Bulletin of Mathematical Biology (2024) 86:121
https://doi.org/10.1007/s11538-024-01350-9

Society for Mathematical Biology

ORIGINAL ARTICLE

Bounding the Softwired Parsimony Score of a Phylogenetic Network

Janosch Döcker[1] · **Simone Linz**[1] · **Kristina Wicke**[2]

Received: 29 May 2024 / Accepted: 16 August 2024

Abstract

In comparison to phylogenetic trees, phylogenetic networks are more suitable to represent complex evolutionary histories of species whose past includes reticulation such as hybridisation or lateral gene transfer. However, the reconstruction of phylogenetic networks remains challenging and computationally expensive due to their intricate structural properties. For example, the small parsimony problem that is solvable in polynomial time for phylogenetic trees, becomes NP-hard on phylogenetic networks under softwired and parental parsimony, even for a single binary character and structurally constrained networks. To calculate the parsimony score of a phylogenetic network N, these two parsimony notions consider different exponential-size sets of phylogenetic trees that can be extracted from N and infer the minimum parsimony score over all trees in the set. In this paper, we ask: What is the maximum difference between the parsimony score of any phylogenetic tree that is contained in the set of considered trees and a phylogenetic tree whose parsimony score equates to the parsimony score of N? Given a gap-free sequence alignment of multi-state characters and a rooted binary level-k phylogenetic network, we use the novel concept of an informative blob to show that this difference is bounded by $k + 1$ times the softwired parsimony score of N. In particular, the difference is independent of the alignment length and the number of character states. We show that an analogous bound can be obtained for the softwired parsimony score of semi-directed networks, while under parental parsimony on the other hand, such a bound does not hold.

Keywords Phylogenetic networks · Level-k · Softwired parsimony · Parental parsimony

✉ Simone Linz
s.linz@auckland.ac.nz

Janosch Döcker
janosch.doecker@gmail.com

Kristina Wicke
kristina.wicke@njit.edu

[1] School of Computer Science, University of Auckland, Auckland, New Zealand

[2] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, USA

Published online: 22 August 2024

1 Introduction

The generalisation of phylogenetic trees to phylogenetic networks goes along with the development of new methods to reconstruct phylogenetic networks from genomic data. Since phylogenetic networks are structurally much more complicated than phylogenetic trees, the algorithms to infer networks are typically computationally expensive. Indeed, several optimisation problems that can be solved efficiently for phylogenetic trees, become computationally difficult for phylogenetic networks. For example, the small parsimony problem, which seeks to find the *parsimony score* of a given phylogenetic tree with character states assigned to its leaves is solvable in polynomial time for phylogenetic trees, e.g., using the well-known Fitch-Hartigan algorithm (Fitch 1971; Hartigan 1973), but becomes NP-hard under different notions of parsimony for phylogenetic networks, even for a single binary character and structurally constrained networks (Fischer et al. 2015; van Iersel et al. 2018).

Despite the popularity of model-based methods to infer phylogenetic trees, maximum parsimony (see, e.g. Felsenstein 2004 and references therein) continues to be widely used in certain areas of evolutionary biology, such as the analysis of morphological data (see, e.g. Sansom et al. 2018; Schrago et al. 2018; Smith 2019). Moreover, since calculating the parsimony score of a phylogenetic tree is computationally less expensive than calculating its likelihood, parsimony trees are often used as starting trees from which a search through tree space is started (Stamatakis et al. 2005) and are also used in Bayesian phylogenetic inference (Zhang et al. 2020).

Recently, different notions for parsimony on rooted phylogenetic networks have been proposed, referred to as *hardwired*, *softwired*, and *parental* parsimony. Softwired (Nakhleh et al. 2005) and parental (van Iersel et al. 2018) parsimony both consider collections of trees that can be extracted from a rooted phylogenetic network (so-called *displayed trees*, respectively *parental trees*) and define the parsimony score as the minimum parsimony score of any tree in the collection. Softwired parsimony is implemented in the popular software package PhyloNetworks (Solís-Lemus et al. 2017) and is the main focus of this paper. We remark, however, that two definitions of softwired parsimony are used in the literature and these are both formally defined in Sect. 2. Informally, in the first definition, every character in a sequence alignment can follow a different displayed tree, whereas in the second, all characters follow the same tree. We will focus on the first definition and then show that our results also hold for the second. Hardwired parsimony (Kannan and Wheeler 2012), on the other hand, calculates the parsimony score of a phylogenetic network by considering character-state transitions along all edges of the network. As the sets of rooted phylogenetic trees that are evaluated when computing the softwired and parental parsimony scores of a phylogenetic network have exponential size, it is of interest to investigate the differences in parsimony scores of elements in these sets. Given a gap-free alignment of multi-state characters and a rooted binary level-k network N (formally defined below), we analyse how different the parsimony score of any phylogenetic tree displayed by N and the softwired parsimony score of N can be. We show that independent of the alignment length and number of character states, this difference is bounded by $k + 1$ times the parsimony score of N. Thus, while computing the softwired parsimony score is in general an NP-hard problem (even for a single binary character and a structurally constrained network (Fischer et al. 2015, Theorem 4.3)), our result implies that an upper bound for the softwired parsimony score of N can be obtained in polynomial time by simply evaluating the parsimony score of an arbitrary phylogenetic tree that is displayed by N. In particular if the level of N is small, this upper bound gives a good indication of the magnitude of the softwired parsimony score of N.

Related to our result, it was shown in Theorem 5.7 of Fischer et al. (2015) that the NP-hard optimisation problem of computing the softwired parsimony of a rooted level-k network for a single multi-state character is, on the positive side, also fixed-parameter tractable, when the parameter is k. If one considers more than a single binary character, the softwired parsimony problem is NP-hard even for a rooted level-1 network (Kelk et al. 2019, Theorem 1). As a consequence of the latter negative result, Kelk et al. (2019) posed the following question. Are there good (i.e. constant-factor) approximation algorithms for computing the softwired parsimony score of a rooted phylogenetic network N and a sequence alignment A without gaps under the following three restrictions: (i) N is level-1, (ii) each biconnected component of N has exactly three outgoing edges, and (iii) A consists of binary characters?

As hinted at above, from an algorithmic perspective, our upper bound result implies a $(k + 1)$-approximation algorithm for computing the softwired parsimony score of a rooted binary level-k network N. Specifically, take an arbitrary phylogenetic tree that is displayed by N, compute its parsimony score, and use this to approximate the softwired parsimony score of N. If the level of N is fixed, this algorithm provides a polynomial-time constant factor approximation. Hence, we answer the aforementioned question by Kelk et al. affirmatively for a much larger class of rooted phylogenetic networks in the sense that our result holds for level-k networks (for a fixed non-negative integer k), it does not require restriction (ii), and it holds for gap-free alignments independent of the number of character states. Our result also complements a recent paper by Frohn and Kelk (2023), in which the authors establish a 2-approximation algorithm for the softwired parsimony problem on binary tree-child networks for a single character.

While softwired parsimony for rooted phylogenetic networks is the main focus of our paper, we additionally show that an analogous upper bound for the softwired parsimony score holds for *semi-directed* networks that are obtained from rooted phylogenetic networks by deleting the root and omitting the direction of all but reticulation edges. Semi-directed networks have recently been central to studying identifiability questions related to phylogenetic networks and to developing phylogenetic network estimation algorithms (e.g. Allman et al. 2019; Gross and Long 2018; Hollering and Sullivant 2021; Solís-Lemus and Ané 2016). We also briefly turn to the notion of parental parsimony (on rooted phylogenetic networks) and show by way of counterexample that an analogous bound for the parental parsimony score does not hold.

The remainder of this paper is organised as follows. We define all relevant concepts related to phylogenetic trees and networks, introduce the notion of softwired parsimony, and state the main result in Sect. 2. In Sect. 3, we revisit the rSPR distance and establish an upper bound on this distance for two phylogenetic trees that are both displayed by a given phylogenetic network. In Sects. 4 and 5, we introduce the notion of an *informative blob* and a *blob reduction*, respectively. Informative blobs are a novel concept that is crucial for obtaining our main result, the upper bound on the softwired parsimony score, which we establish in Sect. 6. Sections 7 and 8 are then devoted to parental parsimony on rooted networks and softwired parsimony on semi-directed networks, respectively. We end the paper with some concluding remarks and directions for future research in Sect. 9.

2 Preliminaries and Statement of Main Result

This section introduces notation and terminology, and states the main result. Throughout this paper, X denotes a non-empty finite set. Let G be a directed graph. We use $V(G)$ and $E(G)$ to denote the vertex set and edge set, respectively, of G. Furthermore, for each edge (u, v) of G, u is called a *parent* of v and v is called a *child* of u. We sometimes also refer to u and v as *neighbours* in G. In a similar vein, for two (not necessarily distinct) vertices s and t of G, we say that s (resp. t) is an *ancestor* (resp. *descendant*) of t (resp. s) if there is a directed path of length zero or more from s to t. Now let G and G' be two directed graphs, and let $e = (u, w)$ be an edge of G. Then *subdividing* e is the operation of replacing e with two new edges (u, v) and (v, w). Furthermore, we call G' a *subdivision* of G if G' can be obtained from G by repeatedly subdividing an edge. We also consider G to be a subdivision of itself.

Phylogenetic Trees and Networks. A *rooted binary phylogenetic network N on X* is a rooted acyclic digraph with no loops and no parallel edges that satisfies the following three properties:

(i) the set of leaves is X,
(ii) the out-degree of the (unique) root ρ is exactly one, and
(iii) every other vertex has either in-degree one and out-degree two, or in-degree two and out-degree one.

The set X is also sometimes called the *label set* of N. Furthermore, a vertex of N is referred to as a *reticulation* if it has in-degree two and as a *tree vertex* if it has in-degree one and out-degree two. Similarly, an edge of N that is directed into a reticulation is referred to as a *reticulation edge*. We denote the number of reticulations in N by $h(N)$.

Let N be a rooted binary phylogenetic network on X. If N has no reticulation, then it is called a *rooted binary phylogenetic X-tree*. Since all phylogenetic trees and networks are rooted and binary throughout this paper except for Sect. 8, we refer to a rooted binary phylogenetic network as a *phylogenetic network on X* and to a rooted binary phylogenetic tree as a *phylogenetic X-tree*.

Let S be a subdivision of a phylogenetic X-tree. We call the directed path from the root of S to its closest degree-three vertex its *root path*. If S is a phylogenetic tree, then the root path consists of a single edge, in which case we sometimes refer to the root path as the *root edge*.

Now let N be a phylogenetic network. A *biconnected component* of N is a maximal subgraph of N that is connected and cannot be disconnected by deleting exactly one of its vertices. Furthermore, a vertex of a biconnected component of N is called a *reticulation* if it is a reticulation in N. With this definition in hand, we say that N is *level-k* if the maximum number of reticulations of a biconnected component of N is

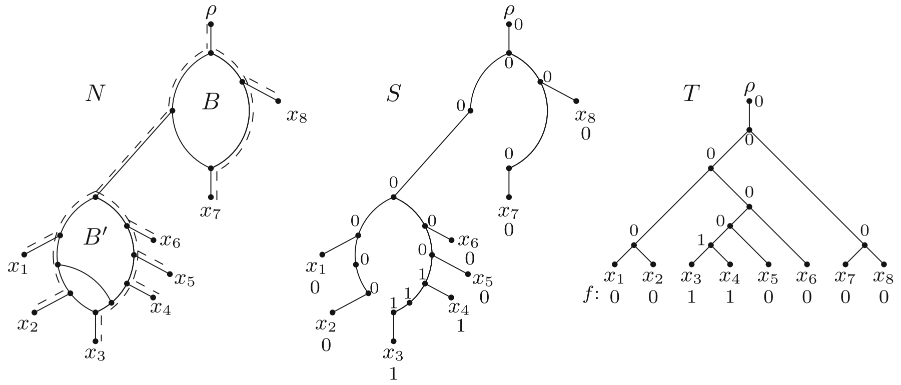

Fig. 1 A phylogenetic network N (left), an embedding S of a phylogenetic X-tree displayed by N together with a binary character f and an extension F (middle; also indicated by the dashed lines in the left panel), and the phylogenetic X-tree T obtained from S by suppressing vertices of in-degree one and out-degree one (right)

at most k. Lastly, we call a biconnected component of N a *blob* if it has at least one reticulation. For a blob B of N, we refer to the unique vertex with in-degree zero and out-degree two in B as the *source* of B. A phylogenetic network N on $\{x_1, x_2, \ldots, x_8\}$ with two blobs B and B' is shown on the left-hand side of Fig. 1.

Clusters. Let M be a subdivision of a phylogenetic network on X, and let Y be a subset of X. We call Y a *cluster* of M if there exists a vertex v in M that has precisely Y as its set of descendant leaves. Note that there may be more than one vertex in M whose cluster is Y and that this may also be the case if M is a subdivision of a phylogenetic X-tree. Furthermore, we use $\mathrm{cl}_M(v)$ or $\mathrm{cl}(v)$ if the subscript is clear from the context to denote the cluster of a given vertex v of M.

Displaying. Let N be a phylogenetic network on X with root ρ, and let T be a phylogenetic X'-tree with $X' \subseteq X$. We say that T is *displayed* by N if there exists a subgraph of N that is a subdivision of T that includes ρ, in which case this subgraph is called an *embedding* of T in N. The set of all phylogenetic X-trees that are displayed by N is referred to as the *display set* of N and denoted by $D(N)$. Ignoring the assignment of 0 and 1 to vertices for the moment, Fig. 1 shows a phylogenetic network N, a phylogenetic tree T that is displayed by N, and an embedding S of T in N. Now, consider a subset R of the reticulation edges of N. We refer to R as a *switching* if, for each reticulation v in N, it contains exactly one of the two edges that are directed into v. By deleting each reticulation edge of N that is not in R, we obtain a connected subgraph G of N with no underlying cycle and, for each leaf $\ell \in X$, there is a directed path from the root of G, which coincides with ρ, to ℓ. If we repeatedly suppress each vertex in G with in-degree one and out-degree one, and delete each vertex in G with out-degree zero that is not in X until no such operation is possible, we obtain a phylogenetic X-tree T_R. We say that R yields T_R. By construction, T_R is displayed by N. Conversely, observe that, for each phylogenetic X-tree T in $D(N)$, there exists at least one switching that yields T. In summary, we have the following observation.

Observation 1 Let N be a phylogenetic network on X, and let T be a phylogenetic X-tree. Then T is displayed by N if and only if there exists a switching of N that yields T.

rSPR Operation. Let T be a phylogenetic X-tree, and let $e = (u, v)$ be an edge of T that is not incident with the root. Let T' be a phylogenetic X-tree obtained from T by deleting e and reattaching the resulting rooted subtree that contains v via a new edge f in the following way: Subdivide an edge of the component that contains the root of T with a new vertex u', join u' and v with f, and suppress u. We say that T' has been obtained from T by a *rooted subtree prune and regraft* (rSPR) operation. The *rSPR distance* between any two phylogenetic X-trees T and T', denoted by $d_{\mathrm{rSPR}}(T, T')$, is the minimum number of rSPR operations that transform T into T'. It is well known that T' can always be obtained from T by a sequence of single rSPR operations and, so, $d_{\mathrm{rSPR}}(T, T')$ is well defined.

Characters. An *r-state character on X* is a surjective function $f\colon X \to C$ from X into a set C of character states with $r = |C| \geq 1$. If $r = 2$, then f is called a *binary* character. Throughout this paper, all results are established for r-state characters with r being fixed and arbitrarily large. For simplicity, we refer to an r-state character on X as a *character* on X.

Let G be an acyclic digraph with leaf set X, and let $f\colon X \to C$ be a character on X. An *extension* of f to $V(G)$ is a function $F\colon V(G) \to C$ such that $F(\ell) = f(\ell)$ for each element $\ell \in X$. For an extension F of f, we set

$$\mathrm{ch}(F, G) = |\{(u, v) \in E(G) : F(u) \neq F(v)\}|,$$

and refer to $\mathrm{ch}(F, G)$ as the *changing number* of F. Intuitively, each edge of G that contributes to the changing number of F requires a character-state transition to explain f on G. Lastly, we say that an extension F of f to $V(G)$ is *minimum* if there exists no extension of f to $V(G)$ whose changing number is strictly smaller than that of F.

In what follows, we often consider a sequence $(f_1, f_2, \ldots, f_n)$ of characters on X instead of a single character. We call such a sequence an *alignment*. Unless stated otherwise, all alignments in this paper are sequences of r-state characters for $r \geq 2$ that do not contain the gap symbol "–". Such an alignment is referred to as *gap-free*. In applied phylogenetics, multiple sequence alignments frequently contain gaps which, intuitively, are placeholders that can take on any of the other r character states. We will see in the last section why the restriction to gap-free alignments is necessary. Lastly, we denote a sequence (f_1) that consists of a single element by f_1 and omit parentheses for simplicity.

Parsimony on Phylogenetic Trees and their Subdivisions. Given an alignment $A = (f_1, f_2, \ldots, f_n)$ of characters on X and an arbitrary rooted tree T with leaf set X, we refer to

$$PS(A, T) = \sum_{i=1}^{n} \min_{F_i}(\mathrm{ch}(F_i, T))$$

as the *parsimony score* of A on T, where the minimum is taken over all extensions of f_i to $V(T)$.

Instead of calculating the parsimony score of a phylogenetic X-tree T, we are often interested in calculating the parsimony score of a subdivision of T in the upcoming sections. The next lemma states that both scores are equal. Its correctness can be established analogously to the proof of Lemma 4.5 in Fischer et al. (2015). In the proof of this lemma, Fischer et al. (2015) have shown that the parsimony score of a character f on a phylogenetic tree T is equal to the parsimony score of f on a particular rooted tree. This latter tree is more general than a subdivision of T in the sense that it may contain unlabelled leaves in addition to the leaves in X.

Lemma 2 *Let f be a character on X, and let S be a subdivision of a phylogenetic X-tree T. Then $PS(f, S) = PS(f, T)$.*

We also have the following observation.

Observation 3 Let f be a character on X, and let S be a subdivision of a phylogenetic X-tree. If F is an extension of f to $V(S)$ such that $F(u) \neq F(v)$ for some edge (u, v) of the root path of S, then there exists an extension F' of f to $V(S)$ such that $F'(u) = F'(v)$ and $\text{ch}(F', T) < \text{ch}(F, T)$.

By Observation 3, we freely assume throughout the remainder of the paper that every extension F of a character to the vertices of a subdivision of a phylogenetic tree has the additional property that there is no character state transition on any edge of its root path.

Parsimony on Phylogenetic Networks. As outlined in the introduction, several notions of parsimony have been introduced that generalise parsimony from phylogenetic trees to phylogenetic networks. In this paper, we are focusing on the notion of softwired parsimony and briefly touch on parental parsimony in Sect. 7. Following Nakhleh et al. (2005), we now formally define softwired parsimony. Let $A = (f_1, f_2, \ldots, f_n)$ be an alignment of characters on X, and let N be a phylogenetic network on X. The *softwired parsimony score* of A on N is defined as

$$\begin{aligned} PS_{\text{sw}}(A, N) = \sum_{i=1}^{n} PS_{\text{sw}}(f_i, N) &= \sum_{i=1}^{n} \min_{T \in D(N)} \min_{F_i}(\text{ch}(F_i, T)) \\ &= \sum_{i=1}^{n} \min_{T \in D(N)} PS(f_i, T), \end{aligned} \tag{1}$$

where, for each character f_i, the first minimum is taken over all phylogenetic trees in the display set of N and the second minimum is taken over all extensions of f_i to $V(T)$. As per Eq. (1), each character in A can follow a different tree in $D(N)$. A slightly more restricted definition of softwired parsimony, which has also appeared in the literature (e.g. see Kelk and Fischer 2017; Kelk et al. 2019), is the following

$$PS_{\text{sw}'}(A, N) = \min_{T \in D(N)} \sum_{i=1}^{n} \min_{F_i}(\text{ch}(F_i, T)) = \min_{T \in D(N)} \sum_{i=1}^{n} PS(f_i, T), \tag{2}$$

where all characters in A follow the same tree in $D(N)$. Clearly, if $n = 1$, then $PS_{\mathrm{sw}}(A, N) = PS_{\mathrm{sw}'}(A, N)$. On the other hand, for $n \geq 1$, it follows from the definition that $PS_{\mathrm{sw}}(A, N) \leq PS_{\mathrm{sw}'}(A, N)$. For the purpose of the upcoming sections, we adopt the softwired parsimony definition as formalised in Eq. (1) and will see later, that our main result holds also under the definition given in Eq. (2). In this context, it is worth mentioning that the hardness result for computing the softwired parsimony score of a level-1 network for an alignment of at least two binary characters (Kelk et al. 2019, Theorem 1) as mentioned in the introduction has been established for the definition given in Eq. (2).

Statement of Main Result. The main result of this paper is the following theorem which we establish in Sect. 6.

Theorem 4 *Let N be a phylogenetic network on X, and let T be a phylogenetic X-tree in $D(N)$. Furthermore, let A be an alignment of characters on X. Then*

$$PS(A, T) \leq (k + 1) \cdot PS_{\mathrm{sw}}(A, N) \text{ and } PS(A, T) \leq (k + 1) \cdot PS_{\mathrm{sw}'}(A, N),$$

where k is the level of N.

For example, if N is a level-1 network, Theorem 4 implies that the parsimony score of an arbitrary tree displayed by N is at most twice the parsimony score of a tree displayed by N whose parsimony score is equal to $PS_{\mathrm{sw}}(A, N)$. Moreover, we show in Sect. 6 that the bound as stated in Theorem 4 is sharp.

The next corollary positively answers the open problem that is detailed in the introduction and that was first posed in Kelk et al. (2019).

Corollary 5 *For a fixed non-negative integer k, let N be a level-k network on X, and let A be an alignment of characters on X. There exists a polynomial $(k+1)$-approximation algorithm to calculate $PS_{\mathrm{sw}}(A, N)$ and $PS_{\mathrm{sw}'}(A, N)$.*

Proof Clearly, we can construct a phylogenetic X-tree T such that $T \in D(N)$ in time that is polynomial in $|V(N)|$. Furthermore, it takes time that is polynomial in $|X|$ to calculate $PS(A, T)$ by applying Fitch's algorithm (Fitch 1971). The result now follows immediately from Theorem 4. □

3 Bounding the rSPR Distance

In this section, we establish an upper bound on the rSPR distance between two phylogenetic trees for when both trees are displayed by a given network. Let N be a phylogenetic network, and let R and R' be two switchings of N. We define

$$d_{\mathrm{switch}}(R, R') = h(N) - |R \cap R'|$$

to be the *switching distance* between R and R'. Intuitively, $d_{\mathrm{switch}}(R, R')$ is the number of reticulations in N for which R and R' contain different reticulation edges.

The following lemma is a generalisation of Lemma 3.1 in Döcker et al. (2024).

Lemma 6 *Let N be a phylogenetic network on X, and let T_R and $T_{R'}$ be two phylogenetic X-trees that are yielded by two switchings R and R', respectively, of N. Then*

$$d_{\mathrm{rSPR}}(T_R, T_{R'}) \leq d_{\mathrm{switch}}(R, R').$$

Proof Let S (resp. S') be an embedding of T_R (resp. $T_{R'}$) in N whose edge set contains each edge in R (resp. R'). Obtain a directed acyclic graph N' from N by deleting each edge that is not contained in $E(S) \cup E(S')$ and, subsequently, applying any of the following two operations until no further operation is possible.

(i) Suppress a vertex of in-degree one and out-degree one.
(ii) If e and e' are two edges in parallel, delete e'.

By construction, N' is a phylogenetic network on X and each reticulation edge that is not contained in $R \cup R'$ is deleted in obtaining N'. Hence, $h(N') \leq d_{\mathrm{switch}}(R, R')$. Furthermore, as S and S' are embeddings of T_R and $T_{R'}$, respectively, T_R and $T_{R'}$ are displayed by N'. Now, let $h(T_R, T_{R'})$ be the minimum number of reticulations of any phylogenetic network that displays T_R and $T_{R'}$. Clearly, $h(T_R, T_{R'}) \leq h(N')$ and, thus

$$d_{\mathrm{switch}}(R, R') \geq h(N') \geq h(T_R, T_{R'}) \geq d_{\mathrm{rSPR}}(T_R, T_{R'}),$$

where the last inequality follows from Equation 10.1 of Semple (2007). □

4 Informative Blobs

To establish the main result of this paper, we introduce the novel concept of informative and non-informative blobs in this section. After giving a formal definition of these blobs, we establish results related to the changing number of character extensions to embeddings of phylogenetic trees that are displayed by phylogenetic networks that consist of a single informative or non-informative blob.

We begin by introducing the notion of children of a blob. Let N be a phylogenetic network on X, and let B be a blob of N. Furthermore, let $C_N(B)$ be the subset of $V(N)$ that contains precisely each vertex that is not in B and that is a child of a vertex of B. We refer to $C_N(B)$, as the *children* of B. As an immediate consequence of the definition of $C_N(B)$, we have the following lemma that we will freely use throughout the remainder of the paper.

Lemma 7 *Let B be a blob of a phylogenetic network N on X, and let v be a vertex of $C_N(B)$. Furthermore, let S be an embedding of a phylogenetic X-tree that is displayed by N. Then, v is a vertex of S. Moreover, if v is a vertex of a blob B' in N, then v is the source of B'.*

Proof We first establish the second assertion of the lemma. Suppose that v is a vertex of a blob B' in N. Since $v \in C_N(B)$, we have $B' \neq B$. Towards a contradiction, assume that v is not the source of B'. It follows that v is a vertex of in-degree two. Let

u be the parent of v in B, and let u' be the other parent of v. Then there is a directed path from the root of N to v that traverses u and a directed path from the root of N to v that traverses u', thereby contradicting that B and B' are two distinct blobs in N. We complete the proof by noting that S contains each edge of N whose deletion disconnects N into more than one connected component and, so, v is a vertex of S.

Now, let s be the source of a blob B in a phylogenetic network N on X. Furthermore, let S be an embedding of a phylogenetic X-tree that is displayed by N. Let f be a character on X, and let F be an extension of f to $V(S)$. We set $\mathrm{ind}(F, B, S) = 0$ if each element in $C_N(B)$ is assigned to the same character state under F and, otherwise, we set $\mathrm{ind}(F, B, S) = 1$. By Lemma 7, recall that each vertex in $C_N(B)$ is also a vertex of S and, thus, $\mathrm{ind}(F, B, S)$ is well defined. Moreover, we say that B is a *non-informative blob* relative to S and f if there exists an extension F of f to $V(S)$ such that $PS(f, S) = \mathrm{ch}(F, S)$ and $\mathrm{ind}(F, B, S) = 0$. Otherwise, we say that B is an *informative blob*. We next extend the concept of a single informative blob to all blobs $B_1, B_2, \ldots, B_m$ of N and set

$$b(f, N, S) = \min_{\substack{F_j \text{ such that} \\ PS(f,S)=\mathrm{ch}(F_j,S)}} \left(\sum_{i=1}^{m} \mathrm{ind}(F_j, B_i, S) \right),$$

where the minimum is taken over all extensions F_j of f to $V(S)$ whose changing number is equal to $PS(f, S)$. Then $b(f, N, S)$ denotes the number of informative blobs relative to S and f in N. If F_j is an extension of f to $V(S)$ such that $b(f, N, S) = \sum_{i=1}^{m} \mathrm{ind}(F_j, B_i, S)$, then we say that F_j *realises* $b(f, N, S)$. See Fig. 1 for an example of a phylogenetic network N on $X = \{x_1, x_2, \ldots, x_8\}$ with two blobs B and B', an embedding S of a phylogenetic X-tree displayed by N, and a binary character f on X such that $b(f, N, S) = 1$. Here, B is non-informative because there exists a minimum extension F of f that assigns character state 0 to all elements of $C_N(B)$, where $C_N(B)$ contains the source of B' and leaves x_7 and x_8. Blob B', on the other hand, is informative, as the elements in $C_N(B') = \{x_1, x_2, \ldots, x_6\}$ are assigned two different states by f and thus by any extension of it. To see that F is indeed minimum, notice that $\mathrm{ch}(F, S) = \mathrm{ch}(F, T) = 1$. This is minimum since f employs two states, and it is well-known and easy to see that, in this case, any extension of f requires at least one change.

Lemma 8 *Let f be a character on X, and let N be a phylogenetic network on X with a single blob B whose source s is the child of the root. Let S and S' be embeddings of two phylogenetic X-trees that are displayed by N. Suppose that B is non-informative relative to S' and f. Let F' be an extension of f to $V(S')$ with* $\mathrm{ch}(F', S') = PS(f, S')$ *that assigns the same character state to each vertex in $C_N(B)$. Then there exists an extension F of f to $V(S)$ such that*

(i) $\mathrm{ch}(F, S) = \mathrm{ch}(F', S')$ *and*
(ii) $F(s) = F'(s)$.

Proof Since B is non-informative, recall that F' exists. Furthermore, by the definition of an embedding, s is the child of the roots of S and S'. Let V be the subset of $V(N)$

that precisely contains each vertex that is not a vertex of B. Since B is the only blob of N, each vertex in V is also a vertex of S and S'. Furthermore, each vertex of S or S' that is not in V, is a vertex of B. Since s is the child of the root of S' and F' assigns the same character state, say α, to each vertex in $C_N(B)$, it follows that F' also assigns α to each vertex in $V(S')$ that is an ancestor of some vertex in $C_N(B)$. In particular, $F'(s) = \alpha$.

Now, consider S. Set $F(u) = F'(u)$ for each vertex $u \in V$ and set $F(u') = \alpha$ for each vertex $u' \in V(S) \setminus V$. By definition of F, we again have that each vertex in $V(S)$ that is an ancestor of some vertex in $C_N(B)$ is assigned to α under F. Hence, as F' is an extension of f to $V(S')$, F is an extension of f to $V(S)$ with $F(s) = F'(s) = \alpha$; thereby satisfying (ii). Moreover, since S and S' are embeddings of two phylogenetic X-trees that are displayed by N, the edges of N satisfy the following property: If $e = (u, v)$ is an edge of S' (resp. S) but not an edge of S (resp. S'), then e is an edge of B and, consequently, $F'(u) = F'(v) = \alpha$ (resp. $F(u) = F(v) = \alpha$). It follows that $\text{ch}(F, S) = \text{ch}(F', S')$ which satisfies (i) and, therefore, completes the proof of the lemma.

Lemma 9 *Let f be a character on X. Let T and T' be two phylogenetic X-trees. Furthermore, let F' be an extension of f to $V(T')$. Then there exists an extension F of f to $V(T)$ such that*

(i) $\text{ch}(F, T) \leq \text{ch}(F', T') + d_{\text{rSPR}}(T', T)$ *and*
(ii) $F(\rho) = F'(\rho')$*, where ρ and ρ' is the root of T and T', respectively.*

Proof We show by induction on $d_{\text{rSPR}}(T', T)$ that there exists an extension F of f to $V(T)$ that satisfies (i)–(ii). Suppose that $d_{\text{rSPR}}(T', T) = 1$. Then there exists a single rSPR operation that transforms T' into T. Given such an rSPR operation, let (u', v') be the edge of T' that is deleted in the pruning part of the operation. Let u'_p and $u'_c \neq v'$ be the parent and other child of u' in T'. Further, let u be the vertex that subdivides an edge, say (u_p, u_c), when reattaching the resulting subtree with root v' such that (u, v') is an edge in T. Noting that each vertex in T except for u is also a vertex of T', we next obtain an extension F of f to $V(T)$ with no character state transition on the root edge of T as follows: For each vertex $w \neq u$, we set $F(w) = F'(w)$. In particular, we have $F(\rho) = F'(\rho')$ and, so, (ii) follows. Moreover, if $u_p = \rho$, we set $F(u) = F(u_p)$. Otherwise, we set $F(u) = \alpha$, where α is a character state that has been assigned to at least one neighbour of u in T under F and there is no other character state that has been assigned to strictly more neighbours of u in T under F. We next show that (i) holds. Considering the edges of T, it follows from the rSPR operation transforming T' into T as described above that, except for the edges (u, v'), (u'_p, u'_c), (u_p, u), and (u, u_c), each edge of T is also an edge of T'. If $F(u'_p) \neq F(u'_c)$, then either $F'(u'_p) \neq F'(u')$ or $F'(u') \neq F'(u'_c)$. Hence, suppressing u' does not increase the changing number. On the other hand, when assigning a character state to u the changing number may increase. More specifically, we consider three cases. First, if $F(u_p) = F(u_c)$, then $F(u) = F(u_p)$ by definition of F. Note that u_p may be ρ. Thus, there is no character state transition on the two edges (u_p, u) and (u, u_c), and at most one such transition on the edge (u, v') under F in T. Second, if $|\{F(u_p), F(u_c), F(v')\}| = 3$, then there is a character state transition on the edge (u_p, u_c) under F' and we have two character state transitions on the three edges (u_p, u), (u, u_c), and (u, v') under F. Third, if $F(u_p) \neq F(u_c)$ and $|\{F(u_p), F(u_c), F(v')\}| = 2$, then there is again a character state transition on the edge (u_p, u_c) under F' and we have one character state transition on the three edges (u_p, u), (u, u_c), and (u, v') under F. Hence, regardless of which case applies

$$\mathrm{ch}(F, T) \leq \mathrm{ch}(F', T') + 1 = \mathrm{ch}(F', T') + d_{\mathrm{rSPR}}(T', T);$$

thereby satisfying (i) for when $d_{\mathrm{rSPR}}(T', T) = 1$.

Now suppose that $d_{\mathrm{rSPR}}(T', T) \geq 2$ and that (i)–(ii) are satisfied for all pairs of phylogenetic trees whose rSPR distance is strictly smaller than $d_{\mathrm{rSPR}}(T', T)$. Let T'' be a phylogenetic X-tree such that $d_{\mathrm{rSPR}}(T', T'') = 1$ and $d_{\mathrm{rSPR}}(T'', T) = d_{\mathrm{rSPR}}(T', T) - 1$. Recalling that F' is an extension of f to $V(T')$, it follows from the induction hypothesis, that there is an extension F'' of f to $V(T'')$ that satisfies (ii) and $\mathrm{ch}(F'', T'') \leq \mathrm{ch}(F', T') + 1$. Again, by the induction hypothesis, there exists an extension F of f to $V(T)$ that satisfies (ii) and $\mathrm{ch}(F, T) \leq \mathrm{ch}(F'', T'') + d_{\mathrm{rSPR}}(T'', T)$. Hence, by combining the two inequalities we obtain

$$\begin{aligned}\mathrm{ch}(F, T) &\leq \mathrm{ch}(F'', T'') + d_{\mathrm{rSPR}}(T'', T)\\ &\leq \mathrm{ch}(F', T') + 1 + d_{\mathrm{rSPR}}(T'', T) = \mathrm{ch}(F', T') + d_{\mathrm{rSPR}}(T', T)\end{aligned}$$

and $F(\rho) = F'(\rho')$. Hence, F satisfies (i)–(ii). This completes the proof of the lemma. □

In the following corollary, we apply Lemma 9 to two phylogenetic X-trees displayed by a phylogenetic network. In this case, the unique root vertex of the network, ρ, is viewed as the root of T (resp. T').

Corollary 10 *Let f be a character on X. Let N be a phylogenetic network on X with root ρ and a single blob, and let T and T' be two phylogenetic X-trees displayed by N. Furthermore, let F' be an extension of f to $V(T')$. Then there exists an extension F of f to $V(T)$ such that*

(i) $\mathrm{ch}(F, T) \leq \mathrm{ch}(F', T') + k$, *where k is the level of N, and*
(ii) $F(\rho) = F'(\rho)$.

Proof Let R and R' be two switchings of N that yield T and T', respectively. By Lemma 6, we have $d_{\mathrm{rSPR}}(T, T') \leq d_{\mathrm{switch}}(R, R')$. Noting that N has a single blob, we have $d_{\mathrm{switch}}(R, R') \leq k$ and the corollary now follows from Lemma 9. □

While Lemma 8 is restricted to phylogenetic networks that consist of a single non-informative blob, the next lemma establishes an analogous result for all phylogenetic networks that consist of a single blob.

Lemma 11 *Let f be a character on X, and let N be a phylogenetic network on X with a single blob B whose source s is the child of the root. Let S and S' be embeddings of two phylogenetic trees that are displayed by N. Furthermore, let F' be an extension of f to $V(S')$. Then there exists an extension F of f to $V(S)$ such that*

(i) $\text{ch}(F, S) \leq \text{ch}(F', S') + k$*, where* k *is the level of* N*, and*
(ii) $F(s) = F'(s)$.

Proof Let T (resp. T') be the two phylogenetic X-trees such that S (resp. S') is an embedding of T (resp. T') in N. Furthermore, let $F'(s) = \alpha$. Observe that each vertex in T' is a unique degree-three vertex in S'. First, let $F'_{T'}$ be the extension of f to $V(T')$ obtained from F' by setting $F'_{T'}(w) = F'(w)$ for each $w \in V(T')$. Then

$$\text{ch}(F'_{T'}, T') \leq \text{ch}(F', S'), \tag{3}$$

and, because there is no character state transition on any edge of the root path of S' that contains s, the root of T' is assigned to α. Second, by Corollary 10, there exists an extension F_T of f to $V(T)$ such that

$$\text{ch}(F_T, T) \leq \text{ch}(F'_{T'}, T') + k \tag{4}$$

and the root of T is also assigned to α. Third, we obtain an extension F of f to $V(S)$ from F_T as follows. Noting that each edge (v, v') in T corresponds to a unique directed path $v = v_1, v_2, \ldots, v_s = v'$ in S whose non-terminal vertices all have degree two, we set $F(v_1) = F(v_2) = \cdots = F(v_{s-1}) = F_T(v)$ and $F(v_s) = F_T(v')$. Then

$$\text{ch}(F, S) = \text{ch}(F_T, T), \tag{5}$$

and, since the root of T is assigned to α, it follows that each vertex on the root path of S, in particular s, is also assigned to α. Hence (ii) is satisfied. Moreover, by combining Eqs. (3)–(5), we have

$$\text{ch}(F, S) = \text{ch}(F_T, T) \leq \text{ch}(F'_{T'}, T') + k \leq \text{ch}(F', S') + k.$$

This concludes the proof of the lemma. □

5 Blob Reduction

In this section, we introduce the notion of a blob reduction. Intuitively, this allows us to decompose a phylogenetic network N into two smaller phylogenetic networks and calculate the parsimony score of an embedding S of a phylogenetic X-tree displayed by N based on these two smaller networks.

Let N be a phylogenetic network on X. A blob B of N with source s is called a *maximal* blob if s is not an ancestor of the source of any blob in N that is distinct from B. Informally, B is maximal if there is no other blob below it. Now, let B be a maximal blob of N with source s, and let $Y = \text{cl}(s)$. For some $y \notin X$, the *blob reduction* of B reduces N to two smaller phylogenetic networks as follows. Let $N(\bar{Y})$ be the phylogenetic network on $(X \setminus Y) \cup \{y\}$ that is obtained from N by replacing the subnetwork of N that is rooted at s with a single new leaf y. Furthermore, let $N(Y)$ be the phylogenetic network on Y that is obtained from the subnetwork of N

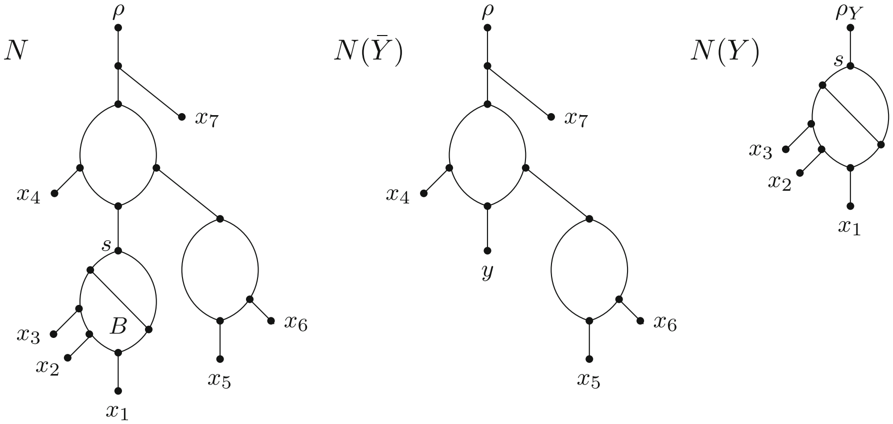

Fig. 2 An example of a blob reduction applied to a phylogenetic network N that replaces the maximal blob B with source s and $\mathrm{cl}(s) = \{x_1, x_2, x_3\}$ with a single new leaf y to obtain $N(\bar{Y})$

that is rooted at s by adding a new vertex ρ_Y and edge (ρ_Y, s). By construction, each of $N(\bar{Y})$ and $N(Y)$ contains at least one leaf. Figure 2 illustrates an example of a blob reduction.

Now, let S be an embedding of a phylogenetic X-tree T that is displayed by N. Recall that s is a vertex of S and $\mathrm{cl}(s) = Y$. Let f be a character on X. Using the aforementioned blob reduction of B as a guide, we next also reduce S to two smaller trees such that one of the resulting trees is an embedding of a subtree of T in $N(\bar{Y})$ and the other one is an embedding of another subtree of T in $N(Y)$. More specifically, let $S(\bar{Y})$ be the tree with leaf set $(X \setminus Y) \cup \{y\}$ that is obtained from S by replacing the subtree of S that is rooted at s with a single new leaf y. Furthermore, let $S(Y)$ be the tree with leaf set Y that is obtained from the subtree of S that is rooted at s by adding a new vertex ρ_Y and edge (ρ_Y, s). We call $(S(Y), S(\bar{Y}))$ the *cluster tree pair* of S relative to B. Let f_Y and $f_{\bar{Y}}$ be a character on Y and on $(X \setminus Y) \cup \{y\}$, respectively, such that $f_Y(\ell) = f(\ell)$ for each $\ell \in Y$ and $f_{\bar{Y}}(\ell') = f(\ell')$ for each $\ell' \in X \setminus Y$. We refer to extensions F_Y of f_Y to $V(S(Y))$ and $F_{\bar{Y}}$ of $f_{\bar{Y}}$ to $V(S(\bar{Y}))$ as a *pair of cluster extensions with respect to* f if $F_{\bar{Y}}(y) = F_Y(\rho_Y)$. Except for $f_{\bar{Y}}(y)$, observe that f uniquely determines the character state of each leaf in $S(Y)$ and $S(\bar{Y})$. Moreover, since the root path of $S(Y)$ contains at least one edge, the definition of a pair of cluster extensions implies that $F_Y(\rho_Y) = F_Y(s)$ by our assumption following Observation 3. An example of a cluster tree pair and a pair of cluster extensions is shown in Fig. 3.

The next lemma shows how the changing number of extensions of characters f, f_Y, and $f_{\bar{Y}}$ to $V(S)$, $V(S(Y))$, and $V(S(\bar{Y}))$, respectively, are related to each other.

Lemma 12 *Let B be a maximal blob of N. Let f be a character on X, and let S be an embedding of a phylogenetic X-tree that is displayed by N. Furthermore, let $(S(Y), S(\bar{Y}))$ be the cluster tree pair of S relative to B. Then, the following two statements hold.*

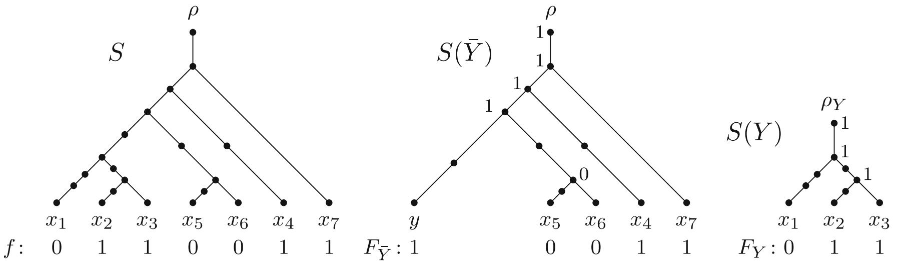

Fig. 3 An embedding S of a phylogenetic X-tree that is displayed by the phylogenetic network N on X as shown in Fig. 2, the cluster tree pair $(S(Y), S(\bar{Y}))$ of S relative to the blob B that is also shown in Fig. 2, and a pair of cluster extensions F_Y and $F_{\bar{Y}}$ with respect to the character f on X. For reasons of clarity, the extensions F_Y and $F_{\bar{Y}}$ to the vertices of $S(Y)$ and $S(\bar{Y})$, respectively, are only shown for some vertices. Observe that $F_{\bar{Y}}(y) = F_Y(\rho_Y)$ and that all leaves of $S(Y)$ and $S(\bar{Y})$, except for y, are assigned to the same character state as in S

(i) If F is an extension of f to $V(S)$, then there exists a pair of cluster extensions $(F_Y, F_{\bar{Y}})$ such that

$$\mathrm{ch}(F, S) = \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(F_{\bar{Y}}, S(\bar{Y})).$$

(ii) If $(F_Y, F_{\bar{Y}})$ is a pair of cluster extensions with respect to f, then there exists an extension F of f to $V(S)$ such that

$$\mathrm{ch}(F, S) = \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(F_{\bar{Y}}, S(\bar{Y})).$$

Proof Let s be the source of B, and let $Y = \mathrm{cl}_N(s)$. By the definition of a cluster tree pair, $S(Y)$ has leaf set Y and root ρ_Y, and $S(\bar{Y})$ has leaf set $(X \setminus Y) \cup \{y\}$ and root ρ, where ρ is also the root of S. Lastly, as s is a vertex of S, it follows from the construction of $S(Y)$ and $S(\bar{Y})$ that s corresponds to the child of ρ_Y in $S(Y)$ and to y in $S(\bar{Y})$, whereas each other vertex of S corresponds to a unique vertex in either $S(Y)$ or $S(\bar{Y})$. To ease reading, we refer to the child of ρ_Y in $S(Y)$ as s_Y. Reversely, the only vertex of $S(Y)$ and $S(\bar{Y})$ that does not correspond to a unique vertex in S is ρ_Y. We next show that (i) and (ii) hold.

First, let F be an extension of f to $V(S)$. Obtain a pair of cluster extensions F_Y and $F_{\bar{Y}}$ of characters f_Y and $f_{\bar{Y}}$ to $V(S(Y))$ and $V(S(\bar{Y}))$, respectively, in the following way. For each vertex w of $V(S(Y))\backslash\{\rho_Y\}$, set $F_Y(w) = F(w')$, where w' is the vertex of S that w corresponds to, and set $F_Y(\rho_Y) = F_Y(s_Y)$. Similarly, for each vertex w of $V(S(\bar{Y}))$, set $F_{\bar{Y}}(w) = F(w')$, where w' is the vertex of S that w corresponds to. Since y and s_Y both correspond to s, it follows that $F_{\bar{Y}}(y) = F_Y(\rho_Y)$. It is is now easily checked that F_Y and $F_{\bar{Y}}$ is a pair of cluster extensions with respect to f and that

$$\mathrm{ch}(F, S) = \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(F_{\bar{Y}}, S(\bar{Y})).$$

Hence, (i) holds.

Now, let F_Y and $F_{\bar{Y}}$ be a pair of cluster extensions with respect to f. In particular, $F_Y(\ell) = f(\ell)$ for each $\ell \in Y$, $F_{\bar{Y}}(\ell) = f(\ell)$ for each $\ell \in X \backslash Y$, and $F_{\bar{Y}}(y) = F_Y(\rho_Y)$. Now obtain an extension F of f to $V(S)$ from F_Y and $F_{\bar{Y}}$ in the following way. For each vertex w' of $V(S)$ that corresponds to a vertex w of $S(Y)$, set $F(w') = F_Y(w)$ and, for each vertex $w'$of $V(S) \setminus \{s\}$ that corresponds to a vertex w of $S(\bar{Y})$, set $F(w') = F_{\bar{Y}}(w)$. Since $F_Y(s_y) = F_Y(\rho_Y)$, it follows that

$$\mathrm{ch}(F, S) = \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(F_{\bar{Y}}, S(\bar{Y})).$$

Thus, (ii) holds as well.

6 Proof of Theorem 4

In this section, we establish the proof of Theorem 4 and show that the bound that is given in the theorem is sharp. Most work in proving Theorem 4 goes into establishing the following lemma.

Lemma 13 *Let f be a character on X. Let N be a phylogenetic network on X, and let S and S' be embeddings of two phylogenetic X-trees that are displayed by N. Furthermore, let F' be an extension of f to $V(S')$ that realises $b(f, N, S')$. Then there exists an extension F of f to $V(S)$ such that*

$$\mathrm{ch}(F, S) \leq \mathrm{ch}(F', S') + k \cdot b(f, N, S'),$$

where k is the level of N.

Proof Let $B_1, B_2, \ldots, B_m$ be the blobs of N. The proof is by induction on m. If $m = 0$, then N is a phylogenetic tree with $k = 0$ and so the result clearly follows since $S = S'$ and, therefore,

$$\mathrm{ch}(F, S) \leq \mathrm{ch}(F', S') + 0$$

when setting $F = F'$. Now assume that $m \geq 1$ and that the statement is true for all phylogenetic networks with at most $m - 1$ blobs. Let B be a maximal blob of N with source s, and let $Y = \mathrm{cl}_N(s)$. Without loss of generality, we assume that $B = B_m$.

For some $y \notin X$, let $N(\bar{Y})$ be the phylogenetic network on $(X \setminus Y) \cup \{y\}$, and let $N(Y)$ be the phylogenetic network on Y and with root ρ_Y resulting from N by applying a blob reduction to B_m. Notice that by construction, $N(Y)$ consists of the single blob B_m with s being the child of ρ_Y, whereas $N(\bar{Y})$ contains precisely $m - 1$ blobs. Moreover, let $(S(Y), S(\bar{Y}))$ be the cluster tree pair of S relative to B_m, and let $(S'(Y), S'(\bar{Y}))$ be the cluster tree pair of S' relative to B_m. Since s is a vertex of S and S', s is also a vertex of $S(Y)$ and $S'(Y)$. Now, by Lemma 12, Part (i), there exists a pair of cluster extensions $(F'_Y, G'_{\bar{Y}})$ with respect to f such that

$$\mathrm{ch}(F', S') = \mathrm{ch}(F'_Y, S'(Y)) + \mathrm{ch}(G'_{\bar{Y}}, S'(\bar{Y})) \tag{6}$$

and $G'_{\bar{Y}}(y) = F'_Y(\rho_Y) = F'_Y(s)$. Let f_Y and $g_{\bar{Y}}$ be the characters on Y and $(X\backslash Y)\cup\{y\}$, respectively, such that F'_Y and $G'_{\bar{Y}}$ are extensions of f_Y and $g_{\bar{Y}}$, respectively.

We next consider $N(\bar{Y})$ and start by making two observations. First, since $\mathrm{ch}(F', S') = PS(f, S')$, it follows from Lemma 12, Parts (i) and (ii), that $\mathrm{ch}(G'_{\bar{Y}}, S'(\bar{Y})) = PS(g_{\bar{Y}}, S'(\bar{Y}))$. Second, by the construction of the pair of cluster extensions in Lemma 12, Part (i), we may assume that, for each vertex w of $V(S'(\bar{Y}))$, we have $G'_{\bar{Y}}(w) = F'(w')$, where w' is the vertex of S' that w corresponds to. Then, as F' realises $b(f, N, S')$, $G'_{\bar{Y}}$ realises $b(g_{\bar{Y}}, N(\bar{Y}), S'(\bar{Y}))$. Noting that $N(\bar{Y})$ has $m-1$ blobs and level at most k, we now apply the induction hypothesis to obtain an extension $G_{\bar{Y}}$ of $g_{\bar{Y}}$ to $V(S(\bar{Y}))$ that satisfies

$$\mathrm{ch}(G_{\bar{Y}}, S(\bar{Y})) \leq \mathrm{ch}(G'_{\bar{Y}}, S'(\bar{Y})) + k \cdot b(g_{\bar{Y}}, N(\bar{Y}), S'(\bar{Y})) \tag{7}$$

such that $G_{\bar{Y}}(y) = g_{\bar{Y}}(y) = G'_{\bar{Y}}(y) = F'_Y(s)$.

To complete the proof, we consider $N(Y)$. Here, we distinguish two cases depending on whether its single blob B_m whose level is at most k is informative or non-informative.

First, assume that B_m is informative. By Lemma 11, there exists an extension F_Y of f_Y to $V(S(Y))$ such that

$$\mathrm{ch}(F_Y, S(Y)) \leq \mathrm{ch}(F'_Y, S'(Y)) + k \quad \text{and} \quad F_Y(s) = F'_Y(s). \tag{8}$$

Since $F_Y(s) = F'_Y(s) = G_{\bar{Y}}(y)$, the pair $(F_Y, G_{\bar{Y}})$ is a pair of cluster extensions with respect to f. Thus, by Lemma 12, Part (ii), there exists an extension F of f to $V(S)$ such that

$$\mathrm{ch}(F, S) = \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(G_{\bar{Y}}, S(\bar{Y})).$$

Now, using Inequalities (7) and (8), we obtain

$$\begin{aligned}\mathrm{ch}(F, S) &= \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(G_{\bar{Y}}, S(\bar{Y}))\\ &\leq \mathrm{ch}(F'_Y, S'(Y)) + k + \mathrm{ch}(G'_{\bar{Y}}, S'(\bar{Y})) + k \cdot b(g_{\bar{Y}}, N(\bar{Y}), S'(\bar{Y}))\\ &= \mathrm{ch}(F'_Y, S'(Y)) + \mathrm{ch}(G'_{\bar{Y}}, S'(\bar{Y})) + k \cdot (1 + b(g_{\bar{Y}}, N(\bar{Y}), S'(\bar{Y})))\\ &= \mathrm{ch}(F', S') + k \cdot b(f, N, S'),\end{aligned}$$

where the last equality follows from Eq. (6) and the fact that B_m is informative.

Second, assume that B_m is non-informative. Then by Lemma 8, there exists an extension F_Y of f_Y to $V(S(Y))$ such that

$$\mathrm{ch}(F_Y, S(Y)) = \mathrm{ch}(F'_Y, S'(Y)) \quad \text{and} \quad F_Y(s) = F'_Y(s). \tag{9}$$

The remainder of the proof is now similar to the first case. In particular, noting that the pair $(F_Y, G_{\bar{Y}})$ is a pair of cluster extensions with respect to f, by Lemma 12, Part(ii), there exists an extension F of f to $V(S)$ such that

$$\mathrm{ch}(F, S) = \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(G_{\bar{Y}}, S(\bar{Y})).$$

Using Inequalities in (7) and (9), we obtain

$$\begin{aligned}\mathrm{ch}(F, S) &= \mathrm{ch}(F_Y, S(Y)) + \mathrm{ch}(G_{\bar{Y}}, S(\bar{Y}))\\ &\leq \mathrm{ch}(F'_Y, S'(Y)) + \mathrm{ch}(G'_{\bar{Y}}, S'(\bar{Y})) + k \cdot b(g_{\bar{Y}}, N(\bar{Y}), S'(\bar{Y}))\\ &= \mathrm{ch}(F', S') + k \cdot b(f, N, S'),\end{aligned}$$

where the last equality follows from Eq. (6) and the fact that B_m is non-informative.

In both cases, we obtain an extension F of f to $V(S)$ such that

$$\mathrm{ch}(F, S) \leq \mathrm{ch}(F', S') + k \cdot b(f, N, S').$$

This concludes the proof of the lemma. □

Corollary 14 *Let f be a character on X. Let N be a phylogenetic network on X, and let S and S' be embeddings of two phylogenetic X-trees displayed by N such that $PS_{\mathrm{sw}}(f, N) = PS(f, S')$. Then*

$$PS(f, S) \leq PS(f, S') + k \cdot b(f, N, S'),$$

where k is the level of N.

Proof Let F' be an extension of f to $V(S')$ that realises $b(f, N, S')$. By Lemma 13, there exists an extension F of f to $V(S)$ such that $PS(f, S) \leq \mathrm{ch}(F, S) \leq PS(f, S') + k \cdot b(f, N, S')$. □

We are finally in a position to establish the main result of this paper, which we restate for convenience.

Theorem 4. Let N be a phylogenetic network on X, and let T be a phylogenetic X-tree in $D(N)$. Let A be an alignment of characters on X. Then

$$PS(A, T) \leq (k + 1) \cdot PS_{\mathrm{sw}}(A, N) \text{ and } PS(A, T) \leq (k + 1) \cdot PS_{\mathrm{sw}'}(A, N),$$

where k is the level of N.

Proof We will establish that

$$PS(A, T) \leq (k + 1) \cdot PS_{\mathrm{sw}}(A, N).$$

Since $PS_{\mathrm{sw}}(A, N) \leq PS_{\mathrm{sw}'}(A, N)$, it immediately follows that $PS(A, T) \leq (k+1) \cdot PS_{\mathrm{sw}'}(A, N)$ also holds.

First, assume that A consists of a single character f. Let S be an embedding of T in N, and let S' be an embedding of a phylogenetic X-tree in N such that $PS_{\text{sw}}(f, N) = PS(f, S')$. By Corollary 14, we have

$$PS(f, T) = PS(f, S) \le PS(f, S') + k \cdot b(f, N, S'),$$

where the first equality follows from Lemma 2. Now, as every blob B in N that is informative relative to S' contributes at least one to $PS(f, S')$, we have

$$b(f, N, S') \le PS(f, S').$$

By Lemma 2 and combining the last two inequalities, we obtain

$$\begin{aligned} PS(f, T) = PS(f, S) &\le PS(f, S') + k \cdot PS(f, S') \\ &= (k+1) \cdot PS(f, S') \\ &= (k+1) \cdot PS_{\text{sw}}(f, N). \end{aligned} \tag{10}$$

Now, assume that $A = (f_1, \ldots, f_n)$ with $n \ge 1$. Then, we apply Inequality (10) to each character, and obtain

$$\begin{aligned} PS(A, T) &= \sum_{i=1}^{n} PS(f_i, T) \\ &\le \sum_{i=1}^{n} (k+1) \cdot PS_{\text{sw}}(f_i, N) \\ &\le (k+1) \cdot \sum_{i=1}^{n} PS_{\text{sw}}(f_i, N) \\ &= (k+1) \cdot PS_{\text{sw}}(A, N). \end{aligned}$$

□

We close this section by presenting, for each $k \ge 0$, a level-k network and a binary character such that the upper bound stated in Theorem 4 is sharp. As level-0 networks on X are phylogenetic X-trees, the bound is sharp for any level-0 network and any binary character on X. For $k \ge 1$, let N be the level-k network on $X = \{x_0\} \cup \{x_i, x_i', x_i'' : 1 \le i \le k\}$ that is depicted in Fig. 4. Further, let $f\colon X \to \{0, 1\}$ be the binary character with $f(x) = 1$ if $x \in \{x_0, x_1, \ldots, x_k\}$ and $f(x) = 0$ otherwise. Let us consider the two phylogenetic X-trees $T, T' \in D(N)$ that are illustrated on the right-hand side of Fig. 4 together with extensions F and F' of f to $V(T)$ and $V(T')$, respectively. By using Fitch's algorithm it is easy to verify that F and F' are minimum. Hence, we have $PS(f, T') = 1$ and, thus, $PS_{\text{sw}}(f, N) = 1$ (since f employs two character states, $PS_{\text{sw}}(f, N) \ge 1$). Moreover, $PS(f, T) = k + 1$. In summary, we have $PS(f, T) = (k+1) \cdot PS_{\text{sw}}(f, N)$. As the construction shown in Fig. 4 involves a single character, the two notions of softwired parsimony on N coincide, and we have $PS(f, T) = (k+1) \cdot PS_{\text{sw}'}(f, N)$ for the same example.

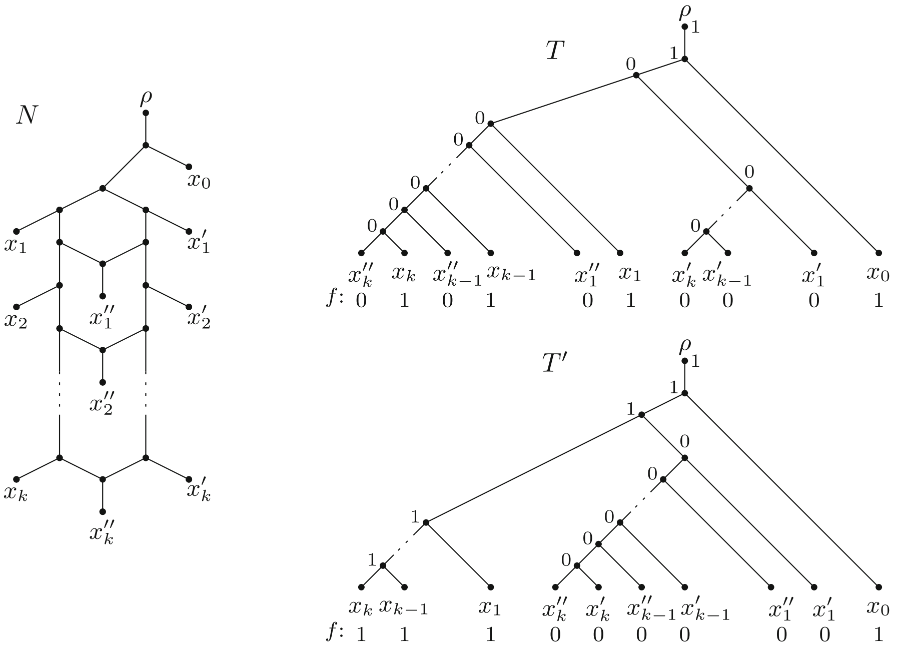
N
ρ
x_0
x_1
x_1'
x_1''
x_2
x_2'
x_2''
x_k
x_k'
x_k''
T
ρ
1
x_k'' x_k x_{k-1}'' x_{k-1} x_1'' x_1 x_k' x_{k-1}' x_1' x_0
f: 0 1 0 1 0 1 0 0 0 1
T′
ρ
1
x_k x_{k-1} x_1 x_k'' x_k' x_{k-1}'' x_{k-1}' x_1'' x_1' x_0
f: 1 1 1 0 0 0 0 0 0 1

Fig. 4 A level-k network N on $X = \{x_0\} \cup \{x_i, x_i', x_i'' : 1 \leq i \leq k\}$ (left) and two phylogenetic X-trees T and T' displayed by N together with a binary character f and extensions F and F' to $V(T)$ and $V(T')$, respectively (right)

7 Parental Parsimony for Phylogenetic Networks

We now briefly consider the notion of *parental parsimony* introduced by van Iersel et al. (2018) as an alternative to softwired and hardwired parsimony. Intuitively, instead of defining the parsimony score of a phylogenetic network N by considering its display set $D(N)$, parental parsimony considers the set of *parental trees* (sometimes also called *weakly displayed trees* (Huber et al. 2016)), which is a superset of $D(N)$.

A *multilabelled tree* on X is a leaf-labelled rooted tree whose root has out-degree one, all other interior vertices have in-degree one and out-degree one, or in-degree one and out-degree two, and, for each element x in X, there exists at least one leaf in T that is labelled x. Now using the same notation as van Iersel et al. (2018), let $U^*(N)$ be the multilabelled obtained from a phylogenetic network N on X as follows: The vertices of $U^*(N)$ are the directed paths in N starting at the root of N, and for each pair of directed paths p, p', there is an edge (p, p') in $U^*(N)$ if and only if p' is an extension of p by one additional edge of N. Furthermore, each vertex in $U^*(N)$ corresponding to a path in N starting at the root of N and ending at $x \in X$ is labelled by x. For an example of the multilabelled tree $U^*(N)$ obtained from a phylogenetic network N, see Fig. 5. Now, a phylogenetic X-tree is called a *parental tree* of N if it can be obtained from this subgraph of $U^*(N)$ by suppressing vertices of in-degree and out-degree one. Notice that we may delete vertices and edges from $U^*(N)$ to obtain a subgraph before suppressing vertices. To denote the set of all parental trees of N,

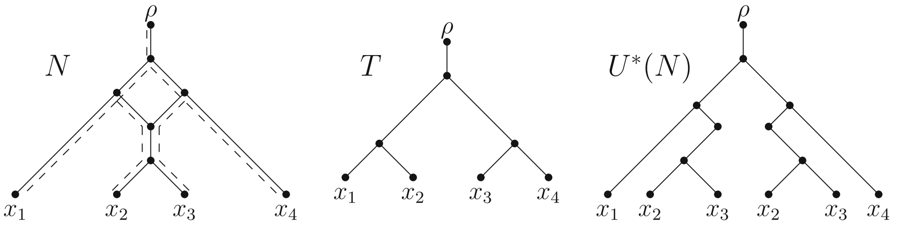

Fig. 5 A phylogenetic network N on $X = \{x_1, x_2, \dots, x_4\}$, a phylogenetic X-tree T such that $T \in P(N)$ and $T \notin D(N)$, and the multilabelled tree $U^*(N)$. The dashed lines in N indicate how T can be drawn inside N

we use $P(N)$. Informally speaking, a tree is a parental tree of a phylogenetic network if it can be drawn inside the network in such a way that the tree vertices of the tree correspond to tree vertices of the network. Importantly, though, a parental tree is not necessarily a displayed tree (see Fig. 5), whereas every displayed tree is also parental.

Given a character f on X and a phylogenetic network N on X, the *parental parsimony score* of f on N is now defined as

$$PS_{\text{pa}}(f, N) = \min_{T \in P(N)} PS(f, T),$$

where the minimum is taken over all parental trees for N.

It was shown by (van Iersel et al. 2018, Theorem 2) that computing the parental parsimony score is NP-hard even if f is a binary character and N is a restricted type of a so-called tree-child network (Cardona et al. 2009). It is thus a natural question if our main result for softwired parsimony (Theorem 4) generalises to parental parsimony. Unfortunately, this is not the case. Suppose that n is an even integer and that N is the level-1 network on $n + 1$ leaves depicted in Fig. 6. Then, both phylogenetic trees T and T' as depicted in the same figure are parental trees of N and $T' \notin D(N)$. Additionally, suppose that f is the binary character that assigns state 0 to leaves $x_2, x_4, x_6, \dots, x_n$, and state 1 to leaves $x_1, x_3, \dots, x_{n+1}$. Crucially, we have $PS(f, T) = n/2$ and $PS(f, T') = 1$. In particular, $PS_{\text{pa}}(f, N) = 1$, and thus,

$$PS(f, T) = n/2 \nleq 2 \cdot PS_{\text{pa}}(f, N) = 2$$

for each $n > 4$ (a similar argument applies to n being odd), which shows that Theorem 4 does not generalise to parental parsimony even if the phylogenetic network is level-1 with a single blob and the alignment consists of a single binary character.

8 Softwired Parsimony for Semi-directed and Unrooted Networks

In this section, we show that binary semi-directed networks have an analogous bound on the softwired parsimony score as the one we established for rooted binary phylogenetic networks. We briefly turn our attention to unrooted binary phylogenetic networks at the end of the section and show that the approach we take to establish the bound for

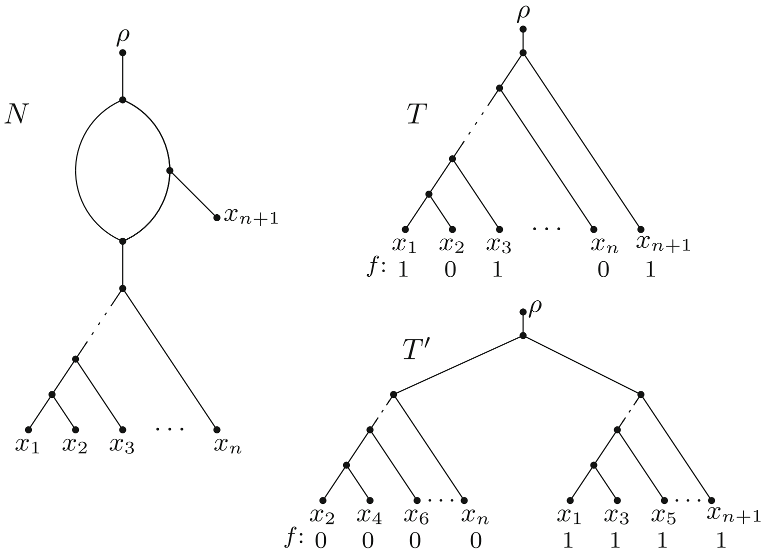
ρ
N
x_{n+1}
x_1 x_2 x_3 $\dots$ x_n
ρ
T
x_1 x_2 x_3 $\dots$ x_n x_{n+1}
f: 1 0 1 0 1
ρ
T'
x_2 x_4 x_6 $\dots$ x_n
f: 0 0 0 0
x_1 x_3 x_5 $\dots$ x_{n+1}
1 1 1 1

Fig. 6 A level-1 network N on $X = \{x_1, x_2, \dots, x_{n+1}\}$ with n being even and two parental trees T and T' of N together with a binary character f on X. Here, $PS_{\mathrm{pa}}(N, f) = PS(f, T') = 1$, whereas $PS(f, T) = n/2$

semi-directed networks does not work in this setting. However, this does not exclude the possibility that similar bounds can be obtained for unrooted phylogenetic networks by other means.

A *binary semi-directed network* N_s on X is a leaf-labelled mixed multigraph without any loops that can be obtained from a rooted binary phylogenetic network N_r by deleting its root, suppressing the child of the root, and omitting the direction of each edge that is not a reticulation edge. We call N_r a *rooted partner* of N_s. By construction, N_s has at most one pair of parallel edges. This is precisely the case when N_r has an underlying 3-cycle that contains the child of the root. Note that N_s may have multiple rooted partners. A vertex v in N_s is called a *reticulation* if N_s contains two edges, referred to as *reticulation edges*, that are directed into v. A *semi-directed level-k network* is a semi-directed network N_s such that a rooted partner of N_s is a rooted binary level-k network. Lastly, an *unrooted binary phylogenetic X-tree T_u* is a connected undirected acyclic graph whose leaf set is X and whose inner vertices all have degree three. Note that an unrooted binary phylogenetic tree is a binary semi-directed network without reticulations. In the following, all rooted (resp. semi-directed) phylogenetic networks and rooted (resp. unrooted) phylogenetic trees are assumed to be binary.

Now, let T_u be an unrooted binary phylogenetic X-tree, and let N_s be a semi-directed network on X. We say that T_u is *displayed* by N_s if there exists a subgraph S of N_s such that S is a subdivision of T_u (omitting the directions of the reticulation edges) and S contains, for each reticulation v in N_s, at most one reticulation edge incident with v. Similar to rooted phylogenetic networks, if T_u is displayed by N_s, we call S an *embedding* of T_u in N_s. Furthermore, we refer to the set of all unrooted phylogenetic

X-trees that are displayed by N_s as the *unrooted display set* of N_s and denote it by $D(N_s)$.

Let $A = (f_1, f_2, \ldots, f_n)$ be an alignment of characters on X, and let N_s be a semi-directed network on X. We define the *softwired parsimony score* of A on N_s as

$$PS_{\text{sw}}(A, N_s) = \sum_{i=1}^{n} \min_{T_u \in D(N_s)} PS(f_i, T_u), \tag{11}$$

where the parsimony score of an unrooted phylogenetic X-tree T is defined as in the rooted case since the corresponding concepts of the changing number and (minimum) extensions naturally translate to undirected graphs.

The next lemma shows how the set of unrooted phylogenetic trees that are displayed by a semi-directed network N_s is related to the set of rooted phylogenetic trees that are displayed by a rooted partner of N_s.

Lemma 15 *Let N_s be a semi-directed network on X, and let N_r be a rooted partner of N_s. Then, the following two statements hold.*

(i) If T_u is an unrooted phylogenetic X-tree that is displayed by N_s, then there exists a rooted phylogenetic X-tree $T \in D(N_r)$ such that T_u can be obtained from T by deleting the root and suppressing its child.

(ii) If T is a rooted phylogenetic X-tree that is displayed by N_r, then there exists an unrooted phylogenetic X-tree $T_u \in D(N_s)$ such that T_u can be obtained from T by deleting the root and suppressing its child.

Proof Throughout the proof, we assume that N_s does not contain a pair of parallel edges. Indeed, if N_s contains such a pair of edges, e_1 and e_2 say, we can obtain a semi-directed network N_s' on X from N_s, by deleting e_1, suppressing the two resulting degree-two vertices, and omitting the direction of e_2. Clearly $D(N_s) = D(N_s')$. Now, let v be the child of the root ρ in N_r, and let w and w' be the children of v. If either w or w' is a reticulation, we assume without loss of generality that w' is a reticulation. Observe that at most one of w and w' is a reticulation. Each edge in N_r that is not incident with v corresponds to exactly one edge in N_s, and each edge in N_s except $e_\rho = \{w, w'\}$ (resp. $e_\rho = (w, w')$ if w' is a reticulation in N_r) corresponds to exactly one edge in N_r. In particular, each reticulation edge in N_r corresponds to exactly one such edge in N_s and vice versa. First, let T_u be an unrooted phylogenetic X-tree that is displayed by N_s. By definition, there exists an embedding S_u of T_u in N_s. If e_ρ is an edge in S_u, we obtain an embedding S of a rooted phylogenetic X-tree T in N_r from S_u as follows. We replace e_ρ with the three directed edges (ρ, v), (v, w) and (v, w') and each edge $e \neq e_\rho$ with its directed counterpart in N_r. By definition, T is displayed by N_r, that is $T \in D(N_r)$. Further, by construction, T_u can be obtained from T by deleting the root and suppressing its child. Now, let us consider the case that e_ρ is not an edge in S_u. Let S' be the connected acyclic subgraph of N_r that is obtained from S_u by replacing each edge in S_u with its directed counterpart in N_r. We next show that there is a unique vertex v_s in S' with in-degree zero and out-degree two. Since S' is acyclic, it follows that S' contains a vertex with in-degree zero. Clearly, the out-degree of this vertex cannot be one or three and, thus, v_s exists. Assume towards a contradiction that $v_s' \neq v_s$ is another vertex with this property. As S' is connected and does not contain a vertex with in-degree two, one of v_s' and v_s is a descendant of the other in S'. But then one of these vertices has in-degree one, a contradiction. It now follows that each vertex in S' is a descendant of v_s. Let $\pi = (\rho = v_1, v_2, \ldots, v_t = v_s)$ be a directed path in N_r. Such a path π exists as there is a directed path from the root to any vertex in N_r. Since each vertex in S' is a descendant of v_s, it follows that v_i with $i < t$ is not in S'. Then, we obtain an embedding S of a rooted phylogenetic X-tree T in N_r from S' by adding the $t-1$ directed edges (v_i, v_{i+1}), $1 \leq i < t$, to S'. Again, we have $T \in D(N_r)$ and, by construction, T_u can be obtained from T by deleting the root and suppressing its child. Hence, (i) holds.

Second, let T be a rooted phylogenetic X-tree that is displayed by N_r. Let S be an embedding of T in N_r, and let $\pi = (\rho = v_1, v_2, \ldots, v_t)$ be the root path of S. Then, we obtain an embedding S_u of an unrooted phylogenetic X-tree T_u in N_s from S by deleting each vertex that lies on π except for v_t, turning directed edges into undirected ones and, if $t = 2$, suppressing v_t. If $t = 2$, then π consists of the single edge (ρ, v), and S contains (v, w) and (v, w'). Hence, S_u contains the edge $\{w, w'\}$ as a result of suppressing v. Otherwise, if $t \geq 3$, then S contains (ρ, v) and exactly one of (v, w) and (v, w'), and v_t is a vertex in N_s. Furthermore, if $t \geq 4$, each edge (v_i, v_{i+1}) with $3 \leq i < t$ that is traversed by π corresponds to a unique edge consisting of the same vertices in N_s. It now follows that S_u is indeed an embedding of T_u in N_s, that is, $T_u \in D(N_s)$. By construction, T_u can be obtained from T by deleting the root and suppressing its child. Hence, (ii) holds as well. □

We next obtain the following corollary from Lemma 15.

Corollary 16 *Let N_s be a semi-directed network on X, let N_r be a rooted partner of N_s, and let A be an alignment of characters on X. Then $PS_{\text{sw}}(A, N_s) = PS_{\text{sw}}(A, N_r)$.*

Proof The corollary follows from Lemma 15 and the fact that, if T is a rooted phylogenetic X-tree and T_u is an unrooted phylogenetic X-tree such that T_u can be obtained from T by deleting the root and suppressing its child, then $PS(A, T) = PS(A, T_u)$. □

We are now in a position to state the main result of the section.

Theorem 17 *Let N_s be a semi-directed network on X, and let T_u be an unrooted phylogenetic X-tree that is displayed by N_s. Furthermore, let A be an alignment of characters on X. Then*

$$PS(A, T_u) \leq (k+1) \cdot PS_{\text{sw}}(A, N_s),$$

where k is the level of N_s.

Proof Let N_r be a rooted partner of N_s, and let $T \in D(N_r)$ be a rooted phylogenetic X-tree such that deleting the root in T and suppressing its child yields T_u. The rooted phylogenetic X-tree T exists by Lemma 15 (i). Since an unrooted phylogenetic tree is a semi-directed network without reticulations, we have (i) $PS(A, T_u) = PS(A, T)$ by Corollary 16. Furthermore, by Theorem 4, we have (ii) $PS(A, T) \leq (k+1) \cdot$

$PS_{\text{sw}}(A, N_r)$. Finally, by Corollary 16, we have (iii) $PS_{\text{sw}}(A, N_r) = PS_{\text{sw}}(A, N_s)$. Combining (i)–(iii) yields $PS(A, T_u) \leq (k+1) \cdot PS_{\text{sw}}(A, N_s)$. □

We now shift our focus from semi-directed networks to *unrooted* phylogenetic networks. An *unrooted binary phylogenetic network* U on X is a connected undirected graph without any loops or edges in parallel, whose leaf set is X, and whose inner vertices all have degree three. As before, we omit the term binary in the following as all unrooted phylogenetic networks considered in this section are binary.

Let U be an unrooted phylogenetic network on X. We say that an unrooted phylogenetic X-tree T_u is *displayed* by U if there exists a subgraph of U that is a subdivision of T_u. Furthermore, we refer to the set of all unrooted phylogenetic X-trees that are displayed by U as the *display set* of U and denote it by $D(U)$. We call U an *unrooted level-k network* if at most k edges have to be deleted in each biconnected component of U such that the resulting graph is acyclic. Lastly, if U can be obtained from a rooted phylogenetic network N by deleting its root, suppressing the child of the root, and omitting all edge directions, we say that N is an *orientation* of U.

Let $A = (f_1, f_2, \ldots, f_n)$ be an alignment of characters on X, and let U be an unrooted phylogenetic network on X. We define the *softwired parsimony score* of A on U as

$$PS_{\text{sw}}(A, U) = \sum_{i=1}^{n} \min_{T \in D(U)} PS(f_i, T). \tag{12}$$

Next, we present an unrooted level-1 network U and a binary character f such that $PS_{\text{sw}}(f, U) \neq PS_{\text{sw}}(f, N)$ for an orientation N of U, that is, we show that Corollary 16 does not translate from semi-directed networks to unrooted phylogenetic networks. Moreover, we give two different orientations N and N' of U with $PS_{\text{sw}}(f, N) \neq PS_{\text{sw}}(f, N')$.

To this end, let U be the unrooted level-1 network on $X = \{a, b, c, d, e\}$, and let N and N' be the two orientations of U as shown in Fig. 7. The display set $D(U)$ of size five is shown in the middle of the same figure. By deleting the root and suppressing its child in each element of $D(N)$ (resp. $D(N')$), we obtain a subset of $D(U)$ as indicated by the two dashed rectangles that each enclose N (resp. N') and two elements of $D(U)$. Now for the single binary character $f : X \to \{0, 1\}$ with

$$f(a) = f(b) = f(c) = 0 \text{ and } f(d) = f(e) = 1,$$

we have $PS_{\text{sw}}(f, U) = 1$, $PS_{\text{sw}}(f, N) = 2$ and $PS_{\text{sw}}(f, N') = 1$. Since the softwired parsimony score of an unrooted phylogenetic network U is not necessarily the same as the score of an orientation of U, we cannot represent an unrooted phylogenetic network by an arbitrary orientation in the way we used a rooted partner of a semi-directed network to obtain Theorem 17. While our example shows that using the same approach as in the semi-directed setting is not viable, it does not exclude the existence of similar bounds for unrooted phylogenetic networks or classes thereof (such as unrooted level-1 networks, for example).

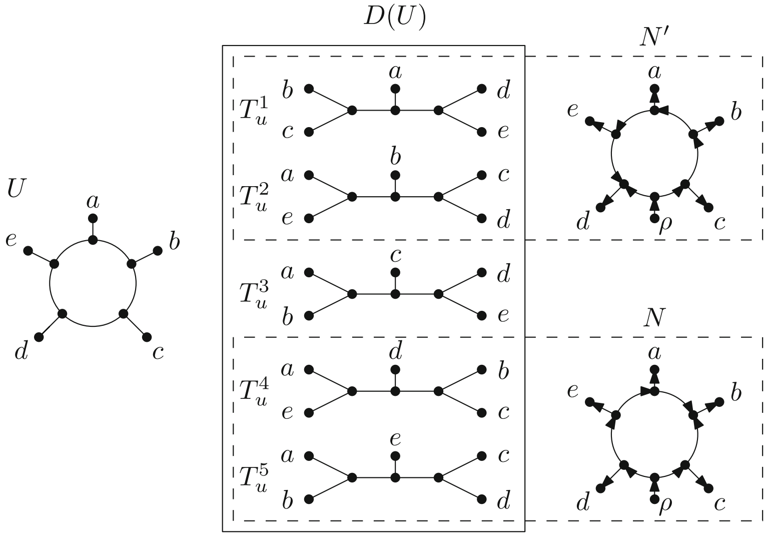

Fig. 7 An unrooted level-1 network U (left) with its display set $D(U)$ (middle), and two orientations N and N' of U (right) with their respective display sets indicated by a dashed rectangle. More precisely, $D(N)$ and $D(N')$ can be obtained by orienting the enclosed unrooted phylogenetic X-trees

9 Concluding Remarks

In this paper, we have obtained a bound on the softwired parsimony score of a gap-free alignment of multistate characters on rooted as well as semi-directed phylogenetic level-k networks. To be precise, we have shown that the maximum difference between the softwired parsimony score of a phylogenetic network N and the parsimony score of any tree displayed by N is bounded by $k + 1$ times the parsimony score of N. Unfortunately, our approximation result as stated in Theorem 4 cannot be generalised to alignments with gaps since it was already shown in Corollary 2 of Kelk et al. (2019) that computing the softwired parsimony score of a level-1 network for an alignment of binary characters that additionally allows gaps is APX-hard.

Extending the notion of softwired parsimony to semi-directed networks and exploiting a connection between the display sets of semi-directed networks and their rooted partners, we have shown that an analogous bound holds for semi-directed networks. For unrooted networks, on the other hand, the approach via rooted partners (more formally, via orientations) does not seem to be viable. Nevertheless, it would be an interesting question for future research to investigate if an analogous or similar bound for the softwired parsimony score can be obtained in some other way for unrooted phylogenetic networks.

Another interesting direction for future research would be to analyse whether our results also apply in the case of *non-binary* phylogenetic networks, i.e., phylogenetic networks that may have vertices of degree strictly greater than three. While there exists a polynomial time algorithm to compute the parsimony score of a given non-binary phylogenetic tree with character states assigned to its leaves, namely the Fitch-Hartigan algorithm (Fitch 1971; Hartigan 1973), other concepts that our results rely on, such as the rSPR distance and its relation to the switching distance, have been studied less for non-binary phylogenetic trees (see Sect. 2 of Fischer and Kelk (2016) for a related discussion).

Acknowledgements The first and second authors thank the New Zealand Marsden Fund for financial support. All authors thank Steven Kelk for helpful discussions. Moreover, we thank two anonymous reviewers for detailed comments on an earlier version of this paper.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability There is no data associated with this publication.

References

Allman ES, Baños H, Rhodes JA (2019) NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms Mol Biol 14:1–25

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

Döcker J, Linz S, Semple C (2024) Hypercubes and Hamilton cycles of display sets of rooted phylogenetic networks. Adv Appl Math 152:102595

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, New York

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Biol 20(4):406–416

Fischer M, Kelk S (2016) On the maximum parsimony distance between phylogenetic trees. Ann Comb 20:87–113

Frohn M, Kelk S (2023) A 2-approximation algorithm for the softwired parsimony problem on binary, tree-child phylogenetic networks (submitted)

Fischer M, Iersel L, Kelk S, Scornavacca C (2015) On computing the maximum parsimony score of a phylogenetic network. SIAM J Discrete Math 29(1):559–585

Gross E, Long C (2018) Distinguishing phylogenetic networks. SIAM J Appl Algebra Geom 2(1):72–93

Hartigan JA (1973) Minimum mutation fits to a given tree. Biometrics 29(1):53

Huber KT, Moulton V, Steel M, Wu T (2016) Folding and unfolding phylogenetic trees and networks. J Math Biol 73(6–7):1761–1780

Hollering B, Sullivant S (2021) Identifiability in phylogenetics using algebraic matroids. J Symb Comput 104:142–158

Kelk S, Fischer M (2017) On the complexity of computing MP distance between binary phylogenetic trees. Ann Comb 21:573–604

Kelk S, Pardi F, Scornavacca C, Iersel L (2019) Finding a most parsimonious or likely tree in a network with respect to an alignment. J Math Biol 78:527–547

Kannan L, Wheeler WC (2012) Maximum parsimony on phylogenetic networks. Algorithms Mol Biol 7:1–10

Nakhleh L, Jin G, Zhao F, Mellor-Crummey J (2005) Reconstructing phylogenetic networks using maximum parsimony. In: 2005 IEEE computational systems bioinformatics conference (CSB'05). IEEE, pp 93–102

Schrago CG, Aguiar BO, Mello B (2018) Comparative evaluation of maximum parsimony and Bayesian phylogenetic reconstruction using empirical morphological data. J Evol Biol 31(10):1477–1484

Sansom RS, Choate PG, Keating JN, Randle E (2018) Parsimony, not Bayesian analysis, recovers more stratigraphically congruent phylogenetic trees. Biol Let 14(6):20180263

Semple C (2007) Hybridization networks. In: Gascuel O, Steel M (eds) Reconstructing evolution: new mathematical and computational advances. Oxford University Press, Oxford, pp 277–314

Solís-Lemus C, Ané C (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet 12(3):1005896

Solís-Lemus C, Bastide P, Ané C (2017) PhyloNetworks: a package for phylogenetic networks. Mol Biol Evol 34(12):3292–3298

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21(4):456–463

Smith MR (2019) Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. Biol Lett 15(2):20180632

van Iersel L, Jones M, Scornavacca C (2018) Improved maximum parsimony models for phylogenetic networks. Syst Biol 67(3):518–542

Zhang C, Huelsenbeck JP, Ronquist F (2020) Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. Syst Biol 69(5):1016–1032

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