Noname manuscript No.
(will be inserted by the editor)

Non-binary treebased unrooted phylogenetic networks and their relations to binary and rooted ones

Mareike Fischer · Michelle Galla · Lina Herbst · Yangjing Long · Kristina Wicke

Received: date / Accepted: date

Abstract Phylogenetic networks are a generalization of phylogenetic trees allowing for the representation of non-treelike evolutionary events such as hybridization. Typically, such networks have been analyzed based on their 'level', i.e. based on the complexity of their 2-edgeconnected components. However, recently the question of how 'treelike' a phylogenetic network is has become the center of attention in various studies. This led to the introduction of *treebased networks*, i.e. networks that can be constructed from a phylogenetic tree, called the *base tree*, by adding additional edges. Here, we initially consider unrooted treebased networks and first revisit some established results known for these networks in case they are binary. We consider them from a more graph-theoretical point of view, before we extend these results to treebased unrooted non-binary networks. While it is known that up to level 4 all binary unrooted networks are treebased, we show that in case of non-binary networks, this result only holds up to level 3. Subsequently, we consider the notion of non-binary *universal treebased networks*, i.e. networks on some taxon set which have every phylogenetic tree on the same taxon set as a base tree. Again, our aim is to understand the unrooted case, but here, we first present a construction for a rooted non-binary universal treebased network, because this easily can be generalized to the unrooted case.

Keywords phylogenetic tree · phylogenetic network · treebased network · universal treebased network · level-k network · Hamiltonian path

Mareike Fischer (✉) E-mail: email@mareikefischer.de
Mareike Fischer, Michelle Galla, Lina Herbst, Kristina Wicke
Institute of Mathematics and Computer Science, University of Greifswald, Germany
Yangjing Long
School of Mathematics and Statistics, Central China Normal University, Wuhan, Hubei, China

1 Introduction

Traditionally, phylogenetic trees have been used in order to represent the evolutionary history of sets of species. However, the evolution of species is not always treelike, in particular if they are subject to reticulation events like hybridization or horizontal gene transfer. In fact, hybridization and horizontal gene transfer occur in a variety of species, ranging from mosquitos (cf. Fontaine et al (2014)), over fish (cf. Cui et al (2013)) to marine mammals (cf. Amaral et al (2014)). Thus, phylogenetic networks have come to the fore as a mathematical generalization of phylogenetic trees, allowing for the representation of non-treelike evolutionary events.

Mathematically, phylogenetic networks are connected graphs that are not necessarily acyclic, as reticulation events may lead to 2-edgeconnected components in the graph. In this regard, Choya et al (2005) introduced the concept of *level-k networks*, which are networks in which in each such component the number of edges which need to be removed to turn them into trees is at most k. In this sense, the level measures the complexity of a network. So the smaller the level, the more 'tree-like' the network.

However, while a network can contain various trees, not all networks are suitable to explain evolution of present-day species to the same extent. This is due to the fact that some networks contain no *support tree* (Francis and Steel (2015)). A support tree is a spanning tree that has the same leaf set as the network, i.e. it represents the evolution of the same present-day species as the given network. Biologically, these networks are most relevant, because the leaves typically represent the species for which one has data (e.g. DNA sequences) and on which the reconstruction of evolution is based.

So given a phylogenetic network, it is of high interest to study its "tree-likeness", and in terms of support trees, this reduces to the question whether the network is merely a tree with additional edges. While Francis and Steel (2015) introduced the concept of treebasedness for binary rooted phylogenetic networks, recently Francis et al (2018) extended it to binary unrooted networks, Jetten and van Iersel (2018) to non-binary rooted networks and Hendriksen (2018) to non-binary unrooted networks.

In the present manuscript, we first focus on unrooted networks and consider both the binary and non-binary case. In particular, we revisit Theorem 1 of Francis et al (2018), which states that all binary unrooted level-k networks with $k \leq 4$ are treebased, and present an alternative proof for this theorem, based on observations from classical graph theory. Moreover, we remark that the example used in Francis et al (2018) to show that binary level-5 networks are not necessarily treebased is unfortunately erroneous (in fact, the level-5 network depicted in Figure 4 of Francis et al (2018) *is* treebased), and present a correct example – i.e. we prove that the corresponding result stated in Francis et al (2018) is still valid. We then generalize these results to unrooted non-binary networks and show that in the non-binary case, level-4 networks are not necessarily treebased. This provides the answer to Question 5.3 posed in Hendriksen (2018), asking whether there are networks of level less than 5 that are not treebased.[1] However, we also show that up to level 3 all non-binary unrooted networks indeed are treebased.

We then turn to the concept of so-called *universal treebased networks*, i.e. networks on some taxon set for which every phylogenetic tree on the same taxon set is a base tree (where a base tree can be obtained from a support tree by suppressing potential degree 2 vertices; cf. Definition 1). Binary universal treebased networks have been introduced and considered by several authors (cf. Francis and Steel (2015); Hayamizu (2016); Zhang (2016); Bordewich and Semple (2016)). However, we focus on the non-binary case, for which such networks have so far not been known, and introduce constructions both for rooted and unrooted non-binary universal treebased networks.

The remainder of this manuscript is organized as follows. In Section 2, we introduce some basic phylogenetic and graph-theoretical concepts and terminology and give an overview of various refinements of treebasedness existing in the literature for unrooted networks. After presenting some general results concerning unrooted networks, we consider treebased unrooted binary and treebased unrooted non-binary networks in more detail. In case of the former we revisit some results obtained by Francis et al (2018) before generalizing these to non-binary networks in Section 3.3. We then turn to the concept of universal treebased networks and show that there exist rooted and unrooted non-binary universal treebased networks on n leaves for all positive integers n. We conclude this manuscript with Section 4, where we discuss our results and indicate possible directions of future research.

2 Preliminaries

2.1 Phylogenetic Concepts

Phylogenetic Networks

Throughout this manuscript we assume that X is a finite set (e.g. of taxa or species) with $|X| \geq 1$. An *unrooted phylogenetic network* N^u (on X) is a connected, simple graph $G = (V, E)$ with $X \subseteq V$ and no vertices of degree 2, where the set of degree 1 vertices (referred to as the *leaves* or *taxa* of the network) is bijectively labelled by and thus identified with X. Such an unrooted network is called *unrooted binary* if every non-leaf vertex $u \in V \setminus X$ has degree 3. In the following, we denote by $\mathring{E}$ the set of inner edges of N^u, i.e. those edges that are *not* incident to a leaf.

A *rooted phylogenetic network* N^r on X is a directed, acyclic graph that contains a single root node of indegree 0 and outdegree at least 1 as well as vertices of indegree 1 and outdegree 0 (called *leaves*), which are bijectively labeled by X, and may additionally contain the following types of vertices:

[1] Note that our definition of treebased networks corresponds to the definition of *loosely treebased* networks in Hendriksen (2018) (see Section 2.1 in the present manuscript), so the question is posed there in a slightly different way.

– vertices of outdegree 1 and indegree 2 or more (called *reticulations*);
– vertices of indegree 1 and outdegree 2 or more (called *tree-vertices*).

For technical reasons, if $|X| = 1$, we allow N^r to consist of a single leaf (which is then at the same time considered to be the root). If the root has outdegree 2, all reticulations are of indegree exactly 2 and additionally all tree-vertices have outdegree exactly 2, the network is called *rooted binary*. When we refer to a network N^u in the following, we always mean an unrooted network, and when we refer to a network N^r, we mean a rooted network. Whenever the rooting is irrelevant for our considerations, we speak of a network N. Moreover, note that an (un)rooted *phylogenetic tree* is an (un)rooted phylogenetic network whose underlying graph structure is a tree.

Edge subdivision and vertex suppression

In order to summarize the various definitions of treebased networks used in the literature later on, we now need to introduce the concepts of *subdividing an edge* and *suppressing a vertex*. Therefore, let N be a phylogenetic network with some edge $e = \{u, v\}$ (note that if N is rooted, e will be directed away from the root – in this case, the following operations all have to keep the direction of the edges). Then, we say that we *subdivide* e by deleting e, adding a new vertex w and adding the edges $\{u, w\}$ and $\{w, v\}$. The new degree 2 vertex w is sometimes also referred to as an *attachment point*. Note that we often also refer to the vertex incident to a leaf x as the attachment point of x, even if this vertex has degree higher than three.

In contrast to adding vertices to a network, given a degree 2 vertex w with adjacent vertices u and v, by *suppressing* w we mean deleting w and its two incident edges $\{u, w\}$ and $\{w, v\}$ and adding a new edge $\{u, v\}$. Note that the resulting graph can be a multigraph (cf. Figure 1). If this is the case, we additionally delete all parallel edges (except for one), such that we obtain a simple graph again and suppress the resulting degree 2 vertices (if any). The reason why we define this is that we will later reduce networks to simpler networks with fewer leaves. However, in doing so it might occur that the resulting graph is not a simple graph anymore and thus *not* a network in the sense of our definition as it might be a multigraph (cf. Figure 1). Thus, we prevent this scenario by suppressing duplicate/parallel edges.

Treebased networks

Edge subdivisions and attachment points play a fundamental role in the concept of treebased networks. When treebasedness was first introduced for binary rooted phylogenetic networks by Francis and Steel (2015), treebased networks were constructed from binary rooted phylogenetic trees by subdividing edges and adding new edges between such pairs of attachment points. However, treebased networks (whether they are rooted or unrooted, binary or non-binary) can be characterized as follows. For technical purposes (e.g. for Lemma 3 in the Appendix), we first define treebasedness for (multi)graphs.

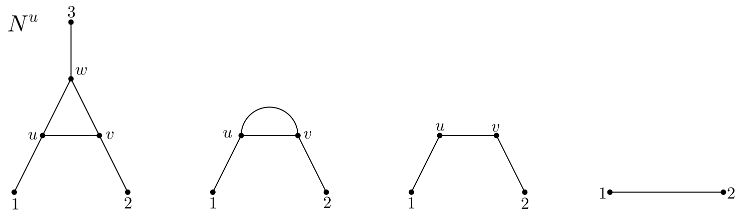

Fig. 1 Unrooted phylogenetic network N^u on three leaves. Deleting leaf 3 and suppressing the resulting degree 2 vertex w, i.e. deleting w and its incident edges $\{u, w\}$ and $\{w, v\}$ and adding a new edge $\{u, v\}$ results in a multigraph, because u and v are already connected by an edge. Thus, subsequently one copy of edge $\{u, v\}$ is deleted and the resulting degree 2 vertices u and v are suppressed.

Definition 1 Let $G = (V, E)$ be a connected (multi)graph without loops and with leaf set V^1, i.e. $V^1 = \{v \in V : deg(v) \leq 1\}$. G is called *treebased* if there is a spanning tree $T = (V, E')$ in G (with $E' \subseteq E$) whose leaf set is equal to V^1. T is then called a *support tree* (for G). Moreover, the tree T' that can be obtained from T by suppressing potential degree 2 vertices is called a *base tree* (for G). If G is a phylogenetic network with leaf set X, i.e. $G = N = (V, E)$ and $V^1 = X$, and G is treebased, then we call N a *treebased network* with support tree T (and base tree T').

Note that the existence of a support tree T' for G implies the existence of a base tree T' for G.

Moreover, note that while we consider only one kind of treebasedness for unrooted non-binary networks, one might want to distinguish between several forms of treebased networks (cf. Hendriksen (2018)). This is due to the fact that while binary treebased networks can only be constructed from a base tree by subdividing edges and adding new edges between pairs of attachment points (in order to keep the network binary), there are more possibilities to construct non-binary treebased networks. We can additionally add edges between attachment points and original vertices of the tree or between two vertices in the base tree. Moreover, we may have more than one additional edge incident to an attachment point.

Thus, different forms of treebased networks are defined in Hendriksen (2018), namely *loosely treebased* networks, *treebased* networks and *strictly treebased* networks, where the notion of loosely treebased networks corresponds to our understanding of treebased networks. In particular, all other definitions of treebasedness given in that manuscript are special cases of our concept, which is why in the present manuscript, we stick to the more general definition.

Cut edges/vertices, blobs, level-k and proper networks

We will see in subsequent sections that it is often useful to decompose a phylogenetic network into simpler pieces, which can then be analyzed individually. Therefore, recall the following definitions from Gambette et al (2012). Let N^u

be an unrooted network. A *cut edge*, or *bridge*, of N^u is an edge e whose removal disconnects the graph, i.e. an edge e such that $N^u - e$ is disconnected. Similarly, we call a vertex v a *cut vertex*, if deleting v and all its incident edges disconnects the graph.

A cut edge is called *trivial* if one of the connected components induced by the removal of the cut edge is a single vertex (which must necessarily be a leaf). We call N^u a *simple* network if all of its cut edges are trivial. A *blob* in a network is a maximal connected subgraph that has no cut edge. If a blob consists only of one vertex, we call the blob *trivial.* Note that this implies that we can consider a network as a "tree" with blobs as vertices (cf. Figure 2). The idea of "blobbed trees" has already been introduced for rooted phylogenetic networks in Gusfield and Bansal (2005) and we use it for unrooted ones in the following. Moreover, note that while in a binary network it can be easily seen that a blob not only contains no cut edges, but it contains no cut vertices either (because in binary networks, every cut vertex is incident to a cut edge, and these are excluded from blobs, cf. Lemma 8 in the Appendix), in the non-binary setting, a blob may contain cut vertices. An example for such a blob can be seen in Figure 2.

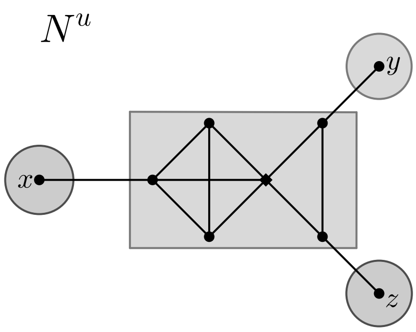

Fig. 2 Unrooted non-binary phylogenetic network N^u on taxon set $X = \{x, y, z\}$. The gray areas correspond to the blobs of N^u. Note that N^u consists of three trivial blobs and one non-trivial blob and this non-trivial blob contains a cut vertex (depicted as a square vertex). Moreover, note that the cut edges and blobs in N^u induce a "tree structure", i.e. N^u can be considered as a tree with blobs as vertices.

Recall that a binary network is called proper if every cut edge induces a *split* of X, that is a bipartition of X into two non-empty subsets. Here, we call a network proper if the removal of any cut edge or cut vertex present in the network leads to connected components containing at least one leaf each. Note that this definition of proper networks generalizes the one given in (Francis et al 2018) to the non-binary case. There, only cut edges were considered for proper networks. However, the alteration of the definition in the non-binary case is needed in order to exclude some networks which cannot be treebased. In particular, we exclude networks where all leaves are attached to the same interior vertex.

Remark 1 A network with $|V| > |X| \geq 1$ such that all leaves are attached to the same interior vertex cannot be treebased. This is due to the fact that any spanning tree will induce additional leaves that are not part of X, i.e. no spanning tree of the network is a support tree (cf. Figure 3). Moreover, note that a network with $|X| = 1$ and $|V| > |X|$ cannot be proper, because if $|X| = 1$ no cut edge or cut vertex can induce a partition of the taxon set.

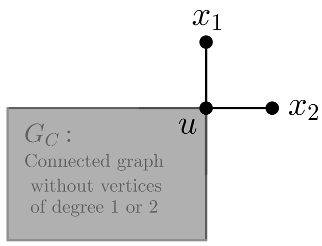

Fig. 3 Network N^u with $|V| > |X| \geq 1$ such that all leaves, i.e. x_1 and x_2, are attached to the same interior vertex u with $deg(u) \geq 4$. Any support tree for N^u would have to cover all vertices of G_C, as well as u, x_1 and x_2. Here, G_C is a connected graph without vertices of degree 1 or 2 (in particular G_C is not a tree). Thus, any spanning tree would induce at least one additional leaf besides x_1 and x_2. In particular, it would not be a support tree and thus N^u cannot be treebased.

Given a network N^u on X and an integer $k \geq 0$, we call N^u a *level-k* network if at most k edges have to be removed from each blob of N^u to obtain a tree.

Moreover, following Francis et al (2018), given a network N^u and a blob B in N^u, we define a simple network B_{N^u} by taking the union of B and all cut edges in N^u incident with vertices in B, where the leaf set of B_{N^u} is simply the set of end vertices of these cut edges that are not already a vertex in B.

2.2 Graph-theoretical concepts

Besides the phylogenetic terminology, we need to introduce some basic concepts from classical graph theory before we can proceed with analyzing binary and non-binary treebased unrooted networks. In particular, we need the notion of *cubic graphs* and *Hamiltonian paths/cycles.*

A *cubic graph* is a graph $G = (V, E)$ such that all vertices have degree 3. Applying the so-called handshaking lemma (Harris et al 2000, Theorem 1.1), which states that $\sum_{v \in V} deg(v) = 2|E|$, we have that three times the number of vertices equals two times the number of edges in any cubic graph, i.e. $3|V| = 2|E|$. In particular, this implies that a cubic graph always has an even number of vertices. Similarly, the handshaking lemma also implies that a binary phylogenetic network always has an even number of vertices. We will need these properties later on (e.g. in the proof of Lemma 6).

Another well-known graph theoretical concept we wish to introduce here is a *Hamiltonian path*. A Hamiltonian path is simply a path in a graph that visits each vertex exactly once. If this path is a cycle, the Hamiltonian path is called a *Hamiltonian cycle*. A graph that contains a Hamiltonian cycle is called a *Hamiltonian graph*. As has been noted in Francis et al (2018) and as we will elaborate in the present manuscript, Hamiltonian paths play an important role concerning the treebasedness of phylogenetic networks.

2.3 The leaf connecting procedure

Before we can present the main results of this manuscript, we need to recall a concept which turns a phylogenetic network into a graph without leaves. This concept is the so-called *leaf connecting* procedure, which was recently introduced in Fischer et al (2018).

Suppose N^u is a phylogenetic network that is not a tree[2] on taxon set X with $|X| \geq 2$, i.e. N^u contains at least two leaves. The aim of the leaf connecting procedure is to turn N^u into a graph without leaves, i.e. without degree 1 vertices. This is achieved in the following way (cf. Fischer et al (2018)):

- Pre-processing: As long as there exists an internal vertex u of N^u such that there is more than one leaf attached to u delete all of them but one. Additionally, suppress potentially resulting degree 2 vertices. In the following, we denote the resulting reduced taxon set of N^u by X^r.[3]
- Leaf connecting:
  - Select two leaves x_1 and x_2 (if they exist) and denote their respective attachment points by u_1 and u_2, respectively. Now, delete x_1 and x_2 as well as their incident edges (i.e. the edges $\{x_1, u_1\}$ and $\{x_2, u_2\}$) and connect their attachment points by introducing a new edge $e := \{u_1, u_2\}$. If this edge is a parallel edge, i.e. if there is another edge $\widetilde{e}$ connecting u_1 and u_2, add two more nodes a and b and replace e by two new edges, namely $e_1 := \{u_1, a\}$ and $e_2 := \{a, u_2\}$. Similarly, replace $\widetilde{e}$ by two new edges, namely $\widetilde{e}_1 := \{u_1, b\}$ and $\widetilde{e}_2 := \{b, u_2\}$. Last, add a new edge $\{a, b\}$.
    Repeat this procedure until no pair of leaves is left.
  - If there is one more leaf x left in the end, remove x and, if its attachment point u then has degree 2, suppress u. If this results in two parallel edges, say $e = \{y, z\}$ and $\widetilde{e} = \{y, z\}$, re-introduce u on edge e and add a new vertex a to the graph, delete $\widetilde{e}$ and introduce two new edges $\widetilde{e}_1 := \{y, a\}$ and $\widetilde{e}_2 := \{a, z\}$. Last, add an edge $\{u, a\}$.

Note that the order in which the leaves are connected may have an impact on the resulting graph. In general, if $|X| > 2$, there might be more than one graph that can be constructed from N^u by the leaf connecting procedure. We denote

[2] Note that for a tree, the pre-processing step would always result in a single edge.

[3] Note that this pre-processing step may have to be repeated several times, but does not influence whether a network is treebased or not (cf. Fischer et al (2018)).

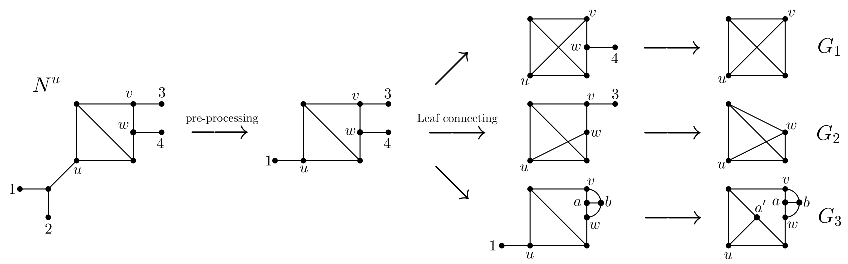

Fig. 4 Network N^u on taxon set $X = \{1, 2, 3, 4\}$ and the graphs resulting from the leaf connecting procedure. In the pre-processing step leaf 2 is removed and the resulting degree 2 vertex is suppressed. Then, first a pair of leaves is chosen and removed from the network, their attachment points are connected and, if necessary, new vertices and edges are introduced. Lastly, the remaining single leaf is removed, and again, if necessary new vertices and edges are introduced. This results in three graphs: G_1, G_2 and G_3. Note, however, that G_1 and G_2 are isomorphic. Thus, $\mathcal{LCON}(N^u)$ consists only of G_1 and G_3. Moreover, note that even though new vertices were introduced to obtain G_3, the total number of vertices in the graph did not increase, i.e. G_3 contains only as many vertices as N^u after the pre-processing step.

the set of all these graphs by $\mathcal{LCON}(N^u)$. An illustration of this concept is given in Figure 4.

3 Results

As a first result, we now present an alternative proof for Theorem 1 in Francis et al (2018), stating that all unrooted binary proper level-4 networks are treebased. We then provide a similar theorem for non-binary networks, stating that all unrooted non-binary proper level-3 networks are treebased. Before we can turn our attention to these extensions of previous results, we have to state some preliminary results from the literature again and show that they also hold for non-binary networks.

3.1 Basic results

In this section, we first state some basic results that will be needed throughout this manuscript. Some of them are mere extensions of the results presented in Francis et al (2018), but here we state them for non-binary networks. We checked all proofs carefully to make sure that higher node degrees do not cause them to fail.

We start by considering the following lemma by Francis et al (2018), which states a close relationship for *binary* networks between the properties of being treebased and being proper.

Lemma 1 (Francis et al (2018)) *If N^u is a binary unrooted treebased network, then every cut edge of N^u induces a split of X, i.e. N^u is proper.*

Note that in the non-binary case, the absence of cut edges need not imply that the network is proper, because it might still contain cut vertices. So we now generalize this lemma to non-binary networks.

Lemma 2 *If N^u is an unrooted treebased network (binary or not), then N^u is proper.*

Note that this implies that a non-proper network cannot be treebased.

Proof It can be easily seen (cf. Lemma 8 in the Appendix) that if N^u is binary and has no cut edge, it also does not have a cut vertex. So for binary networks, our assumptions are identical to those of Francis et al (2018). Anyway, assume N^u is treebased and contains either a cut edge e or a cut vertex v that does not subdivide the taxon set. As N^u is treebased, it contains a spanning tree T whose leaf set coincides with X. As T is a spanning tree of N^u, its vertex set contains any possible cut vertex v, and if e is a cut edge, it must also occur in the spanning tree (otherwise T could not be connected). So removing v or e, respectively, would separate T into at least two connected components, at least one of which does not contain a taxon (as by assumption neither e nor v subdivide the taxon set). So let us consider such a component T_c. Clearly, T_c forms a subtree of T. Moreover, T_c must contain at least one edge. Otherwise, T_c would consist only of one vertex, but this vertex is a leaf of T. As T_c does not contain any leaf labelled by X and as T is a tree with only such leaves, this would be a contradiction. So T_c contains at least one edge and thus at least two vertices, only one of which is adjacent to v or incident to e in T, respectively. This node is one leaf of T_c, but as T_c has an edge, it can be easily seen that T_c must contain at least one more leaf. Again, this contradicts the fact that T_c contains no element of X and all but one leaves of T_c are also leaves of T. This completes the proof.

We next state some general results concerning the decomposition of unrooted networks into simpler parts and the number of vertices in a blob.

Francis et al (2018) prove a decomposition theorem for treebased unrooted binary networks based on the simple networks B_{N^u} associated with blobs introduced above. This can directly be generalized to non-binary networks and we have the following statement.

Proposition 1 *Suppose N^u is an unrooted network. Then N^u is treebased if and only if B_{N^u} is treebased for every blob B in N^u.*

Our proof of this proposition is very similar to the proof of Proposition 1 in Francis et al (2018) and Proposition 2.9 in Hendriksen (2018). However, as we explicitly allow blobs to contain cut vertices, we shortly outline the proof in the Appendix.

Roughly speaking, Proposition 1 states that it is sufficient to analyze all blobs of an unrooted network individually in order to decide whether a network N^u is treebased or not.

As we will show subsequently, it also often makes sense to consider (simple) networks with only two leaves. Therefore, we first recall a useful observation of Francis et al (2018) that is, again, also valid for non-binary networks.

Lemma 3 *Let N^u be a network on X with $|X| \geq 2$. For any $x \in X$ let $N^u - x$ denote the network obtained from N^u by deleting x and its incident edge, and suppressing the potentially resulting degree 2 vertex. Then, if $N^u - x$ is treebased, so is N^u.*

The proof of this lemma is very similar to the proof of Lemma 3 in Francis et al (2018). However, as we are not only considering binary networks, but also non-binary ones, and as we – unlike Francis et al (2018) – explicitly consider the case of parallel edges, we give the proof again in the Appendix. In any case it should be noted that $N^u - x$ might not be a phylogenetic network as it might contain parallel edges. This is why in Definition 1 treebasedness is also defined for (multi)graphs.

Remark 2 Note that the converse does not necessarily hold, i.e. if N^u is treebased, $N^u - x$ does not necessarily have to be treebased, neither if N^u is binary nor if N^u is non-binary. To see this, consider the network depicted in Figure 5. Note that this example is extreme in the following sense: It shows that even if a treebased network N^u is binary and contains only one blob and even if N^u does not contain any cut vertices and no cut edges other than the ones incident to leaves, it might not be possible to remove any leaf and suppress the resulting degree 2 vertex without losing the treebasedness. In particular, there exists no pair of two leaves in this network such that it is still treebased.

Note that this extreme example is based on the graph shown in Figure 6, which we found in Zamfirescu (1976). There, it is proven that any longest path in this graph of 12 vertices has length 10. This immediately implies that in our example depicted in Figure 5, in order for all nodes in each of these two components to be covered, all three 'exits' of each of them need to be used. One of them is just the connection between both components, so this can easily be covered by any tree, but as both of the triangular components need two more exits, it is obvious that all four leaves are needed. Otherwise, no spanning tree would have the same leaf set as the network, so the network would not be treebased anymore.

We now present another observation that will be useful in the following, namely that the number of vertices in a non-trivial blob incident with at most two cut edges is bounded by $2k$ in a level-k network, where $k \geq 2$.

Lemma 4 *Let N^u be an unrooted level-k network (not necessarily binary) with $k \geq 2$ and let B be a non-trivial blob and assume that there are at most two cut edges in N^u incident to B. Let $n = |V(B)|$ denote the number of vertices in B. Then,*

$$3 < n \leq 2k.$$

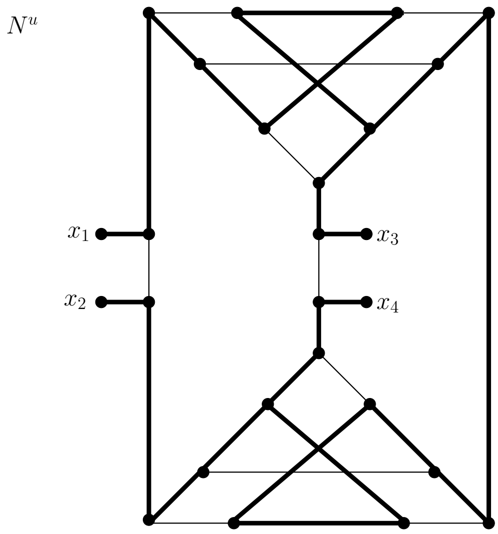

Fig. 5 Binary treebased unrooted phylogenetic network N^u on $X = \{x_1, x_2, x_3, x_4\}$. The corresponding support tree is highlighted in bold. $N^u - x_i$ is not treebased for $i = 1, \ldots, 4$, because there is no spanning tree in $N^u - x_i$ whose leaf set is equal to $X \setminus \{x_i\}$.

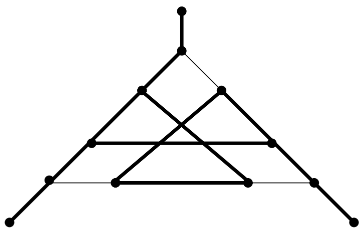

Fig. 6 Graph from Zamfirescu (1976) on which our construction of the example in Figure 5 is based. There, it is proven that any longest path in this graph of 12 vertices has length 10. However, if we consider this graph as a treebased phylogenetic network (the support tree is highlighted in bold), removing a leaf vertex and suppressing the resulting degree 2 node turns this again into a treebased phylogenetic network.

Proof We first show that $n > 3$. Let $m = |E(B)|$ denote the number of edges in B. As we assumed that there are at most two cut edges in N^u incident to B, B can have at most two degree 2 vertices. All other vertices are of degree at least 3 (they can be higher, as we do not assume that the network is binary).

Thus, using the handshaking lemma, we have

$$\begin{aligned} m &= \frac{1}{2} \sum_{v \in V(B)} deg(v) \\ &\geq \frac{1}{2}\big(2 \cdot 2 + (n-2) \cdot 3\big) \\ &\geq \frac{1}{2}\big(3n-2\big) \\ &\geq \frac{3}{2}n - 1. \end{aligned}$$

On the other hand, as we are considering simple graphs, B can have at most $\binom{n}{2}$ edges. Thus,

$$\frac{3}{2}n - 1 \leq m \leq \binom{n}{2}.$$

This implies in particular

$$\frac{3}{2}n - 1 \leq \frac{n(n-1)}{2},$$

which leads to

$$n^2 - 4n + 2 \geq 0.$$

This inequality is fulfilled if $n \geq 2 + \sqrt{2}$ or if $n \leq 2 - \sqrt{2}$. As n is a positive integer, we conclude $n > 3$.

Now, we show that $n \leq 2k$ for $k \geq 2$. Recall that an unrooted tree on n vertices has $n-1$ edges. As N^u is a level-k network, at most k edges have to be removed from B to obtain a tree. Thus, $m - (n-1) \leq k$. In other words, $m \leq k + n - 1$. Using the lower bound for m from above, we derive

$$\frac{3}{2}n - 1 \leq m \leq k + n - 1,$$

and thus in particular $n \leq 2k$. In total, we have

$$3 < n \leq 2k,$$

which completes the proof.

We end this section by establishing a relationship between cubic graphs and Hamiltonian paths, which will be needed subsequently, e.g. in the proof of Lemma 6.

Lemma 5 *Let $G = (V, E)$ be a cubic graph with $|V| \leq 8$. Then, G contains a Hamiltonian path from u to v for all edges $e = \{u, v\} \in E$. In other words, G is Hamiltonian and for every edge $e \in E$, there is a Hamiltonian cycle of G which contains e.*

Proof As the number of nodes in a cubic graph is even by the handshaking lemma and as the smallest cubic graph contains four vertices, we only need to consider the one cubic graph with four vertices, the two cubic graphs with six vertices and the five cubic graphs with eight vertices (all these graphs are depicted in the Appendix in Figures 15, 16 and 17). The fact that they are all Hamiltonian can be found in the literature (cf. Bussemaker et al (1976)), but can also easily be verified by considering all mentioned 8 graphs. We also verified the fact that each edge is contained in at least one Hamiltonian cycle exhaustively. This completes the proof.

3.2 Binary treebased unrooted networks

The main aim of this section is to provide an alternative proof of Theorem 1 of Francis et al (2018) stating that all proper binary unrooted level-4 networks are treebased, while networks of level greater than 4 need not be treebased. While the proof in Francis et al (2018) strongly depends on so-called binary level-k generators of phylogenetic networks, our proof is more basic, as it only uses elementary graph theory.

Theorem 1 (Francis et al (2018)) *All proper binary unrooted level-k networks with $k \leq 4$ are treebased. Moreover, networks of level greater than 4 need not be treebased.*

Before we can proceed with the proof of the first part of Theorem 1, we briefly outline our proof strategy: We first show that it is sufficient to consider the non-trivial blobs of the binary unrooted network N^u and show that all such networks with only two leaves have a close relationship with cubic graphs via the $\mathcal{LCON}(N^u)$ construction, which in this case, where N^u has only two leaves, contains a unique cubic graph which we will call $G(N^u)$. Moreover, we show if such a network is treebased, $G(N^u)$ needs to have a Hamiltonian cycle and thus $N^u - X$ must contain a Hamiltonian path between the two attachment points of its leaves.

Note that if such a network with two leaves is treebased, we can simply attach more leaves by Lemma 3 without losing the treebasedness. So any network that is not treebased can in particular *not* contain a subnetwork with two leaves which are connected by a Hamiltonian path. We can thus investigate Hamiltonian paths in cubic graphs a bit more in-depth and use a simple counting argument based on Lemma 4 to show that the number of nodes necessary to avoid a Hamiltonian path induces a level of $k \geq 5$.

We begin with establishing the required relationship between unrooted binary phylogenetic networks with two leaves and cubic graphs:

Observation 1 *Let N^u be an unrooted binary phylogenetic network on leaf set X with $|X| = 2$ and with $\mathring{E} \neq \emptyset$. Without loss of generality, let $X = \{x, y\}$ and denote the nodes adjacent to x and y by u and v, respectively. Then, $\mathcal{LCON}(N^u)$ contains precisely one graph $G(N^u)$, and this graph is cubic.*

Moreover, by construction the number of vertices of $G(N^u)$ is bounded by the number of vertices of N^u, i.e. we have $|V(G(N^u))| \leq |V(N^u)|$, as in each step, a leaf and its attachment point get deleted, and at most two new vertices get introduced (if otherwise we would have a parallel edge).

Note that the construction of $G(N^u)$ does not require the suppression or deletion of nodes and so, as we require $\mathring{E} \neq \emptyset$, N^u cannot simply be a tree consisting of two nodes connected by a single edge. This implies that the resulting graph $G(N^u)$ is *always* cubic. Moreover, we state the following crucial proposition.

Proposition 2 *Let N^u and $G(N^u)$ be as described in Observation 1. Then, N^u is treebased if and only if $G(N^u)$ contains a Hamiltonian cycle using at least one edge of $G(N^u)$ that is not contained in N^u.*

Before we proceed with the proof of this proposition, note that if $G(N^u)$ contains a Hamiltonian cycle that uses at least one of the new edges (note that there is only one new edge if the introduction of the edge connecting the attachment points of both leaves did not lead to parallel edges), this implies that we could delete these edges from the cycle and thus get a Hamiltonian path from one attachment point to the other one. So in fact, if $G(N^u)$ is Hamiltonian and has a cycle which uses such a new edge, this implies that N^u has a path from x to y visiting all inner nodes of N^u. This path would therefore be a support tree of N^u.

Let us now formally prove the proposition.

Proof Let N^u be an unrooted binary phylogenetic network on leaf set $X = \{x, y\}$ and with $|\mathring{E}| \neq \emptyset$. Let u and v denote the vertices adjacent to x and y, respectively. Consider the graph $G(N^u)$ obtained from the leaf connecting procedure. Note that $G(N^u)$ might contain two new vertices, a and b, if in the construction of $G(N^u)$ parallel edges occured (cf. description of the leaf connecting procedure in Section 2.3). First, assume that $G(N^u)$ contains a Hamiltonian cycle using at least one edge of $G(N^u)$ that is not contained in N^u. We now distinguish between two cases:

- $G(N^u)$ does not contain new vertices a and b:
  As $G(N^u)$ contains a Hamiltonian cycle using the edge $\{u, v\}$ that is not contained in N^u, this implies that there is a Hamiltonian path from u to v in $G(N^u)$. We extend this path to a support tree of N^u by adding x and y as well as the edges $\{x, u\}$ and $\{y, v\}$.
- $G(N^u)$ contains new vertices a and b:
  As $G(N^u)$ contains a Hamiltonian cycle using either the edges $\{u, a\}$, $\{a, b\}$, $\{b, v\}$ or $\{u, b\}$, $\{a, b\}$, $\{a, v\}$, deleting the edge $\{a, b\}$ and suppressing a and b as well as one copy of the parallel edges $e = \tilde{e} = \{u, v\}$ results in a Hamiltonian path from u to v in this modified network. As before, we can now extend this path to a support tree of N^u by adding x and y as well as the edges $\{x, u\}$ and $\{y, v\}$. This completes the first direction of the proof.

On the other hand, if N^u is treebased, this implies that there is a spanning tree whose leaf set is precisely $X = \{x, y\}$. Again, we distinguish between two cases:

- If the construction of $G(N^u)$ does not require adding a and b, it immediately follows that the support tree of N^u leads to a Hamiltonian cycle in $G(N^u)$, as we can go from u to v both via the support tree, which covers all vertices of $G(N^u)$, or via the new edge $\{u, v\}$. Thus, we have a Hamiltonian cycle which uses a new edge.
- If the procedure requires the introduction of a and b, we can obtain a Hamiltonian cycle in $G(N^u)$ by extending the support tree of N^u by the edges $\{u, a\}$, $\{a, b\}$ and $\{b, v\}$ and removing the edges $\{x, u\}$ and $\{v, y\}$ from the support tree. In particular, this cycle uses three new edges.

This completes the proof.

Proposition 2 together with Lemma 5 lead to the following statement, the proof of which can be found in the Appendix.

Lemma 6 *Any minimal proper binary unrooted non-treebased network has 12 vertices (10 internal vertices and 2 leaves), and this bound is tight, i.e. there are proper binary unrooted non-treebased networks with precisely 12 vertices.*

We are now in the position to prove Theorem 1. The first part of the proof is identical to the proof presented in Francis et al (2018), but in the second part we use a different argument, in particular, we do not use so-called level-k generators.

Proof (Theorem 1) We have to show that all proper level-0,1,2,3 and 4 networks are treebased. Note that if X contains only one leaf, either the underlying network consists only of a single node (and is thus trivially treebased) or it is not proper (see Remark 1). This is why we now consider only networks with $|X| \geq 2$. As in Francis et al (2018), we now use the fact that by Proposition 1 it suffices to prove that every simple, level-k network with $k \leq 4$ and two leaves is treebased. This is due to the fact that we can decompose N^u into a collection of simple networks B_{N^u} associated with the non-trivial blobs in N^u each having at least 2 leaves, and if each of these simple networks is treebased, then so is N^u by Proposition 1 (note that trivial blobs are trivially treebased, which is why we only consider non-trivial blobs and their associated networks B_{N^u}). Moreover, if we remove all but 2 leaves for each of these simple networks B_{N^u} and obtain a treebased network, then B_{N^u} must have been treebased due to Lemma 3. Now, for $k = 0$, the statement trivially holds, as a level-0 network is a tree. In particular, a tree is treebased. For $k = 1$, we know that at most one edge has to be removed from each non-trivial blob of N^u to obtain a tree. Removing at most one edge from each such non-trivial blob, however, cannot induce any new leaves (if a former inner vertex of some blob became a leaf after removing one edge from this blob, this would imply that the vertex was a former degree 2 vertex; however, phylogenetic networks do not contain degree 2 vertices). Thus, the tree that we obtain from removing at most one edge of each non-trivial blob in a level-1 network can directly be considered a support tree of N^u. This implies that level-1 networks are always treebased. Therefore, let us now consider the case $k \geq 2$. Due to Lemma 6 we know that any minimal proper non-treebased network has at least 12 vertices, 2 of which are leaves. This implies that any B_{N^u} that is not treebased has at least 12 vertices.[4] Additionally, by Lemma 4 we know that the number of vertices in a non-trivial blob incident to two cut edges (and thus corresponding to a simple network B_{N^u} with two leaves) in a level-k network is bounded from above by $2k$. Now, as we assume that all B_{N^u} have exactly two leaves, any B_{N^u} that is not treebased has to correspond to a non-trivial blob B with at least 10 nodes. Thus,

$$10 \leq n \leq 2k.$$

This immediately implies that $k \geq 5$, thus there cannot be a binary non-treebased unrooted level-k network with $k \leq 4$.

To prove the last statement of the theorem, consider either one of the two level-5 networks depicted in Figure 7. These networks can be seen to not be treebased as follows. If they were treebased, then there would be a path from x to y visiting every vertex exactly once. Any such path must begin with the edge $\{x, a\}$ and end with the edge $\{j, y\}$. Now, for the network at the top of Figure 7, it is straightforward to see that every path visiting both the vertices at the top (i.e. vertices b, c, d, e) and the vertices at the bottom (i.e. vertices f, g, h, i) must visit either vertex a or j twice, which is a contradiction. A similar argument shows that also the second network depicted in Figure 7 cannot be treebased. This completes the proof.

Note that 12 vertices are minimal, as we already know by Lemma 6 that 12 vertices are required for a network to be non-treebased. Actually, the two networks depicted in Figure 7 are the only non-treebased level-5 networks with 12 vertices. We verified this by an exhaustive search with Mathematica Inc. (2017), which was conducted in the following way: First, we obtained a list of all connected simple graphs with 10 vertices (11716571 in total) from the "House of Graphs" database (cf. Brinkmann et al (2013)). These were then analyzed for potential binary networks with 10 inner vertices and 2 leaves by checking whether they contained exactly 8 vertices of degree 3 and 2 vertices of degree 2 (which we called u and v and to which we subsequently attached leaves) using the Mathematica function `VertexDegree[·]`. The resulting 113 graphs were analyzed for treebasedness in the following way:

- We attached one leaf to each of the two degree 2 vertices u and v.
- The two leaves were then connected according to the leaf connecting procedure described in Section 2.3 (note that as there are only 2 leaves, $\mathcal{LCON}(N^u)$ contains only one graph).

[4] Note that by Lemma 8 from the Appendix, as blobs do not contain cut edges and as we are in the binary case, there can also be no cut vertices (as these are always incident to cut edges in this case). So B_{N^u} must be proper, which indeed justifies the usage of Lemma 6.

– We then used the Mathematica function `FindHamiltonianCycle[`$\mathcal{LCON}(N^u)$`, All]` to find all Hamiltonian cycles of $\mathcal{LCON}(N^u)$. We then checked whether one of them used at least one edge in $E(\mathcal{LCON}(N^u)) \setminus E(N^u)$. If so, this Hamiltonian cycle corresponds to a Hamiltonian path from u to v in $\mathcal{LCON}(N^u)$, meaning that N^u is treebased (cf. Proposition 2).

This left us with 10 non-treebased networks, which were then filtered for proper networks. It turned out that 8 of them were not proper, i.e. there are exactly 2 proper binary phylogenetic networks with 12 vertices, both of which are level-5 networks. They are the ones depicted in Figure 7.

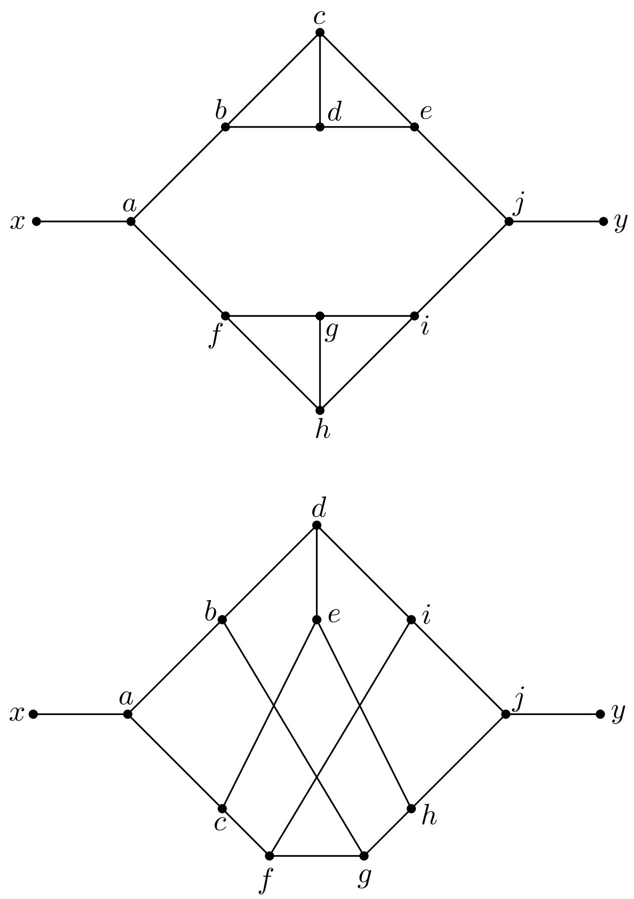

Fig. 7 The only two proper level-5 networks on $X = \{x, y\}$ that are not treebased.

Remark 3 Note that the level-5 network on $X = \{x, y\}$ depicted in Figure 8 is used in Francis et al (2018) to show that level-5 networks need not be treebased. However, this network in fact *is* treebased as there exists a spanning tree in N^u whose leaf set is equal to X. However, as our examples in Figure 7 show, the result presented in Francis et al (2018) is nonetheless valid.

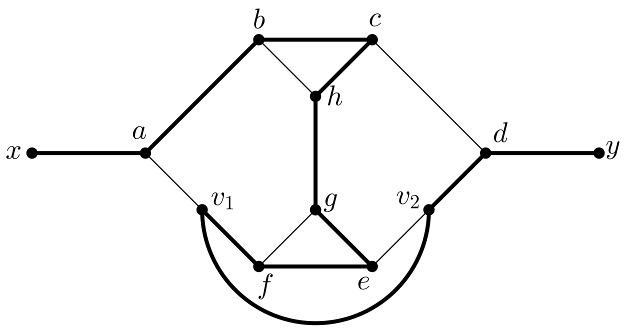

Fig. 8 Level-5 network on $\{x, y\}$ claimed to not be treebased in Francis et al (2018). However, this network is treebased, because there is a spanning tree in N^u whose leaf set is equal to $\{x, y\}$ (depicted in bold). This spanning tree is a path between x and y consisting of the following edges: $\{x, a\}$, $\{a, b\}$, $\{b, c\}$, $\{c, h\}$, $\{h, g\}$, $\{g, e\}$, $\{e, f\}$, $\{f, v_1\}$, $\{v_1, v_2\}$, $\{v_2, d\}$ and $\{d, y\}$.

3.3 Non-binary treebased unrooted networks

The main aim of this section is to generalize Theorem 1 to non-binary networks.

Theorem 2 *All proper non-binary unrooted level-k networks with $k \leq 3$ are treebased. Moreover, non-binary unrooted networks of level greater than 3 need not be treebased.*

In order to prove Theorem 2, we need the following lemma.

Lemma 7 *Any minimal proper non-binary unrooted non-treebased network has 6 internal vertices and 2 leaves, i.e. 8 vertices in total.*

Proof In order to show that any minimal proper non-binary unrooted non-treebased network has 6 internal vertices and 2 leaves, i.e. 8 vertices in total, we need to show that all proper non-binary unrooted networks on less than 8 vertices are treebased. To show this, we performed an exhaustive search using Mathematica (Inc. 2017). In the following we explain the details of this exhaustive search.

First of all, we generated *all* graphs with up to 7 vertices using the Mathematica function `GraphData[·]`. Note that in general, `GraphData[`n`]` does not generate all graphs on n vertices, but only a subset of them. However, for

$n = 1, \ldots, 7$ *all* graphs are generated. We verified this by comparing the numbers of graphs generated by `GraphData[`n`]` with the number of all graphs on n vertices (see for example `https://oeis.org/A000088` for the number of graphs on n vertices). For $n = 1, \ldots, 7$ the numbers of graphs generated by `GraphData[`n`]` coincided with the numbers of all graphs on n vertices, respectively. Thus, `GraphData[`n`]` was used to exhaustively generate all graphs with up to 7 vertices.

We then filtered these graphs for networks and thus excluded the following graphs:

- unconnected graphs
- graphs without leaves (i.e. vertices of degree ≤ 1)
- graphs with vertices of degree 2

We then further excluded all non-proper networks by checking whether the removal of any cut edge or any cut vertex present in the network resulted in connected components containing at least one leaf each.

This left us with 28 proper non-binary unrooted networks, which are all treebased and are depicted in the Appendix (see Catalog of all proper treebased networks with up to 7 vertices in Figures 18 and 19).

This shows that all non-binary unrooted networks with up to 7 vertices are treebased. However, for 8 vertices there exist non-binary unrooted networks which are not treebased. One such example is depicted in Figure 9. This network has 6 inner vertices and 2 leaves and it can be seen to not be treebased as follows. If it were treebased, then there would be a path from x to y visiting each vertex exactly once. Trivially, this path must start with edge $\{x, a\}$ and end with edge $\{f, y\}$. Now, the first three vertices of the path must either be $(x, a, b), (x, a, d)$ or (x, a, e). In all cases, the 4th vertex must be c, because the only other vertex (except from a) that can be reached from b, d or e is f; however, f has to be visited second last. Thus, so far we either have a path starting $(x, a, b, c), (x, a, d, c)$ or (x, a, e, c). Consider the first case, i.e. (x, a, b, c). From c we can either go to d or to e. If we go to d, i.e. if we have (x, a, b, c, d), the only "free" vertex reachable from d is f. This leads to a contradiction as vertex e has not been visited. If we instead go to e, i.e. if we have (x, a, b, c, e), the only "free" vertex reachable from e again f, which causes a contradiction as d has not been visited. Similar contradictions follow for all other cases. This completes the proof.

We are now in the position to prove Theorem 2.

Proof (Theorem 2) The first part of the proof is analogous to the first part of the proof of Theorem 1, i.e. we reduce the analysis to simple networks B_{N^u} with two leaves and show that all proper non-binary level-0,1,2,3 networks are treebased. As in the binary case, it is immediately clear that any level-k network is treebased if $k = 0$ or $k = 1$. For $k = 0$ the network is a tree and is thus treebased. For $k = 1$, at most one edge has to be removed from each non-trivial blob to obtain a tree; this tree is a support tree, because removing at most one edge from each non-trivial blob cannot induce any new leaves (same argument as in the binary case). Thus, let us now consider $k \geq 2$.

By Lemma 7, we know that any simple network B_{N^u} that leads to a proper non-binary non-treebased network has to have at least 6 internal vertices and 2 leaves, i.e. at least 8 vertices in total. Moreover, by Lemma 4 we know that for $k \geq 2$, the number of vertices in a non-trivial blob associated with a network B_{N^u} is bounded from above by $2k$. This implies that there cannot exist a level-2 network that is not treebased.

In order to prove that there also does not exist a non-treebased level-3 network, we exhaustively generated all unrooted proper phylogenetic networks on 8 vertices, analyzed them for treebasedness and computed their level. This exhaustive search was conducted in the following way: We used the “House of Graphs” database (cf. Brinkmann et al (2013)) to obtain a list of all graphs with 8 vertices (12346 in total). These were then filtered for unrooted proper phylogenetic networks in the same way as described in the proof of Lemma 7, resulting in a total of 197 unrooted proper phylogenetic networks. These were then analyzed for treebasedness and it turned out that there are only 8 proper phylogenetic networks on 8 vertices that are not treebased (see Figure 20). However, none of them is a level-3 network, i.e. we can conclude that all proper nonbinary level-3 networks are indeed treebased.

To prove the last statement of the theorem, consider the non-binary unrooted level-4 network depicted in Figure 9. This network is not treebased as we have already seen in the proof of Lemma 7. This completes the proof.

Remark 4 Note that the last part of the proof above answers a question posed in Hendriksen (2018), asking whether there exist networks of level less than 5 that are not loosely treebased. As the definition of loosely treebased in Hendriksen (2018) precisely corresponds to our definition of treebased, the level-4 network depicted in Figure 9 provides such an example.

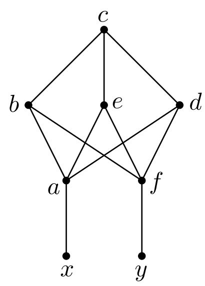

Fig. 9 Non-binary unrooted level-4 network that is not treebased (this network is adapted from Jetten and van Iersel (2018), where it is used in a different context).

3.4 Universal treebased networks

The aim of this section is to show that for all positive integers n, there exist both an unrooted and a rooted non-binary universal treebased network on n leaves, i.e. an (un)rooted network that has every (un)rooted non-binary phylogenetic tree on X as a base tree. For rooted binary networks this has independently been shown by Hayamizu (2016) and Zhang (2016) and has been further refined by Bordewich and Semple (2016). In Francis et al (2018) it was then shown that the existence of a rooted binary universal treebased network generalizes to unrooted binary universal treebased networks for all positive integers n. In the following we will further generalize this to non-binary networks. We will first consider non-binary rooted networks and show that there exists a rooted non-binary universal treebased network for all positive integers n. We will then show that this leads to the existence of an unrooted non-binary universal treebased network for all n.

3.4.1 Rooted universal treebased networks

In the following we will establish the following theorem.

Theorem 3 *For all positive integers n, there exists a rooted non-binary universal treebased network on n leaves.*

In the proof of Theorem 3 we will present a construction of a rooted treebased network for each n. Following the constructions in Hayamizu (2016), Zhang (2016) and Bordewich and Semple (2016), this construction consists of two parts: the upper part, which contains the root, is a network on n leaves that has every non-binary tree shape on n leaves as a support tree; the lower part, which contains the leaves, reorders the leaves of these tree shapes, in order to enable any permutation of leaves and thus, to enable every binary or non-binary phylogenetic tree on n leaves to be a base tree for this network (after suppressing potential degree 2 vertices in the support tree). For the latter we will use the same construction as Bordewich and Semple (2016), namely a so-called *Beneš network* (cf. Beneš (1964a,b)). An example is depicted in Figure 10; for details on the construction of the corresponding phylogenetic network see Bordewich and Semple (2016). Thus, in the following we will only show that the upper part of the construction has every tree shape as a support tree. Analogously to Bordewich and Semple (2016) it then follows that the combination of the upper part with a Beneš network is a universal treebased network.

Proof (Theorem 3) For all positive integers n, we now give a construction of a rooted phylogenetic network U_n on n leaves that has every tree shape on n leaves as support tree. We begin by describing the construction of U_n. First of all, for $n = 1$, U_n consists of a single vertex. Now, let $n \geq 2$. Then, the basic structure of U_n is the rooted star tree on n leaves, whose edges are subdivided and the resulting attachment points are connected by additional edges. To be

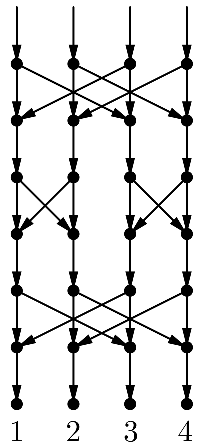

Fig. 10 Beneš network of size four (Figure adapted from Bordewich and Semple (2016)).

precise, we start with a rooted star tree T^* with root ρ on n leaves, where a rooted star tree is a rooted tree such that all leaves are incident to the root. Then:

- Add $n-2$ attachment points to each edge of T^*.
  - For leaf 1 and n, we label them $t_1^1, t_1^2, \ldots, t_1^{n-2}$ and $t_n^1, t_n^2, \ldots, t_n^{n-2}$, respectively (starting the labeling at the attachment point closest to the root; note that these vertices will be tree vertices in the final network);
  - For all leaves $l = 2, 3, \ldots, n-1$, we label them $r_l^1, r_l^2, \ldots, r_l^{n-2}$ (again, starting the labeling at the attachment point closest to the root; note that these vertices will be reticulation vertices in the final network).
- Add the following edges between attachment points
  - (r_i^k, r_{i+1}^k) for $i = 2, \ldots, n-2$ and $k = 1, \ldots, n-2$ (horizontal edges between reticulation vertices);
  - (r_i^k, r_i^{k+1}) for $i = 2, \ldots, n-1$ and $k = 1, \ldots n-3$ (vertical edges between reticulation vertices);
  - (r_i^k, r_j^{k+1}) for $i = 2, \ldots, n-2$, $j = 3, \ldots, n-1$ and $k = 1, \ldots, n-3$ (diagonal edges between reticulation vertices from left to right);
  - (r_i^k, r_j^{k+1}) for $i = 3, \ldots, n-1$, $j = 2, \ldots, n-2$ and $k = 1, \ldots, n-3$ (diagonal edges between reticulation vertices from right to left);
  - (t_i^k, r_j^k) for $i = 1, n$, $j = 2, \ldots, n-1$ and $k = 1, \ldots, n-2$ (diagonal edges between tree and reticulation vertices).

Figure 11 shows the resulting construction for $n = 4$ and $n = 5$. We will now show that – ignoring the leaf labels – U_n contains every tree shape on n leaves as a support tree.

We use induction on n to show that every tree shape on n leaves is a support tree of U_n. Since there is exactly one tree shape for $n = 1$ (consisting

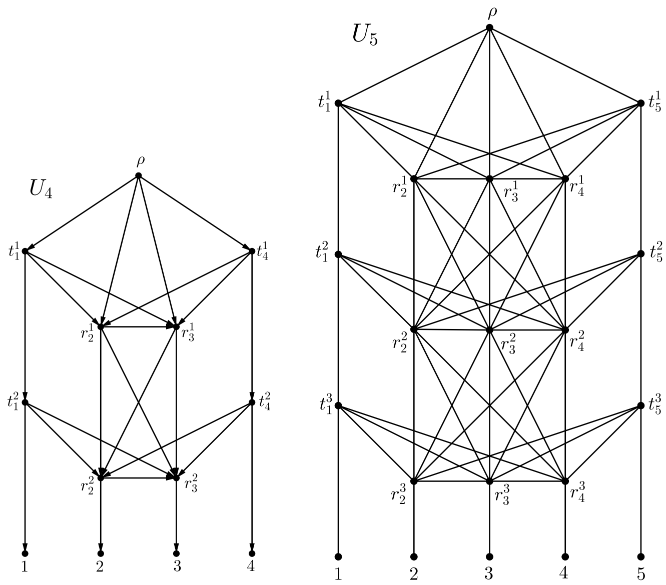

Fig. 11 The construction of U_4 and U_5. In U_5 all horizontal edges, i.e. edges of type (r_i^k, r_{i+1}^k) for $i = 2, 3$ and $k = 1, 2, 3$, are directed left to right; all other edges are directed away from the root.

of a single vertex) and $n = 2$, the base case holds for all $n \leq 2$. Now suppose that $n \geq 3$ and that, for all $m \leq n-1$, any tree shape on m leaves is a support tree of U_m.

Note that as the basic structure of U_m is a star tree on m leaves, the star tree on m trivially is a support tree of U_m. Therefore, we will now show that any other tree shape on m leaves is also a support tree of U_m.

Let T_m be an arbitrary tree shape on m leaves. Let $T_m^1, T_m^2, \ldots, T_m^p$ denote the subtrees of T_m pending at the children of the root ρ and let $m_1, m_2, \ldots, m_p$ denote the number of leaves in these subtrees. We will now show that T_m is a support tree of U_m. We will construct an explicit embedding of T_m into U_m, where we assume for technical reasons that leaves $1, \ldots, m_1$ correspond to T_m^1 and so forth (note that this labeling is just used for technical reasons and is not meant in a phylogenetic sense, i.e. the leaf numbers do not correspond to fixed species). As $m \geq 3$, we know that T_m contains at least one cherry $[u, v]$, i.e. a pair of leaves u and v who share a common parent (cf. Proposition 1.2.5 in Semple and Steel (2003)). Let w denote the parent of u and v and suppose that w has k, $k \geq 2$, children in total (including u and v). Moreover, without loss of generality we may assume that the children of w are labeled $1, \ldots, k$ when enumerating all leaves and are positioned at the outermost left of the tree when drawing it in the plane. We now delete all children of w (which implies that w is now a leaf) and retrieve a tree shape T_{m-k+1} with $m - k + 1$ leaves. As $m - k + 1 < m$, by induction T_{m-k+1} with vertex set $V(T_{m-k+1})$ and edge set $E(T_{m-k+1})$ is a support tree of U_{m-k+1} (see Figure 12 (a) – (c)).

In the following, we will first show that T_{m-k+1} can also be embedded in U_m; we will then re-introduce the deleted children of w and show that this yields a support tree of U_m.

Note that by construction T_{m-k+1} contains leaves labeled with w, $k+1$, $k+2, \ldots, m$. Before we embed T_{m-k+1} into U_m, we relabel some vertices of U_{m-k+1}. To be precise, we have the following renaming (as an example see Figure 12 (c)):

$$\begin{aligned} w &\hookrightarrow 1 \\ t^l_{m-k+1} &\hookrightarrow t^l_m \text{ for } \ l = 1, \ldots, m-k-1 \\ r^l_j &\hookrightarrow r^l_{j+k-1} \text{ for } j = 2, \ldots, m-k \text{ and } l = 1, \ldots, m-k-1. \end{aligned}$$

We now sequentially extend the network U_{m-k+1} to U_m by introducing additional vertices and edges. First of all, we add $k-1$ attachment points on the following edges (cf. Figure 12 (d))

- $(t_1^{m-k-1}, 1)$: attachment points $t_1^{m-k}, t_1^{m-k+1}, \ldots, t_1^{m-2}$
- (t_m^{m-k-1}, m): attachment points $t_m^{m-k}, t_m^{m-k+1}, \ldots, t_m^{m-2}$
- $(r_{j+k-1}^{m-k-1}, j+k-1)$: attachment points $r_{j+k-1}^{m-k}, r_{j+k-1}^{m-k+1}, \ldots r_{j+k-1}^{m-2}$ for $j = 2, \ldots, m-k$.

We add all newly introduced edges to $E(T_{m-k+1})$, i.e. we extend T_{m-k+1} to cover all newly introduced attachment points (as an example see Figure 12 (d)).

We then add $k-1$ edges connecting the root to leaves $2, \ldots, k$ and on each of these edges add $m-2$ attachment points called $r_i^1, \ldots, r_i^{m-2}$ for $i = 2, \ldots, k$ (as an example see Figure 12 (e)). In order to complete the construction of U_m, we add all required edges between newly introduced vertices, i.e. we complete the construction of U_m according to the construction principle presented at the beginning of the proof (see page 23; as an example see Figure 12 (f)).

We now re-introduce the children of w to T_{m-k+1} in order to obtain T_m, i.e. the leaves $2, \ldots, k$ (note that we do not re-introduce leaf 1, as this was already re-introduced in a previous step). We do this in the following way:

These operations transform T_{m-k+1} back to T_m in such a way that all vertices of U_m are also vertices of T_m, thus T_m is a support tree of U_m (as an example see Figure 12 (g)). As T_m was an arbitrary tree shape on m leaves this completes the proof.

3.4.2 Unrooted universal treebased networks

By ignoring the designation of the vertex ρ as root of the network and the orientation of edges, the same construction as in the rooted case can be used to show that for all positive integers $n \geq 3$, there exists an unrooted non-binary universal treebased network with n leaves. For $n = 1$ and $n = 2$, an unrooted universal treebased network trivially exists. For $n = 1$, the only unrooted tree shape is a single vertex, which at the same time is the only treebased unrooted network. For $n = 2$, the only unrooted tree shape is an edge between the two

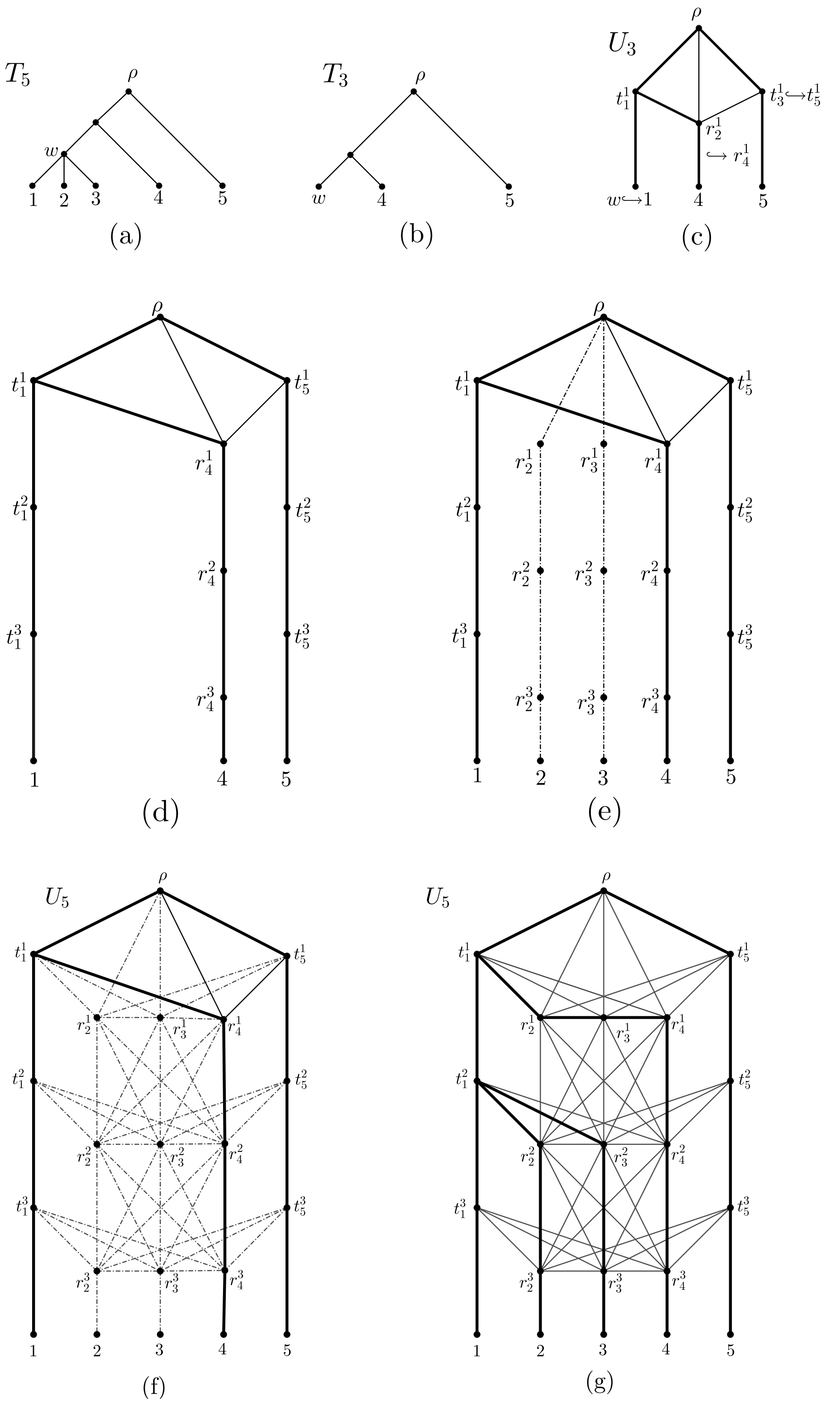

Fig. 12 Illustration of the construction and concepts used in the proof of Theorem 3. T_5 is a non-binary tree shape on 5 leaves (a). We consider vertex w and delete its children, which yields tree shape T_3 on 3 leaves (b). By the inductive hypothesis T_3 is a support tree of U_3; it is depicted in bold (c). After relabeling vertices (c), 2 attachment points are added on the edges $(t_1^1, 1), (r_4^1, 4)$ and $(t_5^1, 5)$, respectively (d). All new edges created in this step, e.g. (t_1^1, t_1^2), are added to the support tree T_3. Then, 2 edges connecting the root to leaves 2 and 3 are added. These edges are subdivided by introducing 3 attachment points on each edge (e). Then, the construction of U_5 is completed by introducing all missing edges between tree vertices and reticulation vertices and between pairs of reticulation vertices (f). In the last step, T_3 is transformed back to T_5 (g): Firstly, the edge (t_1^1, r_4^1) (depicted in bold in (f)) is replaced by the edges $(t_1^1, r_2^1), (r_2^1, r_3^1)$ and (r_3^1, r_4^1) (depicted in bold in (g)). In the last step the edges $(t_1^2, r_2^2), (t_1^2, r_3^2), (r_2^2, r_3^2), (r_2^3, 2), (r_3^2, r_3^3)$ and $(r_3^3, 3)$ are added to the embedding of T_5 into U_5, which results in T_5 being a support tree of U_5 (depicted in bold).

$i = 1$;
while $i \leq m - k$ **do**
  if *edge* (t_1^i, r_{k+1}^i) *is in* $E(T_{m-k+1})$ **then**
    remove edge (t_1^i, r_{k+1}^i) from $E(T_{m-k+1})$;
    add edges $(t_1^i, r_2^i), (r_2^i, r_3^i), \ldots, (r_k^i, r_{k+1}^i)$ to $E(T_{m-k+1})$;
    $i = i + 1$;
  else
    add the following edges to $E(T_{m-k+1})$:
    – $(t_1^i, r_2^i), (t_1^i, r_3^i), \ldots, (t_1^i, r_k^i)$;
    – (r_j^l, r_j^{l+1}) for $j = 2, \ldots, k$ and $l = i, \ldots, m - 3$;
    – (r_j^{m-2}, j) for $j = 2, \ldots, k$.
  end
end

leaves. Thus, any treebased unrooted network on 2 leaves can be considered an unrooted universal treebased network for $n = 2$. Summarizing the above, we have the following statement.

Theorem 4 *For all positive integers n, there exists an unrooted non-binary universal treebased network on n leaves.*

4 Discussion

The main aim of this manuscript was to generalize some of the results presented in Francis et al (2018) for binary treebased unrooted networks to non-binary ones. In particular, we showed that unlike in the binary case, level-4 networks need not necessarily be treebased in the non-binary case. This provides the answer to Question 5.3 in Hendriksen (2018), asking whether there are non-binary networks of level less than 5 that are not treebased.

Additionally, we reproved Theorem 1 of Francis et al (2018), using a different argument. Along the way, we gave a new example showing that binary level-5 networks are not always treebased, as the example given in Francis et al (2018) was unfortunately erroneous.

We concluded our study with the explicit construction of universal non-binary treebased networks both for the rooted and unrooted case. Such a construction for binary networks has been proven useful in the past (cf. Hayamizu (2016); Zhang (2016); Bordewich and Semple (2016)), but so far has been unavailable for non-binary networks. Therefore, we are confident that our constructions will inspire further research. One question concerning universal treebased networks, for instance, is the minimal number of reticulations needed to realize them. Possibly, our construction is more complex than necessary in terms of the number of reticulations. Therefore, it would be interesting to see if this number can be reduced.

Acknowledgement

The third and fifth author thank the state Mecklenburg-Western Pomerania for the Landesgraduierten-Studentship. Moreover, the second author thanks the University of Greifswald for the Bogislaw-Studentship and the fifth author thanks the German Academic Scholarship Foundation for a studentship.

References

Amaral AR, Lovewell G, Coelho MM, Amato G, Rosenbaum HC (2014) Hybrid speciation in a marine mammal: The clymene dolphin (*Stenella clymene*). PLoS ONE 9(1):e83,645, DOI 10.1371/journal.pone.0083645

Beneš VE (1964a) Optimal rearrangeable multistage connecting networks. Bell System Technical Journal 43(4):1641–1656, DOI 10.1002/j.1538-7305.1964.tb04103.x

Beneš VE (1964b) Permutation groups, complexes, and rearrangeable connecting networks. Bell System Technical Journal 43(4):1619–1640, DOI 10.1002/j.1538-7305.1964.tb04102.x

Bordewich M, Semple C (2016) A universal tree-based network with the minimum number of reticulations. URL `https://arxiv.org/abs/1707.08274`, `arXiv:1707.08274`

Brinkmann G, Coolsaet K, Goedgebeur J, Mélot H (2013) House of graphs: A database of interesting graphs. Discrete Applied Mathematics 161(1-2):311–314, DOI 10.1016/j.dam.2012.07.018

Bussemaker F, Cobeljic S, Cvetkovic D, Seidel J (1976) Computer investigation of cubic graphs. EUT report. WSK, Dept. of Mathematics and Computing Science, Technische Hogeschool Eindhoven

Choya C, Jansson J, Sadakane K, Sung WK (2005) Computing the maximum agreement of phylogenetic networks. Theoretical Computer Science 335(01):93–107, DOI 10.1016/j.tcs.2004.12.012

Cui R, Schumer M, Kruesi K, Walter R, Andolfatto P, Rosenthal GG (2013) Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. Evolution 67(8):2166–2179, DOI 10.1111/evo.12099

Fischer M, Galla M, Herbst L, Long Y, Wicke K (2018) Classes of treebased phylogenetic networks. URL `https://arxiv.org/abs/1810.06844`, `arXiv:1810.06844`

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu YC, Smith HA, Love RR, Lawniczak MK, Slotman MA, Emrich SJ, Hahn MW, Besansky NJ (2014) Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347(6217):1258,524–1258,524, DOI 10.1126/science.1258524

Francis A, Steel M (2015) Which phylogenetic networks are merely trees with additional arcs? Systematic Biology 64(5):768–777, DOI 10.1093/sysbio/syv037, URL `http://dx.doi.org/10.1093/sysbio/syv037`

Francis A, Huber KT, Moulton V (2018) Tree-based unrooted phylogenetic networks. Bulletin of Mathematical Biology 80(2):404–416, DOI 10.1007/s11538-017-0381-3, URL `https://doi.org/10.1007/s11538-017-0381-3`

Gambette P, Berry V, Paul C (2012) Quartets and unrooted phylogenetic networks. Journal of Bioinformatics and Computational Biology 10(04):1250,004, DOI 10.1142/S0219720012500047, pMID: 22809417

Gusfield D, Bansal V (2005) A fundamental decomposition theory for phylogenetic networks and incompatible characters. In: Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner PA, Waterman M (eds) Research in Computational Molecular Biology, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 217–232

Harris JM, Hirst JL, Mossinghoff MJ (2000) Combinatorics and Graph Theory. Springer

Hayamizu M (2016) On the existence of infinitely many universal tree-based networks. Journal of Theoretical Biology 396:204–206, DOI 10.1016/j.jtbi.2016.02.023

Hendriksen M (2018) Tree-based unrooted nonbinary phylogenetic networks. Mathematical Bioscience 302:131–138, DOI 10.1016/j.mbs.2018.06.005

Inc WR (2017) Mathematica, Version 10.3. Champaign, IL, 2017

Jetten L, van Iersel L (2018) Nonbinary tree-based phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinformatics 15(1):205–217, DOI 10.1109/TCBB.2016.2615918, URL `https://doi.org/10.1109/TCBB.2016.2615918`

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

Zamfirescu T (1976) On longest paths and circuits in graphs. MATHEMATICA SCANDINAVICA 38:211, DOI 10.7146/math.scand.a-11630

Zhang L (2016) On tree-based phylogenetic networks. Journal of Computational Biology 23(7):553–565, DOI 10.1089/cmb.2015.0228

5 Appendix

5.1 Supplementary results

Lemma 8 *Let N^u be a binary unrooted phylogenetic network. Then every cut vertex is incident to a cut edge.*

Proof Let N^u be a binary unrooted phylogenetic network with cut vertex v. Then, v has degree three, because leaves cannot be cut vertices. We call the three edges incident to v e_1, e_2 and e_3 as depicted in Figure 13.

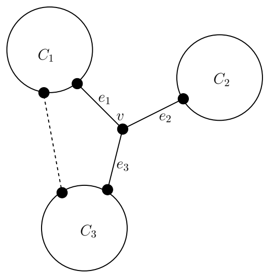

Fig. 13 Binary unrooted phylogenetic network N^u with cut vertex v and edges e_1, e_2 and e_3 leading to components C_1, C_2 and C_3.

It is possible that e_1, e_2 and e_3 lead to three different components C_1, C_2 and C_3 (in this case the dashed line in Figure 13 is excluded). Then, e_1, e_2 and e_3 are all cut edges and thus v is incident to three different cut edges.
Otherwise, two edges lead to the same component. Without loss of generality e_1 and e_3 lead to the same component (dashed line in Figure 13). Then e_2 is a cut edge. Therefore, v is incident to a cut edge.
Note that not all edges can lead to the same component, because if that was the case v would not be a cut vertex. This completes the proof.

Proposition 1 *Suppose N^u is an unrooted network. Then N^u is treebased if and only if B_{N^u} is treebased for every blob B in N^u.*

Proof Suppose N^u is an unrooted treebased network on X. As N^u is treebased there exists a support tree T for N^u, i.e. a spanning tree with leaf set X. As T is a spanning tree, T in particular contains all cut vertices of N^u. Moreover, it has to contain all cut edges of N^u, because otherwise T would not be connected. Thus, any support tree for N^u induces a spanning tree of B_{N^u} and we can conclude that B_{N^u} is treebased. Conversely, suppose that B_{N^u} is treebased for every blob B in N^u. Then by taking a support tree for B_{N^u} for each blob, we can construct a support tree T for N^u by connecting the individual support trees corresponding to the blobs of N^u via the cut-edges that connected the blobs of N^u. Thus, N^u is treebased.

Lemma 3 *Let N^u be a network on X with $|X| \geq 2$. For any $x \in X$ let $N^u - x$ denote the network obtained from N^u by deleting x and its incident edge, and suppressing the potentially resulting degree 2 vertex. Then, if $N^u - x$ is treebased, so is N^u.*

Proof Let N^u be a network on X with $|X| \geq 2$. Let $N^u - x$ be obtained from N^u by deleting leaf x and its incident edge. If this results in a degree 2 vertex v (which for example is the case if N^u is a binary network), v is suppressed. Note that this might imply that $N^u - x$ contains parallel edges (cf. Figure 1) and thus, is not a phylogenetic network anymore. However, note that deleting a leaf and suppressing a degree 2 vertex cannot result in loops. Moreover, it cannot result in an unconnected (multi)graph. In particular, $N^u - x$ is a connected (multi)graph without loops. We will now show that if $N^u - x$ is treebased, so is N^u. At this point, it is important to notice that in this manuscript treebasedness is not only defined for phylogenetic networks, but more generally for (multi)graphs (cf. Definition 1). Let T be a support tree for $N^u - x$. We now distinguish between three cases:

1. $deg(v) = 1$ in N^u:
   If $deg(v) = 1$ in N^u, N^u must consist of a single edge, namely $\{v, x\}$. A single edge is trivially treebased, so there is nothing to show.
2. $deg(v)$ in N^u is strictly greater than 3:
   If $deg(v)$ in N^u is strictly greater than 3, it is strictly greater than 2 in $N - x$. This implies that v is not suppressed in $N^u - x$. Then, we can obtain a support tree for N^u from T by adding the edge $\{x, v\}$ to T.
3. $deg(v) = 3$ in N^u, i.e. $deg(v) = 2$ in $N^u - x$:
   Let $v_1, v_2 \neq x$ denote the other two vertices adjacent to v. Let $\{v_1, v_2\}$ denote the edge that results from suppressing v in $N^u - x$. Note that this might be a parallel edge, if there already is an edge $\{v_1, v_2\}$ in $N^u - x$. Now, there are two cases:
   – If $\{v_1, v_2\}$ is an edge in T (if there are multiple edges between v_1 and v_2, T will only contain one of them), then we can obtain a support tree for N^u by subdividing this edge (i.e. re-introducing the attachment point v) and adding the edge $\{x, v\}$ to T.
   – If $\{v_1, v_2\}$ is not an edge in T, we note the following. As T is a support tree for $N^u - x$, it must contain both v_1 and v_2. Thus, we can obtain a support tree for N^u by re-introducing vertex v and the edges $\{v_1, v\}$ and $\{v, x\}$ (or $\{v_2, v\}$ and $\{v, x\}$) to T.

This completes the proof.

Lemma 6 *Any minimal proper binary unrooted non-treebased network has 12 vertices (10 internal vertices and 2 leaves), and this bound is tight, i.e. there are proper binary unrooted non-treebased networks with precisely 12 vertices.*

Proof First recall that the number of nodes in a binary unrooted phylogenetic network is always even unless if N^u simply consists of only one node. As explained in Section 2.2, this is due to the handshaking lemma.

Now assume that N^u is a proper binary unrooted non-treebased network. As N^u is non-treebased, in particular N^u does not only consist of one node or of two nodes connected by an edge. So as the number of nodes has to be even, the total number of nodes has to be at least four.

Moreover, consider $|X|$. If $|X| = 1$, then N^u cannot be proper (see Remark 1), so we must have $|X| \geq 2$.

In summary, we know so far that N^u has at least four nodes, at least two of which are leaves. First, suppose N^u has exactly four nodes, *two* of which are leaves. This implies that there are two internal nodes, each of them having degree 2. This means that N^u is a path. In particular, N^u is not binary and thus not a binary network. Now, suppose N^u has exactly four nodes, *three* of which are leaves. Then N^u is a tree on 3 leaves. In particular, N^u is treebased, which is a contradiction. Thus, as the number of nodes in a binary unrooted phylogenetic network is even, we can conclude that $|V| \geq 6$.

Now assume that N^u has strictly fewer than 12 vertices in total. As N^u is binary and as we have already seen that binary networks have an even number of nodes, this means that N^u has at most 10 nodes in total. We now distinguish two main cases, which can both be subdivided into two subcases:

- First assume $|X| = 2$. Without loss of generality, let $X = \{x, y\}$ and let u and v denote the nodes adjacent to x and y, respectively. Now, consider the leaf connecting procedure. As $|X| = 2$, $\mathcal{LCON}(N^u)$ consists of precisely one element, which we denote by $G(N^u)$. Note that $G(N^u)$ is a cubic graph as N^u is binary. Moreover, recall that the construction of $G(N^u)$ might have required the introduction of new nodes, say a and b, to avoid parallel edges (see description of the leaf connecting procedure in Section 2.3). We now distinguish between two cases.
  - If in the construction of $G(N^u)$ no nodes a and b had to be added, then $G(N^u)$ contains precisely $|V| - 2$ nodes. As N^u by assumption contains fewer than 12 nodes, $G(N^u)$ contains fewer than 10 nodes. Thus, as $G(N^u)$ is cubic and therefore contains an even number of nodes, $G(N^u)$ contains at most 8 nodes. However, by Lemma 5, we know that up to 8 vertices there is a Hamiltonian path from u' to v' for every edge $\{u', v'\}$ in a cubic graph. So in particular, $G(N^u)$ contains a Hamiltonian path from u to v, i.e. from the attachment point of the first leaf to the attachment point of the second leaf. Adding edge $\{u, v\}$ to this path yields a Hamiltonian cycle using this new edge, which was not contained in N^u, and thus, by Proposition 2, N^u is treebased, which is a contradiction.
  - If in the construction of $G(N^u)$ nodes a and b had to be added, then $G(N^u)$ contains precisely $|V|$ nodes (x and y have been deleted, but a and b have been added). As N^u by assumption contains fewer than 12 nodes, $G(N^u)$ also contains fewer than 12 nodes. As the number of

nodes in any cubic graph is even, $G(N^u)$ contains at most 10 nodes. As above, if $G(N^u)$ contains at most 8 nodes, there is a Hamiltonian path from u to a, which can be extended to a Hamiltonian cycle by adding edge $\{a, u\}$, and thus, by Proposition 2, N^u is treebased, which is a contradiction. So let us consider the case where $G(N^u)$ contains precisely 10 nodes. Note that there are only 19 different cubic graphs with 10 nodes, and only two of them are not Hamiltonian (cf. Bussemaker et al (1976)). Now if $G(N^u)$ is one of the Hamiltonian graphs, it contains a Hamiltonian cycle. We now argue that each such cycle must use edge $\{a, b\}$. Note that by construction of $G(N^u)$ through leaf connection a is only adjacent to u, v and b, and b only to u, v and a. Therefore, the Hamiltonian cycle will connect u and v in two ways, namely with one path visiting all nodes except for a and b, and additionally with a path only visiting a and b. For instance, the path u, a, b, v or u, b, a, v would be possible. In all such cases, edge $\{a, b\}$ is necessarily contained in all Hamiltonian cycles. So deleting edge $\{a, b\}$ leads to a Hamiltonian path from a to b. Subsequently, suppressing a and deleting b as well the edges $\{u, b\}$ and $\{v, b\}$, leads to a Hamiltonian path from u to v. Using the same arguments as above, this implies that N^u is treebased, which would be a contradiction. So $G(N^u)$ has to be one of the two non-Hamiltonian cubic graphs with 10 nodes. These two graphs are depicted in Figure 14.

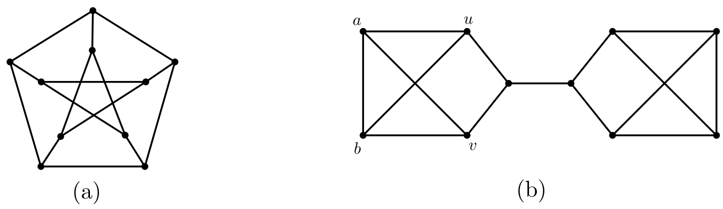

Fig. 14 There are only two non-Hamiltonian cubic graphs with 10 nodes, where (a) is the so-called Petersen graph.

Note that the first one, namely the Petersen graph, contains no pair a and b of vertices such that a and b are adjacent to one another and to the same two other nodes u and v. So as we assume that such nodes a and b were added during the construction of $G(N^u)$, $G(N^u)$ cannot be the Petersen graph. The other one of these two graphs, which is depicted in Figure 14 (b), has two possible positions for the pair a and b, but it also has a cut edge, and the positions for a and b (one of which is depicted in Figure 14 (b)) are such that both a and b would be at the same side of the cut edge. Thus, by the construction procedure leading to $G(N^u)$, also x and y must have been on the same side of the cut edge, but then N^u cannot be proper. This is a contradiction.

In summary, if $|X| = 2$, N^u must contain at least 12 nodes.

- Now assume $|X| > 2$. If necessary, we first perform the pre-processing step of the leaf connecting procedure introduced in Section 2 and denote the resulting reduced taxon set of N^u by X^r. Note that $\mathcal{LCON}(N^u)$ possibly contains more than one graph. So let $G(N^u)$ in the following be an arbitrary element of $\mathcal{LCON}(N^u)$. We now distinguish two cases.
  - Suppose $|X^r|$ is even. Consider $G(N^u)$ and $\widetilde{N^u}$, where $\widetilde{N^u}$ is the second to last graph in the construction of $G(N^u)$ according to the definition of $\mathcal{LCON}(N^u)$. In particular, $\widetilde{N^u}$ is the graph which we get when only two leaves are left, which we would have to connect in order to receive the final graph $G(N^u)$. Thus, by construction $\widetilde{N^u}$ has at most as many nodes as N^u, so by assumption fewer than 12, and it has two leaves. But as N^u is not treebased by assumption, neither is $\widetilde{N^u}$, because any support tree for $\widetilde{N^u}$ would lead to a support tree of N^u. To see this, we distinguish between two cases:
    - If the support tree of $\widetilde{N^u}$ only contains edges that are both present in $\widetilde{N^u}$ and in N^u, we can obtain a support tree for N^u by re-attaching the leaves at their former positions.
    - Otherwise, suppose that the support tree of $\widetilde{N^u}$ contains at least one edge that is not present in N^u. Note that there are two potential types of edges that can be present in $\widetilde{N^u}$ but not in N^u:
      - · Edges that were introduced to avoid multiple edges between two attachment points, say u and v, of leaves, i.e. the edges $\{u, a\}$, $\{u, b\}$, $\{a, v\}$, $\{b, v\}$ and $\{a, b\}$. If the support tree of $\widetilde{N^u}$ uses any of these edges, we can obtain a support tree for N^u as follows: Delete these edges as well as nodes a and b from the support tree. Additionally, add a new edge $e = \{u, v\}$ to the support tree of $\widetilde{N^u}$ (note that this is allowed as $e = \{u, v\}$ must have been contained in N^u, which led to the introduction of a and b) and re-attach the leaves incident to u and v in N^u to u and v, respectively. We can repeat this procedure for all edges of this type.
      - · Edges $e = \{u, v\}$ between two attachment points (u and v) of leaves. If the support tree of $\widetilde{N^u}$ uses such an edge, we can obtain a support tree for N^u as follows: First of all, delete the edge $\{u, v\}$ from the support tree of $\widetilde{N^u}$. Note that this disconnects the support tree. Let T_1 and T_2 denote its two connected components and assume that u is in T_1 and v is in T_2. Moreover, note that both u and v are of degree 3 by construction, thus T_1 and T_2 cannot be single nodes. As N^u is a connected graph, in particular there exists a node $u' \neq u$ in T_1 and a node $v' \neq v$ in T_2 such that u' and v' are connected by an edge $e' = \{u', v'\}$ in N^u (if T_1 and T_2 were only connected via $\{u, v\}$, N^u would not have been connected, as $\{u, v\}$ is only present in $\widetilde{N^u}$ and not in N^u). We now add $e' = \{u', v'\}$ to

the support tree of $\widetilde{N^u}$, and re-attach the leaves incident to u and v in N^u to u and v, respectively. Again, we can repeat this procedure for all edges of this type.

In all cases, we can construct a support tree for N^u from a support tree of $\widetilde{N^u}$. This is a contradiction as N^u is not treebased. Thus, $\widetilde{N^u}$ has to be non-treebased. However, then $\widetilde{N^u}$ would be a non-treebased network with two leaves and strictly fewer than 10 inner nodes, which contradicts the first part of the proof.

– Suppose $|X^r|$ is odd. By construction, as we assume N^u has at most 10 nodes, also $G(N^u)$ can have at most 10 nodes. This is due to the fact that in each step during the construction of $G(N^u)$, either two leaves are deleted or, in the last step, one leaf and its attachment point are deleted. So in all steps, two nodes are deleted and at most two new nodes are added (if parallel edges need to be avoided), so the total number of nodes cannot increase.
Now, if the resulting graph $G(N^u)$ has at most eight nodes, we already know by Lemma 5 that for each edge $e = \{u', v'\}$ it contains a Hamiltonian cycle from u' to v'. Thus, using the same arguments as in the case where $|X| = 2$ combined with Lemma 3, N^u must have been treebased, which is a contradiction[5]. If, on the other hand, G has precisely 10 nodes, then again, as in the case where $|X| = 2$, we only need to consider the two cubic graphs with 10 nodes which are non-Hamiltonian depicted in Figure 14. The Petersen graph as before does not have any pair of nodes a and b that could be suppressed so that we have parallel edges, which implies that no nodes have been added during the deletion of leaves. This means that all 10 nodes of the Petersen graph were already there in N^u, plus at least three leaves. So in total, N^u would have at least 13 vertices, which is a contradiction to the assumption that N^u has fewer than 12 nodes.
The other non-Hamiltonian cubic graph with 10 nodes, however, namely the one depicted in Figure 14 (b) has the property that wherever we attach at least two leaves, the network is immediately treebased whenever it is proper: If the leaf set is such that it is distributed at both sides of the cut edge, the network is treebased, and if all leaves are on the same side of the cut edge, the network is not proper (whether or not we delete the candidate nodes a and b as depicted in Figure 14 (b), which may have been added during the deletion of the leaves, does not matter). Both scenarios contradict the assumption that the network is proper but not treebased.

So in all cases, the result is a contradiction, so every network with fewer than 12 vertices is treebased. This completes the proof.

[5] For instance, we can apply these arguments to the case where, on the way to constructing $G(N^u)$, the resulting network has 3 leaves and the last pair gets connected, before we deal with the last singleton leaf. At least one of the edges resulting from connecting this last pair of leaves must be contained in any Hamiltonian cycle and thus leads to a Hamiltonian path when we disregard it.

5.2 Supplementary figures

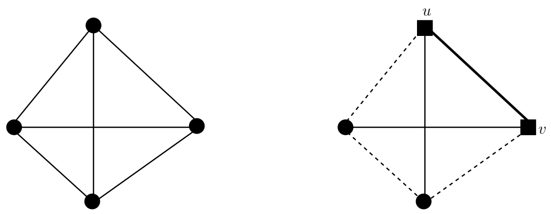

Fig. 15 Left: There exists exactly one cubic graph with 4 vertices. Right: For $e = \{u, v\}$ there exits a Hamiltonian path from u to v indicated by dashed lines.

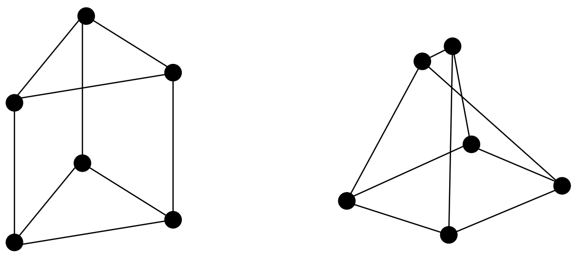

Fig. 16 All cubic graphs with 6 vertices.

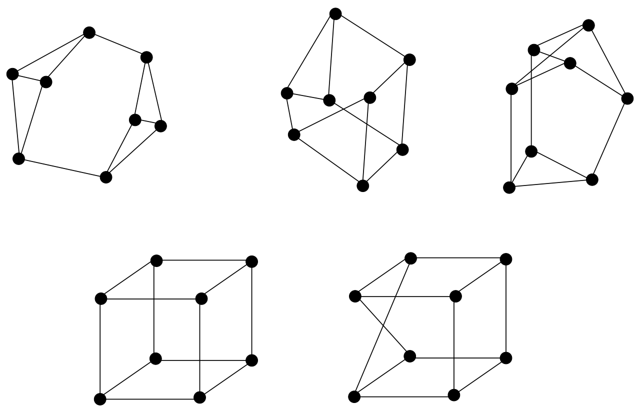

Fig. 17 All cubic graphs with 8 vertices.

5.2.1 Catalog of all proper treebased networks with up to 7 vertices

In the following all proper treebased networks with up to 7 vertices are depicted. For each network, a support tree is shown in bold lines and the additional network edges are given by dashed lines.

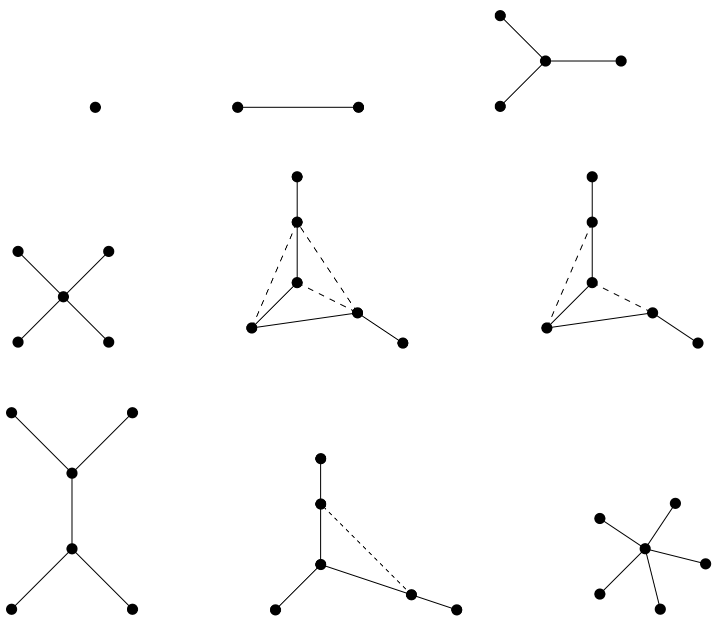

Fig. 18 All proper treebased networks with up to 6 vertices.

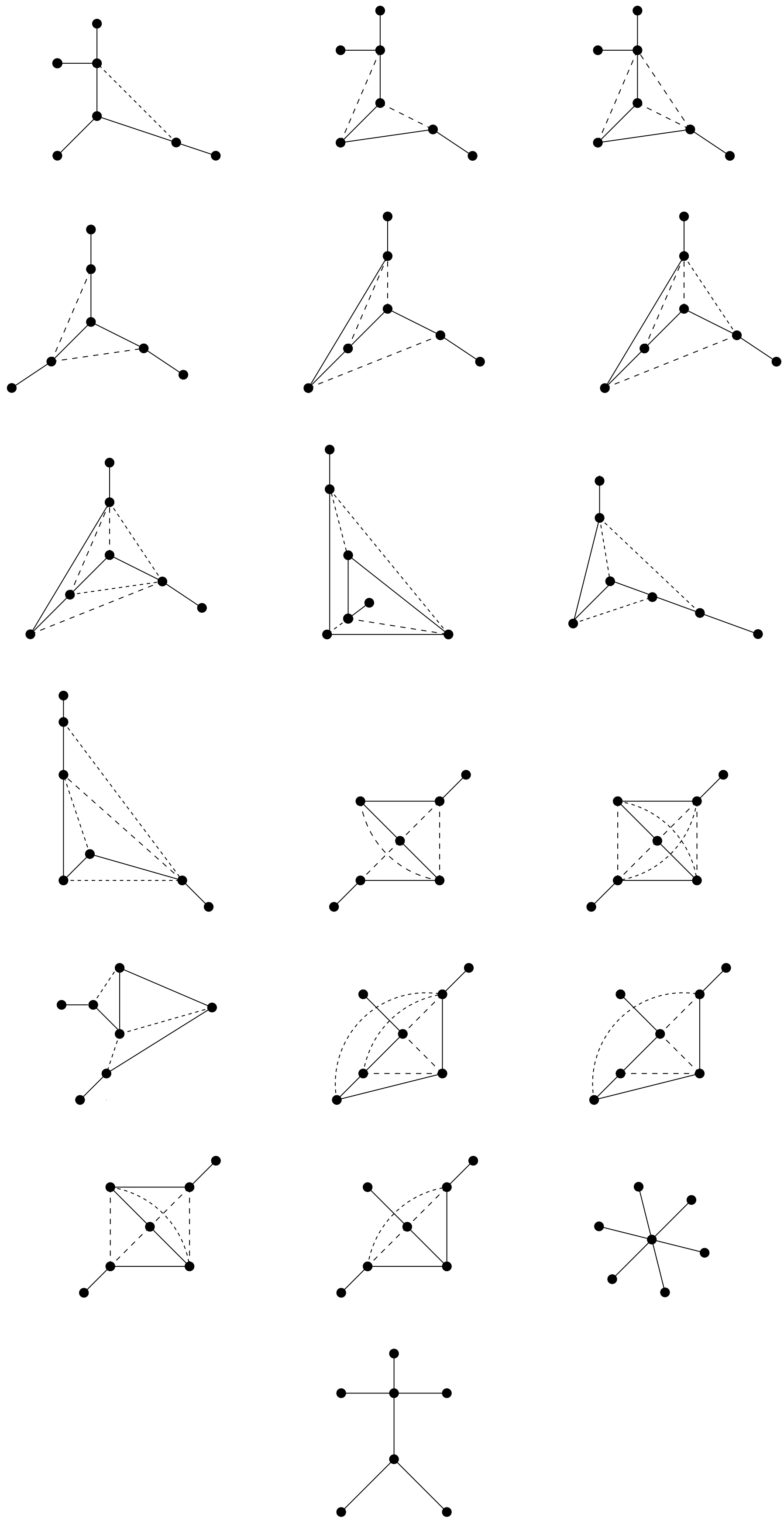

Fig. 19 All proper treebased networks with 7 vertices.

5.2.2 Catalog of all proper non-treebased networks with 6 inner vertices and 2 leaves

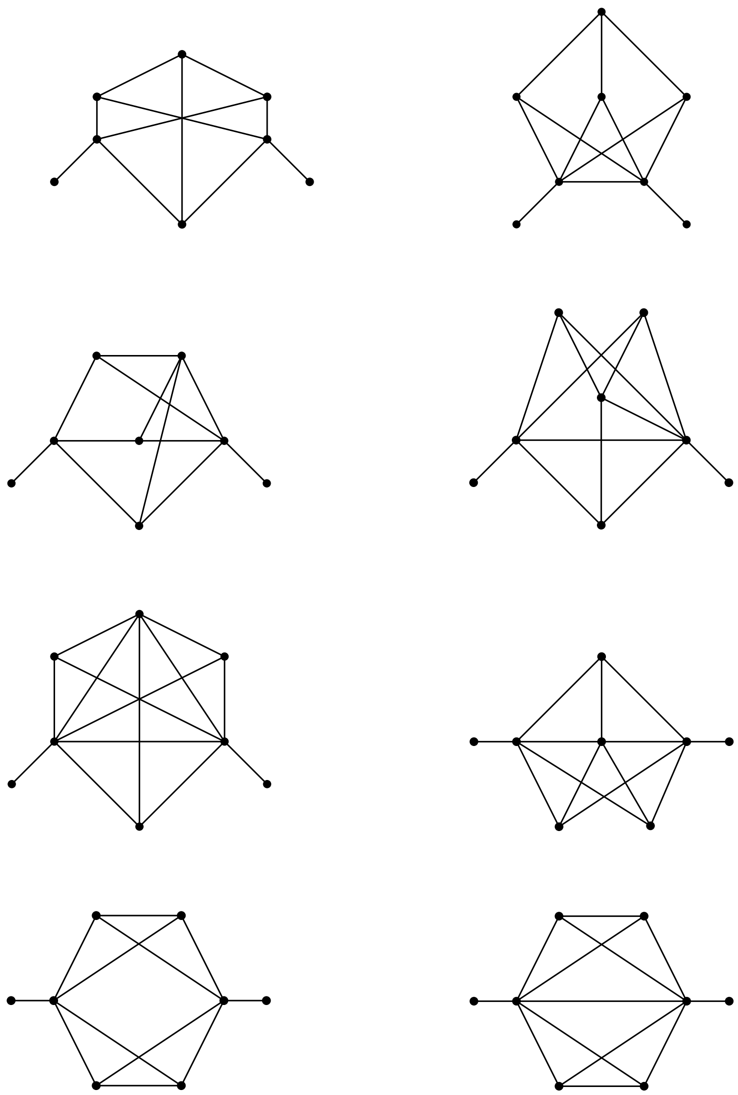

Fig. 20 All proper unrooted non-treebased networks on 8 vertices, 2 of which are leaves.