When Many Trees Go to War: On Sets of Phylogenetic Trees With Almost No Common Structure*

Mathias Weller[1,2]   Norbert Zeh[3†]

[1] *LIGM, Université Gustave Eiffel, Marne la Vallée, France*
[2] *Centre National de Recherche Scientifique, France*
[3] *Faculty of Computer Science, Dalhousie University, Halifax, Canada*

It is known that any two trees on the same n leaves can be displayed by a network with $n-2$ reticulations, and there are two trees that cannot be displayed by a network with fewer reticulations. But how many reticulations are needed to display multiple trees? For any set of t trees on n leaves, there is a trivial network with $(t-1)n$ reticulations that displays them. To do better, we have to exploit common structure of the trees to embed non-trivial subtrees of different trees into the same part of the network. In this paper, we show that, for $t \in o\left(\sqrt{\log n}\right)$, there is a set of t trees with virtually no common structure that could be exploited. More precisely, we show that, for any $t \in o\left(\sqrt{\log n}\right)$, there are t trees such that any network displaying them has $(t-1)n - o(n)$ reticulations. For $t \in o(\log n)$, we obtain a slightly weaker bound. We also prove that already for $t = c \log n$, for any constant $c > 0$, there is a set of t trees that cannot be displayed by a network with $o(n \log n)$ reticulations, matching up to constant factors the known upper bound of $O(n \log n)$ reticulations sufficient to display *all* trees with n leaves. These results are based on simple counting arguments and extend to unrooted networks and trees.

Keywords: Phylogenetic Networks, Displayed Trees, Combinatorial Bounds

1   Introduction

Phylogenetics is the study of evolutionary relationships among a set of taxa, typically represented using tree-like structures that depict patterns of descent from common ancestors. Traditionally, these relationships have been modelled with phylogenetic trees, which are rooted or unrooted trees whose leaves correspond to present-day taxa and whose internal nodes represent ancestral lineages [24]. However, due to processes such as hybridization, horizontal gene transfer, and recombination, tree models often fail to capture the full complexity of evolutionary histories [1, 9, 10, 22]. This has led to the development of phylogenetic networks, which are graphical models that generalize trees by allowing for "reticulate" (non-tree-like) events [13, 23].

Both rooted and unrooted trees and networks are considered in practice, depending on the data and the biological questions at hand. Rooted structures incorporate ancestral/temporal directionality, which is important for inferring evolutionary pathways, while unrooted structures are often used when such directionality is unknown or when analyzing symmetric distance data [14, 24].

A central concept in the analysis of phylogenetic networks is that of a network "displaying" a tree: informally, a network displays a tree if the tree can be embedded in the network in a way that respects its topology [13]. A formal definition will be given in Section 2. This notion is at the heart of computational problems in phylogenetics like comparing phylogenies or parsimoniously reconstructing networks [6]. Indeed, the computational problems TREE CONTAINMENT (decide whether a given tree is embeddable into a given network) [8, 11, 12, 18, 19, 25] and HYBRIDIZATION NUMBER (construct a smallest network displaying all given trees) [2, 3, 7, 17, 20] have been studied extensively.

There is a trivial network that displays any set of t trees without exploiting any common structure (see Figure 1). This network is obtained by simply (1) taking the disjoint union of the t trees, (2) adding a tree "cap" that joins the roots of the trees under a common root and, finally, (3) adding reticulations to "merge" all copies of each leaf in the different trees. This network has $(t-1)n$ reticulations and conveys absolutely no information about the similaries between the t trees. In general, this is far from the minimum number of reticulations necessary to display a set of trees: if multiple input trees share subtrees with the same topology, then these subtrees can be embedded into the same part of the network, which may allow us to save reticulations. Indeed, Wu and Zhang [26] proved that, for three trees, $2n - 2 - \Omega(\log \log n)$ reticulations always suffice and, in general, $(t-1)(n-2) - \Omega(\min\{t^{3/2}, n^3\})$ reticulations always suffice. Yet, the gap between these bounds and the number of reticulations in the trivial network is only $o(tn)$. An interesting question then is whether there exist sets of trees with little enough common structure to make the trivial network the best possible up to such lower order terms.

Baroni, Semple, and Steel [3] answered this question in the affirmative for $t = 2$, by showing that there are two n-caterpillars requiring $n - 2$ reticulations.

Recently, van Iersel, Jones, and Weller [16] proved that even displaying 3 caterpillars without restrictions on the type of network may require $(3/2 - o(1))n$ reticulations. This left open the possibility that the number of reticulations required to display t trees depends sublinearly on t, and they explicity pose the general case of t trees as an open question. This paper settles this question by showing that the dependence on t is linear after all, up to $t \in o(\log n)$.

Another known special case is $t = (2n - 3)!!$, that is, how many reticulations are necessary to display *all* trees on the same n leaves? Bordewich and Semple [5] showed that this number is in $\Theta(n \log n)$ even for tree-based networks, but the $\Omega(n \log n)$ lower bound is derived via a simple counting argument applied to any type of network. In this paper, we generalize this counting argument to the case of t trees. We show that for $t \in o(\log n)$, there is a set of t trees for which the trivial network (see Figure 1) is asymptotically optimal, in the sense that there are t trees that cannot be displayed by a network with fewer than $(t-1)n - o(tn)$ reticulations. For $t \in o\left(\sqrt{\log n}\right)$, we obtain the slightly stronger result that there are t trees that cannot be displayed by a network with fewer than $(t-1)n - o(n)$ reticulations. Thus, we conclude that the trivial network is indeed *asymptotically* optimal in the worst case for displaying t trees. The proof of these two results also shows that, for any constant $c > 0$, there is a set of $c \log n$ trees that cannot be displayed by a network with $o(n \log n)$ reticulations. One way to interpret this result is that "most" of the reticulations in the $\Theta(n \log n)$-reticulation network of Bordewich and Semple [5] that displays all trees on n leaves are caused by a small subset of only $O(\log n)$ trees, and adding the remaining exponentially many trees increases the number of reticulations necessary to display these trees by only a

Fig. 1: The trivial network displaying three trees T_1, T_2, and T_3, each with leaf set $[8]$

constant factor.

As the results by Baroni, Semple, and Steel [3], Bordewich and Semple [5], and van Iersel, Jones, and Weller [16] show, the question of how many reticulations are necessary to display a set of t trees has been the focus of previous work and is interesting in its own right. Our interest in this question stems from its connection to cluster reduction. Cluster reduction is a technique that has been verified experimentally to be the most important technique for computing maximum agreement forests of two trees efficiently [21]. Before this experimental study, Baroni, Semple, and Steel [4] proved that it can also be used to reduce the complexity of finding a network with the minimum number of reticulations that displays two trees, owing to the fact that, for two trees, computing such a network is equivalent to computing a particular type of agreement forest [6]. We say that cluster reduction is "safe" for two trees. But what happens for more than two trees? This question is studied in [15]. One of the main results of that paper is that cluster reduction fails to be safe for four or more trees, that is, the use of cluster reduction may result in computing a network with more than the minimum number of reticulations. The proof is via a reduction that starts with a set of trees that require many reticulations to display and constructs from it a new set of trees for which cluster reduction fails. This proof relies directly on the results shown in this paper.

2 Preliminaries

Phylogenetic networks. An *unrooted phylogenetic network* N^{u} (*unrooted network* for short) is a simple connected undirected graph whose degree-one nodes (called *leaves*) have unique labels and whose non-leaf nodes have degree at least 3 and are unlabelled. If N^{u} does not contain cycles, then we call N^{u} an unrooted phylogenetic tree (*unrooted tree* for short). If the maximum degree in N^{u} is at most three, then we call N^{u} *binary*. A *rooted phylogenetic network* N (*network* for short) is a simple connected directed acyclic graph with a single source, called the *root*. Its out-degree-0 nodes, called *leaves*, have unique labels and its non-leaf, non-root nodes are unlabelled and either have in-degree 1 and out-degree at least 2 (*tree nodes*) or out-degree 1 and in-degree at least 2 (*reticulations*). If N has no reticulations, it is called a *rooted*

phylogenetic tree (*tree* for short). If both the maximum in-degree and the maximum out-degree of N are at most two, then N is called *binary*. Throughout this paper, all considered (unrooted) networks are binary. We use $V(N)$, $E(N)$, and $\mathcal{L}(N)$ to refer to the nodes, edges, and leaves of an (unrooted) network N, respectively. If (u, v) is a directed edge of a network N, then u is called a *parent* of v in N, and v is called a *child* of u in N. Similarly, any in-edge (u, v) of a node v is called a *parent edge* of v, and any out-edge (v, w) of v is called a *child edge* of v.

Labellings.   The labels of the leaves of a tree or network are part of its definition. Whenever we need to refer to the label of a leaf v explicitly, we use the notation $\lambda_L(v)$. We assume that all networks (rooted and unrooted) have their leaves labelled with the integers 1 to n, where n is the number of leaves. We use $[n]$ to refer to the set $\{1, \ldots, n\}$.

For counting networks, it will be useful to also label the edges of networks. If N is an (undirected) network and λ_E is a labelling of its edges, we call the pair (N, λ_E) an (undirected) *edge-labelled network*. When only a subset of the edges is labelled, we extend this to a labelling of all edges by assigning the special label $\bot$ to unlabelled edges. Accordingly, we call the edges with label $\bot$ *unlabelled*. Similarly, it is convenient to extend the leaf labelling λ_L of N to a labelling of all nodes of N, by assigning the label $\bot$ to every non-leaf node.

For two mappings $\phi, \psi : A \to B$ with $\bot \in B$, we say that ψ *extends* ϕ, or ψ is an *extension* of ϕ, if $\psi(a) = \phi(a)$ whenever $\phi(a) \neq \bot$. In other words, ψ may label more elements than ϕ does but ϕ and ψ agree on the labels of all elements that ϕ does label. Viewing ϕ and ψ as relations, this implies that $\phi \setminus (A \times \{\bot\}) \subseteq \psi \setminus (A \times \{\bot\})$. Therefore, we write $\phi \subseteq \psi$ whenever ψ is an extension of ϕ.

Isomorphisms.   An *isomorphism* $\phi : G_1 \to G_2$ between directed graphs G_1 and G_2 is a pair of (set) isomorphisms $\phi_V : V(G_1) \to V(G_2)$ and $\phi_E : E(G_1) \to E(G_2)$ such that, for every edge $(u, v) \in E(G_1)$, $\phi_E((u, v)) = (\phi_V(u), \phi_V(v))$. The definition for undirected graphs is analogous. An isomorphism between (undirected) edge-labelled networks (N, λ_E) and (N', λ'_E) is an (undirected) graph isomorphism $\phi : N \to N'$ with the added conditions that $\lambda_L(v) = \lambda'_L(\phi_V(v))$ for all $v \in V(N)$, and $\lambda_E(e) = \lambda'_E(\phi_E(e))$ for all $e \in E(N)$. Throughout this paper, we consider isomorphic networks to be the same network.

Displaying a tree.   A *subdivision* of a (rooted or unrooted) tree T is a graph T' obtained by replacing each edge $e \in E(T)$ with a path P_e consisting of one or more edges. The paths in $\{P_e \mid e \in E(T)\}$ are internally node-disjoint, and their internal nodes are unlabelled. If T is rooted, then $P_{(u,v)}$ is a directed path from u to v, for all $(u, v) \in E(T)$. Note that such a subdivision T' is a (rooted or unrooted) tree in the graph-theoretic sense but not according to our definition of a phylogenetic tree, as it may have nodes with in-degree 1 and out-degree 1.

A (rooted or unrooted) network N *displays* a (rooted or unrooted) tree T if N has a subgraph T' that is (isomorphic to) a subdivision of T. We call T' an *embedding* of T into N. Note that, in general, T' is not unique, that is, a tree may have multiple embeddings into a network that displays it.

Useful (in)equalities.   We will refer to the base-2 logarithm as $\log$ or $\log_2$ in this paper. Further, we rely on the following equalities and inequalities.

Observation 1. $(2k - 1)!! = \dfrac{(2k)!}{2^k k!}$.

The relations $\binom{n}{k} = \frac{n!}{(n-k)! k!}$ and $\binom{n}{k} \leq 2^n$ immediately imply that

Observation 2. $(a + b)! \leq 2^{a+b}a!b!.$

Lemma 3. $\dfrac{1}{e} < \left(\dfrac{n}{n+1}\right)^n \leq \dfrac{1}{2}$, *for all* $n \in \mathbb{N}^+$.

Proof: For $n = 1$, we have $(n/n+1)^n = 1/2$. For $n \to \infty$, we have

$$\lim_{n\to\infty} (n/n+1)^n = \lim_{n\to\infty} (n/n+1)^{n+1} \cdot (n+1/n) = \lim_{n\to\infty} (1 - 1/n+1)^{n+1} \cdot \lim_{n\to\infty} (1 + 1/n) = 1/e \cdot 1.$$

Since the derivative of $(n/n+1)^n$ is non-positive, for all $n \geq 0$, the lemma follows. $\qquad\square$

Lemma 4. $(n/e)^n < n! < (n/2)^n$, *for all* $n \geq 6$.

Proof: The first inequality holds for all $n \in \mathbb{N}^+$ and follows from Stirling's bound. The proof of the second inequality is by induction on n. It can be verified that the inequality holds for $n = 6$. For all $n \geq 6$,

$$(n+1)! = (n+1) \cdot n! \overset{\text{IH}}{<} (n+1) \cdot \frac{n^n}{2^n} = (n+1) \cdot \frac{(n+1)^n}{2^n} \cdot \left(\frac{n}{n+1}\right)^n$$

$$= \frac{(n+1)^{n+1}}{2^{n+1}} \cdot 2 \left(\frac{n}{n+1}\right)^n \overset{\text{Lem. 3}}{\leq} \frac{(n+1)^{n+1}}{2^{n+1}}.$$

$\qquad\square$

Lemma 5. $(2n - 3)!! > (n/2)^n$, *for all* $n \geq 5$.

Proof: The proof is by induction on n. It can be verified that the claim holds for $n = 5$. For all $n \geq 5$,

$$(2(n+1) - 3)!! = (2n-1)(2n-3)!! \overset{\text{IH}}{>} (2n-1)\left(\frac{n}{2}\right)^n = \frac{2n-1}{n+1}\left(\frac{n+1}{2}\right)^{n+1} \cdot 2\left(\frac{n}{n+1}\right)^n.$$

Since $2n-1/n+1 \geq 3/2$ for all $n \geq 5$ and, by Theorem 3, $(n/n+1)^n > 1/e$ for all $n \in \mathbb{N}^+$, the claim follows. $\qquad\square$

3 Counting Rooted Binary Trees and Networks

Our proof that some sets of trees cannot be displayed by networks with few reticulations uses a simple counting argument. First, we count the number of sets of t trees on n leaves. Then, given some integer r, we count how many such sets can be displayed by any one network with at most r reticulations. If r is the smallest integer such that each set of t trees on n leaves can be displayed by some network with r reticulations, then there is a set of t trees on n leaves that cannot be displayed by a network with fewer than r reticulations. Our goal in this paper is to obtain lower bounds on this number r. We present the proof for rooted trees and networks in this section and consider the unrooted case in the next section.

Proposition 6. *Let* $n, r \in \mathbb{N}^+$, *and let* $\mathcal{N}_{n,r}$ *be the set of all binary networks with* n *leaves and* r *reticulations. Then* $|\mathcal{N}_{n,r}| \leq \frac{(2n+4r-3)!!}{r!} \leq n! \cdot (r-1)! \cdot 2^{2n+6r-3}$.
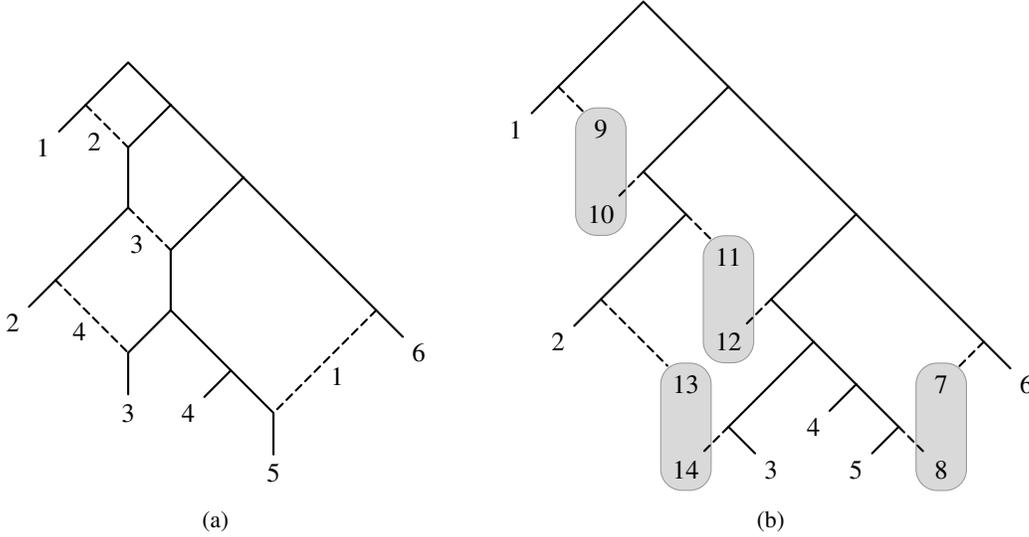
Fig. 2: (a) A reticulation-labelled network (N, λ_E) with 6 leaves and 4 reticulations. Edges labelled 0 are drawn solid. Edges labelled 1 through 4 are drawn dashed. (b) The tree $\tau((N, \lambda_E))$. Pairs of leaves representing the same reticulation of (N, λ_E) are shaded.

Proof: A *switching* $\sigma : E(N) \to \{0, \perp\}$ of N is a labelling of the edges of N such that every non-root node of N has exactly one labelled parent edge. Thus, N has exactly r unlabelled edges, one parent edge per reticulation. A *reticulation labelling* λ_E of N is an extension of a switching σ that bijectively labels all edges left unlabelled by σ with labels in $[r]$. Note that every switching σ of N gives rise to $r!$ reticulation labellings corresponding to all possible permutations of the r edges of N left unlabelled by σ. A *reticulation-labelled network* is an edge-labelled network (N, λ_E), where $N \in \mathcal{N}_{n,r}$ and λ_E is a reticulation labelling of N. Now, for every $N \in \mathcal{N}_{n,r}$, choose an arbitrary but fixed switching σ_N of N, and let

$$\Lambda_{n,r} := \{(N, \lambda_E) \mid N \in \mathcal{N}_{n,r} \land \lambda_E \text{ is a reticulation labelling of } N \land \sigma_N \subseteq \lambda_E\}$$

where, by convention, two edge-labelled networks are considered equal if they are isomorphic. Let $\mathcal{T}_m$ be the set of all binary phylogenetic trees with m leaves. We prove that

$$|\mathcal{N}_{n,r}| \cdot r! \overset{\text{(a)}}{=} |\Lambda_{n,r}| \overset{\text{(b)}}{\leq} |\mathcal{T}_{n+2r}|,$$

implying that $|\mathcal{N}_{n,r}| \leq \frac{(2n+4r-3)!!}{r!}$, as $|\mathcal{T}_{n+2r}| = (2(n+2r)-3)!!$. The bound $\frac{(2n+4r-3)!!}{r!} \leq n! \cdot (r-1)! \cdot 2^{2n+6r-3}$ is obtained using basic algebra then.

(b): To prove (b), we construct an injective mapping $\tau : \Lambda_{n,r} \to \mathcal{T}_{n+2r}$. For each $(N, \lambda_E) \in \Lambda_{n,r}$, we construct $\tau((N, \lambda_E))$ as follows: We replace every edge $e = (u, v)$ such that $\lambda_E(e) \neq 0$ with two edges (u, z) and (v, z'), where z and z' are new leaves. These leaves are assigned labels $\lambda_L(z) := n + 2\lambda_E(e) - 1$ and $\lambda_L(z') := n + 2\lambda_E(e)$. Finally, we drop all edge labels. This construction is illustrated in Figure 2. Since $\tau((N, \lambda_E))$ is a binary phylogenetic tree with $n + 2r$ leaves labelled 1 through $n + 2r$, we have

$\tau((N, \lambda_E)) \in \mathcal{T}_{n+2r}$. This mapping τ is injective because the network (N, λ_E) is easily recovered from $\tau((N, \lambda_E))$ by reversing the above construction: We label all edges of $\tau((N, \lambda_E))$ with label 0. Then, for all $h \in [r]$, let u and v be the parents of the leaves z and z' with labels $\lambda_L(z) = n + 2h - 1$ and $\lambda_L(z') = n + 2h$, respectively. We remove z and z' and their parent edges, and add a new edge (u, v) with label $\lambda_E((u, v)) = h$. It is easily verified that this construction indeed recovers the edge-labelled network (N, λ_E) from $\tau((N, \lambda_E))$.

(a): Since we consider isomorphic networks to be the same network in this paper, the proof that $|\mathcal{N}_{n,r}| = \frac{|\Lambda_{n,r}|}{r!}$ needs to distinguish between the two interpretations of a pair (N, λ_E) as simply a pair or as an edge-labelled network. If $\lambda_E \neq \lambda'_E$, then the two pairs (N, λ_E) and (N, λ'_E) are not the same but the two edge-labelled networks (N, λ_E) and (N, λ'_E) may be isomorphic and, therefore, the same to us. To make this distinction explicit, we refer to (N, λ_E) as $\langle N, \lambda_E \rangle$ when we view it as simply a pair, and as (N, λ_E) when we view it as an edge-labelled network. Let

$$\Gamma_{n,r} := \{\langle N, \lambda_E \rangle \mid N \in \mathcal{N}_{n,r} \wedge \lambda_E \text{ is a reticulation labelling of } N \wedge \sigma_N \subseteq \lambda_E\},$$

where two pairs $\langle N, \lambda_E \rangle$ are equal if their components are equal. Then $|\mathcal{N}_{n,r}| \cdot r! = |\Gamma_{n,r}|$ because, as already observed, there are $r!$ reticulation labellings extending σ_N, for each $N \in \mathcal{N}_{n,r}$. Thus, to prove that $|\mathcal{N}_{n,r}| \cdot r! = |\Lambda_{n,r}|$, we need to prove that $|\Lambda_{n,r}| = |\Gamma_{n,r}|$. Since it is clear that $|\Lambda_{n,r}| \leq |\Gamma_{n,r}|$, it suffices to show the following:

Claim 1. If (N, λ_E) and $(\tilde{N}, \tilde{\lambda}_E)$ are isomorphic, then $\langle N, \lambda_E \rangle = \langle \tilde{N}, \tilde{\lambda}_E \rangle$.

Proof. Consider two pairs $\langle N, \lambda_E \rangle, \langle \tilde{N}, \tilde{\lambda}_E \rangle \in \Gamma_{n,r}$ such that there is an isomorphism $\phi : (N, \lambda_E) \to (\tilde{N}, \tilde{\lambda}_E)$. Then ϕ is also an isomorphism $\phi : N \to \tilde{N}$, so $N = \tilde{N}$. We show that ϕ must be the identity 1_N, which implies that ϕ maps every edge of N onto itself and, therefore, that $\lambda_E = \tilde{\lambda}_E$. This in turn implies that $\langle N, \lambda_E \rangle = \langle \tilde{N}, \tilde{\lambda}_E \rangle$.

Assume for the sake of contradiction that $\phi \neq 1_N$. Since there are no parallel edges in N, there is a lowest node $v \in V(N)$ with $\phi(v) \neq v$, that is, $\phi(w) = w$ holds for all children w of v in N. Since $\phi : N \to N$ is an isomorphism, it respects the labels of all leaves in N. Therefore, $\phi(x) = x$ for all $x \in \mathcal{L}(N)$, so v is not a leaf of N. Let w be a child of v. If w has a single parent edge e, then $\phi(e) = e$, since $\phi(w) = w$ and ϕ is an isomorphism. If w has two parent edges e_1 and e_2, then $\lambda_E(e_1) \neq \lambda_E(e_2)$, since λ_E is a reticulation labelling (one of $\lambda_E(e_1)$ and $\lambda_E(e_2)$ is zero; the other, non-zero). Since $\phi(w) = w$ and ϕ respects edge labels, $\phi(e_1) = e_1$ and $\phi(e_2) = e_2$. This shows that no matter whether w has one or two parents, we have $\phi(e) = e$, for every parent edge e of w. If we choose e to be the parent edge whose top endpoint is v, then this implies that $\phi(v) = v$, contradicting our assumption. Therefore, $\phi = 1_N$ and $\langle N, \lambda_E \rangle = \langle \tilde{N}, \tilde{\lambda}_E \rangle$. ⌟

Having shown (a) and (b), the proposition follows from

$$
\begin{aligned}
|\mathcal{N}_{n,r}| &\overset{\text{(a),(b)}}{\leq} \frac{|\mathcal{T}_{n+2r}|}{r!} = \frac{(2n + 4r - 3)!!}{r!} \\
&\overset{\text{Obs. 1}}{=} \frac{(2n + 4r - 2)!}{2^{n+2r-1}(n + 2r - 1)!r!} \overset{\text{Obs. 2}}{\leq} \frac{2^{2n+4r-2}((n + 2r - 1)!)^2}{2^{n+2r-1}(n + 2r - 1)!r!} = \frac{2^{n+2r-1}(n + 2r - 1)!}{r!} \\
&\overset{\text{Obs. 2}}{\leq} \frac{2^{2n+4r-2}n!(2r - 1)!}{r!} \overset{\text{Obs. 2}}{\leq} \frac{2^{2n+6r-3}n!r!(r - 1)!}{r!} = 2^{2n+6r-3}n!(r - 1)!.
\end{aligned}
$$

$\square$

Since each n-leaf network with r reticulations displays at most 2^r trees, each such network displays at most $\binom{2^r}{t} \leq 2^{rt}$ sets of t trees. Thus, we immediately obtain the following corollary.

Corollary 7. *Let $n, t, r \in \mathbb{N}^+$, and let $\mathcal{M}_{n,t,r}$ be the set of all pairs $(N, \mathcal{T})$ such that $\mathcal{T}$ is a set of t binary trees on n leaves, and N is a binary network with n leaves and r reticulations displaying all trees in $\mathcal{T}$. Then $|\mathcal{M}_{n,t,r}| \leq 2^{rt} \cdot n! \cdot (r-1)! \cdot 2^{2n+6r-3}$.*

Observation 8. *Let $n, t \in \mathbb{N}^+$ with $n \geq 5$, and let $\mathcal{S}_{n,t}$ be the set of all sets of t binary trees on the leaf set $[n]$. Then*

$$|\mathcal{S}_{n,t}| = \binom{(2n-3)!!}{t} \geq \left(\frac{(2n-3)!!}{t}\right)^t \overset{Lem. 5}{>} \left(\frac{n}{2}\right)^{tn} \cdot t^{-t}.$$

Theorem 9. *Let $n, t \in \mathbb{N}^+$ such that $n \geq 6$, and let r be the smallest integer such that every set of t binary trees on the leaf set $[n]$ is displayed by some binary network with r reticulations. Then*

$$r > \frac{(t-1)n\log_2 n - 6tn - t\log_2 t}{\log_2 n + t + \log_2 t} \in (t-1)n - O\left(\frac{nt^2}{\log n + t}\right).$$

Proof: Since every set of t trees on the leaf set $[n]$ is displayed by a network with $(t-1)n$ reticulations (see Section 1), we have $r \leq (t-1)n$. By the choice of r, each tree set $\mathcal{T} \in \mathcal{S}_{n,t}$ has a network N with r reticulations such that $(N, \mathcal{T}) \in \mathcal{M}_{n,t,r}$. Therefore, $|\mathcal{S}_{n,t}| \leq |\mathcal{M}_{n,t,r}|$, so

$$\left(\frac{n}{2}\right)^{tn} \cdot t^{-t} \overset{Obs. 8}{<} |\mathcal{S}_{n,t}| \leq |\mathcal{M}_{n,t,r}| \overset{Cor. 7}{\leq} 2^{rt} \cdot n! \cdot (r-1)! \cdot 2^{2n+6r-3}.$$

By taking the base-2 logarithms on both sides, we obtain that

$$tn\log n - tn - t\log t < rt + \log n! + \log(r-1)! + 2n + 6r - 3$$
$$\overset{Lem. 4}{<} rt + n\log n + r\log r + n + 5r - 3$$
$$= r(t + \log r + 5) + n(\log n + 1) - 3$$
$$\overset{n+r\leq tn}{<} r(t + \log tn) + n(\log n + 5t).$$

Thus,

$$r > \frac{tn\log n - tn - t\log t - n\log n - 5tn}{\log n + t + \log t} = \frac{(t-1)n\log n - 6tn - t\log t}{\log n + t + \log t}$$
$$= (t-1)n - \frac{(t-1)n(t+\log t) + 6tn + t\log t}{\log n + t + \log t} \in (t-1)n - O\left(\frac{nt^2}{\log n + t}\right).$$

$\square$

Corollary 10. *For all $t \in o\left(\sqrt{\log n}\right)$, there is a set $\mathcal{T}$ of t binary trees on the leaf set $[n]$ such that every binary network N that displays the trees in $\mathcal{T}$ has $(t-1)n - o(n)$ reticulations.*

Corollary 11. *For all $t \in o(\log n)$, there is a set $\mathcal{T}$ of t binary trees on the leaf set $[n]$ such that every binary network N that displays the trees in $\mathcal{T}$ has $(t-1)n - o(tn)$ reticulations.*

Corollary 12. *Let $t = c \log_2 n$, for some constant $c > 0$. Then there is a set $\mathcal{T}$ of t binary trees on the leaf set $[n]$ such that every binary network N that displays the trees in $\mathcal{T}$ has*

$$r > \frac{(c \log_2 n - 1)n \log_2 n - O(n \log n)}{(c+1) \log_2 n + o(\log n)} \in \frac{c}{c + 1 + o(1)} n \log_2 n - O(n)$$

reticulations. In particular, as c increases, r approaches $n \log_2 n$.

4 Counting Unrooted Binary Trees and Networks

The argument in the previous section applies to rooted trees but generalizes to unrooted trees fairly naturally. In the unrooted case, there is no concept of reticulations, but we can define the reticulation number of a network $N^{\mathrm{u}} = (V, E)$ as $r(N^{\mathrm{u}}) := |E| - |V| + 1$, that is, as the number of edges that need to be deleted from N^{u} to obtain an unrooted tree. This mirrors the rooted case in the sense that any rooting of an unrooted network N^{u} has $r(N^{\mathrm{u}})$ reticulations.

The proof for rooted networks/trees had three main parts: (a) bounding the number of networks N with reticulation number r, (b) bounding the number of pairs $(N, \mathcal{T})$ such that $\mathcal{T}$ is a set of t trees on n leaves and N is a network with r reticulations that displays the trees in $\mathcal{T}$, and (c) bounding the number of sets of t trees on the leaf set $[n]$. Since there are $(2n-5)!!$ unrooted trees with leaf set $[n]$, the number of sets of t unrooted trees on the leaf set $[n]$ is easy to bound:

Observation 13. *Let $n, t \in \mathbb{N}^+$ with $n \geq 6$, and let $\mathcal{S}^u_{n,t}$ be the set of all sets of t unrooted binary trees on the leaf set $[n]$. Then*

$$|\mathcal{S}^u_{n,t}| = \binom{(2n-5)!!}{t} \geq \left(\frac{(2n-5)!!}{t} \right)^t \overset{\text{Lem. 5}}{>} \left(\frac{n-1}{2} \right)^{t(n-1)} \cdot t^{-t}.$$

Next, we bound the number of unrooted networks with n leaves and reticulation number r. Such a bound is harder to obtain than for rooted networks, as the definition of a (binary) unrooted network given in the preliminaries includes all 3-regular graphs, which clearly do not display any trees. More generally, a network may have pendant 2-edge-connected components without adjacent leaves. These components have no bearing on the set of trees displayed by the network but make it impossible to transfer one of the key ideas in the proof of Theorem 6 to unrooted networks: that the only isomorphism between two reticulation-labelled networks is the identity. For the purpose of this paper, we can restrict our attention to a subclass of unrooted networks, which we call *leaf-connecting*. A leaf-connecting unrooted network has the property that every edge lies on a path between two leaves. It is easy to see that, if a set of unrooted trees $\mathcal{T}^{\mathrm{u}}$ is displayed by an unrooted network N^{u}, then the subnetwork of N^{u} consisting of all edges contained in the embedding of at least one tree in $\mathcal{T}^{\mathrm{u}}$ displays the trees in $\mathcal{T}^{\mathrm{u}}$, is leaf-connecting, and has at most the same reticulation number as N^{u}. Thus, a set of unrooted trees is displayed by an unrooted network with reticulation number r if and only if it is displayed by a leaf-connecting unrooted network with reticulation number at most r.

Proposition 14. *Let $n, r \in \mathbb{N}^+$, and let $\mathcal{N}^u_{n,r}$ be the set of all leaf-connecting unrooted networks with n leaves and reticulation number r. Then $|\mathcal{N}^u_{n,r}| \leq \frac{(2n+4r-5)!!}{r!} \leq n! \cdot (r-2)! \cdot 2^{2n+6r-6}$.*

Proof: The proof follows the same approach as in the rooted case, with adjustments to deal with the absence of edge directions.

For an unrooted network $N^{\mathrm{u}} \in \mathcal{N}_{n,r}^{\mathrm{u}}$, a *switching* of N^{u} is a labelling $\sigma : E(N^{\mathrm{u}}) \to \{0, \bot\}$ such that the labelled edges form a spanning tree of N^{u}. This means that there are r unlabelled edges. As in the rooted case, a *reticulation labelling* λ_E of N^{u} is an extension of a switching σ of N^{u} that bijectively labels all edges left unlabelled by σ with labels in $[r]$, and a reticulation-labelled network is a pair $(N^{\mathrm{u}}, \lambda_E)$, where λ_E is a reticulation labelling. Also as in the rooted case, we fix a particular switching $\sigma_{N^{\mathrm{u}}}$ of N^{u}, for each $N^{\mathrm{u}} \in \mathcal{N}_{n,r}^{\mathrm{u}}$, we define $\mathcal{T}_m^{\mathrm{u}}$ to be the set of unrooted trees with m leaves, and we define the sets

$$\Lambda_{n,r}^{\mathrm{u}} := \{(N^{\mathrm{u}}, \lambda_E) \mid N^{\mathrm{u}} \in \mathcal{N}_{n,r}^{\mathrm{u}} \wedge \lambda_E \text{ is a reticulation labelling of } N^{\mathrm{u}} \wedge \sigma_{N^{\mathrm{u}}} \subseteq \lambda_E\} \text{ and}$$
$$\Gamma_{n,r}^{\mathrm{u}} := \{\langle N^{\mathrm{u}}, \lambda_E \rangle \mid N^{\mathrm{u}} \in \mathcal{N}_{n,r}^{\mathrm{u}} \wedge \lambda_E \text{ is a reticulation labelling of } N^{\mathrm{u}} \wedge \sigma_{N^{\mathrm{u}}} \subseteq \lambda_E\}.$$

Analogously to the proof of Theorem 6, we show that

$$|\mathcal{N}_{n,r}^{\mathrm{u}}| \cdot r! = |\Gamma_{n,r}^{\mathrm{u}}| \overset{(a)}{=} |\Lambda_{n,r}^{\mathrm{u}}| \overset{(b)}{\leq} |\mathcal{T}_{n+2r}^{\mathrm{u}}|.$$

Again, it is obvious that $|\mathcal{N}_{n,r}^{\mathrm{u}}| \cdot r! = |\Gamma_{n,r}^{\mathrm{u}}|$ because every switching $\sigma_{N^{\mathrm{u}}}$ has $r!$ extensions that are reticulation labellings.

(b): The proof that $|\Lambda_{n,r}^{\mathrm{u}}| \leq |\mathcal{T}_{n+2r}^{\mathrm{u}}|$ uses essentially the same mapping $\tau : \Lambda_{n,r}^{\mathrm{u}} \to \mathcal{T}_{n+2r}^{\mathrm{u}}$ as in the rooted case, replacing every edge $e = \{u, v\} \in E(N^{\mathrm{u}})$ with label $\lambda_E(e) > 0$ with two edges $\{u, z\}$ and $\{v, z'\}$, where z and z' are two new leaves labelled $n + 2\lambda_E(e) - 1$ and $n + 2\lambda_E(e)$, respectively. Since it is once again easy to verify that each reticulation-labelled network $(N^{\mathrm{u}}, \lambda_E)$ can be recovered from its image $\tau((N^{\mathrm{u}}, \lambda_E))$, this mapping τ is injective, and we have $|\Lambda_{n,r}^{\mathrm{u}}| \leq |\mathcal{T}_{n+2r}^{\mathrm{u}}|$.

(a): To show that $|\Lambda_{n,r}^{\mathrm{u}}| = |\Gamma_{n,r}^{\mathrm{u}}|$, we prove that the existence of an isomorphism $\phi : (N, \lambda_E) \to (\tilde{N}, \tilde{\lambda}_E)$ between two reticulation-labelled unrooted networks (N, λ_E) and $(\tilde{N}, \tilde{\lambda}_E)$ in $\Lambda_{n,r}^{\mathrm{u}}$ implies that $\langle N, \lambda_E \rangle = \langle \tilde{N}, \tilde{\lambda}_E \rangle$. As in the rooted case, ϕ is also an isomorphism from N to $\tilde{N}$, so $N = \tilde{N}$, but proving that this implies that $\phi = 1_N$ requires more care than in the rooted case.

Since ϕ is an isomorphism from N to itself, it respects the labels of all leaves, so $\phi(x) = x$ for all $x \in \mathcal{L}(N)$. Now, assume for the sake of contradiction that $\phi \neq 1_N$. As there are no parallel edges in N and N is leaf-connecting, there is a path P between two leaves x and y in N that contains an edge $\{u, v\}$ with $\phi(u) = u$ and $v' = \phi(v) \neq v$. Note that this implies that $\phi(\{u, v\}) = \{u, v'\}$ and that neither v nor v' is a leaf of N.

Without loss of generality, suppose that v is on the subpath Q of P from u to y. Then $Q' := \phi(Q)$ is also a path from $u = \phi(u)$ to $y = \phi(y)$. Note that, for every edge $e \in Q$ with $\lambda_E(e) \neq 0$, we have $\phi(e) = e$ because e is the only edge with label $\lambda_E(e)$ and ϕ respects edge labels. Thus, all edges in the symmetric difference $Q \oplus Q'$ of the edge sets of Q and Q' are labelled 0 by λ_E. Moreover, $Q \oplus Q'$ is non-empty because it contains the edges $\{u, v\}$ and $\{u, v'\}$, and every node is easily seen to have even degree in $Q \oplus Q'$. As every node in N has degree at most three, this implies that $Q \oplus Q'$ is a non-empty collection of simple cycles all of whose edges are labelled 0 by λ_E, contradicting the fact that the edges labelled 0 by λ_E form a spanning tree of N. Having obtained the desired contradiction, we conclude that $\phi = 1_N$ and, therefore, that $\langle N, \lambda_E \rangle = \langle \tilde{N}, \tilde{\lambda}_E \rangle$.

Having shown (a) and (b), we conclude that

$$
\begin{aligned}
|\mathcal{N}^{\mathrm{u}}_{n,r}| &\overset{\text{(a),(b)}}{\le} \frac{|\mathcal{T}^{\mathrm{u}}_{n+2r}|}{r!} = \frac{(2n+4r-5)!!}{r!} \\
&\overset{\text{Obs. 1}}{=} \frac{(2n+4r-4)!}{2^{n+2r-2}(n+2r-2)!r!} \overset{\text{Obs. 2}}{\le} \frac{2^{2n+4r-4}((n+2r-2)!)^2}{2^{n+2r-2}(n+2r-2)!r!} = \frac{2^{n+2r-2}(n+2r-2)!}{r!} \\
&\overset{\text{Obs. 2}}{\le} \frac{2^{2n+4r-4}n!(2r-2)!}{r!} \overset{\text{Obs. 2}}{\le} \frac{2^{2n+6r-6}n!r!(r-2)!}{r!} = 2^{2n+6r-6}n!(r-2)!.
\end{aligned}
$$

$\square$

In addition to the bound on the number of networks with reticulation number r established by Theorem 14, we need the following bound on the number of trees displayed by each of them:

Proposition 15. *Let $N^{\mathrm{u}} \in \mathcal{N}^{\mathrm{u}}_{n,r}$ be a binary unrooted network with n leaves and reticulation number r. Then N^{u} displays at most $\binom{n+3r-3}{r} \le 2^{n+3r-3}$ binary unrooted trees with n leaves.*

Proof: For every tree T^{u} displayed by N^{u}, there is a spanning tree of N^{u} that has a subdivision of T^{u} as a subtree. Conversely, for every spanning tree S^{u} of N^{u}, there is exactly one tree displayed by N^{u} such that S^{u} has a subdivision of this tree as a subtree. Thus, the number of unrooted trees displayed by N^{u} is at most the number of spanning trees of N^{u}. We show that this number is at most 2^{n+3r-3}.

As N^{u} is an unrooted binary network with n leaves and reticulation number r, we know that N^{u} has exactly $2n+3r-3$ edges, and every spanning tree of N^{u} is obtained by deleting exactly r of them. Further, exactly n edges of N^{u} are incident to leaves and are, therefore, in every spanning tree of N^{u}. Thus, N^{u} has at most $\binom{n+3r-3}{r} \le 2^{n+3r-3}$ spanning trees. $\square$

Theorem 15 immediately implies that every unrooted network N^{u} displays at most $\binom{2^{n+3r-3}}{t} \le 2^{t(n+3r-3)}$ sets of t unrooted trees with leaf set $[n]$. Together with Theorem 14, this proves the following corollary:

Corollary 16. *Let $n,t,r \in \mathbb{N}^+$ and let $\mathcal{M}^{\mathrm{u}}_{n,t,r}$ be the set of all pairs $(N^{\mathrm{u}}, \mathcal{T}^{\mathrm{u}})$ such that $\mathcal{T}^{\mathrm{u}}$ is a set of t binary unrooted trees on n leaves and N^{u} is a leaf-connecting binary unrooted network with reticulation number r that displays all unrooted trees in $\mathcal{T}^{\mathrm{u}}$. Then $|\mathcal{M}^{\mathrm{u}}_{n,t,r}| \le 2^{(t+2)(n+3r-3)} \cdot n! \cdot (r-2)!.$*

We can show the following analogue of Theorem 9 for unrooted trees now:

Theorem 17. *Let $n,t \in \mathbb{N}^+$ such that $n \ge 6$, and let r be the smallest integer such that every set of t binary unrooted trees on the leaf set $[n]$ is displayed by some binary unrooted network with r reticulations. Then*

$$
r > \frac{(t-1)n\log_2 n - t(8n + \log_2 n + \log_2 t - 1)}{\log_2 n + 3t + \log_2 t} \in (t-1)n - O\left(\frac{nt^2}{\log n + t}\right).
$$

Proof: Similar to the rooted case, we have

$$
\left(\frac{n-1}{2}\right)^{t(n-1)} \cdot t^{-t} \overset{\text{Obs. 13}}{<} \left(\binom{(2n-5)!!}{t}\right) = |\mathcal{S}^{\mathrm{u}}_{n,t}| \le |\mathcal{M}^{\mathrm{u}}_{n,t,r}| \overset{\text{Cor. 16}}{\le} 2^{(t+2)(n+3r-3)} \cdot n! \cdot (r-2)!,
$$

and we can assume that $n + r \leq tn$. Taking base-2 logarithms again gives

$$t(n-1)\log(n-1) - t(n-1) - t\log t < (t+2)(n+3r-3) + \log n! + \log(r-2)!$$

$$\overset{\text{Lem. 4}}{<} (t+1)n + (t+2)(3r-3) + n\log n + r\log r$$

$$= r(3t + \log r + 6) + n(t + \log n + 1) - 3t - 6$$

$$\overset{n+r\leq tn}{<} r(3t + \log tn) + n(\log n + 7t)$$

and, thus,

$$r > \frac{t(n-1)\log n - t(n-1) - t\log t - n\log n - 7tn}{\log n + 3t + \log t} = \frac{(t-1)n\log n - t(8n + \log n + \log t - 1)}{\log n + 3t + \log t}$$

$$= (t-1)n - \frac{(t-1)n(3t + \log t) + t(8n + \log n + \log t - 1)}{\log n + 3t + \log t} \in (t-1)n - O\left(\frac{nt^2}{\log n + t}\right).$$

□

A case analysis similar to the rooted case now proves the following corollaries.

Corollary 18. *For all $t \in o\left(\sqrt{\log n}\right)$, there is a set $\mathcal{T}^u$ of t binary unrooted trees on the leaf set $[n]$ such that every binary unrooted network N^u that displays the trees in $\mathcal{T}^u$ has reticulation number at least $(t-1)n - o(n)$.*

Corollary 19. *For all $t \in o(\log n)$, there is a set $\mathcal{T}^u$ of t binary unrooted trees on the leaf set $[n]$ such that every binary unrooted network N^u that displays the trees in $\mathcal{T}^u$ has reticulation number at least $(t-1)n - o(tn)$.*

Corollary 20. *Let $t = c\log_2 n$, for some constant $c > 0$. Then there is a set $\mathcal{T}^u$ of t binary unrooted trees on the leaf set $[n]$ such that every binary unrooted network N^u that displays the trees in $\mathcal{T}^u$ has reticulation number*

$$r > \frac{(c\log_2 n - 1)n\log_2 n - O(n\log n)}{(3c+1)\log_2 n + o(\log n)} \in \frac{c}{3c+1+o(1)}n\log_2 n - O(n).$$

In particular, as c increases, r approaches $\frac{1}{3}n\log_2 n$.

5 Conclusion

We have shown that, for any sub-logarithmic (in the number of leaves) number t, there is some set of t binary trees for which the trivial binary network with $(t-1)n$ reticulations is the best possible network that displays these trees, up to lower-order terms. This implies that these trees have almost no common structure. In particular, this result holds for any constant number of trees, which, as discussed in the introduction, has the important implication that cluster reduction is not safe for computing the hybridization number of four or more trees. This can be interpreted as a strong argument against phylogenetic reconstruction by parsimony as follows. All input trees admitting the same cluster can reasonably be considered a strong biological signal. Our results imply that no network displaying all trees can represent this signal if it is forced to have a minimum reticulation number. Indeed, it is more reasonable to expect the correct evolutionary history to reside among slightly non-optimal solutions.

Our results also imply that, for any $c > 0$, there are sets of $c \log_2 n$ binary trees that cannot be displayed by a binary network with $o(n \log n)$ reticulations. This suggests that most of the reticulations in the optimal network that displays *all* trees on n leaves, which has $\Theta(n \log n)$ reticulations [5], are due to only a tiny fraction of the exponentially many trees.

Two obvious open questions remain: First, for $t \in o(\log n)$, we believe that there are sets of t binary trees that require $(t-1)n - o(n)$ reticulations to display, even though we were only able to prove the weaker bound of $(t-1)n - o(tn)$. This is probably due to substantial overcounting of the number of sets of t trees that can be displayed by some network with r reticulations. In particular, we bound the number of these sets by the number of pairs $(N, \mathcal{T})$ such that $\mathcal{T}$ is a set of t trees, and N is a network with r reticulations that displays $\mathcal{T}$. We suspect that many such sets $\mathcal{T}$ are displayed by many networks with r reticulations, so we count them often. There are other sources of overcounting in our proofs, but this is most likely the most egregious.

The second open question concerns the bound on the number of reticulations needed to display $c \log_2 n$ unrooted binary trees in Theorem 20. This bound is roughly one third of the bound for rooted trees shown in Theorem 12. We believe that this gap is a caveat of our proof and that there is also a set of $c \log_2 n$ unrooted trees that require close to $n \log_2 n$ reticulations to display as c increases.

References

[1] Eric Bapteste et al. "Prokaryotic evolution and the tree of life are two different things". In: *Biology Direct* 4.34 (2009). DOI: 10.1186/1745-6150-4-34.

[2] Mihaela Baroni, Stefan Grünewald, Vincent Moulton, and Charles Semple. "Bounding the number of hybridisation events for a consistent evolutionary history". In: *Journal of mathematical biology* 51 (2005), pp. 171–182. DOI: 10.1007/s00285-005-0315-9.

[3] Mihaela Baroni, Charles Semple, and Mike Steel. "A framework for representing reticulate evolution". In: *Annals of Combinatorics* 8 (2005), pp. 391–408. DOI: 10.1007/s00026-004-0228-0.

[4] Mihaela Baroni, Charles Semple, and Mike Steel. "Hybrids in Real Time". In: *Systematic Biology* 55.1 (2006), pp. 46–56. DOI: 10.1080/10635150500431197.

[5] Magnus Bordewich and Charles Semple. "A universal tree-based network with the minimum number of reticulations". In: *Discrete Applied Mathematics* 250 (2018), pp. 357–362. ISSN: 0166-218X. DOI: 10.1016/j.dam.2018.05.010.

[6] Magnus Bordewich and Charles Semple. "Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4.3 (2007), pp. 458–466. DOI: 10.1109/TCBB.2007.1019.

[7] Magnus Bordewich and Charles Semple. "Computing the minimum number of hybridization events for a consistent evolutionary history". In: *Discrete Applied Mathematics* 155.8 (2007), pp. 914–928. DOI: 10.1016/j.dam.2006.08.008.

[8] Magnus Bordewich and Charles Semple. "Reticulation-visible networks". In: *Advances in Applied Mathematics* 78 (2016), pp. 114–141. ISSN: 0196-8858. DOI: 10.1016/j.aam.2016.04.004.

[9] James H. Degnan and Noah A. Rosenberg. "Gene tree discordance, phylogenetic inference and the multispecies coalescent". In: *Trends in Ecology & Evolution* 24.6 (2009), pp. 332–340. DOI: 10.1016/j.tree.2009.01.009.

[10] W. Ford Doolittle. "Phylogenetic classification and the universal tree". In: *Science* 284.5423 (1999), pp. 2124–2129. DOI: 10.1126/science.284.5423.2124.

[11] Andreas D. M. Gunawan. "Solving the Tree Containment Problem for Reticulation-Visible Networks in Linear Time". In: *Algorithms for Computational Biology*. Cham: Springer International Publishing, 2018, pp. 24–36. ISBN: 978-3-319-91938-6. DOI: 10.1007/978-3-319-91938-6_3.

[12] Andreas D.M. Gunawan, Bhaskar DasGupta, and Louxin Zhang. "A decomposition theorem and two algorithms for reticulation-visible networks". In: *Information and Computation* 252 (2017), pp. 161–175. ISSN: 0890-5401. DOI: 10.1016/j.ic.2016.11.001.

[13] Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge, UK: Cambridge University Press, 2010. ISBN: 9780521755962. DOI: 10.1017/CBO9780511974076.

[14] Daniel H. Huson and Celine Scornavacca. "Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks". In: *Systematic Biology* 61.6 (2012), pp. 1061–1067. DOI: 10.1093/sysbio/sys062.

[15] Leo van Iersel, Mark Jones, Kaari Landry, Simone Linz, Mathias Weller, Norbert Zeh, and Louxin Zhang. "Cluster Reduction on Multiple Binary Trees". Manuscript in preparation. 2026.

[16] Leo van Iersel, Mark Jones, and Mathias Weller. "When three trees go to war". In: *Peer Community Journal* 4 (2024). article no. e54. DOI: 10.24072/pcjournal.419.

[17] Leo van Iersel and Steven Kelk. "When two trees go to war". In: *Journal of theoretical biology* 269.1 (2011), pp. 245–255. DOI: 10.1016/j.jtbi.2010.10.032.

[18] Leo van Iersel, Charles Semple, and Mike Steel. "Locating a tree in a phylogenetic network". In: *Information Processing Letters* 110.23 (2010), pp. 1037–1043. DOI: 10.1016/j.ipl.2010.07.027.

[19] Iyad A Kanj, Luay Nakhleh, Cuong Than, and Ge Xia. "Seeing the trees and their branches in the network is hard". In: *Theoretical Computer Science* 401.1-3 (2008), pp. 153–164. DOI: 10.1016/j.tcs.2008.04.019.

[20] Steven Kelk, Leo van Iersel, Nela Lekic, Simone Linz, Celine Scornavacca, and Leen Stougie. "Cycle killer... qu'est-ce que c'est? On the comparative approximability of hybridization number and directed feedback vertex set". In: *SIAM journal on discrete mathematics* 26.4 (2012), pp. 1635–1656. DOI: 10.1137/120864350.

[21] Zhijiang Li and Norbert Zeh. "Computing Maximum Agreement Forests without Cluster Partitioning is Folly". In: *25th Annual European Symposium on Algorithms, ESA 2017, September 4-6, 2017, Vienna, Austria*. Vol. 87. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, 56:1–56:14. DOI: 10.4230/LIPICS.ESA.2017.56.

[22] James Mallet. "Hybridization as an invasion of the genome". In: *Trends in Ecology & Evolution* 20.5 (2005), pp. 229–237. DOI: 10.1016/j.tree.2005.02.010.

[23] David A. Morrison. *Introduction to Phylogenetic Networks*. Uppsala, Sweden: RJR Productions, 2011. ISBN: 9789198009903. URL: https://res.slu.se/id/publ/36850.

[24] Charles Semple and Mike Steel. *Phylogenetics*. Oxford, UK: Oxford University Press, 2003. ISBN: 9780198509424.

[25] Mathias Weller. "Linear-Time Tree Containment in Phylogenetic Networks". In: *Proceedings of the 16th International Conference on Comparative Genomics (RECOMB-CG 2018)*. 2018, pp. 309–323. DOI: 10.1007/978-3-030-00834-5_18.

[26] Yufeng Wu and Louxin Zhang. "Bounding the number of reticulation events for displaying multiple trees in a phylogenetic network". In: *Journal of Computer and System Sciences* 152 (2025), p. 103657. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2025.103657.