Journal of Mathematical Biology (2022) 84:47
https://doi.org/10.1007/s00285-022-01746-y

**Mathematical Biology**

# Classes of explicit phylogenetic networks and their biological and mathematical significance

**Sungsik Kong[1] · Joan Carles Pons[2] · Laura Kubatko[1,3] · Kristina Wicke[4]**

Received: 21 September 2021 / Revised: 18 January 2022 / Accepted: 31 March 2022 /
Published online: 3 May 2022

## Abstract

The evolutionary relationships among organisms have traditionally been represented using rooted phylogenetic trees. However, due to reticulate processes such as hybridization or lateral gene transfer, evolution cannot always be adequately represented by a phylogenetic tree, and rooted phylogenetic networks that describe such complex processes have been introduced as a generalization of rooted phylogenetic trees. In fact, estimating rooted phylogenetic networks from genomic sequence data and analyzing their structural properties is one of the most important tasks in contemporary phylogenetics. Over the last two decades, several subclasses of rooted phylogenetic networks (characterized by certain structural constraints) have been introduced in the literature, either to model specific biological phenomena or to enable tractable mathematical and computational analyses. In the present manuscript, we provide a thorough review of these network classes, as well as provide a biological interpretation of the structural constraints underlying these networks where possible.

Sungsik Kong and Joan Carles Pons have contributed equally to this work.

✉ Kristina Wicke
wicke.6@osu.edu

Sungsik Kong
kong.362@osu.edu

Joan Carles Pons
joancarles.pons@uib.es

Laura Kubatko
lkubatko@stat.osu.edu

[1] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH, USA

[2] Department of Mathematics and Computer Science, University of the Balearic Islands, Palma 07122, Spain

[3] Department of Statistics, The Ohio State University, Columbus, OH, USA

[4] Department of Mathematics, The Ohio State University, Columbus, OH, USA

In addition, we discuss how imposing structural constraints on the network topology can be used to address the scalability and identifiability challenges faced in the estimation of phylogenetic networks from empirical data.

**Keywords** Hybridization · Introgression · Lateral gene transfer · Phylogenetic tree · Phylogenetic network

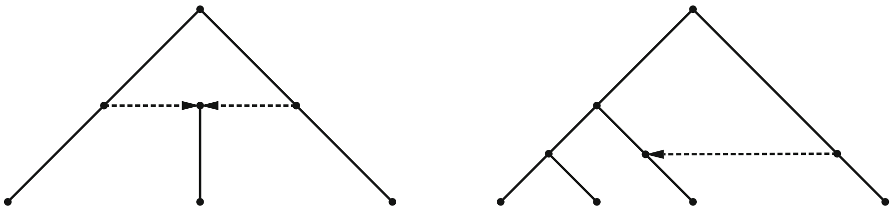
**Mathematics Subject Classification** 05C20 · 05C90 · 92-08 · 92B10

## 1 Introduction

Reconstructing and analyzing the evolutionary relationships among organisms is a central goal in evolutionary biology. Rooted phylogenetic trees (often simply referred to as *phylogenies*) have traditionally been used to represent the evolutionary history for a collection of taxa. In a rooted phylogenetic tree, the leaves (or terminal vertices) usually represent sampled extant taxa, the root corresponds to the most recent common ancestor of the taxa under consideration, and all other interior vertices can be interpreted as split (or speciation) events, where some ancestral taxon evolved into two (or more) distinct taxa. In particular, rooted phylogenetic trees assume vertical inheritance, where genomic material is transmitted from an ancestral species to a descendant species.

However, it is nowadays widely accepted that the evolutionary pathway of an organism is not always tree-like and that many systems in nature experience events in which genetic information is transferred 'horizontally' between taxa rather than 'vertically'. These events include hybridization (which in turn includes hybrid speciation and introgression (Anderson 1953)), lateral gene transfer (LGT; sometimes also called horizontal gene transfer (HGT)), and recombination. Briefly, hybrid speciation refers to the emergence of a novel lineage through interbreeding between two distinct parental lineages. Introgression describes the transmission of genetic information from one lineage (the donor) to another lineage (the recipient) by repeated backcrossing of a hybrid daughter with one of its parents, whereas LGT usually refers to the transmission of genetic material from a donor to a recipient via processes such as transformation, transduction, or conjugation (see Fig. 1 for a schematic representation of hybrid speciation and introgression/LGT). Finally, recombination refers to a rearrangement of genetic material through crossing over of chromosomes during reproduction between a pair of individuals in the same lineage. Note that recombination is thus an intraspecific process, whereas hybrid speciation and LGT are interspecific processes.

Because phylogenetic trees are not adequate to represent non-treelike evolutionary histories such as those described above, rooted phylogenetic networks have been proposed as a generalization of rooted phylogenetic trees in the literature. Rooted phylogenetic networks are often referred to as explicit (directed) networks since they explicitly depict the evolution of a group of organisms from a common ancestor via a combination of splitting and reticulation events. In brief, an explicit phylogenetic network is a rooted directed graph with all edges directed away from the root (see Definition 1 for a formal definition). It is important to distinguish explicit networks



**(a)** Hybrid speciation **(b)** Introgression/LGT

**Fig. 1** **a** Schematic representation of hybrid speciation, where the horizontal dotted lines indicate interbreeding from two distinct ancestral populations that leads to formation of a hybrid taxon. **b** Schematic representation of introgression/LGT, where the right-most lineage transfers genetic information to another lineage at the point in the evolutionary history indicated by the directed horizontal dotted edge

from implicit or abstract networks such as split networks (Bandelt and Dress 1992) or median-joining networks (Bandelt et al 1999) that depict phenetic relatedness based on overall similarity among taxa but do not represent information on the evolutionary history or direction of evolution (Kong et al 2015; Sánchez-Pacheco et al 2020). More recently, an 'intermediate' class of phylogenetic networks, semi-directed phylogenetic networks, was introduced in the literature (e.g., Solís-Lemus and Ané (2016)). Roughly, semi-directed phylogenetic networks are obtained from rooted phylogenetic networks by suppressing the root and ignoring the direction of all edges, except for those involved in a reticulation event, thus keeping information on which vertices are reticulations.

In this paper, we focus on explicit networks as these networks directly model the evolutionary history of a collection of species, rather than simply displaying species similarity, and thus have direct biological relevance. The reconstruction and analysis of these networks is one of the most active areas of research in computational or mathematical phylogenetics and a variety of subclasses of rooted phylogenetic networks have been introduced in the literature. Some of these classes were introduced to explicitly model specific evolutionary scenarios, whereas others seem to have been mainly established for mathematical convenience or computational tractability.

We aim at providing a thorough review of these network classes, as well as an analysis of which are biologically interpretable (in the sense that they depict realistic evolutionary scenarios and are expected to capture features of the true evolutionary history for empirical data sets) and which are merely of mathematical interest. In addition, we discuss how imposing structural constraints on the networks can address mathematical and computational challenges faced when estimating phylogenetic networks from data. Here, we focus in particular on the notions of scalability and identifiability. By *scalability*, we mean how the performance of the inference method, in terms of both computational time and accuracy, changes as the size of the problem (e.g., the number of taxa, the amount of available data, or both; or the complexity of the network, e.g., the number of reticulations in the network) increases. By *identifiability*, we are referring to the study of which features of the phylogenetic network are uniquely determined by the model from which the data arise or are uniquely characterized by certain substructures of the network like 'displayed' trees.

To our knowledge, this study is the first comprehensive review of the more than 20 classes of explicit phylogenetic networks discussed in the literature that not only summarizes their structural properties but also establishes a connection to biological processes (where possible), discusses some scalability challenges faced in network estimation in practice and reviews identifiability results obtained for different classes. We remark, however, that there are some excellent books or book chapters on the general topic of phylogenetic networks and related concepts (e.g., Huson et al 2011; Morrison 2011; Gusfield 2014; Steel 2016; Zhang 2019) as well as some web resources such as the websites "Who is Who in Phylogenetic Networks"[1] (Agarwal et al 2016) and "ISIPhyNC (Information System on Inclusions of Phylogenetic Network Classes)"[2] (Gambette et al 2018b).

Note, however, that we restrict our review to binary phylogenetic networks, in which each speciation event leads to precisely two new species and each reticulation event involves exactly two parental species. On one hand, rooted binary phylogenetic networks are the most common type of explicit phylogenetic networks studied in the literature. On the other hand, we are particularly interested in the biological meaning of the networks under consideration and it is biologically unlikely that a speciation event results in three or more new lineages or that a reticulation event involves three or more distinct species. Nevertheless, we remark that rooted non-binary (or multifurcating) phylogenetic networks can, for example, be useful to reflect uncertainty in the order of speciation or reticulation events.

The remainder of this paper is organized as follows. We begin by introducing and discussing the most important concept of this paper, namely rooted binary phylogenetic networks, in Sect. 2. In addition, we introduce additional terminology and notation used throughout this manuscript. In Sect. 3, we then provide a comprehensive review of more than 20 classes of rooted binary phylogenetic networks currently used in the literature. We first describe classes of networks with an underlying biological interpretation (Sect. 3.1), before discussing additional classes not linked to biology yet (Sect. 3.2). Afterwards, in Sect. 4, we discuss two important mathematical and computational challenges in estimating phylogenetic networks from data. More precisely, we consider the notions of scalability and identifiability. In particular, we describe why and how scalability and identifiability issues affect phylogenetic network estimation in practice. We also review approaches used to address these challenges, and we summarize positive and negative results obtained, highlighting the impact of structural constraints on the estimation of networks. We end with a brief discussion in Sect. 5.

## 2 Rooted binary phylogenetic networks and related concepts

Before we can review and discuss various classes of rooted binary phylogenetic networks and discuss their biological relevance, we need to provide a formal definition and to introduce some additional definitions and concepts.

[1] https://phylnet.univ-mlv.fr/.

[2] https://phylnet.univ-mlv.fr/isiphync/index.php.

### *Phylogenetic networks and phylogenetic trees*

Throughout the paper, $X$ denotes a non-empty finite set (of taxa).

**Definition 1** (Rooted binary phylogenetic network) A *rooted binary phylogenetic network* $\mathcal{N} = (V, E)$ *on* $X$ *with root* $\rho$ is a rooted directed acyclic graph with no parallel arcs satisfying the following properties:

(i) the (unique) root $\rho$ has in-degree zero and out-degree two;
(ii) a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is identified with $X$;
(iii) all other vertices have either in-degree one and out-degree two, or in-degree two and out-degree one.

For technical reasons, if $|X| = 1$, we allow $\mathcal{N}$ to consist of the single vertex in $X$. Moreover, the vertices in $X$ are called *leaves*, the vertices with in-degree one and out-degree two are called *tree vertices*, and the vertices with in-degree two and out-degree one are called *reticulation vertices*. All arcs directed into a reticulation vertex are called *reticulation arcs* and all other arcs are called *tree arcs*. An example of a rooted binary phylogenetic network is depicted in Fig. 2. Here, as in all subsequent figures in the paper, arcs are directed down the page.

Two networks $\mathcal{N}_1 = (V_1, E_1)$ and $\mathcal{N}_2 = (V_2, E_2)$ on $X$ are said to be *isomorphic* if there exists a bijection $\varphi : V_1 \rightarrow V_2$ such that $\varphi(x) = x$ for all $x \in X$, and $(u, v)$ is an arc in $\mathcal{N}_1$ if and only if $(\varphi(u), \varphi(v))$ is an arc in $\mathcal{N}_2$.

A *rooted binary phylogenetic* $X$*-tree* $\mathcal{T}$ is a rooted binary phylogenetic network on $X$ with no reticulation. Unless explicitly stated otherwise, when we refer to phylogenetic networks (trees), we will always mean rooted binary phylogenetic networks (trees).

Finally, note that for most parts of the paper we are interested in structural properties of different phylogenetic networks, i.e., we are interested in network *topologies* or *shapes*. However, when discussing the concepts of scalability and identifiability in Sect. 4, we will also mention branch lengths and inheritance probabilities. In this case, we assume that each arc of $\mathcal{N}$ has a non-negative real-valued length, i.e., we assume that there exists a mapping $w : E \rightarrow \mathbb{R}_{\geq 0}$ under which each arc $e$ of $\mathcal{N}$ is assigned the weight $w(e)$. Additionally, if $v$ is a reticulation vertex with in-coming reticulation arcs $e_1 = (u_1, v)$ and $e_2 = (u_2, v)$, we assume that $e_1$ and $e_2$ are associated with probabilities $\gamma_{e_1} \in (0, 1)$ and $\gamma_{e_2} = 1 - \gamma_{e_1}$ that indicate the proportional contribution of genetic material or features that vertex $v$ inherits from its parents $u_1$ and $u_2$, respectively.

### *Inherent assumptions underlying Definition 1*

Before discussing additional concepts related to phylogenetic networks, we want to point out some inherent assumptions about the evolutionary history of a group of organisms that underlie Definition 1. First, as indicated in the introduction, phylogenetic networks in the sense of this definition depict evolution as a directed process starting at the root of the network and moving towards its leaves. Here, the root corresponds to the most recent common ancestor of the organisms under consideration, and this ancestral organism is assumed to be unique. However, the phylogenetic network does not contain any information about the ancestry of the root itself. Moreover, it

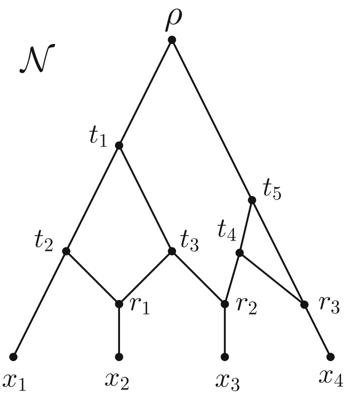


**Fig. 2** Rooted binary phylogenetic network $\mathcal{N}$ on $X = \{x_1, \ldots, x_4\}$. All arcs are directed down the page beginning from the root $\rho$. Vertices labelled $t_*$ are tree vertices of $\mathcal{N}$ and vertices labelled $r_*$ are reticulation vertices. Arcs $(t_*, r_*)$ are reticulation arcs and all others are tree arcs. Apart from $t_3$ and $t_4$ all vertices of $\mathcal{N}$ are visible. Moreover, the arc $(t_5, r_3)$ is a shortcut as there is also a directed path from $t_5$ to $r_3$ in $\mathcal{N}$ (i.e., the path $(t_5, t_4, r_3)$)

is assumed that all taxa under consideration correspond to the leaves of the network, whereas all internal vertices represent hypothetical ancestral species. In particular, it is assumed that all data observed today is observed at the leaves of the network.

Additionally, Definition 1 establishes a bijection between the taxa under consideration (represented by the set $X$) and the set of leaves of the network (since the two sets are identified). This means that a particular leaf of a phylogenetic network represents precisely one species and a particular species is represented by precisely one leaf. In other words, multi-labelling of leaves (i.e., two or more leaves representing the same species) is excluded. However, there is a close relationship between phylogenetic networks and multi-labelled trees (MUL-trees), leaf-labelled trees where more than one leaf may have the same label, that for example arise in the study of polyploids (organisms having multiple complete copies of their genome). In particular, it has been shown that a phylogenetic network can be 'unfolded' to obtain a MUL-tree, and a MUL-tree can under certain conditions be 'folded' into a phylogenetic network that exhibits it (Huber and Moulton 2006; Huber et al 2016; Huber and Scholz 2020).

Third, and again as indicated in the introduction, Definition 1 imposes certain constraints on the degrees of vertices of the network. Tree vertices have an out-degree of precisely two, implying that each speciation events leads to precisely two new species. In other words, *polytomies*, i.e., vertices with an out-degree of three or greater which can either be interpreted as uncertainty in the order of speciation events (soft polytomies) or simultaneous divergence of three or more species (hard polytomies) are excluded from the definition. Similarly, reticulation vertices have an in-degree of exactly two, implying that exactly two parental species are involved in a reticulation event. In addition, all reticulation vertices have an out-degree of one. This can be seen as a way of representing the reticulation *process* with the single child of a reticulation vertex representing the *resulting* taxon. Note that by collapsing both vertices (i.e., the reticulation vertex and its child) into a single vertex representing both the reticulation process and the resulting organism, equivalent mathematical models are obtained. We remark, however, that requiring reticulation vertices to have an out-degree of one does not exclude the single child of the reticulation vertex from being a reticulation vertex itself. As an example, the single child of the reticulation vertex $u$ of the phylogenetic network $\mathcal{N}_2$ depicted in Fig. 4 is itself a reticulation vertex. Last, there are no vertices of in-degree less or equal to one and out-degree one (often referred to as *elementary* vertices). An elementary vertex would represent a species that has only one descendant, and it is impossible to distinguish this ancestral species from its unique descendant through biological information only.

As a final observation note that both arcs directed into a reticulation vertex are commonly referred to as ‘reticulation arcs’. In particular, they are treated symmetrically and are not distinguished from each other. While this assumption is suitable for modelling hybridization or recombination events where both parental organisms play a symmetrical role in producing the resulting organism, it is less suitable when modelling events of LGT or introgression. In this case, it is often assumed that every reticulation vertex has a single (incoming) reticulation arc, whereas the other in-coming arc is a non-reticulation arc. Under these assumptions, a phylogenetic network can be seen as a ‘backbone tree’ composed by non-reticulation arcs representing the main line of evolution with additional arcs, that is reticulation arcs, added to it (Cardona et al (2015); Francis and Steel (2015)). We will elaborate on this idea when introducing LGT networks below.

However, we first need to review some additional concepts related to phylogenetic networks that will be of relevance throughout this manuscript.

### *(Lowest stable) ancestors, descendants, and siblings*

Let $\mathcal{N} = (V, E)$ be a phylogenetic network. If there is an arc $e = (u, v)$ in $\mathcal{N}$, we say that $u$ is a *parent* of $v$, and $v$ is *child* of $u$. Note that $u$ is sometimes also called the *tail* of $e$ and $v$ is called the *head* of $e$. More generally, if there is a directed path between $u$ and $v$, we say that $u$ is an *ancestor* of $v$, and $v$ is a *descendant* of $u$. For vertices $u, v \in V$, we write $u \prec_{\mathcal{N}} v$ if there is a directed path from $u$ to $v$ (and $u \neq v$). In addition, we write $u \preceq_{\mathcal{N}} v$ if $u = v$ or $u \prec_{\mathcal{N}} v$. Given a subset $U \subseteq V$ of the vertices of $\mathcal{N}$, we say that $u \in U$ is a *lowest* vertex in $U$ if there is no $v \in U$ with $u \prec v$. Now, let $U \subseteq V$ be a subset of the vertices of $\mathcal{N}$. Then a *stable ancestor* of $U$ in $\mathcal{N}$ is a vertex $v \in V \setminus U$ such that every path from the root to a vertex in $U$ contains $v$. Moreover, the (unique) *lowest stable ancestor* of $U$ in $\mathcal{N}$ is the lowest such vertex and is denoted by $\mathrm{LSA}_{\mathcal{N}}(U)$. As an example, consider the rooted binary phylogenetic network depicted in Fig. 2 and let $U = \{x_1, x_2\}$. Then, $\mathrm{LSA}_{\mathcal{N}}(U) = t_1$. Finally, two vertices $u$ and $v$ of $\mathcal{N}$ are called *siblings* if they have a common parent.

### *Visible vertices, clusters, and shortcuts*

A vertex $v$ of $\mathcal{N}$ is said to be *visible* if there is a leaf $x \in X$ such that every directed path from the root of $\mathcal{N}$ to $x$ traverses $v$ (if $v$ is a leaf simply take $x = v$). We also say that $x$ *verifies the visibility of* $v$ or $x$ *verifies* $v$ for short (Bordewich and Semple 2016). As an example, all vertices of the network $\mathcal{N}$ depicted in Fig. 2 apart from $t_3$ and $t_4$ are visible. For example, the reticulation vertex $r_1$ is visible because every directed path in $\mathcal{N}$ from the root to leaf $x_2$ traverses $r_1$. Thus, $x_2$ verifies the visibility of $r_1$.

Given a phylogenetic network $\mathcal{N} = (V, E)$ on $X$, the *cluster* associated with a vertex $v \in V$ is the set

$$c_{\mathcal{N}}(v) = \{x \in X : v \preceq_{\mathcal{N}} x\},$$

that is the subset of leaves that can be reached from $v$. As an example, consider vertex $t_1$ of the network $\mathcal{N}$ depicted in Fig. 2. Here, $c_{\mathcal{N}}(t_1) = \{x_1, x_2, x_3\}$.

Lastly, an arc $e = (u, v)$ of a phylogenetic network $\mathcal{N}$ is called *redundant* or a *shortcut* if there is a directed path in $\mathcal{N}$ from $u$ to $v$ that does not use $e$. As an example, arc $(t_5, r_3)$ of the phylogenetic network $\mathcal{N}$ depicted in Fig. 2 is a shortcut as there is also a directed path in $\mathcal{N}$ from $t_5$ to $r_3$ via $t_4$. Note that the shortcut in $\mathcal{N}$ induces an (undirected) cycle of length three, consisting of the vertices $t_5$, $t_4$, and $r_3$. However, shortcuts can also be part of larger cycles. For instance, the network $\mathcal{N}_1$ depicted in Fig. 10 contains the shortcut $(u, r_4)$, which is part of an (undirected) cycle of length four. In fact, a shortcut can be part of a cycle of arbitrary length.

***Edge subdivision and vertex suppression***

Finally, we need to introduce two operations on graphs that will be relevant for various network classes discussed below, namely the concepts of *subdividing an edge* and *suppressing a vertex*. Let $\mathcal{N}$ be a phylogenetic network and let $e = (u, v)$ be an arc of $\mathcal{N}$. Then, we say that we *subdivide* $e$, by deleting $e$, introducing a new vertex $w$, and adding the arcs $(u, w)$ and $(w, v)$. Note that the new (elementary) vertex $w$ of in-degree one and out-degree one is often also referred to as an *attachment point*. Conversely, given a vertex $w$ of in-degree one and out-degree one with adjacent vertices $u$ and $v$, by *suppressing* $w$ we mean deleting $w$ and its two incident arcs $(u, w)$ and $(w, v)$ and introducing a new arc $(u, v)$.

## 3 Classes of rooted binary phylogenetic networks

While Definition 1 is the most general definition of rooted binary phylogenetic networks, numerous additional topological restrictions and subclasses have been defined in the literature in recent years. Often these restrictions were introduced to make certain computational and mathematical problems more tractable, but not necessarily to model particular evolutionary scenarios. Therefore, these subclasses must be distinguished from subclasses of phylogenetic networks often defined in biological studies where the categorization is based on the biological phenomenon (e.g., hybridization networks, ancestral recombination graphs) or on the algorithms and software used to infer the networks (e.g., methods discussed in Sect. 4.1).

### 3.1 Classes of rooted binary phylogenetic networks with a link to biological processes

We begin by introducing various classes of rooted binary phylogenetic networks that have been linked to certain evolutionary processes in the literature and are thought to be biologically meaningful.

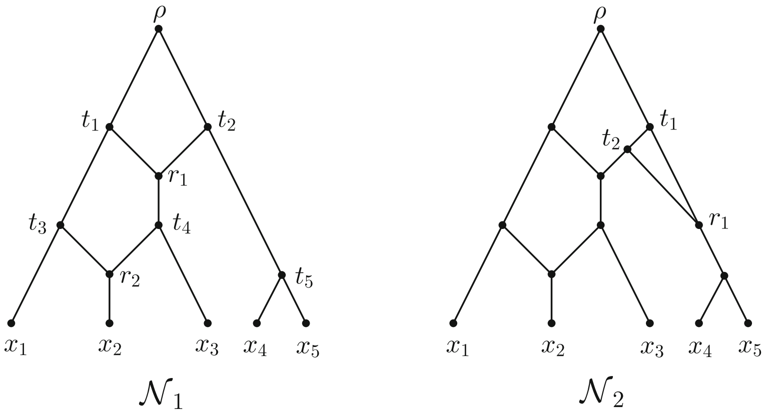


**Fig. 3** Temporal phylogenetic network $\mathcal{N}_1$ and non-temporal network $\mathcal{N}_2$

### *Temporal or time-consistent networks*

A phylogenetic network $\mathcal{N} = (V, E)$ is called *temporal* or *time-consistent* (Baroni et al 2006) if there is a function $t : V \rightarrow \mathbb{N}_{\geq 0}$ such that, for each arc $(u, v)$ of $\mathcal{N}$, the following two properties hold:

(T1) $t(u) = t(v)$ if $(u, v)$ is a reticulation arc;
(T2) $t(u) < t(v)$ if $(u, v)$ is a tree arc.

The map $t$ is called a *temporal labeling* of $\mathcal{N}$. Note that even if $\mathcal{N}$ is temporal, the temporal labeling is not unique. More precisely, if $t$ is a temporal labeling of $\mathcal{N}$, then for each $i \in \mathbb{N}_{\geq 0}$, the map $t' : V \rightarrow \mathbb{N}_{\geq 0}$ defined as $t'(u) = t(u) + i$ for all $u \in V$ also satisfies conditions (T1) and (T2). In particular, while time consistency indicates a potential historical scenario of evolution, it does not represent a unique possible time assignment.

The biological motivation behind this concept, however, is that for a hybridization event to have occurred, the two species involved (along with the hybrid they formed) must have been extant at the same time (T1). Vertical descent from an ancestral species to a descendant species, on the other hand, implies a passage of time (T2).

Note that while all rooted phylogenetic trees have a temporal labeling, this is not necessarily the case for general networks. Consider, for example, the network $\mathcal{N}_2$ depicted in Fig. 3. To see why no temporal labeling exists for this network, suppose for the sake of a contradiction that $\mathcal{N}_2$ has a temporal labeling. Then, condition (T1) implies that $t(r_1) = t(t_1) = t(t_2)$. On the other hand, by condition (T2), we have $t(t_1) < t(t_2)$, a contradiction. Network $\mathcal{N}_1$ depicted in Fig. 3, on the other hand, has a temporal labeling. For instance, it is easily verified that the map $t : V \rightarrow \mathbb{N}_{\geq 0}$ with $t(\rho) = 0$, $t(t_1) = t(t_2) = t(r_1) = 1$, $t(t_3) = t(t_4) = t(t_5) = t(r_2) = 2$, and $t(x_1) = t(x_2) = t(x_3) = t(x_4) = t(x_5) = 3$ satisfies conditions (T1) and (T2).

### *Tree-child networks*

A phylogenetic network $\mathcal{N}$ is a *tree-child* network (Cardona et al 2009b) if every non-leaf vertex is the parent of a tree vertex (see Fig. 4 for examples of tree-child and non-tree-child phylogenetic networks). Equivalently, for every vertex $v$ of a tree-child network $\mathcal{N}$, there is a path from $v$ to a leaf that consists only of tree vertices (except the leaf and possibly $v$ itself). This property is also referred to as the *tree-path* property (Cordue et al 2014). Note that this also implies that every vertex of a tree-child network is visible. Another equivalent definition of tree-child networks is the following: $\mathcal{N}$ is a tree-child network if (i) no tree vertex is the parent of two reticulations and (ii) no reticulation is the parent of another reticulation (Semple 2015).

Biologically, tree-child networks represent a scenario where some portion of each non-extant taxon (that may be the ancestor of contemporary taxa) had some descendants that persisted in the environment via mutation instead of hybridization. Such a scenario is more likely in nature than a 'non-tree-child' scenario, because although the proportion of extant species with hybrid origin is higher than previously thought, due to pre- and/or post-reproductive barriers, hybridization is still considered relatively uncommon compared to traditional speciation events.

When tree child networks in addition satisfy the time-consistency property, they are referred to as *tree-child time consistent* (TCTC) networks (Willson 2007; Cardona et al 2009a). Networks in this class are thought to be biologically meaningful (Willson 2007). In particular, TCTC networks combine the properties of time-consistency (where (i) tree children temporally exist later than their parents, (ii) hybrid children coexist in time with their parents, and (iii) the parents of a hybrid species also coexist in time) and the tree-child condition, which imposes that every non-extant taxon has some descendants that diverged through mutation alone (Cardona et al 2010).

***Normal and regular networks***

A phylogenetic network $\mathcal{N} = (V, E)$ is a *normal* network (Willson 2009) if it is tree-child and has no shortcuts. As an example, both networks depicted in Fig. 3 are tree-child, but only $\mathcal{N}_1$ is normal; $\mathcal{N}_2$ contains a shortcut, arc $(t_1, r_1)$. Note that normal networks 'inherit' the biological meaning of tree-child networks. In addition, they have a number of convenient mathematical and computational properties that make them a popular class of networks in mathematical phylogenetics. As an example, normal networks are uniquely characterized by their displayed caterpillar trees on three and four leaves, whereas arbitrary phylogenetic networks cannot be encoded this way (for details and definitions, see Sect. 4.2.2).

A phylogenetic network $\mathcal{N} = (V, E)$ is called a *regular* network (Baroni et al 2005) if it satisfies the following three properties (adapted from Steel (2016)):

(R1) If $u, v \in V$ are distinct, then $c_{\mathcal{N}}(u) \neq c_{\mathcal{N}}(v)$;
(R2) $u \preceq_{\mathcal{N}} v$ if and only if $c_{\mathcal{N}}(v) \subseteq c_{\mathcal{N}}(u)$;
(R3) $\mathcal{N}$ has no redundant arcs/shortcuts.

Notice that by (R1) a regular network cannot contain an edge $(u, v)$ leading from a vertex $u$ of out-degree one to a vertex $v$ of in-degree one. This might seem very restrictive, but stems from the fact that Baroni et al (2005) introduced this concept for more general networks (so-called 'hybrid phylogenies') than the ones considered in this manuscript. It has been argued by Willson (2007) that under an evolutionary model of "gene aggregation" (essentially, a perfect phylogeny model assuming binary

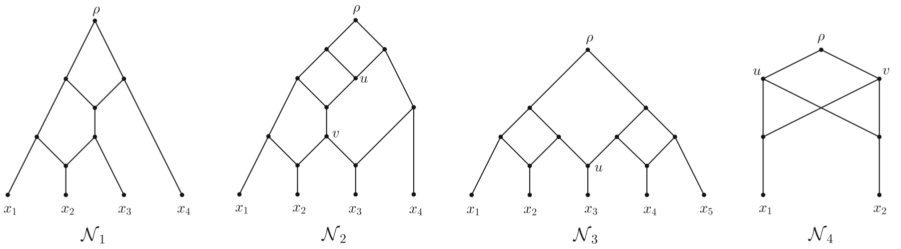


**Fig. 4** Four phylogenetic networks. Network $\mathcal{N}_1$ is tree-child and tree-sibling. Network $\mathcal{N}_2$ is tree-sibling because every reticulation of $\mathcal{N}_2$ has a sibling that is a tree vertex; however, $\mathcal{N}_2$ is not a tree-child network because vertices $u$ and $v$ are not parents of tree vertices. Network $\mathcal{N}_3$ is not tree-sibling (because both siblings of the reticulation vertex $u$ are reticulations) and thus in particular not tree-child, and analogously $\mathcal{N}_4$ is neither tree-sibling nor tree-child. Moreover, networks $\mathcal{N}_1$, $\mathcal{N}_3$, and $\mathcal{N}_4$ are stack-free, whereas $\mathcal{N}_2$ is not (because the reticulation vertex $u$ is the parent of another reticulation vertex). Finally, networks $\mathcal{N}_1$ and $\mathcal{N}_3$ are also FU-stable, whereas $\mathcal{N}_4$ is not (because the two tree vertices $u$ and $v$ have the same set of children)

characters, where each character can mutate only once and is then preserved), only regular networks are meaningful (for further details, see Willson (2007)).

### *Tree-sibling networks*

A phylogenetic network $\mathcal{N}$ is a *tree-sibling* network (Cardona et al 2008) if every reticulation vertex has at least one sibling that is a tree vertex (cf. Fig. 4). Biologically, for each reticulation event in a tree-sibling network, at least one of the species involved in it also has some descendant through mutation, i.e., through (vertical) descent with modification. Notice, however, that the class of tree-sibling networks generalizes tree-child networks by allowing non-leaf vertices to be parents solely of reticulation vertices. For instance, the single child of a reticulation vertex can itself be a reticulation vertex as shown in $\mathcal{N}_2$ in Fig. 4, where the child of the reticulation vertex $u$ is also a reticulation vertex. Biologically, this scenario can be interpreted as the ancestral hybrid taxon $u$ interbreeding with its sibling at the time to produce hybrid taxon $v$ (more precisely, the ancestor of $v$), without persistence of the ancestral taxon $u$ as a distinct lineage. This might occur, for example, when hybrid taxon $v$ is more fit than $u$, leading $u$ to become extinct. Finally, note that tree-sibling networks that additionally satisfy the time-consistent property, are referred to as *tree-sibling time consistent* (TSTC) networks in the literature (Cardona et al 2008).

### *Reticulation-visible and sink-visible networks*

A phylogenetic network $\mathcal{N}$ is called *reticulation-visible* if every reticulation is visible (Gambette et al 2015; van Iersel et al 2010). Furthermore, a reticulation vertex is called a *sink* if it is the parent of a tree vertex. A *sink-visible network* is a phylogenetic network with the property that every sink is visible. Thus, every reticulation-visible network is also sink-visible, whereas the converse is not true. For an illustration of these concepts, see Fig. 5.

Biologically, the visibility of vertices, in particular reticulation vertices, is relevant for reconstructing phylogenetic networks from genomic data observed at the present.

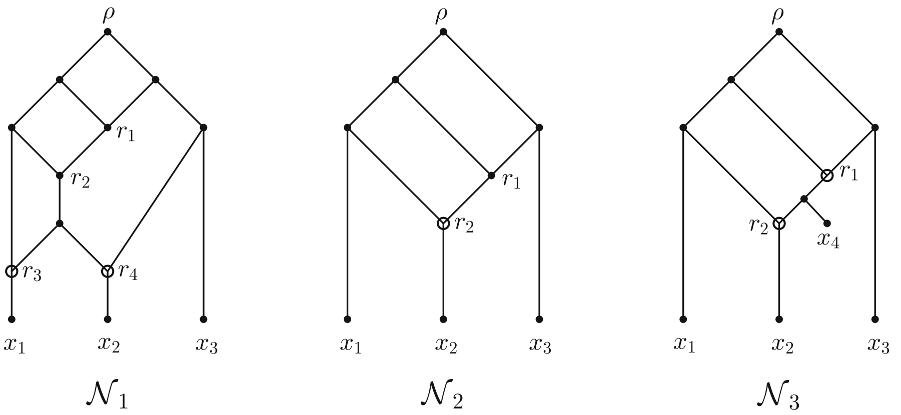


**Fig. 5** Three phylogenetic networks $\mathcal{N}_1, \mathcal{N}_2$, and $\mathcal{N}_3$ whose visible reticulation vertices are marked as unfilled dots. The network $\mathcal{N}_1$ is not sink-visible (since $r_2$ is a sink but is not visible) and thus in particular not reticulation-visible. The network $\mathcal{N}_2$ is sink-visible (reticulation $r_2$ is the only sink of $\mathcal{N}_2$ and is visible) but not reticulation-visible ($r_1$ is not visible). Finally, the network $\mathcal{N}_3$ is reticulation-visible and thus also sink-visible

If a vertex is not visible, there might not be strong evolutionary signal for its presence, since evolutionary information passed from the root to the leaves of the network could have simply bypassed this vertex.

### *Stack-free networks*

A phylogenetic network $\mathcal{N}$ is a *stack-free* network (Semple and Simpson 2018) if $\mathcal{N}$ has no two reticulations, $u$ and $v$ say, such that $u$ is the parent of $v$; that is, there is no reticulation arc in which both end-vertices are reticulations. Such a pair of reticulations are called *stack reticulations*. In contrast, two distinct reticulations are called *sibling reticulations* if they have a common parent. Note that stack-free networks generalize tree-child networks by allowing sibling reticulations. As an example, the networks $\mathcal{N}_1, \mathcal{N}_3$ and $\mathcal{N}_4$ depicted in Fig. 4 are stack-free, whereas the network $\mathcal{N}_2$ in this figure is not since both $u$ and its only child are reticulation vertices.

Biologically, stack-free phylogenetic networks imply that a species resulting from a reticulation event is not directly involved in another reticulation event but rather evolves through vertical descent. Note that stack-free phylogenetic networks are uniquely characterized by the fact that there are two phylogenetic tree embeddings that collectively cover all the arcs of the network (for details see Semple and Simpson (2018)). Stack-free phylogenetic networks can thus also be interpreted as being an 'amalgamation' of two phylogenetic trees (Semple and Simpson 2018).

### *Tree-based networks*

A phylogenetic network $\mathcal{N}$ is called *tree-based* (Francis and Steel 2015) with *base tree* $\mathcal{T}$ if $\mathcal{N}$ can be obtained from $\mathcal{T}$ via the following steps:

(i) Subdivide the arcs of $\mathcal{T}$, i.e., introduce new vertices of in- and out-degree one, so-called *attachment points*.

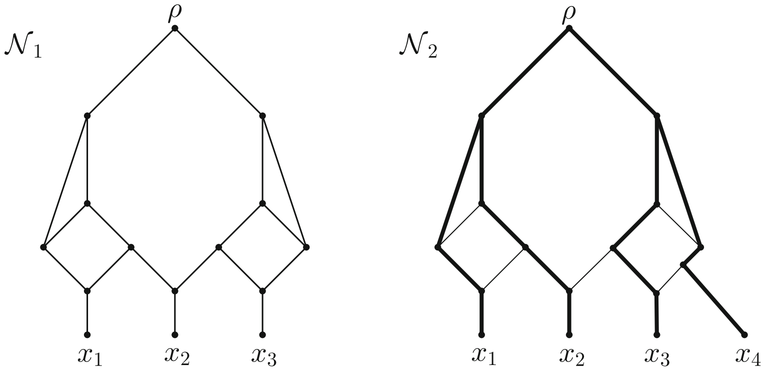


**Fig. 6** Non-tree-based phylogenetic network $\mathcal{N}_1$ and tree-based phylogenetic network $\mathcal{N}_2$. A support tree for $\mathcal{N}_2$ is highlighted in bold

(ii) Add arcs, so-called *linking arcs*, between pairs of attachment points, so that $\mathcal{N}$ remains binary and acyclic.

(iii) Suppress every attachment point that is not incident to a linking arc.

Notice that this procedure allows for parallel edges to be present in a tree-based network (namely, if a linking arc is introduced between two adjacent attachment points). However, these parallel edges can simply be deleted (and the corresponding attachment points suppressed) to be consistent with Definition 1. Alternatively, a phylogenetic network $\mathcal{N}$ on $X$ is tree-based if and only if there exists a rooted spanning tree for $\mathcal{N}$ (that is a rooted tree that contains all vertices and a subset of the arcs of $\mathcal{N}$) whose leaf set is $X$. Such a spanning tree is also called a *support tree* for $\mathcal{N}$. Examples of a tree-based phylogenetic network and a non-tree-based phylogenetic network are shown in Fig. 6.

Moreover, notice that a tree-based network can have different base trees. In fact, there are tree-based networks on $X$, so-called *universal tree-based* networks (Francis and Steel 2015; Hayamizu 2016; Zhang 2016; Bordewich and Semple 2018), that have *every* phylogenetic $X$-tree as a base tree.

Tree-based phylogenetic networks were introduced by Francis and Steel (2015) as a way of quantifying the notion of an 'underlying tree', i.e., as a way to approach the question of whether a phylogenetic network is merely a phylogenetic tree with some additional horizontal edges, or whether a phylogenetic network has little resemblance to a tree and the concept of an underlying tree should be discarded. Tree-based networks are thus relevant to the continuing debate and discussion in the literature on whether evolution is tree-like with occasional non-tree-like events such as horizontal gene transfer (Daubin 2003; Kurland et al 2003), or whether evolution is inherently network-like and has no tree-like similarities at all (Dagan and Martin 2006; Doolittle and Bapteste 2007; Martin 2011; Corel et al 2016).

### *LGT networks and species graphs*

Recall that all arcs directed into a reticulation are referred to as *reticulation arcs*, and are treated symmetrically in the sense that both parents of a reticulation vertex play a symmetrical role. While this is suitable for modelling hybrid speciation and recombination events, it is less suitable for modelling LGT (or introgression) events. In these cases, one of the arcs directed into a reticulation might rather be seen as a 'backbone tree arc' instead of a reticulation arc as in Fig. 1b. This concept is made more precise by the class of *LGT networks* (Cardona et al 2015; Scornavacca et al 2017). While the concept is related to tree-based networks, the approach is specifically designed to model LGT events and to emphasize the asymmetrical role of the parents of a reticulation.

An *LGT network*[3] is a phylogenetic network $\mathcal{N} = (V, E)$ on $X$ along with a bipartition of $E$ in a set of *principal arcs* $E_p$ and a set of *secondary arcs* $E_s$, such that $\mathcal{T}_0 = (V, E_p)$ is a phylogenetic $X$-tree up to suppression of vertices of in-degree one and out-degree one. In this sense, LGT networks are tree-based with a uniquely distinguished base tree. Principal arcs explicitly model the primary (tree-like) line of evolution, and secondary arcs model the LGT events. As an example, for the tree-based network $\mathcal{N}_2$ depicted in Fig. 6, the bold arcs can be seen as corresponding to the primary tree-like evolution of taxa $x_1, \dots, x_4$, whereas the thin arcs can be interpreted as events of LGT.

Note that the distinction of 'principal' and 'secondary' arcs also arises in the context of network inference methods assuming a coalescent process when not only the network topology but also numerical parameters for the inheritance probabilities (p. 5) associated with reticulation arcs are estimated. For example, in PhyloNetworks (Solís-Lemus and Ané 2016; Solís-Lemus et al 2017), the reticulation arc that is assigned the lower probability is referred to as the 'minor hybrid edge', and the other reticulation arc is called the 'major hybrid edge'. It is then also possible to extract the 'major tree', i.e., the tree obtained from deleting the minor hybrid edge at each reticulation (Solís-Lemus and Ané 2016) (see also Jiao et al (2021)).

Another class of networks that explicitly distinguishes between the primary and secondary line of evolution are *species graphs* as defined by Górecki (2004). More restrictive than LGT networks, these are composed of a principal tree and a set of secondary arcs, where the secondary arcs must satisfy a set of more restrictive conditions than in LGT networks. For instance, the resulting network must be time consistent, and the head and source of a secondary arc cannot be connected by a path in the principal tree. We refer the reader to Górecki (2004) for further details.

### *Orchard or cherry-picking networks*

Let $\mathcal{N}$ be a phylogenetic network on $X$, and let $x, y \in X$ be two distinct leaves of $\mathcal{N}$. Let $p_x$ and $p_y$ denote the parents of $x$ and $y$, respectively. If $p_x = p_y$, the pair $\{x, y\}$ is called a *cherry* of $\mathcal{N}$. Furthermore, if one of the parents, say $p_y$, is a reticulation and there is an edge $(p_x, p_y)$ in $\mathcal{N}$, then $\{x, y\}$ is called a *reticulated cherry* of $\mathcal{N}$ with *reticulation leaf* $y$. As an example, consider Fig. 7. Here, the phylogenetic network $\mathcal{N}$ in panel (i) contains a reticulated cherry, namely $\{x_1, x_2\}$, and $x_2$ is the reticulation leaf. Moreover, the phylogenetic network in panel (iii) contains a cherry, namely the pair $\{x_2, x_3\}$.

[3] We are using the definition from Scornavacca et al (2017) for binary networks, which is slightly different from the definition originally considered by Cardona et al (2015) for more general networks.

Now, there are two *cherry reductions* (Erdős et al 2019) (see also Janssen and Murakami (2021)) associated with cherries and reticulated cherries:

- If $\{x, y\}$ is a cherry of $\mathcal{N}$, *reducing* $y$ is the operation of deleting $y$ and suppressing the resulting vertex of in-degree one and out-degree one. If the parent of $x$ and $y$ is the root of $\mathcal{N}$, then reducing $y$ consists of deleting $y$ as well as the root of $\mathcal{N}$, resulting in an isolated vertex $x$.
- If $\{x, y\}$ is a reticulated cherry of $\mathcal{N}$ in which $y$ is the reticulation leaf, *cutting* $\{x, y\}$ is the operation of deleting the reticulation edge $(p_x, p_y)$, and suppressing the resulting two vertices of in-degree one and out-degree one.

Now, if a phylogenetic network $\mathcal{N}$ can be reduced to a single vertex by a sequence of cherry reductions, it is called an *orchard* network (Erdős et al 2019) or a *cherry-picking* network (Janssen and Murakami 2021) (note that Erdős et al (2019) and Janssen and Murakami (2021) independently introduced this class of networks). An example of an orchard network and a possible complete sequence of cherry reductions[4] that reduces it are depicted in Fig. 7.

While orchard networks were originally introduced without any biological justification, van Iersel et al (2021) recently showed that they are characterized by admitting a *HGT-consistent labelling*. Intuitively, this means that orchard networks "are consistent with an evolutionary history in time in which reticulate events represent instantaneous (horizontal) transfers such as LGT events" (van Iersel et al 2021). In this sense, orchard networks can be seen as trees with additional *horizontal* arcs, making them biologically highly relevant. Note that orchard networks are closely related to tree-based networks, except that the additional arcs in tree-based networks do not need to be horizontal. In particular, every orchard network is tree-based (e.g., Huber et al (2019)). They are also closely related to LGT networks, with the difference that orchard networks do not (necessarily) specify which arcs are secondary arcs, i.e., LGT arcs (van Iersel et al 2021).

Orchard networks also exhibit certain desirable mathematical properties. As an example, they are uniquely characterized by their induced subnetworks on three leaves and can also be encoded by certain sets of paths in the network (for details and formal definitions, see Sect. 4.2.2).

### *Galled trees, galled networks, and level-$k$ networks*

Let $\mathcal{N}$ be a phylogenetic network. A *reticulation cycle* of $\mathcal{N}$ is a pair of directed paths with a common start vertex and common end vertex that are vertex-disjoint otherwise (note that the common start vertex is thus necessarily a tree vertex and the common end vertex is necessarily a reticulation). Note that every reticulation of $\mathcal{N}$ lies on at least one reticulation cycle, and this vertex could either be the end vertex of the two paths defining that cycle or an intermediate vertex on one of them. In fact, if a network contains several reticulation cycles, they might be dispersed or they might be more or less interwoven.

[4] Note that there might be several complete sequences of cherry reductions that reduce an orchard network to a single leaf and the order in which cherry reductions are performed does not matter (Erdős et al 2019, Proposition 4.1).

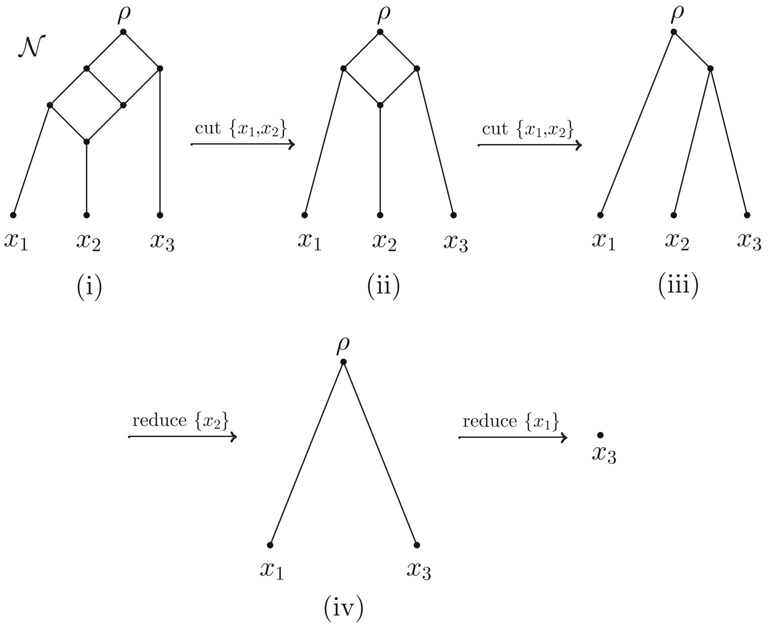


**Fig. 7** Orchard network $\mathcal{N}$ on $X = \{x_1, x_2, x_3\}$ and a complete sequence of cherry reductions that reduces it to a single vertex

Now, if the reticulations cycles of a network $\mathcal{N}$ are all vertex-disjoint (i.e., the reticulation cycles are isolated from each other), $\mathcal{N}$ is said to be a *galled tree* (Gusfield et al 2003)[5]. In this case, no two reticulation vertices can be contained in a common reticulation cycle, and reticulation events can be seen as independent from each other. An example of a galled tree is the phylogenetic network $\mathcal{N}_1$ depicted in Fig. 8. Note that if a galled tree, or more generally a phylogenetic network, only contains one reticulation cycle, it is also called a *unicyclic* network.

A generalization of galled trees are *galled networks* (Huson and Klöpper 2007), where reticulation cycles may share tree vertices (and thus they may overlap on tree arcs). The previous assumption for galled trees is thus relaxed in the sense that the only constraint is that every reticulation vertex is contained in a reticulation cycle formed by paths composed exclusively by tree vertices. As an example, the phylogenetic network $\mathcal{N}_2$ depicted in Fig. 8 is a galled network but not a galled tree.

A slightly more general notion quantifying whether reticulations are widely separated or highly interwoven is the 'level' of a network (Choy et al 2005). To define this, let a *biconnected component* or *block* of a phylogenetic network $\mathcal{N}$ be a subnetwork $\mathcal{N}'$ of $\mathcal{N}$ such that (a) $\mathcal{N}'$ remains connected if any one of its vertices (together with its incident arcs) is deleted; and (b) $\mathcal{N}'$ is maximal with respect to property (a). A phylogenetic network $\mathcal{N}$ is said to be a *level-k network* (Choy et al 2005) if every biconnected component of $\mathcal{N}$ has at most *k* reticulation vertices. Note that a level-0

[5] Note that the property of vertex-disjoint reticulation cycles was first discussed by Wang et al (2001), and later called the "gall property" by Gusfield et al (2003).

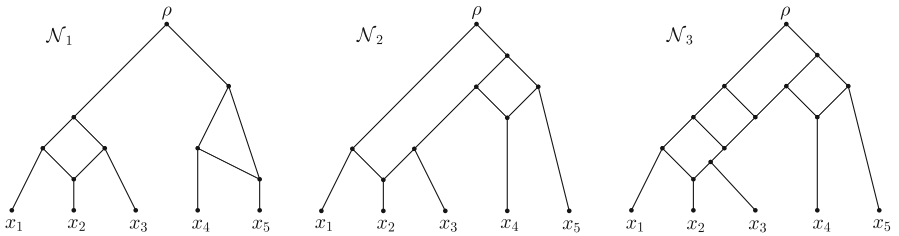


**Fig. 8** Galled tree $\mathcal{N}_1$, galled network $\mathcal{N}_2$ (note that the reticulation cycles of $\mathcal{N}_2$ share tree vertices but no reticulation vertices), and a phylogenetic network $\mathcal{N}_3$ that is neither a galled tree nor a galled network. Note that $\mathcal{N}_1$ is a level-1 network, $\mathcal{N}_2$ is a level-2 network, and $\mathcal{N}_3$ is a level-4 network

network is a phylogenetic tree, and a level-1 network is a galled tree as defined above. An illustration of this concept is given in Fig. 8, where $\mathcal{N}_1$ is a level-1 network, $\mathcal{N}_2$ is a level-2 network, and $\mathcal{N}_3$ is a level-4 network.

Biologically, the level of a network and related concepts provide another way (next to tree-basedness) to assess whether a network can be viewed as mainly tree-like with some local and dispersed reticulations, or whether it is highly tangled and interwoven with little resemblance to a tree-like structure.

Mathematically, restricting the level of a network often leads to more tractable problems. For example, most studies on the statistical identifiability of phylogenetic networks to date have focused on level-1 networks, and similarly many positive results concerning the combinatorial identifiability of phylogenetic networks are restricted to networks of a small level (see Sect. 4.2, where we discuss identifiability questions in more detail). In addition, restricting the level of a network is currently a common way to obtain scalable network inference methods. As an example, both PhyloNetworks (Solís-Lemus and Ané 2016; Solís-Lemus et al 2017) and NANUQ (Allman et al 2019) consider only level-1 networks (note that we discuss this in more detail in Sect. 4.1).

### 3.2 Classes of rooted binary phylogenetic networks not yet linked to biology

In the following, we present additional classes of rooted binary phylogenetic networks that have not been linked to a particular biological process yet and are mostly of mathematical or algorithmic relevance to date.

***FU-stable networks***

A phylogenetic network $\mathcal{N}$ is called an *FU-stable network*[6] (Huber et al 2016) if it is (i) stack-free and (ii) has the property that any two distinct tree vertices of $\mathcal{N}$ have distinct sets of children (Huber et al 2016, Theorem 1). As an example, while the networks $\mathcal{N}_1$, $\mathcal{N}_3$, and $\mathcal{N}_4$ depicted in Fig. 4 are all stack-free, only $\mathcal{N}_1$ and $\mathcal{N}_3$ are FU-stable. Network $\mathcal{N}_4$ is not FU-stable because the two distinct tree vertices $u$ and $v$ have identical sets of children.

[6] Note that Huber et al (2016) originally simply called such networks *stable*. However, as the word “stable” is used in various contexts in the phylogenetic networks literature, we refer to FU-stable networks to avoid any ambiguity.

Note that the class of FU-stable networks arises in the context of 'folding and unfolding' phylogenetic networks that we briefly mentioned in Sect. 2. More precisely, Huber et al (2016) call a phylogenetic network $\mathcal{N}$ FU-stable if $F(U(\mathcal{N}))$ is isomorphic to $\mathcal{N}$, where – informally speaking – $U(\mathcal{N})$ is the MUL-tree obtained from 'unfolding' $\mathcal{N}$ and $F(U(\mathcal{N}))$ is the phylogenetic network obtained from 'folding up' $U(\mathcal{N})$ (for a more rigorous definition and additional details, see Huber et al (2016)). In particular, the definition given here is rather a characterization of the class of FU-stable networks, whereas the original definition of Huber et al (2016) is motivated by the folding–unfolding process applied to a phylogenetic network.

In addition, note that the process of 'unfolding' a phylogenetic network $\mathcal{N}$ into a MUL-tree is used in PhyloNet (Than et al 2008; Wen et al 2018), a popular network inference tool, as one way of calculating the probability of a gene tree given a species network (for details, see Yu et al (2012, 2014)).

### *k-nested networks*

In order to define the class of $k$-nested networks introduced by Jansson and Sung (2004), we need some additional terminology (adapted from Jansson and Sung (2004)). First, let $\mathcal{N} = (V, E)$ be a phylogenetic network and let $r$ be a reticulation vertex. Then, every tree vertex $t$ that is an ancestor of $r$ such that $r$ can be reached using two disjoint directed paths starting at the children of $t$ is called a *split node of* $r$. If $t$ is a split node for $r$, any directed path from $t$ to $r$ is a *merge path of* $r$, and any path from a child of $t$ to a parent of $r$ is a *clipped merge path of* $r$. Jansson and Sung (2004) call a phylogenetic network *nested* if for every two merge paths $P_1$, $P_2$ of two distinct reticulation vertices $r_1$ and $r_2$, $P_1$ and $P_2$ share a common arc if and only if one of these paths is a subpath of the other. Moreover, for a nested phylogenetic network $\mathcal{N} = (V, E)$ and a vertex $v \in V$, the *nesting depth of* $v$, denoted by $d(v)$, is the number of reticulation vertices in $\mathcal{N}$ that have a clipped merge path traversing $v$. Finally, the *nesting depth of* $\mathcal{N}$, denoted by $d(\mathcal{N})$, is the maximum value of $d(v)$ over all $v \in V$, and $\mathcal{N}$ is called *k-nested* if it is nested with nesting depth at most $k$. Note that $d(\mathcal{N}) \leq 1$ if and only if $\mathcal{N}$ is a galled tree. Moreover, note that if $\mathcal{N}$ is a nested phylogenetic network with nesting depth $d$, then we have for the level of $\mathcal{N}$, say $k$, that $k \geq d$, i.e., the nesting depth is a lower bound for the level of $\mathcal{N}$ (Jansson and Sung 2004).

As an example, the phylogenetic network $\mathcal{N}_1$ depicted in Fig. 9 is nested with nesting depth 3, whereas the phylogenetic network $\mathcal{N}_2$ depicted in the same figure is not nested. In case of $\mathcal{N}_1$, $\rho$ is the unique split node for $r_3$, $t_2$ is the unique split node for $r_2$, and $t_4$ is the unique split node for $r_1$. The collection of merge paths for $\mathcal{N}_1$ is given by $\{(\rho, t_1, r_3), (\rho, t_2, t_3, r_2, r_3), (\rho, t_2, t_4, t_5, r_1, r_2, r_3), (\rho, t_2, t_4, t_6, r_1, r_2, r_3), (t_2, t_3, r_2), (t_2, t_4, t_5, r_1, r_2), (t_2, t_4, t_6, r_1, r_2), (t_4, t_5, r_1), (t_4, t_6, r_1)\}$ and it can be easily checked that any two merge paths $P_1$, $P_2$ for two distinct reticulation vertices share a common arc if and only if one of the paths is a subpath of the other. Moreover, to see that the nesting depth of $\mathcal{N}_1$ is equal to 3, notice that vertices $t_5$ and $t_6$ have a nesting depth of $d(t_5) = d(t_6) = 3$ (and as $\mathcal{N}_1$ contains precisely three reticulations this is maximal).

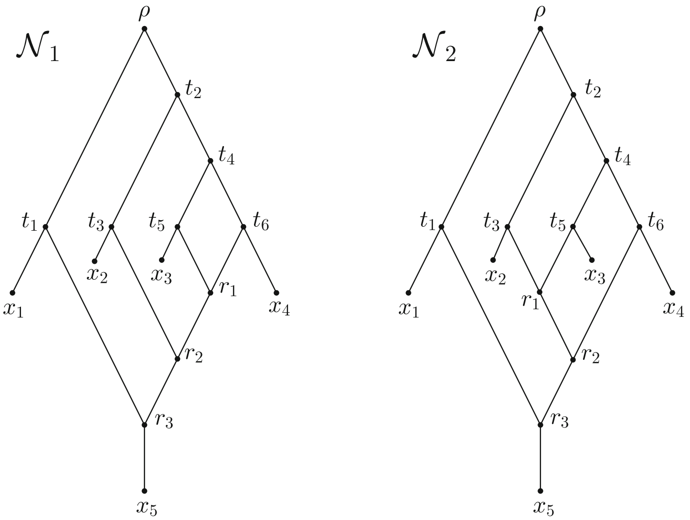


**Fig. 9** Nested phylogenetic network $\mathcal{N}_1$ with nesting depth 3, and phylogenetic network $\mathcal{N}_2$ that is not nested. To see that $\mathcal{N}_2$ is not nested, notice that $t_2$ is a split node for $r_2$ and $\rho$ is a split node for $r_3$. However, the two merge paths $P = (t_2, t_4, t_5, r_1, r_2)$ and $P' = (\rho, t_2, t_4, t_6, r_2, r_3)$ share an arc, namely the arc $(t_2, t_4)$, but $P$ is not a subpath of $P'$ and vice versa

### *k-reticulated networks*

A phylogenetic network $\mathcal{N}$ is called *k-reticulated* (Vu et al 2013) if for any vertex $v$ of in-degree at most one (i.e., for any tree vertex or the root of $\mathcal{N}$), there are at most $k$ reticulation vertices that can be reached from $v$ by at least two directed vertex-disjoint paths from $v$ (i.e., paths whose only shared vertices are the start vertex $v$ and the end vertex). Note that every level-$k$ network is also a $k$-reticulated network, but some level-$k'$ networks are in fact $k$-reticulated networks with $k < k'$ (Vu et al 2013). As an example, the network $\mathcal{N}_1$ depicted in Fig. 8 is a level-1 network and is 1-reticulated. However, the network $\mathcal{N}_2$ depicted in this figure is a level-2 network that is 1-reticulated.

### *Spread-k networks*

The class of *spread-k* networks was introduced by Asano et al (2010) and was motivated by the aim of developing an efficient representation of the clusters of a phylogenetic networks to make the computation of the Robinson-Foulds distance (Robinson and Foulds (1981); see also Cardona et al (2009b)) between networks more efficient. Recall that given a phylogenetic network $\mathcal{N} = (V, E)$ on $X$ and a vertex $v \in V$, the cluster $c_{\mathcal{N}}(v)$ associated with $v$ is the subset of the leaf set $X$ that can be reached from $v$ via a directed path, i.e., $c_{\mathcal{N}}(v) = \{x \in X : v \preceq_{\mathcal{N}} x\}$. Now, following Asano et al (2010), a *leaf numbering function* is a bijection from the leaf set $X$ to the set $\{1, \ldots, n\}$. Moreover, for any leaf numbering function $f$ and a vertex $v \in V$, the *characteristic vector for $v$ under $f$*, denoted $C_f[v]$, is a bit vector (i.e., a vector containing zeros and ones) of length $n$ such that for any $i \in \{1, \ldots, n\}$, the $i^{th}$

bit equals 1 if and only if $f^{-1}(i)$ is contained in the cluster $c_{\mathcal{N}}(v)$ associated with $v$. Note that $C_f[\rho] = 11\ldots11$ for the root $\rho$ of $\mathcal{N}$ and that $C_f[x]$ contains precisely one 1 for each leaf $x \in X$. Furthermore, a maximal consecutive sequence of ones in a bit vector is called an *interval*. Now, given a leaf numbering function $f$ and a vertex $v \in V$, let $I_f(v)$ denote the number of intervals in $C_f[v]$. Then, the *spread of* $f$ is defined as $I_f = \max_{v\in V} I_f(v)$ and the *minimum spread of* $\mathcal{N}$ is the minimum value of $I_f$ taken over all possible leaf numbering functions $f$. Finally, a phylogenetic network $\mathcal{N}$ is a spread-$k$ network if its minimum spread is at most $k$. Note that a level-$k$ network has minimum spread at most $k + 1$ (Asano et al 2010).

As an example, consider the phylogenetic network $\mathcal{N}_1$ depicted in Fig. 9. The minimum spread of $\mathcal{N}_1$ is 2, and is, for example, achieved by the leaf numbering function $f : \{x_1, \ldots, x_5\} \rightarrow \{1, \ldots, 5\}$ with $f(x_i) = i$ for $i = 1, \ldots, 5$. Here, $C_f[t_1] = (1, 0, 0, 0, 1)$, $C_f[t_2] = (0, 1, 1, 1, 1)$, $C_f[t_3] = (0, 1, 0, 0, 1)$, $C_f[t_4] = (0, 0, 1, 1, 1)$, $C_f[t_5] = (0, 0, 1, 0, 1)$, $C_f[t_6] = (0, 0, 0, 1, 1)$, and $C_f[r_1] = C_f[r_2] = C_f[r_3] = (0, 0, 0, 0, 1)$ (as well as $C_f[\rho] = (1, 1, 1, 1, 1)$, $C_f[x_1] = (1, 0, 0, 0, 0)$, $C_f[x_2] = (0, 1, 0, 0, 0)$, $C_f[x_3] = (0, 0, 1, 0, 0)$, $C_f[x_4] = (0, 0, 0, 1, 0)$, and $C_f[x_5] = (0, 0, 0, 0, 1)$). In particular, $I_f[t_1] = I_f[t_3] = I_f[t_5] = 2$ (and $I_f[v] = 1$ for all other vertices $v$), which leads to $I_f = 2$ (as it can easily be checked that there exists no leaf numbering, say $f'$, with $I_{f'} < 2$).

### *Nearly stable, genetically stable, nearly tree-child, and stable-child networks*

A phylogenetic network $\mathcal{N}$ is *nearly stable* (Gambette et al 2015) if for every vertex, either the vertex or its parents are visible. In other words, for each arc $(u, v)$ of $\mathcal{N}$, either $u$ or $v$ (or both) are visible.

$\mathcal{N}$ is called *genetically stable* (Gambette et al 2016) if every reticulation vertex is visible and has at least one visible parent. This means that a reticulation vertex ‘inherits’ the visibility property from one of its parents.

Moreover, $\mathcal{N}$ is called *nearly tree-child* (Gambette et al 2016) if the following two conditions hold: (i) $\mathcal{N}$ is reticulation-visible and (ii) every reticulation of $\mathcal{N}$ has the tree-path property (as introduced on p. 10).

Finally, *stable-child* networks (Gunawan and Zhang 2015) are phylogenetic networks in which every vertex has at least one visible child.

Examples for these types of networks are depicted in Fig. 10.

### *Valid networks*

The class of *valid* networks was very recently introduced by Murakami et al (2019) and requires some additional terminology (all subsequent definitions are adapted from Murakami et al (2019)).

First, given a directed acyclic graph $G = (V, E)$ (possibly containing some labeled vertices) *cleaning up* $G$ is the act of applying the following operations until none is applicable:

1. delete an unlabeled vertex of out-degree zero;
2. suppress a vertex of in-degree one and out-degree one;
3. replace a pair of parallel arcs by a single arc.

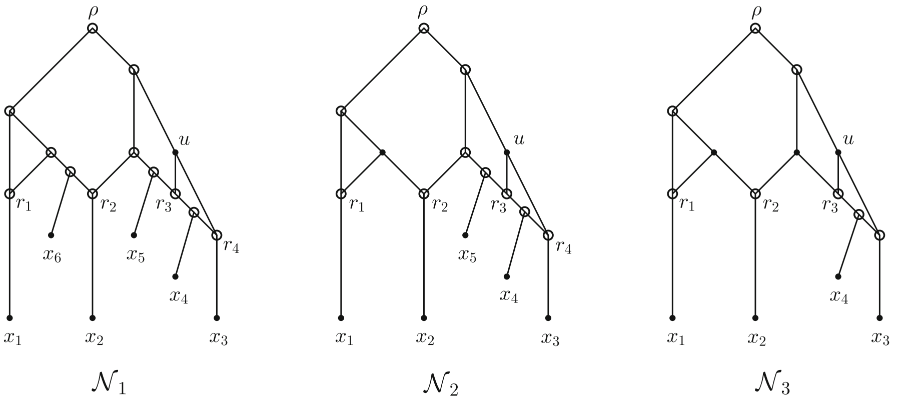


**Fig. 10** Three reticulation-visible but non-tree-child phylogenetic networks $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_3$ for which the visible vertices (except for leaves) are depicted as unfilled dots. The network $\mathcal{N}_1$ is nearly stable, genetically stable, nearly tree-child, and stable-child. The network $\mathcal{N}_2$ is also nearly stable, genetically stable, and stable-child, but not nearly tree-child (because reticulation $r_1$ does not have a parent that is connected to a leaf of $\mathcal{N}_2$ by a tree-path). Finally, the network $\mathcal{N}_3$ is nearly stable, but neither genetically stable (as the reticulation $r_2$ does not have a visible parent), nor nearly tree-child (as reticulations $r_1$ and $r_2$ do not have a parent that is connected to a leaf of $\mathcal{N}_3$ by a tree-path), nor stable-child (since the parent of $u$ does not have a visible child)

Second, given a phylogenetic network $\mathcal{N}$, the deletion of a reticulation arc is *valid* if the resulting subnetwork, after cleaning up, contains exactly two vertices and three arcs fewer than the original network, that is, only the reticulation arc is deleted and its endpoints are suppressed. A reticulation arc is called *valid* if its deletion is valid in this sense; otherwise it is called *invalid*.

Based on this, a phylogenetic network $\mathcal{N}$ is called a *valid* network (Murakami et al 2019) if all its reticulation arcs are valid. An illustration of this concept can be found in Fig. 11.

### 3.3 Relationship among network classes within rooted binary phylogenetic networks

While we have introduced more than 20 different classes of rooted binary phylogenetic networks in the two preceding sections, some of them are closely related in the sense that one class is a proper subset of another class. We summarize these inclusion relationships among network classes within rooted binary phylogenetic networks in Table 1 and additionally visualize them in Fig. 12. An excellent overview of some of these inclusion-relationships among phylogenetic network classes can also be found on the website “ISIPhyNC” (https://phylnet.univ-mlv.fr/isiphync/index.php) (Gambette et al 2018b) mentioned in the introduction.

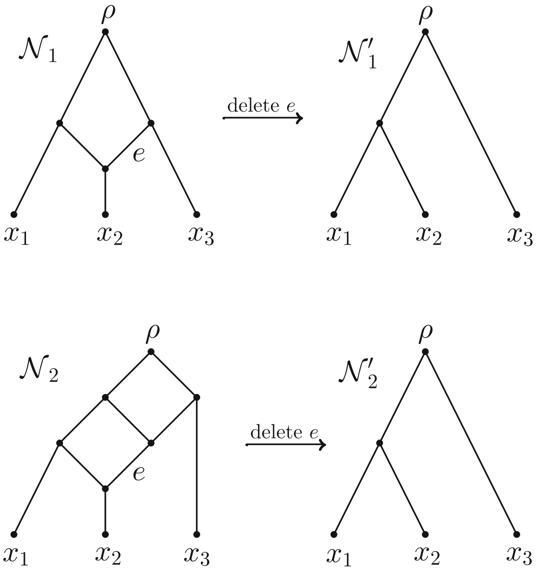


**Fig. 11** Valid phylogenetic network $\mathcal{N}_1$ and non-valid phylogenetic network $\mathcal{N}_2$. To see that $\mathcal{N}_1$ is valid note that the deletion of the reticulation arc *e* is valid since the resulting network $\mathcal{N}_1$ contains precisely two vertices and three fewer arcs than $\mathcal{N}_1$ after cleaning up (by symmetry the second reticulation arc of $\mathcal{N}_1$ is also valid). The network $\mathcal{N}_2$ on the other hand is not valid since the deletion of the reticulation arc *e* is not valid. More precisely, the deletion of *e* results in a network $\mathcal{N}_2'$ with four fewer vertices and six fewer arcs than the original network (Figure adapted from (Murakami et al 2019, Fig. 3))

# 4 Mathematical and computational challenges in estimating phylogenetic networks

In the previous section, we reviewed several classes of rooted binary phylogenetic networks used in the literature. We saw that some of them have direct biological relevance, whereas others lack an immediate biological interpretation. In the following, we discuss two further aspects that need to be taken into account when estimating phylogenetic networks in an empirical setting, namely scalability and identifiability. In particular, we describe how imposing structural constraints on the networks under consideration, i.e., considering specific subclasses instead of arbitrary phylogenetic networks, can help to address challenges related to scalability and identifiability in network inference and estimation. We remark that both the development of scalable methods in network inference and the analysis of network identifiability are currently very active areas of research in phylogenetics. Our aim is thus to convey and summarize the central ideas and results rather than to recapitulate the technical details.

## 4.1 Scalability

In this section, we consider the task of estimating a phylogenetic network given data for a collection of taxa. This leads to two distinct challenges related to scalability. First, we must evaluate the fit of a specified network to a given data set under a chosen model or optimality criterion. Second, we must search the space of possible networks for those that are optimal under the selected model or criterion. While these two challenges are the same as those faced in the inference of phylogenetic trees, they are even more daunting for networks because infinitely many networks are possible for a

**Table 1** Inclusion relationships among network classes within rooted binary phylogenetic networks. Note that we omit trivial inclusions such as "level-$k$ is level-($k+1$)" for the classes of level-$k$, spread-$k$, $k$-reticulated, and $k$-nested networks

| # | Inclusion | Reference |
|---|---|---|
| 1 | Genetically stable is tree-sibling | Gambette et al (2018a, Prop. 3.2(1)) |
| 2 | Tree-child is nearly tree-child | definition |
| 3 | Reticulation-visible is tree-based | Gambette et al (2015, Lemma 1) |
| 4 | Nearly tree-child is genetically stable | definition |
| 5 | Normal is tree-child | definition |
| 6 | Compressed[7]normal is regular | Willson (2009, Theorem 3 .4) |
| 7 | Tree-child is tree-sibling | Cardona et al (2008, p. 2) (see also Steel (2016, Lemma 10.8)) |
| 8 | Tree-child is reticulation-visible | definition |
| 9 | Tree-sibling is tree-based | Francis and Steel (2015, Corollary 2) |
| 10 | A galled network is reticulation-visible | Huson et al (2011, Lemma 6.11.14) |
| 11 | A galled tree is a galled network | definition |
| 12 | Galled trees are tree-sibling | Huson et al (2011, Lemma 6.11.16) |
| 13 | Galled trees are tree-child | Huson et al (2011, Lemma 6.11.11) |
| 14 | Genetically stable is reticulation-visible | definition |
| 15 | Tree-child is nearly stable | definition |
| 16 | Tree child is stable-child | definition |
| 17 | LGT is tree-based | definition |
| 18 | Species graph is LGT | Pons (2016, Section 2.3) |
| 19 | Tree-child is orchard | Janssen and Murakami (2021, p. 138) |
| 20 | Tree-child is valid | Murakami et al (2019, Lemma 2) |
| 21 | Orchard is tree-based | Huber et al (2019, Prop. 2) |
| 22 | Stack-free is tree-based | Semple and Simpson (2018); Zhang (2016) |
| 23 | Valid is stack-free | definition |
| 24 | Reticulation-visible is stack-free | Semple and Simpson (2018, Theorem 2.1) |
| 25 | FU-stable is stack-free | definition |
| 26 | Level-$k$ is spread-($k+1$) | Asano et al (2010, Lemma 5) |
| 27 | Level-$k$ is $k$-reticulated | Vu et al (2013) |
| 28 | Genetically stable is FU-stable | Huber et al (2016, Corollary 1) (using inclusions 1, 14, 24) |

[7] Compressed means that every arc $(u, v)$ leading from a vertex $u$ of out-degree one to a vertex $v$ of in-degree one is contracted

fixed number of taxa unless constraints on possible network topologies are imposed. In the following sections, we discuss each of these challenges in greater detail.

### 4.1.1 Measuring fit to a fixed network

The first challenge in estimating a phylogenetic network lies in developing methods for determining whether a putative phylogenetic network provides a good fit to the

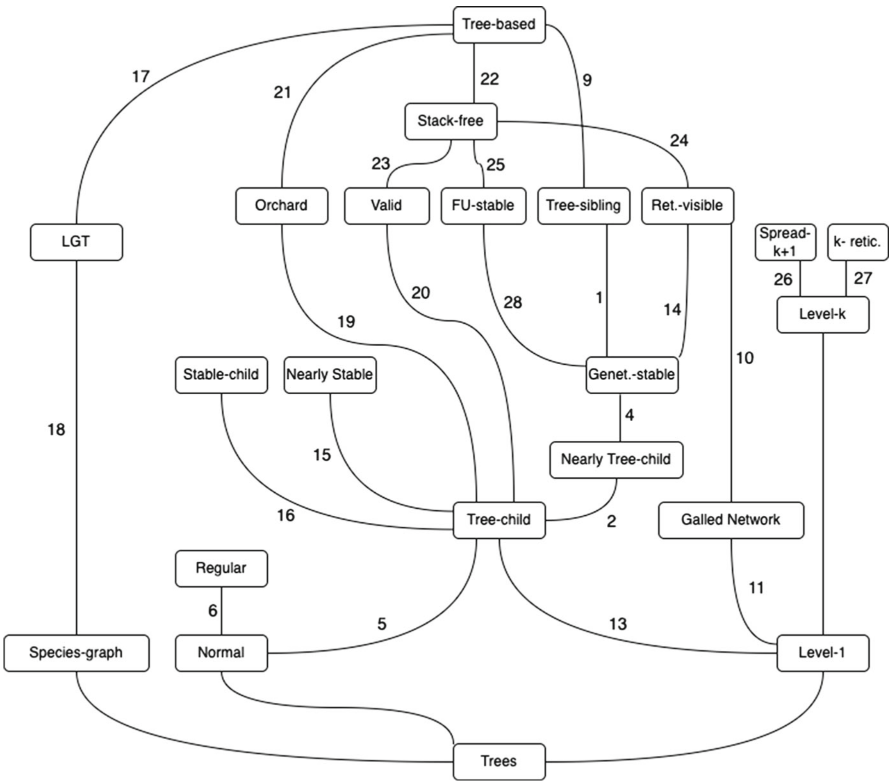


**Fig. 12** Diagram illustrating the inclusion relationships among network classes within rooted binary networks as given in Table 1

data at hand. In order to make this assessment, it is common to define an optimality criterion that quantifies fit of the network to the data, perhaps under the assumption of a specific evolutionary model. Below, we review the common criteria employed for this purpose, indicating the data types to which they are applied, the evolutionary models assumed, and the software implementations in which they are included.

Before doing so, we note that many of the current methods for network inference assume that data from multiple genes sampled throughout the genome (commonly referred to as *multilocus data*) are available. In this case, it is necessary to adopt a model for the relationship between the evolutionary histories of the genes and that of the species from which the genes were sampled. The most common model for this purpose is the coalescent model (see, e.g., Kingman (1982) for a description of the coalescent process for trees, and Meng and Kubatko (2009); Yu et al (2012); Degnan (2018) for the extension to phylogenetic networks). In our descriptions below, we note whether multilocus data are assumed and if so, whether the coalescent model is used for inference.

### *Parsimony*

Nakhleh et al (2005) proposed a method for computing the parsimony score of a phylogenetic network that is formed by adding reticulation events to an existing phylogenetic tree. Their method assumes a single input alignment (and hence a single underlying network), but divides the input data into "blocks" in order to evaluate the parsimony score. While Nakhleh et al (2005) showed that a polynomial time algorithm for computing the maximum parsimony score can be found when the number of reticulation arcs to be added is fixed, it was subsequently shown that the problem is NP-hard in general (Jin et al 2009). Note that further extensions of parsimony to phylogenetic networks and the computational complexity of computing the parsimony score have been studied by various authors in recent years (e.g., Kannan and Wheeler (2012); Fischer et al (2015); Bryant et al (2017); van Iersel et al (2017a)).

### *Minimize deep coalescences*

The mimimize deep coalescences (MDC) method (Yu et al 2013) was one of the first to be proposed for inference of species-level phylogenetic networks from multilocus data under the coalescent. The method takes as input a collection of gene tree topologies and then defines the score for a putative network as the number of deep coalescent events required to explain the observed gene trees (see Yu et al (2013) for details). While the ideas underlying the method are straightforward, it requires that gene trees are first estimated from the alignments for each of the genes. The method is implemented in the PhyloNet software (Than et al 2008; Wen et al 2018). A polynomial-time algorithm for computing the MDC score on level-1 networks has recently been derived (Lemay et al 2021).

### *Likelihood methods*

The maximum likelihood framework is the foundation upon which a wide range of statistical methodology is built. Thus, numerous methods for the inference of phylogenetic networks under the maximum likelihood criterion have been proposed. We provide an overview of several current methods categorized by the data type assumed.

- *Single-locus methods* Lutteropp et al (2021) recently proposed a method, called NetRAX, to infer networks using maximum likelihood under a single-locus model (i.e., a model that assumes that a single network underlies the entire data set) using standard substitution models.
- *Coalescent methods for multilocus data using gene trees as input* Yu et al (2014) proposed a method for computing the likelihood of a specified phylogenetic network given data consisting of gene trees estimated for a collection of loci under the coalescent model (InferNetwork_ML in the PhyloNet package). They discuss the computational complexity associated with computation of the likelihood as well as provide methods for assessing confidence. Using the same model as in Yu et al (2014), Kubatko (2009) used AIC, AICc, and BIC, measures that are based on the likelihood but employ a "penalty" for the addition of parameters to a model, to compare phylogenetic networks with varying numbers of reticulation edges.

***Pseudolikelihood methods***

Due to the complexity of computing the likelihood of a phylogenetic network directly, pseudolikelihood methods have been suggested as an alternative. As in the case of likelihood methods, several methods have been proposed that vary in the type of data they take as input.

- *Coalescent methods for multilocus data using gene trees as input* The PhyloNet package mentioned above includes the InferNetwork_MPL method (Yu and Nakhleh 2015) to estimate phylogenetic networks given an input set of gene trees estimated from multilocus data. The pseudolikelihood calculation uses relationships among rooted triples found in the input gene tree topologies. The other popular method in this category is SNaQ, implemented in the PhyloNetworks package (Solís-Lemus and Ané 2016; Solís-Lemus et al 2017), which computes the pseudolikelihood based on quartet relationships obtained from the input gene trees.
- *Coalescent methods for unlinked biallelic markers* Zhu and Nakhleh (2018) proposed a method that uses pseudolikelihood as a criterion to compare phylogenetic networks using sequence data directly when these data consist of a collection of unlinked biallelic markers, such as single nucleotide polymorphisms (SNPs). The method is implemented in the MLE_BiMarkers module in PhyloNet. This method addresses the fact that using gene trees as input fails to account for variability associated with the initial pre-processing of the data to estimate the gene trees. Because it limits the input data to unlinked biallelic markers, the method is computationally tractable.

***Bayesian methods***

In general, Bayesian methods utilize the likelihood function in computing posterior probabilities of parameters of interest and seek estimators that maximize the posterior probability. In the context of phylogenetic networks, Bayesian methods have an advantage over methods based solely on the likelihood function because in the Bayesian framework (and in particular, through the use of Markov chain Monte Carlo (MCMC) methods to carry out inference; see below), the state space can be augmented to include gene trees for each locus. This allows inference of species-level phylogenetic networks directly from multilocus data without the need to first estimate gene trees for each locus. Methods in this class include SpeciesNetwork in the BEAST2 package (Zhang et al 2017), the MSCi model in BPP (Flouri et al 2019), and MCMC_Seq (Wen and Nakhleh 2017) in PhyloNet.

### 4.1.2 Finding optimal phylogenetic networks

Having described several criteria by which phylogenetic networks can be compared in terms of their fit to empirical data, we now turn our attention to the problem of finding optimal phylogenetic networks under a specified criterion. Below, we discuss several of the associated challenges, including restricting the class of networks considered and development and implementation of methods for efficiently searching the space of networks for those that are optimal.

### *Restricting network space*

There are two types of restrictions on network space that must be considered when the goal is to find an optimal phylogenetic network. We describe each below.

- *Biologically-motivated restrictions* Note that in the context of empirical data, reticulation vertices and arcs are assumed to represent evolutionary events, such as hybrid speciation, gene flow or introgression, LGT, or gene duplication and loss. Thus, the timing and direction of the arcs in a phylogenetic network inferred from empirical data should satisfy the property of time consistency (see p. 9) to ensure that horizontal events occur between lineages that exist contemporaneously. Note, however, that the requirement of time consistency depends on the measurement scale. For example, branch lengths measured in coalescent units (number of generations scaled by twice the effective population size) for a network that satisfies the constraint of time consistency will not necessarily lead to a time consistent network in calendar time, because, for example, generation times among lineages might vary substantially. The biological requirement of time consistency leads to a significant scalability challenge in that parameters must be estimated over a constrained parameter space.
- *Computationally-motivated restrictions* For many of the criteria discussed in the previous section (e.g., parsimony and likelihood), addition of horizontal arcs will always improve the value of the objective function. Intuitively, this occurs because each additional arc can explain some portion of the variability in the observed data, thus providing an improvement in the ‘fit’ of the network to the data. For this reason, algorithms for network inference commonly impose limitations on either the number of horizontal arcs or the class of networks over which to search. For example, PhyloNet (Yu et al 2014) and PhyloNetworks (Solís-Lemus and Ané 2016; Solís-Lemus et al 2017) require the user to specify the number of horizontal arcs (or more precisely, the ‘maximum number of hybridizations’) prior to the search of network space. By running the methods separately with increasing numbers of horizontal arcs, information criteria (i.e., AIC, BIC (Kubatko 2009)) that penalize for the number of parameters can be used to select among networks with varying numbers of reticulation vertices. PhyloNet (Yu et al 2014) also implements a cross-validation approach to deal with this problem.

  In addition to the property of time-consistency and the specification of the number of reticulations mentioned above, it is common to place further restrictions on the class of network considered in order the simplify the search. For example, PhyloNetworks (Solís-Lemus and Ané 2016; Solís-Lemus et al 2017) and NANUQ (Allman et al 2019) consider only level-1 networks. In contrast, PhyloNet does not provide a clear description of the types of networks considered, while for the SpeciesNetwork module of BEAST2, the birth-hybridization prior assumed limits the class of networks considered to those that can be generated under this process (Zhang et al 2017). The extension of the BPP software to handle horizontal events, on the other hand, currently requires that the phylogenetic network be specified in advance, though it can accommodate a relatively broad class of possible networks (Flouri et al 2019). In reviewing these methods, it was often difficult to determine which classes of networks could be inferred with the various methods. One goal in writing this review is to encourage authors of methods for inferring networks to explicitly state the class of networks on which their methods operate. In addition, we encourage authors to discuss the computational limits of their methods, for instance the number of species they can handle, as this is important information for practical purposes. Most methods to-date seem only to be feasible for a handful of species, but it is again difficult to make precise statements here.

***Efficiently searching network space***

- *Criterion-based methods* Numerous approaches to the problem of finding an optimal network under a selected optimality criterion have been taken, many of which are based on methods that have been applied in the case of searching for optimal phylogenetic trees. For instance, heuristic searches operate by proposing a new network by perturbing a current network using generalizations of move strategies such as nearest neighbor interchange (NNI) or subtree prune and regraft (SPR) to networks, or changing the complexity of the current network by adding or deleting reticulations and then evaluating whether the newly proposed network improves the value of the objective function (for more details on network move strategies, see, e.g., Huber et al (2015a); Gambette et al (2017); for their implementation in software packages see, e.g., Than et al (2008); Solís-Lemus and Ané (2016)). If so, the new network becomes the current network and the process is repeated; if not, a different move is attempted. The process continues until all possible networks have been proposed without any being accepted. Such a strategy clearly provides an uphill search that may result in a network that is only locally optimal. To combat this, heuristic algorithms are often run many times from different initial networks. Another possibility is to apply simulated annealing approaches, which have been successfully used to search for optimal phylogenetic trees (see, e.g., Salter and Pearl 2001; Barker 2004; Stamatakis 2005; Kubatko et al 2009; Strobl and Barker 2016).
- *MCMC methods* Markov chain Monte Carlo algorithms are similar to the searches described above in many ways. For example, they also require a strategy for proposing a new network from an existing network; however, they make a probabilistic decision about acceptance of the proposed network based on the posterior probability. Bayesian methods also provide a more elegant way to handle the problem of inferring the number of horizontal events via reversible-jump MCMC (Green 1995; Green et al 2003; Wen and Nakhleh 2017). This approach allows the algorithm to *'jump'* dimensions by adding a new reticulation event or by removing an existing reticulation event in a manner that preserves the theoretical property that the chain samples the desired posterior distribution. While Bayesian algorithms have the appropriate theoretical guarantees, tuning these algorithms to achieve convergence often proves difficult in practice, and such methods are often limited to carrying out inference on only a handful of taxa at a time.

- *Divide-and-conquer methods* Divide-and-conquer methods, which decompose a problem into smaller sub-problems that can be efficiently solved and whose solutions are then combined to form a solution to the original problem, have also been recently proposed for phylogenetic networks. See Zhu et al (2019) for details.
- *Algorithmic methods* Algorithmic methods are those that build a phylogenetic network by applying a sequence of deterministic steps. For example, methods such as NANUQ (Allman et al 2019) and related methods implemented in the MSCquartets package (Rhodes et al 2020) use an algorithmic approach to infer networks under a model that accommodates variation in the evolutionary histories across loci using the coalescent process.

### 4.1.3 Summary

Inference of phylogenetic networks from genomic data is clearly an important endeavor, but one for which important scalability challenges exist. Hejase and Liu (2016) have documented the scalability challenges that arise across a range of potential inference methods, and Elworth et al (2019) provide an excellent review of current methods for inferring species-level phylogenetic networks under the coalescent model. This is an active area of research in which we anticipate important conceptual and computational advances in the years to come. A related challenge is the determination of which network properties can be learned from data, a topic that is discussed in the following section.

## 4.2 Identifiability

While we have seen in the previous section that the task of estimating phylogenetic networks from data imposes certain scalability challenges, in this section we discuss another challenge in network inference, namely identifiability. Strictly speaking, identifiability refers to a statistical model, where a model parameter is said to be identifiable if any probability distribution arising from the model uniquely determines the value of that parameter. In the setting of phylogenetic network inference, it is thus particularly important that the network parameter (i.e., the network topology and possibly its branch lengths and inheritance probabilities) is identifiable. However, another aspect of identifiability (or rather distinguishability) frequently discussed in relation to phylogenetic network inference is the question of whether a given phylogenetic network is uniquely characterized by certain substructures of the network (e.g., by the phylogenetic trees it 'displays') or by certain other structural properties of the network like the distribution of paths in the network or pairwise distances between taxa.

In the following, we discuss both aspects of identifiability (where we distinguish between statistical identifiability and combinatorial identifiability) and relate this to the different phylogenetic network classes introduced earlier.

### 4.2.1 Statistical identifiability

An important question in model-based network estimation is the identifiability of the network parameter, i.e., the question of whether the network topology (and possibly additional properties such as branch lengths or inheritance probabilities) can be uniquely identified from data generated by the network. Here, ‘data generated by the network’ typically refers to genomic sequence data observed at the leaves of the network, where sequence evolution along the branches of the network (or along the branches of a set of gene trees associated with the network) is usually modelled as a Markov process (for an introduction to Markov processes and their application to phylogenetics, see, for instance, Steel (2016)).

While several identifiability results for Markov models and the coalescent model on phylogenetic trees have been established in the literature (e.g., Chang (1996); Allman and Rhodes (2006, 2008); Allman et al (2010); Rhodes and Sullivant (2011); Chifman and Kubatko (2015); Long and Kubatko (2018)), analogous results for phylogenetic networks are much harder to attain. In the following, we summarize some important results obtained in the literature so far. However, we begin by briefly reviewing the notions of identifiability and generic identifiability.

Recall that a model parameter is *identifiable* if any probability distribution arising from the model uniquely determines the value of the parameter. As the notion of identifiability is sometimes too strong for practical purposes, generic identifiability is often considered instead. Here, a model parameter is said to be *generically identifiable* if the set of parameters from which the original parameter cannot be recovered is a set of Lebesgue measure zero in the parameter space, or in other words, if the parameter is identifiable almost surely.

### *Identifiability of semi-directed phylogenetic networks*

While it is natural to model sequence evolution on rooted phylogenetic networks (where sequences evolve from the root of the network towards its leaves), most identifiability results obtained so far focus on *semi-directed phylogenetic networks*, where a semi-directed phylogenetic network is obtained from a (directed) rooted phylogenetic network by suppressing the root vertex and undirecting all tree arcs while keeping the direction of all reticulation arcs. This is simply due to the fact that under common Markov models of sequence evolution the placement of the root is not identifiable due to the time-reversibility of these models.

The identifiability of semi-directed phylogenetic networks has been studied by several authors and under different settings. There are both identifiability results assuming the coalescent model as well as results that do not incorporate a coalescent process.

- *Identifiability of topological properties of semi-directed level-*1 *networks assuming a coalescent process* Solís-Lemus and Ané (2016) and Solís-Lemus et al (2020) studied the detectability of hybridization cycles and the identifiability of numerical parameters such as branch lengths and inheritance probabilities for semi-directed level-1 networks under a pseudolikelihood model based on the coalescent process. Note that while Solís-Lemus and Ané (2016) provided these results, the mathematical proofs were given later by Solís-Lemus et al (2020), who showed that the detectability of hybridization cycles depends on the length of the cycle: cycles containing four or more vertices can be detected from concordance factors (CFs; Baum (2007)), i.e., probabilities of the different quartet topologies displayed on gene trees, under a pseudolikelihood model, cycles containing two vertices are not detectable, and cycles containing three vertices can be detected under certain conditions.

In addition and prior to Solís-Lemus et al (2020), Baños (2018) provided a mathematical justification for the approach taken in Solís-Lemus and Ané (2016) by showing that most topological properties (in particular, each hybridization cycle of length at least four) of a semi-directed level-1 network are generically identifiable from CFs under the network multi-species coalescent model.

- *Identifiability of topological properties of ultrametric level-*1 *networks from log-det distances (assuming a coalescent process)* Building upon work of Baños (2018), Allman et al (2021) showed that most topological properties of ultrametric[7] level-1 networks can be identified from log-det inter-taxon distances computed from aligned genomic-scale sequences using a combination of the network multispecies coalescent model and a mixture of general time-reversible (GTR) Markov processes on gene trees (for details, see Allman et al (2021)).
- *Identifiability of topological properties of semi-directed networks assuming network-based Markov models* While all of the work cited in the previous paragraphs takes into account the coalescent process, other studies have used algebraic approaches to show that certain semi-directed phylogenetic networks can be identified under specific Markov models of sequence evolution. In these studies, it is assumed that all sequence sites have evolved on one of the trees 'displayed' (for a formal definition of the concept of a displayed tree, see Sect. 4.2.2) by a network, i.e., Markov models on phylogenetic trees are extended to phylogenetic networks by considering a convex combination of the corresponding Markov models on the set of trees displayed by the networks (see, e.g., Gross and Long (2018) for a formal definition of network-based Markov models). For network-based Markov models, the following results were obtained in the literature:
  - The network parameter (in this case, the network topology) of a network-based Markov model under the Jukes-Cantor (Gross and Long 2018), Kimura 2-parameter, or Kimura 3-parameter (Hollering and Sullivant 2021) constraints is generically identifiable with respect to the class of models where the network parameter is a semi-directed network on $n$ leaves with exactly one undirected cycle of length at least four.
  - The network parameter of a network-based Markov model under the Jukes-Cantor, Kimura 2-parameter, or Kimura 3-parameter constraints is generically identifiable with respect to the class of models where the network parameter is a triangle-free (i.e., each undirected cycle of the network has length at least 4), level-1 semi-directed network on $n$ leaves with $r \geq 0$ reticulation vertices (Gross et al 2021).
  - Certain semi-directed level-2 networks are identifiable under the Jukes-Cantor, Kimura 2-parameter, and Kimura 3-parameter constraints (for details see Ardiyansyah (2021)).

[7] A rooted edge-weighted phylogenetic network $\mathcal{N}$ is called *ultrametric* if every directed path from the root of $\mathcal{N}$ to any leaf has the same length.

***Identifiability of tree-child networks assuming a probabilistic recombination-mutation model***

Finally, we note that other extensions to tree-based Markov models are possible. For instance, adapting a model used in *pedigree* reconstruction (Thatte 2012), Francis and Moulton (2018) introduced an alternative probabilistic recombination-mutation model and established identifiability for almost the entire class of tree-child networks under this model (and under the mild assumption that the root of the network is not the parent of a reticulation vertex). We refer the reader to Francis and Moulton (2018) for further details on the model assumptions.

### 4.2.2 Combinatorial identifiability

In addition to the question of whether a phylogenetic network is identifiable under a certain evolutionary model, it is also of interest to analyze the question of whether a network is identifiable from certain substructures or other structural properties such as inter-taxon distances.

***Encoding networks by subtrees and subnetworks***

A well-known result in phylogenetics is that a rooted binary phylogenetic tree is uniquely encoded by its set of rooted triples (i.e., the set of induced 3-leaf rooted subtrees) (e.g., Aho et al (1981); Semple and Steel (2003)), and so a natural question to ask is whether a rooted binary phylogenetic network can also be uniquely characterized by certain substructures such as subtrees or subnetworks. It turns out that positive results in this regard can be obtained for some of the network classes introduced earlier, whereas arbitrary rooted binary phylogenetic networks are in general not uniquely characterized by simpler substructures. However, before we can elaborate on this, we need to formally define the notion of displayed trees and (displayed) subnetworks.

*Displayed trees* Let $\mathcal{N}$ be a phylogenetic network on $X$, and let $\mathcal{T}$ be a phylogenetic tree on $Y \subseteq X$ (with $Y \neq \emptyset$). Then we say that $\mathcal{T}$ is displayed by $\mathcal{N}$ if $\mathcal{T}$ can be obtained (up to isomorphism) from $\mathcal{N}$ by deleting arcs and non-root vertices, and suppressing any resulting in-degree one and out-degree one vertices. Note that the roots of $\mathcal{N}$ and $\mathcal{T}$ coincide and thus $\rho$ might have out-degree one in $\mathcal{T}$.[8] The set of all phylogenetic $X$-trees displayed by a phylogenetic network $\mathcal{N}$ on $X$ is denoted as $\mathsf{T}(\mathcal{N})$. As an example, the network $\mathcal{N}$ on $X = \{x_1, x_2, x_3\}$ depicted in Fig. 13a displays the two phylogenetic $X$-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ depicted in panel c of this figure.

*Subnetworks* Let $\mathcal{N}$ be a phylogenetic network on $X$ and let $Y \subseteq X$. Following the notation of van Iersel et al (2017b), the *subnet* of $\mathcal{N}$ on $Y$, denoted by $\mathcal{N}_{|Y}$, is defined as the subgraph obtained from $\mathcal{N}$ by deleting all vertices that are not on any path from the lowest stable ancestor $\mathrm{LSA}_{\mathcal{N}}(Y)$ of $Y$ in $\mathcal{N}$ to elements in $Y$ and subsequently suppressing all in-degree one and out-degree one vertices and parallel arcs until no

[8] We remark that sometimes the notion of display is defined as follows: A phylogenetic network $\mathcal{N}$ displays a phylogenetic tree $\mathcal{T}$ if $\mathcal{T}$ can be obtained from $\mathcal{N}$ be deleting arcs and vertices, and suppressing any resulting in-degree one and out-degree one vertices. In this case, the roots of $\mathcal{N}$ and $\mathcal{T}$ are not required to coincide. Note, however, that the two definitions can be used interchangeably. If a tree $\mathcal{T}$ is displayed by $\mathcal{N}$ in the first sense, it is also displayed by $\mathcal{N}$ in the second sense, and vice versa.

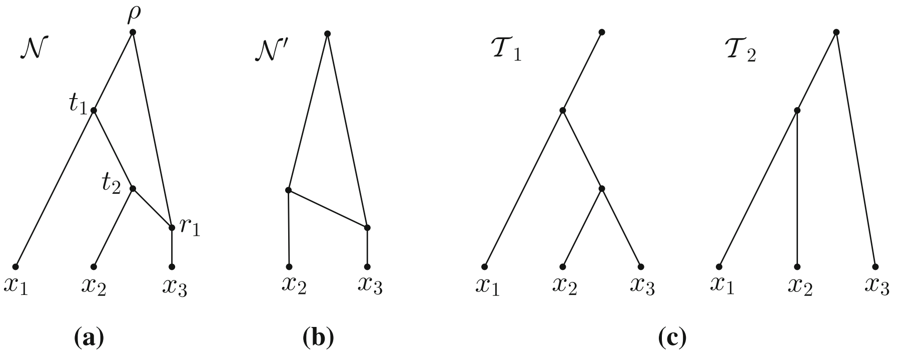


**Fig. 13** **a** Rooted binary phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, x_3\}$. **b** A subnetwork $\mathcal{N}'$ on $Y = \{x_2, x_3\}$ displayed by $\mathcal{N}$. **c** The set $\mathsf{T}(\mathcal{N}) = \{\mathcal{T}_1, \mathcal{T}_2\}$ of phylogenetic $X$-trees displayed by $\mathcal{N}$

such vertices or arcs exist. Now, a network $\mathcal{N}'$ is said to be *displayed* by a network $\mathcal{N}$ if $\mathcal{N}' = \mathcal{N}_{|Y}$ for some $Y \subseteq X$. As an example, the phylogenetic network $\mathcal{N}'$ on $Y = \{x_2, x_3\}$ depicted in Fig. 13b is displayed by the phylogenetic network $\mathcal{N}$ on $X = \{x_1, x_2, x_3\}$ depicted in Fig. 13a. Note that, by definition, $\mathcal{N}_{|X} = \mathcal{N}$ if and only if $\mathrm{LSA}_{\mathcal{N}}(X) = \rho$. In this case, van Iersel et al (2017b) call $\mathcal{N}$ a *recoverable* network.

We are now in a position to summarize some important results on encoding phylogenetic networks by displayed trees and subnetworks.

- *Encodings via displayed trees – arbitrary networks* It was shown by Pardi and Scornavacca (2015) that arbitrary rooted binary phylogenetic networks on $n$ leaves are not encoded by the set of phylogenetic trees on $n$ leaves they display, even if branch lengths are taken into account (Pardi and Scornavacca 2015, Fig. 3). In particular, there exist non-isomorphic phylogenetic networks $\mathcal{N}_1$ and $\mathcal{N}_2$ on $X$ such that $\mathsf{T}(\mathcal{N}_1) = \mathsf{T}(\mathcal{N}_2)$, i.e., $\mathcal{N}_1$ and $\mathcal{N}_2$ display the same set of phylogenetic $X$-trees. Interestingly, the two networks $\mathcal{N}_1$ and $\mathcal{N}_2$ used by Pardi and Scornavacca (2015, Fig. 3) can be distinguished under the network multi-species coalescent model when multiple alleles are sampled per species (Zhu and Degnan 2016). This nicely illustrates that statistical and combinatorial considerations might not always lead to the same conclusions, and more work on combining the two approaches is needed.
- *The NELP property* Pardi and Scornavacca (2015) showed that every phylogenetic network $\mathcal{N}$ can be transformed into a ‘canonical’ form that displays the same set of phylogenetic trees as $\mathcal{N}$ and under mild conditions on the branch lengths of $\mathcal{N}$ (the *no equally long paths* (NELP) property which states that no two directed paths in $\mathcal{N}$ with the same endpoints have the same length, where the length of a path is the sum of the branch lengths assigned to its arcs) this canonical form is unique (up to isomorphism) among all networks satisfying the NELP property. In particular, two networks satisfying the NELP property have the same unique canonical form (up to isomorphism) if and only if they display the same set of trees (with their induced branch lengths). For details, see Pardi and Scornavacca (2015).

  Note that a different canonical form for rooted phylogenetic networks was recently introduced by Francis et al (2021). Specifically, Francis et al (2021) introduced the *normalization* of a phylogenetic network that associates a unique normal network (Willson 2009) $\widetilde{\mathcal{N}}$ with any given phylogenetic network $\mathcal{N}$.

- *Encodings via displayed (caterpillar) trees – normal networks* It was shown by Willson (2011) that normal phylogenetic networks on $n$ leaves are uniquely encoded by the set of phylogenetic trees on $n$ leaves they display. More recently, Linz and Semple (2020) showed that normal phylogenetic networks are in fact uniquely characterized by their sets of displayed caterpillar trees (particular subtrees that contain precisely one cherry) on three and four leaves. Moreover, Linz and Semple (2020) presented a polynomial-time algorithm that takes the set of caterpillar trees on three and four leaves displayed by a rooted binary normal network and reconstructs this network (up to isomorphism). Note that considering caterpillar trees on three and four leaves is essential as there exist two non-isomorphic normal networks that display the same set of rooted triples, i.e., caterpillar trees on three leaves (for an example, see (Linz and Semple 2020, Fig. 1)).
- *Encodings via rooted triples* Phylogenetic networks that are encoded by their rooted triples only seem to be very limited. To our knowledge, the only positive result in this regard was obtained by Gambette and Huber (2011) who showed that level-1 networks are encoded by their sets of displayed triples provided that each reticulation cycle in the network has length at least five.

  We remark, however, that even though phylogenetic networks are in general not uniquely encoded by their rooted triples, several algorithms that reconstruct *a* phylogenetic network consistent with a set of triples (consistent in the sense that the triples are displayed by the resulting network) have been developed. We refer the reader to Poormohammadi and Zarchi (2020) for an overview and comparison of different approaches and recent developments.
- *Encodings via binets, trinets, quarnets, and larger subnetworks* Given the limited number of positive results in encoding phylogenetic networks by tree substructures, several studies have analyzed the question whether a phylogenetic network is uniquely characterized by certain network substructures like binets (Huber et al 2015b; van Iersel et al 2017b), trinets (Huber and Moulton 2012; Huber et al 2015b; Semple and Toft 2021; van Iersel and Moulton 2013; van Iersel et al 2017b), and recently also quarnets (Nipius 2020), i.e., subnetworks on two, three, and four leaves, respectively. Here, the following positive results have been established:

  – Recoverable binary level-2 networks and binary tree-child networks are encoded by their sets of displayed trinets (van Iersel and Moulton 2013).
  – Orchard networks are uniquely encoded by their trinets and can be reconstructed in polynomial time from them (Semple and Toft 2021).
  – Every recoverable binary level-3 network is encoded by its set of displayed quarnets (Nipius 2020).

  However, as far as arbitrary phylogenetic networks are concerned, it has been shown by Huber et al (2014) that even if *all* subnetworks induced on all proper subsets of the leaves of some rooted binary phylogenetic network are given, the network is still not necessarily determined by this information.

- *Encodings via reticulate-edge-deleted subnetworks* Murakami et al (2019) recently showed that level-$k$ tree-child networks with $k \geq 2$ can be determined and reconstructed in polynomial time from their *reticulate-edge-deleted subnetworks*, which are subnetworks obtained by deleting a single reticulation arc. Even stronger, level-$k$ tree-child networks with $k \geq 2$ are encoded by their subnetworks obtained from deleting one reticulation arc from each biconnected component with $k$ reticulations (for details see Murakami et al (2019)).
- *Encodings via tri-LGT-nets* Cardona and Pons (2017) showed that a subclass of time-consistent LGT networks (referred to as time-consistent *BAN-LGT networks*) can be uniquely reconstructed (up to redundant arcs (shortcuts) and isomorphism) from the set of their *tri-LGT-nets*, i.e., from the set of their induced 3-leaf subnetworks (for further details see Cardona and Pons (2017)).

***Encoding networks by sets of paths and pairwise distances***

An alternative approach to encoding a phylogenetic network by substructures is to consider other structural properties of the network such as the distribution of path lengths or pairwise distances between the leaves. In the following, we review different concepts and ideas used in this regard.

- *The $\mu$-representation of phylogenetic networks* The $\mu$-representation of phylogenetic networks relies on the notion of path-multiplicity vectors introduced by Cardona et al (2009b). Using a similar notation as Cardona et al (2009b), let $\mathcal{N} = (V, E)$ be a rooted phylogenetic network on $X = \{x_1, \dots, x_n\}$. For every vertex $v \in V$ and $i \in \{1, \dots, n\}$, let $m_i(v)$ denote the number of different paths from $v$ to leaf $x_i$. Then, the *path-multiplicity vector*, or *$\mu$-vector* for short, of $v \in V$ is defined as $\mu(v) = (m_1(v), \dots, m_n(v))$, i.e., $\mu(v)$ is an $n$-tuple containing the number of paths from $v$ to each leaf of $\mathcal{N}$. Now, the *$\mu$-representation* of a phylogenetic network $\mathcal{N} = (V, E)$ is the multiset $\mu(\mathcal{N})$ of $\mu$-vectors of its vertices. More precisely, the elements of this multiset are the vectors $\mu(v)$ with $v \in V$, and the multiplicity of each element is the number of vertices having this element as its $\mu$-vector.

  Based on this, Cardona et al (2009b) showed that two (not necessarily binary) tree-child networks $\mathcal{N}_1$ and $\mathcal{N}_2$ are isomorphic if and only if they have the same $\mu$-representation, i.e., if and only if $\mu(\mathcal{N}_1) = \mu(\mathcal{N}_2)$. Moreover, given the $\mu$-representation of a phylogenetic network $\mathcal{N}$, the network can be reconstructed in polynomial time. For further details, see Cardona et al (2009b).

  In addition, Cardona et al (2008) showed that the same is true for semi-binary[9] time-consistent tree-sibling phylogenetic networks.

  Recently, these results were extended to the larger class of orchard phylogenetic networks by Erdős et al (2019) and Bai et al (2021). More precisely, Bai et al (2021) showed that any (not necessarily binary) stack-free orchard phylogenetic network is uniquely encoded (up to isomorphism) by its $\mu$-representation (called 'ancestral profile' therein).

  Note that while the results by Cardona et al (2008, 2009b) entail uniqueness

[9] In a semi-binary phylogenetic network all reticulation vertices have in-degree precisely two, but tree vertices may have an out-degree strictly greater than two. Note that every binary phylogenetic network is in particular a semi-binary phylogenetic network.

within the class of time-consistent tree-sibling, respectively tree-child, networks, the result of Bai et al (2021) proves uniqueness among all rooted phylogenetic networks.

In addition, Bai et al (2021) showed that if the 'stack-free' condition is omitted, the ancestral profile of an orchard network $\mathcal{N}$ uniquely encodes $\mathcal{N}$ within the class of orchard networks up to the resolution of vertices of high in-degree. For further details, see Bai et al (2021).

- *Encoding phylogenetic networks by pairwise distances* Some phylogenetic networks can be encoded by considering pairwise distances between the leaves or taxa of the network. Here, distances can either be measured in terms of topological path lengths between leaves, where the length of a path is defined as the number of arcs contained in it, or in terms of the sum of edge weights on these paths when the network is equipped with branch lengths. An important concept in both cases is the notion of *up-down paths* introduced by Bordewich and Semple (2015).

  *Up-down paths.* Let $\mathcal{N}$ be a rooted phylogenetic network on $X$. Then, following the notation of Bordewich and Semple (2015), an *underlying path* of $\mathcal{N}$ is a path of the undirected graph containing undirected edges of arcs of $\mathcal{N}$. Now, for any two elements $x, y \in X$, an *up-down path* from $x$ to $y$ is an underlying path $(x, v_1, v_2, \ldots, v_{k-1}, y)$ in $\mathcal{N}$ such that, for some $i \leq k-1$, the network $\mathcal{N}$ contains the arcs

  $$(v_i, v_{i-1}), (v_{i-1}, v_{i-2}), \ldots, (v_1, x)$$

  and

  $$(v_i, v_{i+1}), (v_{i+1}, v_{i+2}), \ldots, (v_{k-1}, y).$$

  As an example, in Fig. 13a, $(x_1, t_1, t_2, r_1, x_3)$ is an up-down path in $\mathcal{N}$ from $x_1$ to $x_3$.

  *Unweighted phylogenetic networks.* Based on the notion of up-down-paths, Bordewich and Semple (2015) showed that unweighted binary tree-child networks with no arc between the two children of the root can be reconstructed (up to isomorphism) from the *multi-set* of distances between taxa, where the distance between two taxa, say $x$ and $y$, is measured in terms of the number of arcs on the up-down-paths from $x$ to $y$, and so can all binary time-consistent networks with no 'crowns'[10], no arc between the two children of the root, and all reticulation vertices being visible.

  Moreover, Bordewich and Semple (2015) showed that binary time-consistent tree-child networks can be reconstructed (up to isomorphism) from the *set* of distances between taxa in polynomial time.

  *Edge-weighted rooted phylogenetic networks.* In the case of edge-weighted phylogenetic networks, various results have been obtained in the literature.

  First, improving earlier results on the reconstructability of ultrametric galled networks (Chan et al 2005) and networks with a single reticulation cycle

[10] Let $\mathcal{N}$ be a phylogenetic network. Then, a *crown* is an (undirected) cycle in $\mathcal{N}$ consisting only of reticulation arcs.

(Willson 2013)[11], Bordewich and Tokac (2016) introduced a polynomial-time algorithm that reconstructs an ultrametric tree-child network from the set of distances between each pair of taxa, where the set of distances between a pair of taxa, say $x$ and $y$, is the set of the lengths of the up-down-paths from $x$ to $y$ (where the length of any such a path is the sum of branch lengths of the edges in this path rather than the number of edges in the path).

In particular, Bordewich and Tokac (2016) introduced the algorithm NetworkUPGMA that takes a 2-dimensional array of sets of distances (where distances between taxa may for example reflect evolutionary distance estimated from genetic sequence data) and returns an ultrametric tree-child network displaying the distance data if such a network exists.

In a subsequent paper, Bordewich et al (2017) showed that any tree-child network $\mathcal{N}$ on $X$ with an outgroup (i.e., an element $x \in X$ adjacent to the root of $\mathcal{N}$) and strictly positive branch lengths is essentially encoded by the multiset of distances between all pairs of taxa (again, distance refers to sum of branch lengths in up-down-paths) provided that for each reticulation vertex $r$, both reticulation arcs directed into $r$ are of equal length. Note that 'essentially encoded' refers to the fact that this encoding is unique up to re-weighting the edges at the root and at each reticulation (for technical details and examples see Bordewich et al (2017)). Moreover, Bordewich et al (2017) introduced a polynomial-time algorithm for reconstructing edge-weighted tree-child networks from inter-taxa distance data. Note, however, that in a tree-child network, the size of the collection of inter-taxon distances can be exponential in the number of leaves of the network.

In a recent paper, Bordewich et al (2018) showed that for normal phylogenetic networks the same results are obtained with only a quadratic number of inter-taxon distances by using the shortest distance between any pair of taxa (i.e., the sum of branch lengths in a shortest up-down path between the two taxa).

*Edge-weighted semi-directed networks.* Even more recently, Xu and Ané (2021) returned to the identifiability of local and global properties of edge-weighted phylogenetic networks from *average pairwise distances* (Willson (2012); see also Footnote 11). Importantly, the authors show that root location and lengths of reticulation arcs are generally *not* identifiable from average distances, and then focus on the identifiability of "zipped-up semi-directed" networks, where a network is zipped-up if all its reticulation arcs have length zero. For networks of this type, several positive and negative results regarding their identifiability from average distances are obtained and additional conjectures are posed. We refer the reader to Xu and Ané (2021) for further details.

[11] Note that Willson (2013) considered the problem of reconstructing a phylogenetic network $\mathcal{N}$ given the *tree-average distance* (Willson 2012) between any pair of taxa, which is the expected value of their distance in the trees displayed by $\mathcal{N}$, where each displayed tree has a certain probability obtained from assigning inheritance probabilities to the reticulation arcs of $\mathcal{N}$.

### 4.2.3 Summary

In summary, several results concerning the identifiability of phylogenetic networks have been established in recent years and more are likely to be obtained in the near future. However, as the previous paragraphs showed, positive results can mostly only be obtained for restricted network classes, but not for arbitrary phylogenetic networks. While this is to be expected (given the potential complexity of arbitrary phylogenetic networks) it is important to keep these limitations in mind when, for example, devising new network inference methods.

## 5 Concluding remarks

The aim of the present manuscript is to provide a thorough and comprehensive review of the multitude of different classes of rooted binary phylogenetic networks defined in the mathematical literature. We have reviewed and discussed their structural properties and indicated their biological interpretation whenever possible. For some network classes, for instance temporal networks or LGT networks, it is straightforward to provide a biological interpretation, whereas for other network classes, the biological meaning is less evident. This is to be expected, however, as many of the structural constraints that have been considered in the literature were not introduced to model certain biological processes, but rather to simplify mathematical and computational analyses so that problems related to the inference of phylogenetic networks would become tractable.

We have noted that imposing structural constraints on network topologies is often an important step in addressing the scalability and identifiability challenges faced in estimating phylogenetic networks from data, and hope that our review of possible classes of networks will encourage those who develop such methods to carefully describe the class of networks considered by their methods. Even though positive results concerning scalability and identifiability are currently limited to a handful of the more than 20 network classes we have discussed, we expect constant progress in the future as the study and estimation of phylogenetic networks is a growing field of research. In addition, we remark that while we could not find a biological interpretation for all of the network classes discussed, we do not claim that such an interpretation is non-existent. It could well be the case that network types that seem to lack an underlying biological principle in the context of our current understanding of evolution will turn out to be biologically meaningful as this understanding expands.

**Acknowledgements** JCP was supported by the Ministerio de Ciencia e Innovación (MCI), the Agencia Estatal de Investigación (AEI) and the European Regional Development Funds (ERDF); through project PGC2018-096956-B-C43 (FEDER/MICINN/AEI). KW was supported by The Ohio State University's President's Postdoctoral Scholars Program. All authors thank two anonymous reviewers for detailed comments on an earlier version of this manuscript.

**Funding** JCP was supported by the Ministerio de Ciencia e Innovación (MCI), the Agencia Estatal de Investigación (AEI) and the European Regional Development Funds (ERDF); through project PGC2018-096956-B-C43 (FEDER/MICINN/AEI). KW was supported by The Ohio State University's President's Postdoctoral Scholars Program.

**Availability of data and material.** Not applicable.

## Declarations

**Conflicts of interest/Competing interests.** The authors declare that there is no conflict of interest.

**Code availability.** Not applicable.

## References

Agarwal T, Gambette P, Morrison D (2016) Who is Who in phylogenetic networks: articles, authors and programs arXiv e-prints arXiv:1610.01674

Aho AV, Sagiv Y, Szymanski TG et al (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J Comput 10(3):405–421. https://doi.org/10.1137/0210030

Allman ES, Rhodes JA (2006) The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J Comput Biol 13(5):1101–1113. https://doi.org/10.1089/cmb.2006.13.1101

Allman ES, Rhodes JA (2008) Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. Math Biosci 211(1):18–33. https://doi.org/10.1016/j.mbs.2007.09.001

Allman ES, Degnan JH, Rhodes JA (2010) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62(6):833–862. https://doi.org/10.1007/s00285-010-0355-7

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

Allman ES, Baños H, Rhodes JA (2021) Identifiability of species network topologies from genomic sequences using the logDet distance arXiv e-prints arXiv:2108.01765

Anderson E (1953) Introgressive hybridization. Biol Rev 28(3):280–307. https://doi.org/10.1111/j.1469-185x.1953.tb01379.x

Ardiyansyah M (2021) Distinguishing level-2 phylogenetic networks using phylogenetic invariants arXiv e-prints arXiv:2104.12479

Asano T, Jansson J, Sadakane K, et al (2010) Faster computation of the Robinson-Foulds distance between phylogenetic networks In: Combinatorial Pattern Matching Springer Berlin Heidelberg, pp 190–201 https://doi.org/10.1007/978-3-642-13509-5_18

Bai A, Erdős PL, Semple C et al (2021) Defining phylogenetic networks using ancestral profiles. Math Biosci 332:108537. https://doi.org/10.1016/j.mbs.2021.108537

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

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1):37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Baños H (2018) Identifying species network features from gene tree quartets under the coalescent model. Bull Math Biol 81(2):494–534. https://doi.org/10.1007/s11538-018-0485-4

Barker D (2004) LVB: parsimony and simulated annealing in the search for phylogenetic trees. Bioinformatics 20(2):274–275. https://doi.org/10.1093/bioinformatics/btg402

Baroni M, Semple C, Steel M (2005) A framework for representing reticulate evolution. Ann Comb 8(4):391–408. https://doi.org/10.1007/s00026-004-0228-0

Baroni M, Semple C, Steel M (2006) Hybrids in real time. Syst Biol 55(1):46–56. https://doi.org/10.1080/10635150500431197

Baum DA (2007) Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56(2):417–426. https://doi.org/10.1002/tax.562013

Bordewich M, Semple C (2015) Determining phylogenetic networks from inter-taxa distances. J Math Biol 73(2):283–303. https://doi.org/10.1007/s00285-015-0950-8

Bordewich M, Semple C (2016) Reticulation-visible networks. Adv Appl Math 78:114–141. https://doi.org/10.1016/j.aam.2016.04.004

Bordewich M, Semple C (2018) A universal tree-based network with the minimum number of reticulations. Discret Appl Math 250:357–362. https://doi.org/10.1016/j.dam.2018.05.010

Bordewich M, Tokac N (2016) An algorithm for reconstructing ultrametric tree-child networks from inter-taxa distances. Discret Appl Math 213:47–59. https://doi.org/10.1016/j.dam.2016.05.011

Bordewich M, Semple C, Tokac N (2017) Constructing tree-child networks from distance matrices. Algorithmica 80(8):2240–2259. https://doi.org/10.1007/s00453-017-0320-6

Bordewich M, Huber KT, Moulton V et al (2018) Recovering normal networks from shortest inter-taxa distance information. J Math Biol 77(3):571–594. https://doi.org/10.1007/s00285-018-1218-x

Bryant C, Fischer M, Linz S et al (2017) On the quirks of maximum parsimony and likelihood on phylogenetic networks. J Theor Biol 417:100–108. https://doi.org/10.1016/j.jtbi.2017.01.013

Cardona G, Pons JC (2017) Reconstruction of LGT networks from tri-LGT-nets. J Math Biol 75(6–7):1669–1692. https://doi.org/10.1007/s00285-017-1131-8

Cardona G, Llabrés M, Rosselló F et al (2008) A distance metric for a class of tree-sibling phylogenetic networks. Bioinformatics 24(13):1481–1488. https://doi.org/10.1093/bioinformatics/btn231

Cardona G, Llabres M, Rossello F et al (2009) Metrics for phylogenetic networks I: generalizations of the Robinson-Foulds metric. IEEE/ACM Trans Comput Biol Bioinf 6:46–61. https://doi.org/10.1109/TCBB.2008.70

Cardona G, Rossello F, Valiente G (2009) Comparison of tree-child phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinf 6(4):552–569. https://doi.org/10.1109/tcbb.2007.70270

Cardona G, Llabrés M, Rosselló F et al (2010) Path lengths in tree-child time consistent hybridization networks. Inf Sci 180(3):366–383. https://doi.org/10.1016/j.ins.2009.09.013

Cardona G, Pons JC, Rosselló F (2015) A reconstruction problem for a class of phylogenetic networks with lateral gene transfers. Algorithms Mol Biol. https://doi.org/10.1186/s13015-015-0059-z

Chan HL, Jansson J, Lam TW, et al (2005) Reconstructing an ultrametric galled phylogenetic network from a distance matrix In: Mathematical foundations of computer science 2005 Springer Berlin Heidelberg, pp 224–235, https://doi.org/10.1007/11549345_20

Chang JT (1996) Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math Biosci 137(1):51–73. https://doi.org/10.1016/s0025-5564(96)00075-2

Chifman J, Kubatko L (2015) Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J Theor Biol 374:35–47. https://doi.org/10.1016/j.jtbi.2015.03.006

Choy C, Jansson J, Sadakane K et al (2005) Computing the maximum agreement of phylogenetic networks. Theoret Comput Sci 335(1):93–107. https://doi.org/10.1016/j.tcs.2004.12.012

Cordue P, Linz S, Semple C (2014) Phylogenetic networks that display a tree twice. Bull Math Biol 76(10):2664–2679. https://doi.org/10.1007/s11538-014-0032-x

Corel E, Lopez P, Méheust R et al (2016) Network-thinking: graphs to analyze microbial complexity and evolution. Trends Microbiol 24(3):224–237. https://doi.org/10.1016/j.tim.2015.12.003

Dagan T, Martin W (2006) The tree of one percent. Genome Biol 7(10):118. https://doi.org/10.1186/gb-2006-7-10-118

Daubin V (2003) Phylogenetics and the Cohesion of bacterial genomes. Science 301(5634):829–832. https://doi.org/10.1126/science.1086568

Degnan JH (2018) Modeling hybridization under the network multispecies coalescent. Syst Biol 67(5):786–799. https://doi.org/10.1093/sysbio/syy040

Doolittle WF, Bapteste E (2007) Pattern pluralism and the tree of life hypothesis. Proc Natl Acad Sci 104(7):2043–2049. https://doi.org/10.1073/pnas.0610699104

Elworth RAL, Ogilvie HA, Zhu J, et al (2019) Advances in computational methods for phylogenetic networks in the presence of hybridization In: Bioinformatics and Phylogenetics Springer, pp 317–360, https://doi.org/10.1007/978-3-030-10837-3_13

Erdős PL, Semple C, Steel M (2019) A class of phylogenetic networks reconstructable from ancestral profiles. Math Biosci 313:33–40. https://doi.org/10.1016/j.mbs.2019.04.009

Fischer M, van Iersel L, Kelk S et al (2015) On computing the maximum parsimony score of a phylogenetic network. SIAM J Discret Math 29(1):559–585. https://doi.org/10.1137/140959948

Flouri T, Jiao X, Rannala B et al (2019) A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol Biol Evol 37(4):1211–1223. https://doi.org/10.1093/molbev/msz296

Francis A, Moulton V (2018) Identifiability of tree-child phylogenetic networks under a probabilistic recombination-mutation model of evolution. J Theor Biol 446:160–167. https://doi.org/10.1016/j.jtbi.2018.03.011

Francis A, Huson DH, Steel M (2021) Normalising phylogenetic networks. Mol Phylogenet Evol 163:107215. https://doi.org/10.1016/j.ympev.2021.107215

Francis AR, Steel M (2015) Which phylogenetic networks are merely trees with additional arcs? Syst Biol 64(5):768–777. https://doi.org/10.1093/sysbio/syv037

Gambette P, Huber KT (2011) On encodings of phylogenetic networks of bounded level. J Math Biol 65(1):157–180. https://doi.org/10.1007/s00285-011-0456-y

Gambette P, Gunawan ADM, Labarre A, et al (2015) Locating a tree in a phylogenetic network in quadratic time In: Lecture notes in computer science. Springer, pp 96–107, https://doi.org/10.1007/978-3-319-16706-0_12

Gambette P, Gunawan ADM, Labarre A, et al (2016) Solving the tree containment problem for genetically stable networks in quadratic time In: Lecture Notes in Computer Science Springer, pp 197–208, https://doi.org/10.1007/978-3-319-29516-9_17

Gambette P, van Iersel L, Jones M et al (2017) Rearrangement moves on rooted phylogenetic networks. PLoS Comput Biol 13(8):e1005,611. https://doi.org/10.1371/journal.pcbi.1005611

Gambette P, Gunawan AD, Labarre A et al (2018) Solving the tree containment problem in linear time for nearly stable phylogenetic networks. Discret Appl Math 246:62–79. https://doi.org/10.1016/j.dam.2017.07.015

Gambette P, Morgado M, Tavassoli N, et al (2018b) ISIPhyNC, an information system on inclusions of phylogenetic network classes, manuscript in preparation

Górecki P (2004) Reconciliation problems for duplication, loss and horizontal gene transfer In: Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04 ACM Press, https://doi.org/10.1145/974614.974656

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732. https://doi.org/10.1093/biomet/82.4.711

Green PJ, Hjort NL, Richardson S (eds) (2003) Highly structured stochastic systems. Oxford University Press (Oxford Statistical Science Series). ISBN: 978-0-198-51055-0

Gross E, Long C (2018) Distinguishing phylogenetic networks. SIAM J Appl Algebr Geometry 2(1):72–93. https://doi.org/10.1137/17m1134238

Gross E, van Iersel L, Janssen R et al (2021) Distinguishing level-1 phylogenetic networks on the basis of data generated by Markov processes. J Math Biol. https://doi.org/10.1007/s00285-021-01653-8

Gunawan ADM, Zhang L (2015) Bounding the size of a network defined by visibility property arXiv e-prints arXiv:1510.00115

Gusfield D (2014) ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks. The MIT Press, Cambridge, MA

Gusfield D, Eddhu S, Langley C (2003) Efficient reconstruction of phylogenetic networks with constrained recombination In: Computational Systems Bioinformatics. CSB2003 Proceedings of the 2003 IEEE Bioinformatics Conference CSB2003 IEEE Comput Soc https://doi.org/10.1109/csb.2003.1227337

Hayamizu M (2016) On the existence of infinitely many universal tree-based networks. J Theor Biol 396:204–206. https://doi.org/10.1016/j.jtbi.2016.02.023

Hejase HA, Liu KJ (2016) A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinform. https://doi.org/10.1186/s12859-016-1277-1

Hollering B, Sullivant S (2021) Identifiability in phylogenetics using algebraic matroids. J Symb Comput 104:142–158. https://doi.org/10.1016/j.jsc.2020.04.012

Huber K, Moulton V (2006) Phylogenetic networks from multi-labelled trees. J Math Biol 52(5):613–632. https://doi.org/10.1007/s00285-005-0365-z

Huber K, Scholz G (2020) Phylogenetic networks that are their own fold-ups. Adv Appl Math 113:101959. https://doi.org/10.1016/j.aam.2019.101959

Huber KT, Moulton V (2012) Encoding and constructing 1-nested phylogenetic networks with trinets. Algorithmica 66(3):714–738. https://doi.org/10.1007/s00453-012-9659-x

Huber KT, van Iersel L, Moulton V et al (2014) How much information is needed to infer reticulate evolutionary histories? Syst Biol 64(1):102–111. https://doi.org/10.1093/sysbio/syu076

Huber KT, Linz S, Moulton V et al (2015) Spaces of phylogenetic networks from generalized nearest-neighbor interchange operations. J Math Biol 72(3):699–725. https://doi.org/10.1007/s00285-015-0899-7

Huber KT, van Iersel L, Moulton V et al (2015) Reconstructing phylogenetic level-1 networks from nondense binet and trinet sets. Algorithmica 77(1):173–200. https://doi.org/10.1007/s00453-015-0069-8

Huber KT, Moulton V, Steel M et al (2016) Folding and unfolding phylogenetic trees and networks. J Math Biol 73(6–7):1761–1780. https://doi.org/10.1007/s00285-016-0993-5

Huber KT, van Iersel L, Janssen R, et al (2019) Rooting for phylogenetic networks arXiv e-prints arXiv:1906.07430

Huson DH, Klöpper TH (2007) Beyond galled trees - decomposition and computation of galled networks In: Lecture Notes in Computer Science. Springer Berlin, pp 211–225, https://doi.org/10.1007/978-3-540-71681-5_15

Huson DH, Rupp R, Scornavacca C (2011) Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press, New York, NY, USA

Janssen R, Murakami Y (2021) On cherry-picking and network containment. Theoret Comput Sci 856:121–150. https://doi.org/10.1016/j.tcs.2020.12.031

Jansson J, Sung WK (2004) The maximum agreement of two nested phylogenetic networks In: Algorithms and Computation Springer Berlin Heidelberg, pp 581–593, https://doi.org/10.1007/978-3-540-30551-4_51

Jiao X, Flouri T, Yang Z (2021) Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. Nat Sci Rev. https://doi.org/10.1093/nsr/nwab127

Jin G, Nakhleh L, Snir S et al (2009) Parsimony score of phylogenetic networks: hardness results and a linear-time heuristic. IEEE/ACM Trans Comput Biol Bioinf 6(3):495–505. https://doi.org/10.1109/tcbb.2008.119

Kannan L, Wheeler WC (2012) Maximum parsimony on phylogenetic networks. Algorithms Mol Biol. https://doi.org/10.1186/1748-7188-7-9

Kingman J (1982) The coalescent. Stoch Processes Appl 13(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kong S, Sánchez-Pacheco SJ, Murphy RW (2015) On the use of median-joining networks in evolutionary biology. Cladistics 32(6):691–699. https://doi.org/10.1111/cla.12147

Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. Syst Biol 58(5):478–488. https://doi.org/10.1093/sysbio/syp055

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25(7):971–973. https://doi.org/10.1093/bioinformatics/btp079

Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. Proc Natl Acad Sci 100(17):9658–9662. https://doi.org/10.1073/pnas.1632870100

Lemay M, Libeskind-Hadas R, Wu YC (2021) A polynomial-time algorithm for minimizing the deep coalescence cost for level-1 species networks. IEEE/ACM Transactions Comput Biol Bioinform. https://doi.org/10.1109/tcbb.2021.3105922

Linz S, Semple C (2020) Caterpillars on three and four leaves are sufficient to reconstruct binary normal networks. J Math Biol 81(4–5):961–980. https://doi.org/10.1007/s00285-020-01533-7

Long C, Kubatko L (2018) Identifiability and reconstructibility of species phylogenies under a modified coalescent. Bull Math Biol 81(2):408–430. https://doi.org/10.1007/s11538-018-0456-9

Lutteropp S, Scornavacca C, Kozlov AM, et al (2021) NetRAX: accurate and fast maximum likelihood phylogenetic network inference*. bioRxiv https://doi.org/10.1101/2021.08.30.458194

Martin WF (2011) Early evolution without a tree of life. Biol Direct 6(1):36. https://doi.org/10.1186/1745-6150-6-36

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75(1):35–45. https://doi.org/10.1016/j.tpb.2008.10.004

Morrison DA (2011) An introduction to phylogenetic networks RJR Productions, Uppsala, oCLC: 939959509

Murakami Y, van Iersel L, Janssen R et al (2019) Reconstructing tree-child networks from reticulate-edge-deleted subnetworks. Bull Math Biol 81(10):3823–3863. https://doi.org/10.1007/s11538-019-00641-w

Nakhleh L, Jin G, Zhao F, et al (2005) Reconstructing phylogenetic networks using maximum parsimony In: 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05) IEEE https://doi.org/10.1109/csb.2005.47

Nipius L (2020) Rooted binary level-3 phylogenetic networks are encoded by quarnets Master's thesis, Delft University of Technology

Pardi F, Scornavacca C (2015) Reconstructible phylogenetic networks: do not distinguish the indistinguishable. PLoS Comput Biol 11(4):e1004,135

Pons JC (2016) Reconstruction Problems for LGT Networks PhD thesis, University of the Balearic Islands

Poormohammadi H, Zarchi MS (2020) Netcombin: an algorithm for constructing optimal phylogenetic network from rooted triplets. PLoS One 15(9):e0227,842. https://doi.org/10.1371/journal.pone.0227842

Rhodes JA, Sullivant S (2011) Identifiability of large phylogenetic mixture models. Bull Math Biol 74(1):212–231. https://doi.org/10.1007/s11538-011-9672-2

Rhodes JA, Baños H, Mitchell JD et al (2020) MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. Bioinformatics 37(12):1766–1768. https://doi.org/10.1093/bioinformatics/btaa868

Robinson D, Foulds L (1981) Comparison of phylogenetic trees. Math Biosci 53(1–2):131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Salter LA, Pearl DK (2001) Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst Biol 50(1):7–17. https://doi.org/10.1080/106351501750107413

Scornavacca C, Mayol JCP, Cardona G (2017) Fast algorithm for the reconciliation of gene trees and LGT networks. J Theor Biol 418:129–137. https://doi.org/10.1016/j.jtbi.2017.01.024

Semple C (2015) Phylogenetic networks with every embedded phylogenetic tree a base tree. Bull Math Biol 78(1):132–137. https://doi.org/10.1007/s11538-015-0132-2

Semple C, Simpson J (2018) When is a phylogenetic network simply an amalgamation of two trees? Bull Math Biol 80(9):2338–2348. https://doi.org/10.1007/s11538-018-0463-x

Semple C, Steel M (2003) Phylogenetics (Oxford Lecture Series in Mathematics and Its Applications). Oxford University Press, Oxford

Semple C, Toft G (2021) Trinets encode orchard phylogenetic networks. J Math Biol. https://doi.org/10.1007/s00285-021-01654-7

Solís-Lemus C, Ané C (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet 12(3):e1005,896. https://doi.org/10.1371/journal.pgen.1005896

Solís-Lemus C, Bastide P, Ané C (2017) PhyloNetworks: a package for phylogenetic networks. Mol Biol Evol 34(12):3292–3298. https://doi.org/10.1093/molbev/msx235

Solís-Lemus C, Coen A, Ané C (2020) On the identifiability of phylogenetic networks under a pseudolikelihood model arXiv e-prints arXiv:2010.01758

Stamatakis A (2005) An efficient program for phylogenetic inference using simulated annealing In: 19th IEEE international parallel and distributed processing symposium IEEE, https://doi.org/10.1109/ipdps.2005.90

Steel M (2016) Phylogeny: discrete and random processes in evolution Society for Industrial and Applied Mathematics, Philadelphia PA

Strobl MA, Barker D (2016) On simulated annealing phase transitions in phylogeny reconstruction. Mol Phylogenet Evol 101:46–55. https://doi.org/10.1016/j.ympev.2016.05.001

Sánchez-Pacheco SJ, Kong S, Pulido-Santacruz P et al (2020) Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. Proc Natl Acad Sci 117:12,518-12,519. https://doi.org/10.1073/pnas.2007062117

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinform. https://doi.org/10.1186/1471-2105-9-322

Thatte BD (2012) Reconstructing pedigrees: some identifiability questions for a recombination-mutation model. J Math Biol 66(1–2):37–74. https://doi.org/10.1007/s00285-011-0503-8

van Iersel L, Moulton V (2013) Trinets encode tree-child and level-2 phylogenetic networks. J Math Biol. https://doi.org/10.1007/s00285-013-0683-5

van Iersel L, Semple C, Steel M (2010) Locating a tree in a phylogenetic network. Inf Process Lett 110:1037–1043. https://doi.org/10.1016/j.ipl.2010.07.027

van Iersel L, Jones M, Scornavacca C (2017) Improved maximum parsimony models for phylogenetic networks. Syst Biol 67(3):518–542. https://doi.org/10.1093/sysbio/syx094

van Iersel L, Moulton V, de Swart E et al (2017) Binets: fundamental building blocks for phylogenetic networks. Bull Math Biol 79(5):1135–1154. https://doi.org/10.1007/s11538-017-0275-4

van Iersel L, Janssen R, Jones M, et al (2021) Orchard networks are trees with additional horizontal arcs arXiv e-prints arXiv:2110.11065

Vu H, Chin F, Hon WK, et al (2013) Reconstructing k-reticulated phylogenetic network from a set of gene trees In: Bioinformatics Research and Applications Springer Berlin Heidelberg, pp 112–124, https://doi.org/10.1007/978-3-642-38036-5_14

Wang L, Zhang K, Zhang L (2001) Perfect phylogenetic networks with recombination. J Comput Biol 8(1):69–78. https://doi.org/10.1089/106652701300099119

Wen D, Nakhleh L (2017) Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst Biol 67(3):439–457. https://doi.org/10.1093/sysbio/syx085

Wen D, Yu Y, Zhu J et al (2018) Inferring phylogenetic networks using phylonet. Syst Biol 67:735–740. https://doi.org/10.1093/sysbio/syy015

Willson SJ (2007) Restrictions on meaningful phylogenetic networks, Contributed Talk at the EMBO Workshop on Current Challenges and Problems in Phylogenetics, Isaac Newton Inst for Math Sciences, Cambridge, UK

Willson SJ (2009) Properties of normal phylogenetic networks. Bull Math Biol 72(2):340–358. https://doi.org/10.1007/s11538-009-9449-z

Willson SJ (2011) Regular networks can be uniquely constructed from their trees. IEEE/ACM Trans Comput Biol Bioinf 8(3):785–796. https://doi.org/10.1109/tcbb.2010.69

Willson SJ (2012) Tree-average distances on certain phylogenetic networks have their weights uniquely determined. Algorithms Mol Biol. https://doi.org/10.1186/1748-7188-7-13

Willson SJ (2013) Reconstruction of certain phylogenetic networks from their tree-average distances. Bull Math Biol 75(10):1840–1878. https://doi.org/10.1007/s11538-013-9872-z

Xu J, Ané C (2021) Identifiability of local and global features of phylogenetic networks from average distances arXiv e-prints arXiv:2110.11814

Yu Y, Nakhleh L (2015) A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics. https://doi.org/10.1186/1471-2164-16-s10-s10

Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet 8(4):e1002,660. https://doi.org/10.1371/journal.pgen.1002660

Yu Y, Barnett RM, Nakhleh L (2013) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst Biol 62(5):738–751. https://doi.org/10.1093/sysbio/syt037

Yu Y, Dong J, Liu KJ et al (2014) Maximum likelihood inference of reticulate evolutionary histories. Proc Natl Acad Sci 111(46):16,448-16,453. https://doi.org/10.1073/pnas.1407950111

Zhang C, Ogilvie HA, Drummond AJ et al (2017) Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol 35(2):504–517. https://doi.org/10.1093/molbev/msx307

Zhang L (2016) On tree-based phylogenetic networks. J Comput Biol 23(7):553–565. https://doi.org/10.1089/cmb.2015.0228

Zhang L (2019) Clusters, trees, and phylogenetic network classes In: Bioinformatics and Phylogenetics Springer, pp 277–315, https://doi.org/10.1007/978-3-030-10837-3_12

Zhu J, Nakhleh L (2018) Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. Bioinformatics 34(13):i376–i385. https://doi.org/10.1093/bioinformatics/bty295

Zhu J, Liu X, Ogilvie HA et al (2019) A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. Bioinformatics 35(14):i370–i378. https://doi.org/10.1093/bioinformatics/btz359

Zhu S, Degnan JH (2016) Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst Biol. https://doi.org/10.1093/sysbio/syw097

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