Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Characterising tree-based phylogenetic networks

(Karakterisatie van fylogenetische netwerken die een boom als basis hebben)

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

Laura Jetten

Delft, the Netherlands
November 2015

BSc THESIS APPLIED MATHEMATICS

"Characterising tree-based phylogenetic networks"

("Karakterisatie van fylogenetische netwerken die een boom als basis hebben")

Laura Jetten

Delft University of Technology

Supervisor

Dr.ir. L.J.J. van Iersel

Other thesis committee members

Dr.ir. M. Keijzer

Dr. J.L.A. Dubbeldam

November, 2015

Delft, the Netherlands

Abstract

Commonly, the evolutionary history of a set of taxa is described by a phylogenetic tree. However, in certain cases, evolution can best be described by a phylogenetic network. In previous research, the term tree-based was introduced for a phylogenetic network, meaning that it can be drawn as a phylogenetic tree with additional horizontal arcs. In particular, an algorithm was given to determine whether a binary network is tree-based or not. Here we give a simple graph-theoretic classification of all tree-based and non-tree-based binary phylogenetic networks. In addition, we give an upper bound on how many leaves need to be added to make any binary network tree-based. We also give an upper bound for the number of base trees that a tree-based binary phylogenetic network contains. Finally, since there has not been done any previous research on the tree-basedness of non-binary phylogenetic networks, some theorems of the binary case are studied and checked whether they also apply in the non-binary case. We show that some of these theorems apply in the non-binary case and some do not. In particular, we give a classification for non-binary phylogenetic networks that are tree-based.

Contents

Abstract 4
1. Introduction 6
2. Binary phylogenetic networks 8
2.1. Preliminaries 8
2.2. Theorems 11
3. Non-binary phylogenetic networks 24
3.1. Definitions 24
3.2. Theorems 25
4. Conclusion and discussion 27
References 29

1. Introduction

For centuries, evolution has been an important topic in biology. Commonly, the evolutionary history of a set of taxa is described by a phylogenetic tree. However, in certain cases, evolution could best be described by a phylogenetic network. For example, a phylogenetic tree is insufficient to describe the evolutionary history including hybrid species, since a hybrid species has (at least) two parent species. Furthermore, research has shown that bacteria have transferred genes from one species to another, while they did not directly share a common ancestor. These gene transfers can be denoted by horizontal arcs added to a phylogenetic tree. Such a phylogenetic tree with additional horizontal arcs can alternatively be displayed as a phylogenetic network. Mathematically, a rooted binary phylogenetic network is a directed acyclic graph which contains one root, tree-vertices, reticulations and leaves. The root has in-degree 0 and out-degree 1 or 2, tree-vertices have in-degree 1 and out-degree 2, reticulations have in-degree 2 and out-degree 1 and leaves are vertices with in-degree 1 and out-degree 0. A tree contains no reticulations. Therefore, a phylogenetic tree is not always able to fully describe the process of evolution.

In previous research [2] the term tree-based was introduced for a phylogenetic network, meaning that it can be drawn as a phylogenetic tree with additional horizontal arcs. Such arcs can represent gene transfer, in our example between bacterial species. Several theorems were presented that can be used to determine, in certain cases, whether a network is tree-based or not. For example, when a binary phylogenetic network contains a reticulation that has two parents that are reticulations, the network is not tree-based. Additionally, an algorithm was presented to check whether a binary phylogenetic network is tree-based or not. However, although it has been shown that phylogenetic networks are tree-based in some cases and not tree-based in certain other cases, no general classification of tree-based networks is given. In addition, the paper does not discuss any theorems about non-binary phylogenetic networks, in which a vertex can have more than two children or more than two parents.

Therefore, we will thoroughly research the difference between tree-based and non-tree-based binary networks and subsequently study tree-basedness of non-binary networks. First, we will give the most important definitions for binary networks and describe the algorithm from [2] in Section 2.1. An important and new notion is *omnian*, which is a vertex that has only reticulation children. After the preliminaries, some important theorems from previous research are stated in Section 2.2.1, followed by new theorems in Section 2.2.2. First, it will be shown that a binary phylogenetic network is tree-based if and only if there exists a matching in a certain bipartite graph that is associated to the network, where every omnian is covered by the matching. This theorem is used to show that all networks containing at most two reticulations are tree-based. On the other hand, we will show an example of a part of a binary network containing three reticulations that is not tree-based. Then, using Hall's Theorem, we give a simple graph-theoretic classification of tree-based binary networks. We will show that a binary phylogenetic network is tree-based if and only if every subset S of its omnians has at least $|S|$ different children. This is followed by a different graph-theoretic characterization of binary phylogenetic networks that are not tree-based.

In [2] it is stated that every non-tree-based binary pylogenetic network can be expanded to become tree-based by the addition of extra leaves, but it does not state how many additional leaves may be necessary. In Section 2.2.2, we will give an upper bound on the number of additional leaves that need to be added. Francis and Steel [2] also asked how many different base trees a network contains. We will give an upper bound on this number at the end of Section 2.2.2.

Moreover, binary phylogenetic networks are not always as realistic as non-binary networks, because of uncertainty in the order of speciation events, and reticulation events. Therefore, after the binary case, the non-binary case will be studied in Section 3. Since there has not been done any research on the tree-basedness of non-binary networks before, we will look at some of the theorems of the binary case and see if they also hold in the non-binary case. We will show that some of these theorems apply in the non-binary case and some do not. In particular, we will give a classification for non-binary phylogenetic networks that are tree-based. In the last section, there are conclusions and a discussion.

2. Binary phylogenetic networks

2.1. **Preliminaries.** First, some essential concepts around binary phylogenetic networks will be explained. Phylogenetic networks contain vertices and directed edges. Directed edges will called arcs from now on.
A (*rooted*) *binary phylogenetic network* is a directed graph N=(V,A), which is acyclic. It contains 1 unique vertex, the *root*, which has in-degree 0 and out-degree 1 or 2. The other vertices in N are one of the following forms:
- a vertex with out-degree 0, a *leaf* (vertices a,b and c are leaves, coloured blue in Figure 1);
- a *reticulation*, a vertex with in-degree 2 and out-degree 1 (the pink coloured vertices in Figure 1);
- a *tree-vertex*, a vertex with in-degree 1 and out-degree 2.

An example of a binary phylogenetic network is given in Figure 1. A (*rooted*) *binary phylogenetic tree* is a binary phylogenetic network that contains no reticulations. Notice that every arc is drawn as an edge, but they are directed to the lowest vertex. This is the case throughout the rest of the report, unless explicitly mentioned otherwise.

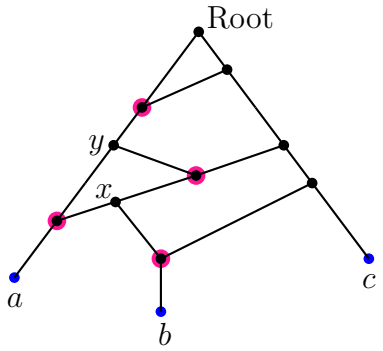
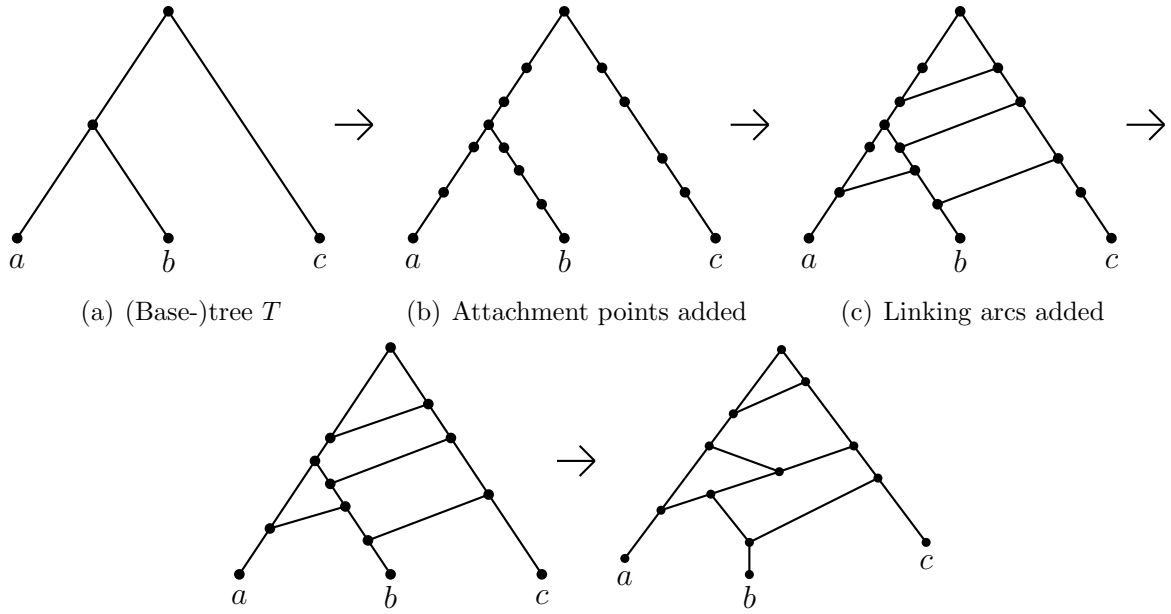


Figure 1. An example of a binary phylogenetic network.

Take $(u, v) = a \in A$, an arc from vertex u to v. Then, a is called an *out-going* arc of u and an *in-coming* arc of v. Vertex u is a *parent* of v and v is called a *child* of u. If there is also an arc $(u, w) \in A$ an arc from vertex u to vertex w, then vertex w and v have a joint parent, so w and v are called *siblings*. When a vertex z has only reticulations as children, then z is called an *omnian*. For example in Figure 1, vertices x and y are omnians, since both children of these vertices are reticulations. Omnians can be reticulations as well, see Figure 4, where both vertices u and v are omnians. The importance of omnians will become clear later on in the report.
A binary phylogenetic network N is *tree-based* with base-tree T, when N can be obtained from T via the following steps:
- Add some vertices to the arcs in T. These vertices, called *attachment points*, have in- and out-degree 1.
- Add arcs, called *linking arcs*, between pairs of attachments points, so that N remains binary and acyclic.
- Suppress every attachment point that is not incident to a linking arc.

An example of the procedure is displayed in Figure 2, in which the tree-basedness of the binary phylogenetic network N of Figure 1 is examined. By definition, we see that N is tree-based, since the last picture in Figure 2 is N and it is obtained from tree T displayed in Figure 2(a).



(a) (Base-)tree T (b) Attachment points added (c) Linking arcs added

(d) Attachment points suppressed (e) N of Figure 1

FIGURE 2. From phylogenetic tree to phyogenetic network in steps (a) to (e), which shows that by definition N is tree-based.

The algorithm to test whether a binary phylogenetic network N is tree-based or not from [2] will be described. There are two steps for the algorithm:
Step 1. Label every outgoing arc of a reticulation and every incoming arc of a tree vertex with t.
Step 2. There are two rules:
R_1. For each reticulation, (i) if one of the incoming arcs has label t then the other incoming arc is assigned label f, and (ii) if one of the incoming arcs has label f then the other incoming arc is assigned label t.
R_2. For each tree-vertex, if one of the outgoing arcs has label f then the other outgoing arc is assigned label t.
Use R_1 and R_2 to label the other arcs in the network.

Now, three cases can occur:
1. Some of the arcs of N have been assigned a label and R_1 and R_2 can no longer be applied to the remaining arcs.
2. All of the arcs of N have been assigned a single label.
3. An arc of N is assigned a label at first and later in the process a different label is assigned by R_1 or R_2.
If case 3 occurs, then N is not tree-based and if case 1 or 2 occurs the network is tree-based. In this report arcs with label t will be coloured green and arcs with label f will be coloured orange.

In the example in Figure 3, Step 1 of the algorithm is executed.

We see that after Step 1, rules R_1 and R_2 can no longer be applied and some arcs remain unlabeled. This means that case 1 occurs, so the binary phylogenetic network N of Figure 1 is again tree-based, now according to the algorithm.

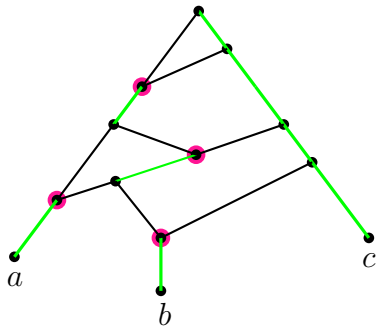
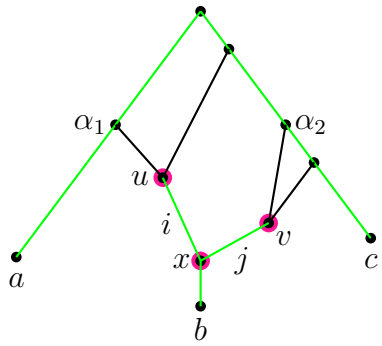
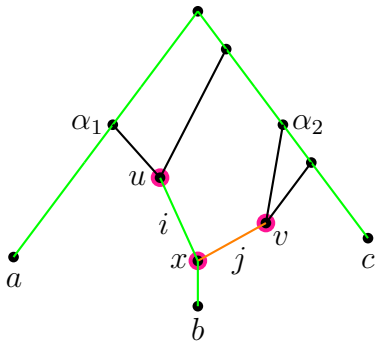


FIGURE 3. The binary phylogenetic network of Figure 1 after Step 1 of the algorithm for tree-basedness.



(a) The network after Step 1 of the algorithm for tree-basedness.



(b) The network after applying R_1 (i) on vertex x in Step 2 of the algorithm for tree-basedness.

FIGURE 4. An example of a non-tree-based binary phylogenetic network.

In Figure 4(a), another example of a binary phylogenetic network is displayed and Step 1 of the algorithm for tree-basedness is applied. We examine incoming arcs i and j of reticulation x. Both arcs are first coloured green, because i is the outgoing arc of reticulation u and j is the outgoing arc of reticulation v. If R_1 (i) is applied on vertex x. It follows that, since arc i is coloured green, that j should be coloured orange, as displayed in 4(b). Therefore, case 3 occurs and it follows that N is not tree-based.

Let $N = (V, A)$ be a binary phylogenetic network. An *antichain* is a set of vertices $K \subseteq V$ for which there is no path from one vertex in K to another vertex in K. Network N satisfies the *antichain-to-leaf property* if for every antichain in N there exists a path from every vertex in K to a leaf, so that these paths are arc-disjoint. Which means, for example, that if there is an antichain of three vertices and there are only two leaves in the network, the network does not satisfy the antichain-to-leaf property.
An example of an antichain can be seen in Figure 4, where vertices α_1 and α_2 form an antichain. The network does not satisfy the antichain-to-leaf property, because when we look at the antichain formed by vertices u and v, there are no arc-disjoint paths to leaves.
A rooted spanning tree τ is a tree that contains all vertices of a phylogenetic network N and a subset of the arcs of N as arcs, so that τ is a tree. A *dummy leaf* of a rooted spanning tree is a vertex that is not a leaf in network N, but is a leaf in τ.
A vertex v is called *stable* if there exists a leaf l for which every path from the root to l passes through v. A network is called *stable* if every reticulation is stable. Let $G = (V, E)$ be a graph. If $v, w \in V$ so that $(v, w) \in E$, then w is a *neighbour* of v. For a set $S \subseteq V$,

the neighbours of S are denoted by $\Gamma(S)$. A *matching* $M \subseteq E$ is a set of edges so that no vertex $v \in V$ is incident with more than one edge in M. A *maximal path* in G is a directed path that is not contained in a larger directed path. When, in graph $G = (V, E)$, there is a path from vertex a to vertex x, without passing through a vertex twice, and there is an edge (a, x), then the path and edge together are called a *circuit*. An example of a circuit can be seen in Figure 11(c). Matchings, maximal paths and circuits will be used in bipartite graphs. Let $N = (V, A)$ be a binary phylogenetic network. Let $B = (U \cup R, E)$ be *the bipartite graph associated to* N. For each vertex $v \in V$, if v is an omnian, put a copy of v in U and if v is a reticulation put a copy of v in R. If v is an omnian as well as a reticulation, we put one copy of v in U and one copy of v in R. There is an edge $\{v, v'\} \in E$ if $(v, v') \in A$, where $v \in U$ and $v' \in R$.

2.2. **Theorems.**

2.2.1. *Previous research.* There were some interesting theorems discovered in earlier analysis. Since we presume that there is yet more to discover around the concepts of stability, antichain and tree-basedness, these will be the main topics that we discuss. First, the relation between a tree-based network and a stable network is considered. From previous research we know the following propositions.

Proposition 2.1. [3] *A stable binary phylogenetic network N has the following property: The child and the parents of a reticulation are tree-vertices.*

Proof. (Adapted from the proof of Proposition 4.1 in [3])
We assume N is stable. The statement is equivalent to that there are no two reticulations in N which have a parent-child relation. Take u and v, two reticulations in N so that u is a parent of v. Because N is stable, every reticulation is stable, so there exists a leaf l for which every path from the root to l goes through u. Let w be the other parent of v. There exists a path P from the root via w and v to leaf l. Then, P does not go through u, which is a contradiction. So there are no two reticulations in N which have a parent-child relation. □

Proposition 2.2. [2] *Consider a binary phylogenetic network N over leaf set X.*
i) If each vertex of N of in-degree 2 has parents that both have out-degree 2, then N is tree-based.
ii) If N has a vertex of in-degree 2 whose parents both have out-degree 1, then N is not tree-based.

Proof. i) Assume that each vertex of N of in-degree 2 has parents that both have out-degree 2. We know by Proposition 2.1 that N is stable. We want to obtain a rooted spanning tree T, so we need to remove incoming arcs of reticulations. Since T is not allowed to contain dummy leaves, we have to make sure that the incoming arcs of reticulations that are removed, are not incident with the same tree-vertex. Take vertex set R containing all reticulations of N, vertex set V containing all tree-vertices of N and E edges that represent the incoming arcs of reticulations of N. Let $S \subseteq R$. The number of edges incident with S is $2\,|S|$ is equal to the number of edges incident with $\Gamma(S)$ which is at most $2\,|\Gamma(S)|$. With Hall's Theorem (stated in Theorem 2.6) it follows that there exists a matching that covers R. We obtain T from N by removing every arc in N that is an edge in the matching. T contains no dummy leaves, because the matching makes sure that no two out-going arcs of a tree-vertex are removed.
ii) Proof can be found in the proof of Proposition 3ii) in [2]. □

Corollary 2.3. *Every binary stable phylogenetic network is tree-based.*

Proof. The proof follows directly from Proposition 2.1 and Proposition 2.2i). □

2.2.2. *New research.* There exist some theorems that help to decide whether a network is tree-based or not. Even an algorithm has been created. Still, there is no simple graph-theoretic characterization of the networks that are tree-based. Therefore, the tree-basedness quality of a network should be explored further, in ways it has not been researched yet. The following theorem presents a different condition for a network to be tree-based.

Theorem 2.4. *Given a binary phylogenetic network N. Let $B = (U \cup R, E)$ be the bipartite graph associated to N. N is tree-based if and only if there exists a matching M in B so that $|U| = |M|$.*

Proof. Assume there exists a matching M in B, so that all omnians are covered by M. Construct a set A of arcs as follows: Add the outgoing arc of every reticulation of N and the incoming arc of all tree-vertices to A. Additionally, add every edge of M as arc to A, that has not yet been added to A. For every reticulation that has not yet been covered, add one of its incomming arcs to A. The tree T, consisting of all vertices of N and the set of arcs A, is a rooted spanning tree, because there is precisely one incoming arc of every vertex contained in T and there are no dummy leaves, because U is covered.
Now, assume that N is tree-based. Let T be a base-tree of N. Colour every edge of B that is an arc in T. When an omnian has out-degree 2 and both arcs are contained in T, decolourise 1 of the 2 arbitrarily in B. T is a base-tree, which means there are no dummy leaves, so all omnians are covered. The fact that T is a base-tree also implies that the vertices in R have in-degree 1. The vertices of out-degree 2 only have 1 coloured edge, so all coloured edges in B form a matching M, so that $|U| = |M|$. □

This theorem can be easily used to verify whether a binary phylogenetic network N is tree-based or not. We will look at an example of a binary phylogenetic network N and the bipartite graph $B = (U \cup R, E)$ associated to N in Figure 5.

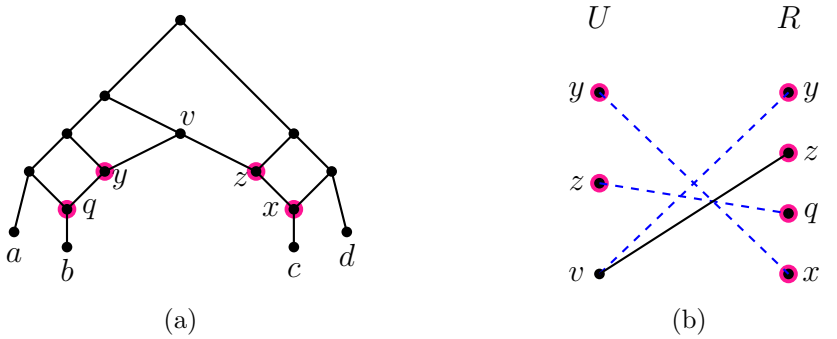


FIGURE 5. Using Theorem 2.4 to show that this is a tree-based binary phylogenetic network.

Since there exists a matching, which is coloured blue and dashed in Figure 5(b), that covers U, the binary phylogenetic network in Figure 5(a) is tree-based. A base-tree T of

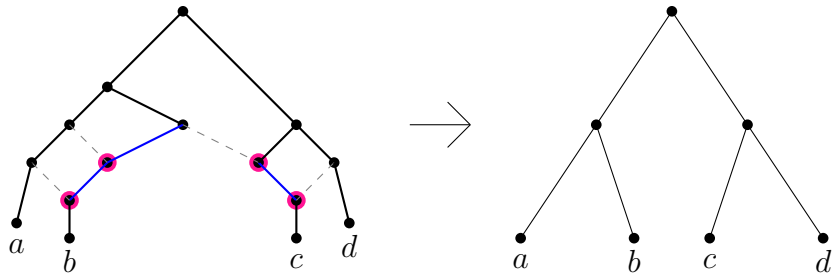


FIGURE 6. A base-tree T of the network in Figure 5(a).

network N can be seen in Figure 6, where the arcs that were edges of the matching are coloured in blue.

Since a binary phylogenetic network that contains no reticulations is a rooted spanning tree, such a network is clearly tree-based. The next theorem shows that this is still the case for all networks with one or two reticulations.

Theorem 2.5. *If a binary phylogenetic network N contains at most 2 reticulations, then N is tree-based.*

Proof. Assume that N contains a at most 2 reticulations.
i) If N contains 1 reticulation, then both parents of this reticulation are tree-vertices and with Proposition 2.2 it follows that N is tree-based.
ii) Consider the case that N contains 2 reticulations x and y. If x and y do not have a parent-child connection, then both parents of x and y are tree-vertices and it follows from Proposition 2.2 that N is tree-based.
Now suppose that x is the parent of y. There are two possibilities, x and y having a joint parent and x and y having a different parent, both displayed in Figure 7.

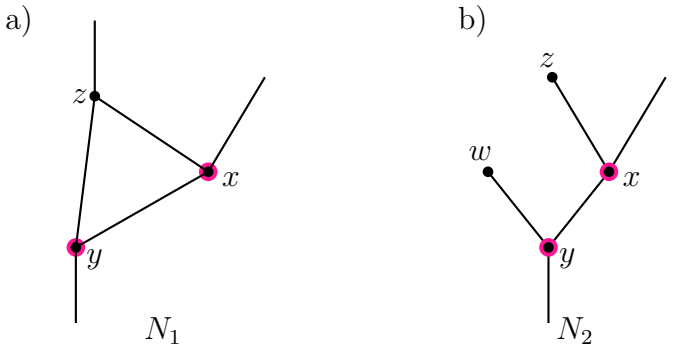
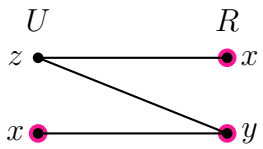


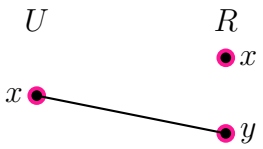
FIGURE 7. The two possibilities that can occur when reticulation x is the parent of reticulation y.

From a) and b) of Figure 7 we create two bipartite graphs, $A = (U \cup R, E)$ associated to N_1 and $B = (U \cup R, E)$ associated to N_2, that are displayed in Figure 8.

a) b)



(a) Bipartite graph A associated to N_1.



(b) Bipartite graph B associated to N_2.

FIGURE 8. The bipartite graphs associated to the partial networks in Figure 7.

In both cases in Figure 8 it is easy to see that there is a matching that covers U. It follows from Theorem 2.4 that N is tree-based. □

We have seen that binary phylogenetic networks containing at most two reticulations are all tree-based. However, Figure 9 shows a part of a network N that contains three reticulations and is not tree-based. So it follows that not all networks with three reticulations are tree-based.

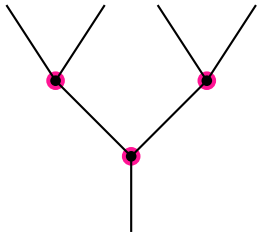


FIGURE 9. Local situation in which N is not tree-based.

In general, we can decide whether the bipartite graph B associated to a network N contains a matching that covers U by using Hall's Theorem, which is stated below.

Theorem 2.6 (Hall's Theorem)**.** [1] *Let $B = (V \cup W, E)$ be a bipartite graph with vertex set V and W. There exists a matching that covers V if and only if for every $V_1 \subseteq V$: $\mid V_1 \mid$ is smaller than or equal to the number of different neighbours of the vertices in V_1.*

Consider Hall's Theorem and Theorem 2.4. Combining those two theorems gives a characterization for a binary phylogenetic network to be tree-based.

Corollary 2.7. *Let N be a binary phylogenetic network and U the set of all omnians of N. Then N is tree-based if and only if for all $S \subseteq U$ the number of different children of S is greater than or equal to the number of omnians in S.*

Proof. Follows directly from Theorem 2.4 and Theorem 2.6. □

An example of how this theorem and corollary can be applied is given in Figure 10, where an example of a binary phylogenetic network N is displayed in (a) and the bipartite graph $B = (U \cup R, E)$ associated to N in (b). The reticulations are coloured in pink, the omnians in blue and the children of the omnians in yellow. For example, vertex f is a reticulation and an omnian, so it is coloured pink and blue.

From the bipartite graph in Figure 10 it follows from Hall's Theorem (Theorem 2.6), with $S = U$, that there exists no matching in B that covers U. Therefore, with Theorem 2.4 it follows that N in Figure 10(a) is not tree-based. Indeed, we can directly see in N that the set $S = \{f, a, i, h, g\}$ of five omnians has only four different children $\{b, c, d, e\}$. Hence this network is not tree-based.

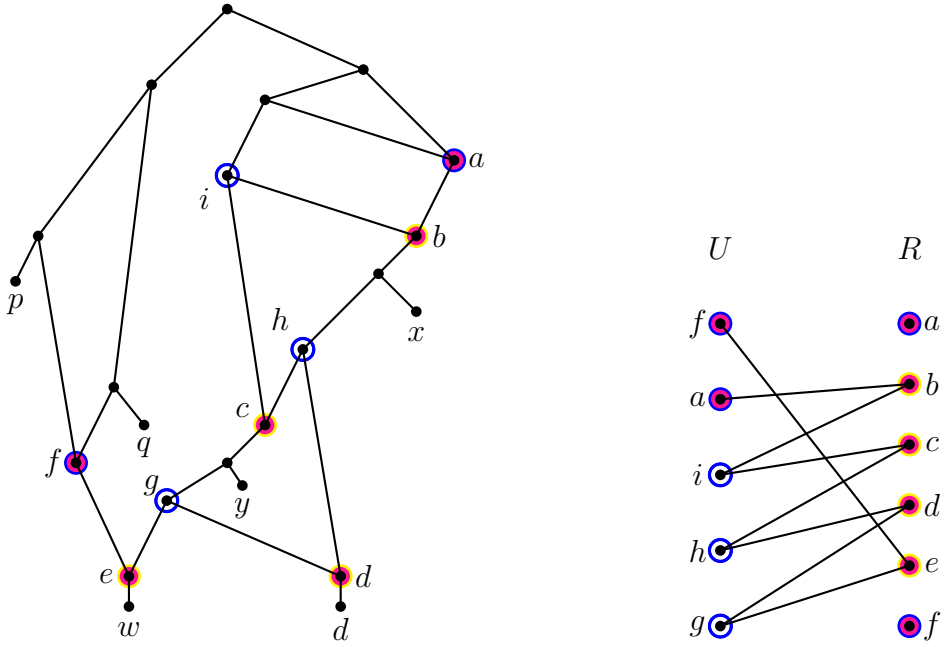


FIGURE 10. Example of a non-tree-based binary phylogenetic network N and the bipartite graph B associated to N.

Proposition 2.2 showed that a binary phylogenetic network is tree-based if for each reticulation both parents are tree-vertices and not tree-based if both parents are reticulations. Therefore, the situation in which a reticulation in N has one parent that is a reticulation and the other a tree-vertex has not yet been specified. The next theorem shows that such networks are tree-based if an additional condition is fulfilled.

Theorem 2.8. *If for every reticulation r in a binary phylogenetic network N, one of the two following cases applies:*
i) Both parents of r are tree-vertices.
ii) One parent of r is a tree-vertex and the sibling of r is a tree-vertex.
Then N is tree-based.

Proof. If every reticulation is of case i) then Proposition 2.2 implies that N is tree-based. Now consider the general case in which N is a binary phylogenetic network containing an arbitrary number of reticulations of cases i) and ii). Let $B = (U \cup R, E)$ be the bipartite graph associated to N. Then, in B, all vertices in the set U of omnians, have degree 1 or 2 and all vertices in the set R of reticulations, have degree 1 or 2. Since in case ii) it is excluded that a reticulation with one parent a reticulation has a sibling that is a reticulation, there are three possibilities that can occur in B.
(a) A reticulation has one reticulation parent and no reticulation sibling. (case ii)
(b) A reticulation r has two parents that are tree-vertices, and with case ii) it follows that siblings of r are tree-vertices or reticulations that also have two parents that are tree-vertices.
(c) Similar to (b) but then reticulations have common parents, which means the number

of reticulations in N is equal to the number of omnians in N, so that a circuit is formed in B.
The possibilities that can occur are displayed in Figure 11. Notice that case (b) and (c) can be infinitely long, this is only an example.

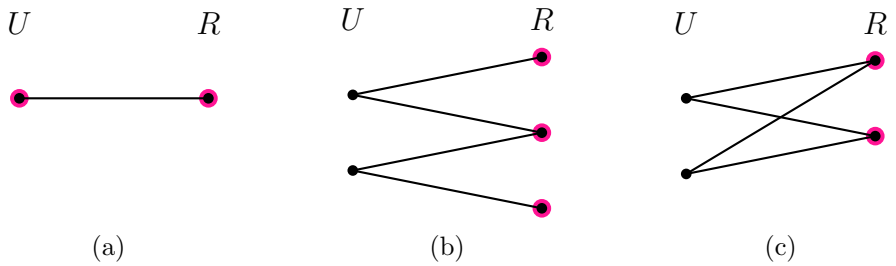


FIGURE 11. Examples of possible maximal paths and circuits in B.

Let $S \subseteq U$. For case (a), the number of edges incident with $S = |S|$ is equal to the number of edges incident with $\Gamma(S)$, which is equal to $|\Gamma(S)|$. For case (b), the number of edges incident with $S = 2\,|S| \leq$ the number of edges incident with $\Gamma(S) \leq 2\,|\Gamma(S)| - 2$. This is equivalent with $|S| = |\Gamma(S)| - 1$. In case (c), the number of edges incident with $S = 2\ |S|$ is equal to the number of edges incident with $\Gamma(S)$, which is equal to $2\ |\Gamma(S)|$. So, when every maximal path in B is examined seperately it follows that for every maximal path in B, $\forall\ S \subseteq U$ the number of neighbours of S is greater than or equal to the number of vertices in S. With Hall's Theorem (Theorem 2.6) the above implies that there exists a matching in B that covers U. With Theorem 2.4 it follows that N is tree-based. □

In previous research, another remarkable proposition was found.

Proposition 2.9. [2] *If a binary phylogenetic network over leaf set X is tree-based, then it satisfies the antichain-to-leaf property.*

On the other hand, if a network is not tree-based, it can still satisfy the antichain-to-leaf property. Yet, until now, only one example had been found, displayed in Figure 12.

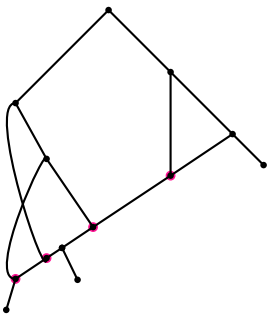


FIGURE 12. Not tree-based binary phylogenetic network satisfying the antichain-to-leaf property [2].

Next, there are two examples displayed in Figure 13(a) and (b), showing a part of a binary phylogenetic network N. These examples are not tree-based, which can be checked using Corollary 2.7.

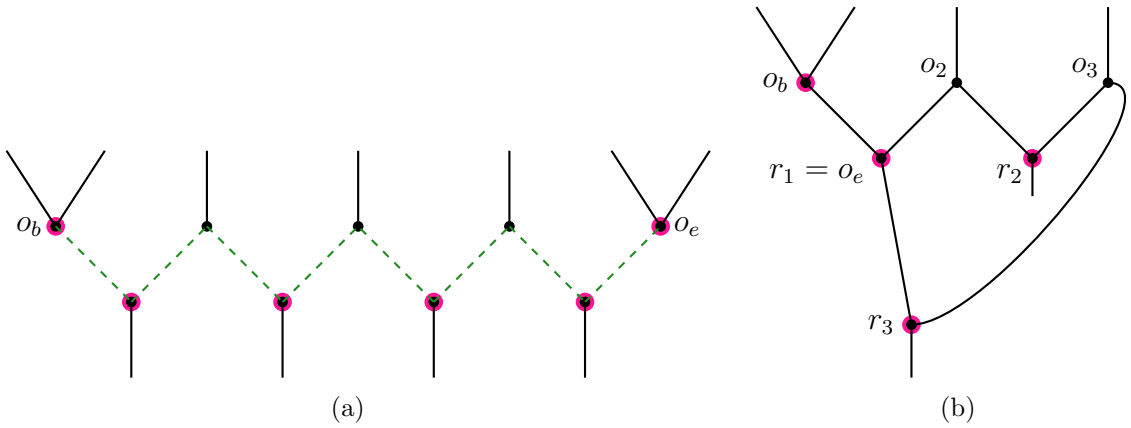


FIGURE 13. Examples of local structures of binary phylogenetic networks that are not tree-based.

However, if there is a directed path in Figure 13(a) from one reticulation to an omnian, so that the network stays acyclic, the network satisfies the antichain-to-leaf property. In addition, if there is a directed path in Figure 13(b) from reticulation r_2 to omnian o_1, we see that the network stays acyclic and also satisfies the antichain-to-leaf property. Both networks that satisfy the antichain-to-leaf property are displayed in Figure 14. This gives us an insight in more partial structures of networks that are not tree-based but do satisfy the antichain-to-leaf property.

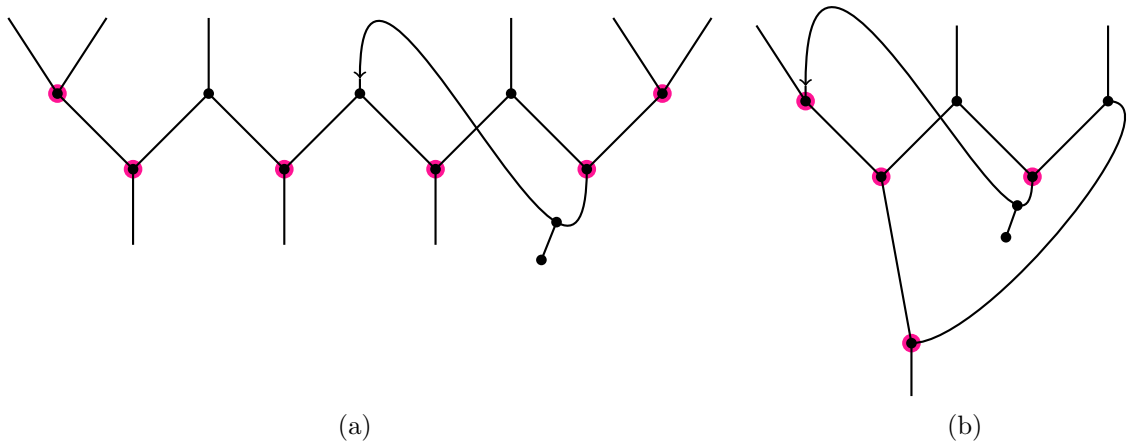


FIGURE 14. Local structures of binary phylogenetic networks that satisfy the antichain-to-leaf property.

While looking at the examples in Figure 13, we see a pattern has emerged in both of them. In (a) the pattern is marked dashed in green. Starting at vertex o_b and ending at vertex o_e, we see a zigzag starting with an omnian, followed by a reticulation, omnian, reticulation (...) and eventually ending with an omnian. The last omnian in the pattern can be a reticulation that is already part of the path, as can be seen in Figure 13(b).

Combining this observation with Theorem 2.4 leads to a graph-theoretic characterization that is stated in the following theorem. The theorem proves that every binary phylogenetic network that is not tree-based has a similar structure as the examples in Figure 13.

Theorem 2.10. *Let N be a binary phylogenetic network and $B = (U \cup R, E)$ the bipartite graph associated to N. N is tree-based if and only if B contains no maximal path which starts and ends in U.*

Proof. Notice that every vertex in B is of degree at most 2, therefore B contains paths and circuits. We distinguish four cases:
i) A maximal path begins and ends in R.
ii) A maximal path begins in U and ends in R.
iii) A maximal path begins and ends in U.
iv) A circuit.

i) All vertices in R are of degree at most 2. Because the maximal path begins and ends in R, all omnians have degree 2. Let $S \subseteq U$. The number of edges incident with $S = 2\,|S| \leq$ the number of edges incident with $\Gamma(S) \leq 2\,|\Gamma(S)|$.
So, $\forall\ S \subseteq U : |\Gamma(S)| \geq |S|$. It follows from Hall's Theorem that there exists a matching in B that covers U.

ii) Let $S \subseteq U$. All vertices in R are of degree 2, except for the reticulation where the maximal path in S ends. All omnians are of degree 2, except for the omnian where the maximal path in S begins. All edges incident with $S = 2\,|S|$, except for the first omnian of the path in S. So, all edges incident with $S = 2\,|S| - 1 =$ all edges incident with $\Gamma\ (S)\ = 2\,|\Gamma(S)| - 1$, because the maximal path in S ends in R, the end-vertex has degree 1. It follows that $|S| = |\Gamma(S)|$. So, $\forall\ S \subseteq U : |\Gamma(S)| \geq |S|$. It follows from Hall's Theorem that there exists a matching in B that covers U.

iii) All omnians in U are of degree 2, except for the omnians where the maximal path begins and ends. All reticulations in R are of degree 2, because the maximal path begins and ends in U. Let $S \subseteq U$, so that $U \subseteq S$. All edges incident with $S = 2\,|S| - 2 =$ all edges incident with $\Gamma(S) = 2\,|\Gamma(S)|$. It follows that $|S| - 1 = |\Gamma(S)|$. So, $|\Gamma(S)| \leq |S|$, from which follows that $\exists\ S \subseteq U : |\Gamma(S)| \leq |S|$. So, it follows from Hall's Theorem that there does not exist a matching in B that covers U.

iv) All vertices in B are of degree 2. Let $S \subseteq U$. The number of edges incident with $S\ =\ 2\ \mid S \mid\ =$ the number of edges incident with $\Gamma(S) = 2\,|\Gamma(S)|$.
So, $\forall\ S \subseteq U : |\Gamma(S)| = |S|$. It follows from Hall's Theorem that there exists a matching in B that covers U.

Hence, there exists a matching in B that covers U precisely if there is no maximal path that starts and ends in U. The theorem now follows from Theorem 2.4. □

From Theorem 2.10 it follows that it is not even necessary to check whether N satisfies the antichain-to-leaf property to see that N is not tree-based.

In [2] it is concluded that every non-tree-based network can be transformed into a tree-based network by adding leaves. Although it has been stated, it is not stated how many leaves should be added. This leads us to the following theorem.

Theorem 2.11. *Let N be a non-tree-based binary phylogenetic network. Let $B = (U \cup R, E)$ be the bipartite graph associated to N. If, for every maximal path in B that begins and ends in U, one tree-vertex and one leaf is added between one of the arcs that is an edge of the maximal path in B, then N becomes tree-based.*

Proof. Assume that N is not tree-based. Then, by Theorem 2.10, there exists at least 1 maximal path in B that begins and ends in U. First, assume there is exactly one maximal path P in B that begins and ends in U. Take omnian $v \in U$ on this path arbitrarily. Then there exists a matching in B that does not cover v. We add a tree-vertex w in N with one child that is a leaf attached to it, between v and a child k, which is a reticulation. Since w is now a child of v and since w is not a reticulation, v is not an omnian anymore. Therefore, $|U| = |U| - 1$. Since v is no longer part of U there are two possibilities for the new situation:
a) 1 edge from v to the child in R is no longer present in B.
b) 2 edges from v to 2 children in R are no longer present in B.
In possibility a), P becomes a path begining in U and ending in R.
In possibility b), P is split into two paths beginning in U and ending in R.
Because there are no maximal paths in B in the new situation that begin and end in U it follows with Theorem 2.10 that N is tree-based.
When there is more than one maximal path that begins and ends in U in B, then for every one of them there should be added a tree-vertex and leaf. It follows in the same way as just described that N is tree-based. □

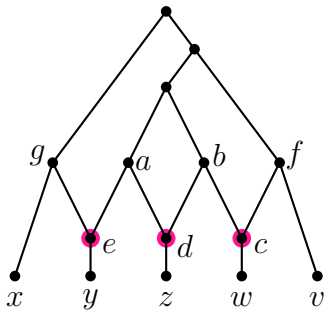


FIGURE 15. An example of a binary phylogenetic network.

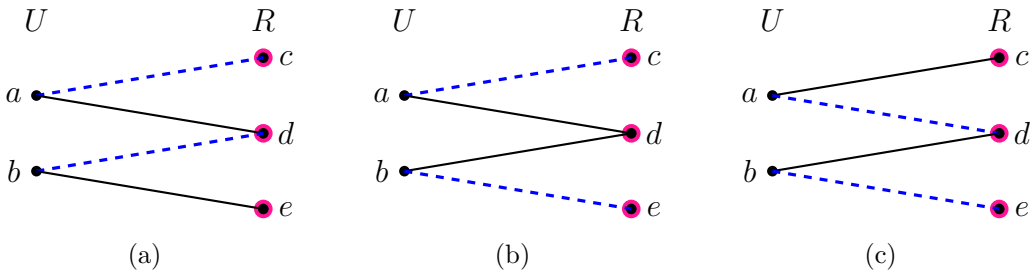


FIGURE 16. A bipartite graph B associated to the network in Figure 15 with its three possible matchings drawn dashed and in blue.

Since we now know that every non-tree-based binary phylogenetic network can be made tree-based, it is interesting to know how many different base trees a tree-based binary phylogenetic network contains. In order to get insight in this number, we will first look

at how many different matchings there exist in the bipartite graph B associated to a tree-based network N. In Figure 15 an example of a binary phylogenetic network N is displayed. Let $B = (U \cup R, E)$ be the bipartite graph associated to N. All possible matchings of B are displayed in Figure 16, where the matchings are dashed and in blue. Notice that every possible matching leaves one reticulation uncovered. This leads us to Theorem 2.13. First, an important observation is done.

Observation 2.12. *Let N be a binary phylogenetic network and B the bipartite graph that is associated to N. Since N is binary, each connected component of B is a path or a circuit.*

Theorem 2.13. *Let N be a binary phylogenetic network and $B = (U \cup R, E)$ the bipartite graph that is associated to N. The number of different matchings in B can be calculated by*

$$2^C \prod_{P \in \mathcal{S}} r(P),$$

with $\mathcal{S}$ the set of maximal paths that begin and end in R, $r(P)$ the number of reticulations contained in a path P and C the number of circuits in the bipartite graph B.

Proof. Since N is tree-based, there exists no maximal path in B that begins and ends in U. For every connected component that is a path in B that begins in U and ends in R, there is only one possible matching that covers all omnians. Therefore, these paths do not have an influence on the total number of different matchings in B.
First, assume there is one connected component that is a path that begins and ends in R. The number of omnians in this path is equal to the number of reticulations minus one ($|U| = |R| - 1$), since the maximal path begins and ends in R. Then, for every reticulation r, there is a matching in B that covers all reticulations except for r. Therefore, the number of different matchings is equal to the number of reticulations in the maximal path. For every circuit in B there are two possible matchings. Now the total number of different matchings in B can be calculated by mutlipying the number of reticulations of every maximal path that begins and ends in R and multiplying this number by two to the power the number of circuits in B. □

Now we we know the number of possible matchings in the bipartite graph B that is associated to binary phylogenetic network N, we want to use this to find the number of possible base-trees of N. Per matching there can be several possible base-trees because, for each reticulation that is not covered by the matching, one of its two incoming edges can be added as arc to the base-tree. Let us return to the example of Figure 15 and the associated figures, Figure 16 and Figure 17. The three different matchings of Figure 16 have been drawn twice in Figure 17 and the different ways of connecting the reticulation that is not covered are coloured dashed and in red. In graphs 1 and 4 in Figure 17, the red dashed line means that the reticulation is connected to the parent that is not an omnian. Notice that the bipartite graphs in Figure 17 with number 2 are the same and the ones with number 3 are the same (except for the colouring of the dashed edges). Therefore, we need to make sure that we do not count any base tree double.

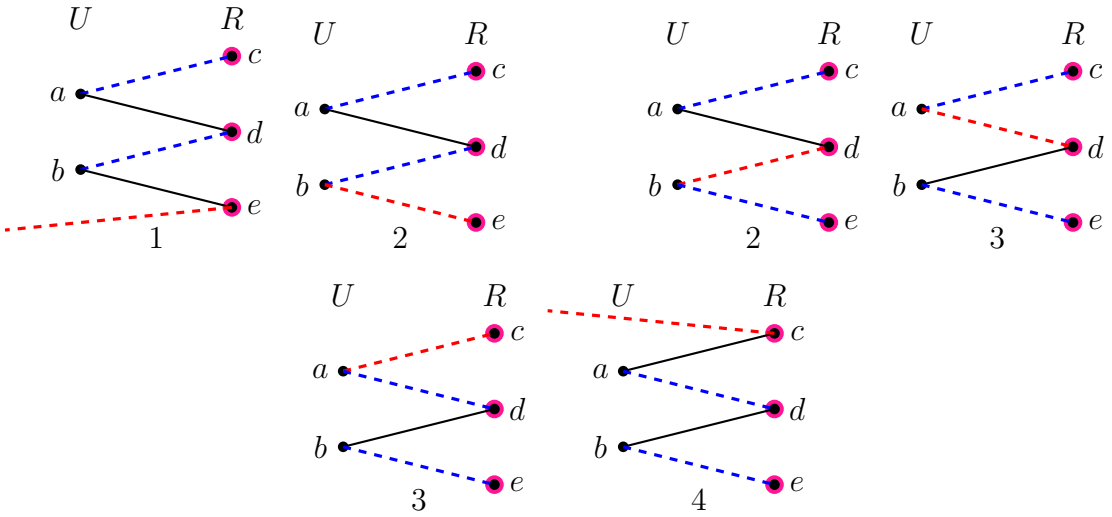


FIGURE 17. The three different matchings in bipartite graph B of the network in Figure 15 with all possible ways of adding the non-covered reticulations.

The previous example gives us an insight in the maximal number of base trees of a tree-based binary phylogenetic network. We will now give an upper bound for the number of base trees that are contained in a tree-based binary phylogenetic network. Notice that a reticulation with degree 0 is actually a maximal path that begins and ends in R.

Theorem 2.14. *Let N be a tree-based binary phylogenetic network and $B = (U \cup R, E)$ the bipartite graph that is associated to N. The maximal number of possible base trees of N is calculated by*

$$2^C \prod_{P \in \mathcal{S}} (r(P) + 1),$$

with $\mathcal{S}$ the set of maximal paths that begin and end in R, $r(P)$ the number of reticulations contained in a path P and C the number of circuits in the bipartite graph B.

Proof. First, assume that there is one maximal path in B that begins and ends in R. From Theorem 2.13 it follows that in that case there are $|R|$ different matchings in B. For the reticulations which have degree 1 in B, the possibility of connecting them to the parent that is not an omnian can only occur when this reticulation is left uncovered in the matching, so these two cases will automatically not be counted double. For the reticulations which have degree 2 in B, if they are left uncoved by the matching, there are two possible ways of connecting each of these reticulations in the base-tree, but then each of these base trees is counted double. Therefore, to calculate the total number of different base trees in the maximal path we get: two times the number of reticulations minus two (the two cases that are automatically not counted double) devided by two because of the double-counting and eventually plus two to count the cases that are automatically not counted double. Which leads to the following calculation:

$$\frac{2\,|R| - 2}{2} + 2 = |R| - 1 + 2 = |R| + 1.$$

When B contains more than one maximal path that begins and ends in R, the above can be applied to every maximal path. Since the number of omnians is equal to the number of reticulations in a circuit, there are two possible matchings, which cover all reticulations and hence leads to two possibilities for the base tree. Therefore, the maximal number

of base-trees of N is calculated by multiplying the number of reticulations plus one of every maximal path that begins and ends in R and multiplying that number by two to the power the number of circuits in B. □

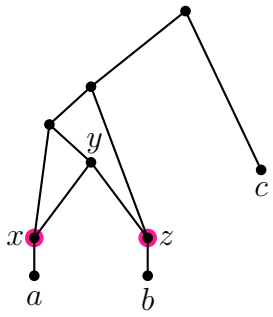


FIGURE 18. An example of a binary phylogenetic network N for counting the number of base trees.

Though it might seem that the upper bound of the number of base trees calculated in the previous theorem is the exact number of base trees, this is not the case. Look at the binary phylogenetic network N in Figure 18. Let B be the bipartite graph associated to N. The number of omnians in N is one. With Theorem 2.13 it follows that there are two possible matchings. In Figure 19, the two matchings are drawn twice, dashed and in blue. In each matching the two possibilities of connecting the non-covered reticulation are drawn in red and dashed.

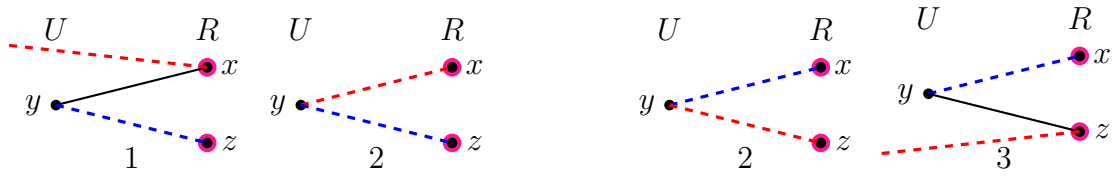


FIGURE 19. The two different matchings in bipartite graph B that is associated to the network in Figure 18 with all possible ways of adding the non-covered reticulations.

Now we draw the three possible base trees from Figure 19 in the network of Figure 18 in blue, the results are displayed in Figure 20. If we simplify the base trees displayed in Figure 20, we get the base tree shown in Figure 21. Notice that this tree represents every base tree of Figure 20, so the binary phylogenetic network of Figure 18 has only one base tree. Which means that Theorem 2.14 indeed gives an upper bound of the number of base trees of a tree-based binary phylogenetic network.

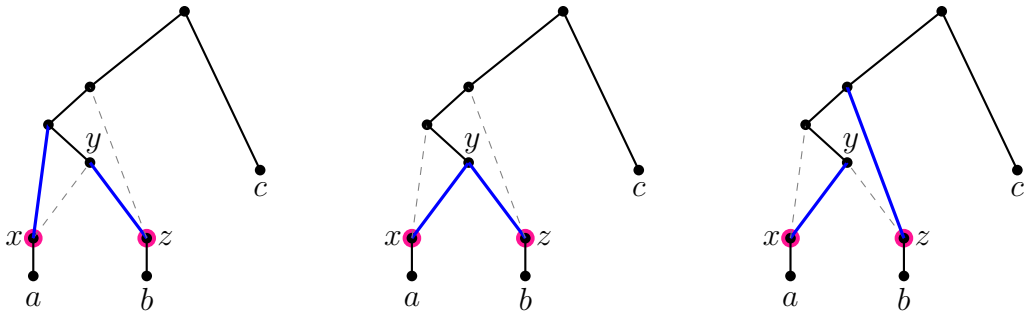


FIGURE 20. The binary phylogenetic network of Figure 18 in which all the different ways of adding the non-covered reticulations are shown.

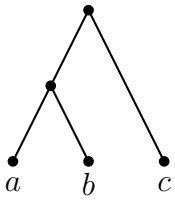


FIGURE 21. The single base tree of the binary phylogenetic network of Figure 18.

3. Non-binary phylogenetic networks

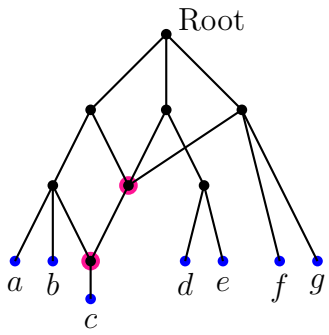


Figure 22. Example of a non-binary phylogenetic network

3.1. **Definitions.** Research has been done about non-binary phylogenetic networks, but not on the tree-basedness of these networks. A (*rooted non-binary*) *phylogenetic network* is a directed, acyclic graph $N = (V, A)$ with the following properties:
- the *root* is a unique vertex with in-degree 0 and out-degree 1 or more;
- vertices with in-degree 1 and out-degree 0, called *leaves*, coloured blue in Figure 22;
- *Reticulations* are vertices with out-degree 1 and in-degree 2 or more, marked in pink in Figure 22;
- vertices with in-degree 1 and out-degree 2 or more, called *tree-vertices*.

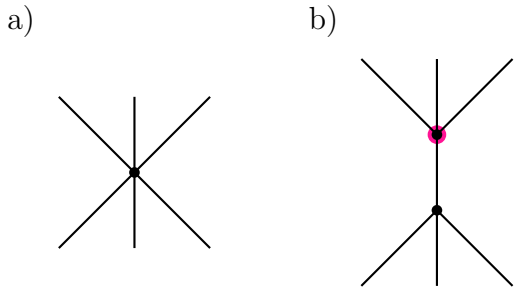


Figure 23. Vertices that are of form a) will be displayed as in b).

Notice that we do not allow vertices like a) in Figure 23 in a phylogenetic network. These kind of vertices will be displayed as b) in Figure 23.

A phylogenetic network N is called *tree-based* with base-tree T, when N can be obtained from T via the following steps:
- Add some vertices to arcs in T. These vertices, called attachment points, have in- and out-degree 1.
- Add arcs, called linking arcs, between pairs of attachments points and from tree-vertices to attachment points, so that N remains binary, acyclic and so that attachment points have in-degree or out-degree 1.
- Supress every attachment point that is not incident to a linking arc.

Let N be a phylogenetic network and $B = (U \cup R, E)$ the bipartite graph that is associated to N. Since vertex $v \in B$ is of degree ≥ 0, there are no paths in B. With the term *maximal path* we refer to every possible maximal path, as defined in Section 2.1, in B seperately. For example, in Figure 25(b), there are two reticulations with degree 0 and three possible maximal paths: $(b, f, g, c, d), (e, c, g, d)$ and (e, f). We call K a *connected component* of B, if K is a maximal path, as just defined, or a circuit in B.
Any definitions from Section 2.1 that have not been mentioned in this section, are defined similarly as in the binary case.

3.2. **Theorems.** We will examine if some of theorems of Section 2.2 hold for non-binary phylogenetic networks as well. First, we look at te stability of networks.

Theorem 3.1. *A stable network N has the following property:*
The child and the parents of every reticulation are tree-vertices.

Proof. The proof of this theorem is nearly equal to the proof of Proposition 2.1. Only w is *another* parent of v instead of *the* other parent. □

So this theorem holds in both the binary and the non-binary case.
Now, the following two questions will be examined:
i) Is every stable phylogenetic network tree-based? (Corollary 2.3 in the binary case)
ii) For phylogenetic network N, is N tree-based if each reticulation of N has parents that are all tree-vertices? (Proposition 2.2 in the binary case)

There is one single example that answers both of the questions displayed in Figure 24.

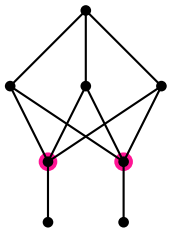


FIGURE 24. Counter example to statements i) and ii).

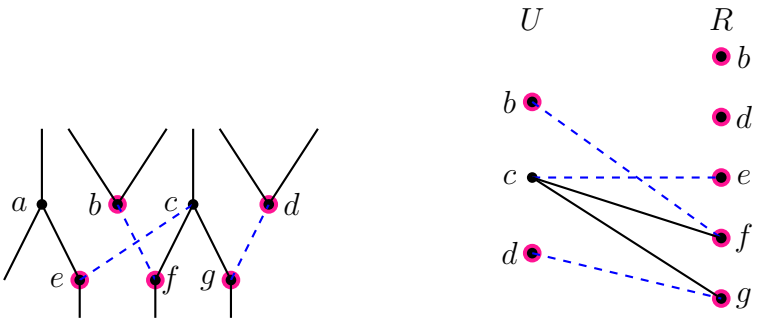
Therefore, questions i) and ii) can be answered with no. They only hold in the binary case. Now we will proof that Theorem 2.4 of the binary case also holds in the non-binary case.

Theorem 3.2. *Given a phylogenetic network N. Let B be the bipartite graph that is associated to N. N is tree-based if and only if there exists a matching M in B so that $|U| = |M|$.*

Proof. Assume there exists a matching M in B, so that all omnians are covered by M. Then it can be proved similarly as in the binary case in Theorem 2.4, that N is tree-based. Assume that N is tree-based. Then it can be proved partially similar as the binary case, that there exists a matching in B that covers all omnians. The only difference is that when an omnian has more than one out-going arc contained in a base-tree T, that only one edge should be coloured and the rest should not be coloured in B. Then all coloured edges in B form a matching M, so that $|U| = |M|$. □

Consider Hall's Theorem (Theorem 2.6) and Theorem 3.2. Combining those two theorems gives a characterization for a non-binary phylogenetic network to be tree-based, which is similar to Corollary 2.7 in the binary case.

Corollary 3.3. *Let N be a phylogenetic network and U the set of all omnians of N. Then N is tree-based if and only if for all $S \subseteq U$ the number of different children of S is greater than or equal to the number of omnians in S.*



(a) The blue dashed lines represent the matching in b)

(b) A matching is B drawn in blue and dashed

FIGURE 25. A partial phylogenetic network and the bipartite graph B that is associated to N.

Theorem 2.10. (binary case) *Let N be a binary phylogenetic network and $B = (U \cup R, E)$ the bipartite graph associated to N. N is tree-based if and only if B contains no maximal path which starts and ends in U.*

When we look at Theorem 2.10 in the binary case, which is restated above, it might be suspected that this would also hold in the non-binary case. We will look at a partial phylogenetic network N, which is displayed in Figure 25(a), and the bipartite graph B that is associated to N, which is displayed in (b). A matching is drawn in B, which is coloured blue and dashed in Figure 25(b). We see that in B there is a maximal path starting and ending in U: starting in b via $f-c-g$ ending in d. Though in the binary case this would mean (with Theorem 2.4) that N is not tree-based, we see in Figure 25(b) that there exists a matching that covers U. With Theorem 3.2 it follows that N is tree-based. Therefore, for a phylogenetic network N and the bipartite graph $B = (U \cup R, E)$ associated to N, if there is a maximal path starting and ending in U, then N can still be tree-based. We see that Theorem 2.4 holds partially in the non-binary case, since the following theorem does hold for the non-binary case.

Theorem 3.4. *Let N be a phylogenetic network and $B = (U \cup R, E)$ the bipartite graph that is associated to N. If B contains no maximal paths which start and end in U, then N is tree-based.*

Proof. Assume that B contains no maximal paths which start and end in U. It follows that B can contain three types of connected components: circuits, maximal paths that start in U and end in R and maximal paths that start and end in R. When we look at every maximal path and circuit separately, it follows similarly as in the proof of Theorem 2.10 that all three types of connected components contain a matching that covers U. With Theorem 3.2 it follows that N is tree-based. □

4. Conclusion and discussion

In the binary case we have seen that a network is tree-based if and only if there exists a matching that covers all omnians in the bipartite graph associated to the network. This theorem, combined with Hall's theorem, gave us an even simpler classification of a tree-based network. It turned out that a binary phylogenetic network N is tree-based if and only if every subset S of the omnians of N has at least $|S|$ different children. Additionally, it was shown that every binary phylogenetic network containing at most two reticulations is tree-based. On the other hand, there is an example of a part of a network containing three reticulations that is not tree-based.
The most important finding is that we have characterised the group of binary phylogenetic networks that are not tree-based. We have shown that all non-tree-based networks contain an alternating path starting with an omnian, reticulation, omnian, ..., reticulation and ending with an omnian. With this outcome, biologists are able to check whether a binary phylogenetic network is equal to a binary phylogenetic tree containing horizontal arcs that represent for example gene-transfer between bacterial species.

Biologists could come across a network that is not tree-based while doing research. We have seen that we can make non-tree-based networks tree-based by adding one tree-vertex with a leaf attached to it for every maximal path in the bipartite graph that begins and ends with an omnian. These leaves were not in the network of the biologist, possibly because a species that should be in the network is already extinct or the sample that the biologist is doing research on might be incomplete, so the sample could miss one or more species. Moreover, our analysis gives an easy way of finding all base-trees, from which biologists are able to check which one is suitable in reality. We have also partially answered the question of Francis and Steel [2] about how many base trees a tree-based network contains, since we have found an upper bound for the number of base trees of a binary phylogenetic network.

In the non-binary case we have also found that a network is tree-based if and only if there exists a matching that covers all omnians in the bipartite graph associated to the network. Some theorems of the binary case did not apply in the non-binary case. For example, not every stable non-binary phylogenetic network is tree-based. In addition, a non-binary phylogenetic network could be not tree-based, even if all parents of every reticulation are tree-vertices. One theorem of the binary case, however, applied partially. Similar to the binary case, if there is no maximal path starting and ending with an omnian in the bipartite graph associated to the non-binary phylogenetic network, then the network is tree-based. However, when there is a maximal path starting and ending with an omnian in the bipartite graph associated to the non-binary phylogenetic network, we have seen that it can still be tree-based.

After the overall process, there still remain some open questions, of which a part originates from [2]. Although we have found an upper bound for the number of base-trees of a tree-based binary phylogenetic network, it is still unknown how many of these base trees are isomorphic. Is there a way to determine which base trees are the same? How many different base trees does a non-binary phylogenetic network have?

We presume that the upper bound that is given for the number of leaves that should be added to a non-tree-based network, in order to make it tree-based, is equal to the minimum of leaves that should be added. But is the given upper bound equal to the minimum of leaves that should be added? We think that this upper bound is equal to the minimum number of leaves that should be added, but this is only a presumption that could be subject for further research.
We have also looked at making non-tree-based networks tree-based by deleting reticulation arcs and suppressing the resulting indegree-1 outdegree-1 vertex (instead of adding leaves). For some networks it looks like this is a sufficient way, but for others it depends if the parent and child of the suppressed vertex are reticulations or tree-vertices. Is there a way to make every non-tree-based network tree-based, by deleting one reticulation arc per maximal path in the associated bipartite graph that begins and ends with an omnian?

Given a binary phylogenetic network and a binary phylogenetic tree, can it be decided in polynomial time whether or not the network is based on the tree? [2] For a given network, we have determined the number of different matchings for every maximal path in the associated bipartite graph B. When B contains a maximal path that begins and ends with an omnian, the network is not tree-based. When B contains a maximal path that begins with an omnian and ends with a reticulation, this gives us one possible matching which covers every omnian and reticulation of this path. For every circuit in B, there are two possible matchings, which again cover every omnian and reticulation of the circuit. For a maximal path in B that begins and ends with a reticulation, a calculation is presented for counting the number of different matchings. Additionally, we have seen how many different ways there are of adding the uncovered reticulation. When there is only one maximal path in B, we can create all the different base-trees and compare every one of them with the given tree. However, if there is more than one maximal path in B, where should we start with deciding which matching is correct to eventually get a base-tree that is similar to the given tree? In my opinion, the research we have done is a great step towards answering the question of [2] entirely. As soon as it can be decided in polynomial time whether or not a given network is based on a given tree, it would be very helpful for biologists, since they can test if a given tree is the correct tree. In addition, they could use this to find out how horizontal transfers in the given tree could run and what that would mean for the evolutionary history of a set of taxa.

References

[1] Peter J Cameron. *Combinatorics: topics, techniques, algorithms*. Cambridge University Press, 1994.

[2] Andrew R. Francis and Mike Steel. Which phylogenetic networks are merely trees with additional arcs? *Systematic Biology*, 64(5):768–777, 2015.

[3] Andreas DM Gunawan, Bhaskar DasGupta, and Louxin Zhang. Stability implies computational tractability: Locating a tree in a stable network is easy. *arXiv preprint arXiv:1507.02119*, 2015.