Journal of Mathematical Biology (2023) 86:10
https://doi.org/10.1007/s00285-022-01838-9

Mathematical Biology

The tree of blobs of a species network: identifiability under the coalescent

Elizabeth S. Allman[1] · Hector Baños[2,3] · Jonathan D. Mitchell[1,4,5] · John A. Rhodes[1]

Received: 6 May 2022 / Revised: 31 August 2022 / Accepted: 17 November 2022

Abstract
Inference of species networks from genomic data under the Network Multispecies Coalescent Model is currently severely limited by heavy computational demands. It also remains unclear how complicated networks can be for consistent inference to be possible. As a step toward inferring a general species network, this work considers its tree of blobs, in which non-cut edges are contracted to nodes, so only tree-like relationships between the taxa are shown. An identifiability theorem, that most features of the unrooted tree of blobs can be determined from the distribution of gene quartet topologies, is established. This depends upon an analysis of gene quartet concordance factors under the model, together with a new combinatorial inference rule. The arguments for this theoretical result suggest a practical algorithm for tree of blobs inference, to be fully developed in a subsequent work.

Keywords Species network · Tree of blobs · Phylogenetics · Phylogenomics · Network multispecies coalescent model

Mathematics Subject Classification 92D15 · 92D20

✉ John A. Rhodes
j.rhodes@alaska.edu

1 Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99775, USA

2 Department of Biochemistry and Molecular Biology, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

3 Department of Mathematics and Statistics, Faculty of Science, Dalhousie University, Halifax, NS, Canada

4 School of Natural Sciences (Mathematics), University of Tasmania, Hobart, TAS 7001, Australia

5 ARC Centre of Excellence for Plant Success in Nature and Agriculture, University of Tasmania, Hobart, TAS 7001, Australia

Published online: 06 December 2022

1 Introduction

Methods for inference of evolutionary relationships between organisms are well-developed provided those relationships can be adequately described by a tree. If hybridization or some form of lateral gene transfer has occurred, tools for data analysis are much more limited. An essential complication is that when such gene transfer has occurred between closely related taxa, the population-genetic effect of incomplete lineage sorting is also likely. Thus individual gene relationships may conflict with the primary tree-like species relationships (if some can be considered to be primary) due to the intermixed effect of these two processes.

The appropriate stochastic model to capture these processes is the Network Multispecies Coalescent (NMSC). Under the NMSC combined with standard sequence substitution models, Bayesian methods for inference of species networks have been implemented: BEAST 2/SpeciesNetwork (Zhang et al. 2017), PhyloNet (Zhu et al. 2016, 2018), BPP (Flouri et al. 2019). However, they are limited by computational demands to small data sets of few taxa and few genes. Pseudolikelihood methods that treat inferred gene trees as data are able to handle larger data sets, PhyloNet (Yu and Nakhleh 2015) and SNaQ (Solís-Lemus and Ané 2016), but require prespecification of the number of reticulation events, with at best heuristic assessment of that number. In addition, to reduce computational effort, inference may be limited to the class of level-1 networks, though a biological justification for that may be lacking. A final approach starting with inferred gene trees combines statistical tests for small networks with combinatorial methods to assemble a large network, NANUQ (Allman et al. 2019). This is considerably faster and offers some insight into model fit, but also is currently limited to level-1 structure.

It is not known how complex a species network can be for its inference from specific data types to be even theoretically possible. This is the question of identifiability of the network (either topological or metric) under the NMSC model: Does the distribution of observations under the NMSC uniquely determine the network? The most complete result in the level-1 topological case comes from Baños' study of identifiability from quartet concordance factors (Baños 2019). Using different notions of data, however, several works have studied the identifiability question for general networks without the coalescent. Researchers have, for instance, investigated what can be determined from average intertaxon distances on a network (Xu and Ané 2021), as well as shortest distances and distance multisets (van Iersel et al. 2020). Identifiability from induced 4-taxon networks (Huber et al. 2018), rooted 3-taxon networks (Semple and Toft 2021), and counts of paths from interior nodes to taxa (Erdős et al. 2019) have also been explored, among other notions.

In this work we approach the network inference problem from a different direction, trying to determine only the tree-like evolutionary relationships for a collection of taxa, hence isolating the parts of their history when more complicated network features are formed. More formally, we study the *tree of blobs* of the network (Gusfield et al. 2007), a tree in which each group of edges in the network describing complex gene transfer, i.e., each blob, has been shrunk to a single node. (A closely related notion appears in Murakami et al. 2019). The tree of blobs thus shows all tree-like parts of

the network, and its inference could be useful to researchers who may subsequently focus on inferring the structure of each blob by other methods.

Our goal here is to show the topology of the unrooted tree of blobs for a network is identifiable from gene quartet data under the NMSC model. That is, the distribution of gene quartet topologies arising under the NMSC on a fixed species network uniquely determines the unrooted tree of blobs of that network. We make no assumptions on blob structure, but do require that numerical parameters lie outside an exceptional set of measure zero. Thus consistent inference of the tree of blobs is theoretically possible.

We first study the probabilities of quartets displayed across independent gene trees under the NMSC, under a generic assumption on numerical parameters. These probabilities—the quartet concordance factors (CFs)—allow for the identification of some sets of 4 taxa that must be collectively related through a blob, while proposing a resolved quartet tree topology for others. A new combinatorial inference rule is then developed that allows this information to be used to identify additional sets of four taxa related through a single blob, even though their CFs suggested otherwise. We show that repeated application of this rule yields all sets of four taxa with blob relationships. Then, with all such blob quartets known, and tree topologies assigned to other sets of four taxa, by treating blob quartets as unresolved we obtain complete information on all quartets displayed on the tree of blobs. This information is enough to determine the tree of blobs (Semple and Steel 2005; Rhodes 2020).

Although rules for inference of large networks from 4-taxon networks have been considered previously (Huber et al. 2018), our rule is different in purpose. It neither assumes knowledge of the full 4-taxon blob structure, nor attempts to infer detailed blob structure on a larger network. Earlier work on quartet closure rules for trees, surveyed in Grünewald and Huber (2007), is also similar in spirit to the rule developed here.

Our approach suggests an algorithm for tree of blobs inference that will be fully developed in a subsequent paper focused on data analysis. First a statistical test can be applied to gene quartet counts to detect blob and tree relationships on induced 4-taxon networks. Then the inference rule is applied repeatedly, until no new blob relationships on the full network are inferred. Finally, the quartet intertaxon distance (Rhodes 2020) is computed treating blob relationships as unresolved. A standard distance-based tree building algorithm, such as Neighbor-Joining (Saitou and Nei 1987), then yields an estimate of the tree of blobs. This is broadly similar to the steps in NANUQ (Allman et al. 2019) for inference of a level-1 network, but the inference rule step is new, and the distance, which in principle should fit a tree, does not require an analysis by NeighborNet (Bryant and Moulton 2004) or construction of a splits graph (Dress and Huson 2004).

Many methods have been developed for a more detailed detection of hybridization or gene transfer than the tree of blobs depicts, e.g. (Blischak et al. 2018; Green et al. 2010; Hamlin et al. 2020; Hibbins and Hahn 2022). Once the tree of blobs has been inferred for a collection of taxa, such methods might be applied to a subset of the taxa in order to explore the structure of a blob through a finer analysis. Unfortunately, these methods are generally restricted to a small number of taxa, and simple scenarios (e.g., level-1). Much work remains to be done to both expand the scope of methodology

for inferring blob structure, and to delineate both theoretical and practical limits to its inference.

Our presentation is structured as follows. Section 2 provides basic definitions and background on the NMSC model. In Sect. 3 we prove the fundamental result that from quartet concordance factors under the NMSC on a 4-taxon network one can determine whether the taxa are related through a single blob (i.e., a 4-blob), or not. If not, then all displayed trees on the 4-network have the same tree topology, which can also be determined. Establishing these facts requires an analysis based in the NMSC model. In Sect. 4, we use combinatorial arguments to show that from such information on the 4-taxon induced subnetworks of a larger network we can, through certain inference rules, gain information on all larger blobs. Section 5 quickly completes the argument for identifiability, and sketches the algorithm for tree of blobs inference suggested by the proof.

2 Networks and models

2.1 Phylogenetic networks

The Network Multispecies Coalescent model of gene tree formation within a species network underlies this work, so we give an appropriate definition of a phylogenetic network for that model.

Definition 1 Solís-Lemus and Ané (2016), Baños (2019) A *topological rooted binary phylogenetic network* $\mathcal{N}^+$ on taxon set X is a connected directed acyclic graph with nodes V and edges E, where V is the disjoint union $V = \{r\} \sqcup V_L \sqcup V_H \sqcup V_T$ and E is the disjoint union $E = E_H \sqcup E_T$, together with a bijective leaf-labeling function $f : V_L \to X$ with the following characteristics:

1. The *root* r has in-degree 0 and out-degree 2.
2. A *leaf* $v \in V_L$ has in-degree 1 and out-degree 0.
3. A *tree node* $v \in V_T$ has in-degree 1 and out-degree 2.
4. A *hybrid node* $v \in V_H$ has in-degree 2 and out-degree 1.
5. A *hybrid edge* $e = (v, w) \in E_H$ is an edge whose child node w is hybrid.
6. A *tree edge* $e = (v, w) \in E_T$ is an edge whose child node w is either a tree node or a leaf.

See Fig. 1L for an example of a rooted binary phylogenetic network. In that figure, and in others throughout this work, red indicates hybrid nodes and the hybrid edges leading to them.

Definition 2 A *cut edge* in a graph is one whose deletion increases the number of connected components.

Note that the notions of cut and non-cut edges are not the same as tree and hybrid edges. Although a hybrid edge is never a cut edge, tree edges may or may not be cut edges. For instance, in Fig. 1L, the child edges of v are both tree edges and non-cut, while the parent edge of v is tree and cut.

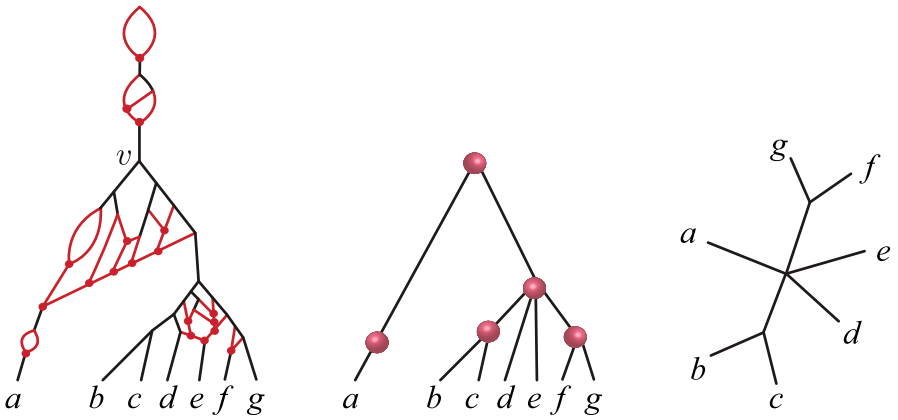

Fig. 1 **L** A species network $\mathcal{N}^+$, with edge lengths in coalescent units. Red indicates hybrid nodes and hybrid edges. The lowest stable ancestor (LSA) of the network is v. This network has 6 non-trivial blobs (a 5-blob, two 3-blobs, and three 2-blobs), and a single trivial 3-blob. **C** The tree-like structure of the LSA network $\mathcal{N}^{\oplus}$, obtained by deleting parts of the network above the LSA v, and showing blobs as red spheres. A sphere is used to suggest an unknown and potentially complicated blob structure. **R** The reduced unrooted tree of blobs, $T_{rd}(\mathcal{N}^-)$, obtained by shrinking blobs in the LSA network to nodes, unrooting, and suppressing degree-2 nodes (color figure online)

Edge directions on a rooted phylogenetic network induce a partial order on its nodes. We say that a node u is *above* or *ancestral to* a node v, or v is *below* or *descended from* u, if there is a directed path in the network from u to v. Thus the root is above all other nodes. We use the same terms to refer to similar relationships between edges, or between edges and nodes.

A topological network is one parameter of the NMSC model. Additional numerical parameters are introduced by giving the network a metric structure. Edge lengths are measured in *coalescent units* (units of generations/population size). In addition, we specify probabilities that a gene lineage at a hybrid node follows one or another hybrid edge as it traces back in time toward the network root.

Definition 3 A *metric rooted binary phylogenetic network* $(\mathcal{N}^+, \{\ell_e\}_{e\in E}, \{\gamma_e\}_{e\in E_H})$ is a topological rooted binary phylogenetic network together with an assignment of weights or *lengths* ℓ_e to all edges and *hybridization parameters* γ_e to all hybrid edges subject to the following restrictions:

1. The length ℓ_e of a tree edge $e \in E_T$ is positive.
2. The length ℓ_e of a hybrid edge $e \in E_H$ is non-negative.
3. The hybridization parameters γ_e and $\gamma_{e'}$ for a pair of hybrid edges $e, e' \in E_H$ with the same child hybrid node are positive and sum to 1.

Our use of the term hybridization parameter does not imply the NMSC model only applies to describing hybridization in any strict biological sense; it is simply a convenient shorthand for a parameter quantifying gene flow. In some works these parameters are called *inheritance probabilities* (Solís-Lemus and Ané 2016).

Note that we require tree edges to have positive length, since lengths of zero would effectively allow networks to be non-binary. Since zero lengths are non-generic in

the parameter space, our formal statements of results holding for generic parameters would need no modification if they were allowed, though perhaps they would be more open to misinterpretation. We do explicitly allow hybrid edges to have length 0, to model possibly instantaneous jumping of a lineage from one population to another. A careful reading of our arguments shows that while such values are also non-generic, they do not lead to additional points in the exceptional set of non-generic points where our claims fail.

The following analog of the most recent common ancestor of taxa on a tree is needed.

Definition 4 (Steel 2016) Let $\mathcal{N}^+$ be a (metric or topological) rooted binary phylogenetic network on X and let $Z \subset V$ be any nonempty subset of the nodes of $\mathcal{N}^+$. Let D be the set of nodes which lie on every directed path from the root r of $\mathcal{N}^+$ to any $z \in Z$. Then the *lowest stable ancestor (LSA) of* Z *on* $\mathcal{N}^+$, denoted $LSA(Z, \mathcal{N}^+)$, is the unique node $v \in D$ such that v is below all $u \in D$ with $u \neq v$.

The LSA of the network, $LSA(\mathcal{N}^+)$, is the LSA of its leaves, $LSA(V_L, \mathcal{N}^+)$.

As shown in Fig. 1L, a rooted phylogenetic network may have a complex structure above its LSA. (If the network is level-1, this is a chain of 2-cycles, as discussed in Banos 2019). Since our methods based on gene quartets do not give us any information about structure above the LSA, we focus only on the structure below the LSA, sometimes with edge direction information lost.

To formalize this, by suppressing a node with both in- and out-degree 1 in a directed graph we mean replacing it and its two incident edges with a single edge from its parent to its child. Suppressing a degree-2 node between two undirected edges means replacing it and its two incident edges with a single undirected edge. Suppressing a node between an undirected edge and a directed out-edge means replacing it and its two incident edges with a single edge with the out-edge direction. Suppressing a node between a directed in-edge and an undirected edge means replacing it and its two incident edges with a single undirected edge. In all these situations, for a metric graph the new edge is assigned a length equal to the sum of lengths of the two replaced. If the out-edge was hybrid, the new edge is also hybrid and retains the hybridization parameter.

Definition 5 (Baños 2019) Let $\mathcal{N}^+$ be a (metric or topological) rooted binary phylogenetic network on X.

1. The *LSA network* $\mathcal{N}^{\oplus}$ induced from $\mathcal{N}^+$ is the network obtained by deleting all edges and nodes above LSA($\mathcal{N}^+$), and designating LSA($\mathcal{N}^+$) as the root node.
2. The *semidirected unrooted network* $\mathcal{N}^-$ is the unrooted network obtained from the LSA network $\mathcal{N}^{\oplus}$ by undirecting all tree edges and suppressing the root, but retaining directions of hybrid edges.

We often need to pass to a network on a subset of taxa from one on a larger set.

Definition 6 Let $\mathcal{N}^+$ be a (metric or topological) rooted binary phylogenetic network on X and let $Y \subset X$. The *induced rooted binary network* $\mathcal{N}^+_Y$ on Y is the network obtained from $\mathcal{N}^+$ by 1) retaining only those nodes and edges ancestral to one or more

taxa in Y, 2) suppressing all nodes with both in- and out-degree 1, and 3) if the root then has outdegree 1, removing it and its descendant edge and designating its child as the root. We then say $\mathcal{N}^+$ *displays* $\mathcal{N}_Y^+$.

2.2 Cycles, blobs, and quartets

Since rooted phylogenetic networks are acyclic by definition, we use the word *cycle* to refer to a sequence of edges in the network which forms a cycle when all edges are undirected. A *k-cycle* is a cycle composed of k edges.

Although we focus on phylogenetic networks, the following definition applies more broadly.

Definition 7 A *blob* on a network is a maximal connected subnetwork that has no cut edges. A blob is *trivial* if it consists of a single node. An edge in the network is said to be *incident* to a blob if exactly one of its incident nodes is in the blob. A blob has *degree m* or is an *m-blob* if (a) it has has exactly m cut edges incident to it and the network's root is not in the blob, or (b) it has exactly $m - 1$ cut edges incident to it and the root is in the blob.

We define an m-blob in this way for two reasons: First, it results in the degree of the blob containing the LSA not changing in passing from a rooted network $\mathcal{N}^+$ to its LSA network $\mathcal{N}^{\oplus}$. Second, the NMSC model considers an "above the root" population of infinite duration in which lineages may coalesce. This is essentially an additional edge, of infinite length, incident to the root. In our terminology if the root of a binary network is a trivial blob, then it is a degree-2 node but forms a degree-3 blob.

A network's blobs can equivalently be defined as the 2-edge-connected components (Xu and Ané 2021), or as the connected components obtained by deleting all cut edges in the network.

On a rooted binary phylogenetic tree, leaves are the only 1-blobs, while the root and internal nodes are 3-blobs. On a non-binary tree, polytomous nodes are k-blobs with $k \geq 4$. Non-tree phylogenetic networks may have k-blobs that are not nodes for any $k > 1$. The simplest blobs have the form of cycles, and a network with only such blobs is level-1. In general, however, blob structure may be much more complicated, with a few simple examples shown in Figs. 1L and 2.

As is well known, on a tree any 3 taxa determine a unique node where undirected paths between each pair of taxa meet, or equivalently a node whose deletion leaves the taxa in distinct connected components. If the tree is not binary, larger sets of taxa may or may not determine a node in this way. The following definition formalizes a similar notion for networks.

Definition 8 A blob is *determined by* a set of leaf labels S with $|S| \geq 3$ if deletion of the cut edges incident to the blob leaves the elements of S in $|S|$ distinct connected components.

On a network $\mathcal{N}^+$ every subset of 3 taxa determines a blob, and every m-blob with $m \geq 3$ that is below the LSA of $\mathcal{N}^+$ is determined by one or more subsets of 3

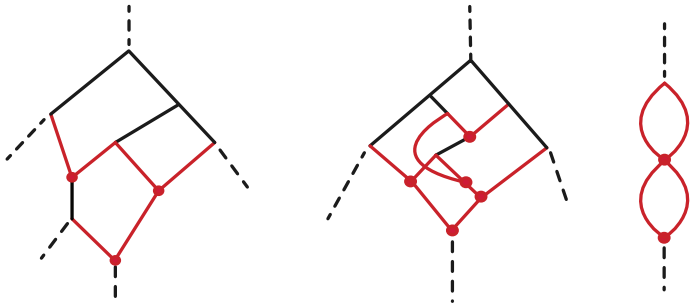

Fig. 2 Examples of blobs in networks. Red indicates hybrid nodes, and hybrid edges above them. Cut edges incident to the blobs are represented by dotted line segments: **L** a planar 5-blob, **C** a non-planar 4-blob, **R** a single 2-blob in a non-binary network, formed from two 2-cycles sharing a single node (color figure online)

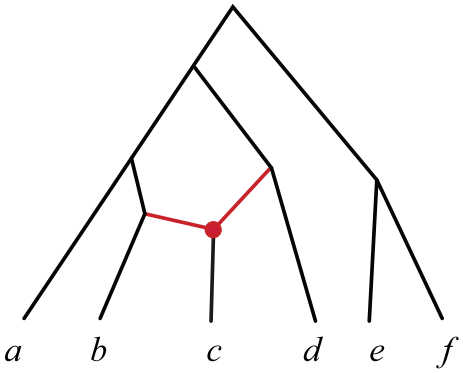

Fig. 3 A network with a 5-blob determined by the sets $\{a, b, c\}$, $\{a, b, d, f\}$, and other sets. The set $\{a, b, e, f\}$, however, does not determine a blob. Both $\{a, b, c, d\}$ and $\{a, b, d, e\}$ are B-quartets on this network. While $\{a, b, c, d\}$ is also a B-quartet on its induced 4-taxon network, $\{a, b, d, e\}$ is a T-quartet on its induced 4-taxon network

taxa. Blobs above the LSA are not determined by any subset of taxa, while an m-blob containing the LSA is determined by 3 taxa if $m \geq 4$.

A set of $k \geq 4$ taxa may or may not determine a blob, but if it does it must be an m-blob with $m \geq k$. For instance, the network of Fig. 3 has a 5-blob determined by the sets $\{a, b, c\}$, $\{a, b, d, f\}$, and others. The set $\{a, b, e, f\}$, however, does not determine a blob.

Note that our definition of blob differs slightly from that given by Gusfield et al. (2007), in which a blob is a maximal set of edges formed by recursively including cycles sharing at least one edge with an earlier cycle. By that definition, if two cycles share only a node as in Fig. 2R, they would be considered to be 2 distinct blobs. In contrast, they form a single blob under our definition. In Gusfield et al. (2007), this situation is handled by inserting an edge to separate two such cycles, joining each at the node they formerly shared, thus making edge-disjoint cycles also node-disjoint. Our restriction to binary networks rules out this possibility regardless.

Definition 9 A *chain of blobs* in a network is a subnetwork composed of a sequence of 1- and 2-blobs connected by their incident edges.

This notion generalizes the chain of 2-cycles defined for level-1 networks in Baños (2019). A chain of blobs will have 1- or 2-blobs at its ends, but all other blobs in

the chain will be 2-blobs. Just as a level-1 phylogenetic network may have a chain of 2-cycles above its LSA, a general phylogenetic network will have a (possibly empty) chain of blobs as the subnetwork between its root and LSA, as in Fig. 1(L).

Definition 10 (Gusfield et al. 2007) The *tree of blobs*, $T(\mathcal{N})$, for a general connected network, $\mathcal{N}$, is the tree obtained by contracting each blob to a node, that is, by removing all of the blob's edges and identifying all its nodes. If the network is rooted, the tree of blobs remains rooted at the same node, or the one arising from identifying the original root with other nodes.

An equivalent construction of a blob tree in Xu and Ané (2021) has nodes for each blob in $\mathcal{N}$, with edges connecting them if there is an edge with endpoints in the two blobs in $\mathcal{N}$.

The tree of blobs is generally not binary, even when the network is. A blob with m incident cut edges in a network produces an m-multifurcation in its tree of blobs. Nodes of degree 4 or more in the tree of blobs indicate non-trivial blobs for a binary network, while those of degree 2 or 3 may correspond to trivial or non-trivial blobs in the network.

While this definition of a tree of blobs applies to an arbitrary connected network, a slight variant is more useful here, as only some of the features of the tree of blobs for a species network may be identified by our methods and data. Any 2-blobs become nodes of degree 2 in the tree of blobs, but we will suppress these since we cannot detect them. Also, while we cannot detect any structure between the root of the network and its LSA, even after suppressing nodes of degree 2 arising from blobs above the LSA, an undetectable edge above the LSA might remain. We therefore discard this as well.

Definition 11 The *reduced rooted tree of blobs*, $T_{rd}(\mathcal{N}^+)$, of a rooted phylogenetic network $\mathcal{N}^+$ is obtained from the tree of blobs $T(\mathcal{N}^{\oplus})$ of the LSA network by suppressing all nodes of both in-degree 1 and out-degree 1. The *reduced unrooted tree of blobs*, $T_{rd}(\mathcal{N}^-)$, of $\mathcal{N}^+$ is obtained from the tree of blobs $T(\mathcal{N}^-)$ of the unrooted semidirected network by suppressing all nodes of degree 2.

See Fig. 1 for an example of a network and its reduced unrooted tree of blobs. The reduced unrooted tree of blobs $T_{rd}(\mathcal{N}^-)$ is undirected since the only directed edges in $\mathcal{N}^-$ are hybrid edges, which are in blobs, and thus lost when passing to its tree of blobs.

The reduced unrooted tree of blobs $T_{rd}(\mathcal{N}^-)$ can also be obtained from the rooted one $T_{rd}(\mathcal{N}^+)$ by undirecting all edges, and either suppressing the root if it has degree 2 (as a node) or dropping its designation as the root if it has larger degree.

Note that if the LSA of the original phylogenetic network $\mathcal{N}^+$ lies in an m-blob, $m \geq 4$, that blob gives only an $(m-1)$-multifurcation in the reduced unrooted tree of blobs. If the LSA lies in a 3-blob, then that blob will be completely suppressed, and not represented by a node.

We next introduce terminology to express the relationships a set of four taxa might have to the blob structure of a network. We follow the standard convention of using the word *quartet* to mean a particular unrooted binary topological tree on four taxa. For instance the quartet $ab|cd$ is the topology with cherries $\{a, b\}$ and $\{c, d\}$ separated

by an internal edge. The *unresolved quartet* is the star topology for the 4-taxon tree, denoted $abcd$. The following additional terminology is also useful in the network setting.

Definition 12 A set of four taxa $Q = \{a, b, c, d\}$ on an n-taxon phylogenetic network is a *Blob quartet*, or *B-quartet*, if there is a blob on the network which is determined by Q.

Equivalent conditions for $Q = \{a, b, c, d\}$ being a B-quartet are (1) the deletion of all edges in a single blob leaves the elements of Q in four distinct connected components, and (2) the unresolved quartet $abcd$ is displayed on the tree of blobs $T_{rd}(\mathcal{N}^-)$. The blob referred to here may be an m-blob for any $m \geq 4$.

If $\{a, b, c, d\}$ is not a B-quartet, then in the tree of blobs there must be an edge whose deletion disconnects two of these taxa from the others. Consequently, the tree of blobs displays a resolved quartet tree for these taxa.

Definition 13 If a set of four taxa is not a B-quartet on an n-taxon phylogenetic network, $n \geq 4$, then it is a *tree-like quartet*, or *T-quartet*. The resolved quartet *associated* to a T-quartet is that displayed on the tree of blobs $T_{rd}(\mathcal{N}^-)$.

Note that the induced 4-taxon network on a T-quartet need not be a tree, since the induced network on the four taxa may contain non-trivial 2-blobs and 3-blobs. However there can be no larger blobs. Nonetheless this induced network is "tree-like" in the sense that it will have a cut edge whose removal disconnects the four taxa into two groups of 2. Equivalently, every tree displayed on the 4-taxon network has the same resolved quartet topology. Thus any T-quartet on a large network is also a T-quartet on the induced quartet network.

In contrast, the induced network on a B-quartet may or may not have a 4-blob, and can even be a tree. In passing from a network to an induced network on fewer taxa, blobs may split into smaller blobs, and in some cases reduce to tree-like relationships. Indeed, this happens even in the level-1 case with a single cycle of k edges, $k \geq 5$. In Fig. 3, for instance, $\{a, b, d, e\}$ is a B-quartet in the full network, yet becomes a T-quartet on the induced 4-taxon network. However, $\{a, b, c, f\}$ is a B-quartet on both the full and the induced networks.

2.3 Coalescent model on networks and quartet concordance factors

The formation of gene trees, tracking the ancestral relationships of individual lineages within populations of ancestral species, is governed not only by the relationships of those species, but also population-genetic effects. Going backwards in time, these lead to gene lineages merging not when they first enter a common ancestral species, but rather further in the past. If they fail to merge before entering an ancestral population with yet other lineages, the gene tree relationships that form may differ from the species relationships. When the species relationships are described by a tree rooted at a common ancestor, the *multispecies coalescent (MSC) model* is the standard probabilistic description of gene tree formation capturing this process (Pamilo and Nei 1988; Liu et al. 2009).

The *network multispecies coalescent (NMSC) model* (Meng and Kubatko 2009; Yu et al. 2012; Zhu et al. 2016) generalizes the MSC, allowing a finite number of hybridization events, or other discrete lateral gene transfer events, between ancestral populations. Its parameters are captured by a metric, rooted phylogenetic network, assumed here to be binary, as in Definition 1. Edge lengths are given in coalescent units (computed as number of generations/population size), so that the rate of coalescence between two lineages is 1. At a hybrid node in the network, a gene lineage may pass into either of two ancestral populations, with probabilities given by the hybridization parameters $\gamma, 1-\gamma$ for the hybrid edges. This differs from other generalizations of the MSC, such as the structured coalescent, where gene flow may be continuous over a time interval.

The NMSC model determines a distribution of binary metric gene trees, and, through marginalization, distributions of binary topological gene trees on subsets of taxa. In this work we use only one type of marginalization, to unrooted binary topological gene trees on subsets of four taxa, or *gene quartets*. The probability of a gene quartet is thus a function of the metric species network parameters under the NMSC. Formulas for these probabilities were obtained in the tree case in Allman et al. (2011), and for level-1 networks in Solís-Lemus and Ané (2016), with further study in Baños (2019). Here we do not restrict to level-1 networks, and without any assumptions on blob structure one cannot obtain precise formulas for gene quartet probabilities. Nonetheless, some features of these probabilities can be analyzed sufficiently for application to determining the tree of blobs of the network.

Definition 14 Let $\mathcal{N}^+$ be a metric rooted binary phylogenetic network on a taxon set X, and $a, b, c, d \in X$ distinct taxa. Then for the gene quartet $ab|cd$, the *quartet concordance factor* $CF_{ab|cd} = CF_{ab|cd}(\mathcal{N}^+)$ is the probability under the NMSC on $\mathcal{N}^+$ that a gene tree displays the quartet $ab|cd$. The *quartet concordance factor for taxa* a, b, c, d, or more simply the *concordance factor*, is the ordered triple

$$CF_{abcd} = CF_{abcd}(\mathcal{N}^+) = (CF_{ab|cd}, CF_{ac|bd}, CF_{ad|bc})$$

of concordance factors of each quartet on the taxa.

Since under the NMSC gene trees are binary, and all gene tree topologies have positive probability, the entries of CF_{abcd} are positive and sum to 1. Note that permuting a, b, c, d permutes the entries of CF_{abcd}. Nonetheless, when a, b, c, d are clear from context, such as when $|X| = 4$, we write CF for CF_{abcd}.

In Allman et al. (2011) it was shown that if the species network is a tree then two of the three entries of CF_{abcd} must be equal, with the third no smaller. We need the following broader notion.

Definition 15 The concordance factor CF_{abcd} is a *cut* CF if two of its entries are equal, and *strictly cut* if in addition the third is distinct. If CF_{abcd} is strictly cut with $CF_{ab|cd} \neq CF_{ac|bd} = CF_{ad|bc}$, then we say CF_{abcd} is *strictly* $(ab|cd)$*-cut*. If CF_{abcd} is not cut, we say it is *non-cut*.

The term “cut" is motivated by Theorem 1 of the next section, which states that for generic parameters a CF is cut exactly when there is a cut edge in the 4-taxon

network whose deletion from the network leaves two connected components each with two taxa.

We emphasize that Definitions 12 and 13 of B- and T-quartets refer to the relationship of 4 taxa through the topology of a specified network, while Definition 15 of cut and non-cut CFs refers to properties of the probability distribution under the NMSC. In passing to an induced network, B-quartets may become T-quartets, although CFs remain unchanged.

Theorem 1 below shows that on 4-taxon networks there is a close correspondence between B-quartets and non-cut CFs. However, these notions are more subtly related on larger networks. For the network of Fig. 3, for instance, $\{a, b, d, e\}$ is a B-quartet yet has a strictly cut CF. This issue is the main obstacle to showing identifiability of the tree of blobs, to be overcome with Theorem 2 below.

3 Blob quartet identifiability on 4-networks

We work under the NMSC model, so that specification of model parameters through a metric rooted binary phylogenetic network determines a distribution of n-taxon gene trees, and by marginalization, the theoretical quartet CFs for each subset of four taxa.

Although our ultimate goal is to identify the reduced unrooted tree of blobs of a rooted phylogenetic network from the CFs, with no assumption on level or other particular network structure, our approach to doing this is by first determining B-quartets. In this section we show that by applying certain inference rules, all B-quartets on 4-taxon networks can be identified from the CFs, assuming generic values of numerical parameters.

By *generic* numerical parameters we mean all those that lie outside of a subset of measure zero in the parameter space. While we do not give an explicit description of such an exceptional set, a good intuitive description that it has measure zero is that if parameter values were chosen at random from an absolutely continuous distribution, then with probability 1 they would not be exceptional. For complex stochastic models it is quite common for identifiability results to depend upon the exclusion of some "small" exceptional subsets of the parameter space (Allman et al. 2009).

A basic combinatorial observation, whose proof we omit, is the following.

Lemma 1 *Let $\mathcal{N}^+$ be a 4-taxon rooted binary phylogenetic network. Then the semidirected unrooted network $\mathcal{N}^-$ must have either*

1. *exactly one 4-blob, or*
2. *exactly two 3-blobs.*

In either case, $\mathcal{N}^-$ may have any number of 2-blobs, but no other non-leaf blobs. In case 1, the reduced unrooted tree of blobs $T_{rd}(\mathcal{N}^-)$ is the unresolved quartet tree and the taxa form a B-quartet. In case 2, $T_{rd}(\mathcal{N}^-)$ is a resolved quartet tree and the taxa form a T-quartet.

As shown in Solís-Lemus and Ané (2016), Baños (2019), for generic parameters on a 4-taxon level-1 network one can detect B-quartets directly from the single CF. We next extend the 4-taxon result for level-1 networks to arbitrary 4-blobs on 4-taxon networks.

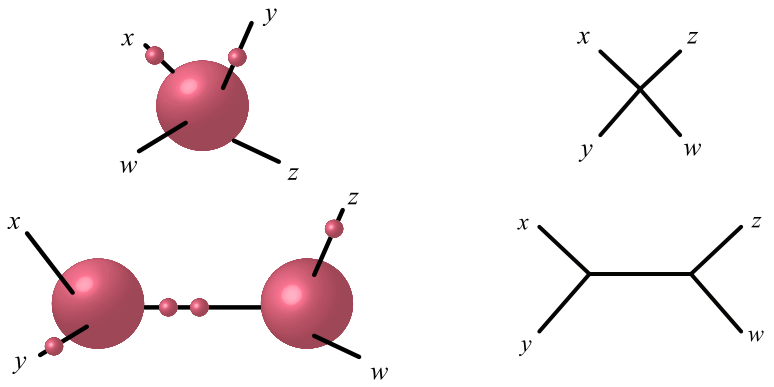

Fig. 4 **L** Schematic depictions of two semidirected unrooted 4-taxon networks $\mathcal{N}^-$, where spheres represent blobs of unspecified structure, and **R** their reduced unrooted trees of blobs $T_{rd}(\mathcal{N}^-)$. Up to taxon labelling, these are the only possible 4-taxon topological reduced unrooted trees of blobs

As illustrated in Fig. 4, we can determine the reduced unrooted tree of blobs of a 4-taxon network by determining if it has a cut edge inducing a non-trivial split. If such a cut edge exists, the tree of blobs is a quartet tree, and if it does not, the tree of blobs is a star tree. That this feature can be detected by quartet concordance factors is the content of the next proposition.

Theorem 1 (CF-detectability of 4-blobs on 4-taxon networks) *Consider a 4-taxon rooted binary phylogenetic network $\mathcal{N}^+$ on taxa $\{a, b, c, d\}$ with quartet concordance factor $CF = CF_{abcd}$ and reduced unrooted tree of blobs $T = T_{rd}(\mathcal{N}^-)$. Then under the NMSC for generic parameters:*

1. *T has the quartet tree topology $ab|cd$ if, and only if, CF_{abcd} is strictly $(ab|cd)$-cut.*
2. *T has the unresolved quartet topology if, and only if, CF_{abcd} is non-cut.*

Proof We prove the following statements, for generic parameters:

(a) If T has the quartet tree topology $ab|cd$, then CF_{abcd} is strictly $(ab|cd)$-cut.
(b) If T has the unresolved quartet topology, then CF_{abcd} is non-cut.

Were it not for the distinction between "cut" and "strictly cut", these statements would immediately yield claims 1 and 2. But since the parameters are assumed to be generic, this issue is easily overcome: Statement (b) implies for generic parameters that if CF_{abcd} is cut, then T has a resolved tree topology, which by (a) implies that CF_{abcd} is strictly cut. Thus for generic parameters CF_{abcd} is cut if, and only if, it is strictly cut.

To establish (a), suppose T is resolved, with topology $ab|cd$. Permuting taxon names if necessary, we may assume that the reduced rooted tree of blobs thus has topology $(((a, b), c), d)$, $((a, b), (c, d))$, or $((a, b), c, d)$.

In the first case, $(((a, b), c), d)$, if a gene tree forms under the NMSC by the a, b lineages coalescing below the 3-blob determined by a, c, d, it contributes to the frequency of unrooted gene quartets with topology $ab|cd$. Otherwise, a, b enter that blob as exchangeable lineages, and $ac|bd$ and $ad|bc$ will be equally probable as

unrooted gene quartets. Thus CF is $(ab|cd)$-cut for all parameters. Moreover, if the cut edges in $\mathcal{N}^+$ are given a sufficiently large length, $CF_{ab|cd}$ can be made as close to 1 as desired, and hence distinct from the other CF entries. Since CF_{abcd} is an analytic function of parameters and one parameter choice leads to its being strictly $(ab|cd)$-cut, generic ones must as well (since any equality of analytic functions either holds everywhere, or only on a lower-dimensional subset of the domain).

The remaining cases, of the reduced rooted trees of blobs $((a, b), (c, d))$ and $((a, b), c, d)$, are similar. Any coalescence below the blob containing the LSA leads to gene trees $ab|cd$. If no such coalescence occurs, then upon entering the blob containing the LSA the lineages from a, b are exchangeable, so $ac|bd$ and $ad|bc$ will be equally probable gene tree topologies, resulting in CF_{abcd} being $(ab|cd)$-cut. Considering sufficiently long cut edges in $\mathcal{N}^+$ again shows the CF is strictly cut generically.

To prove (b), suppose T has the unresolved topology, so that $\mathcal{N}^-$ has a 4-blob. Again using the analyticity of CF_{abcd} it is enough to show there is a single choice of numerical parameters that gives a non-cut CF. We can even choose these parameters to be on the boundary of the stochastic parameter space, since the analytic parametrization of the CFs extends to a larger open set. We now show such a parameter choice exists, with some edge lengths and hybridization parameters 0.

If there are any 2-blobs on $\mathcal{N}^-$, set all edge lengths in $\mathcal{N}^+$ that give rise to them equal to 0, with hybrid parameters arbitrary. Doing so, we have effectively removed these blobs, and may thus assume there are no 2-blobs in $\mathcal{N}^-$. By Lemma 1, the only non-leaf blob in $\mathcal{N}^-$ is a 4-blob.

To further simplify the network, choose some total order for the nodes in $\mathcal{N}^+$ consistent with the partial order arising from the edge directions, with the root highest. Focus on the lowest hybrid node w in this order, and its hybrid edges k_1, k_2. Consider deleting one of the k_i from $\mathcal{N}^+$ and with it all edges from which the only directed path to a taxon leads through k_i, suppressing any degree-2 nodes. If the semidirected unrooted network of the resulting network still has a 4-blob, then set $\gamma_i = 0$ and lengths for the removed edges to be arbitrary, so that we effectively consider a network with one fewer hybrid nodes. Its semidirected unrooted network may have 2-blobs as well as the 4-blob, but after repeatedly 'removing' 2-blobs and one of the lowest hybrid edges in the 4-blob by setting certain parameters to 0, we arrive at a network such that $\mathcal{N}^-$ still has a single 4-blob and no other blobs, but for which removing either of $\mathcal{N}^+$'s lowest hybrid edges h_1, h_2, at lowest hybrid node v, in this way gives a semidirected network with no 4-blobs. We henceforth assume our network $\mathcal{N}^+$ has this property. See Fig. 5 for a schematic depiction of this simplification process, and as an aid in following our subsequent steps.

If v is the lowest hybrid node on $\mathcal{N}^+$, then the subnetwork below v must be a tree. But since $\mathcal{N}^-$ has no 3-blobs, this tree can only have one leaf, and hence there is only a single edge below v. By permuting taxon names, we assume the leaf below v is labelled a. Removing from $\mathcal{N}^+$ either of the h_i, and edges above it as described earlier, gives connected subnetworks N_i which by suppressing degree-2 nodes give phylogenetic networks $\mathcal{N}_i^+$. Moreover, the semidirected unrooted networks $\mathcal{N}_i^-$ each have exactly two 3-blobs, and possibly 2-blobs. By further permuting taxon names we may assume $\mathcal{N}_1^+$ has reduced unrooted tree of blobs topology $ab|cd$.

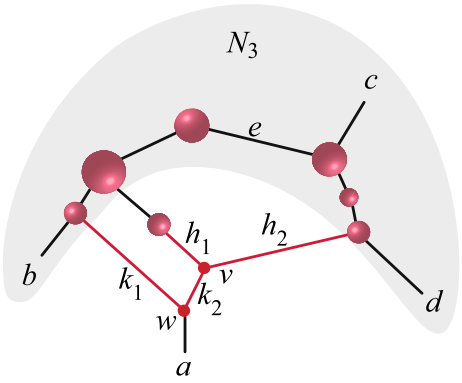

Fig. 5 The schematic form of the 4-taxon network $\mathcal{N}^+$ used to establish Claim (b) in the proof of Theorem 1. The root (unlabelled) could be anywhere in the gray region. Red spheres represent biconnected subgraphs, which may become blobs on induced networks on subsets of taxa. With w the lowest hybrid node in $\mathcal{N}^+$, one of its hybrid edges, k_1, is removed since doing so leaves a 4-blob. With v then the lowest hybrid node, removing either of its hybrid edges, h_1, h_2, would result in no 4-blob. We let $\mathcal{N}_1^+$ be the result of further removing edge h_2 and edges ancestral to it and only the taxon a. The network $\mathcal{N}_2^+$ results similarly from further removing edge h_1 instead of h_2. Removing all edges and nodes which lie above only the taxon a gives network N_3, shown in the gray region. The edge e is the cut edge incident to the single 3-blob in N_3, through which paths from that blob to taxon b pass (color figure online)

For the sake of contradiction, suppose $\mathcal{N}_2^+$'s reduced unrooted tree of blobs also has topology $ab|cd$. Consider the subnetwork N_3 on b, c, d obtained from $\mathcal{N}^+$ by deleting a and all edges above a that are not above any other taxa. Then N_3 is a subnetwork of both N_1 and N_2 which has a blob $\mathcal{B}$ determined by the 3 taxa b, c, d. Let e denote the cut edge of N_3 incident to $\mathcal{B}$ through which undirected paths to b pass. Now e must be a cut edge in both N_1 and N_2, inducing the split $ab|cd$ in both. Thus every edge in $\mathcal{N}^+$ which is incident to N_3 and ancestral to only the taxon a must be attached to N_3 in the b-component of $N_3 \setminus \{e\}$. But this implies that e is a cut edge of $\mathcal{N}^+$ inducing the split $ab|cd$, a contradiction to the existence of a 4-blob on $\mathcal{N}^-$. Thus $\mathcal{N}_2^+$ has a reduced unrooted tree of blobs topology that is resolved, but not $ab|cd$. We henceforth assume this topology is $ac|bd$.

To pick values for the remaining parameters note that since a is the only taxon below the hybrid node v,

$$CF_{abcd}(\mathcal{N}^+) = \gamma_1 CF_{abcd}(\mathcal{N}_2^+) + \gamma_2 CF_{abcd}(\mathcal{N}_1^+).$$

where $\gamma_1, \gamma_2 = 1 - \gamma_1$ are the hybridization parameters for h_1, h_2. Moreover, by (a) we have that $CF(\mathcal{N}_1^+)$ is strictly $(ab|cd)$-cut and $CF(\mathcal{N}_2^+)$ is strictly $(ac|bd)$-cut for generic parameters. Thus by first choosing the numerical parameters other than γ_1, γ_2 on $\mathcal{N}^+$ to yield such generic parameters on the $\mathcal{N}_i^+$, we may then pick values of γ_1, γ_2 so that $CF_{abcd}(\mathcal{N}^+)$ is non-cut. Thus $CF_{abcd}(\mathcal{N}^+)$ is generically non-cut. □

Applying this proposition to quartet CFs from large networks gives the following.

Corollary 1 *Let $\mathcal{N}^+$ be a metric rooted binary phylogenetic network on taxa X, $|X| \geq 4$, with generic numerical parameters. Then under the NMSC, for each 4-taxon subset*

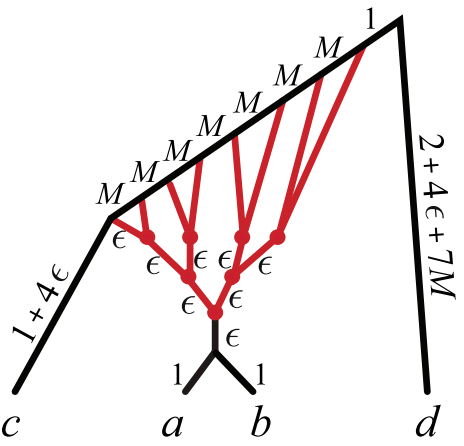

Fig. 6 An instance of the network $\mathcal{N}(k)^+$ used in the proof of Proposition 1, with $k = 3$. All hybridization parameters are $1/2$, while ϵ and M denote variable edge lengths

$Q \subseteq X$, the topology of the reduced unrooted tree of blobs on the induced network $T_{rd}(\mathcal{N}_Q^-)$ is identifiable from CF_Q.

Proof By Theorem 1, for generic numerical parameter values on each induced 4-taxon network we have CF-detectability of a B-quartet or T-quartet. Since the generic conditions only exclude a set of measure zero from the numerical parameter space of each 4-taxon network, they give rise to a generic condition on numerical parameter values on the n-taxon network ensuring that CF-detectability holds on all induced 4-taxon networks. □

We now characterize more fully the set of CFs that arise on 4-networks $\mathcal{N}^+$ whose trees of blobs are resolved. Suppose $\mathcal{N}^+$ has taxa a, b, c, d, and reduced unrooted tree of blobs $T_{rd}(\mathcal{N}^-)$ with quartet topology $ab|cd$. If $\mathcal{N}^+$ is a resolved tree, then Allman et al. (2011) showed $CF_{ab|cd}$ may take on any value in the interval $(1/3, 1)$. If $\mathcal{N}^+$ is level-1, then Baños (2019) showed $CF_{ab|cd}$ may take on any value in $(1/6, 1)$. The following generalizes these results to arbitrary networks.

Proposition 1 *Let $\mathcal{N}^+$ be a 4-taxon rooted binary phylogenetic network whose reduced tree of blobs has quartet topology $ab|cd$. Then under the NMSC the CF is $ab|cd$-cut, with*

$$CF_{abcd} = (CF_{ab|cd}, CF_{ac|bd}, CF_{ad|bc}) = (p, q, q),$$

where $0 < p, q < 1$, $p + 2q = 1$. Conversely, every such triple (p, q, q) arises as the CF from such a network.

Proof By statement 1 of Theorem 1, it only remains to establish the final claim, that every triple (p, q, q) with $p, q > 0$, $p + 2q = 1$, arises as the CF of a network of the sort described. We do this by constructing a sequence of topological networks $\mathcal{N}(k)^+$, $k \in \mathbb{Z}^+$, such that a triple (p, q, q) arises as a CF on $\mathcal{N}(k)^+$ for sufficiently large k and certain numerical parameters.

The form of $\mathcal{N}(k)^+$ is shown in Fig. 6 for $k = 3$. Edges lead from the root to the taxon d and to a 3-blob. The other two edges incident to the 3-blob lead to c, and to a cherry of a, b. The edge leading to the cherry has length ϵ and joins the blob at a node that can be thought of as the 'root' of an inverted binary subtree of hybrid edges (shown in red in Fig. 6), inverted so that its edges are directed toward this node. This binary subtree has 2^k 'leaves', and all internal edges of length ϵ. The 'pendant' edges of this subtree have lengths $\epsilon, \epsilon + M, \epsilon + 2M, \ldots, \epsilon + (2^k - 1)M$, with the subtree 'leaves' connected by a path of edges all of length M. The pendant edges of the network $\mathcal{N}(k)^+$, and the internal edge leading from the root of $\mathcal{N}(k)^+$ to the 3-blob can be given any fixed lengths, but for concreteness, we make the network ultrametric by choosing the remaining internal edge to have length 1 and the pendant edges to a, b, c, d to be of lengths $1, 1, 1 + (k + 1)\epsilon, 2 + (k + 1)\epsilon + (2^k - 1)M$, respectively. We set all hybridization parameters equal to $1/2$.

Note that by Theorem 1, CF_{abcd} is strictly $(ab|cd)$-cut for $\epsilon > 0$.

We next show that under the NMSC on $\mathcal{N}(k)^+$, for $\epsilon \approx 0$ and $k \gg 0$, with high probability the a, b lineages will be on different edges of the network when they reach a height of $1 + (k + 1)\epsilon$ above the taxa. This event is the union of k disjoint events, in which the lineages follow the same path without coalescing to height $1 + \ell\epsilon$ for any $\ell \in \{1, 2, \ldots, k\}$ at which point they diverge on different paths. The probability of this for a specific ℓ is $(p/2)^\ell$, where $p = \exp(-\epsilon)$ is the probability two lineages do not coalesce on an edge of length ϵ. Thus the probability of the full event is

$$\alpha = \sum_{\ell=1}^{k} (p/2)^\ell = \frac{p}{2} \cdot \frac{1 - (p/2)^k}{1 - p/2}.$$

Taking ϵ close to 0 ensures p is as close to 1 as desired. Then choosing k sufficiently large, the probability α can be made as close to $p/(2 - p)$ as desired, and hence arbitrarily close to 1.

Now CF_{abcd} can be expressed as

$$CF_{abcd} = \alpha CF_1 + (1 - \alpha)CF_2,$$

where CF_1 is the CF conditioned on the a, b lineages being on different edges at height $1 + (k + 1)\epsilon$ above the leaves, and CF_2 the CF conditioned on the complementary event. To compute CF_1, note that the conditioning ensures that all coalescent events that can occur will have the same probability that they would if they instead occurred on a species tree with topology $(((c, a), b), d)$ or a species tree with topology $(((c, b), a), d)$, with each of these trees having equal probability. Moreover on these trees the length of the edge ancestral only to the cherry is mM for some $m \in \{1, 2, \ldots, 2^k - 1\}$. Thus by choosing M large enough, we can ensure with probability as close to 1 as we like that gene tree topologies will match the population tree, making CF_1 as close to $(0, 1/2, 1/2)$ as desired. Now since α can be made arbitrarily close to 1, we need not analyze CF_2 (beyond knowing its entries are bounded) to conclude that we can make CF_{abcd} as close to $(0, 1/2, 1/2)$ as desired by choices of $\epsilon \approx 0$ and $k, M \gg 0$.

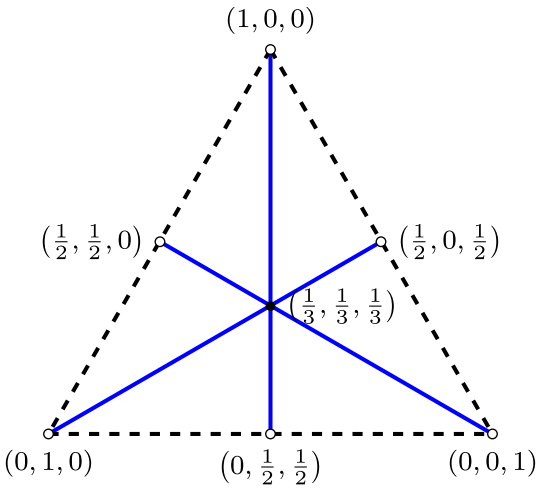

Fig. 7 Geometric view of CFs for 4-taxon network models, with dashed lines outlining the simplex Δ^2. The solid line segments represent CFs arising from species networks whose unrooted reduced trees of blobs are resolved. The vertical line segment corresponds to $ab|cd$, the upward-sloping one to $ac|bd$, and the downward sloping one to $ad|bc$. CFs off of these lines can only arise from networks with unresolved unrooted reduced trees of blobs, and as shown in Baños (2019) all such points arise from level-1 networks. Networks whose unrooted reduced trees of blobs are unresolved may also produce CFs on the line segments, but only for non-generic parameters

Using the same fixed k, so the network topology is still that of $\mathcal{N}(k)^+$, we could instead take $\epsilon \gg 0$, making the probability of coalescence of a, b on the edge above the $\{a, b\}$ cherry as close to 1 as we like, so that CF_{abcd} is arbitrarily close to $(1, 0, 0)$. Since CF_{abcd} lies on the line of points of the form (q, p, p), $q + 2p = 1$ and is a continuous function of numerical parameters, by connectedness of the numerical parameter space for $\mathcal{N}(k)^+$, all intermediate points between the ones we found arise as CF_{abcd} for some parameters. □

The statements of Theorem 1 and Proposition 1 can be made geometric by plotting CFs (Mitchell et al. 2019; Baños 2019; Allman et al. 2019, 2022). A CF is a point in the interior of the 2-dimensional probability simplex,

$$\Delta^2 = \left\{(p_1, p_2, p_3) \mid p_i \geq 0, \sum p_i = 1\right\}.$$

Figure 7 gives a depiction of Δ^2, with the three blue line segments within it showing the locations of cut CFs. If the unrooted reduced tree of blobs of a 4-taxon network is $ab|cd$, then CF_{abcd} lies on the vertical line segment shown in the figure, and every point on this line segment within the simplex arises from some such network. The other line segments in the simplex similarly show values of CF_{abcd} arising from networks with unrooted reduced trees of blobs $ac|bd$ and $ad|bc$. Points in the simplex off these line segments arise as CFs only for networks whose unrooted reduced trees of blobs are unresolved. By Baños (2019), all points off the line segments arise from level-1 networks with 4-cycles. Although a network with a more complicated 4-blob

may produce a CF on the line segments for certain numerical parameters, this cannot happen for generic parameters by Theorem 1.

4 Blob quartet identifiability on large networks

Theorem 1 will be applied to the induced network on four taxa arising from a larger n-taxon network. The CFs computed from the induced 4-taxon networks are the same as gene tree probabilities from the large network marginalized to 4-taxon sets, by the structure of the NMSC model. However, since four taxa which form a B-quartet on a large network may not do so on an induced one, determining B-quartets on a large network generally requires additional arguments, which are developed in this section.

The following lemma leads to one easy deduction of B-quartets from those on induced networks.

Lemma 2 *Consider a network N with degree-1 nodes bijectively labelled by X, and a subnetwork M of N with the restricted labelling of some degree-1 nodes by $Y \subseteq X$. If a set $S \subseteq Y$ determines a blob on M, then S determines a blob on N. Moreover, the incident cut edges of the blob on N leading to elements of S are in M.*

Proof If S determines a blob $\mathcal{B}_0$ on M, then there exist undirected paths in M from $\mathcal{B}_0$ to each $s \in S$, with no edges in common among any pair of paths. But $\mathcal{B}_0$ is contained in a blob $\mathcal{B}$ of N. For each $s \in S$, the path from $\mathcal{B}_0$ to s may include some edges in $\mathcal{B}$, but it has a subpath from $\mathcal{B}$ to s entirely outside of $\mathcal{B}$. Moreover, these subpaths for different s have no edges in common, and must thus pass through distinct cut edges incident to $\mathcal{B}$. Hence S determines $\mathcal{B}$, and the incident cut edges leading to each s are in M. □

To apply this to induced phylogenetic networks on subsets of taxa, observe that induced networks are obtained from subnetworks by suppressing degree-2 nodes. Under this operation, blobs pass to blobs, and cut edges to cut edges. Thus we have the following.

Corollary 2 *Let $\mathcal{N}^+$ be a rooted binary phylogenetic network on X, and $\mathcal{M}^+$ the induced network on $Y \subset X$. Then any B-quartet on $\mathcal{M}^+$ is a B-quartet on $\mathcal{N}^+$.*

To identify additional B-quartets from those identified by Theorem 1 and Corollary 2, we develop an inference rule. To state it concisely, we say taxa a, b are *separated* in a resolved quartet if they lie in different cherries. Thus the taxa a, b are separated in $ac|bd$ and $ad|bc$, but are not separated in $ab|cd$.

Theorem 2 (B-quartet Inference Rule) *Consider a rooted binary phylogenetic network $\mathcal{N}^+$ on n taxa, $n \geq 5$. Suppose that $\{a, b, c, d\}$ and $\{b, c, d, e\}$ are B-quartets on $\mathcal{N}^+$. If on the induced 4-taxon network any one of $\{a, b, c, e\}$, $\{a, b, d, e\}$, or $\{a, c, d, e\}$ is*

(a) a T-quartet, with a, e separated in the reduced unrooted tree of blobs for the induced 4-taxon network, or
(b) a B-quartet,

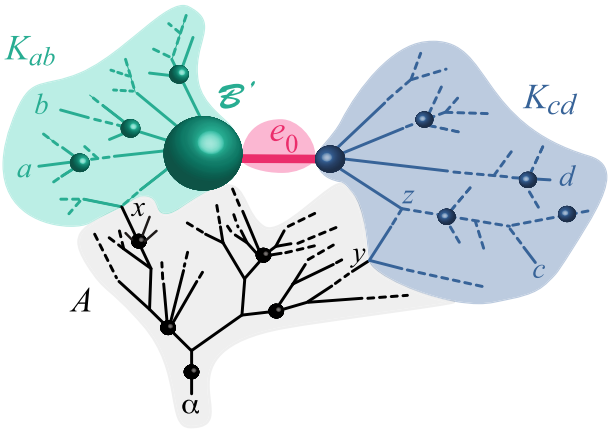

Fig. 8 A schematic of the network $\mathcal{N}^+$, as described in Lemma 3. Edges are partitioned into four color-coded sets. Black edges are ancestral to the taxon α and no other taxa, forming the subnetwork A. Non-black edges form the subnetwork N', in which the blob $\mathcal{B}'$ is determined by $\{a, b, c\}$. The red edge e_0 incident to $\mathcal{B}'$ is a cut edge of $\mathcal{N}'$, separating the connected components K_{ab} and K_{cd}, shown in green and blue, respectively. The root of $\mathcal{N}^+$ might be in either K_{ab} or K_{cd}. The nodes x, y, z are described in the proof of the lemma (color figure online)

then all of $\{a, b, c, e\}$, $\{a, b, d, e\}$, *and* $\{a, c, d, e\}$ *are B-quartets on* $\mathcal{N}^+$.

Proof The taxa b, c, d determine a blob in $\mathcal{N}^+$, corresponding to a node v in its tree of blobs. But since $\{a, b, c, d\}$ and $\{b, c, d, e\}$ are B-quartets, undirected paths in the tree of blobs from the taxa a and e also first meet those from b, c, d at v. The conclusion will follow from showing the paths from a and e to v do not meet each other before v, so that all 5 paths from a, b, c, d, e first meet at v.

Suppose the paths from a, e do meet before v. Then there is an edge in the tree of blobs, and hence a cut edge in the network, that separates a, e from b, c, d. This implies that picking any two of b, c, d, the taxa a, e are not separated in the 4-taxon tree of blobs, nor do they form a B-quartet with a, e. □

For example, for the network of Fig. 3 both $\{a, b, c, d\}$ and $\{b, c, d, e\}$ are CF-detectable B-quartets. While $\{a, b, d, e\}$ is not a CF-detectable B-quartet, since CF_{abde} is strictly $ab|de$-cut, applying Theorem 2 shows that it is a B-quartet.

To show that the previous propositions give sufficient tools to detect all B-quartets for generic parameters, we use the following lemma.

Lemma 3 *Let* $\mathcal{N}^+$ *be a rooted binary phylogenetic network on taxa* X *for which* $\{a, b, c, d\}$ *is a B-quartet, and suppose for some* $\alpha \in X$ *the induced network* $\mathcal{N}'$ *on* $X \smallsetminus \{\alpha\}$ *has a cut edge in* $\mathcal{N}'$ *separating* a, b *from* c, d. *Then* $\{a, b, c, \alpha\}$ *is a B-quartet on the induced network* $\mathcal{M}^+$ *on* $X \smallsetminus \{d\}$.

Proof We define several subnetworks of $\mathcal{N}^+$, with Fig. 8 provided to assist the reader. Let A be the connected subnetwork of $\mathcal{N}^+$ whose edges are those ancestral to the taxon α but no other taxa. Let N' be the connected subnetwork of $\mathcal{N}^+$ whose edges are those ancestral to at least one taxon other than α, and M the connected subnetwork of $\mathcal{N}^+$

whose edges are ancestral to at least one taxon other than d. Note that N', M yield the induced networks $\mathcal{N}'$, $\mathcal{M}^+$ on $X \smallsetminus \{\alpha\}$, $X \smallsetminus \{d\}$ by suppressing degree-2 nodes, and $\mathcal{N}^+ = N' \cup A$.

Let $\mathcal{B}'$ be the blob in N' determined by a, b, and c, and let e_0 be the cut edge of N' incident to $\mathcal{B}'$ through which paths to c pass. Thus e_0 also separates $a, b,$ from c, d in N'. Let K_{ab} (respectively K_{cd}) denote the connected component of $N' \smallsetminus \{e_0\}$ containing a, b (respectively c, d). Then the edges of the four connected subnetworks A, K_{ab}, $\{e_0\}$, and K_{bc} partition the edges of $\mathcal{N}^+$, as shown in Fig. 8. We now construct a cycle in $\mathcal{N}^+$ through these subnetworks with certain features.

First, there is an undirected path P_1 entirely within A from α to a node x in K_{ab}. If this were not the case, then all paths from α to N' within A would end at nodes in K_{cd}. But then e_0 would separate a, b from c, d, α in $\mathcal{N}^+$, contradicting that $\{a, b, c, d\}$ is a B-quartet on $\mathcal{N}^+$.

There is also a path P_2 from x to e_0 in K_{ab}, by the connectedness of K_{ab}.

By a similar argument to that for P_1, there is an undirected path from a node $y \in K_{cd}$ to α within A. Because $y \in K_{cd}$, y is ancestral to a taxon other than α. Since the edge in the path incident to y is ancestral only to α, that edge's parent node must be y and there is a directed path from y to α within A. Choose some such directed path.

The nodes y and c must have a common ancestor in K_{cd}, since either the root of $\mathcal{N}'$ is in K_{cd} or any directed path from the root to any node in K_{cd} passes through e_0 and the child node of e_0 is such an ancestor. Choosing z as a least common ancestor of y, c in K_{cd} (i.e., a common ancestor with no descendent that is a common ancestor), and a directed path from z to y, we form a combined directed path P_4 from z through y to α, with all edges ancestral to α.

If all edges and nodes ancestral only to d are deleted from N', the network remains connected and contains both z and e_0. Thus there is a path P_3 from e_0 to z in K_{cd} with no edges that are ancestral to only the taxon d.

Combining the paths P_1, P_2, the edge e_0, P_3, and P_4, and removing edges to eliminate any self-intersections, yields a cycle C in $\mathcal{N}^+$ which passes through A, K_{ab}, e_0, and K_{cd}. This cycle also lies in M, as none of its edges are ancestral only to the taxon d. Since $\mathcal{B}'$ also lies in M, and the cycle C and $\mathcal{B}'$ intersect, they lie in the same blob $\mathcal{B}$ of M.

It remains to show that $\{a, b, c, \alpha\}$ determines $\mathcal{B}$ in M, and hence is a B-quartet on $\mathcal{M}^+$. Since $\{a, b, c\}$ determines $\mathcal{B}'$ in N' and hence in $M \cap N'$, by Lemma 2 $\{a, b, c\}$ determines $\mathcal{B}$ in M with incident cut edges leading to a, b, c in $M \cap N' \subset N'$. But the initial segment of P_1 gives a path from α to C. Since this path lies entirely in A, the cut edge incident to $\mathcal{B}$ that leads to α must be in A, and is therefore distinct from those to a, b, c. Thus $\{a, b, c, \alpha\}$ determines $\mathcal{B}$ in M. □

We arrive at the main result of this section.

Theorem 3 *On an n-taxon rooted binary phylogenetic network $\mathcal{N}^+$ with generic numerical parameters, all B-quartets can be identified from the quartet CFs using CF-detectability (Theorem 1) and applications of the B-quartet Inference Rule (Theorem 2).*

Proof By Corollary 1, for generic parameters we may identify the topologies of the reduced unrooted trees of blobs of all induced networks on four taxa. Since the B-quartet Inference Rule does not depend on parameters, we have only to show that this information together with the inference rule is enough to identify all B-quartets.

We proceed by induction on the number n of taxa on the network $\mathcal{N}^+$, with the base case of $n = 4$ established. Inductively assume that the result holds for networks with fewer than n taxa, and consider $\mathcal{N}^+$ with $n \geq 5$ taxa.

Suppose $\{a, b, c, d\}$ is a B-quartet on $\mathcal{N}^+$, determining a blob $\mathcal{B}$. Then consider the connected components of the graph obtained by deleting $\mathcal{B}$. Choose one taxon from each component which contains a taxon, with four of these being a, b, c, d. Passing to the induced network on those taxa, all edges in $\mathcal{B}$ are retained. If this network has fewer than n taxa, then the inductive hypothesis gives that $\{a, b, c, d\}$ can be identified as a B-quartet on it, and by Corollary 2 on $\mathcal{N}^+$.

If the number of taxa was not decreased, then $\mathcal{N}^+$ has a relatively simple structure: Its LSA network $\mathcal{N}^{\oplus}$ contains the blob $\mathcal{B}$ with n incident cut edges. If the LSA is in $\mathcal{B}$, then the incident cut edges connect to (possibly empty) chains of 2-blobs leading to leaves. If the LSA is not in $\mathcal{B}$, then $n - 1$ incident cut edges connect to chains of 2-blobs leading to leaves, and one connects through a chain of 2 blobs to a 3-blob containing the LSA, which connects to another chain of 2-blobs leading to a leaf. For a network $\mathcal{N}^+$ of this form if we remove any taxon other than a, b, c, d and pass to the induced network, $\{a, b, c, d\}$ either remains a B-quartet or does not. If $\{a, b, c, d\}$ remains a B-quartet, then we may delete that taxon, and again obtain the result from the inductive hypothesis.

Suppose then that no taxon can be removed from $\mathcal{N}^+$ without $\{a, b, c, d\}$ ceasing to be a B-quartet in the induced network, and fix some $\alpha \in X \smallsetminus \{a, b, c, d\}$. Let $\mathcal{N}'$ be the induced rooted network on $X \smallsetminus \{\alpha\}$. Then the blob $\mathcal{B}$ on $\mathcal{N}^+$ splits into multiple blobs with cut edges joining them on $\mathcal{N}'$, and $\{a, b, c, d\}$ is a T-quartet on $\mathcal{N}'$. There must be a cut edge e_0 in $\mathcal{N}'$ that separates two of a, b, c, d, say a, b, from the others, c, d. Theorem 1 thus shows that CF_{abcd} is $(ab|cd)$-cut.

Applying Lemma 3 twice, we conclude that $\{a, b, c, \alpha\}$ and $\{b, c, d, \alpha\}$ are B-quartets on the networks induced from $\mathcal{N}^+$ by removing d and a respectively. As these are networks on $n - 1$ taxa, the inductive hypothesis ensures that they can be detected as B-quartets. But they must then also be B-quartets on $\mathcal{N}^+$ by Corollary 2. An application of Inference Rule (a) of Theorem 2 then establishes the claim. □

Although the proof of Theorem 3 shows that only part (a) of Theorem 2 is needed to infer all B-quartets from those that are CF-detectable, part (b) is useful in an inference algorithm for reducing computational time.

Figure 9 shows several instructive examples of blobs for understanding the proof and application of Theorem 3. On the left, a simple 7-cycle relates a, b, c, d, e, f and hybrid taxon α. Though $\{a, b, c, d\}$ is a B-quartet, it is not CF-detectable since CF_{abcd} is $ad|bc$-cut. To infer that $\{a, b, c, d\}$ is a B-quartet by the argument of the proof, the taxa e, f can be ignored, as passing to the induced network without them leaves $\{a, b, c, d\}$ a B-quartet. The taxon α will be used, since its deletion would make $\{a, b, c, d\}$ a T-quartet. The CFs for $\{\alpha, a, c, d\}$ and $\{\alpha, b, c, d\}$ show those sets are B-quartets, so using that CF_{abcd} is $ad|bc$-cut in inference rule (a) of Theorem 2 gives

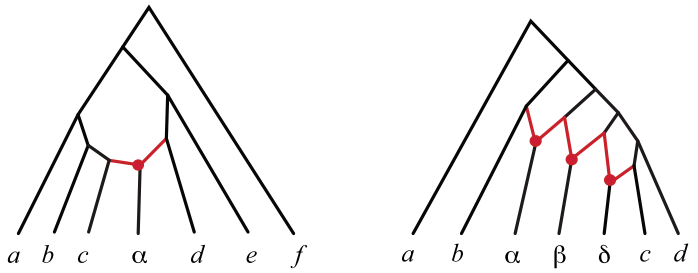

Fig. 9 **L** A 7-blob with a simple cycle structure. While many of its B-quartets are not CF-detectable, each can be inferred from CF-detectable ones by a single application of the B-quartet Inference Rule. For instance, $\{a, b, c, d\}$ is a B-quartet although CF_{abcd} is $ad|bc$-cut. The inference rule shows that it is a B-quartet using the two CF-detectable ones, $\{\alpha, a, c, d\}$ and $\{\alpha, b, c, d\}$. **R** A 7-blob with a more complex structure. The B-quartet $\{a, b, c, d\}$ is not CF-detectable, but three applications of the inference rule allow it to be inferred from those that are

the desired conclusion. Every other B-quartet for this network can be similarly inferred using the inference rule once.

A more complicated example with a 7-blob in Figure 9R illustrates the need for the inductive argument for Theorem 3. Here we explain how to infer that $\{a, b, c, d\}$ is a B-quartet even though CF_{abcd} is $(ab|cd)$-cut. Note that deletion of any one of α, β, δ would give an induced network with $\{a, b, c, d\}$ a T-quartet. We pick any one of these, say α, and find that $\{a, b, c, \alpha\}$ is a B-quartet using its CF. Then, by considering the induced 6-taxon network on $\{b, c, d, \alpha, \beta, \delta\}$, which has a 6-blob when unrooted, we see inductively that $\{\alpha, b, c, d\}$ is a B-quartet, so by the inference rule $\{a, b, c, d\}$ is also. Tracing through the full argument for $\{b, c, d, \alpha, \beta, \delta\}$ to explicitly show $\{\alpha, b, c, d\}$ is a B-quartet requires several more applications of the inference rule.

Of course in an inference algorithm, where the network structure is not yet known, this analysis is done in the opposite order, by first finding all CF-detectable B-quartets, and then using repeated applications of the rule to infer new ones until no more can be produced.

5 Main result

The identifiability of the tree of blobs of a species network now follows easily.

Theorem 4 *Let $\mathcal{N}^+$ be a rooted binary phylogenetic network. Then for generic numerical parameters, the reduced unrooted tree of blobs $T_{rd}(\mathcal{N}^-)$ is identifiable from the distribution of gene quartet topologies under the NMSC model.*

Proof By Theorem 3, for generic numerical parameters on $\mathcal{N}^+$, all B-quartets on $\mathcal{N}^+$ can be identified from the quartet CFs, that is, from the distributions of gene quartet topologies. By Corollary 1, we can additionally identify the topology of each unrooted reduced quartet tree of blobs if it is resolved.

Treating B-quartets on $\mathcal{N}^+$ as unresolved, we thus can identify the topology of every displayed quartet tree on $T_{rd}(\mathcal{N}^-)$. But the collection of displayed quartets determine the tree (Semple and Steel 2005; Rhodes 2020), so the tree of blobs is identifiable. □

This result addresses the theoretical question of whether it is in principle possible to infer $T_{rd}(\mathcal{N}^-)$ from quartet CFs, but its proof also suggests an algorithm for inference of the tree of blobs from data. While the complete development and implementation of the algorithm, together with sample analyses of empirical data, will be treated in another publication, we outline the steps here.

For a set $\{a, b, c, d\}$ of four taxa, a *quartet count concordance factor (qcCF)* is a vector of counts $(n_{ab|cd}, n_{ac|bd}, n_{ad|bc})$ of unrooted topological quartet trees. We assume for each given set $\{a, b, c, d\}$ these counts summarize a sample of independent draws under the NMSC. For instance, these could be displayed quartets on a collection of independent gene trees on the full set X of taxa, or on subsets of X. While gene trees are not empirically observable, given gene sequence data they may be inferred by standard phylogenetic methods, at the price of introducing inference error.

Beginning with a collection of independent gene trees on X, the algorithm proceeds as follows:

1. Tabulate qcCFs for all sets of four taxa.
2. Apply a statistical hypothesis test to each qcCF to judge whether the T-quartet model can be rejected. If so, the taxa form a putative B-quartet on the induced 4-taxon network. If not, infer the resolved quartet tree of blobs topology.
3. Use the B-quartet Inference Rule repeatedly to determine all putative B-quartets on the full network.
4. Treating putative B-quartets as unresolved quartet trees and T-quartets as resolved, estimate the unrooted reduced tree of blobs, using the quartet intertaxon distance (Rhodes 2020) and a tree building method.

This algorithm, is similar in outline to NANUQ (Allman et al. 2019), which provides for statistically-consistent inference of a network provided it is level-1. However, since it does not attempt to infer any details of blob structure, it avoids the complications of interpreting splits graphs. Like NANUQ, it can be shown to provide a statistically consistent estimate of the network features it seeks to infer.

Note that step 2 requires the development of a novel statistical test, since the cut model has a singularity at the point $(1/3, 1/3, 1/3)$ where the 3 model lines intersect. Singularities cause significant problems in statistical analyses, so a new and non-standard distribution must be developed for use in computing p-values associated to qcCFs. Derivation of an appropriate distribution proceeds along the lines of the T3 test of Mitchell et al. (2019).

Step 3 cannot be done naively, since computational complexity needs to be controlled for use on large networks. Multiple cycles of applying the B-quartet inference rule are needed in which any recently inferred B-quartets are compared to known ones. By minimizing the number of pairs of B-quartets considered, this can be accomplished in time $\mathcal{O}(n^5)$, for n taxa.

Finally, for step 4 numerous existing methods for determining a tree from its displayed quartets could be used, provided that they tolerate some amount of noise due

to mistakes in classifying qcCFs as signifying resolved or unresolved quartet trees of blobs. We choose to use the quartet intertaxon distance of Rhodes (2020) combined with a tree building method such as Neighbor-Joining as a means of addressing such noise in the quartets and still achieving reasonable runtimes. In the end, the computational time is $\mathcal{O}((n+m)n^4)$, where m is the number of input gene trees and n the number of taxa. An implementation will be given in the `MSCquartets` R package.

The tree of blobs shown to be identifiable by Theorem 4, and estimated by the algorithm sketched above, is of course a topological tree. Researchers might prefer a metric tree of blobs, indicating (in coalescent units) the distance between blobs. For edges between trivial blobs, it is straightforward to see that edge lengths are identifiable, and heuristics such as those used by ASTRAL (Sayyari and Mirarab 2016) for species tree inference provide a fast estimate of them. However, if either, or both, endpoints of an edge are in non-trivial blobs, then both identifiability and methods for effective estimation are far from obvious. For example, on the 4-taxon network in Fig. 6 if the length ϵ of the edge between the blobs is held fixed but k or M varied, the CF varies over a line segment. This segment overlaps a similarly constructed segment of CFs for a nearby value $\epsilon' \approx \epsilon$. Thus the distance between blobs cannot be identified in this 4-taxon case. This identifiability question for larger networks will also be studied in a future work.

Acknowledgements This work was supported by the National Science Foundation, Grant 2051760, awarded to JR and EA. EA and JM were also supported by NIGMS Institutional Development Award (IDeA), Grant 2P20GM103395. HB was supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation Grant 735923LPI (DOI: https://doi.org/10.46714/735923LPI) awarded to Andrew J. Roger and Edward Susko.

References

Allman ES, Baños H, Rhodes JA (2019) NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms Mol Biol 14(24):1–25

Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62(6):833–862

Allman ES, Matias C, Rhodes JA (2009) Identifiability of parameters in latent structure models with many observed variables. Ann Stat 37(6A):3099–3132

Allman ES, Mitchell JD, Rhodes JA (2022) Gene tree discord, simplex plots, and statistical tests under the coalescent. Syst Biol 71:929–942. https://doi.org/10.1093/sysbio/syaa104

Baños H (2019) Identifying species network features from gene tree quartets. Bull Math Biol 81:494–534

Blischak PD, Chifman J, Wolfe AD, Kubatko LS (2018) HyDe: a Python package for genome-scale hybridization detection. Syst Biol 67:821–829

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265

Dress AWM, Huson DH (2004) Constructing splits graphs. IEEE/ACM Trans Comput Biol Bioinf 1(3):109–115

Erdős PL, Semple C, Steel M (2019) A class of phylogenetic networks reconstructable from ancestral profiles. Math Biosci 313:33–40

Flouri T, Jiao X, Rannala B, Yang Z (2019) A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol Biol Evol 37(4):1211–1223

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz M, Hansen NF, Durand EY, Malaspinas A, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan

P, Brajkovic D, Kucan Ž, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science 328:710–722

Grünewald S, Huber KT (2007) Reconstructing evolution: new mathematical and computational advances, chapter identifying and defining trees. Oxford University Press, pp 217–244

Gusfield D, Bansal V, Bafna V, Song YS (2007) A decomposition theory for phylogenetic networks and incompatible characters. J Comput Biol 14(10):1247–1272

Hamlin JAP, Hibbins MS, Moyle LC (2020) Assessing biological factors affecting postspeciation introgression. Evol Lett 4:137–154

Hibbins MS, Hahn MW (2022) Phylogenomic approaches to detecting and characterizing introgression. Genetics 220(2):173

Huber KT, Moulton V, Semple C, Wu T (2018) Quarnet inference rules for level-1 networks. Bull Math Biol 80(8):2137–2153

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53(1):320–328

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75(1):35–45

Mitchell JD, Allman ES, Rhodes JA (2019) Hypothesis testing near singularities and boundaries. Electron J Stat 13(1):2150–2193

Murakami Y, van Iersel L, Janssen R, Jones M, Moulton V (2019) Reconstructing tree-child networks from reticulate-edge-deleted subnetworks. Bull Math Biol 81(10):3823–3863

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5(5):568–583

Rhodes JA (2020) Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Trans Comput Biol Bioinf 17(6):2107–2118

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sayyari E, Mirarab S (2016) Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol 33(7):1654–1668

Semple C, Steel M (2005) Phylogenetics. Oxford University Press, Oxford

Semple C, Toft G (2021) Trinets encode orchard phylogenetic networks. J Math Biol 83(3):28

Solís-Lemus C, Ané C (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet 12(3):e1005896

Steel M (2016) Phylogeny: discrete and random processes in evolution. SIAM, Philadelphia

van Iersel L, Moulton V, Murakami Y (2020) Reconstructibility of unrooted level-k phylogenetic networks from distances. Adv Appl Math 120:102075

Xu J, Ané C (2021) Identifiability of local and global features of phylogenetic networks from average distances. J Math Biol. https://doi.org/10.1007/s00285-022-01847-8 (to appear)

Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet 8:e1002660

Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16:S10

Zhang C, Ogilvie HA, Drummond AJ, Stadler T (2017) Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol 35(2):504–517

Zhu J, Wen D, Yu Y, Meudt HM, Nakhleh L (2018) Bayesian inference of phylogenetic networks from bi-allelic genetic markers. PLoS Comput Biol 14(1):e1005932

Zhu J, Yu Y, Nakhleh L (2016) In the light of deep coalescence: revisiting trees within networks. BMC Bioinf 5:271–282