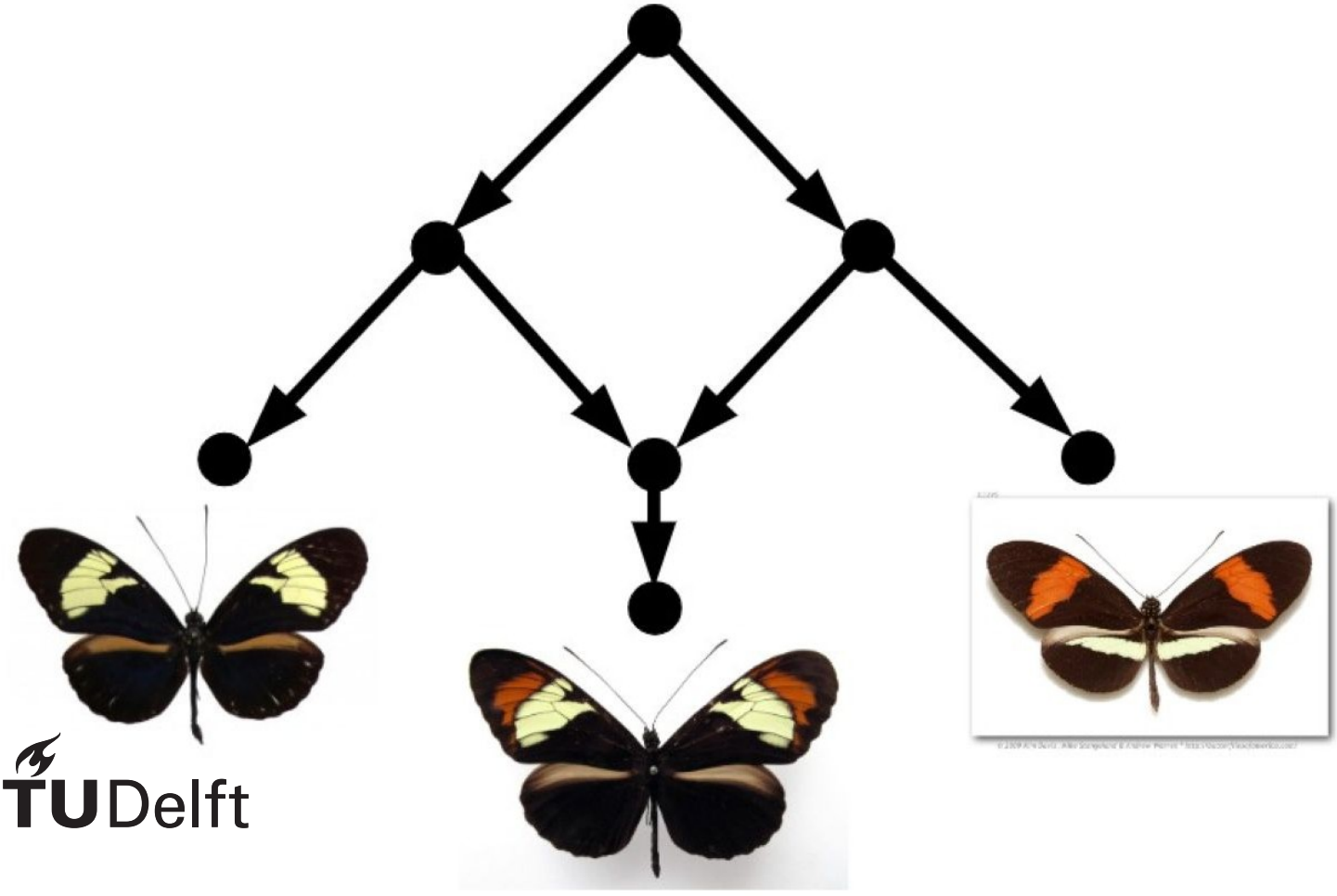
Distinguishing phylogenetic trees from networks

Jari Brits

Bachelor Thesis
Applied mathematics
Delft University of Technology
January 2026



TU Delft

Distinguishing phylogenetic trees from networks

by

Jari Brits

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Tuesday January 20, 2026 at 10:00 AM.

Student number:	5872715	
Project duration:	September 2, 2025 – January 23, 2026	
Thesis committee:	Dr. ir. L. J. J. van Iersel,	TU Delft, supervisor
	ir. N. A. L. Holtgrefe,	TU Delft, supervisor
	Dr. ir. M. Keijzer,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

Evolutionary histories are often described by phylogenetic trees. However, reticulate events such as hybridisation and horizontal gene transfer cannot be described by phylogenetic trees alone. Histories that include these events require phylogenetic networks. For reconstructing evolutionary histories, an important question is whether a network or a tree is needed to describe such a history. Given sufficiently long DNA sequences, it has previously been established that level-1 and level-2 three-leaf networks can be distinguished from three-leaf trees using a phylogenetic invariant.
Here, we generalise this result to networks of level-k, where $k \geq 1$, and show the same distinguishability holds for networks on more than three leaves. This makes it, at least in theory, possible to determine whether the evolutionary history of different taxa happened in a network-like or tree-like way.

Lay summary

Phylogenetic studies investigate the evolutionary relationships among species. Evolution is often thought about as a process in which species repeatedly split into multiple species due to mutation and natural selection. However, not all evolutionary events occur that way. Reticulate events, such as hybridisation and horizontal gene transfer, are examples of different species combining or interchanging genetic material. The common way to visualise evolution is through branching diagrams, but if one wants to include these reticulate events, one should consider phylogenetic networks. In practice, it is often unclear which of these structures to consider when reconstructing the phylogenetic history of certain species, as it is not known whether reticulate events have occurred.

This thesis proves that given sufficiently long DNA sequence data from certain species and assuming a certain model of DNA sequence evolution, it is in theory possible to determine whether reticulate events have taken place between species. This is done on the basis of the frequencies of patterns across the DNA sequences of different species. If no reticulate event has taken place, the evolutionary relationships can be described by a phylogenetic tree. If a reticulate event has occurred, a phylogenetic network is needed. Through different characteristics of these structures and the DNA frequencies, it can be determined whether three species are related by a phylogenetic tree or by the more complex structure, a phylogenetic network. This is achieved using a formula that behaves differently for these mathematical structures, it is zero for a tree and greater than zero for a network. Later, this result will be extended to any number of species.

These results show that under the considered model, it is in theory possible to determine whether reticulate events have taken place in the evolutionary history of species. This means that it is in principle possible to distinguish tree-like evolution from network-like evolution using genetic information. This can play an important role in reconstructing evolutionary relationships.

Contents

1 Introduction

Evolution has become one of the main research areas in biology. Evolutionary biology investigates changes in taxa, which are scientifically classified groups of organisms, over time. Evolution is based on the principle of natural selection and biological variance, which is often related to mutations and sexual recombination between two individuals of the same taxon [11]. However, biological variance can also come from reticulate events, which are events where two taxa share genetic material with each other in a different way than in a parent-offspring relationship. Horizontal gene transfer, hybridisation, symbiosis, and infectious heredity are examples of mechanisms that can produce reticulate events [6]. These reticulate events are often linked to bacteria, fungi, and plants, but horizontal gene transfer seems to occur in many different lineages of eukaryotes [9, 14], including humans [12].

Phylogenetic relationships are often described by phylogenetic trees, these branching diagrams give a visual representation of the taxa and their evolutionary relationships over time. However, phylogenetic trees lack the ability to represent these reticulate events. A reticulate event creates a cycle by two ancestors coming together in one new taxon. One of the characteristics of a tree is that it does not contain any cycles. Phylogenetic networks can be used to describe evolutionary history that includes reticulate events, since they are allowed to contain cycles. Figure 1.1 shows an example of a phylogenetic network with the red dashed lines representing reticulate events, often referred to as reticulations. On the right of Figure 1.1 the associated semi-directed network (explained below) is given, here the dashed arrows represent the reticulated events.

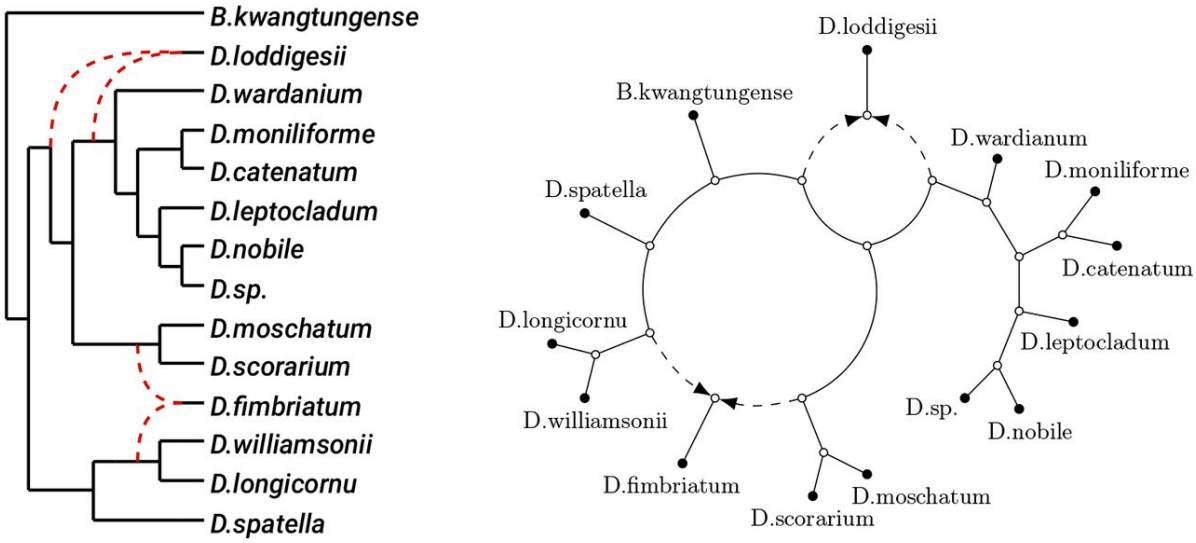


Figure 1.1: Left: A directed phylogenetic network of Orchids, the directions go from left to right. Right: The semi-directed network obtained by suppressing the root and omitting the direction of non-reticulation edges. The images are from Englander et al. [5] with data from Wang et al. [18]

Phylogenetic trees and networks are graph-based structures that describe evolutionary relationships among taxa. A *graph* consists of vertices which can be leaves and internal vertices. Leaves represent the taxa and have observed data, the internal vertices represent the ancestors of the leaves. The other component in graphs are edges, these show how evolution connects the vertices, and they may be directed, typically reflecting the direction of time. If two directed edges are directed into the same vertex, it is a reticulation vertex, which represents a reticulate event that occurred.
Evolutionary histories that include reticulate events can be described by *directed phylogenetic networks*, which have only directed edges and a root vertex representing a common ancestor of all leaves. Here we work with the associated *semi-directed phylogenetic networks*, obtained by suppressing the root and only the edges representing reticulation events keep their direction; the other edges are undirected, see Figure 1.1.
To draw conclusions about how taxa are related, data is needed to compare the taxa. It is assumed that all taxa have a DNA sequence that has evolved from a common ancestor. By aligning these DNA sequences of the different taxa and studying the differences or similarities, the relationship between the taxa and their common ancestor can sometimes be deduced. The two main methods are the distance based method which looks at how much two DNA sequences differ and calculates the evolutionary distance from there, and the character based methods which look at how often the combinations of nucleotides appear over the different DNA strings of the taxa [13]. Here, we will focus on character based methods.
With a sufficient amount of DNA data on different taxa, the goal is often to identify the phylogenetic history between these taxa. Given a collection of networks, sufficiently long DNA sequences, and a model for DNA evolution, a network is *identifiable* if it can be uniquely determined among the networks in the collection. Sometimes it is not possible to uniquely identify a network among the networks, but it is possible to be almost certain from the data, in which case we call a network *generically identifiable* [7].
The identifiability of networks is crucial for assessing whether reconstructive algorithms and software, such as `SQUIRREL` [8] and `TINNiK` [1], yield the correct network. Since larger networks can get significantly more complicated to reconstruct directly, often smaller sub-networks, such as 3-leaf networks (trinets) and 4-leaf networks (quarnets), are analysed first. They can then be used to puzzle out larger networks as in these algorithms.
For distinguishing phylogenetic networks or trees, phylogenetic invariants are sometimes used, phylogenetic invariants are formulas that behave differently for DNA distribution data generated on different trees or on different networks. Phylogenetic invariants have been used to establish identifiability results for quarnets with one reticulation [7] and in some networks with two reticulations [3, 5].
When reconstructing phylogenetic history between taxa it is important to know if network-like evolution or tree-like evolution should be considered. This is the same as determining if reticulate events happened in the phylogenetic history between taxa.
This thesis focuses on identifying whether reticulate events (reticulations) have taken place in semi-directed networks. A semi-directed network is said to be *level-k* if it contains at most k reticulations per *blob*, which is a maximal subnetwork that cannot be disconnected by deleting a single edge. A *sub-blob* is subnetwork that cannot be disconnected by deleting a single edge but does not have to be maximal. If a blob consist of one vertex it is a *trivial* blob. In Figure 1.1, the union of the two cycles is a non-trivial blob. Since this is the only non-trivial blob and it contains two reticulations, the network is level-2. The vertex connected to both D.sp and D.nobile is, for example, a trivial blob.
Blobs with exactly two incident edges are called *2-blobs*, and sub-blobs with two incident edges *2-sub-blobs*. Inserting or suppressing 2-(sub-)blobs does not change the DNA site-pattern frequencies (see Section 2.2.3). Hence, it is not possible to identify such structures under the model considered here. The *tree-of-blobs* is the tree obtained by contracting each blob to a vertex, the algorithm `TINNiK` [1] reconstructs this tree-of-blobs. The algorithm `SQUIRREL` [8] is able to consistently reconstruct triangle-free level-1 networks. Figure 1.2 illustrates examples of a tree and of level-1 and level-2 networks on three leaves.
Englander et al. [5] showed that under the Jukes-Cantor model(a certain model of DNA evolution on a phylogenetic tree or network, see Section 2.2.1), it is possible to distinguish a three-leaf tree from a level-1 or level-2 trinet, that has at least one reticulation and no 2-blobs. They proved this through a case by case analysis of all eight level-2 trinets using a phylogenetic invariant. They also conjectured that this result can be generalized to level-k for all $k > 0$.
This thesis will confirm this conjecture by extending the above mentioned result from [5] to level-k trinets and show that it is even possible to extend to level-k networks on any number of leaves using the same phylogenetic invariant. Hence, under the Jukes-Cantor model, a phylogenetic network with a non-trivial n-blob with $n > 2$ is distinguishable from a phylogenetic tree. This is achieved by splitting the network into subnetworks

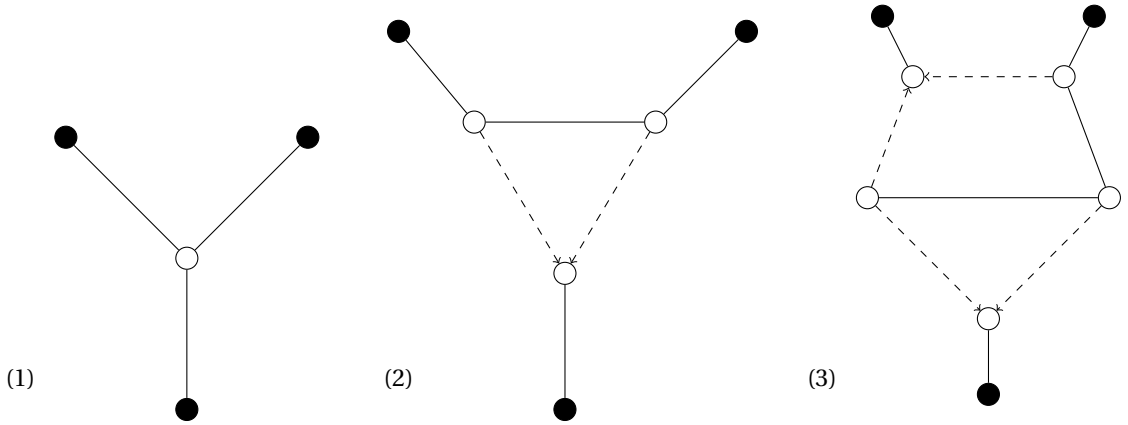


Figure 1.2: *Left*: A tree. *Middle*: A level-1 network. *Right*: A level-2 network.

and decomposing the same phylogenetic invariant as used for the level-1 and level-2 networks. This result and the lemmas required for the proof will be proven in Chapter 3. Throughout this introduction, concepts and definitions are described on an intuitive level, they will be formalised in Chapter 2.

2

Preliminaries and definitions

This chapter introduces some definitions and concepts required for studying phylogenetic networks. This includes definitions from graph theory and from phylogenetic networks and trees, mostly taken from [3–5, 17].

Before going into phylogenetics, it is important to know some basics of graph theory. A *graph* is a mathematical way to show relations between objects. A graph $G = (V, E)$ is made up of the set of *vertices* V that represent the objects and the set E which is a collection of *edges* that are unordered pairs of vertices. These *edges* represent a relationship between two objects. Edges can have a direction, they then point towards one of the vertices, which is usually denoted with an arrow instead of a line. If all edges have a direction, a graph is *directed*, if no edges have a direction, a graph is *undirected*. When a graph contains both edges with and without a direction, it is called *partially directed*. Each undirected edge $e \in E$ is a set of two vertices $\{u, v\}$, *e connects* u and v, and u and v are said to be *incident* to e. The *degree* of a vertex is the number of edges incident to that vertex. The *in-degree* and *out-degree* of vertex are defined the similarly but for only incoming and outgoing edges. We only consider graphs without loops, which is an edge that connects twice to the same vertex. We also only allow at most one edge between two vertices.
A *path* is a sequence with alternating vertices and edges, such that the edges are incident to the vertices they are in between. The sequence begins and ends with a vertex. The path is said to connect these two end vertices. All edges and vertices can be used at most once in a path. If in a path, the first vertex is the same as the last vertex, this path is also known as a *cycle*. A graph is *acyclic* if it does not contain a cycle, and a directed graph is *acyclic* if there is no directed cycle. A path is an *up-down path* if the first k edges are all either undirected or directed towards the first vertex of the path and the last $n-k$ edges are all either undirected or directed towards the last vertex of the path, for some $0 \leq k \leq n$.

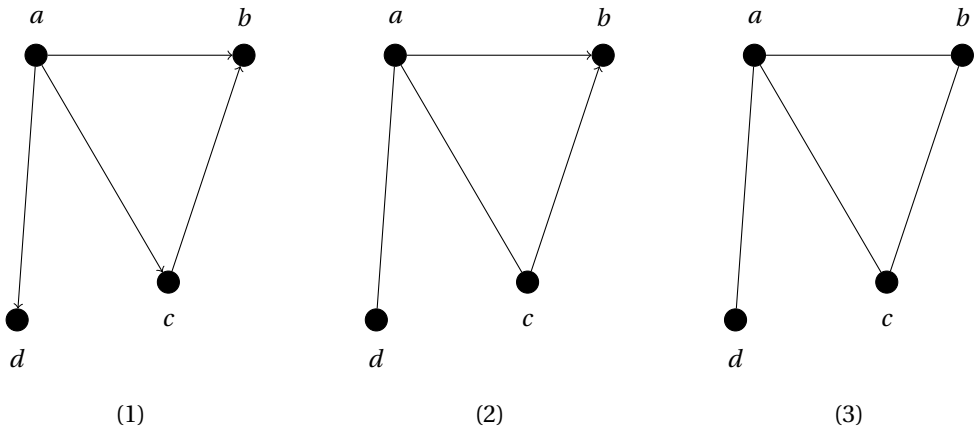


Figure 2.1: From left to right: A directed graph, a partially directed graph and an undirected graph

In Figure 2.1 Network (1), (d, a, c, b) is an up-down path since the first edge is directed to d and the remaining edges to b. In Network (3), (a, c, b, a) is a cycle but in Network (1) and (2) it is not since the since the directions

of the edges do not point in the right way. A graph $G = (V, E)$ is said to be *connected* if for every pair of vertices $u, v \in V$ there is a path between u and v. If there exists a pair of vertices that are not connected, these vertices and the graph are said to be *disconnected*. If the removal of an edge disconnects the graph or increases the number of connected components, this edge is called a *cut edge*. Furthermore, the vertices of this graph can be partitioned into two disjoint subsets A and B such that any path between $a \in A$ and $b \in B$ uses this cut edge e, and this graph thus contains an $A-B$ *split*. In graphs in Figure 2.1, the edges between a and d are a cut-edges.

2.1. Phylogenetic networks and trees

Definition 2.1. A *directed (binary phylogenetic) network* $\mathcal{N}^+$ on a set of taxa $\mathcal{X}$ is a directed acyclic graph that has the following properties.

(i) There is a unique *root*, which has out-degree two and in-degree zero.

(ii) The elements of the set $\mathcal{X}$ are the *leaves* which have in-degree one and out-degree zero.

(iii) The other vertices either have in-degree one and out-degree two, and are so called *tree vertices*, or they have have in-degree two and out-degree one, and are so called *reticulation vertices*.

In an evolutionary perspective, the root can be seen as the *last common ancestor*, or the most recent ancestor, such that over time the taxa $\mathcal{X}$ evolved from this ancestor. All edges therefore point away from the root. These leaves, taxa $\mathcal{X}$, can be seen as taxa of which there is data. The non-leaf vertices are known as *internal* vertices. The tree and reticulation vertices represent taxa that were involved in the evolutionary events. Edges directed towards reticulation vertices are called *reticulation edges*. Two reticulation vertices are *stacked* if the outgoing edge of one reticulation vertex points to the other. Now, a leaf below a reticulation is called a *reticulation leaf*, and this reticulation is called a *leaf reticulation*.
A *directed (binary phylogenetic) tree* is a directed binary phylogenetic network without reticulation vertices. This makes sure there are no cycles in the tree which distinguishes a tree from a network in graph theory.

Definition 2.2. A *semi-directed (binary phylogenetic) network* on a set of taxa $\mathcal{X}$ is a partially directed graph obtained by altering a directed binary phylogenetic network. Namely, removing edge directions of non-reticulation edges and suppressing the root.

Reticulation vertices, reticulation edges, stacks, leaf reticulations, reticulation leaves and internal vertices are defined the same way for semi-directed phylogenetic networks. A semi-directed network without reticulation vertices is an *unrooted phylogenetic tree*, again without a cycle. If in a semi-directed network one incoming reticulation edge per reticulation vertex is chosen, and the network is simplified by omitting the direction of the other incoming edges and then exhaustively deleting non-leaf out-degree zero vertices and suppressing degree-2 nodes, an unrooted phylogenetic tree is *displayed*.
When drawing semi-directed phylogenetic networks leaves will be drawn as filled vertices and internal vertices as unfilled vertices. Reticulation edges will be dashed edges with a direction and network edges are regular lines. Figure 2.2 shows a directed phylogenetic network and the semi-directed network obtained by unrooting the directed phylogenetic network. It also shows two trees obtained by deleting one of the incoming edges of the reticulation. Multiple directed phylogenetic networks can give the same semi-directed network.

A *blob* of a directed or semi-directed network is a maximal connected subgraph without any cut edges. A n-blob is a blob connected to n vertices not in the blob. We call a blob *trivial* if it contains a single vertex, this implies that all leaves are trivial blobs. Three-leaf and four-leaf networks are called trinets and quarnets.
Now, in a semi-directed network $\mathcal{N} = (V, E)$ let G be a subgraph of $\mathcal{N}$ induced by a set $W \subset V$ of at least two vertices. The *boundary* ∂G of G in $\mathcal{N}$ is the set of nodes $w \in W$ adjacent to a node $z \notin W$. The graph G is a *sub-blob of degree n*, or a *n-sub-blob* in $\mathcal{N}$ if it has no cut edge in G, and its boundary in $\mathcal{N}$ has exactly n nodes. A 2-sub-blob can be seen as single-entry single-exit subgraph. This definition slightly deviates from Allman et al. [2]. They require a sub-blob to be hybrid-closed. A sub-graph is hybrid-closed if for every reticulation edge, the paired reticulation edge is also in the subgraph.

Definition 2.3. The *reticulation number* is the number of reticulations in a phylogenetic network. A phylogenetic network is said to be level-k if all blobs have at most k reticulations, a network is even *strict* level-k if it is not level-$(k-1)$.

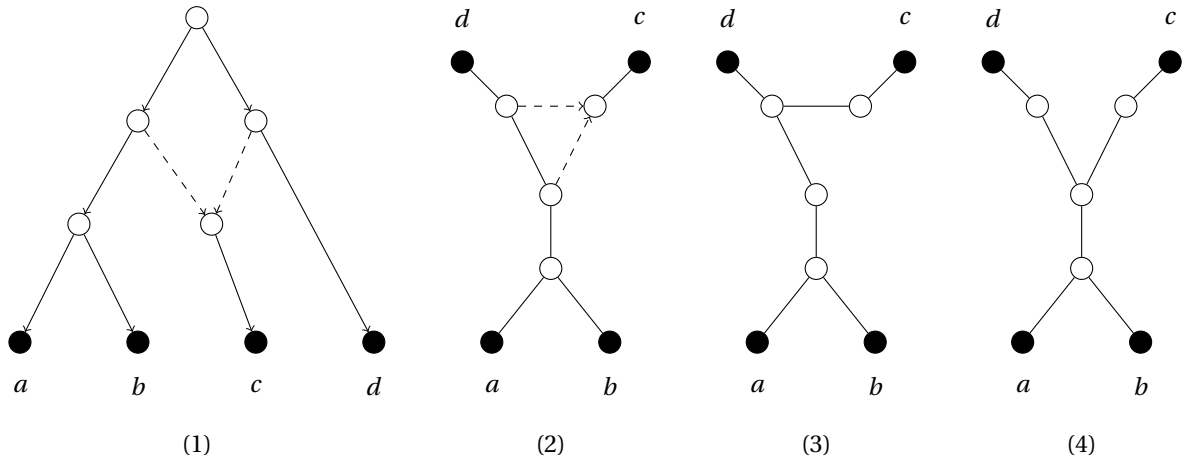


Figure 2.2: Four networks, from left to right: A directed phylogenetic network, the associated semi-directed phylogenetic network, and the two trees obtained by deleting reticulation edges in the semi-directed network.

A phylogenetic network is *simple* if it has only one non-leaf blob. If a phylogenetic network is simple and strict level-k then it contains exactly k reticulations. Given a reticulation vertex r with reticulation edges $e' = (v', r)$ and $e'' = (v'', r)$ in $\mathcal{N}$, we denote by $\mathcal{N} - e'$ and $\mathcal{N} - e''$ the subnetworks of $\mathcal{N}$ obtained by removing e' and e'', respectively, and removing edges not on up-down paths between leaves.

Definition 2.4. Let $\mathcal{N}$ be a phylogenetic network on leaves $\mathcal{X}$, given a subset $X \subset \mathcal{X}$ with $|X| \geq 2$. The *restriction of the network* $\mathcal{N}$ *to the set of leaves* X is denoted as $\mathcal{N}|X$ and defined as the union of the up-down paths in $\mathcal{N}$ between the leaves of X.

Definition 2.5. Let $\mathcal{N}$ be a semi-directed network with vertex set V and let $X \subseteq V$ be a subset of vertices. The subgraph $\mathcal{N}[X]$ is *k-leaf-connected* if the maximum number of edge-disjoint paths between a leaf and a vertex of $\mathcal{N}[X]$ is k in $\mathcal{N}$.

Lemma 2.6. *If a k-leaf-connected subgraph G does not contain a cut-edge, it is a k'-sub-blob with $k' \geq k$. Moreover, if G is a blob, then $k = k'$.*

Proof. Let G be a k-leaf-connected subgraph of a network $\mathcal{N}$ without a cut edge. The subgraph G has k boundary points with a disjoint paths to a leaf. There could also be more boundary points that do not have a disjoint path to a leaf. Therefore, G is a k'-sub-blob with $k' \geq k$.
Suppose that the k-leaf-connected subgraph G, has no cut edge and is maximal. Now also suppose that two boundary points of G have a path with a common edge. Then adding the vertices and the edges of both paths until the last common vertex to G also gives a subgraph with no cut edge. This contradicts the maximality, and thus there exists the same number of disjoint paths between leaves and boundary points as there are boundary points. Therefore, G is a maximal k-leaf-connected k-sub-blob which is a k-blob. □

In Figure 2.3 there is an example where the inequality $k' \geq k$ is strict since subgraph B is a 4-sub-blob and 3-leaf-connected.

2.2. Markov models on networks

Markov models are stochastic models in which it is assumed that the future states depend only on the current state, and not on any states of the past. Markov DNA evolution models therefore assume that the DNA of a new taxon only depends on the DNA of the taxa it evolved from. We will now use Markov evolution models along the edges of networks to describe the change in DNA between the vertices. Since we will only be using the Jukes-Cantor model, which is time reversible, we are allowed to move between rooted networks and semi-directed networks.
Evolution models on networks describe the observed probability distribution of DNA nucleotides at the leaves of a network. If we have a n-leaf network with vertices V and edges E, every vertex $v \in V$ has an associated random variable X_v with the state space $\Sigma = \{A, G, C, T\}$, which represents the four DNA bases. Every edge e has a transition matrix M^e which gives the probability of changing from one DNA base to

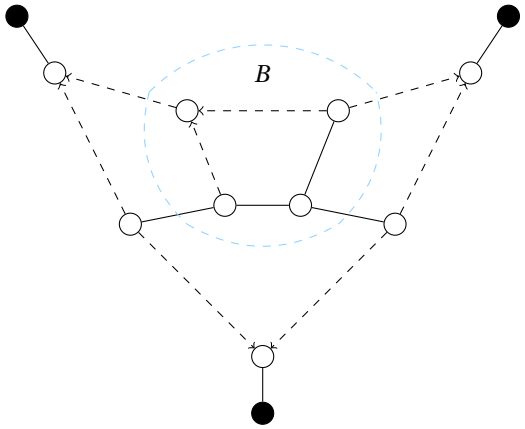


Figure 2.3: A semi-directed network where subgraph B is a 4-sub-blob while being 3-leaf-connected.

another along that edge. The transition matrix M^e, for an edge e from vertex u to vertex v is a stochastic matrix with entries $m^e_{ij} = P(X_v = j | X_u = i)$. This is the probability that the DNA base of v is j given that it evolved from u having the DNA base i. Let Δ^d be the d-th dimensional probability simplex $\Delta^d := \{p \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} p_i = 1, p_i \geq 0 \text{ for } 1 \leq i \leq d+1\}$. All vertices evolved from the root ρ with the root distribution given by $\pi = (\pi_A, \pi_G, \pi_C, \pi_T) \in \Delta^3$.

First, consider a tree $\mathcal{T}$ with vertex set $V(\mathcal{T})$ and edge set $E(\mathcal{T})$. If for a certain position in the DNA sequences, the map $\phi : V(\mathcal{T}) \to \Sigma$ gives the DNA bases for every vertex, the probability of observing that assignment ϕ in tree $\mathcal{T}$ is

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

Usually, only data on the DNA sequences of the leaves are known. Therefore, we want to know the probability of observing a *site pattern* $\omega \in \Sigma^n$, which is the DNA bases of all n leaves of $\mathcal{T}$. By marginalising the probability over all maps that follow the site pattern at the leaves, the probability of observing ω in $\mathcal{T}$ is given by

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

For networks, a reticulation vertex can receive only one DNA base (or a changed one) per position on the DNA sequence. Thus, for every position, only one of the incoming edges is used per reticulation. Every reticulation v_i, with edges e^0_i and e^1_i, has a *reticulation parameter* $\delta_i \in (0,1)$ which gives the probability of the site going through the edge e^1_i and $1-\delta_i$ gives the probability of the site going through the edge e^0_i. The probability of observing a site pattern at the leaves of a network is therefore the probability of observing that site pattern in every tree multiplied by the probability of obtaining that tree with the reticulation parameters. Given a network $\mathcal{N}$ with $k \geq 1$ reticulation vertices $v_1, ..., v_k$. Each reticulation v_i has incoming edges e^0_i and e^1_i with reticulation parameters δ_i and $1-\delta_i$, respectively. There are 2^k possible combinations of reticulation edges and therefore possible 2^k different trees displayed by choosing one edge per reticulation. Let the binary vector $\sigma_j \in \{0,1\}^k$ represent the trees $\mathcal{T}_j$ indicating whether e^0_i or e^1_i was used for every reticulation v_i. The probability of observing a site pattern $\omega \in \Sigma^n$ in the network $\mathcal{N}$ is given by the following formula.

$$P_\omega(\mathcal{N}) = \sum_{\sigma \in \{0,1\}^k} \left[\prod_{i=1}^{k} \delta_i^{1-\sigma_i} (1-\delta_i)^{\sigma_i} \right] P_\omega(\mathcal{T}_\sigma)$$

If we take Θ_N as the stochastic parameter space with transition matrices M^e, reticulation parameters δ_i and the root distribution π, we can view Jukes-Cantor model on $\mathcal{N}$ as the image of a polynomial map $\psi_\mathcal{N}$ defined as follows:

$$\psi_\mathcal{N} : \Theta_\mathcal{N} \to \Delta_{4^n-1}$$

$$(\pi, M^{e_1}, \ldots, M^{e_m}, \delta_1, \ldots, \delta_k) \mapsto (p_\omega(\mathcal{N}))_{\omega \in \Sigma^n}$$

The image of $\psi_\mathcal{N}$ is called the *phylogenetic model* associated to $\mathcal{N}$ and is denoted by $\mathcal{M}_\mathcal{N}$ [3, 5].

2.2.1. Jukes-Cantor model

The evolution Markov model we will use is the Jukes-Cantor model (JC model). This substitution model assumes that the probability of changing from one DNA base to another is the same for all bases. The probability of mutating is $\beta \in (0, 1/4)$ and therefore the probability of not mutating is $1-3\beta$. This gives the following transition matrix M^e.

$$\begin{pmatrix} 1-3\beta & \beta & \beta & \beta \\ \beta & 1-3\beta & \beta & \beta \\ \beta & \beta & 1-3\beta & \beta \\ \beta & \beta & \beta & 1-3\beta \end{pmatrix}$$

The JC model also assumes that the root distribution is uniform. The K2P and K3P models are similar models that allow more parameters to assign different probabilities for different DNA mutations [16].

2.2.2. Fourier transform

The discrete Fourier transform is a linear transformation that simplifies the probability of observing a site pattern from a polynomial parametrisation to a monomial parametrisation [3]. This is possible since the state space can be identified with the elements of the Klein four-group $\mathbb{Z}_2 \times \mathbb{Z}_2$ in the following way $A = (0,0), G = (1,0), C = (0,1), T = (1,1)$. For every edge $e \in E$ the associated Fourier parameters are indicated by a^e_A, a^e_G, a^e_C and a^e_T. Since the probability of mutation is the same for every DNA base, we have $a^e_G = a^e_C = a^e_T$ and $a^e_A = 1$ for biological parameters. All other parameters are in the interval $(0,1)$, because a zero or infinite branch length are excluded. Let $\omega = (g_1, g_2, ..., g_n)$ be a site pattern observed at the leaves of a n-leaf tree $\mathcal{T}$. Let $\Sigma(\mathcal{T})$ be the set of splits $A|B$ induced by the edges of $\mathcal{T}$. Let $a^{A|B}_g = a^e_g$ be the parameter of the edge e that induced the split, with $g \in \mathbb{Z}_2 \times \mathbb{Z}_2$. Then, the Fourier transform of $P_\omega(\mathcal{T})$ is given by

$$q_\omega(\mathcal{T}) = \begin{cases} \prod_{e \in \Sigma(\mathcal{T})} a^e_{\sum_{i \in A} g_i} & \text{if } \sum_{i=1}^{n} g_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

The *Fourier coordinates* can describe the probability distribution in a phylogenetic tree by a monomial with one parameter per edge. In a network, every reticulation doubles the amount of possible displayed trees, therefore in a network with k reticulations can be parametrised by the sum of 2^k monomials weighted by the reticulation parameters of their corresponding displayed trees.
In this thesis, we simplify the notation for the Fourier parameters in the following way. We defined the parameter a^e_g with e the edge and g the DNA base and thus the group the group element. But under the JC model $a^e_A = 1$ and since $a^e_C = a^e_G = a^e_T$ are the same value in $(0,1)$ they can be denoted by one parameter per edge. We will use the letters $a, b, c, ...$ to represent the edges and also use these letters as the names of that edge.

2.2.3. Distinguishability

In order to distinguish different networks, the maps of their network models may not map into each other's image too much, for the formal definition we follow Definition 2.5 in [5] by Englander et al.

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

In order to show distinguishability there, a tool often used is *ideals of phylogenetic invariants*. Phylogenetic invariants are polynomials that vanish on the set of site-pattern probabilities associated with a given phylogenetic network model. To link invariants to distinguishability, the ideal of phylogenetic invariants is defined as the following set of polynomials:

$$\mathcal{I}(\mathcal{M}_N) = \left\{ f \in \mathbb{C}[p_\omega] \;\middle|\; f(p_\omega) = 0 \text{ for all } p_\omega \in \mathcal{M}_N \right\}.$$

Here ψ_N extends to a complex polynomial map ψ'_N by considering the polynomial ring $\mathbb{C}[p_\omega]$ over the indeterminates p_ω with $\omega \in \Sigma^n$ [5, 15]. If a phylogenetic invariant belongs to the ideal of one phylogenetic network model and not in the ideal of another, it can be used to show that the network structures are distinguishable. This was shown by Sullivant [16] by the proposition below.

Proposition 2.8. *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 exists a polynomial $f \in \mathcal{I}(\mathcal{M}_{N_1}) \setminus \mathcal{I}(\mathcal{M}_{N_2})$, then $\mathcal{M}_{N_1}$ and $\mathcal{M}_{N_2}$ are distinguishable.*

This proposition makes it possible to distinguish two networks by showing that a phylogenetic invariant vanishes for one network and does not vanish for the other network.
In order to distinguish two networks on the same set of leaves it is sufficient to show that it is possible to distinguish the restrictions of the networks. This was shown by Englander et al. in Proposition 2.8 in [5].

Proposition 2.9. *Let N_1 and N_2 be two semi-directed phylogenetic networks on the same leaf set X. Let $S \subseteq X$, and suppose that the restrictions $N_1|_S$ and $N_2|_S$ are distinguishable under the JC model. Then N_1 and N_2 are distinguishable under the JC model.*

Multiple networks can give the same site pattern probability distribution, which means they are indistinguishable. Degree 2 vertices do not change the site pattern probability distribution since under the JC model the transition matrices of the edges can be multiplied, and the two edges and the vertex can thus be replaced by one edge. Sullivant [17] showed that 2-blobs do not change the site pattern probability distribution. However, since the maximality of a 2-blob is not needed for the proof this result can be extended to 2-sub-blobs.

Proposition 2.10. *Let G be a DAG with a 2-blob or 2-sub-blob, let B be set of vertices of the 2-blob or 2-sub-blob, and suppose that $a \to b_1$ is the edge pointing into B and $b_2 \to c$ is the edge pointing out of B. Let G' be the graph obtained from G by deleting all the vertices in $\mathcal{B}$ and their incident edges and adding the edge $a \to c$. The two graphs G and G' produce the same probability distributions under the Jukes-Cantor model.*

Proof. Since the maximality is not needed it is possible to apply the proof of Proposition 5.4 in [17] to both 2-blobs and 2-sub-blobs. Sullivant [17] also proves that the JC model satisfies the model requirements for the proof. □

This shows that reticulations in a 2-blob or a 2-sub-blob do not change the site pattern probability distribution. If a reticulation is in a 2-blob or a 2-sub-blob it is not distinguishable from the network with the 2-blob or 2-sub-blob replaced by an edge. Since 2-blobs and 2-sub-blob do not change the site pattern probability distribution they are sometimes *suppressed* which means they are replaced by an edge.

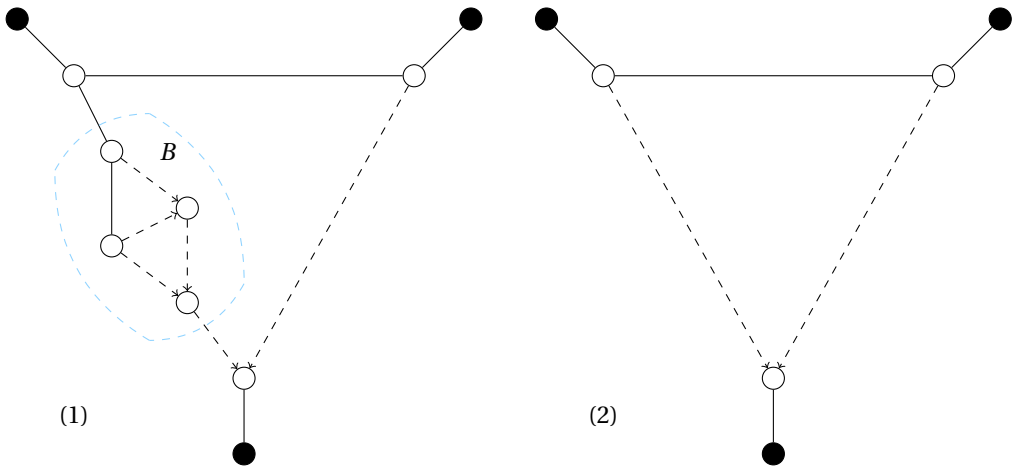


Figure 2.4: Two networks that produce the same site pattern probability distribution, while network (1) contains a 2-sub-blob B.

Figure 2.4 shows two networks that produce the same site pattern probability distribution. If in network 1 the 2-sub-blob B is suppressed network 2 is obtained.

2.3. Lower level networks

For level-1 and level-2 networks it has already been proven that it is possible to distinguish them from a tree. This was done using a phylogenetic invariant expressed in Fourier coordinates. For the Fourier coordinates q of a trinet under the Jukes-Cantor model, $q_{CCA} = q_{GGA} = q_{TTA}$ are the same and denoted by q_{110}. We define

q_{101} and q_{011} in the same way. We write q_{111} for any of the following values $q_{CGT} = q_{CTG} = q_{GCT} = q_{GTC} = q_{TGC} = q_{TCG}$, which are again the same. Englander et al. [5] proved the following proposition.

Proposition 2.11. *Consider the Fourier coordinates q associated to a binary, semi-directed level-2 trinet $\mathcal{N}$ under the JC model. Then the invariant*

$$q_{011}q_{101}q_{110} - q_{111}^2 \tag{2.1}$$

evaluates to zero if N is a three-leaf tree and is strictly positive if N is a strict level-1 or strict level-2 trinet with a non-trivial 3-blob.

Here, we will only show, as an example of the technique, that proposition 2.11 holds for the three-leaf tree and the strict level-1 trinet with a non-trivial 3-blob, both are shown in Figure 2.5. First we will show that the

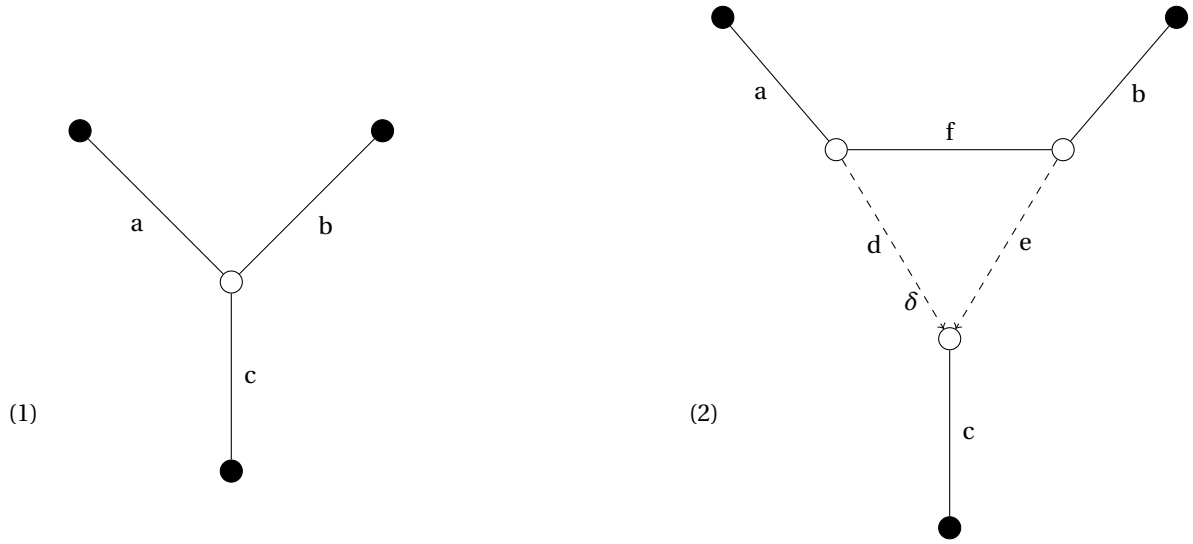


Figure 2.5: (1) is the three-leaf tree and (2) is the strict level-1 trinet. Both are unique up to labelling of the leaves. The edge labels are used in to prove distinguishability.

invariant of the three-leaf tree evaluates to zero. Using the parameters in Figure 2.5, the Fourier coordinates are $q_{111} = abc$, $q_{110} = ab$, $q_{101} = ac$, $q_{011} = bc$. Using these values in the invariant 2.1, it evaluates to zero.
For the strict level-1 network let the edge labelling be as in Figure 2.5. The reticulation parameter δ is in $(0,1)$. The Fourier coordinates are as follows:

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

Substituting these values into the invariant gives the following expression:

$$q_{011}q_{101}q_{110} - q_{111}^2 = a^2b^2c^2f\,\delta(1-\delta)\,e(1-f)^2. \tag{2.2}$$

Using that all parameters are in $(0,1)$ we can conclude that the invariant is strictly positive.
Up to labelling of the leaves there are exactly six strict level-2 networks. For each of them Englander et al. [5] showed the strict positivity of the invariant on a case by case basis, in a similar way as the level-1 network above.

3

Identifiability

In this chapter, we present an extension of the theorem given by Englander et al. [5] about the distinguishability of phylogenetic networks and trees. Englander et al. proved that strict level-1 or strict level-2 networks trinets are distinguishable from a three-leaf tree under the Jukes-Cantor model.
We generalise this result by proving that every strict level-k network containing a non-trivial q-blob with $q \geq 3$ on at least three leaves is distinguishable from a tree on the same leaf set, again under the Jukes–Cantor model. In order to prove this distinguishability Theorem 3.4, Corollary 3.2 and Proposition 3.3 are needed. These results will first be proven.

3.1. Structural results on level-k phylogenetic networks

Lemma 3.1. *Let $\mathcal{N}$ be a binary semi-directed phylogenetic network on at least three leaves with a non-trivial k-blob B with $k \geq 3$, Then there exists a reticulation in B incident to a cut edge. If the network has three leaves, this cut-edge is incident to a leaf.*

Proof. We prove this statement for directed rooted binary phylogenetic networks, and afterwards conclude that it is also true for binary semi-directed phylogenetic networks. Let $\mathcal{M}$ be a directed rooted binary phylogenetic network and replace the root with an edge. Let $\mathcal{B}$ be the blob in $\mathcal{M}$. All internal vertices in $\mathcal{B}$ are of one of the two following types: a tree vertex or a reticulation vertex.
A tree vertex has an in-degree of 1 and out-degree of 2, while a reticulation vertex has an in-degree of 2 and an out-degree of 1. Suppose only tree vertices are incident to vertices outside of $\mathcal{B}$, then those vertices have an out-degree of at least 1 in $\mathcal{B}$, and all the other vertices also have an out-degree of at least 1. This leads to a directed cycle in $\mathcal{B}$, which is a contradiction since $\mathcal{B}$ is the blob in the directed phylogenetic network $\mathcal{M}$ and directed phylogenetic networks do not contain directed cycles. Thus, there must exist at least one reticulation incident to a cut edge.
Since this holds for all directed rooted binary phylogenetic networks $\mathcal{M}$, it is also true for the unrooted semi-directed versions of these networks, which are binary semi-directed phylogenetic networks $\mathcal{N}$. Note that every binary semi-directed phylogenetic network has a rooting, and thus every binary semi-directed phylogenetic network has a reticulation incident to a cut edge.
If the network has three leaves and a non-trivial 3-blob, then the cut-edge incident to the reticulation must also be incident to a leaf. □

Corollary 3.2. *Let $\mathcal{N}$ be a binary semi-directed phylogenetic network on three leaves with a non-trivial 3-blob. For each leaf reticulation r there exists a pair of leaves for which r is not on any up-down path between these leaves.*

Proof. Lemma 3.1 tells us that there is a reticulation directly above a leaf. In the binary semi-directed phylogenetic network $\mathcal{N}$ there is an up-down path between every pair of leaves. Take the other two leaves, distinct from the one identified in the lemma. If the up-down path went through the reticulation above the leaf, then it would not be an up-down path, or one would be stuck in this reticulation. Thus, there exists a pair of leaves such that the reticulation is not on an up-down path between these two leaves. □

Proposition 3.3. *Let $\mathcal{N}$ be a binary semi-directed phylogenetic network on three leaves X. If $\mathcal{N}$ has 3-blob and is level-k with $k \geq 2$ after suppression of 2-sub-blobs, then for every leaf reticulation there exists an incoming edge e of that reticulation such that $\mathcal{N} - e$ has a non-trivial 3-blob.*

Proof. This proof starts with a case distinction and then proves each case has at least one resulting subnetwork containing a non-trivial 3-blob. Since $\mathcal{N}$ is level-k with $k \geq 2$ it has at least two reticulations, by Lemma 3.1 it is known that at least one of these two reticulations has to be above a leaf, we call this reticulation leaf l_r. Let r_l be this leaf reticulation and e' and e'' the incoming edges of r_l coming from v' and v'', respectively. We divide into three cases, and for each case we prove that at least one of $\mathcal{N}' := \mathcal{N} - e'$ and $\mathcal{N}'' := \mathcal{N} - e''$ has a non-trivial 3-blob.
The first possibility, case (i), is that v' and v'' are not reticulation vertices, thus r_l is not in a stack. The second possibility, case (ii), is that one of the parents of r_l is a reticulation vertex. The third possibility, case (iii), is that both v' and v'' are reticulation vertices. Note that it could be possible that the parents of v' and v'' are also reticulation vertices, there could be an arbitrarily large stack. The cases are illustrated in Figure 3.1.

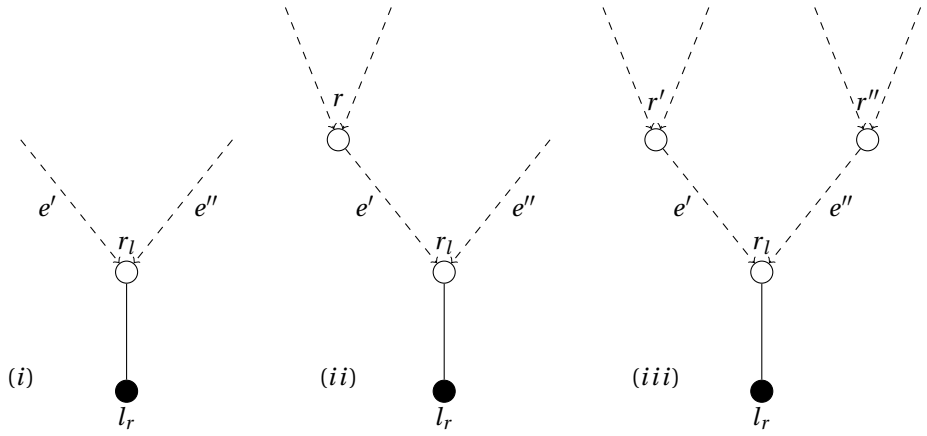


Figure 3.1: The three possible cases for reticulation r_l

We prove that for every case at least one of the subnetworks $\mathcal{N}'$ and $\mathcal{N}''$ contains a non-trivial 3-blob

(i) Both v' and v'' are not reticulation vertices. Let r be a reticulation different from r_l in $\mathcal{N}$. Suppose r is in a 2-sub-blob B of $\mathcal{N}'$ given by $B = \mathcal{N}'[X]$ with $X \subset V$ where V is the vertex set of $\mathcal{N}$. Since only the path from l_r to B uses r_l the sub-blob B is still 2-leaf-connected with the two leaves other than l_r in $\mathcal{N}'$, these paths remain in both $\mathcal{N}'$ and $\mathcal{N}''$. The sub-blob B was 3-leaf-connected in $\mathcal{N}$. This implies there is a path through e' from l_r to B, this path was disconnected in $\mathcal{N}'$. This path remains present in $\mathcal{N}''$, this implies that B is 3-leaf-connected in $\mathcal{N}''$. This proves that at least one of the subnetworks $\mathcal{N}'$ and $\mathcal{N}''$ contains a non-trivial 3-blob. This case is illustrated in the first drawing of Figure 3.2.

(ii) Without loss of generality assume v' is a reticulation vertex with reticulation r and v'' is not a reticulation vertex. In $\mathcal{N}'$ r gets deleted because it is not on any up-down path between two leaves, so $\mathcal{N}'$ can become a tree. Since v'' is a tree vertex no vertices are deleted and only e'' is deleted in $\mathcal{N}''$.
Let B be the blob containing r in $\mathcal{N}''$, and suppose B is not 3-leaf-connected in $\mathcal{N}''$. Since B is a blob in $\mathcal{N}''$ and not 3-leaf-connected it is a 2-blob in $\mathcal{N}''$. If $v'' \notin B$, then B is a 2-sub-blob in $\mathcal{N}$ which is a contradiction. If $v'' \in B$ then $B \cup r_l$ is a 2-sub-blob in $\mathcal{N}$ which is again a contradiction. Therefore the assumption that B was not 3-leaf-connected in $\mathcal{N}''$ is incorrect. This proves that t least one of the subnetworks $\mathcal{N}'$ and $\mathcal{N}''$ contains a non-trivial 3-blob. This case is illustrated in the second drawing of Figure 3.2.

(iii) Both v' and v'' are reticulation vertices, and the stack is possibly higher than v' and v''. Let r' and r'' be the reticulations of v' and v'' respectively. In $\mathcal{N}'$ the incoming edges of r' get deleted because it is not on any up-down path between two leaves and the same is true for $\mathcal{N}''$ and r''. Suppose r'' is in a 2-sub-blob B in $\mathcal{N}'$. B was 3-leaf-connected in $\mathcal{N}$, this implies that a path to a leaf was cut by the deletion of an incoming edge of r' or an incoming edge of a reticulation higher in the stack. Call this leaf l_1, call the other leaf l_2 and call the boundary point of B in $\mathcal{N}$ that connects to leaf l_1 c. In $\mathcal{N}$ the sub-blob B had

a path p_1 to leaf l_1 and a path p_2 to leaf l_2. The path p_1 goes through one or more reticulations of the stack of r', since p_1 was cut in $\mathcal{N}'$. This is illustrated in third drawing in Figure 3.2.
At some reticulation in the stack of r', the path p_1 must traverses an edge against its direction, otherwise the path would end in leaf l_r. Denote this reticulation r, this could possibly coincide with r'.
We prove r is 3-leaf-connected in $\mathcal{N}''$. The path p_1 connects leaf l_1 to r. The path p_2, some of B and the part between of p_1 between c and r connects r to l_2. Since p_1 and p_2 were paths that 3-leaf-connect B in $\mathcal{N}$ they are disjoint. This together with that no overlapping parts of p_1 are being used tells us that these new paths to r also are disjoint. Finally r has a disjoint path to leaf l_r though the stack of r', this makes r 3-leaf-connected. Since r is 3-leaf-connected there is a blob with r that also is 3-leaf-connected in $\mathcal{N}''$. This proves that at least one of the subnetworks $\mathcal{N}'$ and $\mathcal{N}''$ contains a non-trivial 3-blob.

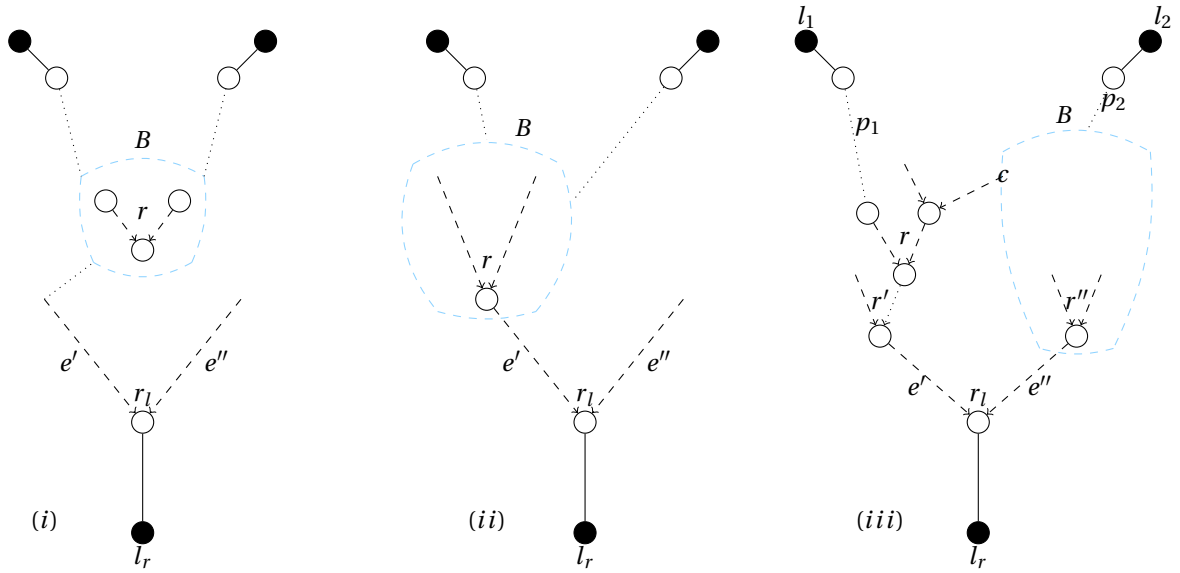


Figure 3.2: Illustration of the proof for each case, from left to right case (i), (ii), (iii). The blue dashed area is the blob or sub-blob defined per case and the dashed lines represent paths.

In all three cases, at least one of the sub-networks obtained by deleting a reticulation edge still contains a non-trivial 3-blob. This completes the proof of Proposition 3.3. □

3.2. Distinguishing Trees and Networks

We now prove the main result, showing that 3-leaf trees are distinguishable from 3-leaf networks. This will then be generalized to $n \geq 3$ leaves in Corollary 3.7.

Theorem 3.4. *Let $\mathcal{T}$ and $\mathcal{N}$ be two binary semi-directed phylogenetic networks on the same three leaves. Suppose that $\mathcal{T}$ does not have a non-trivial 3-blob and $\mathcal{N}$ has a non-trivial 3-blob. Then, under the JC model and if $\mathcal{N}$ has no trivial parameter values, the polynomial invariant*

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

evaluates to zero for $\mathcal{T}$ and is strictly positive for $\mathcal{N}$. Hence, $\mathcal{T}$ and $\mathcal{N}$ are distinguishable under the JC model.

Proof. For trees and for networks without non-trivial 3-blobs, it is already known from Englander et al. [5] that the invariant is equal to zero. We now prove that for every level-k network $\mathcal{N}$ with $k \geq 1$ the invariant is strictly positive. If $k = 1$ then the invariant is strictly positive, this was also proven by Englander et al. [5].
Assume $k \geq 2$, we prove that the invariant is strictly positive using induction on k, where $k = 1$ is the base case. From Lemma 3.1 it is known that there is a leaf-reticulation r . Note that both Corollary 3.2 and Proposition 3.3 apply to leaf reticulations, so we can use both on r. Let e' and e'' be the incoming edges of this reticulation associated with reticulation parameter $\bar{\delta} := 1 - \delta$ and δ. We essentially split the network into the two subnetworks $\tilde{\mathcal{N}}' := \tilde{\mathcal{N}} - e'$ and $\tilde{\mathcal{N}}'' := \tilde{\mathcal{N}} - e''$. These networks from the induction step.

Without loss of generality, assume that the leaf below r is the first leaf in the ordering of the site-pattern distribution. Corollary 3.2 tells us that the up-down path between the other two leaves does not use the leaf reticulation, and therefore $q_{011} = q'_{011} = q''_{011}$, where q_{011}, q'_{011} and q''_{011} correspond to the Fourier coordinates of $\mathcal{N}$, $\mathcal{N}'$ and $\mathcal{N}''$, respectively. Using this, we can split the invariant p of $\mathcal{N}$ in the following way.

$$\begin{aligned}
p &= q_{011}\, q_{101}\, q_{110} - q_{111}^2 \\
&= (\delta q'_{011} + \bar{\delta} q''_{011}) \cdot (\delta q'_{101} + \bar{\delta} q''_{101}) \cdot (\delta q'_{110} + \bar{\delta} q''_{110}) - (\delta q'_{111} + \bar{\delta} q''_{111})^2 \\
&= q_{011} \cdot \big(\delta^2 q'_{101} q'_{110} + \delta\bar{\delta}\, q'_{101} q''_{110} + \delta\bar{\delta}\, q''_{101} q'_{110} + \bar{\delta}^2 q''_{101} q''_{110}\big) \\
&\quad - \big(\delta^2 (q'_{111})^2 + 2\delta\bar{\delta}\, q'_{111} q''_{111} + \bar{\delta}^2 (q''_{111})^2\big) \qquad (3.1) \\
&= \delta^2 \big(q_{011}\, q'_{101} q'_{110} - (q'_{111})^2\big) + \bar{\delta}^2 \big(q_{011}\, q''_{101} q''_{110} - (q''_{111})^2\big) \\
&\quad + \delta\bar{\delta}\, q_{011} \big(q'_{101} q''_{110} + q''_{101} q'_{110}\big) - 2\delta\bar{\delta}\, q'_{111} q''_{111} \qquad (\text{using } q_{011} = q'_{011} = q''_{011}) \\
&= \delta^2 p' + \bar{\delta}^2 p'' + \delta\bar{\delta}\, q_{011} \big(q'_{101} q''_{110} + q''_{101} q'_{110}\big) - 2\delta\bar{\delta}\, q'_{111} q''_{111}
\end{aligned}$$

Here, p' and p'' are the invariants corresponding to $\mathcal{N}'$ and $\mathcal{N}''$. Using Proposition 3.3 on $\mathcal{N}$ it is possible to conclude that $\mathcal{N}'$ or $\mathcal{N}''$ still contains a 3-blob. Using the induction hypothesis gives $p' > 0$ or $p'' > 0$. Thus, one of the first two terms is strictly positive, and the other one is non-negative . Without loss of generality, assume that $p' > 0$, this is used to prove that the cross-term is strictly positive.

$$\delta\bar{\delta}\, q_{011} \big(q'_{101} q''_{110} + q''_{101} q'_{110}\big) - 2\delta\bar{\delta}\, q'_{111} q''_{111} > 0 \qquad (3.2)$$

Using $p' > 0$ and $p'' \geq 0$ we can conclude the following inequalities.

$$q'_{011}\, q'_{101}\, q'_{110} > (q'_{111})^2$$
$$q''_{011}\, q''_{101}\, q''_{110} \geq (q''_{111})^2$$

Taking square roots on both sides and using the fact that every term is positive, it is possible to deduce the following.

$$q'_{111} < \sqrt{q'_{011}\, q'_{101}\, q'_{110}}$$
$$q''_{111} \leq \sqrt{q''_{011}\, q''_{101}\, q''_{110}}$$

Multiplying both inequalities and again using $q_{011} = q'_{011} = q''_{011}$, the following is obtained.

$$q'_{111} q''_{111} < q_{011} \sqrt{q'_{101}\, q'_{110}\, q''_{101}\, q''_{110}}$$

Now using an inequality called the inequality of arithmetic and geometric mean (AM-GM inequality), this inequality holds for positive numbers. This inequality becomes an equality if $x = y$.

$$\sqrt{xy} \leq \frac{x+y}{2} \qquad (3.3)$$

Finally, multiplying both sides by two and applying the AM-GM inequality, we get the following.

$$2q'_{111} q''_{111} < 2q_{011} \sqrt{q'_{101}\, q'_{110}\, q''_{101}\, q''_{110}} \leq q_{011} (q'_{101} q''_{110} + q''_{101} q'_{110})$$

Using this, we can conclude that Equation (3.2) holds and the cross-term is positive. Combined with the fact that $p' > 0$ and $p'' \geq 0$, we can conclude that $p > 0$. □

Remark 3.5. If the decomposition in Equation (3.1) of the proof of Theorem 3.4 is applied to the strict level-1 trinet, $\mathcal{N}'$ and $\mathcal{N}''$ become trees. This implies that $p' = 0$ and $p'' = 0$. In this case, the cross-term in Equation (3.2) becomes strictly positive. This is worked out in Appendix A, showing an independent proof for the strict level-1 case.

Remark 3.6. In the decomposition of Equation (3.1) there is, apart from Proposition 3.3, no rule on how much the level decreases when going from the original network to the subnetworks. In Figure 3.3 a network of an arbitrary high level is shown, within the blue sub-blob B any number of repetitions of the block with reticulations can be present. On the right, it is shown that the two subnetworks can still become a tree and a level-1 network.

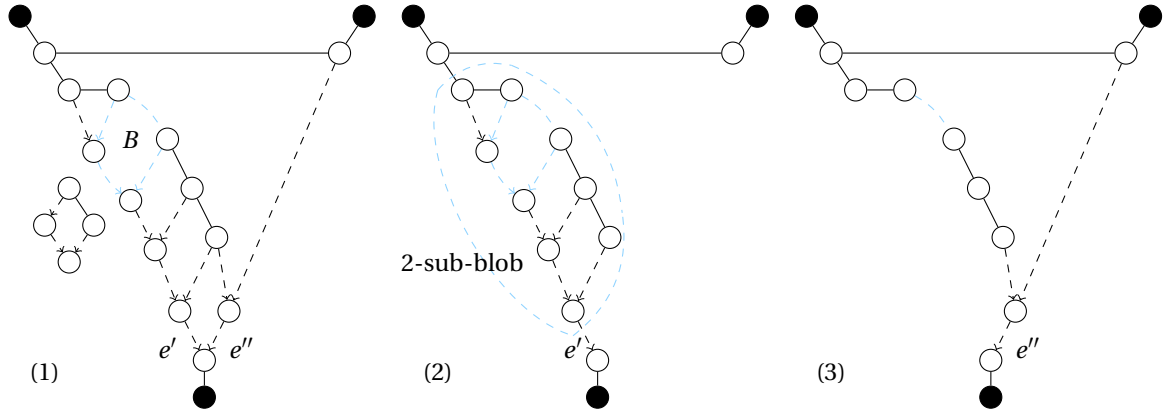


Figure 3.3: A network of arbitrary level where the two subnetworks are a tree and a level-1 network.

Corollary 3.7. *Let $\mathcal{N}$ be a binary semi-directed phylogenetic network containing a non-trivial k-blob B with $k \geq 3$ and let $\mathcal{T}$ be a phylogenetic tree on the same set of leaves $\mathcal{X}$. Then, under the JC model, $\mathcal{T}$ and $\mathcal{N}$ are distinguishable.*

Proof. Lemma 3.1 tells us that there is a reticulation in B incident to a cut-edge. This reticulation is 3-leaf-connected. Let $\tilde{X}$ be the corresponding three leaves, then the induced subnetwork $\mathcal{N}|\tilde{X}$ has a non-trivial 3-blob. Using Theorem 3.4 we can distinguish $\mathcal{N}|\tilde{X}$ from $\mathcal{T}|\tilde{X}$. Using Proposition 2.9, it is possible to distinguish $\mathcal{T}$ and $\mathcal{N}$. □

4

Discussion

In Chapter 3, we have shown that it is possible to distinguish a level-k network from a tree on the same set of leaves under the Jukes-Cantor model. This is one of the first identifiability results for higher-level networks. Given some taxa, it can, in theory, be used to determine whether networks or trees should be considered when constructing their phylogenetic history. As part of our proof, we establish that for any trinet of level-2 or higher, there exists a reticulation such that deleting one of its incoming reticulation edges produces a trinet that remains reticulated and does not reduce to a tree after suppression. This follows from the fact that at least one reticulation is not eliminated by 2-blob or 2-sub-blob suppression. This result could be useful for other proofs involving higher-level trinets.

Our main result is theoretical and it remains to be studied how the invariant reacts to finite data. For example Barton et al. [4] investigated the interpretation of residuals from invariants of level-1 quarnets when observed site-pattern frequencies are substituted into the invariants. They further applied statistical learning methods to determine whether the residuals can be used to infer differences between these networks. Future research could look into applying a similar approach for evaluating the invariant studied in this thesis using observed site-pattern frequencies.

Interestingly, the identifiability result established in this thesis does not directly extend to more general Markov models for evolution. Kouwenhoven [10] showed that a seemingly similar invariant for K2P could evaluate to zero for a level-1 network with a non-trivial 3-blob. Therefore, extending this result to more general Markov models requires a different invariant than the one given by Kouwenhoven, or a different approach.

Another possibility for future research is to apply Theorem 3.4 to induced trinets of different subsets of leaves to determine where reticulations are located in a network. By analysing which subsets of leaves give a trinet that have a non-trivial 3-blob, it could reveal between which taxa reticulations are present. This would suggest that invariant-based tests could serve as a preprocessing step in network reconstruction algorithms.

Bibliography

[1] Elizabeth S. Allman, Héctor Baños, Jonathan D. Mitchell, and John A. Rhodes. Tinnik: inference of the tree of blobs of a species network under the coalescent model. *Algorithms for Molecular Biology*, 19 (1):23, 2024. ISSN 1748-7188. doi: 10.1186/s13015-024-00266-2. URL https://doi.org/10.1186/s13015-024-00266-2.

[2] Cécile Ané, John Fogg, Elizabeth S. Allman, Héctor Baños, and John A. Rhodes. Anomalous networks under the multispecies coalescent: theory and prevalence. *Journal of Mathematical Biology*, 88(3): 29, 2024. ISSN 1432-1416. doi: 10.1007/s00285-024-02050-7. URL https://doi.org/10.1007/s00285-024-02050-7.

[3] Muhammad Ardiyansyah. Distinguishing level-2 phylogenetic networks using phylogenetic invariants, 2021. URL https://arxiv.org/abs/2104.12479.

[4] Travis Barton, Elizabeth Gross, Colby Long, and Joseph Rusinko. Statistical learning with phylogenetic network invariants, 2022. URL https://arxiv.org/abs/2211.11919.

[5] Aviva K. Englander, Martin Frohn, Elizabeth Gross, Niels Holtgrefe, Leo van Iersel, Mark Jones, and Seth Sullivant. Identifiability of phylogenetic level-2 networks under the jukes-cantor model. *bioRxiv*, 2025. doi: 10.1101/2025.04.18.649493. URL https://www.biorxiv.org/content/early/2025/10/09/2025.04.18.649493.

[6] Nathalie Gontier, editor. *Reticulate Evolution*. Interdisciplinary Evolution Research. Springer International Publishing, Cham, 2015. ISBN 978-3-319-16345-1. doi: 10.1007/978-3-319-16345-1. Chapter 1, pp. 1–40.

[7] Elizabeth Gross, Leo van Iersel, Remie Janssen, Mark Jones, Colby Long, and Yukihiro Murakami. Distinguishing level-1 phylogenetic networks on the basis of data generated by markov processes. *Journal of Mathematical Biology*, 83(3):32, 2021. doi: 10.1007/s00285-021-01653-8. URL https://doi.org/10.1007/s00285-021-01653-8.

[8] Niels Holtgrefe, Katharina T Huber, Leo van Iersel, Mark Jones, Samuel Martin, and Vincent Moulton. Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments. *Molecular Biology and Evolution*, 42(4):msaf067, 03 2025. ISSN 1537-1719. doi: 10.1093/molbev/msaf067. URL https://doi.org/10.1093/molbev/msaf067.

[9] Patrick J. Keeling. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nature Reviews Genetics*, 25(6):416–430, 2024. ISSN 1471-0064. doi: 10.1038/s41576-023-00688-5. URL https://doi.org/10.1038/s41576-023-00688-5.

[10] Guuske Anne Kouwenhoven. Identifiability of phylogenetic trinets, 2025. URL https://repository.tudelft.nl/record/uuid:5aa152f2-e99c-4bb8-bb27-c112ab50b5f4#metadata. Bachelor Thesis.

[11] Ulrich Kutschera and Karl J. Niklas. The modern theory of biological evolution: an expanded synthesis. *Naturwissenschaften*, 91(6):255–276, 2004. ISSN 1432-1904. doi: 10.1007/s00114-004-0515-y. URL https://doi.org/10.1007/s00114-004-0515-y.

[12] Nick Patterson, Daniel J. Richter, Sante Gnerre, Eric S. Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108, 2006. ISSN 1476-4687. doi: 10.1038/nature04789. URL https://doi.org/10.1038/nature04789.

[13] S. Roy, R. Dasgupta, and A. Bagchi. A review on phylogenetic analysis: A journey through modern era. *Computational Molecular Bioscience*, 4:39–45, 2014. doi: 10.4236/cmb.2014.43005. URL https://doi.org/10.4236/cmb.2014.43005.

[14] Shannon M. Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015. ISSN 1471-0064. doi: 10.1038/nrg3962. URL https://doi.org/10.1038/nrg3962.

[15] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants, 2004. URL https://arxiv.org/abs/q-bio/0402015.

[16] Seth Sullivant. *Algebraic Statistics*, volume 194 of *Graduate Studies in Mathematics*, chapter 15, pages 329–358. American Mathematical Society, 2018.

[17] Seth Sullivant. Phylogenetic network models as graphical models, 2025. URL https://arxiv.org/abs/2507.23056.

[18] Xiao-Xiao Wang, Chien-Hsun Huang, Diego F. Morales-Briones, Xiang-Yu Wang, Ying Hu, Na Zhang, Pu-Guang Zhao, Xiao-Mei Wei, Kun-Hua Wei, Xinya Hemu, Ning-Hua Tan, Qing-Feng Wang, and Ling-Yun Chen. Phylotranscriptomics reveals the phylogeny of asparagales and the evolution of allium flavor biosynthesis. *Nature Communications*, 15(1):9663, 2024. doi: 10.1038/s41467-024-53943-6.

A

An alternative proof for level-1

In this appendix, we work out the details of Remark 3.5, which can be used as an alternative proof for the distinguishability between a 3-leaf tree and a strict level-1 trinet with a non-trivial 3-blob.
The proof of Theorem 3.4 assumes that the leaf reticulation we split is first in the labelling. Therefore, the labelling of the leaves in the Fourier coordinates is as follows $q_{g_3g_2g_1}$. Now, since the network is level-1 both the subnetworks are trees and therefore, $p' = p'' = 0$. In Figure A.1 the level-1 network is given with edge labelling.

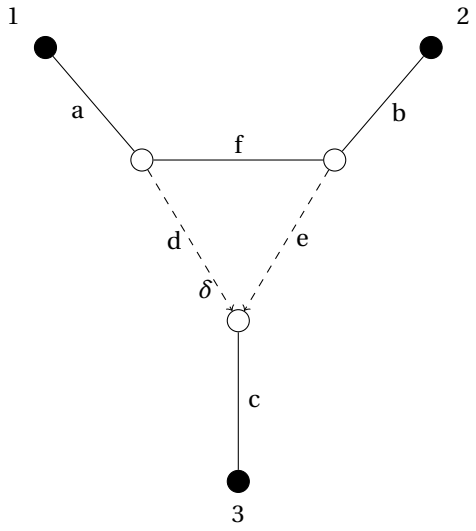


Figure A.1: Level-1 network with edge labelling

Now, the Fourier coordinates of the level-1 network are given by q, and the Fourier coordinates of the subnetworks are given by q' for $\mathcal{N}' := \mathcal{N} - e$ and q'' for $\mathcal{N}'' := \mathcal{N} - d$.

$$
\begin{aligned}
q_{111} &= cbaf\big(\delta d + (1-\delta)e\big), & q'_{111} &= cbafd, & q''_{111} &= cbafe,\\
q_{110} &= cb\big(\delta df + (1-\delta)e\big), & q'_{110} &= cbdf, & q''_{110} &= cbe,\\
q_{101} &= ca\big(\delta d + (1-\delta)ef\big), & q'_{101} &= cad, & q''_{101} &= caef,\\
q_{011} &= bfa, & q'_{011} &= bfa, & q''_{011} &= bfa.
\end{aligned}
$$

$$
\begin{aligned}
p &= \delta^2 p' + \bar{\delta}^2 p'' + \delta\bar{\delta}(bfa\cdot cad\cdot cbe + bfa\cdot caef\cdot cbdf - 2cbafd\cdot cbafe)\\
&= \delta\bar{\delta}(c^2b^2a^2def(1+f^2-2f))\\
&= \delta\bar{\delta}(c^2b^2a^2def(1-f)^2) > 0
\end{aligned}
$$

Filling these values into decomposition in Equation (3.1) only the cross term from Equation (3.2) remains. Factoring out the common components, one can see that the cross term is strictly positive, which gives an alternative proof for the strict level-1 network.