Bridging between Deviation Indices for Non-Tree-Based Phylogenetic Networks

Takatora Suzuki¹, Han Guo¹, and Momoko Hayamizu¹

¹Affiliation not available

May 25, 2024

Abstract

Phylogenetic networks are a useful model that can represent reticulate evolution and complex biological data. In recent years, mathematical and computational aspects of tree-based networks have been well studied. However, not all phylogenetic networks are tree-based, so it is meaningful to consider how close a given network is to being tree-based; Francis–Steel–Semple (2018) proposed several different indices to measure the degree of deviation of a phylogenetic network from being tree-based. One is the minimum number of leaves that need to be added to convert a given network to tree-based, and another is the number of vertices that are not included in the largest subtree covering its leaf-set. Both values are zero if and only if the network is tree-based. Both deviation indices can be computed efficiently, but the relationship between the above two is unknown, as each has been studied using different approaches. In this study, we derive a tight inequality for the values of the two measures and also give a characterisation of phylogenetic networks such that they coincide.
Bridging between Deviation Indices for Non-Tree-Based Phylogenetic Networks
Takatora Suzuki, Han Guo and Momoko Hayamizu

Abstract—Phylogenetic networks are a useful model that can represent reticulate evolution and complex biological data. In recent years, mathematical and computational aspects of tree-based networks have been well studied. However, not all phylogenetic networks are tree-based, so it is meaningful to consider how close a given network is to being tree-based; Francis–Steel–Semple (2018) proposed several different indices to measure the degree of deviation of a phylogenetic network from being tree-based. One is the minimum number of leaves that need to be added to convert a given network to tree-based, and another is the number of vertices that are not included in the largest subtree covering its leaf-set. Both values are zero if and only if the network is of phylogenetic trees, also known as ‘subdivision trees’ or ‘support trees’, represent reticulate evolution and complex biological data. In re-

Index Terms—Phylogenetic networks, Maximum Covering Subtree Problem, Deviation from tree-based networks, Maximal zig-zag trail decomposition.

I. INTRODUCTION

PHYLOGENETIC trees are a standard model of evolution, which represent branching histories of species. Because of their simplicity, trees are mathematically and computationally tractable, and many methods have been developed and implemented for various computational problems. However, the process of evolution cannot be explained by branching alone when reticulation events such as horizontal gene transfer and hybridisation occur. It is therefore important to explore a more general model beyond trees in order to more accurately describe complex evolution.

Phylogenetic networks provide a more flexible model of evolution than phylogenetic trees because they allow lineages to merge and diverge. Phylogenetic networks can be used not only in situations such as hybridisation, where two lineages merge, but also to deal with complex information, such as inconsistencies between trees inferred from different genes. If it is possible to integrate multiple trees together into a single phylogenetic network by some method, the interpretation of evolutionary scenarios can be made easier by searching for a single probable tree contained in that network. This idea can be seen, for example, in the notion of ‘statistical tree of life’ by O’Malley and Koonin (2011) [1].

Although there are several ways to define phylogenetic trees within a phylogenetic network, there has been much discussion about what might be called ‘phylogenetic spanning trees’, also known as ‘subdivision trees’ or ‘support trees’, of phylogenetic networks [2]–[10]. Following this line of research, Hayamizu (2021) [5] established a canonical way to decompose any rooted binary phylogenetic network (and, more generally, any almost-binary network) into a unique set of subgraphs called maximal zig-zag trails. In [5], Hayamizu used it to obtain the structure theorem for phylogenetic networks and provided linear-time and linear-delay algorithms for many computational problems related to subdivision trees, such as counting, listing, and optimisation. In addition, Hayamizu and Makino (2022) [6] gave a linear-delay algorithm for the more general problem of generating subdivision trees sequentially from an optimal one to a k-th optimal one, according to the values of a certain type of objective function, such as likelihood or log-likelihood. Thus, the structure theorem has led to the development of such fast algorithms, and this has been made possible because it explicitly characterises the set of all phylogenetic spanning trees of a given network N. We note that it is not possible to obtain similar algorithms for the set of all phylogenetic subtrees of N (i.e. the so-called ‘displayed’ trees of N) rather than spanning trees. In fact, the problem of just counting the subtrees displayed by N is known to be #P-complete [11] (see e.g. [12], [13] for other hardness results related to displayed trees).

Phylogenetic networks with a phylogenetic spanning tree are called tree-based networks, but not all networks are tree-based. Francis–Semple–Steel (2018) [14] introduced several indices to measure how far a given network N deviates from being tree-based, such as the minimum number δ∗(N) of extra leaves that need to be added to make N tree-based, and the minimum difference η∗(N) between the number of vertices of N and the number of vertices of a subtree of N with the same root and the same set of leaves as N. Hayamizu [5] proved δ∗(N) = |W|, where |W| is the number of maximal W-fences of N, and provided a linear-time algorithm for computing δ∗(N). The problem of computing η∗(N) is called the Maximum Covering Subtree (MAX-CST), and Davidov et al. [15] showed that MAX-CST can be solved in polynomial time even for non-binary N by using an algorithm for solving the minimum cost flow problem. The proposed method converts N into a flow network where some edges have negative weights and uses for example Bellman-Ford algorithm. Although this takes cubic time even if N is binary, an algorithm for solving the minimum cost flow problem in almost linear-time has been recently developed [16], so the computational complexity of their method can be improved over that described in [15].

In this paper, in order to investigate the relationship between the two deviation measures δ∗(N) and η∗(N), we study the number η*(N) of vertices of N that are not covered by a maximum covering subtree, using a different approach from
the one used in [15]. Assuming that $N$ is binary or almost-binary, we use the above-mentioned canonical decomposition of $N$ into the set of unique maximal zig-zag trails [5]. Focusing on the number $|W_N|$ of W-fences in $N$, we prove that $|W_N|$ gives a lower bound on $\eta^*(N)$, i.e. $\eta^*(N) \geq |W_N| = \delta^*(N)$. This lower bound is tight, i.e. there exist phylogenetic networks $N$ for which $\eta^*(N) = |W_N|$ (e.g. tree-based networks). We provide a necessary and sufficient condition for $N$ to satisfy $\eta^*(N) = |W_N|$, thus introducing a class of phylogenetic networks for which $\eta^*(N)$ can be computed in linear time. Such phylogenetic networks $N$ can be seen as a generalisation of tree-based networks.

II. PRELIMINARIES

A. Graph theoretical terminology

In this paper, we consider only (weakly) connected, finite, simple, directed acyclic graphs which we now define. A directed graph is defined as an ordered pair $(V,E)$ of a set $V$ of vertices and a set $E$ of directed edges, whereas an undirected graph consists of a set of vertices and a set of undirected edges. An undirected graph is connected if there exists a path between any pair of vertices, and a directed graph $G$ is (weakly) connected if the undirected graph obtained by ignoring the direction of each edge of $G$ is connected. From now on, we will use the terms ‘graph’ and ‘edge’ to mean a directed graph and directed edge, respectively. Given a graph $G$, its vertex-set and edge-set are denoted by $V(G)$ and $E(G)$, respectively. A graph $G$ is said to be finite if $V(G)$ and $E(G)$ are finite sets. An edge oriented from vertex $u$ to vertex $v$ is denoted by $(u,v)$. A graph is said to be simple if it does not contain any loops (i.e. edges that start and end at the same vertex) or multiple edges (i.e. multiple directed edges between the same pair of vertices).

For an edge $(u,v)$ of a directed graph with $u \neq v$, we say that $u$ is a parent of $v$ and $v$ a child of $u$. Given an edge $e = (u,v)$, we often use the notation tail$(e)$ and head$(e)$ to mean the starting point $u$ and the ending point $v$, respectively. For a vertex $v$ of a graph $G$, the in-degree of $v$ (in $G$), denoted by indeg$_G(v)$, is defined to be the number of edges of $G$ with head$(e) = v$. Similarly, the out-degree of $v$ (in $G$), denoted by outdeg$_G(v)$, is defined to be the number of edges of $G$ with tail$(e) = v$.

A (directed) path is a directed graph $G$ that can be represented by an alternating sequence of vertices and consecutive edges $v_1,(v_1,v_2),v_2,...,(v_{k-1},v_k),v_k$, where all vertices are distinct and we have $(\text{indeg}_G (v_1), \text{outdeg}_G (v_1)) = (0,1)$, $(\text{indeg}_G (v_k), \text{outdeg}_G (v_k)) = (1,0)$ and $(\text{indeg}_G (v_i), \text{outdeg}_G (v_i)) = (1,1)$ for any vertex $v_i$ other than $v_1,v_k$. A (directed) cycle is a directed graph $G$ that can be represented by an alternating sequence of vertices and consecutive edges $v_1,(v_1,v_2),v_2,...,(v_{k-1},v_k),v_k$, where all vertices are distinct, $v_1 = v_k$, and $(\text{indeg}_G (v_i), \text{outdeg}_G (v_i)) = (1,1)$ for any vertex $v_i$ of $G$. For two graphs $G$ and $H$, $H$ is called a subgraph of $G$ if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. A graph is acyclic if it does not contain any directed cycles. A (directed) tree is a directed acyclic graph $G$ such that there exists a unique vertex $v$ of $G$ with indeg$_G(v) = 0$ and for any vertex $v$ of $G$ other than $v$, we have indeg$_G(v) = 1$. A subgraph $H$ of a graph $G$ is called a subtree of $G$ if $H$ is a directed tree. A subtree $H$ of $G$ with $V(H) = V(G)$ is called a spanning tree of $G$. For a vertex $v$ of a graph $G$, removing $v$ means deleting $v$ and all the edges starting or ending at $v$, and we write $G - \{v\}$ to represent the subgraph of $G$ obtained by removing $v$ from $G$.

For a graph $G = (V,E)$ and any subset $E' \subseteq E$, $E'$ is said to induce the subgraph $(V',E')$ of $G$, where $V' := \{v \in V \mid v = \text{head}(e) \text{ or } v = \text{tail}(e) \text{ for some } e \in E'\}$. For notational simplicity, we write $G[E']$ to represent the edge-induced subgraph $(V',E')$ of $G$. For a graph $G = (V,E)$ and any partition $\{E_1,...,E_n\}$ of $E$, the collection $\{G[E_1],...,G[E_n]\}$ of edge-induced subgraphs is called a decomposition of $G$.

Two directed graphs $G$ and $H$ are said to be isomorphic if there exists a bijection between $V(G)$ and $V(H)$ that yields a bijection between $E(G)$ and $E(H)$. In other words, $G$ and $H$ are isomorphic if there exists a labelling map $f : V(G) \rightarrow V(H)$ such that there is a one-to-one correspondence between $V(G)$ and $V(H)$ and for any $v_i$ and $v_j$ in $V(G)$, $(v_i,v_j)$ is in $E(G)$ if and only if $(f(v_i),f(v_j))$ is in $E(H)$, including the loops and parallel edges if they exist. For an edge $(u,v)$ of a graph $G$, subdividing edge $(u,v)$ refers to the operation of introducing a new vertex $p$ and replacing $(u,v)$ with the new consecutive edges $(u,p)$ and $(p,v)$. Any graph that can be obtained by subdividing each edge of $G$ zero or more times is called a subdivision of $G$. Conversely, for a vertex $p$ of a graph $G$ with indeg$_G(p) = \text{outdeg}_G(p) = 1$ and two consecutive edges $(u,p)$ and $(p,v)$, smoothing vertex $p$ refers to the operation of removing the vertex $p$ together with the edges $(u,p)$ and $(p,v)$ and then adding a new edge $(u,v)$. Two graphs are said to be homeomorphic if they are isomorphic after smoothing all vertices of in-degree one and out-degree one.

B. Phylogenetic networks

Throughout this paper, $X = \{1,...,n\}(n \geq 1)$ represents a non-empty finite set, which can be biologically interpreted as a set of $n$ present-day species. In this paper, we slightly generalise the notion of ‘binary’ phylogenetic $X$-networks and consider ‘almost-binary’ phylogenetic $X$-networks. This is because, as discussed in [5], the structure theorem for rooted binary phylogenetic $X$-networks can be applied not only to binary but also to almost-binary $X$-networks. Note that almost-binary networks are in the class of ‘semi-binary’ networks introduced in [17].

**Definition 1** (See also Figure 1). A rooted almost-binary phylogenetic $X$-network is defined to be any simple directed acyclic graph $N = (V,E)$ with the following properties:

1) There exists a unique vertex $\rho \in V$ with indeg$_N(\rho) = 0$, outdeg$_N(\rho) \in \{1,2\}$;

2) There exists a bijection between $X$ and $L := \{v \in V \mid \text{indeg}_N(v) = 1, \text{outdeg}_N(v) = 0\}$;

3) For any $v \in V \setminus (X \cup \{\rho\})$, indeg$_N(v) \in \{1,2\}$ and outdeg$_N(v) \in \{1,2\}$ hold.

According to Definition 1, the vertex set $V$ of a rooted almost-binary phylogenetic $X$-network consists of three types
of vertices: a unique vertex \( \rho \) called the root of \( N \), which represents the most recent common ancestor of the species in the set \( X \); the vertices in \( L \), called leaves of \( N \), that have no descendants in \( N \) and thus can be identified with the present-day species in \( X \); and the remaining non-root, non-leaf vertices, which have one or two incoming edges and one or two outgoing edges.

As illustrated Figure 1, a rooted almost-binary phylogenetic \( X \)-network \( N \) is called a rooted binary phylogenetic \( X \)-network if \( N \) contains neither a vertex \( v \) with \( \text{indeg}_N(v) = \text{outdeg}_N(v) = 2 \) nor a vertex \( v \) with \( \text{indeg}_N(v) = \text{outdeg}_N(v) = 1 \). Moreover, if \( N \) does not also contain a vertex \( v \) with \( (\text{indeg}_N(v),\text{outdeg}_N(v)) = (2,1) \), then \( N \) is called a rooted binary phylogenetic \( X \)-tree. A vertex \( v \) with \( (\text{indeg}_N(v),\text{outdeg}_N(v)) = (1,2) \) is called a tree vertex of \( N \), whereas a vertex \( v \) with \( (\text{indeg}_N(v),\text{outdeg}_N(v)) = (2,1) \) is called a reticulation of \( N \). Intuitively, a tree vertex is a branching point in an evolutionary history, whereas a reticulation can be viewed as a point where two different lineages merge.

![Fig. 1. Examples of a rooted binary phylogenetic X-tree (a), a rooted binary phylogenetic X-network (b), a rooted almost-binary phylogenetic X-network (c) where \( X = \{1, \ldots, 6\} \).](image)

Fig. 1. Examples of a rooted binary phylogenetic X-tree (a), a rooted binary phylogenetic X-network (b), a rooted almost-binary phylogenetic X-network (c) where \( X = \{1, \ldots, 6\} \).

III. MAXIMUM COVERING SUBTREE PROBLEM AND KNOWN RESULTS

In this section, we formally describe the Maximum Covering Subtree Problem, which was originally described by Francis–Semple–Steel [14], and summarise relevant known results from the work by Davidov et al. [15].

We first recall the definition of tree-based networks and subdivision trees, as maximum covering subtrees are closely related to these concepts. A rooted almost-binary phylogenetic \( X \)-network \( N \) is called a tree-based network (on \( X \)) if \( N \) has a spanning tree (i.e. a subtree \( T \) with \( V(T) = V(N) \)) that is a subdivision of some rooted binary phylogenetic \( X \)-tree, and in this case \( T \) is called a subdivision tree of \( N \) (see Figure 2 for an illustration). Note that this definition implies that a subdivision tree \( T \) of \( N \) not only contains all vertices of \( N \), but also shares the root and leaf-set \( X \) with \( N \).

A covering subtree of a rooted (not necessarily binary) phylogenetic \( X \)-network \( N \) is a subtree whose root and leaf-set is the same as those of \( N \). A covering subtree of \( N \) with the largest number of vertices is called a maximum covering subtree of \( N \). We now state the problem as follows. See also Figure 3 for an illustration.

**Maximum Covering Subtree Problem (MAX-CST)**

Given a rooted (not necessarily binary) phylogenetic \( X \)-network \( N \), compute the number \( \eta^*(N) := |V(N)| - |V(T)| \), where \( T \) denotes a maximum covering subtree of \( N \).

![Fig. 3. An illustration of MAX-CST where \( \eta^*(N) = 1 \). (a): A rooted almost-binary phylogenetic X-network \( N \) with 12 vertices, which is not tree-based. (b): A covering subtree with 10 vertices. (c): A covering subtree with 11 vertices, which is maximum.](image)

Fig. 3. An illustration of MAX-CST where \( \eta^*(N) = 1 \). (a): A rooted almost-binary phylogenetic X-network \( N \) with 12 vertices, which is not tree-based. (b): A covering subtree with 10 vertices. (c): A covering subtree with 11 vertices, which is maximum.

MAX-CST asks for the number \( \eta^*(N) \) of vertices of \( N \) that are not covered by a maximum covering subtree of \( N \). We note that \( \eta^*(N) = 0 \) holds if and only if \( N \) is a tree-based network and thus \( \eta^*(N) \) can be interpreted as quantifying how far \( N \) deviates from being tree-based. Davidov et al. [15] provided an algorithm that solves MAX-CST (and also finds a maximum covering subtree \( T \)) in \( O(|V(N)|^2|E(N)|) \) time,
based on a linear-time reduction to a minimum cost flow problem. Under the assumption that the input network \( N \) is almost-binary, the algorithm runs in \( O(|V(N)|^3) = O(|E(N)|^3) \) time if Bellman–Ford algorithm is used to handle negative edge-weights. As mentioned in Section I, there was a recent advancement [16] that could improve this time complexity.

IV. CANONICAL DECOMPOSITION OF ROOTED ALMOST-BINARY PHYLOGENETIC NETWORKS AND ITS IMPLICATIONS FOR TREE-BASED NETWORKS

We recall the relevant definitions and results that will be used in this paper from [5] without providing proofs, thus ensuring our exposition is self-contained. Although the results in [5] were originally intended for rooted binary phylogenetic \( X \)-networks, all of them apply to rooted almost-binary phylogenetic \( X \)-networks as well, as pointed out in Section 6 of [5].

**Definition 2.** For a rooted almost-binary phylogenetic \( X \)-network \( N \), a *zig-zag trail* in \( N \) is a connected subgraph \( Z \) of \( N \) with \( |E(Z)| \geq 1 \), such that there is a labelling \( (e_1, \ldots, e_m) \) of the edges of \( Z \) where any two consecutive edges \( e_i \) and \( e_{i+1} \) of \( Z \) satisfy either \( \text{head}(e_i) = \text{head}(e_{i+1}) \) or \( \text{tail}(e_i) = \text{tail}(e_{i+1}) \). A zig-zag trail \( Z \) in \( N \) is said to be *maximal* if \( N \) contains no zig-zag trail \( Z' \) such that \( Z \) is a proper subgraph of \( Z' \).

Any zig-zag trail \( Z \) in a rooted binary network \( N \) can be denoted by an alternating sequence of (not necessarily distinct) vertices and distinct edges. Note that a maximal zig-zag trail may consist of a single edge. However, if we drop the edges from the notation, we can more concisely express \( Z \) as a sequence of vertices \( v_1 > v_2 < v_3 > v_4 < \cdots > v_m-1 < v_m \) (or in reverse order, \( v_m > v_{m-1} < \cdots > v_4 < v_3 > v_2 < v_1 \)). We sometimes even write \( (e_1, \ldots, e_m) \) when there is no confusion without specifying the edge directions. Every maximal zig-zag trail falls into one of the four types defined in Definition 3 (see also Figure 4). A similar definition is also found in [4].

**Definition 3.** Let \( N \) be a rooted almost-binary phylogenetic \( X \)-network and let \( Z \) be a maximal zig-zag trail in \( N \). The four types of maximal zig-zag trails are defined as follows:

- **Z** is a *crown* if \( Z \) has an even number \( |E(Z)| \geq 4 \) of edges and \( Z \) can be represented as
  \[
  v_1 < v_2 > v_3 < \cdots < v_{2k-1} < v_{2k} > v_{2k+1} = v_1.
  \]
- **Z** is a *M-fence* if \( Z \) has an even number \( |E(Z)| \geq 2 \) of edges and \( Z \) can be represented as
  \[
  v_1 < v_2 > v_3 < \cdots < v_{2k-1} < v_{2k} > v_{2k+1} = v_1.
  \]
- **Z** is an *N-fence* if \( Z \) has an odd number \( |E(Z)| \geq 1 \) of edges and \( Z \) can be represented as
  \[
  v_1 < v_2 > v_3 < \cdots < v_{2k-1} < v_{2k} = v_1.
  \]
- **Z** is a *W-fence* if \( Z \) has an even number \( |E(Z)| \geq 2 \) of edges and \( Z \) can be represented as
  \[
  v_1 > v_2 < v_3 > \cdots < v_{2k-1} < v_{2k} > v_{2k+1} = v_1.
  \]

**Remark 4.** We note that fences do not necessarily have to have the shapes of the letters \( M, N, \) and \( W \) as described in Figure 4. In other words, maximal zig-zag trails are not necessarily bipartite graphs, i.e., a directed graph \( G \) whose vertex-set can be partitioned into two non-empty disjoint sets \( V_1 \) and \( V_2 \) such that every edge \((u, v)\) of \( G \) satisfies \( u \in V_1 \) and \( v \in V_2 \). For instance, see the M-fence \( Z_3 \) in the network on the right of Figure 5(b).

In this paper, our entire argument will be based on Theorem 5, which provides a canonical way of decomposing any rooted almost-binary phylogenetic \( X \)-network into a unique set of maximal zig-zag trails. Theorem 5 was originally proved for binary networks in [5], but as was noted in Section 6 of [5], all the arguments, statements and algorithms in [5] hold true for any almost-binary networks. Therefore, we here do not repeat the proof of Theorem 5 although it is short and straightforward. The interested reader may refer to the proof of Theorem 4.2 in [5].

**Theorem 5** ([5]). For any rooted almost-binary phylogenetic \( X \)-network \( N \), there exists a unique decomposition \( \mathcal{Z} = \{Z_1, \ldots, Z_k\} \) of \( N \) such that each \( Z_i \in \mathcal{Z} \) is a maximal zig-zag trail in \( N \) (see Figure 5 for an illustration).

The unique decomposition of \( \mathcal{Z} \) of \( N \) in Theorem 5 is called the maximal zig-zag trail decomposition of \( N \). Using such a canonical decomposition, Hayamizu [5] provided an explicit characterisation of the set of all subdivision trees (spanning trees with leaf-set \( X \)) in a given tree-based phylogenetic network \( N \) on \( X \) as a direct product of families of sets of possible edge-choices within each subgraph \( Z_i \) of \( N \).

Such a unique decomposition and characterisation in direct product form is in the spirit of various structure theorems in mathematics, such as the structure theorem for finite Abelian groups, which states that every finite Abelian group can be uniquely decomposed as a direct product of cyclic groups. In light of this, when we speak of the “structure theorem" for rooted binary (or almost-binary) phylogenetic networks, it would be appropriate to refer to both the uniqueness of the decomposition (Theorem 5) and the direct-product characterisation of the set of all subdivision trees (Theorem 4.8 in [5]). For convenience, however, we often use the name “structure theorem" to refer only to Theorem 5 and this was the case even in the original paper [5]. In this paper we only need Theorem 5, so we do not restate the full statement of the structure theorem here. For readers interested in the rest of
the structure theorem, we refer to Theorem 4.8 in [5].

Fig. 5. An illustration of the maximal zig-zag trail decomposition of a rooted binary tree-based X-network (a) and a rooted almost-binary phylogenetic X-network (b).

Proposition 6 has been stated and proved in independent studies using different proofs and terminology, Hayamizu [5] described that this proposition still holds for almost-binary phylogenetic networks.

**Proposition 6** ([Zhang (2016) [4], Jetten and van Iersel (2016) [10], Hayamizu (2021) [5]]) Let $\mathcal{N}$ be a rooted binary phylogenetic X-network and $\mathcal{Z} = \{Z_1, \ldots, Z_k\}$ be the maximal zig-zag trail decomposition of $\mathcal{N}$. Then, $\mathcal{N}$ is a tree-based network on $X$ if and only if no element $Z_i \in \mathcal{Z}$ is a W-fence.

Proposition 6 is just one of many interesting consequences of Theorem 5 (see [5], [6] for many other results and various linear-time algorithms derived from Theorem 5) but is worth recalling before considering the Maximum Covering Subtree Problem (MAX-CST). This is because Proposition 6 means that $\eta^*(\mathcal{N}) = 0$ holds if and only if the number $|\mathcal{W}|$ of W-fences in $\mathcal{N}$ is exactly zero, where $\mathcal{W}$ denotes the set of W-fences in $\mathcal{N}$. This implies that if $\mathcal{N}$ is a tree-based network, then we have $|\mathcal{W}| = \eta^*(\mathcal{N}) = 0$.

Since the number $|\mathcal{W}|$ of W-fences in $\mathcal{N}$ can be computed in linear time [5], a natural question arises as to when $|\mathcal{W}| = \eta^*(\mathcal{N})$ holds. In Section V, we will provide a necessary and sufficient condition for when $|\mathcal{W}| = \eta^*(\mathcal{N})$ holds (Theorem 19).

V. Results: New Approach for MAX-CST Using Canonical Decomposition of Phylogenetic Networks

In this section, we will use the notation $\mathcal{W}$ and $\mathcal{M}$ to represent the sets of W-fences and M-fences in $\mathcal{N}$. Also, since we will only consider maximal zig-zag trails, when no confusion arises, the word “maximal” may be omitted from the description of crowns, M-fences, N-fences and W-fences in $\mathcal{N}$. In Section V-A, we prove $|\mathcal{W}| \leq \eta^*(\mathcal{N})$ (Theorem 11). In Section V-B, we provide a necessary and sufficient condition for when $|\mathcal{W}| = \eta^*(\mathcal{N})$ holds (Theorem 19).

A. Bounding $\eta^*(\mathcal{N})$ from below by the number $|\mathcal{W}|$ of W-fences

As noted earlier in Remark 4, a zig-zag trail is not necessarily a bipartite graph as depicted in Figure 4. In the case of $Z_3$ of Figure 5(b), for example, there is no obvious vertical relationship between the vertices of $Z_3$. However, since any maximal zig-zag trail $Z_t$ can be expressed as in Definition 3, we can consider ‘upper’ and ‘lower’ vertices of $Z_t$ that are formally defined in Definition 7.

**Definition 7.** Let $Z$ be a maximal zig-zag trail in a rooted almost-binary phylogenetic network $\mathcal{N}$ as described in Definition 3. If $Z$ is a W-fence, we call each odd-labelled vertex $v_{2i-1}$ (resp. even-labelled vertex $v_{2i}$) of $Z$ an upper vertex (resp. lower vertex) of $Z$. Otherwise, we call each even-labelled vertex $v_{2i}$ (resp. odd-labelled vertex $v_{2i-1}$) of $Z$ an upper vertex (resp. lower vertex) of $Z$. The set of upper vertices and the set of lower vertices of $Z$ is denoted by $V_u(Z)$ and by $V_l(Z)$, respectively.

By Definition 7, each upper vertex $v$ of a maximal zig-zag trail $Z$ satisfies $\text{outdeg}_Z(v) \geq 1$, and each lower vertex $v$ of $Z$ satisfies $\text{indeg}_Z(v) \geq 1$. Therefore, an upper vertex $v$ of a maximal zig-zag trail $Z$ can be a lower vertex of $Z$ at the same time (see the above-mentioned $Z_3$ in Figure 5). The next proposition follows from Definition 3.

**Proposition 8.** Let $Z$ be a maximal zig-zag trail in a rooted almost-binary phylogenetic network $\mathcal{N}$. Then,

$$|V_u(Z)| - |V_l(Z)| = \begin{cases} 1 & \text{if } Z \text{ is an M-fence;} \\ 0 & \text{if } Z \text{ is a crown or an N-fence;} \\ -1 & \text{if } Z \text{ is a W-fence.} \end{cases}$$ (1)

Fig. 6. An illustration of Proposition 8: upper and lower vertices in a crown (a), an M-fence (b), an N-fence (c) and a W-fence (d).

**Proposition 9.** For a rooted almost-binary phylogenetic X-network $\mathcal{N}$, let $T$ be a maximum covering subtree of $\mathcal{N}$. Then, for any maximal zig-zag trail $Z$, $|V_u(Z) \cap V(T)| \leq |V_l(Z) \cap V(T)|$ holds.

**Proof.** On the contrary, suppose that $|V_u(Z) \cap V(T)| > |V_l(Z) \cap V(T)|$ holds. Let $G$ be a subgraph $G := (V(Z) \cap V(T), E(Z) \cap E(T))$ of $Z$, which is a forest (i.e. a collection of trees). By a slight abuse of notation, let $V_l(G) := V_l(Z) \cap V(T)$ and $V_u(G) := V_u(Z) \cap V(T)$, respectively. Since we have $\text{outdeg}_G(v) \geq 1$ for any $v \in V_u(G)$, $\sum_{v \in V_u(G)} \text{outdeg}_G(v) \geq \sum_{v \in V_u(G)} 1 = |V_u(G)|$ holds. By the assumption, $|V_u(G)| >
|V_e(G)|. By the hand-shaking lemma, \( \sum_{v \in V_e(G)} \text{outdeg}_G(v) = \sum_{v \in V_e(G)} \text{indeg}_G(v) \), which means that there exists a vertex \( v \in V_e(G) \) such that \( \text{indeg}_G(v) \geq 2 \), which contradicts that \( G \) is a forest. This completes the proof.

Proposition 9 yields the following useful result.

**Lemma 10.** Let \( Z \) be a W-fence in a rooted almost-binary phylogenetic X-network \( N \) and let \( T \) be a maximum covering subtree of \( N \). Then, there exists at least one upper vertex \( v \in V_u(Z) \) that is not included in \( T \).

**Proof.** Since \( Z \) is a W-fence, the equation 1 in Proposition 8 gives \( |V_u(Z)| = |V_u(Z)| - 1 \). Proposition 9 gives \( |V_u(Z) \cap V(T)| \leq |V_u(Z) \cap V(T)| \). Clearly, \( |V_u(Z) \cap V(T)| \leq |V_u(Z)| \). Thus, \( |V_u(Z) \cap V(T)| \leq |V_u(Z)| - 1 \), which means that \( T \) does not include at least one upper vertex of \( Z \). This completes the proof.

**Theorem 11.** For any rooted almost-binary phylogenetic X-network \( N \), \( \eta^*(N) \geq |\mathcal{W}_N| \) holds.

**Proof.** Applying Lemma 10 to each W-fence \( Z \) in \( N \), we obtain \( \eta^*(N) \geq |\mathcal{W}_N| \).

**B. Necessary and Sufficient Condition for when \( \eta^*(N) = |\mathcal{W}_N| \) holds**

In this section, we will give a characterisation of rooted almost-binary phylogenetic X-networks \( N \) with \( \eta^*(N) = |\mathcal{W}_N| \) and thus show that the inequality in Theorem 11 gives a tight lower bound for \( \eta^*(N) \). We now give the necessary definitions.

**Definition 12.** Let \( \mathscr{Z} = \{Z_1, \ldots, Z_k\} \) be the maximal zig-zag trail decomposition of a rooted almost-binary phylogenetic X-network \( N \). For any distinct \( Z_i, Z_j \in \mathscr{Z} \), we say that \( Z_i \) is above \( Z_j \) or \( Z_j \) is below \( Z_i \) if there exists a vertex \( v \in V_i(Z_i) \cap V_u(Z_j) \). An M-W pair of \( N \) is an ordered pair \( (Z_i, Z_j) \) such that \( Z_i \in \mathcal{M}_N \), \( Z_j \in \mathcal{W}_N \), and \( Z_i \) above \( Z_j \).

With the ‘above-below’ relation between maximal zig-zag trails in \( N \) introduced in Definition 12, we can formally describe whether and how any pair of maximal zig-zag trails share a vertex. As defined in Definition 12 and as Proposition 13 states, particularly important for MAX-CST is an M-W pair.

**Proposition 13.** Let \( T \) be a maximum covering subtree of a rooted almost-binary X-network \( N \). Suppose \( Z \) is a W-fence in \( N \) that is not below any M-fence in \( N \). Then there exists a maximal zig-zag trail \( Z' \) (not an M-fence) above \( Z \) such that there exists an upper vertex \( v \in V_u(Z') \) with \( v \notin V(T) \).

**Proof.** We prove that there exists a maximal zig-zag trail \( Z' \) above \( Z \) such that \( Z' \) is not an M-fence and \( |V_u(Z') \cap V(T)| \leq |V_u(Z')| - 1 \) holds. By Lemma 10, there exists an upper vertex \( u \) of \( Z \) such that \( u \notin V(T) \). Let \( Z' \) be a maximal zig-zag trail in \( N \) with \( u \in V_u(Z') \). By Definition 7, \( Z' \) is above \( Z \). Then, by \( u \notin V(T) \), \( |V_u(Z') \cap V(T)| \leq |V_u(Z')| - 1 \) holds. As \( Z' \) is not an M-fence, by equation 1 in Proposition 8, we have \( |V_u(Z')| \leq |V_u(Z)| \). Therefore, we can conclude that \( |V_u(Z') \cap V(T)| \leq |V_u(Z)| - 1 \). As Proposition 9 implies \( |V_u(Z') \cap V(T)| \leq |V_u(Z) \cap V(T)| \), we obtain \( |V_u(Z') \cap V(T)| \leq |V_u(Z)| - 1 \). Thus, \( Z' \) has the desired property. This completes the proof.

**Definition 14.** Let \( N \) be a rooted almost-binary phylogenetic X-network that has at least one maximal W-fence, and let \( \mathcal{M}_N \) and \( \mathcal{W}_N \) be the sets of M-fences and W-fences in \( N \), respectively. An M-W matching in \( N \) is a non-empty set \( S = \{(M_1, W_1), \ldots, (M_p, W_p)\} \) of M-W pairs in \( N \) such that \( M_i \neq M_j \) and \( W_i \neq W_j \) hold for any \( i \neq j \in [1, p] \), where \( p \geq 1 \). An M-W matching is said to be \( \mathcal{W}_N \)-saturated if \( |S| = |\mathcal{W}_N| \).

See Figure 7(a) for an illustration of an M-W pair \((Z_1, Z_2)\). See also Figure 7(b) for an illustration of a \( \mathcal{W}_N \)-saturated matching \( \{(Z_1, Z_2), (Z_2, Z_4)\} \). We note that there exists no \( \mathcal{M}_N \)-saturated M-W matching for any \( N \), as in Remark 15.

![Figure 7](image1.png)

**Fig. 7.** (a): a rooted almost-binary phylogenetic X-network \( N \) where there exists no \( \mathcal{W}_N \)-saturated M-W matching. (b): a rooted almost-binary phylogenetic X-network \( N \) that has a \( \mathcal{W}_N \)-saturated M-W matching.

![Figure 8](image2.png)

**Fig. 8.** (a): An M-W pair \((Z_1, Z_2)\) in a rooted almost-binary phylogenetic X-network, where the vertices of the M-fence \( Z_1 \) and those of the W-fence \( Z_2 \) are indicated using \( m_1 \) and \( m_2 \), respectively. Resolving \((Z_1, Z_2)\) at vertex \( m_2 = w_1 \) deletes the vertex (gray) and its three incident edges, resulting in the graph depicted in (b).

**Remark 15.** The number \( |\mathcal{M}_N| \) of M-fences is greater than or equal to the number \( |\mathcal{W}_N| \) of W-fences.

**Proof.** The proof will be completed if we can show that \( |\mathcal{M}_N| - |\mathcal{W}_N| = |X| - 1 \) for any \( N \) with \( |X| \geq 2 \). For any non-leaf vertex \( v \) of \( N \), there exists a maximal zig-zag trail \( Z \) of \( N \) such that \( v \) is an upper vertex of \( Z \). Therefore, the number of vertices of \( N \) that are contained in \( V_u(Z) \) of some maximal zig-zag trail \( Z \) of \( N \) equals \( |V(N)| - |X| \). On the other hand, for any non-root vertex \( v \) of \( N \), there exists a maximal zig-zag trail \( Z \) of \( N \) such that \( v \) is a lower vertex of \( Z \), and thus the number of vertices of \( N \) that are contained in \( V_u(Z) \) of some maximal zig-zag trail \( Z \) of \( N \) equals \( |V(N)| - 1 \). Then, we
have $\sum_{Z \subseteq \mathcal{Z}} (|V(Z)| - |V_a(Z)|) = |X| - 1$. By the equation (1) in Proposition 8, $|X| - 1 = 1 \times |\mathcal{M}_N| + 0 \times (|\mathcal{Z}| - |\mathcal{M}_N| - |\mathcal{W}_N|) + (-1) \times |\mathcal{W}_N|$. Thus, $|\mathcal{M}_N| - |\mathcal{W}_N| = |X| - 1$, which completes the proof. \hfill \square

To consider MAX-CST, we define the operation of eliminating an M-W pair in $N$ as follows. The resulting network $N'$ is still a rooted almost-binary phylogenetic $X$-network, as ensured by Lemma 17.

Definition 16. For an M-W pair $(Z_i, Z_j)$ in a rooted almost-binary phylogenetic $X$-network $N$ and any vertex $v \in V_i(Z_i) \cap V_a(Z_j)$, we say resolving $(Z_i, Z_j)$ at $v$ to mean the operation of creating the subgraph $N - \{v\}$ of $N$ (as illustrated in Figure 8b).

Lemma 17. Let $N$ be a rooted almost-binary phylogenetic $X$-network that contains an M-W pair $(Z_i, Z_j)$. Let $N'$ be the graph obtained by resolving $(Z_i, Z_j)$ at any vertex $v$ in $V_i(Z_i) \cap V_a(Z_j)$. Then, $N'$ is also a rooted almost-binary phylogenetic $X$-network. Moreover, $N'$ satisfies $|\mathcal{W}_{N'}| = |\mathcal{W}_N| - 1$ and $|\mathcal{M}_{N'}| = |\mathcal{M}_N| - 1$.

Proof. Suppose $(Z_i, Z_j)$ is an M-W pair with $Z_i : m_1 < m_2 < \cdots < m_{2p-1} < m_{2p} > m_{2p+1}$ and $Z_j : w_1 < w_2 < w_3 < \cdots < w_{2q-1} < w_2q < w_{2q+1}$ ($p, q \in \mathbb{N}$). It is not hard to see that $N'$ is still an almost-binary phylogenetic $X$-network.

Then, Theorem 5 ensures that there exists a unique maximal zig-zag trail decomposition, $\mathcal{Z}$ and $\mathcal{Z}'$, of $N$ and $N'$, respectively. We now examine the following cases for $v \in V_i(Z_i) \cap V_a(Z_j)$ and will prove that both $|\mathcal{W}_{N'}| = |\mathcal{W}_N| - 1$ and $|\mathcal{M}_{N'}| = |\mathcal{M}_N| - 1$ holds in each case.

Case 1: when $v$ is a vertex with $\text{indeg}(v) = 1$ and $\text{outdeg}(v) = 2$ (see also Figure 9). In this case, the edge incoming to $v$ is either $(m_2, m_1)$ or $(m_{2p}, m_{2p+1})$; without loss of generality, we may assume $v = m_1$. We can also see that $v = w_k$ ($k \neq 1, 2q + 1$). Then, the resolution of $(Z_i, Z_j)$ at $v$ deletes the three edges of $N$, that is, $(m_2, m_1), (w_{2k+1}, w_{2k}), (w_{2k+1}, w_{2k+2})$. Let $E' := E(Z_i) \setminus \{(m_2, m_1)\}$ and $E'' := E(Z_j) \setminus \{(w_{2k+1}, w_{2k}), (w_{2k+1}, w_{2k+2})\}$. Also, let $G[E']$ and $G[E'']$ be the subgraphs of $N'$ induced by $E'$ and $E''$, respectively. Since $(m_2, m_1)$ is a terminal edge of $Z_i$, we easily see that $G[E']$ is a maximal zig-zag trail in $N'$. Moreover, $G[E']$ is an N-fence since $G[E']$ has an odd number of edges. Regarding the consecutive edges $(w_{2k+1}, w_{2k})$ and $(w_{2k+1}, w_{2k+2})$ of $Z_j$, $G[E'']$ consists of two N-fences of $N'$ since neither of the two edges is a terminal edge of $Z_j$. Thus, resolving $(Z_i, Z_j)$ at $v$ turns the M-fence $Z_i$ of $N$ into an N-fence of $N'$ and the W-fence $Z_j$ of $N$ into two N-fences of $N'$. Hence, $|\mathcal{W}_{N'}| = |\mathcal{W}_N| - 1$ and $|\mathcal{M}_{N'}| = |\mathcal{M}_N| - 1$.

Case 2: when $v$ is a vertex with $\text{indeg}(v) = 2$ and $\text{outdeg}(v) = 1$ (see also Figure 10). In this case, the edge outgoing from $v$ is either $(w_1, w_2)$ or $(w_{2q+1}, w_{2q+2})$; without loss of generality, we may assume $v = w_1$. Then, by an argument similar to Case 1, we obtain $|\mathcal{W}_{N'}| = |\mathcal{W}_N| - 1$ and $|\mathcal{M}_{N'}| = |\mathcal{M}_N| - 1$.

Case 3: when $v$ is a vertex with $\text{indeg}(v) = \text{outdeg}(v) = 1$ (see also Figure 11). In this case, the edge incoming to $v$ (resp. the edge outgoing from $v$) is a terminal edge of $Z_i$ (resp. $Z_j$). Similarly to the above, we can see that resolving $(Z_i, Z_j)$ turns the M-fence $Z_i$ (resp. $Z_j$) into an N-fence of $N'$. Thus, we obtain $|\mathcal{W}_{N'}| = |\mathcal{W}_N| - 1$ and $|\mathcal{M}_{N'}| = |\mathcal{M}_N| - 1$.

Case 4: when $v$ is a vertex with $\text{indeg}(v) = \text{outdeg}(v) = 2$ (see also Figure 12). In this case, the two edges incoming to $v$ (resp. the two edge outgoing from $v$) are non-terminal edges of $Z_i$ (resp. $Z_j$). Similarly to the above, we can see that resolving $(Z_i, Z_j)$ breaks the M-fence $Z_i$ (resp. $Z_j$) into two N-fences of $N'$. Thus, we obtain $|\mathcal{W}_{N'}| = |\mathcal{W}_N| - 1$ and $|\mathcal{M}_{N'}| = |\mathcal{M}_N| - 1$. This completes the proof. \hfill \square

![Fig. 9: Illustrations of Case 1 in the proof of Lemma 17.](image1)

![Fig. 10: Illustrations of Case 2 in the proof of Lemma 17.](image2)

![Fig. 11: Illustrations of Case 3 in the proof of Lemma 17.](image3)

We now prove a special case of the main result.
Proposition 18. In the case of \( |\mathcal{W}_N| = 1 \), \( \eta^*(N) = |\mathcal{W}_N| = 1 \) holds if and only if there exists an \( \mathcal{W}_N \)-saturated M-W matching of \( N \).

Proof. Let \( Z_j \) be the only maximal W-fence in \( N \). We first show that if there exists exactly one \( \mathcal{W}_N \)-saturated M-W matching of \( N \), then \( \eta^*(N) = |\mathcal{W}_N| = 1 \) holds. Let \( (Z_i, Z_j) \) be the (only) M-W pair in \( N \), and let \( N' \) be the subgraph of \( N \) obtained by resolving \((Z_i, Z_j)\) at vertex \( v \in V_i(Z_i) \cap V_j(Z_j) \). Lemma 17 says that \( N' \) is a rooted almost-binary phylogenetic X-network and that \( N' \) has no W-fence. Then, \( N' \) has a subdivision tree \( T' \) by Proposition 6. Since \( N' := N - \{v\} \) is a subgraph of \( N \), there exists a subtree \( T \) of \( N \) such that \( V(N') \setminus V(T) = \{v\} \) and \( T \) is isomorphic to \( T' \). As \( T \) shares the same root and leaf-set \( X \) with \( N \), \( T \) is a covering subtree of \( N \) with \( |V(N)| - |V(T)| = 1 \). As Theorem 11 gives \( 1 = |\mathcal{W}_N| \leq \eta^*(N) \), we see that \( T \) is a maximum covering subtree of \( N \). Thus, \( \eta^*(N) = 1 \) holds.

To prove the converse, assume that there exists no \( \mathcal{W}_N \)-saturated M-W matching of \( N \). This means that \( N \) contains no M-W pair \( (Z_i, Z_j) \) for the W-fence \( Z_j \). By Lemma 10, there exists a vertex \( v \in V_i(Z_i) \) that is not included in a maximum covering subtree \( T \) in \( N \). Then, by Proposition 13, there exists a maximal zig-zag trail \( Z \) above \( Z_j \) with \( v \in V_i(Z) \), where \( Z \) is not a M-fence. Proposition 13 also states that there exists \( u \in V_i(Z) \) with \( u \notin V(T) \). Thus, \( T \) contains neither \( u \) nor \( v \). Hence, we obtain \( \eta^*(N) \geq 2 > 1 \) as desired.

Finally, we give a characterisation of rooted almost-binary phylogenetic X-networks \( N \) with \( \eta^*(N) = |\mathcal{W}_N| \) as follows.

Theorem 19. For a rooted almost-binary phylogenetic X-network \( N \), \( \eta^*(N) = |\mathcal{W}_N| \) holds if and only if there is no W-fence or there exists an \( \mathcal{W}_N \)-saturated M-W matching of \( N \).

Proof. As the case of \( |\mathcal{W}_N| = 0 \) is trivial, we may assume \( |\mathcal{W}_N| \geq 1 \). The goal is to show that \( \eta^*(N) = |\mathcal{W}_N| \) holds if and only if there exists an \( \mathcal{W}_N \)-saturated M-W matching of \( N \). The proof is by induction on the number \( |\mathcal{W}_N| \) of W-fences of \( N \). In the case of \( |\mathcal{W}_N| = 0 \), the assertion holds by Proposition 6. When \( |\mathcal{W}_N| = 1 \), the assertion holds by Proposition 18.

Assuming that Theorem 19 holds for any \( N' \) with \( |\mathcal{W}_{N'}| \leq i \), we will show that the assertion holds for any \( N \) with \( |\mathcal{W}_N| = i + 1 \). Suppose there exists a \( \mathcal{W}_N \)-saturated M-W matching of \( N \). Let \( (Z_i, Z_j) \) be an element of \( \mathcal{W}_N \)-saturated M-W matching of \( N \) and let \( N' \) be an almost-binary phylogenetic X-network with \( |\mathcal{W}_{N'}| = i \) obtained by the resolution of \((Z_i, Z_j)\) (Lemma 17 is used here). Then, by the induction hypothesis, we have \( \eta^*(N') = i \). Let \( T' \) be a maximum covering subtree of \( N' \). Similarly to the argument in the proof of Proposition 18, \( N \) contains a covering subtree \( T \) that is isomorphic to \( T' \). As \( V(N) \setminus V(N') = \{v\} \), the number of vertices of \( N \) that are not covered by \( T \) equals \( i + 1 \). As Theorem 11 guarantees \( i + 1 = |\mathcal{W}_N| \leq \eta^*(N) \), we can conclude that \( T \) is a maximum covering subtree of \( N \) and \( \eta^*(N) = i + 1 \) holds.

To prove the converse, assuming that there is no \( \mathcal{W}_N \)-saturated M-W matching of \( N \), we will show that \( \eta^*(N) > |\mathcal{W}_N| = i + 1 \) holds. Let \( Z_j \) be a W-fence that is not in any M-W pair in \( N \). Then, by Lemma 10, there exists an upper vertex \( v \) of \( Z_j \) with \( v \notin V(T) \) for any maximum covering subtree \( T \) of \( N \). Let \( Z \) be the maximal zig-zag trail in \( N \) that is above \( Z_j \) and \( v \) as its lower vertex. When \( Z \) is not an M-fence, Proposition 13 says that there exists a vertex \( u \in V_u(Z) \) with \( u \notin V(T) \), which implies \( \eta^*(N) \geq i + 2 \). Next, we must consider the case when \( Z \) is an M-fence. In fact, even if there is no \( \mathcal{W}_N \)-saturated M-W matching, it is possible that there exists a M-fence above each W-fence. This is because an M-fence can be above two or more W-fences (see the left panel of Figure 7 for an example). Let \( Z \) be a maximal M-fence that is above two distinct maximal W-fences \( Z_j \) and \( Z_j' \) in \( N \) and let \( k \) be the number of upper vertices of \( Z \). According to Proposition 8, the number of lower vertices of \( Z \) is given by \( k + 1 \). Lemma 10 ensures the existence of two distinct vertices \( p \in V_u(Z_j) \) and \( q \in V_u(Z_j') \) such that neither \( p \) nor \( q \) is in \( V(T) \). Since \( p, q \in V_i(Z) \), we have \( |V_i(Z) \cap V(T)| \leq k + 1 - 2 = k - 1 \). Proposition 9 yields \( |V_u(Z) \cap V(T)| \leq |V_i(Z) \cap V(T)| \), leading to \( |V_u(Z) \cap V(T)| \leq k - 1 \). Thus, at least one upper vertex of \( Z \) is not in \( V(T) \). Consequently, we deduce that \( \eta^*(N) \geq i + 2 \). This completes the proof.

VI. CONCLUSION

In this paper, we have considered the number \( \eta^*(N) \) of uncovered vertices in a phylogenetic network \( N \) by a new approach using a canonical decomposition of phylogenetic networks called the maximal zig-zag trail decomposition. We defined an above/below relationship between maximal zig-zag trails by defining the concept of upper and lower vertices of each maximal zig-zag trail, and then introduced the new concept of M-W pairs and \( \mathcal{W}_N \)-saturated M-W matchings. We gave a tight lower bound \( |\mathcal{W}_N| \leq \eta^*(N) \) and proved that \( |\mathcal{W}_N| = \eta^*(N) \) holds if and only if there exists a \( \mathcal{W}_N \)-saturated M-W matching in \( N \).

The results in this paper have many different implications. First, as \( |\mathcal{W}_N| \) can be computed in linear time [5], Theorem 19 means that MAX-CST can be solved in linear time for (binary or) almost-binary phylogenetic networks with \( \mathcal{W}_N \)-saturated M-W matching. Second, since tree-based networks \( N \) are a trivial instance of MAX-CST (because \( |\mathcal{W}_N| = \eta^*(N) = 0 \)), phylogenetic networks with \( \mathcal{W}_N \)-saturated M-W matchings can be understood as a generalisation of tree-based networks. Recalling \( \delta^*(N) = |\mathcal{W}_N| \) [5], we have thus clarified the re-
lationship between the two indices $\delta^*(N)$ and $\eta^*(N)$ for measuring the deviation of $N$ from being tree-based, which were originally proposed by Francis–Semple–Steel [14]. For future work, it would be interesting to investigate the mathematical and computational properties of phylogenetic networks with or without $\mathcal{W}_N$-saturated M-W matching.

ACKNOWLEDGMENTS

M.H. is supported by JST FOREST Program (Grant Number JPMJFR2135, Japan). T.S. is grateful to participants at the 6th JST Math Open Problem Workshop for helpful discussions.

REFERENCES


Takatora Suzuki

Takatora Suzuki obtained a B.S. in applied mathematics from Waseda University in 2023. He is currently an M.S. student at Waseda University. His current research topics are combinatorial problems in phylogenetics.

Han Guo

Han Guo obtained a B.S. in applied mathematics from Waseda University in 2022. He was an M.S. student when this research was conducted.

Momoko Hayamizu

Momoko Hayamizu received her M.D. from the School of Medicine, The University of Tokyo. She completed her residency and fellowship as a radiologist. She received her Ph.D. in statistical science from the Graduate University for Advanced Studies for the research of discrete mathematical problems in biology and life sciences. She is currently an associate professor at the Department of Applied Mathematics, Waseda University. Her research interests include combinatorial and algorithmic problems related to phylogenetics. She was awarded the Young Scientists’ Prize of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology for 2022.