IDENTIFIABILITY OF PHYLOGENETIC LEVEL-2 NETWORKS UNDER THE JUKES-CANTOR MODEL

AVIVA K. ENGLANDER[1], MARTIN FROHN[2], ELIZABETH GROSS[3], NIELS HOLTGREFE[4], LEO VAN IERSEL[4], MARK JONES[4], AND SETH SULLIVANT[5]

Abstract. We consider the fundamental question of which evolutionary histories can potentially be reconstructed from sufficiently long DNA sequences, by studying the identifiability of phylogenetic networks from data generated under Markov models of DNA evolution. This topic has previously been studied for phylogenetic trees and for phylogenetic networks that are level-1, which means that reticulate evolutionary events were restricted to be independent in the sense that the corresponding cycles in the network are non-overlapping.

In this paper, we study the identifiability of phylogenetic networks from DNA sequence data under Markov models of DNA evolution for more general classes of networks that may contain pairs of tangled reticulations. Our main result is generic identifiability, under the Jukes-Cantor model, of binary semi-directed level-2 phylogenetic networks that satisfy two additional conditions called triangle-free and strongly tree-child.

We also consider level-1 networks and show stronger identifiability results for this class than what was known previously. In particular, we show that the number of reticulations in a level-1 network is identifiable under the Jukes-Cantor model.

Moreover, we prove general identifiability results that do not restrict the network level at all and hold for the Jukes-Cantor as well as for the Kimura-2-Parameter model. We show that any two binary semi-directed phylogenetic networks are distinguishable if they do not display exactly the same 4-leaf subtrees, called quartets. This has direct consequences regarding the blobs of a network, which are its reticulated components. We show that the tree-of-blobs of a network, the global branching structure of the network, is always identifiable, as well as the circular ordering of the subnetworks around each blob, for networks in which edges do not cross and taxa are on the outside.

Keywords. Phylogenetic network; identifiability; reticulate evolution; tangled reticulations; level-2; Markov model; phylogenetic invariant.

1. Introduction

Reticulate events such as hybridization, horizontal gene transfer, and recombination are thought to be crucial in understanding the evolutionary histories of many classes of organisms (Arnold, 1997; Barton, 2001; Mallet, 2005, 2007; Keeling and Palmer, 2008; Soucy et al., 2015), including bacteria (e.g. Diop et al., 2022), viruses (e.g. Worobey et al., 2008; Pekar et al., 2021; Jiao et al., 2024), plants (e.g. Rieseberg, 1997; Wang et al., 2024), fish (e.g. Meier et al., 2019), birds (e.g. Taylor and Larson, 2019), and primates (e.g. Patterson et al., 2006; Green et al., 2010; Raghavan et al., 2014). Phylogenetic networks have emerged as a fundamental framework for modeling evolutionary relationships that involve reticulate events (Nakhleh, 2010; Bapteste et al., 2013;

Date: April 18, 2025.

[1] *Department of Mathematics, University of Wisconsin-Madison, Madison, WI, USA.*
[2] *Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands.*
[3] *Department of Mathematics, University of Hawai'i at Mānoa, Honolulu, HI, USA.*
[4] *Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands.*
[5] *Department of Mathematics, North Carolina State University, Raleigh, NC, USA.*

Kong et al., 2022). Unlike phylogenetic trees, which enforce a strict branching structure, phylogenetic networks allow for the representation of more complex evolutionary processes, including reticulate events where lineages merge. However, as network structures increase in complexity, determining the underlying network that produced a given dataset becomes significantly more challenging. A major question of study in this area is what network structures are identifiable from which types of data. This question is important for the development of consistent phylogenetic network reconstruction methods, such as the recently developed Nanuq (Allman et al., 2019), Nanuq$^+$ (Allman et al., 2024b), Squirrel (Holtgrefe et al., 2025) and TINNIK (Allman et al., 2024a) methods.

Unfortunately, so far almost all identifiability results and consistent reconstruction methods for phylogenetic networks have been restricted to the class of level-1 networks. This class is more general than phylogenetic trees, as it allows for reticulate nodes in which lineages merge. However, it is still severely restricted since these reticulate events can not be tangled. This can be made more precise by using the term *blob*, which is a subnetwork that contains at least one reticulate node and cannot be disconnected by deleting a single edge. In a level-1 network, each blob may contain only one reticulate node and hence consists of a single cycle. Hence, any two cycles in a level-1 network must not overlap. Unfortunately, this restriction excludes many realistic evolutionary scenarios (e.g. Ge et al., 1999; Westenberger et al., 2005; Sessa et al., 2012; Raghavan et al., 2014; Szöllősi et al., 2015; Wang et al., 2024). For example, consider the phylogenetic networks for orchids depicted in Figure 1. The network to the left is directed. Since the root location is not identifiable under many models, we will consider *semi-directed* networks, in which the root is suppressed (such networks can often be rooted in practice using an outgroup (Solís-Lemus and Ané, 2016)). The underlying semi-directed network for this case is depicted to the right. These networks are not level-1, since they contain overlapping cycles. In fact, these networks are level-2 since each network has a single blob with 2 reticulation nodes. In general, a binary network in which each blob contains at most 2 reticulate nodes is called *level*-2. Hence, level-2 networks can be seen as networks with isolated reticulations and pairs of tangled reticulations. Although the difference between level-1 and level-2 seems small, tangled reticulations lead to major mathematical and computational challenges especially when considering identifiability.

When studying phylogenetic networks beyond level-1 (and beyond level-2), a first crucial step is to determine the identifiability of the *tree-of-blobs*, the tree obtained from a network by contracting each blob to a node (Gusfield et al., 2007; Allman et al., 2023, 2024a). This tree describes the global branching structure of a phylogenetic network, see Figure 2. It is important to note that reconstructing a tree-of-blobs is fundamentally different from reconstructing a phylogenetic tree. Although the tree-of-blobs is a tree, this approach does not assume that the underlying evolutionary history is purely treelike, but uses a model of evolution including reticulation.

1.1. **Related work.** There is a substantial body of literature on the identifiability of phylogenetic networks, with studies differing in the types of models and data considered.

A prominent line of research focuses on network inference from gene tree distributions, where gene trees arise under the multispecies coalescent process (Yu et al., 2011). In this context, the problem of identifiability is analyzed using gene tree probabilities and related summary statistics such as quartet concordance factors (Baños, 2019). In early work, Solís-Lemus and Ané (2016) explored when reticulation could be detected from quartet concordance factors by showing that an n-taxon binary level-1 network with a cycle of length k is distinguishable from a network with the same topology but one reticulation removed if $n \geq 5$ and $k \geq 4$. Later, Baños (2019) showed that for binary level-1 networks, cycles of size at least four and reticulate nodes in cycles of size

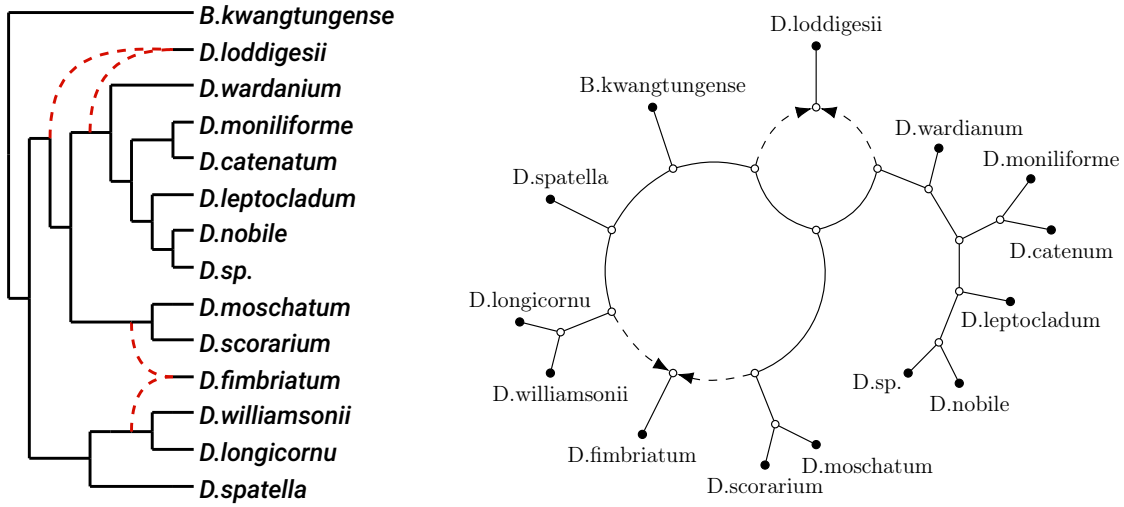

FIGURE 1. *Left:* A directed phylogenetic level-2 network on 13 orchids from the genus *Dendrobium* and the orchid *Bulbophyllum kwangtungense*, where the directions go from left to right (Wang et al., 2024). The complete name of *D.catenatum* is *D.catenatum SRR17026697* but was shortened for readability. *Right:* The semi-directed phylogenetic level-2 network obtained from the directed network by suppressing its root and only letting the dashed reticulation edges keep their direction.

at least five are identifiable from quartet concordance factors. This work was followed up in the literature by Allman et al. (2023), who showed that the tree-of-blobs is recoverable from quartet concordance factors. Moreover, Rhodes et al. (2025) showed that for outer-labeled planar blobs (blobs that can be drawn with no crossing edges and pendant edges on the outside) the circular ordering of the pendant edges around the blob is also recoverable from quartet concordance factors. Most recently, Allman et al. (2024c) investigated triangles (3-cycles) and showed that the presence of a triangle can be detected, but identification of the reticulate node depends on the numerical parameters.

On the other hand, there have been a number of papers which specifically consider Markov models of evolution of aligned sequence data on a network. The Markov models that were considered usually assume that each column of the alignment evolves independently down a tree displayed by the network. This contrasts with coalescent-based models, which can generate any tree under a given network. The complexity of the Markov models also depends on the constraints on the transition matrices. The simplest models use the Cavendar-Farris-Neyman (CFN) and Jukes-Cantor (JC) constraints, while the the generalised time-reversible (GTR) model and the general Markov model are more complex.

Gross and Long (2018) proved that binary large-cycle networks, which are binary semi-directed phylogenetic networks with a single cycle of length greater than three, are generically identifiable under the JC constraints. They proved their result by considering the algebraic varieties that come from the parameterizations of the network. Hollering and Sullivant (2021) showed that pairs of phylogenetic trees are generically identifiable for 2-tree CFN and Kimura 3-Parameter (K3P) mixtures, where each column evolves down one of two trees. They also showed that binary large-cycle networks are generically identifiable under each of the K2P and K3P constraints. Their work is focused on a method that uses algebraic matroids which are computed using network parametrizations. Gross et al. (2021) showed that binary semi-directed level-1 networks are

generically identifiable under each of the JC, K2P and K3P constraints, assuming the network is triangle-free and the number of reticulation vertices is known. While most results have focused on level-1 networks, there has been some initial work on level-2 networks by Ardiyansyah (2021), who considers the algebraic varieties of a restricted class of binary level-2 networks that satisfy a number of combinatorial conditions on their reticulation vertices to establish identifiability for the restricted class under the JC, K2P and K3P constraints.

Furthermore, Francis and Moulton (2018) considered a different extension of the tree-based Markov model to networks, where blocks of columns of the alignment evolve down the same displayed tree and there is a fixed probability to start a new block evolving down another displayed tree. For this model they showed an identifiability result for the class of rooted binary tree-child networks with the same number of reticulation vertices, at the cost of fixing the parameter values of the evolutionary model.

1.2. **Our contribution.** Despite the range of different modeling scenarios that have been considered, the complexity of previous identifiability results for phylogenetic network models has been constrained to either level-1 networks, or the tree-of-blobs of a network. In our work, we were interested in extending results beyond level-1 networks. Our main result is the following:

Theorem 3.2. *The network parameter of a network-based Markov model under the Jukes-Cantor constraints is generically identifiable with respect to the class of models where the network parameter is an n-leaf binary, triangle-free, strongly tree-child, level-2 semi-directed phylogenetic network.*

The main breakthrough in Theorem 3.2 is that it allows for tangled pairs of reticulated events by relaxing the level-1 restriction. In addition, it does not require the number of reticulate events to be known. On the other hand, since the networks are restricted to level-2, more complicated structures are not allowed, such as triples of reticulations that are all tangled in a single blob. There are two additional restrictions on the networks. The first is the triangle-free assumption, which forbids reticulate events between a node and a direct ancestor of that node. Previous results show that these scenarios are often hard to identify (Gross and Long, 2018; Baños, 2019; Allman et al., 2022; Xu and Ané, 2023). The second is the strongly tree-child condition which disallows a reticulate event that directly follows another reticulate event (which is also known to be hard to identify (Sullivant, 2025)) as well as nodes that are parents of two reticulate events (less likely scenarios for certain types of data (Kong et al., 2022)).

The proof requires a mix of techniques using tools from algebra and combinatorics. On the algebra side, we find phylogenetic invariants that can be used to distinguish between certain pairs of small networks, and use the matroid-based approach (Hollering and Sullivant, 2021) to rule out other pairs. In addition, we derive a new inequality for trinets (3-leaf networks), which can rule out further pairs of networks under JC. Once we have established a set of distinguishing results for pairs of small networks on 3, 4 and 5 leaves, we are able to use the combinatorial structure of level-2 networks to prove Theorem 3.2. We note that in previous work by Huber et al. (2024), it was shown that all binary semi-directed level-2 networks are uniquely encoded by their induced quarnets (4-leaf networks). However, for the case of binary semi-directed level-2 triangle-free tree-child networks, it is not true that they can be uniquely recovered from their induced triangle-free tree-child quarnets, as we will later see in Figure 5. Thus, this requires us to prove some distinguishing results beyond just quarnets and beyond just triangle-free tree-child quarnets.

Along the way, as part of our proofs, we see a number of other results that are interesting in their own right and hold for networks of any level. First of all, we show that under the JC and K2P constraints, if two networks have different sets of displayed quartet trees, then they are

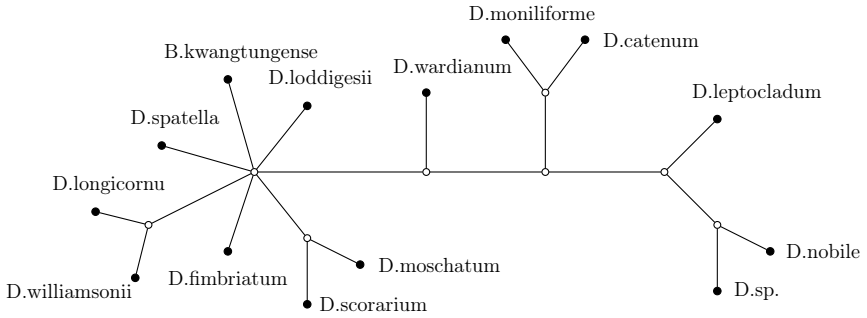

FIGURE 2. The tree-of-blobs of the semi-directed level-2 network on 14 orchids from Figure 1.

distinguishable. By combining our result with results from Allman et al. (2023); Huber et al. (2024); Rhodes et al. (2025), it follows that under JC and K2P constraints, the tree-of-blobs of any binary semi-directed network is identifiable as well as the circular ordering of the subnetworks around outer-labeled planar blobs.

We also improve the results of Gross et al. (2021), by strengthening the identifiability results for level-1 networks. We show that, under JC, we do not need to assume that the networks are triangle free, as long as we do not need to identify the reticulate node in a triangle. Moreover, under both JC and K2P, we lift the assumption, used by Gross et al. (2021), that the number of reticulate nodes is fixed. This means that we can even distinguish between level-1 networks with different numbers of reticulate events. This could be important since the number of reticulate events is in practice usually not known in advance. In particular, it means that under JC we can distinguish between tree-like and level-1 network-like evolution. We show that this even holds for level-2 and conjecture that it can be generalized to all levels.

The outline of this paper is as follows. In the next section, we cover preliminary background on networks, Markov models on trees and networks, and identifiability including the algebraic approaches to identifiability we will use. In Section 2.2, we use linear invariants and inequalities to show that under the JC and K2P constraints, two networks that have different displayed quartet trees are distinguishable. In Section 2.3 we use phylogenetic invariants and matroids to distinguish certain pairs of small quarnets which will be used in later proofs. In Section 2.4, we derive a new trinet inequality that can distinguish a 3-leaf tripod tree from a level-1 or level-2 trinet under the JC model. In Section 2.5, we use combinatorial techniques to distinguish level-2 networks with any number of leaves within two specific classes. Then, in the Results section, we put all these results together to strengthen the identifiability results for level-1 and to prove our main result for identifiability of a family of level-2 networks under JC. We end with a Discussion section.

2. Materials & Methods

2.1. **Preliminaries.**

2.1.1. *Phylogenetic networks.* In this section, we introduce concepts from graph theory that we utilize to study phylogenetic networks. Before we formally describe what a phylogenetic network is, we first define a *blob* of a (partially) directed graph as a maximal connected subgraph without any cut-edges, i.e. edges whose removal disconnect the graph. Such a blob is an m-blob for some positive integer m if the blob is incident to exactly m vertices not in the blob. A blob is *trivial* if it consists of a single vertex.

Definition 2.1. A *directed (binary phylogenetic) network* N^+ on a set of leaves $\mathcal{X}$ is a directed acyclic graph without parallel edges, such that the out-degree and in-degree of every vertex sum at most to three, where $\mathcal{X}$ is the set of vertices with out-degree zero, which all have in-degree one, there is a unique vertex with in-degree zero, which has out-degree two and is called the root, and there are no 2-blobs and no non-leaf 1-blobs.

We call a vertex in N^+ with in-degree one and out-degree two a *tree vertex* and a vertex in N^+ with in-degree two and out-degree one a *reticulation vertex*. The network N^+ is a *rooted (binary) phylogenetic tree* if the network does not contain any reticulation vertices. We call a non-leaf vertex in N^+ *internal* and an edge directed into a reticulation vertex a *reticulation edge*. A pair of reticulation vertices v, w is called *stacked*, if one is a child of the other. Note that every cut-edge e of a phylogenetic network N induces a bi-partition $A \cup B$ of its leaves $\mathcal{X}$, in which case we say that e induces the *split* $A|B$. The split is *nontrivial* if both $|A|, |B| \geq 2$.

Definition 2.2. A *semi-directed (binary phylogenetic) network* N on a set of leaves $\mathcal{X}$ is a partially directed graph that can be obtained from a directed binary phylogenetic network on $\mathcal{X}$ by removing edge directions of non-reticulation edges, suppressing all vertices of degree two, and identifying parallel edges.

Reticulation vertices, reticulation edges, internal vertices, splits, and leaves are defined analogously for semi-directed networks. A semi-directed network without reticulation vertices is called an *unrooted (binary) phylogenetic tree*. We say that a vertex v is *below* a reticulation vertex r in a semi-directed network N if all shortest paths between r and v do not contain reticulation edges directed into r. The vertex v is *directly below* r if v and r are also adjacent. Because of the correspondence between reticulation vertices in directed networks and semi-directed networks, we can define the following properties. We note that some of these are independent of the network being a directed network or a semi-directed network, while others are not.

Definition 2.3. Let N^+ be a directed phylogenetic network (respectively, N a semi-directed phylogenetic network), and let k be some non-negative integer. Then, we call the network N^+ (respectively, N)

(i) *level-k* if there exists a maximum of k reticulation vertices in each blob of the network;
(ii) *strict level-k* if the network is level-k but not level-$(k-1)$;
(iii) *stack-free* if none of the reticulation vertices of the network are stacked;
(iv) *simple* if the network has a single non-leaf blob;
(v) *triangle-free* if the network contains no cycles of length three, where edge directions are disregarded.

Furthermore, N^+ is *tree-child* if every internal vertex has at least one child that is either a tree vertex or a leaf. We say that N is *weakly tree-child* if there exists a directed network from which N can be obtained that is tree-child, whereas N is *strongly tree-child* if every such directed network is tree-child.

To illustrate some of these definitions, consider Figure 3, which depicts a tree-child directed network N^+ and the strongly tree-child semi-directed network N that can be obtained from N^+. Both networks are strict level-2, simple, triangle-free and stack-free. Note that, while the directed network N^+ is not the unique directed phylogenetic network from which N can be obtained, all possible rootings of N result in a tree-child network.

We obtain the *tree-of-blobs* of a semi-directed network N by contracting every blob to a single vertex. Observe that a tree-of-blobs is an unrooted (possibly non-binary) phylogenetic tree on $\mathcal{X}$ because semi-directed networks do not contain 2-blobs.

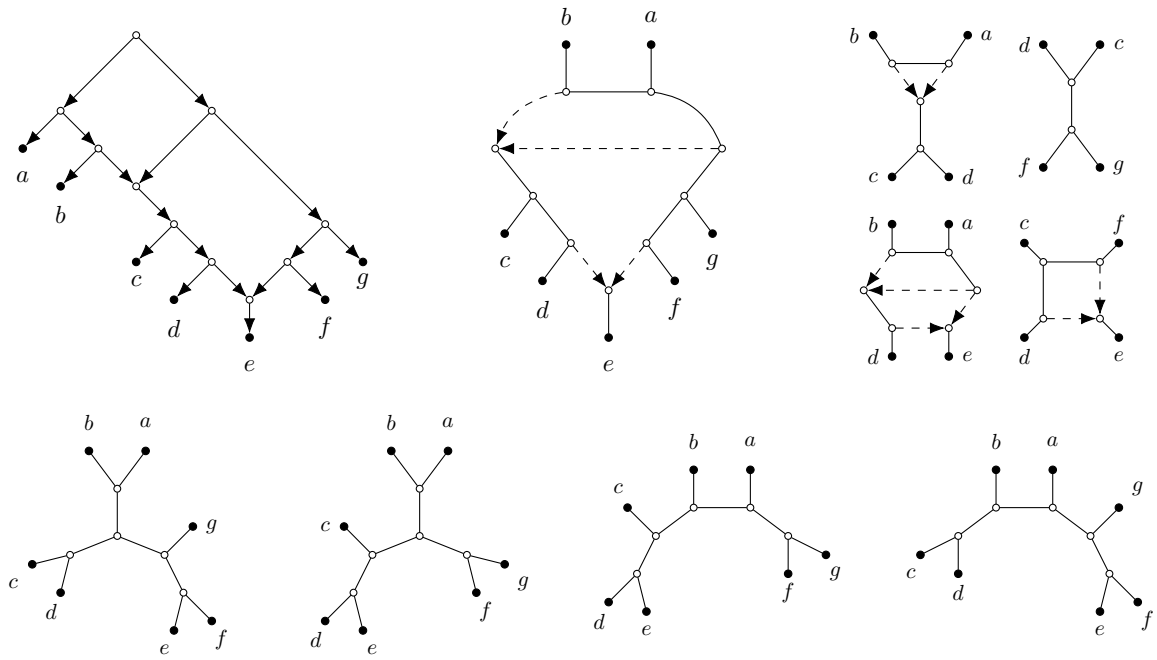

FIGURE 3. *Top left:* A directed phylogenetic network N^+ on $\mathcal{X} = \{a, b, c, d, e, f, g\}$. *Top middle:* The semi-directed phylogenetic network N obtained from N^+. *Top right:* Four of the quarnets induced by N. *Bottom:* The four unrooted phylogenetic trees on $\mathcal{X}$ displayed by N.

A *switching* of a directed network N^+ is obtained from N^+ by deleting exactly one incoming edge per reticulation node. Similarly, a *switching* of a semi-directed network N is obtained from N by deleting exactly one incoming edge per reticulation node and omitting the direction of the other incoming edge. We say that a (rooted or unrooted) phylogenetic tree T on X' is *displayed* by a (directed or semi-directed, respectively) network N on $\mathcal{X}$ if $X' \subseteq \mathcal{X}$ and T can be obtained from a subtree of a switching of N by edge contractions. See Figure 3 for an example of all unrooted phylogenetic trees on $\mathcal{X}$ displayed by a semi-directed network on $\mathcal{X}$. Note that a switching of N may have degree-2 vertices and leaves that are not in $\mathcal{X}$, while a displayed tree cannot have degree-2 vertices and must have leaf set $X' \subseteq \mathcal{X}$.

An *up-down path* between two leaves x_i and x_j of a semi-directed network is a path of k edges where the first l edges are directed towards x_i and the last $k - l$ edges are directed towards x_j. Here, we consider undirected edges to be bidirected. Using this notion, we can formally define the *restriction* of a semi-directed network.

Definition 2.4. Given a semi-directed network N on $\mathcal{X}$ and $X \subset \mathcal{X}$ with $|X| \geq 2$, we define the *restriction of N to X* as the semi-directed network $N|_X$ obtained from N by taking the union of all up-down paths between leaves in X, followed by exhaustively suppressing all vertices of degree two, identifying parallel edges and suppressing 2-blobs.

We suppress vertices of degree two, identify parallel edges, and suppress 2-blobs, because the resulting model network yields the same distributions as the network with those other features (see Sullivant (2025)). Intuitively, the restriction of a semi-directed network can be viewed as is its induced subnetwork. If $|X| = 3$, $|X| = 4$ or $|X| = 5$, then we call $N|_X$ a *trinet*, *quarnet* or

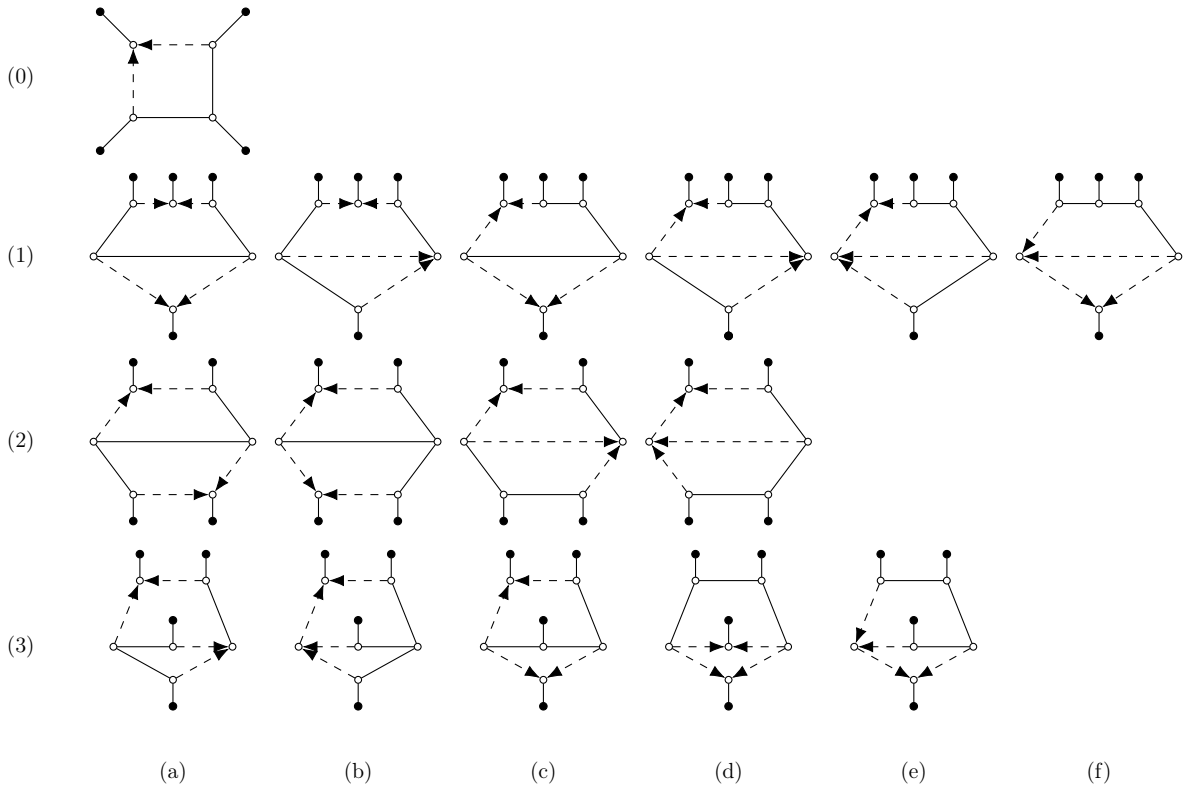

FIGURE 4. All sixteen possible semi-directed simple level-2 quarnets up to labeling the leaves (indicated by the filled black vertices).

quinnet induced by N, respectively. See Figure 3 for an example of four quarnets induced by a semi-directed network.

Proving identifiability by restricting to (level-2) trinets, quarnets, and quinnets will be a key idea in our proofs. Figure 4 shows all sixteen possible semi-directed simple level-2 quarnets, up to labeling the leaves. Whenever we refer to a quarnet of a specific type, we use the annotations as in this figure. For example, the quarnet of type 0 is the only simple quarnet that is strict level-1. Furthermore, the quarnets of type 0, 2 and 3 are triangle-free, while the quarnets of type 1 contain a triangle. The quarnets of type 0, 1a, 1b, 2a and 3a are the only strongly tree-child quarnets in the figure.

Definition 2.5. Let N be a semi-directed network. Then, we classify N as

(i) *outer-labeled planar* if there exists a planar representation of N in which all leaves of N lie in the unbounded face;
(ii) *strongly outer-labeled planar* if N is outer-labeled planar and simple.

A *circular order* of $\mathcal{X}$ is an order of its elements up to reversal and cyclic permutations. For an outer-labeled planar, semi-directed network N with labeled leaves $\mathcal{X}$ and a circular order σ of $\mathcal{X}$, we say that N is *congruent with* σ if there exists an outer-labeled planar representation of N that induces the circular order σ of the leaves in $\mathcal{X}$. As shown by Rhodes et al. (2025), a strongly outer-labeled planar, semi-directed network is congruent with exactly one circular order of its leaves. For instance, the semi-directed network N in Figure 3 is outer-labeled planar, because in the planar representation shown, all of the leaves lie in the unbounded face. Moreover, N is

strongly outer-labeled planar and induces only the circular order (a, b, c, d, e, f, g) of its leaves. Note that the quarnets of type 3 in Figure 4 are not outer-labeled planar.

To finish this subsection we introduce another useful concept. We define a *signature* $s = (x_i, x_j)$ for all simple semi-directed level-2 quarnets (see Figure 4) as the ordered tuple of taxa directly below its reticulation vertices (with possibly $x_i = x_j$). Precisely, if a quarnet only has one leaf x_i directly below a reticulation, its signature is (x_i, x_i). If instead a quarnet has two taxa x_i and x_j directly below a reticulation, the ordering in the signature is determined by the planar representation of the quarnet in Figure 4, where we consider the taxa from left to right and top to bottom. For example, if the quarnets 1c and 2b in Figure 4 are labeled with taxa $\mathcal{X} = \{a, b, c, d\}$ (starting at the top-left and then going clockwise), their signature is (a, d). On the other hand, quarnets 0, 1d and 2d (again labeled with $\mathcal{X} = \{a, b, c, d\}$, starting at the top-left and then going clockwise) have signature (a, a). Note that the signature of a simple semi-directed level-2 quarnet is unique for any fixed labeling of leaves.

2.1.2. *Network-based Markov models.* In this paper, we consider network-based Markov models, which are described most naturally via a directed phylogenetic network. However, for reversible models, such as the Jukes-Cantor and Kimura 2-parameter models considered here, any two directed networks N_1^+ and N_2^+ that have the same semi-directed network N will yield the same set of probability distributions (see proof of Proposition 1 in Gross et al. (2021)). A similar argument holds for networks with 2-blobs as well (Sullivant, 2025). Thus, we can use directed networks and semi-directed networks interchangeably when describing model parametrizations.

Let $\mathcal{S}_4$ be the set of 4×4 stochastic matrices and let $\Delta_{d-1} = \{p \in \mathbb{R}^d_{\geq 0} : \sum_{i=1}^d p_i = 1\}$ be the $(d-1)$-dimensional probability simplex. Given a directed phylogenetic network $N = (V, E)$, for each $v \in V$ we associate a random variable X_v with state space $\Sigma = \{A, G, C, T\}$, corresponding to the four DNA bases. We define the general Markov model by equipping each edge $e = u \to v \in E$ with a 4×4 transition matrix $M^e \in \mathcal{S}$ such that $M^e_{ij} = \Pr[X_v = j \mid X_u = i]$ for row i and column j indexed by Σ, and equipping the root ρ of N with a root distribution $\pi = (\pi_A, \pi_G, \pi_C, \pi_T) \in \Delta_3 \subset \mathbb{R}^4$.

For a switching T of N with vertex set $V(T)$ and edge set $E(T)$, let ϕ denote an assignment of $V(T)$ to states Σ. Then, the probability that assignment ϕ can be observed for switching T under our Markov model is

$$p(\phi) = \pi_{\phi(\rho)} \prod_{e \in E(T)} M^e_{\phi(u),\phi(v)}.$$

Now, we can calculate the probability of a site pattern $\omega \in \Sigma^n$ at the leaves of T in $\mathcal{X}$ by marginalizing probabilities $p(\phi)$. Specifically, let $\phi(\mathcal{X})$ denote the restriction of assignment ϕ to the leaves of T in $\mathcal{X}$. Then, the probability of observing site pattern ω in T is given by

$$p_\omega(T) = \sum_{\phi\,:\,\phi(\mathcal{X})=\omega} p(\phi) = \sum_{\phi\,:\,\phi(\mathcal{X})=\omega} \pi_{\phi(\rho)} \prod_{e \in E(T)} M^e_{\phi(u),\phi(v)}.$$

Finally, to derive the site pattern probability distribution of N, let k be the number of reticulation vertices of N and consider the set $\{T_1, \ldots, T_{2^k}\}$ of switchings we obtain from N by independently deleting exactly one edge e_k in each pair (e_i, e_j) of reticulation edges corresponding to one reticulation vertex with probability $\lambda_k \in (0, 1)$ such that $\lambda_i = 1 - \lambda_j$ and removing edge subdivisions. For $j \in \{1, \ldots, 2^k\}$, let $\zeta_j \in \{0, 1\}^m$ denote the indicator variable for the

edges from E present in T_j. Then, the probability of observing site pattern ω in N is given by

$$p_\omega(N) = \sum_{j=1}^{2^k} \left(\prod_{i=1}^{m} \lambda_i^{1-(\zeta_j)_i} (1-\lambda_i)^{(\zeta_j)_i} \right) p_\omega(T_j).$$

Thus, for the stochastic parameter space Θ_N of transition matrices M^e and root distribution π, we can view the general Markov model on N as the image of the polynomial map

$$\begin{aligned} \psi_N : \Theta_N &\to \Delta_{4^n-1} \\ (\pi, M^{e_1}, \ldots, M^{e_m}) &\mapsto (p_\omega(N))_{\omega \in \Sigma^n} \end{aligned}$$

The image of map ψ_N is called the *phylogenetic model associated to* N or a *phylogenetic model* if N is clear from the context, and is denoted $\mathcal{M}_N$. We extend ψ_N to be a complex polynomial map ψ'_N by considering the polynomial ring $\mathbb{C}[p_\omega]$ over indeterminates p_ω, $\omega \in \Sigma^n$.

Definition 2.6. The vanishing ideal or phylogenetic ideal of the model $\mathcal{M}_N$ is the set of polynomials

$$\mathcal{I}(\mathcal{M}_N) = \{f \in \mathbb{C}[p] : f(p) = 0 \ \ \forall p \in \mathcal{M}_N\}.$$

This is also called the *ideal of phylogenetic invariants* or the *phylogenetic ideal*.

Clearly, each phylogenetic invariant vanishes on all probability distributions in a given phylogenetic model. In addition, we call the Zariski closure of the image of ψ'_N the *phylogenetic variety*, denoted by V_N.

The model we have described up to this point is the general Markov model, because the parameter space allows for arbitrary 4×4 Markov transition matrices associated to each edge in the network. In this paper, we consider two submodels of the general Markov model, obtained by placing restrictions on both the transition matrices that are used, and the root distribution in the model. Under the *Jukes-Cantor model* (JC), we restrict the transition probabilities M^e_{ij} to be identical for all $i, j \in \Sigma$, $i \neq j$, $e \in E$. Since the transition matrices are stochastic, this means the Markov model is parameterized by exactly one parameter for each edge in the network. Under the Kimura 2-parameter model (K2P), we add one more parameter to only differentiate between probabilities of transitions (changes between either states A and G or C and T) and transversions (all other state changes). Hence a general K2P transition matrix will appear as follows:

$$\begin{pmatrix} \alpha & \beta & \gamma & \beta \\ \beta & \alpha & \beta & \gamma \\ \gamma & \beta & \alpha & \beta \\ \beta & \gamma & \beta & \alpha \end{pmatrix}$$

where $\alpha + 2\beta + \gamma = 1$. The general JC transition matrix is obtained by setting $\gamma = \beta$. In both models, we require the root distribution to be the uniform distribution.

At first sight it seems daunting to find phylogenetic invariants p for any phylogenetic model because we need to verify that p vanishes in all site pattern probabilities $p_\omega(N)$ of the phylogenetic model which are parameterized by a sum of $2^k 4^{n+2r-1}$ terms. However, a basis change into Fourier coordinates can simplify this task significantly for group-based models, including the Jukes-Cantor and Kimura 2-parameter models. The Fourier transform (or Hadamard transform), is a linear change of coordinates that simplifies the parametrization and hence the calculation of phylogenetic invariants.

Now, we describe the Fourier parameterizations of site pattern probabilities $p_\omega(T)$ following Sturmfels and Sullivant (2005). First, we identify our state space Σ with the Klein four-group $\mathbb{Z}_2 \times \mathbb{Z}_2$ as follows: $A = (0,0)$, $G = (1,0)$, $C = (0,1)$, $T = (1,1)$. We associate four Fourier

parameters with each edge $e \in E$, denoted by a_A^e, a_G^e, a_C^e and a_T^e. In addition, under the JC model we have that $a_C^e = a_G^e = a_T^e$, and under the K2P model we have that $a_G^e = a_T^e$. Furthermore, $a_A^e = 1$ for biological parameters. All other parameters are in the interval $(0, 1)$, explicitly excluding parameters being equal to 1, which corresponds to a zero branch length. Then, for site pattern $\omega = (g_1, g_2, \ldots, g_n)$, the Fourier parametrization of $p_\omega(T)$ is given by

$$q_\omega(T) = \begin{cases} \prod\limits_{\substack{e \in E(T) \\ e \text{ induces the split } A|B}} a^e_{\sum_{i \in A} g_i} & \text{if } \sum\limits_{i=1}^{n} g_i = 0, \\ 0 & \text{otherwise} \end{cases}$$

where addition is in the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. For any edge $e \in E$ which induces the split $A|B$, we also write Fourier parameter a_s^e as $a_s^{A|B}$ for all $s \in \Sigma$.

We conclude with the fact that for the JC and K2P models, all probability distributions in a phylogenetic model of a tree can be parameterized by a monomial with one parameter per edge in the tree. Therefore, distributions from a phylogenetic model of a network can be parameterized by a sum of monomials, one for each tree T_j, $j \in \{1, \ldots, 2^k\}$, weighted by the corresponding reticulation edge probabilities.

2.1.3. *Generic identifiability and distinguishability.* Our paper seeks to answer questions about the distinguishability and generic identifiability of level-2 phylogenetic networks. Below we will discuss the definitions and the connection between these two concepts.

Definition 2.7. Let $\{\mathcal{M}_N\}_{N \in \mathcal{N}}$ be a class of phylogenetic network models and $\psi_N : \Theta_N \to \Delta_{4^n-1}$ be the parameterization map for a phylogenetic network model. Given two distinct n-leaf networks N_1 and N_2, we say they are *distinguishable* if the set of numerical parameters in Θ_{N_1} that ψ_{N_1} maps into $\mathcal{M}_{N_2}$ is a set of Lebesgue measure zero and the set of numerical parameters Θ_{N_2} that ψ_{N_2} maps into $\mathcal{M}_{N_1}$ is a set of Lebesgue measure zero.

Since our models are expressed as the images of polynomial maps, the set of parameter values that map $\Gamma_{N_1} \subseteq \Theta_{N_1}$ into $\mathcal{M}_{N_2}$ will be a semialgebraic set. That set will have measure zero inside Θ_{N_1} if and only if it is a semialgebraic set of strictly smaller dimension than Θ_{N_1}. So taking the image of Γ_{N_1} by ψ_{N_1} must also give a subset of $\mathcal{M}_{N_2}$ of strictly smaller dimension. Hence, for algebraic phylogenetic models, we have the following equivalent definition of distinguishability.

Definition 2.8. Given a class of algebraic phylogenetic network models $\{\mathcal{M}_N\}_{N \in \mathcal{N}}$ and two distinct n-leaf networks N_1 and N_2, we say they are *distinguishable* if

$$\dim(\mathcal{M}_{N_1} \cap \mathcal{M}_{N_2}) < \min(\dim(\mathcal{M}_{N_1}), \dim(\mathcal{M}_{N_2})).$$

Below we give a definition, which we use to connect distinguishability and generic identifiability.

Definition 2.9. The network parameter of a phylogenetic network model is *generically identifiable* if given any two distinct n-leaf networks $N_1, N_2 \in \mathcal{N}$, the networks N_1 and N_2 are distinguishable.

In summary, distinguishability concerns a pair of network models and generic identifiability tells us if we always have pairwise distinguishability within a class of models.

One of our main tools for proving identifiability results for phylogenetic network models will be to look at polynomial invariants for those models and the vanishing ideals these invariants generate. Here we briefly introduce phylogenetic invariants and explain how those algebraic methods can be useful for proving identifiability results. In particular, we have the following results that relate the existence of phylogenetic invariants to distinguishability.

Proposition 2.10 (Prop. 16.1.12, Sullivant (2023))**.** *Let $\mathcal{M}_{N_1}$ and $\mathcal{M}_{N_2}$ be two phylogenetic network models on the same number of leaves. If there exist polynomials f_1 and f_2 such that $f_1 \in \mathcal{I}(\mathcal{M}_{N_1}) \backslash \mathcal{I}(\mathcal{M}_{N_2})$ and $f_2 \in \mathcal{I}(\mathcal{M}_{N_2}) \backslash \mathcal{I}(\mathcal{M}_{N_2})$ then*

$$\dim(\mathcal{M}_{N_1} \cap \mathcal{M}_{N_2}) < \min(\dim(\mathcal{M}_{N_1}), \dim(\mathcal{M}_{N_2})).$$

In particular, networks N_1 and N_2 are distinguishable.

If the two models $\mathcal{M}_{N_1}$ and $\mathcal{M}_{N_2}$ have the same dimension, it is sufficient to find a single polynomial to distinguish the two networks.

Proposition 2.11. *Let $\mathcal{M}_{N_1}$ and $\mathcal{M}_{N_2}$ be two phylogenetic network models on the same number of leaves. Suppose that $\dim(\mathcal{M}_{N_1}) = \dim(\mathcal{M}_{N_2})$. If there exist a polynomial $f_1 \in \mathcal{I}(\mathcal{M}_{N_1}) \backslash \mathcal{I}(\mathcal{M}_{N_2})$ then*

$$\dim(\mathcal{M}_{N_1} \cap \mathcal{M}_{N_2}) < \min(\dim(\mathcal{M}_{N_1}), \dim(\mathcal{M}_{N_2})).$$

In particular, networks N_1 and N_2 are distinguishable.

Sometimes it can be difficult or computationally expensive to actually find phylogenetic invariants, and an alternate strategy based on algebraic matroids is useful for proving the distinguishability of two networks. Matroids provide us with a generalization of linear independence for polynomial matrices. The matroid method can be performed by considering the ranks of submatrices of the Jacobians of two models. This strategy is described in Section 3 of Hollering and Sullivant (2021).

Definition 2.12. Let $\psi_N : \Theta_N \to \Delta_{4^n-1}$ be the parameterization for a phylogenetic network model. We label the model parameters via the vector $(\theta_1, \dots \theta_k)$. The Jacobian of ψ_N is the matrix $J(\psi_N)$ whose (i, j)-th entry is the derivative of the ith coordinate of ψ_N with respect to θ_j.

For a Jacobian matrix $J(\psi_N)$ and a subset S of Σ^n, let $J(\psi_N)_S$ denote the submatrix of $J(\psi_N)$ consisting of the rows indexed by S.

Proposition 2.13. *Let $\mathcal{M}_{N_1}$ and $\mathcal{M}_{N_2}$ be two phylogenetic network models on the same number of leaves. Suppose there are subsets of $S_1, S_2 \subseteq \Sigma^n$ such that*

$$\operatorname{rank} J(\psi_{N_1})_{S_1} < \operatorname{rank} J(\psi_{N_2})_{S_1} \qquad \operatorname{rank} J(\psi_{N_2})_{S_2} < \operatorname{rank} J(\psi_{N_1})_{S_2}.$$

Then

$$\dim(\mathcal{M}_{N_1} \cap \mathcal{M}_{N_2}) < \min(\dim(\mathcal{M}_{N_1}), \dim(\mathcal{M}_{N_2})).$$

In particular, networks N_1 and N_2 are distinguishable.

Similarly to finding phylogenetic invariants, if the two networks have models of the same dimension, then it suffices to check only one direction.

Proposition 2.14. *Let $\mathcal{M}_{N_1}$ and $\mathcal{M}_{N_2}$ be two phylogenetic network models on the same number of leaves, and suppose that $\dim(\mathcal{M}_{N_1}) = \dim(\mathcal{M}_{N_2})$. Suppose there is a subset $S \subseteq \Sigma^n$ such that*

$$\operatorname{rank} J(\psi_{N_1})_S < \operatorname{rank} J(\psi_{N_2})_S.$$

Then

$$\dim(\mathcal{M}_{N_1} \cap \mathcal{M}_{N_2}) < \min(\dim(\mathcal{M}_{N_1}), \dim(\mathcal{M}_{N_2})).$$

In particular, networks N_1 and N_2 are distinguishable.

Propositions 2.13 and 2.14 are a combination of Proposition 3.1 and Algorithm 3.3 in Hollering and Sullivant (2021).

A major tool for proving distinguishability of a pair of networks is that we can restrict to subsets of the leaves, prove distinguishability of the restricted networks on those subsets, and then deduce distinguishability of the larger networks. This is summarized in the following proposition (see Proposition 4.3 of Gross and Long (2018) for a related result on distinguishing phylogenetic varieties).

Proposition 2.15. *Let N_1 and N_2 be two semi-directed networks on the same leaf set $\mathcal{X}$ under a certain phylogenetic model. Let $S \subset \mathcal{X}$, and suppose that $N_1|_S$ and $N_2|_S$ are distinguishable. Then N_1 and N_2 are distinguishable.*

Proof. Restriction of a network model to a set of leaves S gives a linear map from the probability simplex $\Delta_{4^{|\mathcal{X}|}-1}$ to the probability simplex $\Delta_{4^{|S|}-1}$, that maps $\mathcal{M}_N$ onto $\mathcal{M}_{N|_S}$. This is because the restriction network is the network that is produced by summing out over the states of the leaves that are not in S. Since the maps are linear and the models are irreducible semialgebraic sets, if

$$\dim(\mathcal{M}_{N_1|_S} \cap \mathcal{M}_{N_2|_S}) < \min(\dim(\mathcal{M}_{N_1|_S}), \dim(\mathcal{M}_{N_2|_S}))$$

then also

$$\dim(\mathcal{M}_{N_1} \cap \mathcal{M}_{N_2}) < \min(\dim(\mathcal{M}_{N_1}), \dim(\mathcal{M}_{N_2}))$$

(otherwise a full dimensional intersection would be Zariski dense in the image, and hence would map to a full-dimensional subset of the intersection in the restriction). □

With Proposition 2.15 in mind, our strategy for proving identifiability of the network parameter for level-2 strongly tree-child triangle-free semi-directed networks becomes clear. We will show distinguishability of many pairs of small networks, using different strategies including computing phylogenetic invariants, matroid calculations, and finding inequalities that can distinguish the small networks. Then we will show that arbitrary size networks can be distinguished by considering restrictions to small sets of leaves.

2.2. **Distinguishing Networks with Different Displayed Quartets.** In this section we consider the linear invariants and inequalities that hold for phylogenetic mixture models and how they might be applied for phylogenetic network identifiability. These tools can be used for the JC and K2P Models. These invariants and inequalities for the JC and K2P model appeared in Allman et al. (2010, 2012) for proving identifiability of mixture models. However, since distributions in the network model are mixtures of distributions from phylogenetic tree models, those results can also be applied to deduce distinguishability results for network models. The main result of this section is to show that two networks that have different displayed quartets must be distinguishable under the JC and K2P models. This result holds regardless of the level of the networks. We use that as a tool to distinguish between many pairs of level-2 semi-directed networks as a first step to prove our main results. As a corollary, we additionally show that the tree-of-blobs of a network of any level is identifiable under these models (see Corollary 2.23).

Proposition 2.16. *Consider the quartet tree with the split* $12|34$. *Under the JC or K2P model, any distribution for the q in Fourier coordinates satisfies the invariant*

$$q_{GGGG} - q_{GGTT} = 0.$$

If q is a distribution from the quartet tree with split $13|24$ *or* $14|23$, *then*

$$q_{GGGG} - q_{GGTT} > 0.$$

Proof. In the first case, we have that

$$\begin{aligned} q_{GGGG} &= a_G^1 a_G^2 a_G^3 a_G^4 a_A^{12|34} \\ q_{GGTT} &= a_G^1 a_G^2 a_T^3 a_T^4 a_A^{12|34} \end{aligned}$$

However, since $a_G^3 = a_T^3$ and $a_G^4 = a_T^4$ for JC and K2P we see that $q_{GGGG} = q_{GGTT}$. On the other hand, for the tree with split $13|24$ we have

$$\begin{aligned} q_{GGGG} &= a_G^1 a_G^2 a_G^3 a_G^4 a_A^{13|24} \\ q_{GGTT} &= a_G^1 a_G^2 a_T^3 a_T^4 a_C^{13|24} \end{aligned}$$

Then, since $a_G^3 = a_T^3$ and $a_G^4 = a_T^4$, and $a_A^{13|24} = 1$ for JC and K2P we see that

$$q_{GGGG} - q_{GGTT} = a_G^1 a_G^2 a_G^3 a_G^4 (1 - a_C^{13|24}).$$

Since $a_C^{13|24} \in (0,1)$ we obtain that $q_{GGGG} - q_{GGTT} > 0$. A similar calculation shows that $q_{GGGG} - q_{GGTT} > 0$ for distributions that come from the tree with split $14|23$. □

Corollary 2.17. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two phylogenetic models for 4 taxa. Suppose that every probability distribution in $\mathcal{M}_1$ is a mixture of JC or K2P distributions that come from the tree with split $12|34$. Suppose that every probability distribution in $\mathcal{M}_2$ is a mixture of JC or K2P distributions from any four leaf trees and that in each such distribution either a distribution from the tree with split $13|24$ or a distribution from the tree with split $14|23$ appears with positive mixing weight (or both). Then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.*

Proof. This uses the linearity of mixture distributions and the fact that the conditions from Proposition 2.16 are linear. Indeed, according to the conditions, any distribution in $\mathcal{M}_1$ must satisfy the condition $q_{GGGG} - q_{GGTT} = 0$ whereas any distribution in $\mathcal{M}_2$ must satisfy $q_{GGGG} - q_{GGTT} > 0$. Hence $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$. □

Proposition 2.18. *Let q be any distribution in Fourier coordinates from either the quartet tree with split $12|34$ or $14|23$ under the JC or K2P model. Then q satisfies the linear relation*

$$q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} = 0.$$

On the other hand, if q is a distribution from the quartet tree with split $13|24$ then

$$q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} > 0.$$

Proof. This is along similar lines to the proof of Proposition 2.16. Under the tree $12|34$ we have

$$\begin{aligned} q_{GGGG} &= a_G^1 a_G^2 a_G^3 a_G^4 a_A^{12|34} \\ q_{GGTT} &= a_G^1 a_G^2 a_T^3 a_T^4 a_A^{12|34} \\ q_{GTTG} &= a_G^1 a_T^2 a_T^3 a_G^4 a_C^{12|34} \\ q_{GTGT} &= a_G^1 a_T^2 a_G^3 a_T^4 a_C^{12|34} \end{aligned}$$

Since $a_G^e = a_T^e$ for all edges e, we see that $q_{GGGG} = q_{GGTT}$ and $q_{GTTG} = q_{GTGT}$, so

$$q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} = 0.$$

Similarly, under the tree 14|23 we have

$$\begin{aligned} q_{GGGG} &= a_G^1 a_G^2 a_G^3 a_G^4 a_A^{14|23} \\ q_{GGTT} &= a_G^1 a_G^2 a_T^3 a_T^4 a_C^{14|23} \\ q_{GTTG} &= a_G^1 a_T^2 a_T^3 a_G^4 a_A^{14|23} \\ q_{GTGT} &= a_G^1 a_T^2 a_G^3 a_T^4 a_C^{14|23} \end{aligned}$$

so $q_{GGGG} = q_{GTTG}$ and $q_{GGTT} = q_{GTGT}$ and

$$q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} = 0.$$

On the other hand, under the tree 13|24 we have that

$$\begin{aligned} q_{GGGG} &= a_G^1 a_G^2 a_G^3 a_G^4 a_A^{13|24} \\ q_{GGTT} &= a_G^1 a_G^2 a_T^3 a_T^4 a_C^{13|24} \\ q_{GTTG} &= a_G^1 a_T^2 a_T^3 a_G^4 a_C^{13|24} \\ q_{GTGT} &= a_G^1 a_T^2 a_G^3 a_T^4 a_A^{13|24} \end{aligned}$$

so

$$q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} = 2a_G^1 a_G^2 a_G^3 a_G^4 (1 - a_C^{13|24})$$

since $a_G^e = a_T^e$ and $a_A^e = 1$ for all edges e. Since $a_C^{13|24} \in (0,1)$, this shows that

$$q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} > 0.$$

□

Corollary 2.19. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two phylogenetic models for 4 taxa. Suppose that every probability distribution in $\mathcal{M}_1$ is a mixture of JC or K2P distributions that come from the trees 12|34 and 14|23. Suppose that every probability distribution in $\mathcal{M}_2$ is a mixture of JC or K2P distributions from any four leaf trees and that each such distribution includes a distribution from 13|24 with positive mixing weight. Then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.*

Proof. This uses the linearity of mixture distributions and the fact that the conditions from Proposition 2.18 are linear. Indeed, according to the conditions, any distribution in M_1 must satisfy the condition $q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} = 0$ whereas any distribution in M_2 must satisfy $q_{GGGG} - q_{GGTT} - q_{GTTG} + q_{GTGT} > 0$. Hence $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$. □

Theorem 2.20. *Let N_1 and N_2 be two binary semi-directed phylogenetic networks on the same leaf set $\mathcal{X}$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of probability distributions produced by the Markov models on these networks under either the JC or K2P models. If there is a 4 element subset $S \subseteq \mathcal{X}$ such that the quarnets $N_1|_S$ and $N_2|_S$ have different sets of displayed quartets, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.*

Proof. It suffices to assume that N_1 and N_2 are themselves quarnets, since if the marginalizations are disjoint then so are the initial models $\mathcal{M}_1$ and $\mathcal{M}_2$. We will use Corollaries 2.17 and 2.19, because in the network Markov model, every distribution is a mixture of distributions from the switchings, or equivalently, from the displayed trees. We also assume that all mixing weights are always positive.

Let D_1 and D_2 be the sets of displayed quartets from networks N_1 and N_2 respectively, and we assume that $D_1 \neq D_2$. Suppose that one of D_1 or D_2 is a singleton. After symmetry, we can assume that this singleton is $D_1 = \{12|34\}$ and that D_2 contains one of the other quartet trees. Then Corollary 2.17 shows that $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.

So now assume that D_1 has two quartet trees, and D_2 has two or three trees. After symmetry, we can assume that $D_1 = \{12|34, 14|23\}$ and that $\{13|24\} \subseteq D_2$. Then Corollary 2.19 shows that $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$. □

Corollary 2.21. *Let N_1 and N_2 be two binary semi-directed level-2 phylogenetic networks on the same leaf set $\mathcal{X}$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of probability distributions produced by the Markov models on these networks under either the JC or K2P models. If there is a 4 element subset $S \subseteq \mathcal{X}$ such that the quarnet $N_1|_S$ is of type 3a, 3b, 3c or 3d and the quarnet $N_2|_S$ is not, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.*

Proof. First note that $N_1|_S$ displays exactly three quartet trees. On the other hand, $N_2|_S$ displays either one or two quartet trees. The result then follows from Theorem 2.20. □

The following result is presented in a slightly different form in Rhodes et al. (2025). For completeness, we prove part (a) in detail here as well, since the original source of the proof (Allman et al., 2023) presents only an implicit proof in a different context.

Proposition 2.22. *Let N be a binary semi-directed phylogenetic network on 4 leaves $\{a, b, c, d\}$. Then,*

- *(a) N induces the nontrivial split $ab|cd$ if and only if N displays solely the quartet with nontrivial split $ab|cd$;*
- *(b) if N is strongly outer-labeled planar (congruent with circular order (a, b, c, d)), then N displays exactly the two quartets with nontrivial splits $ab|cd$ and $ad|bc$.*

Proof. The only-if direction of part (a) is trivial since every quartet displayed by N has the cut-edge induced by the split $ab|cd$ in its embedding. Thus, all the quartets displayed by N have the split $ab|cd$.

For the if-direction of part (a), we use induction on the level k of the network and prove the result more generally for networks with parallel edges. Assume that N does not induce any nontrivial split. For $k = 0$ this is not possible and for $k = 1$ we can check manually that N then displays two quartets with different splits, so we are done. For $k \geq 2$, pick a reticulation r with a leaf x below it. Create two level $k - 1$ networks N_1, N_2 by deleting one of the two reticulation edges entering r, undirecting the other reticulation edge and exhaustively deleting degree-2 reticulations and suppressing non-reticulate degree-2 nodes. If at least one of N_1, N_2 does not induce a nontrivial split, then it displays at least two quartets by induction. Hence, N also displays at least two quartets. Now consider the remaining case, that both N_1 and N_2 induce a nontrivial split. Then it is sufficient to show that N_1 and N_2 display different nontrivial splits.

Suppose that N_1 and N_2 display the same nontrivial split, say $xy|wz$. Let N' be the mixed graph obtained from N by deleting x and exhaustively deleting resulting degree-2 reticulations. Note that N' is connected since otherwise N could not be oriented as a directed phylogenetic network. Let e be the cut-edge of N' inducing the split $y|wz$ with maximum distance from y. Let N'' be the connected component of the mixed graph obtained from N' by deleting e, such that N'' contains w and z. Since N does not have a cut-edge inducing the split $xy|wz$, it contains a directed path from N'' to the neighbour of x, which does not use e. Hence, N_i contains a directed path from N'' to the neighbour of x for some $i \in \{1, 2\}$ that does not use e. However, this implies that N_i does not have a cut-edge inducing the split $xy|wz$, a contradiction. □

In the following corollary we show that the tree-of-blobs of a semi-directed network of any level is identifiable under the JC and K2P models.

Corollary 2.23. *Let N_1 and N_2 be two binary semi-directed phylogenetic networks on the same leaf set X. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of probability distributions produced by the Markov models on these networks under either the JC or K2P constraints. If the trees-of-blobs of N_1 and N_2 are not isomorphic, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.*

Proof. In Huber et al. (2024) it is shown that the tree-of-blobs of a semi-directed network of any level is uniquely determined by the nontrivial splits of its quarnets. Since N_1 and N_2 have non-isomorphic trees-of-blobs, for some 4 element set $S \subseteq \mathcal{X}$ the quarnets $N_1|_S$ and $N_2|_S$ do not induce the same nontrivial split (with possibly one of them not inducing a nontrivial split at all). The result then follows from Theorem 2.20 and Proposition 2.22a. □

Corollary 2.24. *Let N_1 and N_2 be two binary semi-directed phylogenetic networks on the same leaf set $\mathcal{X}$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of probability distributions produced by the Markov models on these networks under either the JC or K2P constraints. If N_1 and N_2 are both outer-labeled planar and they have different sets of congruent circular orders, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.*

Proof. By the previous corollary, we may assume without loss of generality that N_1 and N_2 have the same trees-of-blobs. Then, for the two networks to have different sets of congruent circular orders, there must be a k-blob with $k \geq 4$ in the two networks which induces a different ordering of the subnetworks around it in N_1 than in N_2. Hence, we may assume without loss of generality that N_1 and N_2 are both simple networks, i.e. they consists of a single non-leaf blob. In Theorem 5.3 of Rhodes et al. (2025) it is shown that the circular order of a binary, simple, outer-labeled planar semi-directed phylogenetic network is uniquely determined by the circular orders of all its strongly outer-labeled planar quarnets. Since we assume that N_1 and N_2 are congruent with a different circular order, there is some 4-leaf set $A \subseteq \mathcal{X}$ such that $N_1|_A$ is congruent with a unique circular order and $N_2|_A$ is congruent with a different unique circular order or it is not congruent with a unique circular order. The result then follows from Theorem 2.20 and Proposition 2.22. □

2.3. **Distinguishing Small Networks using Invariants and Matroids.** In this section, we describe some results we will need to distinguish pairs of small networks, so that we can use them to prove identifiability results of the network parameters in later sections. The results in this section are computational: Lemma 2.25 is verified by a sequence of Macaulay2 calculations (Grayson and Stillman). Some of these results are produced by using the matroid method of comparing the ranks of submatrices of the Jacobians from the two networks. The other results are found by finding low degree phylogenetic invariants. Direct calculations of the vanishing ideals of these models seem out of reach at present, but we were able to use the *MultigradedImplicitization* toolbox in Macaulay2 to compute invariants of our networks (Cummings and Hollering, 2023). Parameterizations were computed using the algorithm available through (Grindstaff et al., 2025).

Lemma 2.25. *Let N_1 and N_2 be two binary level-2 semi-directed phylogenetic networks on the same leaf set $\mathcal{X}$. Let S be a 4 leaf subset of $\mathcal{X}$. Then, under the JC model:*

(a) if $N_1|_S$ and $N_2|_S$ are distinct quarnets of type 1b with respective signatures (x_i, x_j) that differ in x_i, then N_1 and N_2 are distinguishable.
(b) if $N_1|_S$ and $N_2|_S$ are distinct quarnets of type 2a, then N_1 and N_2 are distinguishable;
(c) if $N_1|_S$ and $N_2|_S$ are distinct quarnets of type 3a, 3c or 3d, then N_1 and N_2 are distinguishable.

Let $S = \{a, b, c, d, e\}$ be a 5 leaf subset of X. Then, under the JC model:

(d) if $N_1|_S$ and $N_2|_S$ are the two quinnets from Figure 5, then N_1 and N_2 are distinguishable.

Proof. For each of these results, it suffices to consider the case that S is the set of leaves of the networks under consideration, simply by marginalizing (see Proposition 2.15). Hence we do not need to write the $|_S$ in the calculations that we describe. For example, to prove (a) it suffices to show that if N_1 and N_2 are type 1b quarnets with respective signatures (x_i, x_j) that differ in x_i, then the networks are distinguishable.

The proof of all four of these results depend on computations that were carried out in Macaulay2 (Grayson and Stillman) using different strategies. For each part of the proposition there is a separate file verifying the result as an exact computation in Macaulay2. We explain how each of these computations work:

(a) This computation is verified in the file `Lemma225a.m2`. First of all, we only need to look at the different labelings of the leaves of quarnet 1b that have the same displayed trees. With the leaves labeled in circular order starting in the upper left, this amounts to looking at the eight circular orders

$$1234, 2341, 3412, 4123, 4321, 1432, 2143, 3214.$$

Here we want to show that each of these circular orders is distinguished from the others except in the case that the top middle leaf gets the same label, e.g. we do not need to distinguish the pair $1234, 3214$. In fact, separate computations show that that those pairs have the same ideal, and we suspect they cannot be distinguished.

For each of the pairs we want to distinguish, there is a subset of the variables that has the property that the corresponding columns of the Jacobian matrices under the networks have different ranks. These subsets are listed in "successList" in the file. For example, for the first pair (corresponding to the two orders $1234, 2341$, the set of columns of the Jacobian matrix

$$\{0, 1, 3, 4, 5, 6, 7, 9, 10\}$$

yield different ranks in the two networks. These correspond to the indeterminates

$$q_{AAAA}, q_{AACC}, q_{ACAC}, q_{ACGT}, q_{CACA}, q_{CAAC}, q_{CAGT}, q_{CGTA}, q_{CGTA}.$$

Running the code and producing the term "listofrankpairs" shows the ranks of the corresponding submatrices of the Jacobians for these two networks. It has rank 8 for ordering 1234 and rank 9 for ordering 2341.

(b) The proof of (b) follows a similar strategy as (a), and the corresponding computation is in the file `Lemma225b.m2`.

(c) The proof of (c) is established by finding phylogenetic invariants that distinguish each pair of networks. In the file `InvariantForThreeA.m2`, we give a degree 6 invariant with 53 terms that distinguishes quarnet 3a with the labeling 1234 from all other labeled 3a, 3b, and 3d quarnets. In the file `InvariantForThreeC.m2`, we give a degree 5 invariant with 34 terms that distinguishes the 3c quarnet with the labeling 1234 from all other labeled 3a, 3b, and 3d quarnets, and in the file `InvariantForThreeD.m2`, we give a degree 5 invariant with 50 terms that distinguishes the 3d quarnet with labeling 1234 from all other labeled 3a, 3b, and 3d quarnets. The computation verifying the statements is in the file `Lemma225c.m2`.

(d) The proof of (d) follows a similar strategy as (a), and the corresponding computation is in the file `Lemma225d.m2`. In this case there is just one pair of networks to compare. Considering the set of indeterminates

$$q_{AACCA}, q_{ACAAC}, q_{ACACA}, q_{ACCAA}, q_{CACAA}, q_{ACAGT},$$

$$q_{ACGTA}, q_{CAGAT}, q_{CAGTA}, q_{CGAAT}, q_{CGATA}, q_{CGTAA}$$

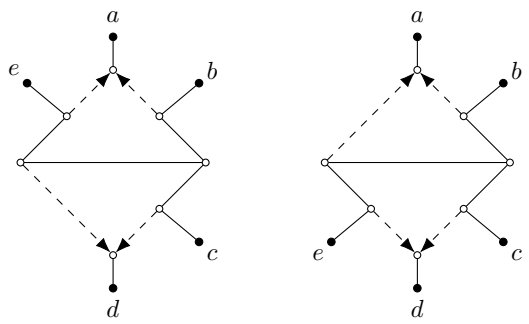

FIGURE 5. Two distinct simple, level-2 quinnets (semi-directed networks with 5 leaves) on leaf set $S = \{a, b, c, d, e\}$ that are distinguishable by Lemma 2.25d.

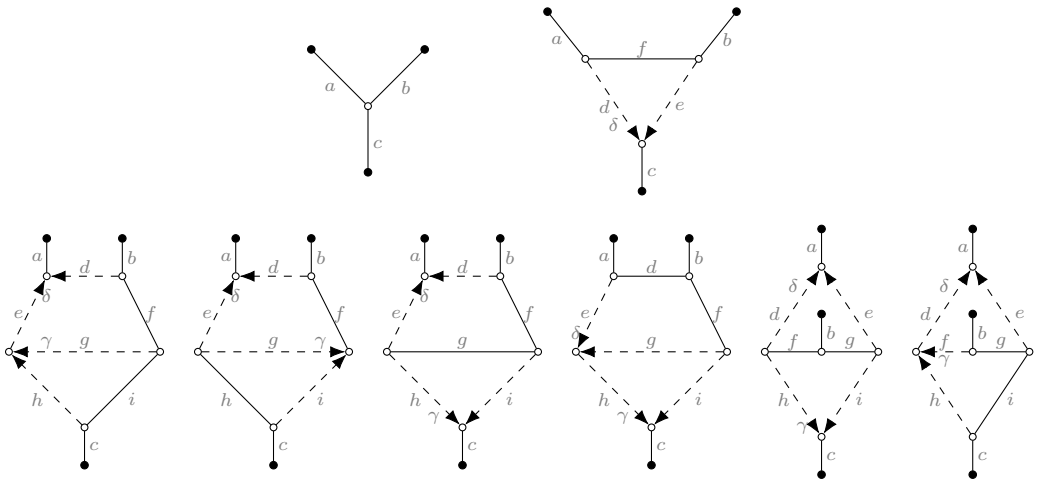

FIGURE 6. All eight possible simple level-2 trinets (semi-directed networks with 3 leaves) up to labeling the leaves (indicated by the black filled vertices). The gray edge labels are used for the Fourier parametrization in the proof of Proposition 2.26.

the code shows that for the left network, the corresponding rank of the submatrix of the Jacobian is 12, and for the right network it is 11. □

2.4. **Distinguishing Three-Leaf Trees from Trinets.** In this section, we describe an inequality for the JC model that allows us to distinguish a three-leaf tree from a level-1 or level-2 trinet. This allows us to prove distinguishability results for many pairs of level-2 networks, since many level-2 networks have induced subnetworks that are trees.

For the Fourier coordinates q under the Jukes-Cantor model for a trinet, we write q_{110} as shorthand for any of the values $q_{CCA} = q_{GGA} = q_{TTA}$, since all of these values are the same under the Jukes-Cantor. We define q_{101} and q_{011} similarly. We write q_{111} to denote any of $q_{CGT} = q_{CTG} = q_{GCT} = q_{GTC} = q_{TGC} = q_{TCG}$, as again all of these values are the same under the Jukes-Cantor model. This makes it visually easier to see the differences between the different q coordinates. We also simplify the notation for the Fourier parameters used as part of the proof. In the description in Section 2.1.2, the parameters were notated as a_g^e, where e is an edge and g is a group element. However, in the Jukes-Cantor model, we have that $a_A^e = 1$, and $a_C^e = a_G^e = a_T^e$.

So to simplify the expressions in our calculations we use a single letter to denote the parameter value $a^e_C = a^e_G = a^e_T$, and we use a different letter $a, b, c, \ldots$ for each edge in the network.

Proposition 2.26. *Let N_1 and N_2 be two binary semi-directed level-2 phylogenetic networks on the same three leaves. Suppose that N_1 is a three-leaf tree and N_2 is a strict level-1 or strict level-2 trinet. Then, under the JC model and if N_2 has no trivial parameter values, the polynomial invariant*

$$q_{011}q_{101}q_{110} - q_{111}^2$$

evaluates to zero for N_1 and is strictly positive for N_2. Hence, N_1 and N_2 are distinguishable under the JC model.

For this proof, we explicitly show that the inequality is satisfied for each of the level-2 trinets as displayed in Figure 6. We provide this proof for two of the trinets here, the other trinets follow a similar argument and their proof is shown in the Appendix.

Proof. We first show that $q_{011}q_{101}q_{110} - q_{111}^2$ equals 0 for the three-leaf tree N_1. To this end, let a, b and c be the nontrivial Fourier parameters associated to the three edges of N_1 (see also Figure 6). Then, under the JC model, the considered Fourier parametrizations of N_1 are $q_{111} = abc$, $q_{110} = ab$, $q_{101} = ac$ and $q_{011} = bc$. Clearly, we then get that $q_{011}q_{101}q_{110} - q_{111}^2 = 0$ for N_1.

Note that N_2 can either be the strict level-1 trinet, or one of the six strict level-2 trinets (see Figure 6). For reference, we categorize the strict level-2 trinets as types 1 through 6, ordered from left to right in Figure 6. We now show that the invariant is strictly positive if N_2 is the strict level-1 trinet or if it is the strict level-2 trinet of type 1. We defer the very similar calculations for the strict level-2 trinets of types 2, 3, 4, 5 and 6 to Appendix A.

Trinet 0: *N_2 is the strict level-1 trinet.* Let a-f be the nontrivial Fourier parameters associated to the edges of N_2 (see also Figure 6). Furthermore, let δ be the reticulation parameter. Then, N_2 has the following Fourier parametrizations under the JC model:

$$\begin{aligned} q_{111} &= abcf(\delta d + (1-\delta)e) \\ q_{110} &= ac(\delta d + (1-\delta)ef) \\ q_{101} &= abf \\ q_{011} &= bc(\delta df + (1-\delta)e). \end{aligned}$$

We compute that

$$\begin{aligned} q_{011}q_{101}q_{110} &= a^2b^2c^2f \cdot (\delta df + (1-\delta)e)(\delta d + (1-\delta)ef) \\ &= a^2b^2c^2f \cdot \left(\delta^2d^2f + \delta(1-\delta)de(1+f^2) + (1-\delta)^2e^2f\right), \end{aligned}$$

while

$$\begin{aligned} q_{111}^2 &= a^2b^2c^2f \cdot f(\delta d + (1-\delta)e)^2 \\ &= a^2b^2c^2f \cdot \left(\delta^2d^2f + \delta(1-\delta)de(2f) + (1-\delta)^2e^2f\right). \end{aligned}$$

Thus,

$$\begin{aligned} q_{011}q_{101}q_{110} - q_{111}^2 &= a^2b^2c^2f\delta(1-\delta)de(1+f^2-2f) \\ &= a^2b^2c^2f\delta(1-\delta)de(1-f)^2. \end{aligned}$$

Since all parameters are in $[0,1]$ in general, $q_{011}q_{101}q_{110} - q_{111}^2$ is strictly positive if $a, b, c, d, e, f > 0$ and $f, \delta \neq 1$.

Trinet 1: *N_2 is the strict level-2 trinet of type 1.* As depicted in Figure 6, let a-i be the nontrivial Fourier parameters associated to the edges of N_2. Furthermore, let δ and γ be the two reticulation parameters. Under the JC model, N_2 has the following Fourier parametrizations:

$$\begin{aligned} q_{111} &= abcfi(\delta \cdot e \cdot (\gamma g + (1-\gamma)h) + (1-\delta) \cdot d) \\ q_{110} &= ab(\delta \cdot ef \cdot (\gamma g + (1-\gamma)h) + (1-\delta)d) \\ q_{101} &= ac(\delta \cdot e \cdot (\gamma gi + (1-\gamma)h) + (1-\delta) \cdot dfi)) \\ q_{011} &= bcfi. \end{aligned}$$

We now compute $q_{110}q_{101}q_{011} - q_{111}^2$ and look at the coefficients for δ^2, $(1-\delta)^2$ and $\delta(1-\delta)$. The coefficient for δ^2 is

$$\begin{aligned} &a^2b^2c^2(\gamma g + (1-\gamma)h)\Big(ie^2f^2(\gamma gi + (1-\gamma)h) - f^2i^2e^2(\gamma g + (1-\gamma)h)\Big) \\ &= a^2b^2c^2(\gamma g + (1-\gamma)h)\Big(e^2f^2i(1-i)(1-\gamma)h\Big). \end{aligned}$$

The coefficient for $(1-\delta)^2$ is as follows:

$$a^2b^2c^2\Big(d^2f^2i^2 - d^2f^2i^2\Big) = 0.$$

Lastly, the coefficient for $(1-\delta)\delta$ becomes

$$\begin{aligned} &a^2b^2c^2\Big(defi \cdot [f^2i(\gamma g + (1-\gamma)h) + (\gamma gi + (1-\gamma)h)] - def^2i^2(\gamma g + (1-\gamma)h)\Big) \\ &= a^2b^2c^2defi\Big((\gamma gi + (1-\gamma)h) - fi(1-f)(\gamma g + (1-\gamma)h)\Big) \\ &= a^2b^2c^2defi\Big(if^2(\gamma g + (1-\gamma)h) + \gamma gi(1-f) + (1-\gamma)h(1-if)\Big). \end{aligned}$$

Provided that the parameter values are nontrivial (i.e. they are in $(0,1)$) the coefficients for δ^2 and $(1-\delta)\delta$ are positive, while the coefficient for $(1-\delta)^2$ was equal to zero. All in all, this shows that $q_{110}q_{101}q_{011} - q_{111}^2$ is positive for N_2. □

As a main application of the trinet inequality, we show that many quarnets are distinguished from each other.

Corollary 2.27. *Let N_1 and N_2 be two binary level-2 semi-directed phylogenetic networks on the same leaf set $\mathcal{X}$. Let S be a 4 leaf subset of $\mathcal{X}$. Then, under the JC model:*

(a) if $N_1|_S$ is a type 0, 1c, 1d, 1e, 1f, 2d, 3b or 3e quarnet and $N_2|_S$ is a type 1a, 1b, 2a, 2b, 2c, 3a, 3c or 3d quarnet, then N_1 and N_2 are distinguishable;

(b) if $N_1|_S$ and $N_2|_S$ are distinct quarnets of type 0, 1c, 1d, 1e, 1f, 2d, 3b or 3e with respective signatures (x_i, x_j) that differ in x_i, then N_1 and N_2 are distinguishable.

Proof. For any combination of $N_1|_S$ and $N_2|_S$ in the above two cases, there exists some set $T \subset S$ with $|T| = 3$ such that $N_1|_T$ is a three-leaf tree and $N_2|_T$ is not a three-leaf tree (i.e. it is a nontrivial trinet). The result then follows from Proposition 2.26. □

While we have only been able to find a proof of Proposition 2.26 for level-2 networks, and only by analyzing each network on a case by case basis, we conjecture that the analog of this result should hold for trinets of arbitrary level.

Conjecture 2.28. *Let N be a strict level-k, semi-directed trinet with $k > 0$. Then, under the JC model and if N has no trivial parameter values,*

$$q_{011}q_{101}q_{110} - q_{111}^2 > 0.$$

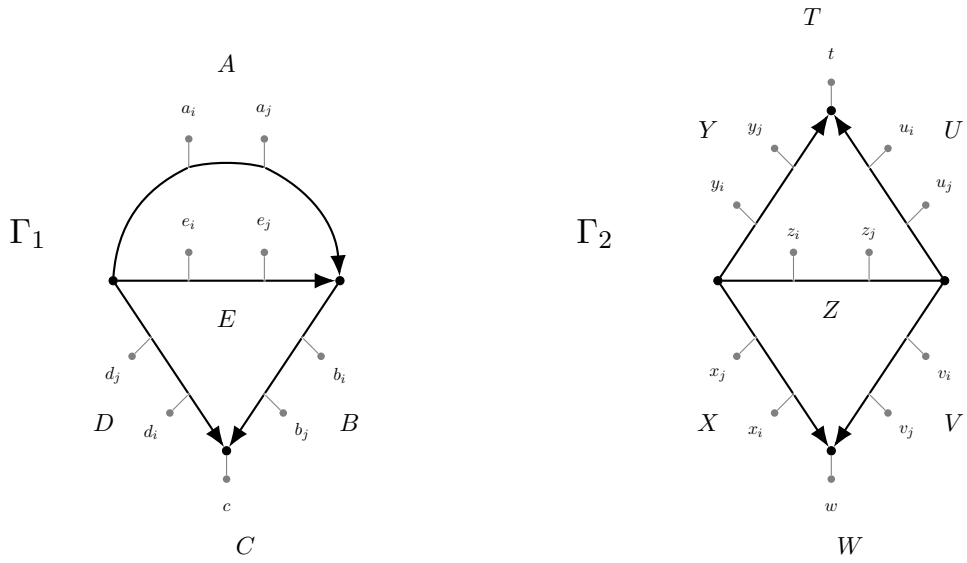

FIGURE 7. The two types of simple, not outer-labeled planar, semi-directed level-2 generators Γ_1 and Γ_2 in thick black (Huber et al., 2024, Lem. 4.2). The sides $\{A, B, C, D, E\}$ and $\{T, U, V, W, X, Y, Z\}$ of the two generators are the places where the leaves (in gray) are attached in the corresponding semi-directed network.

2.5. **Distinguishing Large Level-2 Networks.** For the remainder of this section we focus on simple semi-directed level-2 networks with any number of leaves. In particular, to prove the main result of this paper, we first prove two separate lemmas that show distinguishability of semi-directed level-2 networks under the JC constraints: one that treats those networks that are not outer-labeled planar, and another that treats the outer-labeled planar networks. Our strategy in both lemmas will be to argue that for every pair of networks N_1 and N_2 in the considered class, there will always be some set of leaves S such that the restrictions $N_1|_S$ and $N_2|_S$ are distinguishable, using one of our previous results. We will then implicitly use Proposition 2.15 to show that the two full networks N_1 and N_2 are also distinguishable.

Lemma 2.29. *Let N_1 and N_2 be distinct binary, stack-free, simple, semi-directed level-2 phylogenetic networks on the same leaf set. If N_1 and N_2 are both not outer-labeled planar, then N_1 and N_2 are distinguishable under the JC constraints.*

Proof. Let $\mathcal{X}$ be the leaf set of the two networks. First note that both N_1 and N_2 are strict level-2, otherwise they would be outer-labeled planar. The only strict level-2 quarnets that are stack-free and not outer-labeled planar are of type 3a, 3c or 3d. Hence, if $|\mathcal{X}| = 4$, N_1 and N_2 are distinguishable by Lemma 2.25c and we may thus assume that $|\mathcal{X}| \geq 5$.

Let G be the *generator* of N_1, i.e., the mixed multi-graph obtained from N_1 by deleting all leaves and suppressing degree-2 vertices. Similarly, we let H be the generator of N_2. By Lemma 4.2 in Huber et al. (2024), there are exactly two possibilities for G and H, depicted in Figure 7: the generators of type Γ_1 and Γ_2. The edges and degree-2 reticulations of Γ_1 and Γ_2 are called *sides*. We can identify these sides by sets of leaves of the corresponding network, since the sides are exactly where the leaves of the original network were attached (see again Figure 7). We now consider three cases. We will implicitly use that for every $S \subseteq \mathcal{X}$ with $|S| = 4$ the two quarnets $N_1|_S$ and $N_2|_S$ either both have the same nontrivial split or both do not have a

nontrivial split, otherwise the two quarnets have a distinct tree-of-blobs and the networks are already distinguishable by Corollary 2.23.

Case 1: $G \ncong H$. Without loss of generality we assume that G is a generator of type Γ_1 and H a generator of type Γ_2. In N_1, let a be any leaf on side A, b any leaf on side B, c the leaf on side C, and e any leaf on side E. Note that a and e exist since N_1 is not outer-labeled planar, and that b exists since N_1 is stack-free. Then, the quarnet $N_1|_S$ with $S = \{a, b, c, e\}$ is of type 3a. Now note that $N_2|_S$ can not be of type 3a, since then N_2 would have two reticulations above each other, contradicting that its generator is Γ_2. Thus, $N_2|_S$ is distinct from $N_1|_S$. By a similar argument, $N_2|_S$ can also not be of type 3b or 3e, so it is of type 1, 2, 3c or 3d. Then, if $N_2|_S$ is of type 1 or 2, N_1 and N_2 are distinguishable by Corollary 2.21. Otherwise, if $N_2|_S$ is a quarnet of type 3c or 3d, N_1 and N_2 are distinguishable by Lemma 2.25c.

Case 2: $G \cong H$ *and both are of type* Γ_1. Let $S = \{a, b, c, e\}$ be an arbitrary set of leaves as in Case 1, such that $N_1|_S$ is a quarnet of type 3a. Similar reasoning as in Case 1 then shows that N_1 and N_2 are distinguishable if for some choice of S the quarnet $N_1|_S$ is distinct from $N_2|_S$. Hence, we may assume that $N_1|_S$ and $N_2|_S$ are isomorphic for any valid choice of S. It follows that N_1 and N_2 share the same leaves on the sides A, B, C and E of their generator (modulo swapping sides A and E), and therefore also on side D. Then, since by assumption N_1 and N_2 are distinct and $|\mathcal{X}| \geq 5$, we have that for some side $\Sigma \in \{A, B, D, E\}$ (with $|\Sigma| \geq 2$) two leaves $\sigma_i \in \Sigma$ and $\sigma_j \in \Sigma$ are ordered differently in the two networks. Now, let c be the leaf on side C of N_1 and N_2. First suppose that $\Sigma \in \{A, D, E\}$ and let b be a leaf on side B of N_1 and N_2 (which exists since the networks are stack-free). Then, the quarnets $N_1|_S$ and $N_2|_S$ with $S = \{\sigma_i, \sigma_j, b, c\}$ are both strongly outer-labeled planar, i.e. both are congruent with a single circular ordering of S. But since σ_i and σ_j are ordered differently on side Σ in the two networks, these circular orders cannot be the same. Hence, N_1 and N_2 are distinguishable by Corollary 2.24, which states that two strongly outer-labeled planar quarnets are distinguishable if they are not congruent with the same circular ordering. The case where $\Sigma = B$ follows a similar argument, instead considering the subset of leaves $S = \{\sigma_i, \sigma_j, a, c\}$, where a is a leaf from side A of both networks (which exists since the networks are not outer-labeled planar).

Case 3: $G \cong H$ *and both are of type* Γ_2. In N_1, let t be the leaf on side T, w the leaf on side W, z any leaf on side Z, and σ any other leaf. Note that t, w and z exist since N_1 is not outer-labeled planar. Then, the quarnet $N_1|_S$ with $S = \{t, w, z, \sigma\}$ is of type 3c or 3d. If for some valid choice of S the quarnet $N_2|_S$ is of type 0, 1 or 2, we get that N_1 and N_2 are distinguishable by Corollary 2.21. Now note that if $N_2|_S$ is not of type 0, 1 or 2, it can only be of type 3c or 3d. By Lemma 2.25c, two distinct quarnets of type 3c or 3d are distinguishable, so we may now assume that $N_1|_S$ and $N_2|_S$ are isomorphic for all valid choices of S. It follows that N_1 and N_2 share the same leaves on the sides T, $Y \cup U$, Z, $X \cup V$ and W of their generators (modulo swapping $\{Y, T, U\}$ with $\{X, W, V\}$).

If $|Y \cup U \cup X \cup V| \leq 1$ for N_1 and N_2, then N_1 and N_2 share the same leaves on every side (modulo swapping Y with U and swapping X with V). If $|Y \cup U \cup X \cup V| \geq 2$ for N_1 and N_2, let σ and π be two arbitrary leaves in this set. Furthermore, let t be the leaf on side T of N_1 and N_2 and w the leaf on side W of N_1 and N_2. Then, $N_1|_S$ and $N_2|_S$ with $S = \{\sigma, \pi, t, w\}$ are both strongly outer-labeled planar quarnets, i.e. both congruent with a single circular order of S. If for some valid choice of S the induced orderings are not the same, then N_1 and N_2 are distinguishable by Corollary 2.24. Thus, we may assume that these circular orders are the same for any such set S, again implying that N_1 and N_2 share the same leaves on every side (modulo swapping Y with U and swapping X with V). Furthermore, the leaves on these four sides must be ordered the same.

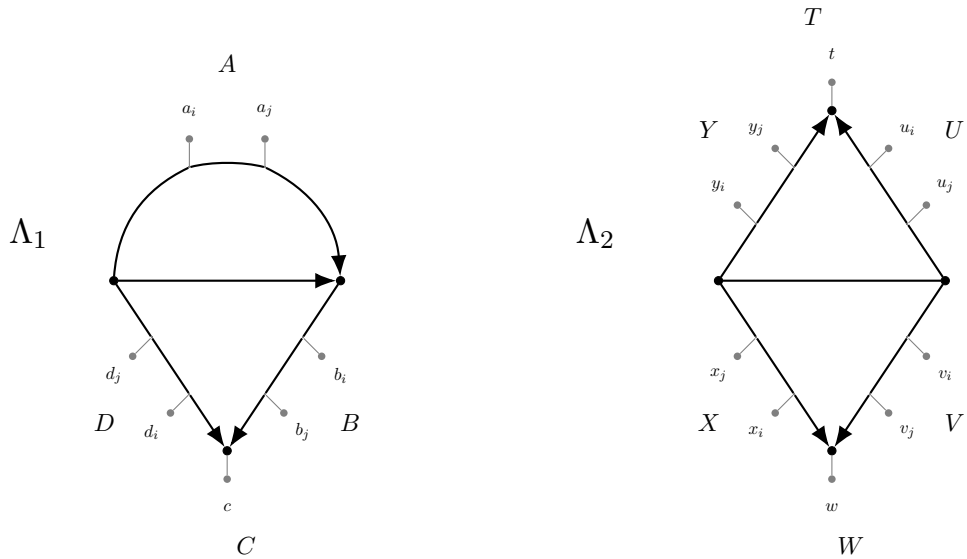

FIGURE 8. The two types of simple, outer-labeled planar, semi-directed level-2 generators Λ_1 and Λ_2 in thick black (Huber et al., 2024, Lem. 4.2). The sides $\{A, B, C, D\}$ and $\{T, U, V, W, X, Y\}$ of the two generators are the places where the leaves (in gray) are attached in the corresponding semi-directed network.

The only remaining way for N_1 and N_2 to be distinct is then that there are two leaves z_i and z_j on side Z of N_1 and N_2 that are not ordered the same in the two networks. Let σ be a leaf unequal to z_i or z_j from some side $\Sigma \in \{U, V, X, Y, Z\}$ (which exists since $|\mathcal{X}| \geq 5$). If $\Sigma \in \{U, Y, Z\}$ then we consider the set $S = \{\sigma, t, z_i, z_j\}$ with t the leaf on side T of both networks. Otherwise, if $\Sigma \in \{V, X\}$, we instead pick the other leaf below the reticulation w (on side W of both networks) and consider the set $S = \{\sigma, w, z_i, z_j\}$. In both cases, the quarnets $N_1|_S$ and $N_2|_S$ are of type 0 (i.e. they are level-1 quarnets with a 4-cycle) and they induce a unique circular ordering of S. But since z_i and z_j are ordered differently on side Z in the two networks, these induced orders cannot be the same. Hence, N_1 and N_2 are distinguishable by Corollary 2.24. □

Lemma 2.30. *Let N_1 and N_2 be distinct binary, triangle-free, strongly tree-child, simple, semi-directed level-2 phylogenetic networks on the same leaf set. If N_1 and N_2 are both outer-labeled planar, then N_1 and N_2 are distinguishable under the JC constraints.*

Proof. Let $\mathcal{X}$ be the leaf set of the two networks. If N_1 and N_2 are both level-1 networks, they are distinguishable by Gross and Long (2018). If exactly one of them (say N_1) is a strict level-2 network, then there will be some $S \subseteq \mathcal{X}$ with $|S| = 4$ such that $N_1|_S$ is a type 1b or 2a quarnet (since N_1 is triangle-free, strongly tree-child and outer-labeled planar). Because N_2 is level-1, $N_2|_S$ is either a type 0 quarnet or a quarnet inducing a nontrivial split. In the first case, $N_1|_S$ and $N_2|_S$ are distinguishable by Corollary 2.27, whereas in the second case the two quarnets have a distinct tree-of-blobs and are distinguishable by Corollary 2.23. Thus, we may assume that N_1 and N_2 are both strict level-2. Now note that the only outer-labeled planar strict level-2 quarnets that are strongly tree-child and triangle-free are of type 2a. Hence, if $|\mathcal{X}| = 4$, the two networks are distinguishable by Lemma 2.25b, which states that two distinct type 2a quarnets are distinguishable. Thus, we will also assume that $|\mathcal{X}| \geq 5$.

As in the proof of Lemma 2.29, we let G be the generator of N_1 and H the generator of N_2. In the case of outer-labeled planar networks there are again two types of generators (Huber et al., 2024, Lem. 4.2): Λ_1 and Λ_2, depicted in Figure 8. We again consider three cases. Note that no quarnets of N_1 or N_2 can be of type 3, since both networks are outer-labeled planar. As in the proof of the previous lemma, we will implicitly use that for every $S \subseteq \mathcal{X}$ the two quarnets $N_1|_S$ and $N_2|_S$ with $|S| = 4$, either both have the same nontrivial split or both do not have a nontrivial split, otherwise the two quarnets have a distinct tree-of-blobs and the networks are already distinguishable by Corollary 2.23. Furthermore, by Corollary 2.24, we can assume that N_1 and N_2 share the same circular ordering of their leaf set $\mathcal{X}$, otherwise the two networks are again distinguishable.

Case 1: $G \ncong H$. Without loss of generality we assume that G is a generator of type Λ_1 and H a generator of type Λ_2. In N_1, let $\{a_i, a_j\}$ be any pair of leaves on side A, c the leaf on side C and d any leaf on side D. Note that a_i and a_j exist since N_1 is triangle-free and that d exists since N_1 is strongly tree-child. Then, the quarnet $N_1|_S$ with $S = \{a_i, a_j, c, d\}$ is of type 2d with the leaf c below the reticulation, i.e. it has a signature of the form $(c, \bullet)$. If $N_2|_S$ is of type 1a, 1b, 2a, 2b or 2c, the two quarnets are distinguishable by Corollary 2.27a, whereas the two quarnets are distinguishable by Corollary 2.27b if $N_2|_S$ is of type 0, 1c, 1d, 1e, 1f or 2d with a signature of the form $(\sigma, \bullet)$ where $\sigma \neq c$. It follows that we may assume that $N_2|_S$ is a quarnet of type 0, 1c, 1d, 1e, 1f or 2d with a signature of the form $(c, \bullet)$. This means that c is below a reticulation in $N_2|_S$, and therefore, c has to be on side T or W of N_2.

Let $\{a_i, a_j, d\}$ be as before and pick b as an arbitrary leaf on side B of N_1 (which exists since N_1 is stack-free). Then, the quarnet $N_1|_S$ with $S = \{a_i, a_j, b, d\}$ is of type 0 with the leaf b below the reticulation, i.e. it has signature (b, b). A similar argument as before then shows distinguishability, unless b is on side T or W of N_2. We now show that this is impossible by deriving a contradiction, which will then prove Case 1. If b is on side T or W of N_2, we can assume that the leaves on sides $B \cup C$ of N_1 coincide with the leaves on sides $T \cup W$ of N_2 (which also implies that $|B| = 1$ for N_1). Now note that the leaves b and c are neighbors in the circular ordering of $\mathcal{X}$ induced by N_1. By our assumption, this is then also the case for N_2 (where b and c are on sides T and W). Therefore, the sides X and Y (or U and V) of N_2 are both empty. But then N_2 is not strongly tree-child: a contradiction.

Case 2: $G \cong H$ *and both are of type* Λ_1. In N_1, let $\{a_i, a_j\}$ be any pair of leaves on side A, d any leaf on side D, and σ any leaf on side $\Sigma \in \{B, C\}$. Note that a_i and a_j exist since N_1 is triangle-free and that d exists since N_1 is strongly tree-child. The quarnet $N_1|_S$ with $S = \{a_i, a_j, d, \sigma\}$ is then of type 0 or 2d, in both cases with the leaf σ below the reticulation (i.e. with a signature of the form $(\sigma, \bullet)$). Similar reasoning (based on Corollary 2.27) as in the previous case then shows that N_1 and N_2 are distinguishable, unless $N_2|_S$ is a quarnet of type 0, 1c, 1d, 1e or 1f with signature $(\sigma, \bullet)$. Thus, from now on we may assume that for any valid choice of σ, the quarnet $N_2|_S$ has σ below a reticulation, meaning that σ is also on sides $B \cup C$ of N_2. Hence, N_1 and N_2 share the same leaves on sides $B \cup C$ (and therefore also on sides $A \cup D$).

Now, in N_1, let a be a leaf on side A, b a leaf on side B, c the leaf on side C and d a leaf on side D. Note that these leaves all exist since N_1 is stack-free and strongly tree-child. Furthermore, pick these four leaves such that $\{a, b\}$ and $\{c, d\}$ are both neighboring pairs of leaves in the circular ordering of X induced by N_1 (and hence also by N_2). The quarnet $N_1|_S$ with $S = \{a, b, c, d\}$ is now of type 1b (with signature of the form $(c, \bullet)$). Recall that N_1 and N_2 share the same circular ordering of their leaf set X and that they share the same leaves on sides $B \cup C$ and $A \cup D$. By our specific choice of the leaves in S, we then have that $N_2|_S$ is either isomorphic to $N_1|_S$, or it is a type 1b quarnet with the same circular ordering as $N_1|_S$ but with signature $(b, \bullet)$ instead. In the latter case, N_1 and N_2 are distinguishable by Lemma 2.25a.

Hence, we can assume that $N_1|_S$ and $N_2|_S$ are isomorphic, also implying that N_1 and N_2 share the same leaf on side C (and therefore also have the same leaves on side B).

Lastly, let c be the leaf on side C of N_1 and N_2. Now consider the trees-of-blobs of the networks $N_1|_S$ and $N_2|_S$ with $S = \mathcal{X} \setminus \{c\}$. First, assume towards a contradiction that these two trees-of-blobs are isomorphic. Then, since the two networks already share the same leaves on their side B and both networks have at least one leaf on their side D, they must share the same number of leaves on their side D (and therefore also on their side A) for the trees-of-blobs to be isomorphic. However, we already know that N_1 and N_2 share the same circular ordering of their leaf set $\mathcal{X}$, showing that N_1 and N_2 share the same leaves on every side and that they are also ordered the same. Therefore, N_1 and N_2 are isomorphic: a contradiction. Thus, we may instead assume that the two trees-of-blobs are not isomorphic. But then $N_1|_S$ and $N_2|_S$ (and therefore also the two networks N_1 and N_2) are distinguishable by Corollary 2.23.

Case 3: *$G \cong H$ and both are of type Λ_2.* Let t be the leaf on side T of N_1 and let w be the leaf on side W of N_1. Note that t and w can not be neighbors in the circular ordering of N_1 and N_2, otherwise N_1 would not be strongly tree-child. Now consider the set $S = \mathcal{X} \setminus \{t, w\}$. Then, $N_1|_S$ is a tree and all its induced trinets are triplet trees. If this is not the case for $N_2|_S$, the two networks are distinguishable by Proposition 2.26, which can be used to distinguish triplet trees from trinets with a reticulation. Thus, we can assume that $N_2|_S$ also induces only triplet trees, i.e. $N_2|_S$ is a tree. We now show that t and w are then also the leaves on sides T and W of N_2. In particular, if neither t nor w is on sides $T \cup W$ of N_2, we clearly have that $N_2|_S$ is not a tree: a contradiction. On the other hand, assume w.l.o.g. that t is on side T of N_2 and w is not on side W of N_2. Then, since t and w are not neighbors in the circular ordering of $\mathcal{X}$ induced by N_2 and since N_2 is strongly tree-child, there will be some $\sigma_i \notin \{t, w\}$ on sides $X \cup Y$ and some $\sigma_j \notin \{t, w\}$ on sides $U \cup V$ of N_2. Let $w' \notin \{t, w, \sigma_i, \sigma_j\}$ then be the leaf on side W of N_2. The trinet $N_2|_{\{\sigma_i, \sigma_j, w'\}}$ with $\{\sigma_i, \sigma_j, w'\} \subseteq A$ will then not be a triplet tree: a contradiction. Thus, t and w are on sides $T \cup W$ of N_2. Therefore, N_1 and N_2 share the same leaves on sides T and W (modulo swapping the two sides). Since the two networks also induce the same circular ordering of $\mathcal{X}$, they also share the same leaves on sides $X \cup Y$ and $U \cup V$ (modulo swapping these two sets of sides). We now consider two subcases.

Case 3.1: $|\Sigma| \leq 1$ for all $\Sigma \in \{U, V, X, Y\}$ for N_1 and N_2. In this case, we have that $|\mathcal{X}| \leq 6$, while we already showed that we may assume that $|\mathcal{X}| \geq 5$. When $|\mathcal{X}| = 6$, N_1 and N_2 can not be distinct while sharing the same leaves below the reticulations, sharing the same leaves on sides $U \cup V$ and $X \cup Y$ (modulo swapping these two sets of sides), and sharing the same circular ordering. Hence, we get that $|\mathcal{X}| = 5$. Then, the networks N_1 and N_2 are the two distinct quinnets from Figure 5 (up to relabeling the leaves). Thus, N_1 and N_2 are distinguishable by Lemma 2.25d.

Case 3.2: $|\Sigma| \geq 2$ for some $\Sigma \in \{U, V, X, Y\}$ for N_1 or N_2. Without loss of generality, assume that $|U| \geq 2$ for N_1. Now let t be the leaf on side T of N_1 and set $S = \mathcal{X} \setminus \{t\}$. Then, the tree-of-blobs of the network $N_1|_S$ will induce a nontrivial split that cuts off the leaves on side U of N_1 from the rest. If the tree-of-blobs of $N_2|_S$ is not isomorphic, the two networks are distinguishable by Corollary 2.23. Thus, we can assume they are isomorphic and the tree-of-blobs of $N_2|_S$ also induces the same split. But since N_1 and N_2 already share the same leaves on sides $U \cup V$, this means that they share the same leaves on side U and on side V. If $|X| \geq 2$ or $|Y| \geq 2$ for N_1 or N_2, we can repeat this argument to obtain distinguishability between the two networks. Hence, we may assume that $|X| \leq 1$ and $|Y| \leq 1$ for both networks. Recall that the two networks already share the same leaves on sides $X \cup Y$ and that they are ordered the same. If both sets have size 1, it follows that the two networks are equal: a contradiction. Thus, for the two networks to be distinct, we assume without loss of generality that $|X| = 1$ and $|Y| = 0$

for N_1, while $|X| = 0$ and $|Y| = 1$ for N_2. Now, let σ be the leaf that is on the sides $X \cup Y$ of the two networks. Furthermore, let t be the leaf on side T of the networks, w the leaf on side W of the networks, u a leaf on side U of the networks, and v a leaf on side V of the networks (which both exist, since N_1 is triangle-free and its side Y is empty, while N_2 is triangle-free and its side X is empty). Then, $N_1|_S$ and $N_2|_S$ with $S = \{\sigma, t, u, v, w\}$ are the two distinct quinnets from Figure 5 (up to relabeling the leaves). Thus, N_1 and N_2 are distinguishable by Lemma 2.25d. □

3. Results

3.1. **Identifiability of Level-1 Networks.** Building on the results of the previous subsections we are now ready to prove the main identifiability result for level-2 networks. Before we do this, we first strengthen the best existing identifiability results for semi-directed level-1 networks under a network-based Markov model from Gross et al. (2021). They proved that the network parameter of a network-based Markov model with respect to the class of n-leaf triangle-free, level-1 semi-directed networks with a fixed number of $r \geq 0$ reticulation vertices is identifiable under the JC, K2P and K3P constraints. Using our previous results, we are able to considerably strengthen these results in the case of JC and K2P constraints. In particular, the following theorem shows that we can not only identify such level-1 networks if they have the same number of reticulation vertices but also in case the number of reticulation vertices differs. Furthermore, for the JC constraints we can relax the restriction that the networks need to be triangle-free.

Theorem 3.1.

(a) *The network parameter of a network-based Markov model under the JC constraints is generically identifiable up to the orientation of triangles, with respect to the class of models where the network parameter is an n-leaf binary level-1 semi-directed phylogenetic network.*

(b) *The network parameter of a network-based Markov model under the K2P constraints is generically identifiable with respect to the class of models where the network parameter is an n-leaf binary, triangle-free, level-1 semi-directed phylogenetic network.*

Proof. The tree-of-blobs of a network is generically identifiable under both the JC and K2P constraints by Corollary 2.23. By Proposition 2.26, triangles can also be distinguished from trivial blobs under the JC constraints. The results then follow from the fact that simple level-1 networks (or, *sunlets*) with at least four leaves are identifiable under the JC (Gross and Long, 2018) and K2P (Hollering and Sullivant, 2021) constraints. □

3.2. **Identifiability of Level-2 Networks.** Recall that Lemmas 2.29 and 2.30 showed distinguishability under the JC constraints of classes of outer-labeled planar and not outer-labeled planar simple level-2 networks, respectively. These lemmas now allow us to prove our main result in the next theorem. Note that Theorem 3.1a is not a special case of this theorem, since here we also assume that the network is triangle-free.

Theorem 3.2. *The network parameter of a network-based Markov model under the Jukes-Cantor constraints is generically identifiable with respect to the class of models where the network parameter is an n-leaf binary, triangle-free, strongly tree-child, level-2 semi-directed phylogenetic network.*

Proof. Let N_1 and N_2 be two distinct n-leaf binary, triangle-free, strongly tree-child, level-2 semi-directed phylogenetic networks on the same leaf set $\mathcal{X}$. If N_1 and N_2 have non-isomorphic trees-of-blobs, they are distinguishable by Corollary 2.23. Note that N_1 and N_2 can not contain

nontrivial 3-blobs since those are not triangle-free or not strongly tree-child (see Figure 6). Hence, it will be enough to assume that N_1 and N_2 are both simple networks with at least 4 leaves. If N_1 and N_2 are both not outer-labeled planar, then the strongly tree-child assumption forces both networks to be stack-free. Then, they are distinguishable by Lemma 2.29. If N_1 and N_2 are both outer-labeled planar, they are distinguishable by Lemma 2.30. Lastly, without loss of generality, assume that N_1 is outer-labeled planar and N_2 is not. Then, N_2 is strict level-2 and there is some $S \subseteq \mathcal{X}$ with $|S| = 4$ such that the quarnet $N_2|_S$ is stack-free and not outer-labeled planar and thus is of type 3a, 3b, 3c or 3d. On the other hand, $N_1|_S$ must be outer-labeled planar and can thus not be of type 3a, 3b, 3c or 3d. But the type 3a, 3b, 3c or 3d quarnets are the only level-2 quarnets that display exactly three quartet trees. Therefore, N_1 and N_2 are distinguishable by Theorem 2.20, which lets us distinguish quarnets on the basis of their displayed trees. □

3.3. **Identifiability Beyond Level-2.** In Section 2.2 we proved that semi-directed networks with different displayed quartets are identifiable (Theorem 2.20). Consequently, we also obtained two interesting identifiability results regarding the tree-of-blobs of semi-directed networks (Corollary 2.23) and the congruent circular orderings of outer-labeled planar semi-directed networks (Corrolary 2.24), which we later used to obtain our proof of identifiability for level-2 networks. We summarize these three results here in the Results section since they are also of independent interest (see e.g. Allman et al. (2023); Rhodes et al. (2025) for similar results under coalescent based models), especially because the results pertain to networks of any level. Note that these results hold under both the JC and K2P models.

Theorem 3.3. *Let N_1 and N_2 be two binary semi-directed phylogenetic networks on the same leaf set $\mathcal{X}$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of probability distributions produced by the Markov models on these networks under either the JC or K2P constraints.*

(a) If there is a 4 element subset $S \subseteq \mathcal{X}$ such that the quarnets $N_1|_S$ and $N_2|_S$ have different sets of displayed quartets, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.
(b) If the trees-of-blobs of N_1 and N_2 are not isomorphic, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.
(c) If N_1 and N_2 are both outer-labeled planar and they have different sets of congruent circular orders, then $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$.

4. Discussion

We have shown that the class of strongly tree-child, triangle-free, level-2, binary semi-directed networks is identifiable under the Jukes-Cantor model of evolution. Our results indicate that, under this model, nucleotide sequence data provides enough information to infer semi-directed networks from this class. This is the first complete identifiability result of a large class of networks beyond the level-1 framework. As detailed by Kong et al. (2022), a tree-child scenario is more likely to appear in nature than a non-tree-child scenario, since hybridization is considered to be an uncommon evolutionary event, further supporting the biological relevance of our identified class.

The results in this paper do not strictly pertain to level-2 networks. Specifically, we have also shown that the tree-of-blobs of any binary semi-directed network and the circular ordering of the subnetwork around each outer-labeled planar blob are identifiable under both the Jukes-Cantor and the Kimura-2 parameter models, thereby extending existing results in Allman et al. (2023); Rhodes et al. (2025) to other classes of models. Identifiability of the tree-of-blobs also allowed us to considerably strengthen previous identifiability results for level-1 networks under Markov models (Gross et al., 2021).

It remains an open problem to try to prove identifiability of level-2 (and higher level) networks for broader classes of networks than we have shown in this paper. However, there are already

challenges such that for the Markov models, it is not possible to identify a network structure in full generality. For example, stacked reticulations always lead to non-identifiability issues, as shown in Sullivant (2025).

The combinatorial component of our proof, which relies on algebraic distinguishability proofs for small networks, may provide a blueprint to prove similar level-2 identifiability results under other models of evolution. For instance, to lift our level-2 identifiability result to the Kimura 2-parameter model, it would suffice to generalize Lemma 2.25 and Proposition 2.26, which establish distinguishability for specific combinations of small networks. Going beyond the level-2 assumption will likely require further improvements to the existing algebraic tools. Notably, it is already known that semi-directed level-3 networks are not encoded by their quarnets (Huber et al., 2024). As a result, achieving identifiability of (a subclass of) level-3 networks may necessitate analyzing numerous combinations of induced level-3 quinnets (5-leaf networks), leading to a combinatorial explosion of combinations compared to level-2 quarnets.

On the algebraic side, there are many open questions related to level-2 networks, and even quarnets. For example, while we computed some phylogenetic invariants for these networks in some cases, in most cases it remains open to try to determine the vanishing ideal of the model, or other basic algebraic properties of the models. We found instances of pairs of networks that have the same underlying algebraic matroid but different phylogenetic ideals. There were examples of networks that had a lower dimension than would be expected based on a simple parameter count. Developing tools to explain these phenomena leads to interesting algebraic challenges, and may provide new tools for proving deeper identifiability results.

Finally, on the algorithmic side, our findings open several promising research directions. Previous identifiability proofs for level-1 networks (Gross et al., 2021) have resulted in theoretical algorithms that puzzle together quarnets to construct a complete level-1 network (Huebler et al., 2019; Frohn et al., 2025), ultimately contributing to a quarnet-based practical tool for inference of level-1 networks (Holtgrefe et al., 2025). Building on our results, similar theoretical algorithms could serve as a foundation for consistent level-2 network inference methods. However, such algorithms, which rely on assembling multiple quarnets, also require consistent methods for inferring these quarnets from biological data. Two existing approaches, both based on invariants, have been developed for level-1 quarnets (Barton et al., 2022; Martin et al., 2023). To establish a complete pipeline from sequence data to level-2 network reconstruction, analogous techniques may be necessary for level-2 quarnets.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1929284 while the authors were in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the semester program on "Theory, Methods, and Applications of Quantitative Phylogenomics". AE was funded by the Wisconsin Alumni Research Foundation and the Georgia Benkart Legacy Fund. EG was supported by the National Science Foundation under Grant No. DMS-1945584. MF, NH, LvI & MJ were supported by the Dutch Research Council (NWO) under Grant No. OCENW.M.21.306, and LvI & MJ also under Grant No. OCENW.KLEIN.125.

Data Availability Statement

All the *Macaulay2* scripts necessary to reproduce the proof of Lemma 2.25 can be accessed at: https://doi.org/10.5281/zenodo.15228012.

References

E. S. Allman, S. Petrović, J. A. Rhodes, and S. Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM transactions on computational biology and bioinformatics*, 8:710 –722, 2010.

E. S. Allman, J. A. Rhodes, and S. Sullivant. When do phylogenetic mixture models mimic other phylogenetic models? *Systematic Biology*, 61:1049–1059, 2012.

E. S. Allman, H. Baños, and J. A. Rhodes. Nanuq: a method for inferring species networks from gene trees under the coalescent model. *Algorithms for Molecular Biology*, 14:1–25, 2019.

E. S. Allman, H. Baños, and J. A. Rhodes. Identifiability of species network topologies from genomic sequences using the logdet distance. *Journal of Mathematical Biology*, 84(5):35, 2022.

E. S. Allman, H. Baños, J. D. Mitchell, and J. A. Rhodes. The tree of blobs of a species network: identifiability under the coalescent. *Journal of Mathematical Biology*, 86(1):10, 2023.

E. S. Allman, H. Baños, J. D. Mitchell, and J. A. Rhodes. Tinnik: inference of the tree of blobs of a species network under the coalescent model. *Algorithms for Molecular Biology*, 19(1):23, 2024a.

E. S. Allman, H. Baños, J. A. Rhodes, and K. Wicke. Nanuq$^+$: A divide-and-conquer approach to network estimation. *bioRxiv*, pages 2024–10, 2024b.

E. S. Allman, H. Baños, M. Garrote-López, and J. A. Rhodes. Identifiability of level-1 species networks from gene tree quartets. *Bulletin of Mathematical Biology*, 86(9), 2024c.

M. Ardiyansyah. Distinguishing level-2 phylogenetic networks using phylogenetic invariants. arXiv:2104.12479, 2021.

M. L. Arnold. *Natural hybridization and evolution*. Oxford University Press, 1997.

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

E. Bapteste, L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie, et al. Networks: expanding evolutionary thinking. *Trends in Genetics*, 29(8):439–441, 2013.

N. H. Barton. The role of hybridization in evolution. *Molecular ecology*, 10(3):551–568, 2001.

T. Barton, E. Gross, C. Long, and J. Rusinko. Statistical learning with phylogenetic network invariants. arXiv:2211.11919, 2022.

J. Cummings and B. Hollering. Computing implicitizations of multi-graded polynomial maps. arXiv:2311.07678, 2023.

A. Diop, E. L. Torrance, C. M. Stott, and L.-M. Bobay. Gene flow and introgression are pervasive forces shaping the evolution of bacterial species. *Genome Biology*, 23(1):239, 2022. doi: 10.1186/s13059-022-02809-5.

A. Francis and V. Moulton. Identifiability of tree-child phylogenetic networks under a probabilistic recombination-mutation model of evolution. *Journal of Theoretical Biology*, 446:160–167, 2018.

M. Frohn, N. Holtgrefe, L. van Iersel, M. Jones, and S. Kelk. Reconstructing semi-directed level-1 networks using few quarnets. *Journal of Computer and System Sciences*, 152:103655, 2025. doi: 10.1016/j.jcss.2025.103655.

S. Ge, T. Sang, B.-R. Lu, and D.-Y. Hong. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proceedings of the National Academy of Sciences*, 96(25):14400–14405, 1999.

D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at http://www2.macaulay2.com.

R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Željko Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo. A draft sequence of the neandertal genome. *Science*, 328(5979):710–722, 2010. doi: 10.1126/science.1188021.

G. Grindstaff, E. Gross, M. Hill, R. Huang, N. Riasat, and N. Vaduthala. Computing ideals of level-k jukes-cantor networks. *forthcoming, code available at* `https://github.com/Rick3yHuang/Phylogenetics-Identifiability`, 2025.

E. Gross and C. Long. Distinguishing phylogenetic networks. *SIAM Journal on Applied Algebra and Geometry*, 2(1):72–93, 2018.

E. Gross, L. van Iersel, R. Janssen, M. Jones, C. Long, and Y. Murakami. Distinguishing level-1 phylogenetic networks on the basis of data generated by markov processes. *Journal of Mathematical Biology*, 83:1–24, 2021.

D. Gusfield, V. Bansal, V. Bafna, and Y. S. Song. A decomposition theory for phylogenetic networks and incompatible characters. *Journal of Computational Biology*, 14(10):1247–1272, 2007.

B. Hollering and S. Sullivant. Identifiability in phylogenetics using algebraic matroids. *Journal of Symbolic Computation*, 104:142–158, 2021.

N. Holtgrefe, K. T. Huber, L. van Iersel, M. Jones, S. Martin, and V. Moulton. Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments. *Molecular Biology and Evolution*, 42(4):msaf067, 2025. doi: 10.1093/molbev/msaf067.

K. T. Huber, L. van Iersel, M. Jones, V. Moulton, and L. Veenema-Nipius. When are quarnets sufficient to reconstruct semi-directed phylogenetic networks? arXiv:2408.12997, 2024.

S. Huebler, R. Morris, J. Rusinko, and Y. Tao. Constructing semi-directed level-1 phylogenetic networks from quarnets. arXiv:1910.00048, 2019.

Y. Jiao, M. An, N. Zhang, H. Zhang, C. Zheng, L. Chen, H. Li, Y. Zhang, Y. Gan, J. Zhao, H. Shang, and X. Han. Multiple third-generation recombinants formed by crf55_01b and crf07_bc in newly diagnosed hiv-1 infected patients in shenzhen city, china. *Virology Journal*, 21(1):306, 2024. doi: 10.1186/s12985-024-02563-z.

P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008.

S. Kong, J. C. Pons, L. Kubatko, and K. Wicke. Classes of explicit phylogenetic networks and their biological and mathematical significance. *Journal of Mathematical Biology*, 84(6):47, 2022.

J. Mallet. Hybridization as an invasion of the genome. *Trends in ecology & evolution*, 20(5): 229–237, 2005.

J. Mallet. Hybrid speciation. *Nature*, 446(7133):279–283, 2007.

S. Martin, V. Moulton, and R. M. Leggett. Algebraic invariants for inferring 4-leaf semi-directed phylogenetic networks. *bioRxiv*, pages 2023–09, 2023.

J. I. Meier, R. B. Stelkens, D. A. Joyce, S. Mwaiko, N. Phiri, U. K. Schliewen, O. M. Selz, C. E. Wagner, C. Katongo, and O. Seehausen. The coincidence of ecological opportunity with hybridization explains rapid adaptive radiation in lake mweru cichlid fishes. *Nature*

communications, 10(1):5391, 2019. doi: 10.1038/s41467-019-13278-z.

L. Nakhleh. Evolutionary phylogenetic networks: models and issues. In *Problem solving handbook in computational biology and bioinformatics*, pages 125–158. Springer, 2010.

N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108, 2006. doi: 10.1038/nature04789.

J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, and J. O. Wertheim. Timing the sars-cov-2 index case in hubei province. *Science*, 372(6540):412–417, 2021. doi: 10.1126/science.abf800.

M. Raghavan, P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T. W. Stafford, Jr, L. Orlando, E. Metspalu, M. Karmin, K. Tambets, S. Rootsi, R. Mägi, P. F. Campos, E. Balanovska, O. Balanovsky, E. Khusnutdinova, S. Litvinov, L. P. Osipova, S. A. Fedorova, M. I. Voevoda, M. DeGiorgio, T. Sicheritz-Ponten, S. Brunak, S. Demeshchenko, T. Kivisild, R. Villems, R. Nielsen, M. Jakobsson, and E. Willerslev. Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, Jan. 2014.

J. A. Rhodes, H. Baños, J. Xu, and C. Ané. Identifying circular orders for blobs in phylogenetic networks. *Advances in Applied Mathematics*, 163:102804, 2025.

L. H. Rieseberg. Hybrid origins of plant species. *Annual review of Ecology and Systematics*, 28 (1):359–389, 1997.

E. B. Sessa, E. A. Zimmer, and T. J. Givnish. Unraveling reticulate evolution in north american dryopteris (dryopteridaceae). *BMC Evolutionary Biology*, 12:1–24, 2012.

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

S. M. Soucy, J. Huang, and J. P. Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015.

B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12(4):457–481, 2005.

S. Sullivant. *Algebraic statistics*, volume 194. American Mathematical Society, 2023.

S. Sullivant. Phylogenetic network models as directed acyclic graphs. arXiv, 2025.

G. J. Szöllősi, A. A. Davín, E. Tannier, V. Daubin, and B. Boussau. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140335, 2015.

S. A. Taylor and E. L. Larson. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature ecology & evolution*, 3(2):170–177, 2019. doi: 10.1038/s41559-018-0777-y.

X.-X. Wang, C.-H. Huang, D. F. Morales-Briones, X.-Y. Wang, Y. Hu, N. Zhang, P.-G. Zhao, X.-M. Wei, K.-H. Wei, X. Hemu, N.-H. Tan, Q.-F. Wang, and L.-Y. Chen. Phylotranscriptomics reveals the phylogeny of asparagales and the evolution of allium flavor biosynthesis. *Nature Communications*, 15(1):9663, 2024.

S. J. Westenberger, C. Barnabé, D. A. Campbell, and N. R. Sturm. Two hybridization events define the population structure of trypanosoma cruzi. *Genetics*, 171(2):527–543, 2005.

M. Worobey, M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J.-J. Muyembe, J.-M. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. P. Gilber, and S. M. Wolinsky. Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. *Nature*, 455(7213):661–664, 2008. doi: 10.1038/nature07390.

J. Xu and C. Ané. Identifiability of local and global features of phylogenetic networks from average distances. *Journal of Mathematical Biology*, 86(1):12, 2023.

Y. Yu, C. Than, J. H. Degnan, and L. Nakhleh. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2): 138–149, 2011.

Appendix A. Remainder of the Proof of Proposition 2.26

In Proposition 2.26 we showed that the invariant $q_{011}q_{101}q_{110} - q_{111}^2$ evaluates to zero under the JC model for the three-leaf tree. Furthermore, we showed that the invariant is positive on the strict level-1 trinet and the strict level-2 trinet of type 1 (see Figure 6), assuming the parameter values are in $(0, 1)$. In this appendix we provide the calculations to show that the invariant $q_{011}q_{101}q_{110} - q_{111}^2$ is also positive on the strict level-2 trinets of types 2-6. Recall that we numbered the categories of strict level-2 trinets from left to right in Figure 6. For each of the five trinets we will provide the Fourier parametrizations under the JC model (using the Fourier parameters as given in Figure 6) and we write out the invariant $q_{011}q_{101}q_{110} - q_{111}^2$ in terms of these parameters to show that it is positive.

Trinet 2. The Fourier parametrizations of Trinet 2 are as follows:

$$\begin{aligned}
q_{111} &= abc(\delta efh(\gamma g + (1-\gamma)i) + (1-\delta)df(\gamma gh + (1-\gamma)i)) \\
q_{110} &= ab(\delta ef(\gamma g + (1-\gamma)hi) + (1-\delta)d) \\
q_{101} &= ac(\delta eh + (1-\delta)df(\gamma gh + (1-\gamma)i)) \\
q_{011} &= bc(\delta f(\gamma gh + (1-\gamma)i) + (1-\delta)f(\gamma gh + (1-\gamma)i)).
\end{aligned}$$

For the invariant we get that

$$\begin{aligned}
q_{011}q_{101}q_{110} - q_{111}^2 &= \delta^2 \cdot a^2b^2c^2e^2f^2i(1-\gamma)(hi(1-h) + gh^2(2-\gamma) + \gamma g) \\
&+ \delta(1-\delta) \cdot a^2b^2c^2e^2f^2i(1-\gamma)(hi(1-h) + gh^2(2-\gamma) + \gamma g),
\end{aligned}$$

which is positive if all parameter values are in $(0, 1)$.

Trinet 3. The Fourier parametrizations for Trinet 3 are

$$\begin{aligned}
q_{111} &= abc(\delta efg(\gamma h + (1-\gamma)i) + (1-\delta)df(\gamma gh + (1-\gamma)i)) \\
q_{110} &= ab(\delta efg + (1-\delta)d) \\
q_{101} &= ac(\delta e(\gamma h + (1-\gamma)gi) + (1-\delta)df(\gamma gh + (1-\gamma)i)) \\
q_{011} &= bcf(\gamma gh + (1-\gamma)i).
\end{aligned}$$

Then, one obtains that

$$\begin{aligned}
q_{011}q_{101}q_{110} - q_{111}^2 &= \delta^2 \cdot a^2b^2c^2e^2f^2ghi(1-g)^2\gamma(1-\gamma) \\
&+ \delta(1-\delta) \cdot a^2b^2c^2def(gh\gamma + i(1-\gamma))(h(1-fg)^2\gamma + gi(1-f)^2(1-\gamma)),
\end{aligned}$$

which is again positive if all parameter values are in $(0, 1)$.

Trinet 4. The Fourier coordinates of Trinet 4 are:

$$\begin{aligned}
q_{111} &= abc(\gamma dh(\delta e + (1-\delta)fg) + (1-\gamma)dfi) \\
q_{110} &= abd \\
q_{101} &= ac(\gamma h(\delta e + (1-\delta)dfg) + (1-\gamma)dfi) \\
q_{011} &= bc(\gamma h(\delta de + (1-\delta)fg) + (1-\gamma)fi).
\end{aligned}$$

The invariant for Trinet 4 is

$$\begin{aligned}
q_{011}q_{101}q_{110} - q_{111}^2 &= \gamma^2 \cdot a^2b^2c^2defgh^2\delta(1-\delta)(1-d)^2 \\
&+ (1-\gamma)\gamma \cdot a^2b^2c^2(d(1-d)^2efhi\delta + (1-\delta)d^2f^2hi(1-g)),
\end{aligned}$$

which is positive under our parameter constraints.

Trinet 5. The Fourier coordinates are now as follows:

$$\begin{aligned} q_{111} &= abc(\delta df(\gamma h + (1-\gamma)ig) + (1-\delta)eg(\gamma hf + (1-\gamma)i)) \\ q_{110} &= ab(\delta df + (1-\delta)eg) \\ q_{101} &= ac(\delta d(\gamma h + (1-\gamma)ifg) + (1-\delta)e(\gamma hfg + (1-\gamma)i)) \\ q_{011} &= bc(\gamma fh + (1-\gamma)gi). \end{aligned}$$

The invariant then becomes

$$\begin{aligned} q_{011}q_{101}q_{110} - q_{111}^2 &= \delta^2 \cdot (abce)^2(1-\gamma)\gamma fghi(g-1)^2 + (1-\delta)^2 \cdot (abcd)^2 fhi(1-\gamma)(f-1)^2 \\ &+ (1-\delta)\delta \cdot (abc)^2 de(\gamma^2 fgh^2(f-1)^2 \\ &+ (1-\gamma)^2 fgi^2(g-1)^2 + (1-\gamma)\gamma hi(f-g)^2). \end{aligned}$$

This can again be seen to be positive if all parameter values are in $(0,1)$.

Trinet 6. The Fourier parametrizations of Trinet 6 are

$$\begin{aligned} q_{111} &= abc(\delta dgi(\gamma f + (1-\gamma)h) \\ q_{110} &= ab(\delta d(\gamma f + (1-\gamma)ghi) + (1-\gamma)eg) \\ q_{101} &= ac(\delta d(\gamma fgi + (1-\gamma)h) + (1-\delta)ei) \\ q_{011} &= bcgi. \end{aligned}$$

After simplification, the invariant becomes

$$\begin{aligned} q_{011}q_{101}q_{110} - q_{111}^2 &= \delta^2 \cdot a^2b^2c^2d^2 fghi\gamma(1-\gamma)(1-gi)^2 \\ &+ (1-\delta)\delta \cdot a^2b^2c^2degi(\gamma fi(1-g)^2 + (1-\gamma)gh(1-i)^2). \end{aligned}$$

This is positive provided all parameters are positive and smaller than 1. □