\gamma)\ell_1 - \left(1 + \frac{1}{s}\right)\gamma\ell_2\right).
\end{aligned}$$

Hence, it is left to show that

$$\left(2 - \frac{1}{s}\right)(1-\gamma)\ell_1 \geq \left(1 + \frac{1}{s}\right)\gamma\ell_2$$

Equivalently, we show that

$$\frac{\gamma}{1-\gamma} \cdot \frac{\ell_2}{\ell_1} \leq \frac{2s-1}{s+1}. \tag{18}$$

Since inequality (17) can be rewritten as

$$\frac{\gamma}{1-\gamma} \cdot \frac{\ell_2}{\ell_1} > \frac{2s-1}{2s+1} \text{ or } \frac{1-\gamma}{\gamma} \cdot \frac{\ell_1}{\ell_2} < \frac{2s+1}{2s-1},$$

we know that inequality (18) can hold only if

$$\frac{2s-1}{s+1} - \frac{2s-1}{2s+1} > 0.$$

In other words, we require $s > 1/2$. Moreover, we observe that inequality (18) holds if

$$\underbrace{\frac{\gamma}{1-\gamma} \cdot \frac{\ell_2}{\ell_1}}_{=:L(\gamma,\ell_1,\ell_2)} - \frac{1-\gamma}{\gamma} \cdot \frac{\ell_1}{\ell_2} \leq \underbrace{\frac{\gamma}{1-\gamma} \cdot \frac{\ell_2}{\ell_1} - \frac{2s-1}{2s+1}}_{>0} + \underbrace{\frac{2s-1}{s+1} - \frac{2s+1}{2s-1}}_{=:R(s)} \tag{19}$$

$$\Leftrightarrow \quad \frac{1}{L(\gamma,\ell_1,\ell_2)} \geq \frac{2s-1}{2s+1} - R(s). \tag{20}$$

Then for $L(\gamma,\ell_1,\ell_2) < 1$, we have

$$\lim_{s\to\infty} R(s) = \lim_{s\to\infty} \frac{2s^2 - 7s}{2s^2 + s - 1} = 1 > L(\gamma,\ell_1,\ell_2),$$

implying inequality (19). Otherwise, inequality (20) follows from

$$\lim_{s\to\infty} \frac{2s-1}{2s+1} - R(s) = \lim_{s\to\infty} \frac{12s^2+4s+1}{4s^3+4s^2-s-1} = 0 < \frac{1}{L(\gamma,\ell_1,\ell_2)}.$$

Thus, since

$$L(\gamma,\ell_1,\ell_2) \leq \frac{\gamma}{1-\gamma} \cdot \frac{\log \min_{e\in E} \theta(e)}{\log \max_{e\in E} \theta(e)},$$

our claim follows from the definition of s. □

References

[1] Elizabeth S Allman, Hector Baños, Jonathan D Mitchell, and John A Rhodes. The tree of blobs of a species network: identifiability under the coalescent. *Journal of Mathematical Biology*, 86(1):10, 2023.

[2] Elizabeth S Allman, Hector Baños, Jonathan D Mitchell, and John A Rhodes. TINNiK: inference of the tree of blobs of a species network under the coalescent model. *Algorithms for Molecular Biology*, 19(1):23, 2024.

[3] Elizabeth S Allman, Hector Baños, and John A Rhodes. Nanuq: a method for inferring species networks from gene trees under the coalescent model. *Algorithms for Molecular Biology*, 14(1):24, 2019.

[4] Elizabeth S Allman, Hector Baños, John A Rhodes, and Kristina Wicke. Nanuq$^+$: A divide-and-conquer approach to network estimation. *Algorithms for Molecular Biology*, 20(1):14, 2025.

[5] N Alon and J H Spencer. *The Probabilistic Method*. John Wiley & Sons, New York, 1992.

[6] Hans-Jürgen Bandelt and Andreas Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7(3):309–343, 1986.

[7] Hector Baños. Identifying species network features from gene tree quartets under the coalescent model. *Bulletin of Mathematical Biology*, 81:494–534, 2019.

[8] Eric Bapteste, Leo van Iersel, Axel Janke, Scot Kelchner, Steven Kelk, James O McInerney, David A Morrison, Luay Nakhleh, Mike Steel, Leen Stougie, et al. Networks: expanding evolutionary thinking. *Trends in Genetics*, 29(8):439–441, 2013.

[9] Peter Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50, 1974.

[10] Hans Colonius and Hans-Hennig Schulze. Tree structures for proximity data. *British Journal of Mathematical and Statistical Psychology*, 34(2):167–180, 1981.

[11] Mary Cryan, Leslie Ann Goldberg, and Paul W Goldberg. Evolutionary trees can be learned in polynomial time in the two-state general markov model. *SIAM Journal on Computing*, 31(2):375–397, 2001.

[12] Miklos Csűrös and Ming-Yang Kao. Recovering evolutionary trees through harmonic greedy triplets. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms, Baltimore, Md*, pages 17–19, 1999.

[13] Aviva K Englander, Martin Frohn, Elizabeth Gross, Niels Holtgrefe, Leo van Iersel, Mark Jones, and Seth Sullivant. Identifiability of phylogenetic level-2 networks under the jukes-cantor model. *bioRxiv*, pages 2025–04, 2025.

[14] Péter L Erdős, Michael A Steel, László A Székely, and Tandy J Warnow. A few logs suffice to build (almost) all trees (i). *Random Structures & Algorithms*, 14(2):153–184, 1999.

[15] Péter L Erdős, Michael A Steel, László A Székely, and Tandy J Warnow. A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science*, 221(1-2):77–118, 1999.

[16] Martin Frohn, Niels Holtgrefe, Leo van Iersel, Mark Jones, and Steven Kelk. Reconstructing semi-directed level-1 networks using few quarnets. *Journal of Computer and System Sciences*, 152:103655, 2025.

[17] Elizabeth Gross and Colby Long. Distinguishing phylogenetic networks. *SIAM Journal on Applied Algebra and Geometry*, 2(1):72–93, 2018.

[18] Dan Gusfield, Satish Eddhu, and Charles Langley. Efficient reconstruction of phylogenetic networks with constrained recombination. In *Computational Systems Bioinformatics. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 363–374. IEEE, 2003.

[19] Michael D Hendy and David Penny. Spectral analysis of phylogenetic data. *Journal of classification*, 10(1):5–24, 1993.

[20] Max Hill and Sebastien Roch. Lower bounds on the sample complexity of species tree estimation when substitution rates vary across loci. *Bulletin of Mathematical Biology*, 87(152), 2025.

[21] Niels Holtgrefe, Katharina T Huber, Leo van Iersel, Mark Jones, Samuel Martin, and Vincent Moulton. Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments. *Molecular Biology and Evolution*, 42(4):msaf067, 2025.

[22] Katharina T Huber, Vincent Moulton, Charles Semple, and Taoyang Wu. Quarnet inference rules for level-1 networks. *Bulletin of Mathematical Biology*, 80:2137–2153, 2018.

[23] Daniel H Huson, Scott M Nettles, and Tandy J Warnow. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology*, 6(3-4):369–386, 1999.

[24] Joshua A. Justison and Tracy A. Heath. Exploring the distribution of phylogenetic networks generated under a birth-death-hybridization process. *Bulletin of the Society of Systematic Biologists*, 2:1–22, 3 2024.

[25] JCM Keijsper and RA Pendavingh. Reconstructing a phylogenetic level-1 network from quartets. *Bulletin of Mathematical Biology*, 76(10):2517–2541, 2014.

[26] Valerie King, Li Zhang, and Yunhong Zhou. On the complexity of distance-based evolutionary tree. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, page 444, 2003.

[27] Sungsik Kong, Joan Carles Pons, Laura Kubatko, and Kristina Wicke. Classes of explicit phylogenetic networks and their biological and mathematical significance. *Journal of Mathematical Biology*, 84(6):47, 2022.

[28] Sungsik Kong, David L Swofford, and Laura S Kubatko. Inference of phylogenetic networks from sequence data using composite likelihood. *Systematic Biology*, 74(1):53–69, 2025.

[29] James A Lake. Reconstructing evolutionary trees from dna and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences*, 91(4):1455–1459, 1994.

[30] Samuel Martin, Niels Holtgrefe, Vincent Moulton, and Richard M Leggett. Algebraic invariants for inferring 4-leaf semi-directed phylogenetic networks. *Systematic Biology*, page syaf071, 2025.

[31] Elchanan Mossel. On the impossibility of reconstructing ancestral data and phylogenies. *Journal of Computational Biology*, 10(5):669–676, 2003.

[32] Elchanan Mossel. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, 356(6):2379–2404, 2004.

[33] Vincent Moulton and Taoyang Wu. Planar rooted phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1289–1297, 2022.

[34] Jerzy Neyman. Molecular studies of evolution: a source of novel statistical problems. In *Statistical Decision Theory and Related Topics*, pages 1–27. Elsevier, 1971.

[35] JMS Simoes Pereira. A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, 6(3):303–310, 1969.

[36] John A Rhodes, Hector Baños, Jingcheng Xu, and Cécile Ané. Identifying circular orders for blobs in phylogenetic networks. *Advances in Applied Mathematics*, 163:102804, 2025.

[37] Charles Semple and Mike Steel. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.

[38] Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3):e1005896, 2016.

[39] Michael Steel. Recovering a tree from the leaf colourations it generates under a markov model. *Applied Mathematics Letters*, 7(2):19–23, 1994.

[40] Tandy Warnow, Bernard ME Moret, and Katherine St. John. Absolute convergence: true trees from short sequences. In *Symposium on Discrete Algorithms:*

Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, volume 7, pages 186–195, 2001.

[41] Tandy Warnow, Yasamin Tabatabaee, and Steven N Evans. Advances in estimating level-1 phylogenetic networks from unrooted snps. *Journal of Computational Biology*, 32(1):3–27, 2025.

[42] Jingcheng Xu and Cécile Ané. Identifiability of local and global features of phylogenetic networks from average distances. *Journal of Mathematical Biology*, 86(1):12, 2023.

[43] Qiuyi Richard Zhang, Satish Rao, and Tandy Warnow. New absolute fast converging phylogeny estimation methods with improved scalability and accuracy. In *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*, pages 8–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2018.

Martin Frohn and Steven Kelk
Department of Advanced Computing Sciences
Maastricht University
6229 EN Maastricht
The Netherlands
e-mail: martin.frohn@maastrichtuniversity.nl

Niels Holtgrefe and Leo van Iersel
Delft Institute of Applied Mathematics
Delft University of Technology
2628 CD Delft
The Netherlands

Mark Jones
Department of Computer Science
Middlesex University
Hendon Town Hall, The Burroughs
London NW4 4BT
UK

Communicated by Simone Linz

Received: 27 November 2025.

Accepted: 29 April 2026.