Discrete Optimization 36 (2020) 100570

Contents lists available at ScienceDirect

Discrete Optimization

www.elsevier.com/locate/disopt

# On the Balanced Minimum Evolution polytope

Daniele Catanzaro [a,b,*], Raffaele Pesenti [c], Laurence Wolsey [a]

[a] *Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium*
[b] *Luxembourg Institute of Socio-Economic Research (LISER), 11 Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg*
[c] *Department of Management, Università Ca' Foscari, San Giobbe, Cannaregio 837, I-30121 Venezia, Italy*

ARTICLE INFO

*Article history:*
Received 19 March 2019
Received in revised form 2 February 2020
Accepted 3 February 2020
Available online xxxx

*Keywords:*
Balanced Minimum Evolution
Polyhedral combinatorics
Facet-defining inequalities
Manifold of unrooted binary trees
Kraft's equality
Enumeration of trees

ABSTRACT

Recent advances on the polyhedral combinatorics of the *Balanced Minimum Evolution Problem* (BMEP) enabled the characterization of a number of facets of its convex hull (also referred to as the *BMEP polytope*) as well as the discovery of connections between this polytope and the permutoassociahedron. In this article, we extend these studies, by presenting new results concerning some fundamental characteristics of the BMEP polytope, new facet-defining inequalities in the case of six or more taxa, a number of valid inequalities, and a polynomial time oracle to recognize its vertices. Our aim is to broaden understanding of the polyhedral combinatorics of the BMEP with a view to developing new and more effective exact solution algorithms.

© 2020 Elsevier B.V. All rights reserved.

## 1. Introduction

Consider a set $\Gamma = \{1, 2, \ldots, n\}$ of $n \geq 4$ vertices, hereafter referred to as *taxa*. A *phylogeny* $T$ of $\Gamma$ is an *Unrooted Binary Tree* (UBT) having $\Gamma$ as leaf-set. By definition, a phylogeny of $\Gamma$ has $(n-2)$ *internal vertices* having degree 3 and an overall number of $(2n-3)$ edges, $n$ of which are *external*, i.e., incident to a taxon, and $(n-3)$ are *internal*, i.e., incident to two internal vertices. The labeling of the internal vertices of a phylogeny is usually less important than the labeling of the leaves, hence it is generally omitted. However, whenever the context requires us to specify a particular vertex under study, we will use the convention to label it as an integer in $\{n+1, n+2, \ldots, 2n-2\}$. Fig. 1(a) shows a possible phylogeny of a set $\Gamma$ of seven taxa, shaped as a caterpillar. Fig. 1(b) shows a phylogeny with a more complex topology together with a terminology of its components that we will formally define in the next section. In Fig. 1(a) the internal

* Corresponding author at: Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium.
*E-mail addresses:* daniele.catanzaro@uclouvain.be (D. Catanzaro), pesenti@unive.it (R. Pesenti), laurence.wolsey@uclouvain.be (L. Wolsey).

https://doi.org/10.1016/j.disopt.2020.100570
1572-5286/© 2020 Elsevier B.V. All rights reserved.

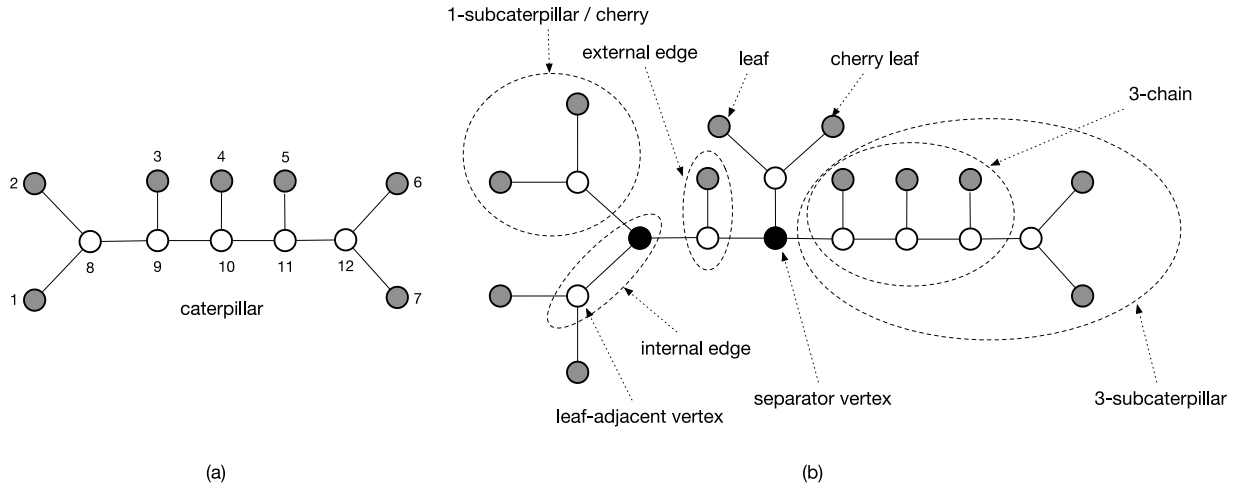



**Fig. 1.** (a) An example of a phylogeny (in particular, a *caterpillar*) of the set $\Gamma = \{1, 2, 3, 4, 5, 6, 7\}$; (b) An example of a phylogeny with a more complex topology; leaves are colored gray, internal vertices are colored black or white; the subfigure also shows a terminology of the components of a phylogeny that will prove useful throughout the article.

vertices are labeled as indicated above. On the other hand, in Fig. 1(b), the labeling of the vertices of this phylogeny is omitted to better highlight the terminology.

Let $\mathcal{T}$ be the set of the possible distinct phylogenies of $\Gamma$. The cardinality of $\mathcal{T}$ is known to be a function of the cardinality of $\Gamma$ and can be easily shown to be equal to $(2n-5)!! = 1 \times 3 \times 5 \cdots \times (2n-5)$ [1]. Given a phylogeny $T$ of $\Gamma$ and a taxon $i \in \Gamma$, let $\Gamma_i$ be the set $\Gamma \setminus \{i\}$ and let $\tau_{ij}$ be the *topological distance* between taxa $i, j \in \Gamma$, $i \neq j$, i.e., the number of edges belonging to the (unique) path from taxon $i$ to taxon $j$ in $T$. For example, by referring to the phylogeny shown in Fig. 1(a), $\tau_{13} = 3$, $\tau_{15} = 5$, and $\tau_{35} = 4$.

Let $\mathbf{D}$ be a given $n \times n$ symmetric distance matrix, whose generic entry $d_{ij} \in \mathbb{R}_{0+}$ represents a measure of dissimilarity between the corresponding pair of taxa $i, j \in \Gamma$. Then, the *Balanced Minimum Evolution Problem* (BMEP) is to find a phylogeny $T$ of $\Gamma$ that minimizes the following *length function* [2,3]:

$$L(T) = \sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \frac{d_{ij}}{2^{\tau_{ij}}}. \tag{1}$$

The BMEP is a version of the *network design problem* [4] defined over UBTs. It was introduced in the literature on phylogenetics in 2000 by Pauplin [5] and systematically investigated from a biological perspective in Desper and Gascuel [6] and Gascuel and Steel [7]. The problem is $\mathcal{NP}$-hard and inapproximable within $c^n$, for some constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$ [8]. This fact has justified the development of a number of implicit enumeration algorithms to exactly solve the problem [3,9,10] as well as a number of heuristics to approximate its optimal solution [10–12].

The current state-of-the-art exact solution algorithm for the BMEP, described in Catanzaro et al. [3], is based on an Integer Linear Programming (ILP) model that exploits some combinatorial connections between the BMEP and the *Huffman Coding Problem* [13–15]. Such connections proved fundamental to identify a number of properties that characterize phylogenies, such as the *Kraft equalities* or the *Third Equality* (see Section 2 and [3]), and had a central role in the study of tight lower bounds on the optimal solution to the problem. The algorithm is however unable to solve instances of the BMEP containing more than two dozen taxa within one hour computing time. Thus, in the attempt to develop algorithms able to exactly solve increasingly large instances of the problem, particular attention has been recently devoted to the study of its polyhedral combinatorics. The earliest attempts were the works Eickmeyer et al. [16] and Haws et al. [17]. Eickmeyer et al. [16] introduced the convex hull of the BMEP (hereafter also referred to as the

*BMEP polytope*) and characterized its vertices. Haws et al. [17] characterized instead some faces of the polytope in the space of the topological distances. Neither studies led to results directly applicable in implicit enumeration approaches. Moreover, due to the nonlinearity of some specific equations that characterize phylogenies (see Section 2), it is difficult to carry out the analysis of the polytope in the chosen space. Hence, Forcey et al. [18] investigated alternative spaces to perform such analysis and proposed the use of Polymake [19] to assist in this task. Their empirical approach proved successful in unveiling connections between the BMEP polytope and the *permutoassociahedron* [20–23] as well as in characterizing the BMEP polytope in low and fixed dimensions. In particular, they succeeded in characterizing all of its facets for five taxa and some of the facets for six taxa. Unfortunately, the very large number of facets to analyze in the latter case (over 90 000, see Section 6.2) prevented the complete description of the polytope.

In this article, we extend the results of Eickmeyer et al. [16] and Forcey et al. [18,21], by presenting both alternative and novel proofs concerning the polyhedral combinatorics of the BMEP as well as algorithms that may inspire the development of new exact solution approaches to the problem based on implicit enumeration. Some alternative proofs (e.g., the characterization of the dimension of the BMEP polytope) are interesting per se because they enable the analysis of aspects not trivially deducible from previous studies [16,18,21]. We also characterize new facets of the BMEP polytope for six or more taxa as well as valid inequalities. We address the problem of consistently generating the vertices of the BMEP polytope as well as the problem of recognizing its vertices. We present a set of necessary and sufficient conditions to address the vertex generation problem and we show how to translate such conditions into a set of nonlinear constraints in order to develop new exact solution approaches to the problem similar to those already proposed in [3] and [10]. We will show that these conditions also suggest a possible polynomial-time oracle to solve the recognition problem as well. The article is organized as follows. In Section 2 we introduce some notation, definitions, and fundamental properties of the topological distances that will prove useful throughout the article. In Section 3, we review Forcey et al. [18] space. In Section 4, we show a possible basis for the space of Forcey et al. [18]. In Sections 5 and 6 we discuss some fundamental characteristics of the BMEP polytope and we extend the analysis of Eickmeyer et al. [16] and Forcey et al. [18]. Finally, in Section 7, we address the problem of recognizing and consistently generating the vertices of the BMEP polytope.

## 2. Nomenclature of phylogenetics and properties of the topological distances

In this section, we introduce some specific notation and nomenclature of phylogenetics that will prove useful throughout the article. We will give particular emphasis to the concept and the properties of a "path-length sequence" of a phylogeny because it has a central role in the combinatorics of the BMEP.

By referring to the phylogeny shown in Fig. 1(b), we call

- a *leaf adjacent vertex* any internal vertex of a phylogeny that is adjacent to a leaf;
- a *separator vertex* any internal vertex of a phylogeny that is not leaf adjacent;
- a *cherry* any connected subtree of a phylogeny induced by two leaves and their common leaf adjacent vertex;
- a *cherry leaf* one of the two leaves of a cherry;
- a *caterpillar* any phylogeny whose subgraph induced by its internal vertices is a path-graph;
- a *d-subcaterpillar* any connected subtree of a phylogeny $T$ induced by $d \geq 1$ leaf adjacent vertices and their leaves, which is connected to the rest of $T$ through a single bridge whose extreme vertex is both a separator vertex and is not in the $d$-subcaterpillar. As an example, Fig. 1(b) includes both a 1-subcaterpillar and a 3-subcaterpillar.
- a *d-chain* any connected subtree of a phylogeny $T$ induced by $d \geq 2$ leaf adjacent vertices whose subgraph induced by its internal vertices (i) is a path-graph and (ii) is connected to the rest of $T$ through two bridges. As an example, Fig. 1(b) includes a 3-chain.

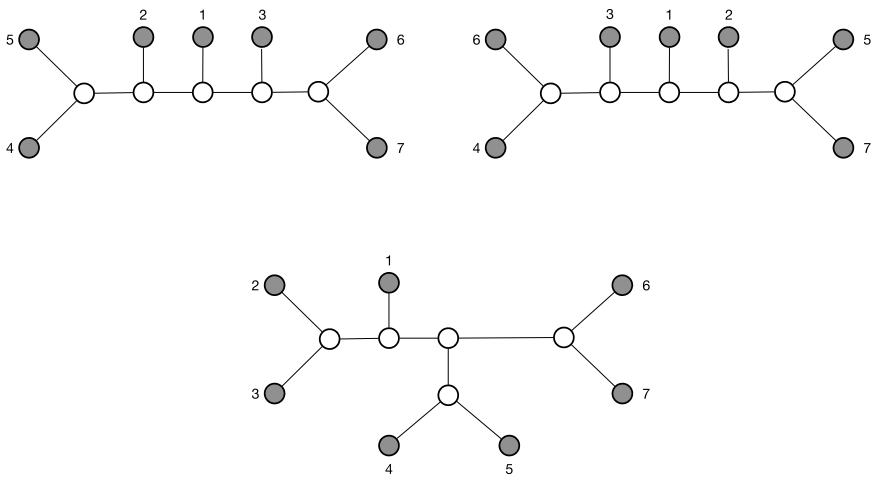



**Fig. 2.** An example of the three distinct phylogenies that can be reconstructed from the path-length sequence $\tau_1 = [3, 3, 4, 4, 4, 4]$.

Moreover, we refer to a phylogeny having three cherries as a *1-branch caterpillar* (see e.g., the third phylogeny in Fig. 2).

Given two integers $\alpha$ and $\beta$, with $\alpha < \beta$, let $[\alpha, \beta]$ denote the discrete interval constituted by the integers included between $\alpha$ and $\beta$. For a nonempty subset $S \subseteq \mathbb{R}^n$, we denote by $\text{Lin}(S)$ and $\text{Aff}(S)$, the linear and affine hulls of the element of $S$, respectively. Given a rectangular matrix $A$ having generic entry $a_{ij} \in \mathbb{R}$, we define

$$A_{(r_1,c_1),(r_2,c_2)} = \{a_{i,j} : i \in [r_1, r_2],\ j \in [c_1, c_2]\}.$$

In other words, $A_{(r_1,c_1),(r_2,c_2)}$ is the sub-matrix obtained from $A$ by considering the consecutive rows between $r_1$ and $r_2$ (included) and the consecutive columns between $c_1$ and $c_2$ (included).

Similarly to Parker and Ram [14], we define a *sequence* as a collection of nonnegative real values such as $\mathbf{s} = [s_1, s_2, \ldots, s_m]$, $s_j \in \mathbb{R}_{0+}$. Repetition of values in the sequence is permitted: the values $s_j$ need not to be distinct.

Consider a phylogeny $T$ of a set $\Gamma$ of $n$ taxa and a taxon $i \in \Gamma$. We define the *path-length sequence* $\tau_i = [\tau_{ij} \in [2, n-1] : j \in \Gamma_i]$ as the sequence of the topological distances relative to the $n-1$ (unique) paths in $T$ from taxon $i$ to each taxon $j \in \Gamma_i$. For example, consider the phylogeny showed in Fig. 1(a). Then, $\tau_1 = [2, 3, 4, 5, 6, 6]$ and $\tau_4 = [4, 4, 3, 3, 4, 4]$. It is worth noting that the path-length sequence $\tau_i$ associated to a phylogeny $T$ of $\Gamma$ describes the UBT $T$ from the "perspective" of taxon $i \in \Gamma$. Hence, with an abuse of nomenclature, we will say that $\tau_i$ describes the phylogeny $T$ *rooted in taxon* $i$. Fixing a taxon $i \in \Gamma$, we denote $\Theta_i$ as the set of the path-length sequences $\tau_i$ encoding the phylogenies of $\Gamma$ rooted in $i$.

The following extension of the concept of a path-length sequence proves particularly useful to model the BMEP: we define the *path-length matrix* $\tau$ associated to a phylogeny $T$ of $\Gamma$ as a $n \times n$ integer matrix having as generic entry $\tau_{ij}$, for all $i, j \in \Gamma$. For example, the following path-length matrix is associated to the phylogeny shown in Fig. 1(a) under the assumption that the relative taxa are ordered according to their

labels:

$$\tau = \begin{pmatrix} 0 & 2 & 3 & 4 & 5 & 6 & 6 \\ 2 & 0 & 3 & 4 & 5 & 6 & 6 \\ 3 & 3 & 0 & 3 & 4 & 5 & 5 \\ 4 & 4 & 3 & 0 & 3 & 4 & 4 \\ 5 & 5 & 4 & 3 & 0 & 3 & 3 \\ 6 & 6 & 5 & 4 & 3 & 0 & 2 \\ 6 & 6 & 5 & 4 & 3 & 2 & 0 \end{pmatrix}.$$

Observe that, apart from the diagonal entries, each row (or each column) of $\tau$ is a path-length sequence of the considered phylogeny, rooted in a taxon $i \in \Gamma$. In particular, the first row refers to $\tau_1$, the second row to $\tau_2$, and so on. In analogy to $\Theta_i$, we denote $\Theta$ as the set of the path-length matrices $\tau$ associated to all the possible phylogenies of $\Gamma$. A question that naturally arises and that proves useful to study the polyhedral combinatorics of the BMEP is whether it is possible to characterize $\Theta$ and the sets $\Theta_i$, for all $i \in \Gamma$. The following two sections address this issue.

### 2.1. Characterizing $\Theta_i$

As observed in Catanzaro et al. [3], a characterization of the set $\Theta_i$, for a fixed $i \in \Gamma$, can be obtained from the analogies between phylogenies and *Huffman trees* [14]. Specifically, Huffman trees are rooted binary trees used in coding theory to represent symbols belonging to a given alphabet $\Psi$. The leaves of a Huffman tree correspond to the symbols in $\Psi$ and the tree itself can be described by means of a path-length sequence $\rho = [\rho_j : j \in \Psi]$ whose generic entry $\rho_j$ represents the topological distance of the shortest path from the root of the tree to the symbol $j \in \Psi$. In this context, the following well-known necessary and sufficient condition relates rooted binary trees and path-length sequences:

**Proposition 1** (*Kraft Equality [14]*)**.** *Consider a set $\Psi$ of $n$ symbols. Then, $\rho = [\rho_j, \rho_2, \dots, \rho_n]$ is the sequence of topological distances of a rooted binary tree having $\Psi$ as leafset if and only if*

$$\sum_{j \in \Psi} 2^{-\rho_j} = 1. \tag{2}$$

Interestingly, Proposition 1 can be used to provide a characterization of the set $\Theta_i$. Specifically, consider a phylogeny $T$ of $\Gamma$ and a taxon $i \in \Gamma$. Denote $\hat{i}$ as the "father" of taxon $i$, i.e., as the only internal vertex adjacent to $i$ in $T$. For example, by referring to the phylogeny shown in Fig. 1, if $i = 1$ then $\hat{i} = 6$. We observe that if we disregard the edge $(i, \hat{i})$ then the remaining tree can be seen as a Huffman tree rooted in $\hat{i}$ and coding the symbols in $\Psi = \Gamma_i$. Thus, Proposition 1 can be restated as follows:

**Proposition 2** (*Kraft Equality for Phylogenies [3]*)**.** *Let $\Gamma$ be a set of $n$ taxa, and let $i \in \Gamma$. A sequence of positive integers $\tau_i = [\tau_{ij} \in [2, n-1] : j \in \Gamma_i]$ is a path-length sequence of a phylogeny $T$ of $\Gamma$ if and only if the entries of $\tau_i$ satisfy the following condition:*

$$\sum_{j \in \Gamma_i} 2^{-\tau_{ij}} = \frac{1}{2}. \tag{3}$$

A phylogeny corresponding to a given path-length sequence $\tau_i = [\tau_{ij} \in [2, n-1] : j \in \Gamma_i]$, for some $i \in \Gamma$, can be easily reconstructed, e.g., by sorting $\tau_i$ in ascending order and by drawing the path-lengths from $i$ to all of the remaining taxa in $\Gamma_i$. However, it is worth noting that no bijective relation between the set $\Theta_i$ and either the set of phylogenies or the set of the UBTs can be defined, as a path-length sequence may correspond to multiple distinct phylogenies. For example, Fig. 2 shows that there exists three possible phylogenies that can be reconstructed from the path-length sequence $\tau_1 = [3, 3, 4, 4, 4, 4]$.

### 2.2. Characterizing $\Theta$

Providing a characterization of the set $\Theta$ is less trivial than characterizing $\Theta_i$. As a first attempt, we may observe that, as phylogenies are (non-oriented) acyclic graphs (trees), any path-length matrix $\tau \in \Theta$ must satisfy the following properties:

$$\tau_{ii} = 0 \qquad \forall\, i \in \Gamma \tag{4}$$

$$\tau_{ij} = \tau_{ji} \qquad \forall\, i, j \in \Gamma : i < j \tag{5}$$

$$\tau_{ij} + \tau_{jk} - \tau_{ik} \geq 2 \qquad \forall\, i, j, k \in \Gamma : i \neq j \neq k \tag{6}$$

$$\tau_{ij} \in \{2, 3, 4, \ldots, n-1\} \qquad \forall\, i, j \in \Gamma : i \neq j. \tag{7}$$

As the tree encoded by any path-length matrix $\tau \in \Theta$ has to be a phylogeny, the rows of $\tau$ must satisfy Kraft's equalities, i.e.,

$$\sum_{j \in \Gamma_i} 2^{-\tau_{ij}} = \frac{1}{2} \qquad \forall\, i \in \Gamma. \tag{8}$$

Moreover, as shown in Catanzaro et al. [3], the entries of any path-length matrix $\tau \in \Theta$ must satisfy the following equation:

$$\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \tau_{ij} 2^{-\tau_{ij}} = (2n - 3). \tag{9}$$

We refer to the manifold described by (9) as *the phylogenetic manifold*.

Finally, as any path-length matrix $\tau \in \Theta$ must encode a tree, it must satisfy Buneman's *additive property*, i.e., exactly one of the following conditions must hold on the entries of $\tau$ [24–26]:

$$\begin{cases} \tau_{ij} + \tau_{pq} + 2 \leq \tau_{iq} + \tau_{jp} = \tau_{ip} + \tau_{jq} \\ \tau_{iq} + \tau_{jp} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{ip} + \tau_{jq} \\ \tau_{ip} + \tau_{jq} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{iq} + \tau_{jp} \end{cases} \qquad \forall\, i, j, p, q \in \Gamma : i \neq j \neq p \neq q. \tag{10}$$

It is easy to see that conditions (4)–(9) are independent. As an example, the following minimal dimension matrix

$$\begin{pmatrix} 0 & 3 & 3 & 3 & 3 \\ 3 & 0 & 3 & 3 & 3 \\ 3 & 3 & 0 & 3 & 3 \\ 3 & 3 & 3 & 0 & 3 \\ 3 & 3 & 3 & 3 & 0 \end{pmatrix}$$

satisfies all the considered conditions but (9). In fact, $\sum_{i \in \Gamma} \sum_{j \in \Gamma_i} \tau_{ij} 2^{-\tau_{ij}} = \frac{15}{2} \neq 7$. Similarly, the following minimal dimension matrix

$$\begin{pmatrix} 0 & 3 & 5 & 5 & 2 & 4 \\ 3 & 0 & 3 & 4 & 4 & 3 \\ 5 & 3 & 0 & 2 & 5 & 4 \\ 5 & 4 & 2 & 0 & 5 & 3 \\ 2 & 4 & 5 & 5 & 0 & 3 \\ 4 & 3 & 4 & 3 & 3 & 0 \end{pmatrix} \tag{11}$$

satisfies all of the considered conditions but (6). In fact, $\tau_{12} + \tau_{12} - \tau_{13} = 1 < 2$. It is currently unclear whether conditions (4)–(10) are all independent and sufficient to characterize $\Theta$. The presence of nonlinear relationships in the space of the topological distances, such as Kraft equalities (8) or the phylogenetic manifold (9), anyhow suggests the search for alternative spaces to carry out the study of the polyhedral combinatorics of the BMEP, in which at least one between (8) or (9) can be linearized. In the next section, we will consider the space proposed by Forcey et al. [18]. Before proceeding, we summarize in Table 1 the most important notation that is necessary to retain throughout the article.

**Table 1**
Table of notation.

| Symbol | Description |
|---|---|
| $\Gamma$ | Set of taxa |
| $\Gamma_i$ | $\Gamma \setminus \{i\}$, for some $i \in \Gamma$ |
| $\mathcal{T}$ | Set of the phylogenies of $\Gamma$ |
| $\tau_{ij}(T)$ | Length of the (unique) path between taxa $i, j \in \Gamma$ on a fixed phylogeny $T \in \mathcal{T}$ |
| $\tau_i(T)$ | Sequence of path-lengths associated to a phylogeny $T \in \mathcal{T}$ rooted in taxon $i$ |
| $\tau$(T) | Matrix of path-lengths associated to a given phylogeny $T \in \mathcal{T}$ |
| $\Theta_i$ | Set of path-lengths sequences associated to the phylogenies of $\Gamma$ rooted in taxon $i$ |
| $\Theta$ | Set of path-length matrices associated to the phylogenies in $\mathcal{T}$ |
| $x_{ij}(T)$ | Quantity equal to $2^{n-1-\tau_{ij}}$, for some given $\tau(T)$, $T \in \mathcal{T}$ |
| $X(T)$ | Matrix whose generic entry is $x_{ij}(T)$, for some given $\tau(T)$, $T \in \mathcal{T}$ |
| $\mathcal{X}$ | Set of the matrices $X(T)$ corresponding to the phylogenies $T \in \mathcal{T}$ |
| Space($\mathcal{X}$) | Vector space of minimal dimension that includes $\mathcal{X}$ |
| $\hat{T}$ | Fundamental caterpillar phylogeny (see Fig. 4) |
| $T^{r,s}$ | Caterpillar phylogeny obtained from $\hat{T}$ by swapping taxa $r$ and $s$ |
| $\mathcal{C}^s$ | The set $\{X(T^{r,s}) : r \in [1, s-1]\}$ for a fixed natural $s \in [2, n-1]$ |
| $\mathcal{D}^k$ | The set $\bigcup_{s=2}^{k} \mathcal{C}^s$ for a fixed natural $k \in [2, n-1]$ |
| $\begin{cases} \text{Lin}(S) \\ \text{Aff}(S) \\ \text{Conv}(S) \end{cases}$ | $\begin{cases} \text{Linear} \\ \text{Affine} \\ \text{Convex} \end{cases}$ hull of a given set $S$ |

## 3. The $\mathcal{X}$ space

Given a set $\Gamma$ of $n \geq 3$ taxa, a phylogeny $T$ of $\Gamma$, and a topological distance $\tau_{ij}$ between two not necessarily distinct taxa $i, j \in \Gamma$ in $T$, we define

$$x_{ij} = 2^{n-1-\tau_{ij}}. \tag{12}$$

Note that by definition $\tau_{ij}$ is a positive integer such that $2 \leq \tau_{ij} \leq n-1$ if $i \neq j$, and 0 otherwise. Hence, by definition, $x_{ij}$ is a positive integer such that $1 \leq x_{ij} \leq 2^{n-3}$ if $i \neq j$, and $2^{n-1}$ otherwise. Given the path-length matrix $\tau$ associated to $T$, we denote $X$ as the $n \times n$ symmetric matrix whose generic entry is $x_{ij}$ and $\mathcal{X}$ as the set of the $X$ matrices corresponding to all of the possible path-length matrices $\tau \in \Theta$. To avoid ambiguous situations in which it may be unclear whether a matrix $X$ is associated to a specific phylogeny or to another, hereafter we will write $X(T)$ whenever we will need to specify that $X$ is associated to the phylogeny $T$, and we will write $x_{ij}(T)$ to denote the corresponding generic entry. For example, consider a set $\Gamma$ of four taxa; then, the corresponding set $\Theta$ includes the following three path-length matrices encoding the only three possible phylogenies of $\Gamma$ shown in Fig. 3:

$$\tau(T_1) = \begin{pmatrix} 0 & 2 & 3 & 3 \\ 2 & 0 & 3 & 3 \\ 3 & 3 & 0 & 2 \\ 3 & 3 & 2 & 0 \end{pmatrix}; \qquad \tau(T_1) = \begin{pmatrix} 0 & 3 & 2 & 3 \\ 3 & 0 & 3 & 2 \\ 2 & 3 & 0 & 3 \\ 3 & 2 & 3 & 0 \end{pmatrix}; \qquad \tau(T_3) = \begin{pmatrix} 0 & 3 & 3 & 2 \\ 3 & 0 & 2 & 3 \\ 3 & 2 & 0 & 3 \\ 2 & 3 & 3 & 0 \end{pmatrix}.$$

Then, the corresponding set $\mathcal{X}$ includes the following three matrices

$$X(T_1) = \begin{pmatrix} 8 & 2 & 1 & 1 \\ 2 & 8 & 1 & 1 \\ 1 & 1 & 8 & 2 \\ 1 & 1 & 2 & 8 \end{pmatrix}; \qquad X(T_2) = \begin{pmatrix} 8 & 1 & 2 & 1 \\ 1 & 8 & 1 & 2 \\ 2 & 1 & 8 & 1 \\ 1 & 2 & 1 & 8 \end{pmatrix}; \qquad X(T_3) = \begin{pmatrix} 8 & 1 & 1 & 2 \\ 1 & 8 & 2 & 1 \\ 1 & 2 & 8 & 1 \\ 2 & 1 & 1 & 8 \end{pmatrix}.$$

By construction, each matrix $X \in \mathcal{X}$ is symmetric, diagonal dominant and doubly stochastic up to a constant $3 \cdot 2^{n-2}$. The bijection (12) induces the following $n(n+3)/2$ linear independent conditions on the entries of each $X \in \mathcal{X}$, which are analogous to conditions (4), (5), and (8) in the $\Theta$ space:

$$x_{ii} = 2^{n-1} \qquad \forall\, i \in \Gamma \tag{13}$$

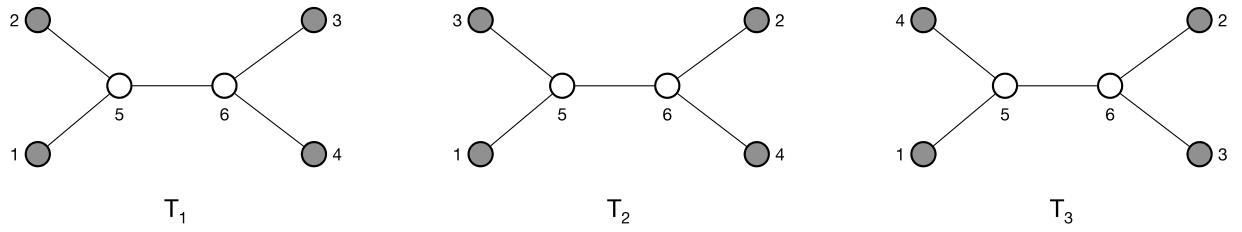



**Fig. 3.** The set of all of the possible phylogenies for the set $\Gamma = \{1, 2, 3, 4\}$.

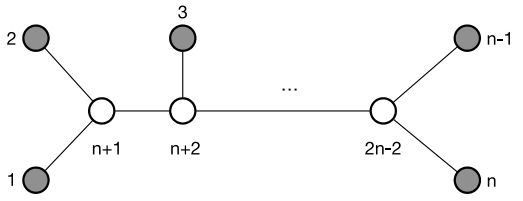



**Fig. 4.** The fundamental caterpillar phylogeny of a set $\Gamma$ of $n \geq 4$ taxa.

$$x_{ij} = x_{ji} \qquad \forall\, i, j \in \Gamma : i \neq j \tag{14}$$

$$\sum_{j \in \Gamma_i} x_{ij} = 2^{n-2} \qquad \forall\, i \in \Gamma. \tag{15}$$

Note, in particular, how Kraft equalities for phylogenies (8) becomes linear in (15). Note also that, due to the symmetry conditions (14), once that $n(n-3)/2$ out of $n(n-1)/2$ upper triangular entries of a matrix $X \in \mathcal{X}$ are fixed, the remaining $n(n+3)/2$ entries are uniquely determined.

We denote Space($\mathcal{X}$) as the vector space of minimal dimension that includes $\mathcal{X}$ and we propose, as already done in [18], to study the BMEP polytope in such a space. We will show in the next section that there exists a basis for Space($\mathcal{X}$) having cardinality $n(n-3)/2 + 1 = (n-1)(n-2)/2$. This result will prove useful to characterize some properties of the BMEP polytope in Space($\mathcal{X}$).

## 4. A basis for Space($\mathcal{X}$)

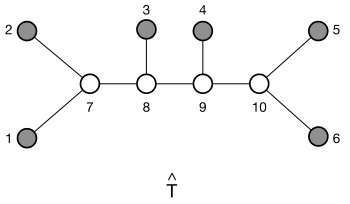
Given a set $\Gamma$ of $n \geq 4$ taxa, we define the *fundamental caterpillar phylogeny* of $\Gamma$ as the caterpillar phylogeny $\hat{T}$ such that

- the first taxon is assigned to the first of the four leaves of the caterpillar that have a sibling;
- the $i$th taxon is assigned to the only leaf of the caterpillar $\hat{T}$ at topological distance $\tau_{1,i} = i$, for $i \in [2, n-2]$;
- the last two taxa are assigned to the only two leaves at topological distance $n-1$ from the first taxon. Fig. 4 shows an example of $\hat{T}$.

We denote $T^{r,s}$ as the caterpillar phylogeny that can be obtained from $\hat{T}$ by swapping the positions of taxa $r$ and $s$. As an example, Fig. 5 shows the fundamental caterpillar phylogeny $\hat{T}$ for the set $\Gamma = [1, 5]$ as well as the caterpillar phylogenies $T^{1,4}$, $T^{2,4}$, and $T^{3,4}$.

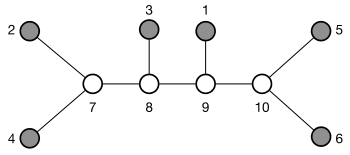
We define the following sets

$$\mathcal{C}^s = \{X(T^{r,s}) : r \in [1, s-1]\} \quad \forall\, s \in [2, n-1]$$

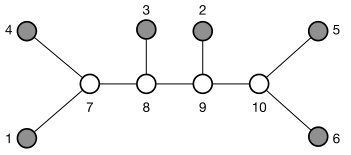



$\hat{T}$

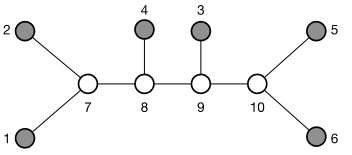
$$X(\hat{T}) = \begin{pmatrix} 32 & 8 & 4 & 2 & 1 & 1 \\ 8 & 32 & 4 & 2 & 1 & 1 \\ 4 & 4 & 32 & 4 & 2 & 2 \\ 2 & 2 & 4 & 32 & 4 & 4 \\ 1 & 1 & 2 & 4 & 32 & 8 \\ 1 & 1 & 2 & 4 & 8 & 32 \end{pmatrix}$$




$T^{1,4}$




$T^{2,4}$




$T^{3,4}$

$$X(T^{1,4}) = \begin{pmatrix} 32 & 2 & 4 & 2 & 4 & 4 \\ 2 & 32 & 4 & 8 & 1 & 1 \\ 4 & 4 & 32 & 4 & 2 & 2 \\ 2 & 8 & 4 & 32 & 1 & 1 \\ 4 & 1 & 2 & 1 & 32 & 8 \\ 4 & 1 & 2 & 1 & 8 & 32 \end{pmatrix}$$

$$X(T^{2,4}) = \begin{pmatrix} 32 & 2 & 4 & 8 & 1 & 1 \\ 2 & 32 & 4 & 2 & 4 & 4 \\ 4 & 4 & 32 & 4 & 2 & 2 \\ 8 & 2 & 4 & 32 & 1 & 1 \\ 1 & 4 & 2 & 1 & 32 & 8 \\ 1 & 4 & 2 & 1 & 8 & 32 \end{pmatrix}$$

$$X(T^{3,4}) = \begin{pmatrix} 32 & 8 & 2 & 4 & 1 & 1 \\ 8 & 32 & 2 & 4 & 1 & 1 \\ 2 & 2 & 32 & 4 & 4 & 4 \\ 4 & 4 & 4 & 32 & 2 & 2 \\ 1 & 1 & 4 & 2 & 32 & 8 \\ 1 & 1 & 4 & 2 & 8 & 32 \end{pmatrix}$$

**Fig. 5.** Top: the fundamental caterpillar phylogeny $\hat{T}$ of a set $\Gamma$ of six taxa and the associated matrix $X(\hat{T})$. Middle: the caterpillar phylogenies $T^{1,4}$, $T^{2,4}$, and $T^{3,4}$. Bottom: the set $\mathcal{C}^4 = \{X(T^{1,4}), X(T^{2,4}), X(T^{3,4})\}$.

and we observe that, by definition,

- only matrix $X(T^{1,2})$ in $\mathcal{C}^2$ is equal to $X(\hat{T})$ as leaves 1 and 2 are siblings;
- $X(T^{r,s}) \neq X(\hat{T})$, for all the remaining values of $r$ and $s$;
- $X(T^{r,s}) \neq X(T^{\bar{r},\bar{s}})$ if and only if $\bar{r} \neq r$ or $\bar{s} \neq s$.

For example, Fig. 5 shows the matrix $X(\hat{T})$ associated to $\hat{T}$ as well as the matrices $X(T^{1,4})$, $X(T^{2,4})$, and $X(T^{3,4})$ associated to $T^{1,4}$, $T^{2,4}$, and $T^{3,4}$. The matrices $X(T^{1,4})$, $X(T^{2,4})$, and $X(T^{3,4})$ are elements of the set $\mathcal{C}^4$.

For a fixed positive integer $k \in [2, n-1]$, we define

$$\mathcal{D}^k = \bigcup_{s=2}^{k} \mathcal{C}^s.$$

We shall prove now that the set $\mathcal{D}^{n-1}$ constitutes a basis for Space($\mathcal{X}$). As a first step, we observe that, for a given positive integer $s \in [2, n-1]$, each set $\mathcal{C}^s$ includes $s-1$ matrices and, by construction, has empty intersection with any other set in $\mathcal{D}^{n-1}$, i.e., $|\mathcal{C}^s| = s-1$ and $\mathcal{C}^s \cap \mathcal{C}^t = \emptyset$, for $s, t \in [2, n-1]$, $s \neq t$. As a consequence, we have that

$$\left|\mathcal{D}^{n-1}\right| = \left|\bigcup_{r=2}^{n-1} \mathcal{C}^s\right| = \sum_{r=2}^{n-1} |\mathcal{C}^s| = \sum_{s=2}^{n-1} (s-1) = \frac{(n-1)(n-2)}{2} = \frac{n(n-3)}{2} + 1. \tag{16}$$

We also observe that the entries of each $X$ matrix in the set $\mathcal{D}^{n-1}$ can be expressed in function of the entries of the path-length matrix associated to $\hat{T}$. Specifically, letting $(\alpha)^+ = \max\{0, \alpha\}$, $\alpha \in \mathbb{R}$, the topological distance between taxa $i$ and $j$ in $\hat{T}$ can be written as

$$\tau_{i,j}(\hat{T}) = \begin{cases} n - 1 - (i-2)^+ - (n-1-j)^+ & \text{if } j > i \\ 0 & \text{if } j = i \\ n - 1 - (j-2)^+ - (n-1-i)^+ & \text{if } j < i. \end{cases}$$

Then, for all $i, j \in \Gamma$, the generic entry of matrix $X(\hat{T})$ is

$$x(\hat{T})_{i,j} = \begin{cases} 2^{(i-2)^+ + (n-1-j)^+} & \text{if } j > i \\ 2^{n-1} & \text{if } j = i \\ 2^{(j-2)^+ + (n-1-i)^+} & \text{if } j < i \end{cases} \tag{17}$$

and the generic entry of each matrix $X(T^{r,s})$ is

$$x(T^{r,s})_{i,j} = \begin{cases} x(\hat{T})_{r,j} & \text{if } i = s \text{ and } j \neq r \\ x(\hat{T})_{i,r} & \text{if } j = s \text{ and } i \neq r \\ x(\hat{T})_{s,j} & \text{if } i = r \text{ and } j \neq s \\ x(\hat{T})_{i,s} & \text{if } j = r \text{ and } i \neq s \\ x(\hat{T})_{i,j} & \text{otherwise.} \end{cases} \tag{18}$$

In other words, each matrix $X(T^{r,s})$ can be written as $X(T^{r,s}) = P^{r,s} X(\hat{T}) P^{r,s}$ where, $P^{r,s} = (P^{r,s})^{-1} = (P^{r,s})^T$ is the $n \times n$ permutation matrix that swaps the $r$th row (column) with the $s$th row (column). It is easy to show that the set of permutation matrices $\{P^{r,s} : r \in [2, n-1], s \in [r+1, n]\}$ is linearly independent. However, this result does not have any direct implication on the linear independence of the set $\mathcal{D}^{n-1}$.

The following intermediate result will prove useful to show that $\mathcal{D}^{n-1}$ constitutes a basis for Space($\mathcal{X}$).

**Proposition 3.** *Consider a matrix $\bar{X} \in \mathcal{X}$ and let $\Phi = \left\{X^k \in \mathcal{X}, k \in [1, \kappa]\right\}$, for some positive integer $\kappa > 1$, be a set of distinct matrices in $\mathcal{X}$. If $\bar{X} = \sum_{k=1}^{\kappa} \lambda_k X^k$, then*

$$\sum_{k=1}^{\kappa} \lambda_k = 1.$$

**Proof.** As (15) holds on the entries of $\bar{X}$ we have that $\sum_{j=1}^{n} \bar{x}_{1,j} = 3 \cdot 2^{n-2} = \sum_{j=1}^{n} x^k_{1,j}$, for $k \in [1, \kappa]$. Moreover, as $\bar{X} = \sum_{k=1}^{\kappa} \lambda_k X^k$, the condition $\bar{x}_{1,j} = \sum_{k=1}^{\kappa} \lambda_k x^k_{1,j}$ also holds, for $j \in [1, n]$. Then, we have

$$3 \cdot 2^{n-2} = \sum_{j=1}^{n} \bar{x}_{1,j} = \sum_{j=1}^{n} \sum_{k=1}^{\kappa} \lambda_k x^k_{1,j} = \sum_{k=1}^{\kappa} \lambda_k \sum_{j=1}^{n} x^k_{1,j} = \sum_{k=1}^{\kappa} \lambda_k \left(3 \cdot 2^{n-2}\right),$$

hence $\sum_{k=1}^{\kappa} \lambda_k = 1$. □

In the light of Proposition 3, we can now prove that the matrices in $\mathcal{D}^{n-1}$ are linearly independent.

**Proposition 4.** *The set $\mathcal{D}^s$ is linearly independent, for any $s \in [2, n-1]$.*

**Proof.** We prove the statement by induction.

*Base case.* The set $\mathcal{D}^2 = \mathcal{C}^2 = \{X(T^{1,2})\}$ includes the single matrix $X(T^{1,2})$ which is linearly independent as it is different from the null matrix 0.

*Inductive step.* By assuming that the matrices in $\mathcal{D}^k$ are linearly independent, we prove that the matrices in the set $\mathcal{D}^k \cup \mathcal{C}^{k+1}$ are linearly independent as well. To this end, we observe that each matrix $X(T^{r,s}) \in \mathcal{D}^k$ is characterized by the entries

$$x(T^{r,s})_{i,j} = \begin{cases} 2^{(i-2)^+ + (n-1-j)^+} & \text{if } i \neq r, s \\ 2^{s-2+(n-1-j)^+} & \text{if } i = r \\ 2^{(r-2)^+ + (n-1-j)^+} & \text{if } i = s \end{cases} \qquad i \in [1, k],\ j \in [k+1, k+2]. \tag{19}$$

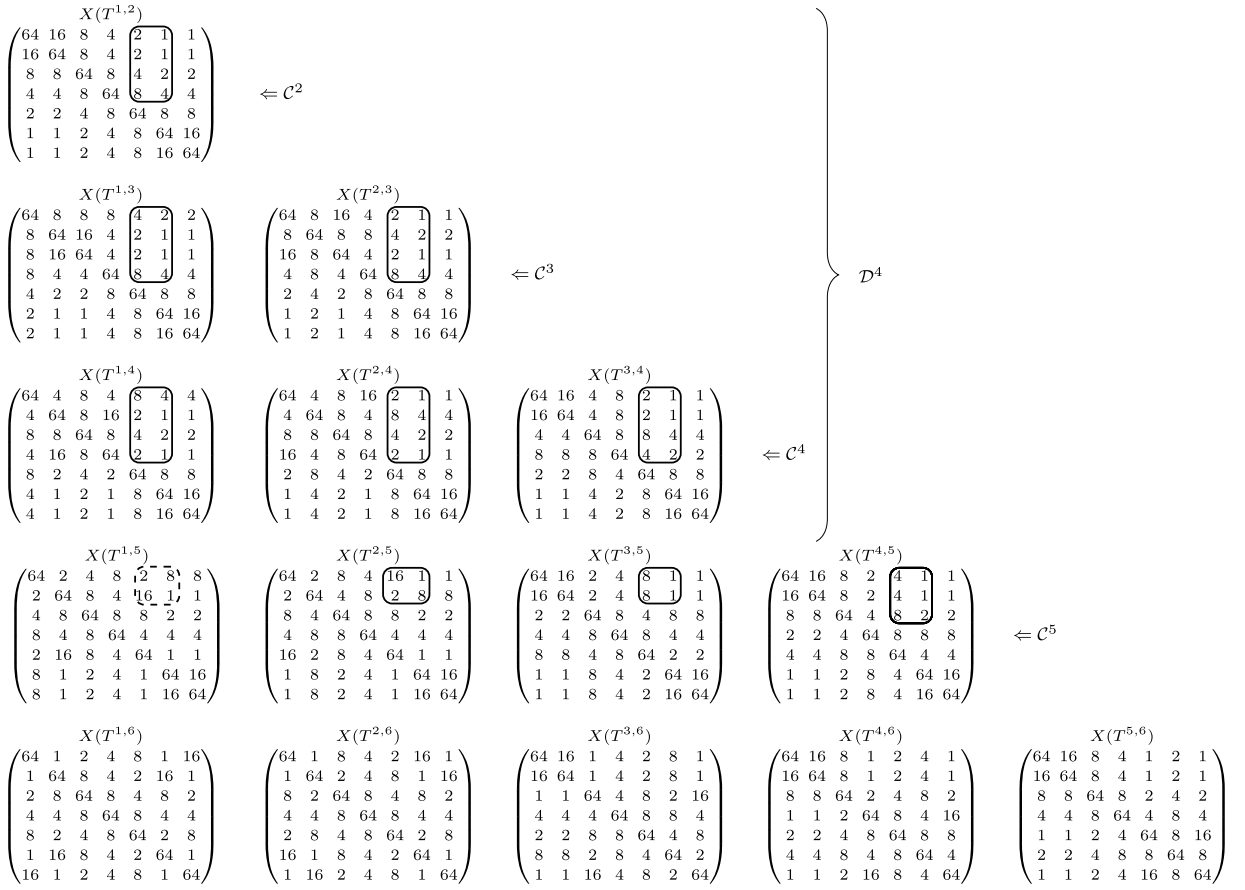



**Fig. 6.** Sets $\mathcal{D}^4$ and $\mathcal{C}^5$ of a set $\Gamma$ of seven taxa. The entries of the matrices that satisfy conditions (20) are framed within a solid line. The entries of the matrices $X(T^{1,5})$ and $X(T^{2,5})$ that are used to define the corresponding matrices of type (22) are framed within a dashed line.

Note that in (19) the index $j$ is always different from the index $s$ because indices $r$ and $s$ are smaller than or equal to $k$ whereas $j$ is strictly greater than $k$. Also note that each matrix $X(T^{r,k+1}) \in \mathcal{C}^{k+1}$ is characterized by the entries

$$x(T^{r,k+1})_{i,k+1} = \begin{cases} 2^{(i-2)^+ + n-1-r} & \text{if } i < r \\ 2^{(r-2)^+ + n-1-i} & \text{if } i > r \\ 2^{(i-2)^+ + n-2-k} & \text{if } i = r \end{cases} \quad i \in [1, k-1]$$

and

$$x(T^{r,k+1})_{i,k+2} = \begin{cases} 2^{(i-2)^+ + (n-3-k)^+} & \text{if } i \neq r, \\ 2^{n-4+(k+3-n)^+} & \text{if } i = r \end{cases} \quad i \in [1, k-1].$$

Now, given a generic matrix $X \in \mathcal{D}^k \cup \mathcal{C}^{k+1}$, consider the following linear condition parametrized in $t \in [2, k]$:

$$x_{t,k+1} + 2^{1-(k+3-n)^+} \sum_{i=1}^{t-1} x_{i,k+2} = 2^{1-(k+3-n)^+} x_{t,k+2} + \sum_{i=1}^{t-1} x_{i,k+1}. \tag{20}$$

It is easy to see that the matrices $X(T^{r,s}) \in \mathcal{D}^k$ satisfy conditions (20) for $t \in [2, k]$. In contrast, the matrices $X(T^{r,k+1}) \in \mathcal{C}^{k+1}$ satisfy conditions (20) if $r \in [3, k]$ and $t \in [2, r-1]$ (see, e.g., Fig. 6). Finally, if $r \in [1, 2]$, none of the matrices $X(T^{r,k+1}) \in \mathcal{C}^{k+1}$ satisfy condition (20). We can take advantage of condition (20) to

prove that the matrices in the set

$$\mathcal{D}^k \cup \left\{X(T^{r,k+1}) : r \in [2,k]\right\} = \mathcal{D}^k \cup \left(\mathcal{C}^{k+1} \setminus \{X(T^{1,k+1})\}\right)$$

are linearly independent. To this end, we note that $X(T^{k,k+1}) \in \mathcal{C}^{k+1}$ does not belong to the linear hull $\text{Lin}(\mathcal{D}^k)$, since $X(T^{k,k+1})$ does not satisfy condition (20) for $t = k$. We also note that $X(T^{k-1,k+1}) \notin \text{Lin}(\mathcal{D}^k \cup \{X(T^{k,k+1})\})$, since $X(T^{k-1,k+1})$ does not satisfy conditions (20) for $t = k-1$. In general, for $r \in [2,k]$, $X(T^{r,k+1}) \notin \text{Lin}(\mathcal{D}^k \cup \{X(T^{\hat{r},k+1}) : \hat{r} \in [r+1,k]\})$, since $X(T^{r,k+1})$ does not satisfy conditions (20) for $t = r$.

It remains to prove that $X(T^{1,k+1}) \notin \text{Lin}(\mathcal{D}^k \cup \{X(T^{r,k+1}) : r \in [2,k]\})$. To this end, assume by contradiction that there exist two sets of scalars $\{\lambda_r : r \in [2,k]\}$ and $\{\mu_{r,s} : r \in [1,k-1],\ s \in [r+1,k]\}$ such that

$$X(T^{1,k+1}) = \underbrace{\sum_{r=2}^{k} \lambda_r X(T^{r,k+1})}_{\text{matrices in } \mathcal{C}^{k+1} \setminus \{X(T^{1,k+1})\}} + \underbrace{\sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} X(T^{r,s})}_{\text{matrices in } \mathcal{D}^k}. \tag{21}$$

For a fixed matrix $X \in \mathcal{X}$, consider the $2 \times 2$ matrix

$$Y(X) = \begin{pmatrix} x_{1,k+1} & 2^{1-(k-n+3)^+} x_{1,k+2} \\ x_{2,k+1} & 2^{1-(k-n+3)^+} x_{2,k+2} \end{pmatrix} \tag{22}$$

(see, e.g., Fig. 6) and the scalar

$$\xi(X) = x_{2,k+1} + 2^{1-(k-n+3)^+} x_{1,k+2} - 2^{1-(k-n+3)^+} x_{2,k+2} - x_{1,k+1}.$$

Eq. (21) implies that

$$\xi(X(T^{1,k+1})) = \sum_{r=2}^{k} \lambda_r \xi(X(T^{r,k+1})) + \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} \xi(X(T^{r,s})).$$

It is worth noting that $\xi(X(T^{1,k+1})) = \xi(X(T^{2,k+1}))$ and $\xi(X(T^{r,s})) = 0$ for $\mathcal{D}^k \cup \{X(T^{r,k+1}) : r \in [3,k]\}$ as the matrices in this latter set satisfy condition (20) for $t = 2$. Thus, we have that $\lambda_2 = 1$ and we can rewrite (21) as

$$X(T^{1,k+1}) - X(T^{2,k+1}) = \sum_{r=3}^{k} \lambda_r X(T^{r,k+1}) + \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} X(T^{r,s})$$

which implies

$$Y(X(T^{1,k+1}) - X(T^{2,k+1})) = \sum_{r=3}^{k} \lambda_r Y(X(T^{r,k+1})) + \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} Y(X(T^{r,s})).$$

Now, observe that the following conditions hold:

$$Y(X(T^{1,s})) = \begin{pmatrix} 2^{s-2+n-2-k} & 2^{s-2+n-2-k} \\ 2^{n-2-k} & 2^{n-2-k} \end{pmatrix} \quad \forall\ X(T^{1,s}) \in \mathcal{D}^k$$

$$Y(X(T^{2,s})) = \begin{pmatrix} 2^{n-2-k} & 2^{n-2-k} \\ 2^{s-2+n-2-k} & 2^{s-2+n-2-k} \end{pmatrix} \quad \forall\ X(T^{2,s}) \in \mathcal{D}^k$$

$$Y(X(T^{r,s})) = \begin{pmatrix} 2^{n-2-k} & 2^{n-2-k} \\ 2^{n-2-k} & 2^{n-2-k} \end{pmatrix} \quad \forall\ X(T^{r,s}) \in \mathcal{D}^k,\ r \geq 3$$

$$Y(X(T^{r,k+1})) = \begin{pmatrix} 2^{n-1-r} & 2^{n-2-k} \\ 2^{n-1-r} & 2^{n-2-k} \end{pmatrix} \quad \forall\ X(T^{r,k+1}) \in \mathcal{C}^{k+1},\ r \geq 3$$

and

$$Y(X(T^{1,k+1}) - X(T^{2,k+1})) = \begin{pmatrix} 2^{n-2-k} - 2^{n-3} & 2^{n-3} - 2^{n-2-k} \\ 2^{n-3} - 2^{n-2-k} & 2^{n-2-k} - 2^{n-3} \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \forall\, k \geq 2.$$

Then, the following system

$$2^{n-2-k} - 2^{n-3} = \sum_{r=3}^{k} \lambda_r 2^{n-1-r} + \sum_{s=2}^{k} \mu_{1,s} 2^{s-2+n-2-k} + \sum_{s=3}^{k} \mu_{2,s} 2^{n-2-k} + \sum_{r=3}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} 2^{n-2-k} \tag{23}$$

$$2^{n-3} - 2^{n-2-k} = \sum_{r=3}^{k} \lambda_r 2^{n-2-k} + \sum_{s=2}^{k} \mu_{1,s} 2^{s-2+n-2-k} + \sum_{s=3}^{k} \mu_{2,s} 2^{n-2-k} + \sum_{r=3}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} 2^{n-2-k} \tag{24}$$

$$2^{n-3} - 2^{n-2-k} = \sum_{r=3}^{k} \lambda_r 2^{n-1-r} + \sum_{s=2}^{k} \mu_{1,s} 2^{n-2-k} + \sum_{s=3}^{k} \mu_{2,s} 2^{s-2+n-2-k} + \sum_{r=3}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} 2^{n-2-k} \tag{25}$$

$$2^{n-2-k} - 2^{n-3} = \sum_{r=3}^{k} \lambda_r 2^{n-2-k} + \sum_{s=2}^{k} \mu_{1,s} 2^{n-2-k} + \sum_{s=3}^{k} \mu_{2,s} 2^{s-2+n-2-k} + \sum_{r=3}^{k-1} \sum_{s=r+1}^{k} \mu_{r,s} 2^{n-2-k} \tag{26}$$

is infeasible. Specifically, by subtracting (24) from (23) we get:

$$2(2^{n-2-k} - 2^{n-3}) = \sum_{r=3}^{k} \lambda_r 2^{n-1-r} - \sum_{r=3}^{k} \lambda_r 2^{n-2-k}. \tag{27}$$

By subtracting (26) from (25) we get:

$$2(2^{n-2-k} - 2^{n-3}) = -\sum_{r=3}^{k} \lambda_r 2^{n-1-r} + \sum_{r=3}^{k} \lambda_r 2^{n-2-k}. \tag{28}$$

Now, by adding (27) and (28) we get $2^{n-2-k} - 2^{n-3} = 0$ which leads to a contradiction. Thus, the statement follows. □

Let $\text{Lin}(\mathcal{D}^{n-1})$ denote the linear hull of set $\mathcal{D}^{n-1}$. Moreover, let $\text{Conv}(\mathcal{X})$ denote the convex hull of the matrices $X \in \mathcal{X}$ and let $\text{Dim}(\text{Conv}(\mathcal{X}))$ be its dimension. Then, the following proposition holds.

**Proposition 5.**

*(i)* $\text{Dim}(\text{Conv}(\mathcal{X})) = n(n-3)/2$;
*(ii)* *Space*$(\mathcal{X})$ = *Lin*$(\mathcal{D}^{n-1})$.

**Proof.**

(i) First note that $\mathcal{D}^{n-1}$ is a set of linearly independent vectors included in $\text{Conv}(\mathcal{X})$ and having cardinality $n(n-3)/2 + 1$, hence $\text{Dim}(\text{Conv}(\mathcal{X})) \geq n(n-3)/2$. Now, let $\text{Aff}(\mathcal{X}) \subseteq \mathbb{R}^{n \times n}$ denote the set of the $n \times n$ real matrices satisfying conditions (13)–(15). By definition, $\text{Aff}(\mathcal{X})$ is an affine space that includes $\text{Conv}(\mathcal{X})$. Moreover, by the linear independence of conditions (13)–(15) it holds that $\text{Dim}(\text{Aff}(\mathcal{X})) = n(n-3)/2$. Because $n(n-3)/2 \leq \text{Dim}(\text{Conv}(\mathcal{X})) \leq \text{Dim}(\text{Aff}(\mathcal{X}))$, the statement follows.

(ii) As shown in (i), $\mathcal{D}^{n-1}$ is a basis for $\text{Aff}(\mathcal{X})$. Now, observe that the affine space induced by the vectors in $\text{Conv}(\mathcal{X})$ does not include the null matrix, hence $\text{Dim}(\text{Space}(\mathcal{X}))$ must be strictly greater than $\text{Dim}(\text{Conv}(\mathcal{X}))$. Because $\text{Space}(\mathcal{X})$ is the vector space of minimal dimension that includes $\text{Conv}(\mathcal{X})$ and because $\text{Dim}(\text{Lin}(\mathcal{D}^{n-1})) = \text{Dim}(\text{Conv}(\mathcal{X})) + 1$, the statement trivially follows. □

Eickmeyer et al. [16] first identified the dimension of the BMEP polytope in the space of the topological distances. The above proposition states that this result is valid also for Space($\mathcal{X}$). We conclude this section by observing that Space($\mathcal{X}$) includes an exponential number of bases made of caterpillars. For example, by using an approach similar to the one used in this section, it is easy to show that the set $\mathcal{D}_{\pi}^{n-1}$, obtained from the fundamental caterpillar phylogeny $\hat{T}$ when applying a permutation $\pi$ to the assignment of the taxa in $\Gamma$ to the leaves of $\hat{T}$, is linearly independent.

## 5. On the convex hull of the set $\mathcal{X}$

A question that naturally arises is whether it is possible to characterize some properties of the vertices of Conv($\mathcal{X}$). To address this question, we first observe that the bijection (12) implies that $\tau_{ij} = (n-1-\log_2 x_{ij})$ for any phylogeny $T \in \mathcal{T}$ and any pair $i, j \in \Gamma$. Then, we can rewrite (9) as

$$\sum_{\substack{i,j\in\Gamma \\ i\neq j}} x_{ij}(n-1-\log_2 x_{ij}) = (2n-3)2^{n-1}. \tag{29}$$

As (15) implies

$$\sum_{\substack{i,j\in\Gamma \\ i\neq j}} x_{ij} = n2^{n-2}, \tag{30}$$

for all $X \in \mathcal{X}$ we have that

$$\sum_{\substack{i,j\in\Gamma \\ i\neq j}} x_{ij}\log_2 x_{ij} = \sum_{\substack{i,j\in\Gamma \\ i\neq j}} x_{ij}(n-1-\tau_{ij}) = (n^2-5n+6)2^{n-2}. \tag{31}$$

In the light of this result, the following proposition holds:

**Proposition 6.** *For any fixed phylogeny $T \in \mathcal{T}$, the unique optimal solution to the problem*

$$\begin{aligned} z^* = \min & \sum_{\substack{i,j\in\Gamma \\ i\neq j}} x_{ij}\tau_{ij}(T) \\ X \in & \ Conv(\mathcal{X}) \end{aligned}$$

*is $X(T)$.*

**Proof.** We first observe that (30)–(31) imply that, for all $X \in \mathcal{X}$, the following equalities hold

$$\sum_{\substack{i,j\in\Gamma \\ i\neq j}} x_{ij}\tau_{ij} = (2n-3)2^{n-1} \tag{32}$$

$$\sum_{\substack{i,j\in\Gamma \\ i\neq j}} (x_{ij}\log_2 x_{ij} - x_{ij}) = (n^2-6n+6)2^{n-2}. \tag{33}$$

Now, consider a function $g : \mathbb{R}_+^{n\times n} \to \mathbb{R}$ defined as

$$g(X) = \sum_{\substack{i,j\in\Gamma \\ i\neq j}} (x_{ij}\log_2 x_{ij} - x_{ij}).$$

It is easy to see that $g$ is differentiable, strictly convex, and that the sublevel set

$$L = \{X \in \mathbb{R}^{n\times n} : g(X) \leq (n^2-6n+6)2^{n-2}\}$$

is strictly convex and includes $Conv(\mathcal{X})$. We note that, for each fixed phylogeny $T \in \mathcal{T}$, the tangent plane to $L$ in $X(T)$ is

$$\sum_{\substack{i,j\in\Gamma\\ i\neq j}} \frac{g(X(T))}{\partial x_{ij}}(x_{ij} - x_{ij}(T)) = 0 \Leftrightarrow \sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}\log_2(x_{ij}(T)) = (n^2 - 5n + 6)2^{n-2}.$$

Then, the strict convexity of $g(x)$ over $L \supset Conv(\mathcal{X})$ implies that, for all $X \in Conv(\mathcal{X})$,

$$\sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}(T)\log_2(x_{ij}(T)) = (n^2 - 5n + 6)2^{n-2} > \sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}\log_2(x_{ij}(T)).$$

This last condition, together with equalities (30), (31) and (32), in turn implies that, for all $X \in Conv(\mathcal{X})$,

$$\sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}(T)\tau_{ij}(T) = (2n-3)2^{n-1} < \sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}\tau_{ij}(T)$$

thus, the statement follows. □

The following two corollaries derive from Proposition 6. It is worth noting that Corollary 1 was first claimed by [16]. We provide here a proof of their result by deriving it from the phylogenetic manifold.

**Corollary 1.** *The set $\mathcal{X}$ coincides with the set of the vertices of $Conv(\mathcal{X})$.*

**Proof.** We recall from Section 2, that any phylogeny $T$ of $\Gamma$ must satisfy the equation of the phylogenetic manifold (9):

$$\sum_{\substack{i,j\in\Gamma\\ i\neq j}} \tau_{ij}2^{-\tau_{ij}} = 2n - 3.$$

As the bijection (12) implies that $\tau_{ij} = (n - 1 - \log_2 x_{ij})$, we can rewrite (9) as

$$\sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}(n - 1 - \log_2 x_{ij}) = (2n-3)2^{n-1}.$$

Hence, defining the function $f : \mathbb{R}_+^{n\times n} \to \mathbb{R}$ as

$$f(X) = \sum_{\substack{i,j\in\Gamma\\ i\neq j}} x_{ij}(n - 1 - \log_2 x_{ij}), \tag{34}$$

all of the matrices $X \in \mathcal{X}$ are points in which the manifold (34) satisfies $f(X) = (2n-3)2^{n-1}$. We also observe that the function $f$ is strictly concave as the entries of its Hessian are such that

$$H_{ij,lk} = \frac{\partial^2 f}{\partial x_{ij}\partial x_{lk}} = \begin{cases} \frac{\partial^2 x_{ij}(n-1-\log_2 x_{ij})}{\partial x_{ij}^2} = -\frac{1}{x_{ij}\log_2 e} < 0 & \text{if } i = l \text{ and } j = k,\\ 0 & \text{otherwise.}\end{cases} \tag{35}$$

Now, assume by contradiction that, for some positive constant $\kappa > 1$, there exist both a matrix $\bar{X} \in \mathcal{X}$ and a subset $\{X_k : k \in \{1,\dots,\kappa\}\} \subseteq \mathcal{X} \setminus \{\bar{X}\}$ such that

$$\bar{X} = \sum_{k=1}^{\kappa} \lambda_k X_k \quad \text{with} \quad \sum_{k=1}^{\kappa} \lambda_k = 1 \quad \text{and} \quad 0 < \lambda_k < 1, \quad \forall\, k \in \{1,\dots,\kappa\}.$$

For such a matrix, the strict concavity of function $f$ implies that

$$f(\bar{X}) = f(\sum_{k=1}^{\kappa} \lambda_k X_k) > \sum_{k=1}^{\kappa} \lambda_k f(X_k) = (2n-3)2^{n-1}.$$

However, as $\bar{X} \in \mathcal{X}$, this fact contradicts the equation of the phylogenetic manifold (9) that requires $f(\bar{X}) = (2n-3)2^{n-1}$. Thus, the statement follows. □

**Corollary 2.** *Given a set of positive reals $c_{ij}$ for all $i, j \in \Gamma$, $i \neq j$, a lower bound on the optimal solution to the problem*

$$z^* = \min \sum_{\substack{i,j \in \Gamma \\ i \neq j}} c_{ij} x_{ij}$$
$$X \in Conv(\mathcal{X})$$

*is $\lambda(n^2-5n+6)2^{n-2}$, where $\lambda$ is the value such that*

$$\sum_{\substack{i,j \in \Gamma \\ i \neq j}} c_{ij} 2^{\frac{c_{ij}}{\lambda}} = \lambda(n^2-5n+6)2^{n-2}.$$

**Proof.** As the level set $L = \{X \in \mathbb{R}^{n \times n} : g(x) \leq (n^2-6n+6)2^{n-2}\}$ is strictly convex and includes $Conv(\mathcal{X})$, then the optimal solution of

$$z_{LB} = \min_{X \in L} \sum_{\substack{i,j \in \Gamma \\ i \neq j}} c_{ij} x_{ij}$$

is a lower bound for $z^*$ and lies on the frontier of $L$. The Karush-Kuhn–Tucker conditions applied to this latter problem impose $c_{ij} = \lambda \log_2(x_{ij})$ and

$$\sum_{\substack{i,j \in \Gamma \\ i \neq j}} x_{ij} \log_2(x_{ij}) = (n^2-5n+6)2^{n-2},$$

thus the statement follows. □

In the next sections, we will briefly review the facet-defining inequalities of Conv($\mathcal{X}$) currently described in the literature on the BMEP and we will present new families of valid inequalities. We will see that for specific values of $n$ some of these families are also facet-defining.

## 6. Valid inequalities and facets of Conv($\mathcal{X}$)

### *6.1. Known facets of Conv($\mathcal{X}$)*

Forcey et al. [18] provided a complete description of the BMEP polytope for $n \leq 5$ (i.e., when assuming an input set $\Gamma$ including up to five taxa) and characterized a number of facet-defining inequalities of Conv($\mathcal{X}$) for $n \geq 6$ (i.e., when assuming an input set $\Gamma$ including six taxa). For the sake of completeness, in this subsection we briefly report and comment upon their results. In the next subsection, we will present a new set of valid and facet-defining inequalities that will contribute to the description of Conv($\mathcal{X}$) for $n = 6$ or more.

1. Let $\Gamma$ be a set of $n \geq 4$ taxa. For each pair of taxa $i, j \in \Gamma$ such that $i < j$, the *caterpillar inequality*

$$x_{ij} \geq 1 \tag{36}$$

provides a set of $\binom{n}{2}$ facet-defining inequalities for $\text{Conv}(\mathcal{X})$. In particular, four caterpillar inequalities hold as equality in the case the optimal solution to the BMEP is a caterpillar.

2. Let $\Gamma$ be a set of $n \geq 4$ taxa. For any distinct triplet of taxa $i, j, k \in \Gamma$, the *intersecting cherries inequality*

$$x_{ij} + x_{jk} - x_{ik} \leq 2^{n-3} \tag{37}$$

provides a set of $\binom{n}{2}(n-2)$ facet-defining inequalities for $\text{Conv}(\mathcal{X})$. We note that in some circumstances there exists an analogy between this family of inequalities and the family of triangular inequalities (6) in the $\tau$-space (i.e., $\tau_{ij} + \tau_{jk} - \tau_{ik} \geq 2$) introduced in Catanzaro et al. [3]. Specifically, as

$$\tau_{ij} = n - 1 - \log_2(x_{ij}),$$

we can rewrite (6) as

$$n - 1 - \log_2(x_{ij}) - \log_2(x_{jk}) + \log_2(x_{ik}) \geq 2 \tag{38}$$

or analogously as

$$\log_2(x_{ij}) + \log_2(x_{jk}) - \log_2(x_{ik}) \leq n - 3. \tag{39}$$

Raising both sides of (39) to power 2, we get

$$\frac{x_{ij}x_{jk}}{x_{ik}} = 2^{\log_2(x_{ij})+\log_2(x_{jk})-\log_2(x_{ik})} \leq 2^{n-3}. \tag{40}$$

Now, it is worth noting that when the triangular inequality (6) holds tightly, i.e., $\tau_{ik} = \tau_{ij} + \tau_{jk} - 2$ and either both $\tau_{ik} \leq \tau_{ij}$ and $\tau_{ik} \leq \tau_{ij}$ or both $\tau_{ik} \geq \tau_{ij}$ and $\tau_{ik} \geq \tau_{ij}$, then the following inequality holds:

$$x_{ij} + x_{jk} - x_{ik} \leq \frac{x_{ij}x_{jk}}{x_{ik}}. \tag{41}$$

Specifically, under the considered assumptions, we can rewrite (41) as $(x_{ik} - x_{ij})(x_{ik} - x_{jk}) \geq 0$, or analogously as

$$(2^{n-1-\tau_{ik}} - 2^{n-1-\tau_{ij}})(2^{n-1-\tau_{ik}} - 2^{n-1-\tau_{jk}}) = 2^{n-1}(2^{-\tau_{ik}} - 2^{-\tau_{ij}})(2^{-\tau_{ik}} - 2^{-\tau_{jk}}) \geq 0. \tag{42}$$

In particular, when $\tau_{ik} = \tau_{ij} + \tau_{jk} - 2$, (42) becomes:

$$\begin{aligned} 2^{n-1}(2^{-\tau_{ik}} - 2^{-\tau_{ij}})(2^{-\tau_{ik}} - 2^{-\tau_{jk}}) &= 2^{n-1}(2^{-\tau_{ij}-\tau_{jk}+2} - 2^{-\tau_{ij}})(2^{-\tau_{ij}-\tau_{jk}+2} - 2^{-\tau_{jk}}) \\ &= 2^{n-1}\left[2^{-\tau_{ij}}(2^{2-\tau_{jk}} - 1)\right]\left[2^{-\tau_{jk}}(2^{2-\tau_{ij}} - 1)\right] \geq 0 \end{aligned} \tag{43}$$

which trivially holds for any value of $\tau_{ij}$ and $\tau_{jk}$. These arguments show that if a triangular inequality (6) is tight for a given taxa assignment then the corresponding intersecting cherry inequality (37) is obviously feasible but not necessarily tight. In contrast, if an intersecting cherry inequality is tight then the corresponding triangular inequality is necessarily tight as condition (37) holds as an equality only when either $\tau_{jk} = 2$ or $\tau_{ij} = 2$ [18].

The intersecting cherries inequalities also show that

$$x_{ij} \leq 2^{n-3} \tag{44}$$

is not facet-defining for $\text{Conv}(\mathcal{X})$. In fact, (44) can be seen as a sum of the intersecting cherries inequalities $x_{ij} + x_{jk} - x_{ik} \leq 2^{n-3}$ and $x_{ij} + x_{ik} - x_{jk} \leq 2^{n-3}$, hence it is redundant for the description of $\text{Conv}(\mathcal{X})$.

We also note that, as any non-diagonal entry of a matrix $X \in \mathcal{X}$ satisfies $x_{ij} \leq 2^{n-3}$ and at least four non-diagonal entries of $X$ satisfy $x_{ij} = 2^{n-3}$, then at least $4(n-2)$ intersecting cherries inequalities hold as equalities at any optimal solution to the problem.

Finally, we note that the intersecting-cherry inequalities are tight when either a cherry $\{i, j\}$ or a cherry $\{j, k\}$ is present on a phylogeny.

3. Let $\Gamma$ be a set of $n = 5$ taxa. Then, the *cyclic ordering inequality*

$$x_{ij} + x_{jk} + x_{kp} + x_{pq} + x_{qi} \leq 13 \tag{45}$$

for all distinct $i, j, k, p, q \in \Gamma$ provides a set of 12 facet-defining inequalities for $\text{Conv}(\mathcal{X})$, for $n = 5$. In particular, the cyclic ordering inequalities together with the caterpillar inequalities and the cherry inequalities provide a complete description of $\text{Conv}(\mathcal{X})$.

4. Let $\Gamma$ be a set of $n \geq 6$ taxa. Consider a bipartition of $\Gamma$ into two subsets, say $S_1$ and $S_2$, such that $|S_1| = 3$ and $|S_2| = m \geq 3$. Then, for any distinct triplet of taxa $i, j, k \in S_1$, the $(m, 3)$-*split inequality*

$$x_{ij} + x_{jk} + x_{ik} \leq 2^{n-2} \tag{46}$$

provides a set of $\binom{n}{3}$ facet-defining inequalities for $\text{Conv}(\mathcal{X})$. More in general, consider a bipartition of the set $\Gamma$ into two subsets, say $S_1$ and $S_2$, such that $|S_1| = k \geq 3$ and $|S_2| = m \geq 3$. Then, the $(m, k)$-*split inequality*

$$\sum_{\substack{i,j \in S_1 \\ i<j}} x_{ij} \leq (k-1)2^{n-3} \tag{47}$$

provides a set of $2^{n-1} - \binom{n}{2} - n - 1$ facet-defining inequalities for $\text{Conv}(\mathcal{X})$. We note that the $(m, k)$-split inequality generated by a given bipartition $(S_1, S_2)$ of the set $\Gamma$ imposes that the subtree of a phylogeny $T$ having $S_1$ as leaf-set must be rooted and binary. We observe that these inequalities are equivalent to cut inequalities

$$\sum_{i \in S_1, j \in S_2} x_{ij} \geq 2^{n-2}$$

that can be separated by standard max-flow/min cut algorithms.

### 6.2. New valid inequalities for Conv($\mathcal{X}$)

In this section, we provide a new set of valid inequalities for $\text{Conv}(\mathcal{X})$. If not stated otherwise, we will intend that all of them are at least face-inducing for each $n$ and facet-inducing, at least for some values of $n$. Specifically, we will start by considering the case $n = 6$ (i.e., a set $\Gamma$ of six taxa) as in Forcey et al. [18] and, where possible, we will generalize the result for larger number of taxa.

One way to determine the facet-defining attribute of a given inequality consists in showing a set of $\text{Dim}(\text{Conv}(\mathcal{X})) - 1$ affinely independent $X$ matrices that satisfy the inequality at equality [27]. As seen in the previous sections, this task can be quite long and tedious, mainly due to a lack of general properties to assess the linear independence of a system of matrices in the $\text{Space}(\mathcal{X})$. Hence, we will proceed here as follows. We will first check whether a given inequality is valid. Subsequently, as suggested in Forcey et al. [18], we will make use of technical computing systems and general purpose software for polyhedral analyses, such as Mathematica [28] and Polymake [19], respectively, to verify whether there exist $\text{Dim}(\text{Conv}(\mathcal{X})) - 1$ $X$ matrices that are both affinely independent and satisfy the inequality at equality. A fact that will be recurrently used throughout this section is that the topology of a phylogeny of six taxa can only be either a *caterpillar* or a *balanced* phylogeny (see Fig. 7). Hence, the $X$ matrix of a balanced phylogeny of six taxa includes no entry equal to 1, whereas the $X$ matrix of a caterpillar of six taxa includes four entries equal to 1. We will also recurrently use the following simple facts (whose proofs are omitted):

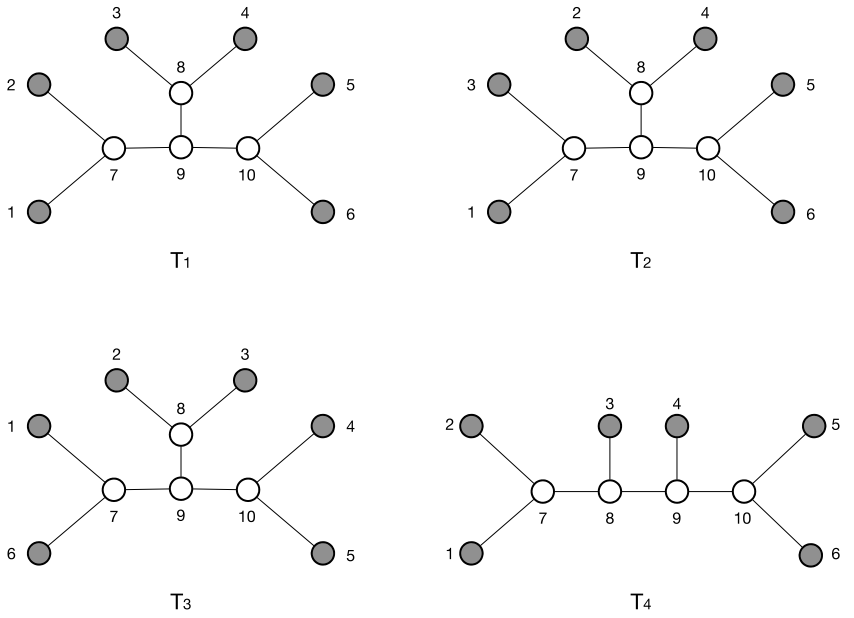



**Fig. 7.** An example of possible phylogenies of a set $\Gamma$ of six taxa. The topology of such phylogenies can be either balanced (e.g., as in $T_1$, $T_2$ and $T_3$) or a caterpillar (e.g., as in $T_4$).

**Observation 1.** *Let $T$ denote a phylogeny of $\Gamma$. Then,*

*(i) $T$ has at least two cherries.*

*(ii) $T$ is a caterpillar if and only if*

$$\max_{i,j\in\Gamma}\{\tau_{ij}\} = n-1.$$

*As a consequence, $X(T)$ contains an entry equal to one if and only if $T$ is a caterpillar.*

*(iii) $T$ is a 1-branch caterpillar if and only if*

$$\max_{i,j\in\Gamma}\{\tau_{ij}\} = n-2.$$

*As a consequence, if in a matrix $X(T)$ there exists an entry such that $x_{ij}(T) = 2$ for some $i, j \in \Gamma$, then $T$ is either a caterpillar or a 1-branch caterpillar.*

*(iv) If $T$ is neither a caterpillar nor a 1-branch caterpillar, then*

$$\max_{i,j\in\Gamma}\{\tau_{ij}\} \leq n-3.$$

*In this situation, the corresponding entries of $X(T)$ are such that $x_{ij}(T) \geq 4$ for all $i, j \in \Gamma$.*

*(v) A path between two taxa of $T$ is maximal only if the taxa belong to distinct cherries.*

*(vi) If $T$ has $k+2$ cherries, for some natural $k \geq 1$, then the maximum length of a path in $T$ is $n-1-k$. As a consequence, if $T$ has $k+2$ cherries then the corresponding entries of $X(T)$ are such that $x_{ij}(T) \geq 2^k$.*

In the light of the above facts, the following proposition holds.

**Proposition 7.** *Let $\Gamma$ be a set of $n \geq 6$ taxa. Then, for any triplet of distinct taxa $i_1, i_2, i_3 \in \Gamma$, the following inequalities*

$$2^{n-5}x_{i_1i_2} + 2^{n-5}x_{i_1i_3} + x_{i_2i_3} \geq 5 \cdot 2^{(n-5)} \tag{48}$$

$$2^{n-4}x_{i_1i_2} + 2^{n-4}x_{i_1i_3} + x_{i_2i_3} \geq 2^{(n-2)} \tag{49}$$

$$x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} \geq (3+\rho)2^{k-1}, \quad \text{where } k = \lfloor \frac{n}{3} \rfloor \text{ and } \rho = n - 3k, \tag{50}$$

*are valid for Conv($\mathcal{X}$).*

**Proof.**

**Validity of (48).** Consider the phylogeny $T$ encoded by a generic matrix $X \in \mathcal{X}$. If $x_{i_1i_2} + x_{i_1i_3} \geq 5$ then (48) is trivially valid for Conv($\mathcal{X}$). Otherwise, the following three cases may occur:

**Case 1:** $x_{i_1i_2} = x_{i_1i_3} = 1$. In this situation, $T$ is a caterpillar and its two cherries are constituted by the pairs of taxa $i_2$, $i_3$ and $i_1$, $i_\alpha$, for some $\alpha \neq \{1,2,3\}$. Thus, it holds that $x_{i_2i_3} = 2^{n-3}$ and the inequality is valid.

**Case 2:** $x_{i_1i_2} = 1$ and $x_{i_1i_3} = 2$. Also in this situation $T$ is a caterpillar and its two cherries are constituted by the pairs of taxa $i_1$, $i_\alpha$ and $i_2$, $i_\beta$, for some $\alpha, \beta \neq \{1,2\}$. As $x_{i_1i_3} = 2$, the topological distance from $i_2$ to $i_3$ is $\tau_{23} = 3$ and $x_{i_2i_3} = 2^{n-4}$, so the inequality is again valid.

**Case 3:** $x_{i_1i_2} = x_{i_1i_3} = 2$. In this situation, $T$ can be either a caterpillar or a 1-branch caterpillar. If $T$ is a caterpillar, the hypothesis $x_{i_1i_2} = x_{i_1i_3} = 2$ imposes that $i_2$ and $i_3$ define a cherry, hence $x_{i_2i_3} = 2^{n-3}$ and as result the inequality is valid. If $T$ is a 1-branch caterpillar, we have that $i_1$, $i_2$ and $i_3$ belong to the three different cherries. In particular, the internal edges of the leaf adjacent vertices of the cherries that include $i_2$ and $i_3$ have to share the same separator vertices. The topological distance from $i_2$ to $i_3$ is then $\tau_{23} = 4$, $x_{i_2i_3} = 2^{n-5}$, and so the inequality is valid again.

**Validity of (49).** Consider the phylogeny $T$ encoded by a generic matrix $X \in \mathcal{X}$. If $x_{i_1i_2} + x_{i_1i_3} \geq 4$, then (49) is trivially valid for Conv($\mathcal{X}$). Otherwise, the following two cases may occur:

**Case 1:** $x_{i_1i_2} = x_{i_1i_3} = 1$. In this situation, $T$ is a caterpillar with cherries $i_2$, $i_3$ and $i_1$, $i_\alpha$, for some $\alpha \neq \{1,2,3\}$. Then, $x_{i_2i_3} \geq 2^{n-3}$ and the inequality is valid.

**Case 2:** $x_{i_1i_2} = 1$ and $x_{i_1i_3} = 2$ (the case $x_{i_1i_2} = 2$ and $x_{i_1i_3} = 1$ is analogous). In this situation, $T$ is a caterpillar with cherries $i_1$, $i_\alpha$, and $i_2$, $i_\beta$, for some $\alpha, \beta \neq \{1,2,3\}$, and $i_3$ is at most the closest vertex to $i_2$ on the path linking the cherries. Thus, $x_{i_2i_3} \geq 2^{n-4}$ and the inequality is valid.

**Validity of (50).** We prove the validity of (50) by showing that the minimum value of its left-hand side always satisfies the right-hand side. To this end, let $j$ denote the only internal vertex which is common to the three paths that join taxa $i_1, i_2, i_3$ on a phylogeny $T$. Then, it holds that

$$\begin{aligned} \tau_{i_1i_2} &= \tau_{i_1j} + \tau_{i_2j} \\ \tau_{i_1i_3} &= \tau_{i_1j} + \tau_{i_3j} \\ \tau_{i_2i_3} &= \tau_{i_2j} + \tau_{i_3j} \end{aligned}$$

hence the left-hand side of (50) can be rewritten as

$$x_{12} + x_{13} + x_{23} = 2^{n-1-(\tau_{i_1j}+\tau_{i_2j})} + 2^{n-1-(\tau_{i_1j}+\tau_{i_3j})} + 2^{n-1-(\tau_{i_2j}+\tau_{i_3j})}.$$

Now, without loss of generality, suppose that taxa $i_1$, $i_2$, $i_3$ are assigned to the leaves of $T$ such that $\tau_{i_1j} \leq \tau_{i_2j} \leq \tau_{i_3j}$ and $\tau_{i_1i_2} \leq \tau_{i_1i_3} \leq \tau_{i_2i_3}$. Then, we first claim that if the left-hand side of (50) is minimum then $\tau_{i_1j} + \tau_{i_2j} + \tau_{i_3j} = n$. To prove this claim, suppose by contradiction that the minimum of the left-hand side of (50) is achieved when $\tau_{i_1j} + \tau_{i_2j} + \tau_{i_3j} = n - s > 0$, for some positive integer $s$. Then, we can construct a new phylogeny $\tilde{T}$ having a smaller value of the left-hand side of (50), thus contradicting the minimality hypothesis. Specifically, construct 1-branch caterpillar phylogeny $\tilde{T}$ by

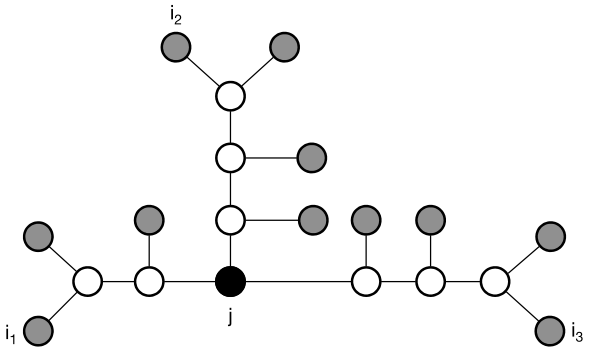



**Fig. 8.** An example of phylogenies of a set $\Gamma$ of eleven taxa in which the value $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3}$, left-hand side of (50), assumes a minimal value.

joining a single separator vertex $j$ to three $d$-subcaterpillars such that each of the taxa $i_1$, $i_2$, $i_3$ belongs to the cherry of a different subcaterpillar (see e.g., Fig. 8) and

$$\begin{aligned}\tilde{\tau}_{i_1j} &= \tau_{i_1j} + s\\ \tilde{\tau}_{i_2j} &= \tau_{i_2j}\\ \tilde{\tau}_{i_3j} &= \tau_{i_3j}.\end{aligned}$$

In this situation, it holds that $\tilde{\tau}_{i_1j} > \tau_{i_1j}$, which in turn implies that

$$\begin{aligned}2^{n-1-(\tilde{\tau}_{i_1j}+\tilde{\tau}_{i_2j})} + 2^{n-1-(\tilde{\tau}_{i_1j}+\tilde{\tau}_{i_3j})} + 2^{n-1-(\tilde{\tau}_{i_2j}+\tilde{\tau}_{i_3j})} <& 2^{n-1-(\tau_{i_1j}+\tau_{i_2j})} + 2^{n-1-(\tau_{i_1j}+\tau_{i_3j})}\\ &+ 2^{n-1-(\tau_{i_2j}+\tau_{i_3j})}\end{aligned}$$

thus leading to the mentioned contradiction of the minimality hypothesis. We also note that requiring the minimality of the left-hand side of (50) further implies that the condition $\tau_{i_3j} - \tau_{i_1j} \leq 1$ (hence $\tau_{i_2i_3} - \tau_{i_1i_2} \leq 1$) must hold as well. In particular, it is easy to see that if we shorten $\tau_{i_3j}$ by one unit and we lengthen $\tau_{i_1j}$ by the same quantity we have that

$$\begin{aligned}2^{n-1-(\tau_{i_1j}+\tau_{i_2j}+1)} + 2^{n-1-(\tau_{i_1j}+\tau_{i_3j})} + 2^{n-1-(\tau_{i_2j}+\tau_{i_3j}-1)} \leq& 2^{n-1-(\tau_{i_1j}+\tau_{i_2j})} + 2^{n-1-(\tau_{i_1j}+\tau_{i_3j})}\\ &+ 2^{n-1-(\tau_{i_2j}+\tau_{i_3j})}\end{aligned}$$

when $\tau_{i_1j} + 1 \leq \tau_{i_3j}$. Then, in the light of the above considerations, the following three cases hold:

**Case 1:** $n = 3k$. In this situation, the minimum value of $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3}$ is $3 \cdot 2^{k-1}$ obtained when $\tau_{i_1i_2} = \tau_{i_1i_3} = \tau_{i_2i_3} = 2k$.

**Case 2:** $n = 3k+1$. In this situation, the minimum value of $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3}$ is $4 \cdot 2^{k-1} = (3+1)2^{k-1}$ obtained when $\tau_{i_1i_2} = 2k$ and $\tau_{i_1i_3} = \tau_{i_2i_3} = 2k + 1$.

**Case 3:** $n = 3k+2$. In this situation, the minimum value of $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3}$ is $= 5 \cdot 2^{k-1} = (3+2)2^{k-1}$ obtained when $\tau_{i_1i_2} = \tau_{i_1i_3} = 2k + 1$ and $\tau_{i_2,i_3} = 2k + 2$.

All of the three cases satisfy the right-hand side of (50), thus this inequality is valid for $\mathrm{Conv}(\mathcal{X})$. □

It is worth noting that (48) is facet-defining for $n = 6, 7, 8$. In particular, for $n = 6$ this property can be easily verified by enumerating the set of vertices of $\mathrm{Conv}(\mathcal{X})$ that satisfy (48) as an equality. This set, shown in Table 2, defines a matrix whose rank is $n(n-3)/2 = 9$. With a similar enumerative approach, the facet-defining property of (48) can be verified also for $n = 7$ and $n = 8$. We observe that also (49) is

**Table 2**
The complete set of vertices of Conv($\mathcal{X}$) that satisfy $2^{n-5}x_{i_1i_2} + 2^{n-5}x_{i_1i_3} + x_{i_2i_3} \geq 5 \cdot 2^{(n-5)}$ at equality when assuming $n = 6$. The vertices are encoded by concatenating one after another the rows of the strictly upper triangular parts of a matrix $X$. By replacing e.g., the index $i_1$ with 1, $i_2$ with 2, and $i_3$ with 3 it is easy to realize that the corresponding entries in the table satisfy the considered inequality. The tables that follow provide a similar information for other classes of inequalities.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 8 | 2 | 2 | 2 |
| 02 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 2 |
| 03 | 2 | 2 | 2 | 8 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 |
| 04 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 8 | 2 | 2 | 2 | 2 | 2 |
| 05 | 2 | 2 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 |
| 06 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 2 | 2 |
| 07 | 2 | 1 | 8 | 4 | 1 | 4 | 2 | 4 | 4 | 1 | 2 | 8 | 4 | 1 | 2 |
| 08 | 1 | 2 | 8 | 4 | 1 | 4 | 1 | 2 | 8 | 2 | 4 | 4 | 4 | 1 | 2 |
| 09 | 2 | 1 | 8 | 1 | 4 | 4 | 2 | 4 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 10 | 1 | 2 | 8 | 1 | 4 | 4 | 1 | 8 | 2 | 2 | 4 | 4 | 1 | 4 | 2 |
| 11 | 2 | 1 | 4 | 8 | 1 | 4 | 4 | 2 | 4 | 2 | 1 | 8 | 4 | 2 | 1 |
| 12 | 1 | 2 | 4 | 8 | 1 | 4 | 2 | 1 | 8 | 4 | 2 | 4 | 4 | 2 | 1 |
| 13 | 2 | 1 | 1 | 8 | 4 | 4 | 4 | 2 | 4 | 8 | 1 | 2 | 1 | 2 | 4 |
| 14 | 1 | 2 | 1 | 8 | 4 | 4 | 8 | 1 | 2 | 4 | 2 | 4 | 1 | 2 | 4 |
| 15 | 2 | 1 | 4 | 1 | 8 | 4 | 4 | 4 | 2 | 2 | 8 | 1 | 2 | 4 | 1 |
| 16 | 1 | 2 | 4 | 1 | 8 | 4 | 2 | 8 | 1 | 4 | 4 | 2 | 2 | 4 | 1 |
| 17 | 2 | 1 | 1 | 4 | 8 | 4 | 4 | 4 | 2 | 8 | 2 | 1 | 2 | 1 | 4 |
| 18 | 1 | 2 | 1 | 4 | 8 | 4 | 8 | 2 | 1 | 4 | 4 | 2 | 2 | 1 | 4 |

**Table 3**
Vertices of Conv($\mathcal{X}$) that satisfy $2^{n-4}x_{i_1i_2} + 2^{n-4}x_{i_1i_3} + x_{i_2i_3} \geq 2^{(n-2)}$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 1 | 8 | 4 | 1 | 4 | 2 | 4 | 4 | 1 | 2 | 8 | 4 | 1 | 2 |
| 02 | 1 | 2 | 8 | 4 | 1 | 4 | 1 | 2 | 8 | 2 | 4 | 4 | 4 | 1 | 2 |
| 03 | 1 | 1 | 8 | 4 | 2 | 8 | 1 | 2 | 4 | 1 | 2 | 4 | 4 | 2 | 4 |
| 04 | 2 | 1 | 8 | 1 | 4 | 4 | 2 | 4 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 05 | 1 | 2 | 8 | 1 | 4 | 4 | 1 | 8 | 2 | 2 | 4 | 4 | 1 | 4 | 2 |
| 06 | 1 | 1 | 8 | 2 | 4 | 8 | 1 | 4 | 2 | 1 | 4 | 2 | 2 | 4 | 4 |
| 07 | 2 | 1 | 4 | 8 | 1 | 4 | 4 | 2 | 4 | 2 | 1 | 8 | 4 | 2 | 1 |
| 08 | 1 | 2 | 4 | 8 | 1 | 4 | 2 | 1 | 8 | 4 | 2 | 4 | 4 | 2 | 1 |
| 09 | 1 | 1 | 4 | 8 | 2 | 8 | 2 | 1 | 4 | 2 | 1 | 4 | 4 | 4 | 2 |
| 10 | 2 | 1 | 1 | 8 | 4 | 4 | 4 | 2 | 4 | 8 | 1 | 2 | 1 | 2 | 4 |
| 11 | 1 | 2 | 1 | 8 | 4 | 4 | 8 | 1 | 2 | 4 | 2 | 4 | 1 | 2 | 4 |
| 12 | 1 | 1 | 2 | 8 | 4 | 8 | 4 | 1 | 2 | 4 | 1 | 2 | 2 | 4 | 4 |
| 13 | 2 | 1 | 4 | 1 | 8 | 4 | 4 | 4 | 2 | 2 | 8 | 1 | 2 | 4 | 1 |
| 14 | 1 | 2 | 4 | 1 | 8 | 4 | 2 | 8 | 1 | 4 | 4 | 2 | 2 | 4 | 1 |
| 15 | 1 | 1 | 4 | 2 | 8 | 8 | 2 | 4 | 1 | 2 | 4 | 1 | 4 | 4 | 2 |
| 16 | 2 | 1 | 1 | 4 | 8 | 4 | 4 | 4 | 2 | 8 | 2 | 1 | 2 | 1 | 4 |
| 17 | 1 | 2 | 1 | 4 | 8 | 4 | 8 | 2 | 1 | 4 | 4 | 2 | 2 | 1 | 4 |
| 18 | 1 | 1 | 2 | 4 | 8 | 8 | 4 | 2 | 1 | 4 | 2 | 1 | 4 | 2 | 4 |

facet-defining for $n = 6, 7, 8$. In particular, Table 3 shows the set of vertices of Conv($\mathcal{X}$) that satisfy (49) as an equality when $n = 6$. Finally, (50) is facet-defining for five and eight taxa, that is for $(k, \rho) = (1, 2)$ and $(k, \rho) = (2, 2)$, but not for six and seven taxa. For eight taxa, the vertices of the facets are associated only to 1-branch caterpillar phylogenies. We note that the left-hand side of (50) is equal to the left-hand side of an $(m, k) - split$ facet.

**Proposition 8.** *Consider a set $\Gamma$ of $n \geq 6$ taxa and denote $k = \lfloor \frac{n}{3} \rfloor$ and $\rho = n - 3k$. Then, the following inequalities are valid for Conv($\mathcal{X}$):*

$$2^{n-4}x_{i_1i_2} - x_{i_3i_4} \geq 0 \qquad \forall\, i_1, i_2, i_3, i_4 \in \Gamma : i_1 \neq i_2 \neq i_3 \neq i_4; \tag{51}$$

$$x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5} \geq (4 + \rho)2^{k-1} \qquad \forall\, i_1, i_2, i_3, i_4, i_5 \in \Gamma : i_1 \neq i_2 \neq i_3 \neq i_4 \neq i_5. \tag{52}$$

**Proof.**

**Validity of (51).** As $1 \leq x_{i_1 i_2} \leq 2^{n-3}$ and $1 \leq x_{i_3 i_4} \leq 2^{n-3}$, the inequality always holds true, unless $x_{i_1 i_2} = 1$ and $x_{i_3 i_4} = 2^{n-3}$ at the same time. This last situation, however, cannot occur. In fact, assume by contradiction that there exists a phylogeny $T$ of $\Gamma$ whose corresponding matrix $X(T)$ is such that $x_{i_1 i_2} = 1$. Then, $T$ is a caterpillar in which one of its two pairs of cherries includes taxon $i_1$ and the other includes taxon $i_2$. Now assume that $x_{i_3 i_4} = 2^{n-3}$ also occurs. Then, $T$ has a cherry that includes taxa $i_3$ and $i_4$. As taxa $i_1$, $i_2$, $i_3$, and $i_4$ are distinct by hypothesis, a contradiction arises. Thus, the statement follows.

**Validity of (52).** We prove the validity of (52) by following an approach similar to the one used to show the validity of (50), i.e., by determining the minimum value that its left-hand side may achieve. To this end, we denote $j$ as the only internal vertex common to the three paths that join taxa $i_1$, $i_2$ and $i_3$ on a phylogeny $T$. Then, the following relationships hold:

$$\begin{aligned} \tau_{i_1 i_2} &= \tau_{i_1 j} + \tau_{i_2 j} \\ \tau_{i_1 i_3} &= \tau_{i_1 j} + \tau_{i_3 j} \\ \tau_{i_2 i_3} &= \tau_{i_2 j} + \tau_{i_3 j}. \end{aligned}$$

Without loss of generality, suppose that taxa $i_1$, $i_2$, $i_3$ are assigned to the leaves of $T$ such that: $\tau_{i_1 j} \leq \tau_{i_2 j} \leq \tau_{i_3 j}$; $\tau_{i_1 i_2} \leq \tau_{i_1 i_3} \leq \tau_{i_2 i_3}$; and taxa $i_2$ and $i_3$ are leaves of two different cherries. Then, we claim that there is no loss of generality in assuming also that $i_4$, respectively $i_5$, belongs to the same cherry of $i_2$, respectively $i_3$. To prove this claim, consider $\tau_{i_4 i_5}$ and observe that the following two cases may theoretically occur:

(i) $\tau_{i_4 i_5}$ shares all of its internal edges with one of $\tau_{i_1 i_2}$, $\tau_{i_1 i_3}$, or $\tau_{i_2 i_3}$;
(ii) $\tau_{i_4 i_5}$ has $s > 0$ internal edges not in common with $\tau_{i_1 i_2}$, $\tau_{i_1 i_3}$, and $\tau_{i_2 i_3}$.

Consider the case (i). Because we are interested in the minimality of the left-hand side of (52) and because $\tau_{i_1 i_2} \leq \tau_{i_1 i_3} \leq \tau_{i_2 i_3}$ holds, we may assume that $\tau_{i_4 i_5} = \tau_{i_2 i_3}$. Hence, we can deduce that $i_4$, respectively $i_5$, belongs to the same cherry of $i_2$, respectively $i_3$.

Concerning the case (ii), it is easy to see that it cannot realize, as it would contradict the minimality of the left-hand side of (52). Specifically, assume that $\tau_{i_4 i_5}$ has $s > 0$ internal edges not in common with $\tau_{i_1 i_2}$, $\tau_{i_1 i_3}$, and $\tau_{i_2 i_3}$. Then, we can construct, as in the proof of (50), a 1-branch caterpillar phylogeny $\tilde{T}$ by joining a single separator vertex $j$ to three $d$-subcaterpillars such that

(i) each of the taxa $i_1$, $i_2$, $i_3$ belong to the cherry of a different subcaterpillar (see again Fig. 8);
(ii) taxon $i_4$, respectively $i_5$, belongs to the same cherry of $i_2$, respectively $i_3$;
(iii) and the following relationships hold:

$$\begin{aligned} \tilde{\tau}_{i_1 j} &= \tau_{i_1 j} \\ \tilde{\tau}_{i_2 j} &= \tilde{\tau}_{i_4 j} = \tau_{i_2 j} \\ \tilde{\tau}_{i_3 j} &= \tilde{\tau}_{i_5 j} = \tau_{i_3 j} + s. \end{aligned}$$

These latter topological distances imply that $\tilde{\tau}_{i_4 i_5} > \tau_{i_4 i_5}$, hence that the left-hand side of (52) achieves a strictly lower value in $\tilde{T}$ than in $T$, by leading to a contradiction.

We note that requiring the minimality of the left-hand side of (52) further implies that the condition $\tau_{i_3 j} - \tau_{i_1 j} \leq 2$ (hence $\tau_{i_2 i_3} - \tau_{i_1 i_2} \leq 2$) holds as well. In particular, it can be easily verified that if we

**Table 4**
Vertices of Conv($\mathcal{X}$) that satisfy $2^{n-4}x_{i_1i_2} - x_{i_3i_4} \geq 0$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 8 | 2 | 2 | 2 | 2 | 2 |
| 02 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 2 | 2 |
| 03 | 1 | 8 | 4 | 2 | 1 | 1 | 2 | 4 | 8 | 4 | 2 | 1 | 4 | 2 | 4 |
| 04 | 1 | 8 | 4 | 1 | 2 | 1 | 2 | 8 | 4 | 4 | 1 | 2 | 2 | 4 | 4 |
| 05 | 1 | 4 | 8 | 2 | 1 | 2 | 1 | 4 | 8 | 4 | 4 | 2 | 2 | 1 | 4 |
| 06 | 1 | 4 | 8 | 1 | 2 | 2 | 1 | 8 | 4 | 4 | 2 | 4 | 1 | 2 | 4 |
| 07 | 1 | 4 | 2 | 8 | 1 | 2 | 4 | 1 | 8 | 4 | 4 | 2 | 2 | 4 | 1 |
| 08 | 1 | 2 | 4 | 8 | 1 | 4 | 2 | 1 | 8 | 4 | 2 | 4 | 4 | 2 | 1 |
| 09 | 2 | 1 | 1 | 8 | 4 | 4 | 4 | 2 | 4 | 8 | 1 | 2 | 1 | 2 | 4 |
| 10 | 1 | 2 | 1 | 8 | 4 | 4 | 8 | 1 | 2 | 4 | 2 | 4 | 1 | 2 | 4 |
| 11 | 1 | 1 | 2 | 8 | 4 | 8 | 4 | 1 | 2 | 4 | 1 | 2 | 2 | 4 | 4 |
| 12 | 1 | 4 | 2 | 1 | 8 | 2 | 4 | 8 | 1 | 4 | 2 | 4 | 4 | 2 | 1 |
| 13 | 1 | 2 | 4 | 1 | 8 | 4 | 2 | 8 | 1 | 4 | 4 | 2 | 2 | 4 | 1 |
| 14 | 2 | 1 | 1 | 4 | 8 | 4 | 4 | 4 | 2 | 8 | 2 | 1 | 2 | 1 | 4 |
| 15 | 1 | 2 | 1 | 4 | 8 | 4 | 8 | 2 | 1 | 4 | 4 | 2 | 2 | 1 | 4 |
| 16 | 1 | 1 | 2 | 4 | 8 | 8 | 4 | 2 | 1 | 4 | 2 | 1 | 4 | 2 | 4 |
| 17 | 2 | 4 | 4 | 2 | 4 | 1 | 1 | 8 | 4 | 8 | 1 | 2 | 1 | 2 | 4 |
| 18 | 2 | 4 | 4 | 4 | 2 | 1 | 1 | 4 | 8 | 8 | 2 | 1 | 2 | 1 | 4 |

shorten $\tau_{i_3j}$ by one unit and we lengthen $\tau_{i_1j}$ by the same quantity, we have that

$$2^{n-1-(\tau_{i_1j}+\tau_{i_2j}+1)} + 2^{n-1-(\tau_{i_1j}+\tau_{i_3j})} + 2 \cdot 2^{n-1-(\tau_{i_2j}+\tau_{i_3j}-1)} \leq 2^{n-1-(\tau_{i_1j}+\tau_{i_2j})} + 2^{n-1-(\tau_{i_1j}+\tau_{i_3j})} + 2 \cdot 2^{n-1-(\tau_{i_2j}+\tau_{i_3j})}$$

when $\tau_{i_1j} + 2 \leq \tau_{i_3j}$. Then, in the light of the above considerations, the following three cases hold:

**Case 1:** $n = 3k$. In this situation, the minimum value of $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5}$ is $4 \cdot 2^{k-1}$ obtained when either $\tau_{i_1i_2} = \tau_{i_1i_3} = \tau_{i_2i_3} = \tau_{i_4i_5} = 2k$ or $\tau_{i_1i_2} = 2k - 1$, $\tau_{i_1i_3} = 2k$, $\tau_{i_2i_3} = \tau_{i_4i_5} = 2k + 1$.

**Case 2:** $n = 3k + 1$. In this situation, the minimum value of $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5}$ is $5 \cdot 2^{k-1} = (4 + 1)2^{k-1}$ obtained when either $\tau_{i_1i_2} = 2k$, $\tau_{i_1i_3} = \tau_{i_2i_3} = \tau_{i_4i_5} = 2k + 1$ or $\tau_{i_1i_2} = \tau_{i_1i_3} = 2k$, $\tau_{i_2i_3} = \tau_{i_4i_5} = 2k + 2$.

**Case 3:** $n = 3k + 2$. In this situation, the minimum value of $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5}$ is $6 \cdot 2^{k-1} = (4 + 2)2^{k-1}$ obtained when $\tau_{i_1i_2} = \tau_{i_1i_3} = 2k + 1$, $\tau_{i_2,i_3} = \tau_{i_4,i_5} = 2k + 2$. Differently, note that when $\tau_{i_1i_2} = 2k$, $\tau_{i_1i_3} = \tau_{i_2i_3} = \tau_{i_4i_5} = 2k + 2$ we have $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_2i_4} = 7 \cdot 2^{k-1}$.

All of the three cases satisfy the right-hand side of (52), thus this inequality is valid for Conv($\mathcal{X}$). □

It is worth noting that (51) holds as an equality, e.g., when $x_{i_1i_2} = 2$ and $x_{i_3i_4} = 2^{n-3}$. Moreover, it is facet-defining for $n = 6, 7, 8$. For example, Table 4 shows the set of vertices of Conv($\mathcal{X}$) that satisfy (51) as an equality when $n = 6$. Inequality (52) is facet-defining for $n = 6$, that is for $(k, \rho) = (2, 0)$. However, enumerative analyses of the facets of the BMEP polytope carried out by Polymake showed that (52) is not facet-defining for $n = 7$ and $n = 8$. When $n = 6$, the vertices of the facets defined by (52) are associated to the only two topologies (caterpillar and balanced) that can be obtained with six taxa. Such vertices are shown in Table 5.

**Proposition 9.** *Let $\Gamma$ be a set of $n \geq 6$ taxa. Then, for all distinct taxa $i_1, i_2, i_3, i_4, i_5, i_6 \in \Gamma$. the following inequalities are valid for Conv($\mathcal{X}$):*

$$2x_{i_1i_2} + 2x_{i_3i_4} + x_{i_5i_6} \geq 8 \tag{53}$$

$$x_{i_1i_2} + x_{i_3i_4} - 2^{n-4}x_{i_5i_6} \leq 2^{n-4}. \tag{54}$$

**Table 5**
Vertices of Conv($\mathcal{X}$) that satisfy $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5} \geq (4+\rho)2^{k-1}$ at equality when assuming $n = 6$ and $k = 2$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 8 | 2 | 2 | 2 |
| 02 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 2 |
| 03 | 2 | 2 | 2 | 8 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 |
| 04 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 8 | 2 | 2 | 2 | 2 | 2 |
| 05 | 2 | 2 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 |
| 06 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 2 | 2 |
| 07 | 4 | 1 | 8 | 1 | 2 | 2 | 4 | 2 | 4 | 1 | 8 | 4 | 1 | 2 | 4 |
| 08 | 1 | 4 | 8 | 1 | 2 | 2 | 1 | 8 | 4 | 4 | 2 | 4 | 1 | 2 | 4 |
| 09 | 2 | 1 | 8 | 1 | 4 | 4 | 2 | 4 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 10 | 1 | 2 | 8 | 1 | 4 | 4 | 1 | 8 | 2 | 2 | 4 | 4 | 1 | 4 | 2 |
| 11 | 4 | 1 | 1 | 8 | 2 | 2 | 2 | 4 | 4 | 8 | 1 | 4 | 1 | 4 | 2 |
| 12 | 1 | 4 | 1 | 8 | 2 | 2 | 8 | 1 | 4 | 2 | 4 | 4 | 1 | 4 | 2 |
| 13 | 2 | 1 | 1 | 8 | 4 | 4 | 4 | 2 | 4 | 8 | 1 | 2 | 1 | 2 | 4 |
| 14 | 1 | 2 | 1 | 8 | 4 | 4 | 8 | 1 | 2 | 4 | 2 | 4 | 1 | 2 | 4 |
| 15 | 4 | 2 | 4 | 2 | 4 | 1 | 8 | 1 | 2 | 1 | 8 | 4 | 1 | 2 | 4 |
| 16 | 2 | 4 | 2 | 4 | 4 | 1 | 8 | 1 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 17 | 4 | 2 | 2 | 4 | 4 | 1 | 1 | 8 | 2 | 8 | 1 | 4 | 1 | 4 | 2 |
| 18 | 2 | 4 | 4 | 2 | 4 | 1 | 1 | 8 | 4 | 8 | 1 | 2 | 1 | 2 | 4 |

**Proof.**

**Validity of (53).** Consider the phylogeny $T$ encoded by a generic matrix $X \in \mathcal{X}$. If $T$ is not a caterpillar then $\min_{i,j \in \Gamma}\{x_{ij}\} = 2$, hence the inequality is trivially valid. If $T$ is a caterpillar then two cases may occur:

**Case 1:** $x_{i_1i_2} = x_{i_3i_4} = 1$. In this situation, the pair of taxa $i_1$ and $i_2$ are leaves of the two opposite cherries of $T$, whose other leaves are necessarily $i_3$ or $i_4$. This fact trivially implies that $i_5$ and $i_6$ cannot be cherry leaves. Hence, $\tau_{i_5i_6} \leq n-3$ that implies $x_{i_5i_6} \geq 2^2$, and the inequality is valid.

**Case 2:** $x_{i_1i_2} = x_{i_5i_6} = 1$ (the case $x_{i_3i_4} = x_{i_5i_6} = 1$ is analogous and is omitted). As for Case 1, in this situation $i_1$ and $i_2$ are leaves of the two opposite cherries of $T$, whose other leaves are necessarily $i_5$ or $i_6$. As $i_3$ and $i_4$ cannot be cherry leaves and $\tau_{i_3i_4} \leq n-3$, then $x_{i_3i_4} \geq 2^2 = 4$, i.e., $2x_{i_3i_4} \geq 8$, thus the inequality is valid.

**Validity of (54).** Consider the phylogeny $T$ encoded by a generic matrix $X \in \mathcal{X}$. As $x_{i_5i_6} \geq 1$, (54) is trivially valid for Conv($\mathcal{X}$) if both $x_{i_1i_2}$ and $x_{i_3i_4}$ are smaller than or equal to $2^{n-4}$. Otherwise, the following two cases may occur:

**Case 1:** $x_{i_1i_2} = x_{i_3i_4} = 2^{n-3}$. In this situation, $T$ has cherries $i_1$, $i_2$ and $i_3$, $i_4$; moreover, it also holds that $\tau_{56} \leq n-3$. This last nontrivial fact can be proved by contradiction. Specifically, suppose that $\tau_{56} > n-3$, then $T$ is either a caterpillar or a 1-branch caterpillar. If $T$ is a caterpillar, then at least one of $i_5$ and $i_6$ belongs to a cherry; if $T$ is a 1-branch caterpillar, then $i_5$ and $i_6$ belong to two different cherries. Both situations contradict the fact that $T$ has cherries $i_1$, $i_2$ and $i_3$, $i_4$. Hence, $\tau_{56} \leq n-3$, i.e., $x_{i_5i_6} \geq 4$, and the inequality is valid.

**Case 2:** $x_{i_1i_2} = 2^{n-3}$ and $x_{i_3i_4} = 2^{n-4}$. In this situation, $i_1$ and $i_2$ form a cherry in $T$. If $T$ is a caterpillar, then $x_{i_5i_6} \geq 2$ as every path of topological length $n-1$ on $T$ must have taxa $i_1$ or $i_2$ as an endpoint. If $T$ is not a caterpillar, then the topological distance between any pair of vertices is less than $n-1$ and then $x_{i_5i_6} \geq 2$. In both cases $x_{i_5i_6} \geq 2$ holds, so the inequality is valid. □

Inequality (53) is facet-defining for $n = 6$. In particular, Table 6 shows a set of vertices of Conv($\mathcal{X}$) that satisfy (53) as an equality when $n = 6$. Empirical analyses of the facets of the BMEP polytope carried out by Polymake suggest that (53) may not be facet defining for $n > 6$. In particular, when $|\Gamma| = 8$ the most

**Table 6**
Vertices of Conv($\mathcal{X}$) that satisfy $2x_{i_1i_2} + 2x_{i_3i_4} + x_{i_5i_6} \geq 8$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 8 | 1 | 4 | 1 | 2 | 4 | 4 | 4 | 1 | 4 | 1 | 2 | 8 | 2 |
| 02 | 1 | 8 | 2 | 4 | 1 | 1 | 4 | 2 | 8 | 2 | 4 | 1 | 4 | 4 | 2 |
| 03 | 1 | 8 | 1 | 4 | 2 | 1 | 8 | 2 | 4 | 1 | 4 | 2 | 2 | 4 | 4 |
| 04 | 2 | 8 | 1 | 1 | 4 | 2 | 4 | 4 | 4 | 1 | 1 | 4 | 8 | 2 | 2 |
| 05 | 1 | 8 | 2 | 1 | 4 | 1 | 4 | 8 | 2 | 2 | 1 | 4 | 4 | 4 | 2 |
| 06 | 1 | 8 | 1 | 2 | 4 | 1 | 8 | 4 | 2 | 1 | 2 | 4 | 4 | 2 | 4 |
| 07 | 2 | 1 | 8 | 4 | 1 | 4 | 2 | 4 | 4 | 1 | 2 | 8 | 4 | 1 | 2 |
| 08 | 1 | 2 | 8 | 4 | 1 | 4 | 1 | 2 | 8 | 2 | 4 | 4 | 4 | 1 | 2 |
| 09 | 1 | 1 | 8 | 4 | 2 | 8 | 1 | 2 | 4 | 1 | 2 | 4 | 4 | 2 | 4 |
| 10 | 2 | 1 | 8 | 1 | 4 | 4 | 2 | 4 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 11 | 1 | 2 | 8 | 1 | 4 | 4 | 1 | 8 | 2 | 2 | 4 | 4 | 1 | 4 | 2 |
| 12 | 1 | 1 | 8 | 2 | 4 | 8 | 1 | 4 | 2 | 1 | 4 | 2 | 2 | 4 | 4 |
| 13 | 1 | 4 | 1 | 8 | 2 | 2 | 8 | 1 | 4 | 2 | 4 | 4 | 1 | 4 | 2 |
| 14 | 1 | 1 | 4 | 8 | 2 | 8 | 2 | 1 | 4 | 2 | 1 | 4 | 4 | 4 | 2 |
| 15 | 1 | 4 | 1 | 2 | 8 | 2 | 8 | 4 | 1 | 2 | 4 | 4 | 4 | 1 | 2 |
| 16 | 1 | 1 | 4 | 2 | 8 | 8 | 2 | 4 | 1 | 2 | 4 | 1 | 4 | 4 | 2 |
| 17 | 2 | 2 | 4 | 4 | 4 | 8 | 1 | 4 | 1 | 1 | 4 | 1 | 2 | 8 | 2 |
| 18 | 2 | 2 | 4 | 4 | 4 | 8 | 1 | 1 | 4 | 1 | 1 | 4 | 8 | 2 | 2 |
| 19 | 2 | 4 | 2 | 4 | 4 | 1 | 8 | 4 | 1 | 1 | 2 | 8 | 4 | 1 | 2 |
| 20 | 2 | 4 | 2 | 4 | 4 | 1 | 8 | 1 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |

**Table 7**
Vertices of Conv($\mathcal{X}$) that satisfy $x_{i_1i_2} + x_{i_3i_4} - 2^{n-4}x_{i_5i_6} \leq 2^{n-4}$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 8 | 2 | 1 | 4 | 1 | 2 | 1 | 4 | 1 | 4 | 4 | 4 | 2 | 8 | 2 |
| 02 | 8 | 1 | 2 | 4 | 1 | 1 | 2 | 4 | 1 | 4 | 2 | 8 | 4 | 4 | 2 |
| 03 | 8 | 2 | 1 | 1 | 4 | 2 | 1 | 1 | 4 | 4 | 4 | 4 | 8 | 2 | 2 |
| 04 | 8 | 1 | 2 | 1 | 4 | 1 | 2 | 1 | 4 | 4 | 8 | 2 | 4 | 4 | 2 |
| 05 | 4 | 2 | 1 | 8 | 1 | 4 | 2 | 4 | 2 | 4 | 2 | 4 | 1 | 8 | 1 |
| 06 | 4 | 1 | 2 | 8 | 1 | 2 | 4 | 4 | 2 | 4 | 1 | 8 | 2 | 4 | 1 |
| 07 | 4 | 1 | 1 | 8 | 2 | 2 | 2 | 4 | 4 | 8 | 1 | 4 | 1 | 4 | 2 |
| 08 | 4 | 2 | 1 | 1 | 8 | 4 | 2 | 2 | 4 | 4 | 4 | 2 | 8 | 1 | 1 |
| 09 | 4 | 1 | 2 | 1 | 8 | 2 | 4 | 2 | 4 | 4 | 8 | 1 | 4 | 2 | 1 |
| 10 | 4 | 1 | 1 | 2 | 8 | 2 | 2 | 4 | 4 | 8 | 4 | 1 | 4 | 1 | 2 |
| 11 | 4 | 4 | 2 | 4 | 2 | 2 | 1 | 8 | 1 | 4 | 2 | 4 | 1 | 8 | 1 |
| 12 | 4 | 2 | 4 | 4 | 2 | 1 | 2 | 8 | 1 | 4 | 1 | 8 | 2 | 4 | 1 |
| 13 | 4 | 2 | 2 | 4 | 4 | 1 | 1 | 8 | 2 | 8 | 1 | 4 | 1 | 4 | 2 |
| 14 | 4 | 4 | 2 | 2 | 4 | 2 | 1 | 1 | 8 | 4 | 4 | 2 | 8 | 1 | 1 |
| 15 | 4 | 2 | 4 | 2 | 4 | 1 | 2 | 1 | 8 | 4 | 8 | 1 | 4 | 2 | 1 |
| 16 | 4 | 2 | 2 | 4 | 4 | 1 | 1 | 2 | 8 | 8 | 4 | 1 | 4 | 1 | 2 |

similar inequalities that are facet-defining for Conv($\mathcal{X}$) are:

$$4x_{i_1i_2} + 2x_{i_3i_4} + 2x_{i_5i_6} + x_{i_7i_8} \geq 24$$
$$2x_{i_1i_2} + 2x_{i_1i_3} + 2x_{i_1i_4} + x_{i_2i_3} + x_{i_2i_4} + x_{i_3i_4} \geq 48.$$

Inequality (54) is facet-defining for $n = 6, 7, 8$. In particular, Table 7 shows the set of vertices of Conv($\mathcal{X}$) that satisfy (54) as an equality when $n = 6$.

Enumerative analyses of the facets of the BMEP polytope carried out by Polymake showed that the following propositions hold:

**Proposition 10.** *Let $\Gamma$ be a set of $n = 6$ taxa. Then, for any distinct $i_1, i_2, i_3, i_4, i_5, i_6 \in \Gamma$, the inequalities*

$$2x_{i_1i_2} + x_{i_2i_3} + x_{i_3i_4} + x_{i_4i_5} + x_{i_5i_6} + x_{i_1i_6} + x_{i_2i_4} \geq 16 \tag{55}$$
$$-2x_{i_1i_2} + 7x_{i_1i_3} + 7x_{i_1i_4} + 7x_{i_2i_5} + 7x_{i_2i_6} \geq 40 \tag{56}$$

*are facet-defining for Conv($\mathcal{X}$).*

**Table 8**
Vertices of Conv($\mathcal{X}$) that satisfy $2x_{i_1i_2} + x_{i_2i_3} + x_{i_3i_4} + x_{i_4i_5} + x_{i_5i_6} + x_{i_1i_6} + x_{i_2i_4} \geq 16$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 8 | 2 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 2 | 8 | 2 |
| 02 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 8 | 2 | 2 | 2 |
| 03 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 2 | 8 | 2 | 8 | 2 | 2 | 2 | 2 |
| 04 | 2 | 8 | 1 | 4 | 1 | 2 | 4 | 4 | 4 | 1 | 4 | 1 | 2 | 8 | 2 |
| 05 | 1 | 8 | 2 | 4 | 1 | 1 | 4 | 2 | 8 | 2 | 4 | 1 | 4 | 4 | 2 |
| 06 | 1 | 4 | 8 | 2 | 1 | 2 | 1 | 4 | 8 | 4 | 4 | 2 | 2 | 1 | 4 |
| 07 | 1 | 4 | 8 | 1 | 2 | 2 | 1 | 8 | 4 | 4 | 2 | 4 | 1 | 2 | 4 |
| 08 | 1 | 2 | 8 | 4 | 1 | 4 | 1 | 2 | 8 | 2 | 4 | 4 | 4 | 1 | 2 |
| 09 | 1 | 2 | 8 | 1 | 4 | 4 | 1 | 8 | 2 | 2 | 4 | 4 | 1 | 4 | 2 |
| 10 | 1 | 4 | 2 | 8 | 1 | 2 | 4 | 1 | 8 | 4 | 4 | 2 | 2 | 4 | 1 |
| 11 | 2 | 4 | 4 | 4 | 2 | 1 | 4 | 1 | 8 | 2 | 8 | 1 | 2 | 4 | 1 |

**Table 9**
Vertices of Conv($\mathcal{X}$) that satisfy $-2x_{i_1i_2} + 7x_{i_1i_3} + 7x_{i_1i_4} + 7x_{i_2i_5} + 7x_{i_2i_6} \geq 40$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 2 | 2 | 8 |
| 02 | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 8 | 2 | 2 | 8 | 2 |
| 03 | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 8 | 8 | 2 | 2 |
| 04 | 8 | 4 | 2 | 1 | 1 | 4 | 2 | 1 | 1 | 4 | 2 | 2 | 4 | 4 | 8 |
| 05 | 8 | 4 | 1 | 2 | 1 | 4 | 1 | 2 | 1 | 2 | 4 | 2 | 4 | 8 | 4 |
| 06 | 8 | 4 | 1 | 1 | 2 | 4 | 1 | 1 | 2 | 2 | 2 | 4 | 8 | 4 | 4 |
| 07 | 8 | 2 | 4 | 1 | 1 | 2 | 4 | 1 | 1 | 4 | 4 | 4 | 2 | 2 | 8 |
| 08 | 8 | 1 | 4 | 2 | 1 | 1 | 4 | 2 | 1 | 2 | 4 | 8 | 4 | 2 | 4 |
| 09 | 8 | 1 | 4 | 1 | 2 | 1 | 4 | 1 | 2 | 2 | 8 | 4 | 2 | 4 | 4 |
| 10 | 8 | 2 | 1 | 4 | 1 | 2 | 1 | 4 | 1 | 4 | 4 | 4 | 2 | 8 | 2 |
| 11 | 8 | 1 | 2 | 4 | 1 | 1 | 2 | 4 | 1 | 4 | 2 | 8 | 4 | 4 | 2 |
| 12 | 8 | 1 | 1 | 4 | 2 | 1 | 1 | 4 | 2 | 8 | 2 | 4 | 2 | 4 | 4 |
| 13 | 8 | 2 | 1 | 1 | 4 | 2 | 1 | 1 | 4 | 4 | 4 | 4 | 8 | 2 | 2 |
| 14 | 8 | 1 | 2 | 1 | 4 | 1 | 2 | 1 | 4 | 4 | 8 | 2 | 4 | 4 | 2 |
| 15 | 8 | 1 | 1 | 2 | 4 | 1 | 1 | 2 | 4 | 8 | 4 | 2 | 4 | 2 | 4 |
| 16 | 1 | 2 | 1 | 8 | 4 | 4 | 8 | 1 | 2 | 4 | 2 | 4 | 1 | 2 | 4 |
| 17 | 1 | 1 | 2 | 8 | 4 | 8 | 4 | 1 | 2 | 4 | 1 | 2 | 2 | 4 | 4 |
| 18 | 1 | 2 | 1 | 4 | 8 | 4 | 8 | 2 | 1 | 4 | 4 | 2 | 2 | 1 | 4 |
| 19 | 1 | 1 | 2 | 4 | 8 | 8 | 4 | 2 | 1 | 4 | 2 | 1 | 4 | 2 | 4 |

**Proposition 11.** *Let $\Gamma$ be a set of at most 8 taxa and denote $k = \lfloor \frac{n}{2} \rfloor$ and $\rho = n - 2k$. Then, for any distinct taxa $i_1, i_2, i_3, \ldots, i_n \in \Gamma$, the inequality*

$$\rho x_{i_1 i_n} + x_{i_1 i_2} + x_{i_3 i_4} + \cdots + x_{i_{n-\rho-1} i_{n-\rho}} = \rho x_{i_1 i_n} + \sum_{r=1}^{k} x_{i_{2r-1} i_{2r}} \geq 2^{n-k-3}(k + 2 - \rho) + \rho \qquad (57)$$

*is facet-defining for Conv($\mathcal{X}$).*

Tables 8–10 show three sets of vertices that define facets of Conv($\mathcal{X}$) and satisfy (55), (56) and (57) at equality, respectively, when assuming $n = 6$. We also observe that (57) are equivalent to Forcey et al. [21]'s intersecting cherry inequalities for $n = 5$. Moreover, enumerative analyses showed that (55) and (56) are not facet-defining for $n > 6$, whereas (57) are facet-defining also for $n = 7$ and $n = 8$.

To conclude this section, we summarize in Table 11 the families of known and new inequalities that are facet-defining or just valid for Conv($\mathcal{X}$) when assuming $n = 6$.

## 7. Characterizing and recognizing elements of $\mathcal{X}$

In this section, we investigate two fundamental issues concerning the set $\mathcal{X}$, namely the problem of consistently generating its elements and the problem of deciding whether a given matrix is an element of

**Table 10**
Vertices of Conv($\mathcal{X}$) that satisfy $x_{i_1i_2} + x_{i_3i_4} + x_{i_5i_6} \geq 5$ at equality when assuming $n = 6$.

| Vertices | Entries | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ | $x_{23}$ | $x_{24}$ | $x_{25}$ | $x_{26}$ | $x_{34}$ | $x_{35}$ | $x_{36}$ | $x_{45}$ | $x_{46}$ | $x_{56}$ |
| 01 | 2 | 8 | 1 | 4 | 1 | 2 | 4 | 4 | 4 | 1 | 4 | 1 | 2 | 8 | 2 |
| 02 | 1 | 8 | 2 | 4 | 1 | 1 | 4 | 2 | 8 | 2 | 4 | 1 | 4 | 4 | 2 |
| 03 | 2 | 8 | 1 | 1 | 4 | 2 | 4 | 4 | 4 | 1 | 1 | 4 | 8 | 2 | 2 |
| 04 | 1 | 8 | 2 | 1 | 4 | 1 | 4 | 8 | 2 | 2 | 1 | 4 | 4 | 4 | 2 |
| 05 | 2 | 1 | 8 | 4 | 1 | 4 | 2 | 4 | 4 | 1 | 2 | 8 | 4 | 1 | 2 |
| 06 | 1 | 2 | 8 | 4 | 1 | 4 | 1 | 2 | 8 | 2 | 4 | 4 | 4 | 1 | 2 |
| 07 | 2 | 1 | 8 | 1 | 4 | 4 | 2 | 4 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 08 | 1 | 2 | 8 | 1 | 4 | 4 | 1 | 8 | 2 | 2 | 4 | 4 | 1 | 4 | 2 |
| 09 | 2 | 4 | 1 | 8 | 1 | 4 | 4 | 2 | 4 | 2 | 4 | 2 | 1 | 8 | 1 |
| 10 | 1 | 4 | 1 | 8 | 2 | 2 | 8 | 1 | 4 | 2 | 4 | 4 | 1 | 4 | 2 |
| 11 | 2 | 1 | 4 | 8 | 1 | 4 | 4 | 2 | 4 | 2 | 1 | 8 | 4 | 2 | 1 |
| 12 | 1 | 1 | 4 | 8 | 2 | 8 | 2 | 1 | 4 | 2 | 1 | 4 | 4 | 4 | 2 |
| 13 | 2 | 4 | 1 | 1 | 8 | 4 | 4 | 4 | 2 | 2 | 2 | 4 | 8 | 1 | 1 |
| 14 | 1 | 4 | 1 | 2 | 8 | 2 | 8 | 4 | 1 | 2 | 4 | 4 | 4 | 1 | 2 |
| 15 | 2 | 1 | 4 | 1 | 8 | 4 | 4 | 4 | 2 | 2 | 8 | 1 | 2 | 4 | 1 |
| 16 | 1 | 1 | 4 | 2 | 8 | 8 | 2 | 4 | 1 | 2 | 4 | 1 | 4 | 4 | 2 |
| 17 | 2 | 2 | 4 | 4 | 4 | 8 | 1 | 4 | 1 | 1 | 4 | 1 | 2 | 8 | 2 |
| 18 | 2 | 2 | 4 | 4 | 4 | 8 | 1 | 1 | 4 | 1 | 1 | 4 | 8 | 2 | 2 |
| 19 | 2 | 4 | 2 | 4 | 4 | 1 | 8 | 4 | 1 | 1 | 2 | 8 | 4 | 1 | 2 |
| 20 | 2 | 4 | 2 | 4 | 4 | 1 | 8 | 1 | 4 | 1 | 8 | 2 | 1 | 4 | 2 |
| 21 | 2 | 4 | 4 | 2 | 4 | 4 | 1 | 8 | 1 | 2 | 4 | 2 | 1 | 8 | 1 |
| 22 | 2 | 4 | 4 | 2 | 4 | 1 | 4 | 8 | 1 | 2 | 1 | 8 | 4 | 2 | 1 |
| 23 | 2 | 4 | 4 | 4 | 2 | 4 | 1 | 1 | 8 | 2 | 2 | 4 | 8 | 1 | 1 |
| 24 | 2 | 4 | 4 | 4 | 2 | 1 | 4 | 1 | 8 | 2 | 8 | 1 | 2 | 4 | 1 |

**Table 11**
Summary of known and new valid inequalities for Conv($\mathcal{X}$) when assuming $n = 6$.

| Facet-defining inequalities | | Reference |
|---|---|---|
| $x_{ij} \geq 1$<br>$x_{ij} + x_{jk} - x_{ik} \leq 2^{n-3}$<br>$\sum_{\substack{i,j \in S_1 \\ i<j}} x_{ij} \leq (k-1)2^{n-3}$ | For all distinct $i, j \in \Gamma$<br>For all distinct $i, j, k \in \Gamma$<br>For any $S_1, S_2 \subset \Gamma$ :<br>$S_1 \cap S_2 = \emptyset$, and $\lvert S_1 \rvert = 3$ | Forcey et al. [21](see also Section 6.1) |
| $2^{n-5}x_{i_1i_2} + 2^{n-5}x_{i_1i_3} + x_{i_2i_3} \geq 5 \cdot 2^{(n-5)}$<br>$2^{n-4}x_{i_1i_2} + 2^{n-4}x_{i_1i_3} + x_{i_2i_3} \geq 2^{(n-2)}$ | For all distinct $i_1, i_2, i_3 \in \Gamma$ | Proposition 7 - (48)<br>Proposition 7 - (49) |
| $2^{n-4}x_{i_1i_2} - x_{i_3i_4} \geq 0$<br>$x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} + x_{i_4i_5} \geq (4+\rho)2^{k-1}$ | For all distinct $i_1, i_2, i_3, i_4 \in \Gamma$<br>For all distinct $i_1, i_2, i_3, i_4, i_5 \in \Gamma$<br>$k = \lfloor \frac{n}{3} \rfloor$, $\rho = n - 3k$ | Proposition 8 - (51)<br>Proposition 8 - (52) |
| $2x_{i_1i_2} + 2x_{i_3i_4} + x_{i_5i_6} \geq 8$<br>$x_{i_1i_2} + x_{i_3i_4} - 2^{n-4}x_{i_5i_6} \leq 2^{n-4}$ | For all distinct $i_1, i_2, i_3, i_4, i_5, i_6 \in \Gamma$ | Proposition 9 - (53)<br>Proposition 9 - (54) |
| $2x_{i_1i_2} + x_{i_2i_3} + x_{i_3i_4} + x_{i_4i_5} + x_{i_5i_6} + x_{i_1i_6} + x_{i_2i_4} \geq 16$<br>$-2x_{i_1i_2} + 7x_{i_1i_3} + 7x_{i_1i_4} + 7x_{i_2i_5} + 7x_{i_2i_6} \geq 40$ | For all distinct $i_1, i_2, i_3, i_4, i_5, i_6 \in \Gamma$ | Proposition 10 - (55)<br>Proposition 10 - (56) |
| $\rho x_{i_1i_n} + \sum_{r=1}^{k} x_{i_{2r-1}i_{2r}} \geq 2^{n-k-3}(k+2-\rho) + \rho$ | For all distinct $i_1, i_2, \ldots, i_n \in \Gamma$<br>$k = \lfloor \frac{n}{2} \rfloor$, $\rho = n - 2k$ | Proposition 11 - (57) |
| Valid but not facet-defining inequalities | | |
| $x_{i_1i_2} + x_{i_1i_3} + x_{i_2i_3} \geq (3+\rho)2^{k-1}$ | $k = \lfloor \frac{n}{3} \rfloor$, $\rho = n - 3k$ | Proposition 7 - (50) |

$\mathcal{X}$ (hence, a vertex of Conv($\mathcal{X}$)). We present a set of necessary and sufficient conditions to accomplish the first task and we show how to translate such conditions into a set of nonlinear constraints. Moreover, we present a polynomial-time oracle to decide whether a given $n \times n$ symmetric matrix having generic entry $2^\alpha$, for some $\alpha \in \{0, \ldots, n-1\}$ is an element of $\mathcal{X}$ (hence encoding a phylogeny of $\Gamma$). Before proceeding, we introduce a number of definitions that will prove useful throughout the section.

Given a set $\Gamma$ of $n$ taxa and a subset $S \subseteq \Gamma$, $|S| \geq 3$, we define a *partial phylogeny of* $\Gamma$ as any phylogeny of $S$. We denote $S_k$ as a subset of $\Gamma$ such that $|S| = k$. By analogy, we denote $T_k$ as a partial phylogeny of $\Gamma$ having $S_k$ as leafset. Given a subset $S_k \subset \Gamma$, the corresponding partial phylogeny $T_k$, and a taxon $i$ in $T_k$, we say that we *insert* a taxon $j \in \Gamma \setminus S_k$ in taxon $i$ when we generate from $T_k$ a new partial phylogeny

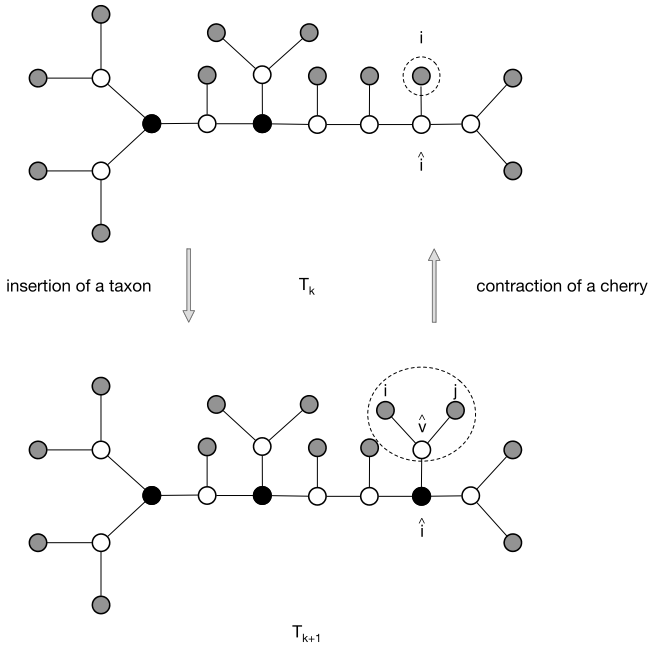



**Fig. 9.** The down arrow shows an example of an insertion of a taxon $j$ in the edge $(\hat{\imath}, i)$. Vice versa, the up arrow shows a contraction of a cherry with leaves $i$ and $j$ into a leaf $i$.

$T_{k+1}$ of $\Gamma$ by replacing taxon $i$ in $T_k$ with an internal vertex, say $\hat{v}$, and by adding to $T_k$ the edges $(\hat{v}, i)$ and $(\hat{v}, j)$. Fig. 9 shows an example of an insertion of a taxon on a (partial) phylogeny with $k$ leaves.

### *7.1. Generating elements of $\mathcal{X}$*

We recall the following observation from Catanzaro et al. [3]:

**Observation 2.** *All of the phylogenies of $\Gamma$ can be generated by using a constructive procedure based on the following steps: (i) construct the subset $S_3$ by selecting the first three taxa in $\Gamma$; (ii) construct the partial phylogeny $T_3$; and (iii) iteratively insert, one at a time, each of the remaining taxa in $\Gamma \setminus S_3$ in each taxon of the successive partial phylogenies $T_k$, $3 \leq k \leq n-1$, until a phylogeny of $\Gamma$ is obtained.*

Observation 2 proves useful to characterize any matrix $X \in \mathcal{X}$. In particular, it is worth noting that a given $n \times n$ symmetric matrix having generic entry $2^\alpha$, for some $\alpha \in \{0, \ldots, n-1\}$, encodes a phylogeny of $\Gamma$ if and only if we can characterize it in terms of a sequence of insertions of taxa in a partial phylogeny. To this end, consider a subset $S_k$ of $\Gamma$, $k \geq 3$, the corresponding partial phylogeny $T_k$, and a $n \times n$ symmetric matrix $X(k)$ whose generic entries are defined as follows:

$$x_{rs}(k) = \begin{cases} 2^{k-1} & \text{if } r, s \in \Gamma,\ r = s; \\ 2^{k-1-\tau_{rs}(k)} & \text{if } r, s \in S_k,\ r \neq s; \\ 0 & \text{otherwise} \end{cases} \quad \forall\ r, s \in \Gamma \tag{58}$$

where $\tau_{rs}(k)$ is the topological distance between taxa $r$ and $s$ on the partial phylogeny $T_k$. Then, the following proposition holds:

**Proposition 12.** *Let $T_{k+1}$ be the partial phylogeny obtained from $T_k$ by inserting a taxon $j \in \Gamma \setminus S_k$ in taxon $i$ of $T_k$. Then, the following relations exist between the entries of matrices $X(k)$ and $X(k+1)$:*

$$x_{rs}(k+1) = \begin{cases} 2^k & \text{if } r = s; \\ 2^{k-2} & \text{if } (r = i \text{ and } s = j) \text{ or } (r = j \text{ and } s = i); \\ x_{is}(k) & \text{if } r = j \text{ and } s \notin \{i, j\}; \\ x_{ri}(k) & \text{if } r \notin \{i, j\} \text{ and } s = j; \\ x_{rs}(k) & \text{if } r \neq s \text{ and } (r = i \text{ or } s = i); \\ 2x_{rs}(k) & \text{otherwise} \end{cases} \qquad \forall\, r, s \in \Gamma : r < s \tag{59}$$

**Proof.** Conditions (59) derive from the following facts:

1. $\tau_{rr}(k+1) = 0$ on $T_{k+1}$, hence $x_{rr}(k+1) = 2^k$;
2. $\tau_{ij}(k+1) = 2$ on $T_{k+1}$, hence $x_{ij}(k+1) = x_{ji}(k+1) = 2^{k-2}$;
3. $\tau_{is}(k+1) = \tau_{js}(k+1) = 2\tau_{is}(k)$ on $T_{k+1}$, for $s \notin \{i, j\}$, hence $x_{is}(k+1) = x_{js}(k+1) = x_{is}(k)$ and $x_{si}(k+1) = x_{sj}(k+1) = x_{si}(k)$;
4. $\tau_{rs}(k+1) = \tau_{rs}(k)$ on $T_{k+1}$, for $r, s \notin \{i, j\}$, hence $x_{rs}(k+1) = 2x_{rs}(k+1)$. □

Now, consider a phylogeny $T$ of $\Gamma$, the associated matrix $X$, the subset $S_3$ constituted by the first three taxa in $\Gamma$, and an arbitrary $n \times n$ symmetric matrix $X(3)$. In particular, assume, without loss of generality, that

$$x_{rs}(3) = \begin{cases} 4 & \text{if } r, s \in \Gamma,\ r = s; \\ 1 & \text{if } r, s \in S_3,\ r \neq s; \\ 0 & \text{otherwise} \end{cases} \qquad \forall r, s \in \Gamma. \tag{60}$$

Then, in the light of Observation 2 and Proposition 12, we can define a set of necessary and sufficient nonlinear relationships between matrices $X(3)$ and $X$ that allows us to describe how we can iteratively construct the phylogeny $T$ encoded by $X$ starting from the knowledge of the partial phylogeny $T_3$. Specifically, for all $i, j \in \Gamma$ and $k \in \mathbb{N}$, $3 \leq k \leq n-1$, let $y_{ij}(k)$ be a decision variable defined as follows:

$$y_{ij}(k) = \begin{cases} 1 & \text{if phylogeny } T_{k+1} \text{ is obtained from phylogeny } T_k \text{ by inserting } j \text{ in the leaf of } i; \\ 0 & \text{otherwise.} \end{cases} \tag{61}$$

Then, for all $k \in \mathbb{N}$, $3 \leq k \leq n-1$, the following conditions hold:

$$x_{rr}(k+1) = 2^k, \qquad \forall\, r \in \Gamma \tag{62}$$

$$\begin{aligned} x_{rs}(k+1) =& 2^{k-2}(y_{rs}(k) + y_{sr}(k)) + x_{rs}(k) \left( \sum_{j \in \Gamma, j \neq r,s} y_{rj}(k) + \sum_{i \in \Gamma, i \neq r,s} y_{is}(k) \right) \\ &+ 2x_{rs}(k) \left( 1 - y_{rs}(k) - y_{sr}(k) - \sum_{j \in \Gamma, j \neq r,s} y_{rj}(k) - \sum_{i \in \Gamma, i \neq r,s} y_{is}(k) \right) \\ & \forall\, r \in \Gamma,\ s \in \Gamma \setminus S_3,\ r \neq s \end{aligned} \tag{63}$$

$$\sum_{k=3}^{n-1} \sum_{r \in \Gamma} y_{rs}(k) = 1 \qquad \forall\, s \in \Gamma \setminus S_3 \tag{64}$$

$$\sum_{r \in \Gamma} \sum_{s \in \Gamma \setminus S_3} y_{rs}(k) = 1 \tag{65}$$

$$0 \leq y_{rs}(k) \leq \sum_{l=3}^{k-1} \sum_{p \in \Gamma} y_{pr}(l) \qquad \forall\, r, s \in \Gamma \setminus S_3 \tag{66}$$

**Algorithm 1:** A membership oracle for the set $\mathcal{X}$

**Input** : A candidate matrix $X = \{x_{ij}\}$
**Output** : True if $X \in \mathcal{X}$, False otherwise
**Data**: $X'$: a $(2n-1) \times (2n-1)$ symmetric matrix used to support computation
ProcessedTaxa: a set of positive integers
InternalVertex: a positive integer

Set $x'_{ij} = x_{ij}$, for all $i, j \in \{1, \dots, n\}$, and $x'_{ij} = 0$ otherwise
Set ProcessedTaxa $= \{1, \dots, n\}$
Set InternalVertex $= n + 1$
**while** |ProcessedTaxa| $> 3$ **do**
  Find a pair of indices $\hat{i}$ and $\hat{j}$ such that $x'_{\hat{i},\hat{j}} = 2^{|\text{ProcessedTaxa}|-3}$; if such pair does not exists **return** False
  **for** *all* $k$ *in* ProcessedTaxa **do**
    $x'_{\text{InternalVertex},k} = x'_{\hat{i},k}$
    $x'_{k,\text{InternalVertex}} = x'_{\text{InternalVertex},k}$
  **for** *all* $i, j$ *in* ProcessedTaxa *such that* $i < j$ **do**
    $x'_{i,j} = x'_{i,j}/2$
    $x'_{j,i} = x'_{i,j}$
  ProcessedTaxa $= (\text{ProcessedTaxa} \setminus \{\hat{i}, \hat{j}\}) \cup \{\text{InternalVertex}\}$
  InternalVertex $=$ InternalVertex $+ 1$
  **for** *all* $i$ *in* ProcessedTaxa **do**
    **if** $\sum_{\substack{j \in \text{ProcessedTaxa} \\ j \neq i}} x'_{ij} \neq 2^{|\text{ProcessedTaxa}|-2}$ **then**
      **return** False
**if** $x'_{ij} = 1$, *for all* $i, j$ *in* ProcessedTaxa *such that* $i \neq j$ **then**
  **return** True
**else**
  **return** False

$$y_{rs}(k) \in \{0, 1\} \quad \forall\, r \in \Gamma, s \in \Gamma \setminus S_3. \tag{67}$$

Constraints (62)–(63) are equivalent to conditions (59). Constraints (64) impose that each taxa $i \in \Gamma$ must belong to $T_n = T$. Constraint (65) imposes that a taxon $i \in \Gamma \setminus S_3$ can be inserted only one time in the successive partial phylogenies. Constraints (66) impose that a taxon $s \in \Gamma \setminus S_3$ can be inserted in taxon $r$ of a partial phylogeny $T_k$, only if $r$ belongs to $T_k$.

If appropriately exploited, the above conditions could inspire implicit enumeration algorithms based on iterative insertions on a partial phylogeny such as the ones described in [3,10].

### *7.2. Recognizing elements of $\mathcal{X}$*

Consider a $n \times n$ symmetric matrix having diagonal entries $x_{ii} = 2^{n-1}$, for $i \in \{1, \dots, n\}$, and generic non-diagonal entry $x_{ij} = 2^{\alpha}$, for some $\alpha \in \{0, \dots, n-3\}$, in the following referred to as *candidate matrix*. A question that may naturally arise is whether this candidate matrix is or not an element of $\mathcal{X}$ (hence, a vertex of Conv($\mathcal{X}$)). In this subsection we address this problem and we present a possible recognition algorithm able to provide in polynomial-time a yes/no answer to this question. Before proceeding, we recall that a *full binary tree* is a tree in which every vertex has either zero or two children [29]. Moreover, because the recognition problem inquires whether a given candidate matrix is encoding a phylogeny of $\Gamma$ or not, we introduce the following equivalent recursive definition of a phylogeny that will prove useful to our ends: a phylogeny is either a single vertex or an acyclic graph constituted by three full binary trees, a new vertex, and three edges connecting the new vertex to the root of each full binary tree.

Consider a phylogeny $T$ with $n$ leaves. We define a *contraction* as a topological operation on $T$ that consists in (i) removing a cherry, say $i$ and $j$, from $T$ together with the edges connecting the two leaves with their immediate ancestor, say $\hat{v}$, and (ii) replacing $\hat{v}$ with leaf $i$. Fig. 9 shows an example of a contraction on a given (partial) phylogeny with $k+1$ leaves. We observe that the following proposition holds:

**Proposition 13.** *A contraction is a topological operation that preserves the degree constraint on the internal vertices of a given phylogeny $T$ with $n+1$ leaves, i.e., it operates on $T$ so as to provide a new phylogeny with $n$ leaves.*

**Proof.** Consider any cherry of $T$ on which we want to perform a contraction. Denote the cherry leaves by $i$ and $j$; let $\hat{v}$ be their immediate ancestor, and let $\hat{i}$ the ancestor of $\hat{v}$. Then, the statement trivially follows by observing that (i) the graph resulting from a contraction on $T$ is still a tree (indeed, after a contraction two edges of $T$ are removed and no new edge is added), and (ii) such a graph still satisfies the recursive definition of a phylogeny because after a contraction one of the three full binary trees adjacent to $\hat{i}$, namely the one rooted in $\hat{v}$, is replaced with a new full binary tree constituted by the leaf $i$. □

The recursive definition of a phylogeny, together with the definition of a contraction, suggest the following iterative algorithm to decide whether a candidate matrix $X$ is or not an element of $\mathcal{X}$:

1. if $X$ is a $3 \times 3$ matrix then

    1.1 if such a matrix maps a star tree with three leaves then $X \in \mathcal{X}$; return yes;
    1.2 otherwise return false;

2. otherwise

    2.1 find a cherry in $X$; if no cherry is found, then $X \notin \mathcal{X}$; return false;
    2.2 perform a contraction of such a cherry, and let $\bar{X}$ the resulting candidate matrix;
    2.3 check if $\bar{X}$ satisfies Kraft equalities; if there are violations then $X \notin \mathcal{X}$; return false;
    2.4 set $X = \bar{X}$ and go to step 1.

Algorithm 1 shows the pseudo code implementing the above iterative algorithm. Specifically, Algorithm 1 takes as input a candidate matrix $X$ and returns as an output a boolean equal to true if $X \in \mathcal{X}$, and false otherwise. The algorithm uses as a support three data structures, namely a $(2n-1) \times (2n-1)$ symmetric matrix called $X'$ used to store the sequence of contractions, a set of positive integer called `ProcessedTaxa` used to track the leaves that must be still processed, and a temporary integer called `InternalVertex`. The first three lines of Algorithm 1 initialize the support data structures. Specifically, at line 1 the entries of $X$ are copied into the corresponding ones of $X'$, while the remaining entries of $X'$ are set to 0. At line 2 `ProcessedTaxa` is set to $\{1, \dots, n\}$ and at line 3 `InternalVertex` is set to $n+1$. The subsequent lines implement the sequence of contractions described above. In particular, lines 4–16 take into account the case in which the phylogeny encoded by $X'$ has at least four leaves, while lines 17–20 take into account the case in which the leaves are 3. More in detail, line 5 takes care of identifying a cherry; lines 6–13 perform a contraction, by appropriately storing the entries of matrix $\bar{X}$ into $X'$; and lines 14–16 check if the contracted matrix $\bar{X}$ satisfies Kraft equalities. Finally, lines 17–20 check whether the final $3 \times 3$ matrix $\bar{X}$ maps a star tree. It is easy to realize that the computational complexity of Algorithm 1 is $O(n^3)$. As verifying conditions (4)–(9) takes $O(n^3)$, the algorithm is optimal.

**Acknowledgments**

The authors are very grateful to the reviewers for their careful and meticulous reading of the article as well as their valuable comments. The authors are also in debt with Dr. Roberto Ronco for helpful insights about the structure of the matrix (11). The first author acknowledges support from the Belgian National Fund for Scientific Research (FRS-FNRS) via the research grant "Crédit de Recherche" ref. S/25-MCF/OL J.0026.17; the Université Catholique de Louvain, Belgium via the "Fonds Spéciaux de Recherche" (FSR) 2017–2021; and the Fondation Louvain via the research grant COALESCENS of the funding program "Le

numérique au service de l'humain". Part of this work has been developed when D. Catanzaro was Invited Professor at the Department of Management of University Ca Foscari of Venice, Italy, during the academic year 2018.

## References

[1] J. Felsenstein, Inferring Phylogenies, Sinauer Associates, Sunderland, MA, 2004.
[2] D. Catanzaro, Estimating phylogenies from molecular data, in: R. Bruni (Ed.), Mathematical Approaches To Polymer Sequence Analysis and Related Problems, Springer, NY, 2011, pp. 149–176.
[3] D. Catanzaro, M. Labbé, R. Pesenti, J.J. Salazar-Gonzáles, The balanced minimum evolution problem, INFORMS J. Comput. 24 (2) (2012) 276–294.
[4] A.H.G. Rinnooy Kan, D.S. Johnson, J.K. Lenstra, The complexity of the network design problem, Networks 8 (1978) 279–285.
[5] Y. Pauplin, Direct calculation of a tree length using a distance matrix, J. Mol. Evol. 51 (2000) 41–47.
[6] R. Desper, O. Gascuel, Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting, Mol. Biol. Evol. 21 (3) (2004) 587–598.
[7] O. Gascuel, M. Steel, Neighbor-joining revealed, Mol. Biol. Evol. 23 (11) (2006) 1997–2000.
[8] S. Fiorini, G. Joret, Approximating the balanced minimum evolution problem, Oper. Res. Lett. 40 (1) (2012) 31–35.
[9] R. Aringhieri, D. Catanzaro, M. Di Summa, Optimal solutions for the balanced minimum evolution problem, Comput. Oper. Res. 38 (2011) 1845–1854.
[10] F. Pardi, Algorithms on Phylogenetic Trees (Ph.D. thesis), University of Cambridge, UK, 2009.
[11] D. Catanzaro, The minimum evolution problem: Overview and classification, Networks 53 (2) (2009) 112–125.
[12] O. Gascuel, Mathematics of Evolution and Phylogeny, Oxford University Press, New York, 2005.
[13] D.A. Huffman, A method for the construction of minimum redundancy codes, in: Proceedings of the IRE, 1952.
[14] D.S. Parker, P. Ram, The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees, SIAM J. Comput. 28 (5) (1996) 1875–1905.
[15] K. Sayood, Introduction to Data Compression, fifth ed., Morgan Kaufmann, San Francisco, CA, 2017.
[16] K. Eickmeyer, P. Huggins, L. Pachter, R. Yoshida, On the optimality of the neighbor-joining algorithm, Algorithms Mol. Biol. 3 (5) (2008) 1–11.
[17] D.C. Haws, T.L. Hodge, R. Yoshida, Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope, Bull. Math. Biol. 73 (11) (2011) 2627–2648.
[18] S. Forcey, L. Keefe, W. Sands, Facets of the balanced minimal evolution polytope, J. Math. Biol. 73 (2) (2016) 447–468.
[19] E. Gawrilow, M. Joswig, Polymake: A framework for analyzing convex polytopes, in: G. Kalai, G.M. Ziegler (Eds.), Polytopes - Combinatorics and Computation, Birkhäuser, 2000, pp. 43–74.
[20] L.J. Billera, S.P. Holmes, K. Vogtmann, Geometry of the space of phylogenetic trees, Adv. Appl. Math. 27 (4) (2001) 733–767.
[21] S. Forcey, L. Keefe, W. Sands, Split-facets for balanced minimal evolution polytopes and the permutoassociahedron, Bull. Math. Biol. 79 (5) (2017) 975–994.
[22] M.M. Kapranov, The permutoassociahedron, Mac Lane's coherence theorem and asymptotic zones for the KZ equation, J. Pure Appl. Algebra 85 (2) (1993) 119–142.
[23] V. Reiner, G.M. Ziegler, Coxeter-associahedra, Technical Report SC-93-11, Zuse Institute Berlin (ZIB), Berlin, Takustrasse 7, 14195, Germany, 1993.
[24] P. Buneman, The recovery of trees from measure of dissimilarities, in: F.R. Hodson, D.G. Kendall, P. Tautu (Eds.), Archaeological and Historical Science, Edinburgh University Press, Edinburgh, UK, 1971, pp. 387–395.
[25] P.L. Erdös, M. Steel, L.A. Székély, T. Warnow, A few logs suffice to build (almost) all trees: Part I, Random Struct. Algorithms 14 (2) (1999) 153–184.
[26] M.S. Waterman, T.F. Smith, M. Singh, W.A. Beyer, Additive evolutionary trees, J. Theoret. Biol. 64 (1977) 199–213.
[27] G.L. Nemhauser, L.A. Wolsey, Integer and Combinatorial Optimization, Wiley-Interscience, New York, 1999.
[28] Wolfram Research, Inc. Mathematica, Version 11.3. Champaign, IL, 2018.
[29] D. Reinhard, Graph Theory, Springer-Verlag, New York, 2005.