OPTIMIZING OVER PATH-LENGTH MATRICES OF UNROOTED BINARY TREES

Daniele Catanzaro, Raffaele Pesenti, Allan Sapucaia, Laurence Wolsey

REPRINT | 3321



UCLouvain

CORE | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email:lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/core/core-reprints

Mathematical Programming A manuscript No.
(will be inserted by the editor)

Optimizing over Path-Length Matrices of Unrooted Binary Trees

Daniele Catanzaro · Raffaele Pesenti · Allan Sapucaia · Laurence Wolsey

Received: date / Accepted: date

Abstract *Path-Length Matrices* (PLMs) form a tree encoding scheme that is often used in the context of optimization problems defined over *Unrooted Binary Trees* (UBTs). Determining the conditions that a symmetric integer matrix of order $n \geq 3$ must satisfy to encode the PLM of a UBT with n leaves is central to these applications. Here, we show that a certain subset of known necessary conditions is also sufficient to characterize the set Θ_n of PLMs induced by the set of UBTs with n leaves. We also identify key polyhedral results as well as facet-defining and valid inequalities for the convex hull of these matrices. We then examine how these results can be used to develop a new Branch-&-Cut algorithm for the *Balanced Minimum Evolution Problem* (BMEP), a $\mathcal{NP}$-hard nonlinear optimization problem over Θ_n much studied in the literature on molecular phylogenetics. We present a new valid integer linear programming formulation for this problem and we show how to refine it, by removing redundant variables and constraints. We also show how to strengthen the reduced formulation by including lifted versions of the previously identified facet-defining and valid inequalities. We provide separation oracles for these inequalities and we exploit the tight lower bound provided by the linear programming relaxation of this formulation to design a new primal heuristic for the BMEP. The results of extensive computational experiments show that the Branch-&-Cut algorithm derived from this study outperforms the current state-of-the-art exact solution algorithm for the problem.

Keywords combinatorial optimization · integer programming · polyhedral combinatorics · branch-&-cut · path-length matrices · unrooted binary trees · tree realization · Kraft's equality · Buneman's four-point conditions · balanced minimum evolution · distance methods · phylogenetics.

PACS 90C10, 90C11, 90C27

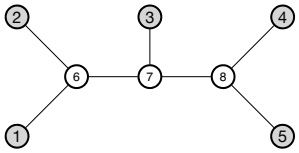
1 Introduction

An *Unrooted Binary Tree* (UBT) is a tree graph that has $n \geq 3$ leaves, $n-2$ internal nodes, each of degree 3, and $2n-3$ edges. UBTs can be encoded in many ways [8,17]. The encoding that we consider here is based on *Path-Length Matrices* (PLMs) [16]. Specifically, let $\mathcal{U}_n$ denote the set of UBTs with n leaves and let P_{ij}^T denote the unique path in a UBT $T \in \mathcal{U}_n$ between leaves i and j. Then, the PLM $\tau(T)$ of T is a symmetric integer matrix of order n whose generic entry $\tau(T)_{ij} \in \{2, \ldots, n-1\}$ represents the number of edges on P_{ij}^T. Figure 1 provides examples of a UBT and its induced PLM.

Characterizing the conditions that a symmetric integer matrix τ of order n must satisfy in order to belong to the set Θ_n of the PLMs induced by the UBTs in $\mathcal{U}_n$ is central in several applications and

D. Catanzaro, A. Sapucaia, and L. Wolsey
Center for Operations Research and Econometrics, Université Catholique de Louvain, Voie du Roman Pays 34, 1348, Louvain-la-Neuve, Belgium.
Emails: daniele.catanzaro@uclovain.be; allan.sapucaia@uclouvain.be; laurence.wolsey@uclouvain.be

R. Pesenti
Department of Management, University Ca' Foscari, San Giobbe, Cannaregio 837, Venice I-30121, Italy.
Email: pesenti@unive.it



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

Fig. 1 An example of an Unrooted Binary Tree (UBT) with 5 leaves and the induced Path-Length Matrix (PLM) τ.

optimization contexts. Previous research efforts have identified a number of necessary conditions for belonging to Θ_n [19]. Here, we show that a certain subset of these known necessary conditions is also sufficient. We achieve this result by strengthening Buneman's results on the *Tree-Realization Problem* (TRP) [7], recalled in Section 2 for the sake of completeness. We extend our study to $Conv(\Theta_n)$, the convex hull of Θ_n, by presenting several polyhedral results and families of facet-defining and valid inequalities. We then investigate how these results can be used to develop a new Branch-&-Cut algorithm for the *Balanced Minimum Evolution Problem* (BMEP), a nonlinear $\mathcal{NP}$-hard problem over Θ_n much studied in the literature on molecular phylogenetics [12]. Given a symmetric distance matrix $\mathbf{D} = \{d_{ij}\}$ of order n with a null diagonal and positive entries elsewhere, the BMEP is the problem

$$\min_{\tau \in \Theta_n} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} d_{ij} 2^{-\tau_{ij}}. \tag{1}$$

Based on our characterization of Θ_n and working in an extended space that enables the linearization of the objective function of the problem, we derive a new valid integer linear programming formulation for the BMEP. We refine this formulation, by removing redundant variables and constraints, and we strengthen it by including lifted versions of the previously identified families of facet-defining and valid inequalities for $Conv(\Theta_n)$. We provide separation oracles for these inequalities and we exploit the tight lower bound provided by the linear programming relaxation of this formulation to design a new primal heuristic for the problem. The results of extensive computational experiments show that the Branch-&-Cut algorithm derived from this study outperforms the current state-of-the-art exact solution algorithm for the BMEP recently presented in [13].

This article is organized as follows. In the remainder of this section, we provide some background and pointers to relevant earlier works on the BMEP. In Section 2, we introduce the notation and we briefly recall known results on trees that will be later used in several proofs. These results include, among others, Buneman's results on the TRP and Kraft's equality for UBTs. In Section 3, we present the sufficient conditions that characterize Θ_n. In Section 4, we refine this characterization, by reducing by a factor of n the number of conditions that a symmetric integer matrix τ must satisfy to belong to Θ_n. In Section 5, we study $Conv(\Theta_n)$. In Section 6, we focus on the BMEP and we present a new valid Integer Linear Programming (ILP) formulation for the problem based on our characterization of Θ_n. In Section 7, we refine and strengthen this formulation by obtaining a final ILP formulation for the BMEP that does not directly rely on Buneman's results on the TRP. In Section 8, we study the computational performance of this ILP formulation and in Section 9 we discuss some open questions that may inspire future developments.

Some background on the BMEP

A *phylogeny* is a weighted tree that represents evolutionary relationships among species, commonly referred to as *taxa*. Its structure and interpretation can vary based on the application or context. In systematics, particularly relevant here, phylogenies are typically UBTs, where leaves represent taxa, internal nodes indicate speciation events, edges show evolutionary relationships, and edge weights are measures of the similarity between vertices [12]. Since the evolutionary process of a set Γ of $n \geq 3$ taxa is in general unobservable as a whole, the phylogeny of Γ is usually estimated from observed traits (e.g., molecular data extracted from taxa, such as DNA, RNA, codon sequences, or genomes) by solving a network design problem which is usually highly nonlinear and $\mathcal{NP}$-hard in almost all of its versions described in the literature [10,12,31]. The BMEP is one possible version of such a problem (see [12] for a comprehensive

Taxa	Sequence		
$Taxon\ 1$	A	A	A
$Taxon\ 2$	A	C	C
$Taxon\ 3$	C	G	C
$Taxon\ 4$	C	C	G
$Taxon\ 5$	G	A	G

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2 & 3 & 3 & 2 \\ 2 & 0 & 2 & 2 & 3 \\ 3 & 2 & 0 & 2 & 3 \\ 3 & 2 & 2 & 0 & 2 \\ 2 & 3 & 3 & 2 & 0 \end{bmatrix} \end{array}$$

Fig. 2 A simple example of an instance of the BMEP characterized by a distance matrix $\mathbf{D}$ of order 5 obtained by computing the Hamming distances between the pairs of DNA sequences shown on the left. The PLM shown on the right of Figure 1 is optimal for this instance.

introduction on the problem).=Introduced by Desper and Gascuel [21] in 2004 and based on Pauplin's seminal work [45], the BMEP estimates the phylogeny of Γ based on similarity measures (i.e., *distances*) between pairs of taxa [14,18,23,42,49]. Figure 2 illustrates a simple example of an instance of the BMEP consisting of a distance matrix of order 5 obtained by computing the Hamming distances for a fictitious set of DNA sequences. The optimal solution τ^* to this instance and the associated UBT are shown in Figure 1.

The BMEP is $\mathcal{NP}$-hard and inapproximable within c^n, for some constant $c > 1$, unless $\mathcal{P} = \mathcal{NP}$ [24]. This negative result holds even in the case where the structure of the UBT is fixed and only the assignment of the n taxa to the n leaves of the UBT needs to be identified [28].

Several authors investigated the combinatorics of the problem, with the earliest attempts dating back to the works of Eickmeyer et al. [22] and Haws et al. [35]. The former focused on studying basic combinatorial properties of path-lengths and the latter on identifying some faces of $Conv(\Theta_n)$. Forcey et al. [25, 26] studied the polyhedral combinatorics of the BMEP in a space in which its objective function can be linearized. This space, called $\mathcal{X}$-space in [19], consists of the matrices $\mathbf{X}$ of order n associated with the UBTs in $\mathcal{U}_n$ and having as generic entry $x_{ij} = 2^{n-1-\tau_{ij}}$. The authors highlighted the existence of connections between the BMEP polytope in the $\mathcal{X}$-space and the *permutoassociahedron* [3,20,26,37,47] and succeeded in (i) describing the entire BMEP polytope for $n \leq 5$ and (ii) identifying some of its facets for $n = 6$. More recently, Catanzaro et al. [19] extended Forcey et al.'s study, by presenting new facet-defining and valid inequalities for the BMEP polytope in the $\mathcal{X}$-space and by addressing the problem of recognizing its vertices in polynomial-time. Unfortunately, the lack of a characterization of the matrices $\mathbf{X}$ limited the use of the results of Forcey et al. [25,26] and Catanzaro et al. [19] in the context of implicit enumeration solution approaches to the BMEP.

On the algorithmic front, Pardi [43] first proposed in 2007 a Branch-&-Bound algorithm for the problem based on a set of combinatorial lower bounds, which proved to be fast to compute and tight enough to consistently solve benchmark instances of the BMEP (namely, the Easy-Set instances discussed in Section 8.1) up to $n = 15$, within the time limit of one hour, and under the same hardware, computational settings and environment described in Section 8.2 (hereafter assumed as reference).

Aringhieri et al. [2] proposed in 2011 an alternative exact solution algorithm for the BMEP that exploited the tree-isomorphism between UBTs and proved capable of consistently solving benchmark instances of the BMEP up to $n = 15$ within one hour. The algorithm enumerated by brute-force of all the possible non-isomorphic unlabeled UBTs with n leaves and, for each of them, solved a quadratic problem consisting in finding an assignment of taxa to the leaves of a fixed non-isomorphic unlabeled UBT so as to minimize (1). The enumeration of the PLMs induced by the unlabeled UBTs with n leaves was independent of the specific instance of the BMEP, could be always carried out offline, and had a runtime of less than a minute up to $n = 27$. Moreover, the resolution of the assignment problem for each unlabeled UBT could be processed independently of the others, thereby enabling a natural parallelization of the algorithm.

In 2012, Catanzaro et al. [16] identified a set of necessary conditions that any symmetric integer matrix of order n must satisfy to belong to Θ_n. Based on these conditions, the authors developed the earliest valid ILP formulation for the BMEP. While this formulation did not outperform Pardi's Branch-&-Bound algorithm, the lower bounds derived from solving its linear programming relaxation proved to be significantly tighter than Pardi's combinatorial ones. This fact motivated the authors to mix Pardi's implicit enumeration strategy with these lower bounds, thereby deriving a Branch-&-Bound algorithm for the BMEP capable of consistently solving benchmark instances up to $n = 17$ within one hour.

In 2017, Fortz et al. [27] observed that Catanzaro et al.'s ILP formulation could be tightened by replacing the edge variables encoding the structure of a UBT with flow variables and conservation con-

Symbol	Legend
UBT	Unrooted Binary tree
$\mathcal{U}_n$	The set of UBTs with $n \geq 3$ leaves
P_{ij}^T	The unique path in a UBT $T \in \mathcal{U}_n$ between nodes i and j
$\tau(T)_{ij}$	The number of edges (or *path-length*) in the (unique) path P_{ij}^T in a UBT $T \in \mathcal{U}_n$ between nodes i and j
PLM	Path-Length Matrix, i.e., a symmetric integer matrix of order n having $\tau(T)_{ij}$ as generic entry, for some $T \in \mathcal{U}_n$
Θ_n	The set of PLMs induced by the UBTs in $\mathcal{U}_n$
C_T	The set of the circular orderings on a UBT $T \in \mathcal{U}_n$
$[\alpha, \beta]$	The discrete interval $\{\alpha, \alpha+1, \ldots, \beta\}$ included between the two integers α and β
$[\alpha]$	The discrete interval $\{1, 2, \ldots, \alpha\}$, for some integer $\alpha \geq 1$
$G([\alpha])$	The complete graph having the set $[\alpha]$ as set of nodes

Table 1 Summary of the basic notation used in the article.

straints. However, the solution times of Fortz et al.'s formulation proved to be very slow, being unable to solve benchmark instances of order $n \geq 12$ within one hour.

In 2023, Catanzaro et al. [13] presented an improved version of the Branch-&-Bound algorithm developed in [16], including both algorithmic refinements and a massive parallelism, which, under the same computational settings and environment, allowed to consistently solve benchmark instances of order $n = 22$ within one hour. The results on the characterization of Θ_n presented in the next sections enable the development of an ILP formulation for the BMEP that outperforms Catanzaro et al.'s algorithm and allow, for the first time, to solve benchmark instances of order up to $n = 30$ within one hour (see Section 8).

2 Notation, definitions, and known results on UBTs

In this section, we introduce the notation (summarized in Table 1) that we will use in this article and we recall known results about trees that we will use in several proofs.

Given two integers α and β, with $\alpha \leq \beta$, we denote by $[\alpha, \beta]$ the discrete interval $\{\alpha, \alpha+1, \alpha+2, \ldots, \beta\}$ and by $[\alpha, \beta] \setminus i$ the following union of two discrete intervals $[\alpha, \alpha+i-1] \cup [\alpha+i+1, \ldots, \beta]$. For the sake of conciseness, we denote by $[\alpha]$ the interval $[1, \alpha]$. For the same reason, whenever we use the universal quantifier $\forall$ on different elements of a given set (or indices of an interval), we always understand that these elements (or indices) are distinct. For example, when we write $\forall\ i, j \in S$, for some set S, (or $\forall\ i, j \in [n]$), we intend that i and j are distinct.

Given a tree T with $n \geq 3$ leaves, we label its leaves by using positive integers in $[n]$, and its internal nodes by using positive integers that are greater than or equal to $n+1$. A *weighted tree* (T, w) is a pair consisting of a tree graph $T = (V, E)$, where V is the node-set, E is the edge-set, and $w = \{w_e : \forall e \in E\}$ is a set of weights associated with the edges of T. We say that an edge of a tree is *external* if one of its endpoints is a leaf; otherwise, we say that it is *internal*. We recall that every UBT in $\mathcal{U}_n$ has $2n-3$ edges, n of which are external and $n-3$ are internal.

The problem of identifying the properties that a symmetric integer matrix τ of order n must satisfy to encode the PLM induced by a UBT is a special case of the *Tree Realization Problem* (TRP) [6,7,34, 44,46,49,51,52]. Given a *metric* matrix $\mathbf{D} = \{d_{ij}\}$ on a target set S of n points, i.e., a matrix of order n satisfying

$$
\begin{aligned}
d_{ii} &= 0 && \forall\ i \in S && \text{(null diagonal conditions)}\\
d_{ij} &\geq 0 && \forall\ i, j \in S && \text{(nonnegativity conditions)}\\
d_{ji} &= d_{ij} && \forall\ i, j \in S && \text{(symmetry conditions)}\\
d_{ij} + d_{jk} &\geq d_{ik} && \forall\ i, j, k \in S && \text{(triangle inequalities)}
\end{aligned}
$$

the TRP asks whether there exists a *tree realization* of $\mathbf{D}$, i.e., a weighted tree (T, w) such that $S \subset V$, $w_e \geq 0$, for all $e \in E$, and

$$
\sum_{e \in P_{ij}^T} w_e = d_{ij} \quad \forall\ i, j \in S. \qquad \text{(tree-fitting conditions)}
$$

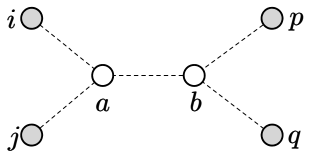


Fig. 3 An example of a UBT T such that the paths P_{ij}^T and P_{pq}^T are disjointed and connected by the path P_{ab}^T. Dashed lines indicate parts of the tree that are omitted and contain at least one edge.

If there is a tree realization of $\mathbf{D}$, then we say that $\mathbf{D}$ *fits* (T, w). In 1974, Buneman characterized the properties that ensure a tree realization for metrics by means of the following result:

Proposition 1 *(Buneman, 1974 [7])*
Let $\mathbf{D}$ *be a metric on a target set* S *of* n *points. Then, there exists a tree realization for* $\mathbf{D}$ *if and only if the following condition holds:*

$$d_{ij} + d_{pq} \leq \max\{d_{ip} + d_{jq}, d_{iq} + d_{jp}\} \qquad \forall\ i, j, p, q \in S. \qquad \text{(four-point conditions)}$$

Determining the properties that a integer metric matrix τ must satisfy to encode the PLM induced by a UBT is equivalent to determining the properties that ensure a *UBT realization* for τ when imposing the additional requirement that all the weights w_e must be equal to 1. This last requirement makes the four-point conditions necessary but not sufficient for a UBT realization.

The combinatorial properties of UBTs can be exploited to strengthen both the triangle inequalities and Buneman's four-point conditions. The fact that the paths between each pair of leaves of a UBT are unique, and that some of these paths must necessarily share one or more edges, implies the following propositions [19]:

Proposition 2 *Consider a UBT* $T \in \mathcal{U}_n$ *and three distinct leaves* i, j *and* k *of* T*. Then, the path-lengths* τ_{ij}, τ_{jk}, *and* τ_{ik} *of* T *satisfy the following inequality:*

$$\tau_{ij} + \tau_{jk} - \tau_{ik} \geq 2. \tag{2}$$

Proof Because T is a tree, there exists a unique internal node a that belongs to P_{ij}^T, P_{jk}^T, and P_{ik}^T. Then

$$\tau_{ij} + \tau_{jk} - \tau_{ik} = 2\tau_{aj} \geq 2$$

and the statement follows. □

Proposition 3 *Consider a UBT* $T \in \mathcal{U}_n$ *and any quartet of leaves* i, j, p, *and* q *of* T*. Then, the path-lengths of* T *relative to these leaves satisfy exactly one of the following three conditions:*

$$\tau_{ij} + \tau_{pq} + 2 \leq \tau_{ip} + \tau_{jq} = \tau_{iq} + \tau_{jp} \tag{3a}$$
$$\tau_{ip} + \tau_{jq} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{iq} + \tau_{jp} \tag{3b}$$
$$\tau_{iq} + \tau_{jp} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{ip} + \tau_{jq}. \tag{3c}$$

Proof Denote by $T_{\alpha\beta}$ the minimum subtree of T containing leaves α and β. Since T is a tree, there must be a leaf $v \in \{j, p, q\}$ such that T_{iv} is edge-disjoint from the subtree containing the other two leaves. Without loss of generality, let j be that leaf, so that the subtrees T_{ij} and T_{pq} are edge-disjoint (see Figure 3). Let $a \in P_{ij}^T$ and $b \in P_{pq}^T$ be the endpoints of the internal path joining P_{ij}^T and P_{pq}^T. In this scenario, we have that $\tau_{ip} + \tau_{jq} = \tau_{iq} + \tau_{jp}$ and $(\tau_{ip} + \tau_{jq}) - (\tau_{ij} + \tau_{pq}) = 2\tau_{ab}$. Note that a and b must be distinct, otherwise T would have a node of degree 4. So $\tau_{ab} \geq 1$ and the statement follows. The other two cases are similar, and are obtained when the pairs of disjoint subtrees are T_{ip} and T_{jq}, or T_{iq} and T_{jp}. □

We refer to (2) and (3) as the *strong triangle inequality* and the *strong four-point condition*, respectively. Observe that, since two paths can share only an integer number of edges, (2) and (3) can be further strengthened by replacing the term $+2$ by $+2s$ for some integer $s \in [n-1]$.

A well-known result on UBTs, derived from coding theory, states that:

Proposition 4 *(Kraft, 1949 [38]; McMillan, 1956 [40])*
A sequence $\tau_i = \{\tau_{ij}, j \in [n]\}$ of positive integers in $[2, n-1]$ encodes the path-length sequence of a UBT $T \in \mathcal{U}_n$ rooted in the i-th leaf of T if and only if its entries satisfy

$$\sum_{j \in [n] \setminus i} 2^{-\tau_{ij}} = \frac{1}{2}. \qquad \text{(Kraft's equality)}$$

Another result on UBTs, introduced in the literature by Semple and Steel [50], relates the PLM τ of a UBT T to certain undirected Hamiltonian tours on the leaves of T. Let $G([n])$ denote the complete graph having $[n]$ as the set of nodes and let $\mathcal{H}([n])$ denote the set of Hamiltonian tours on $G([n])$. In the following, we will represent an undirected Hamiltonian tour $H \in \mathcal{H}([n])$ either by the permutation $\pi = (i_1, \ldots, i_n)$, indicating the order in which the nodes are visited, or by the ordered set of edges $\{(i_k, i_{k+1})\}_{k=1}^{n}$ of the tour with the assumption $i_{n+1} = i_1$.

Definition 1 *A tour $H = (i_1, \ldots, i_n) \in \mathcal{H}([n])$ is a* circular ordering *on a UBT $T \in \mathcal{U}_n$ if there exists a planar representation of T in which the leaves, visited clockwise, give H.*

For example, consider the UBT $T \in \mathcal{U}_5$ in Figure 1. Its planar representation in the figure implies that $H = (1, 2, 3, 4, 5)$ is a circular order. The Hamiltonian tour $(1, 2, 3, 5, 4)$ is another circular order $(1, 2, 3, 5, 4)$, while $(1, 2, 4, 3, 5)$ is not.

Let $\mathcal{C}_T \subset \mathcal{H}([n])$ denote the set of circular orderings of a UBT $T \in \mathcal{U}_n$. Then, by adapting the results of Proposition 3.1 in [50] to UBTs, we have that:

Proposition 5

- *The number of circular orderings on a UBT $T \in \mathcal{U}_n$ is 2^{n-3}, i.e., $|\mathcal{C}_T| = 2^{n-3}$.*
- *The fraction of the circular orderings of $T \in \mathcal{U}_n$ in which two leaves i and j are adjacent (i.e., in which (i, j) is an edge of the tour) is $2^{-(\tau_{ij}-2)}$. Therefore, i and j are adjacent in exactly $2^{n-1-\tau_{ij}}$ circular orderings.*

3 Characterizing Θ_n

We observed in Section 2.2 of [19] that the integrality and the symmetry conditions of the path-lengths together with the strong triangle inequalities, the strong four-point conditions, Kraft's equalities, and the following equality, referred to as *$\mathcal{U}_n$-manifold*,

$$\sum_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} \tau_{ij} 2^{-\tau_{ij}} = 2n - 3 \qquad (4)$$

are necessary conditions to ensure that an integer matrix τ of order n can encode the PLM of a UBT. The following theorem indicates which set of the above conditions is not only necessary but also sufficient to characterize an integer matrix as a PLM. In the proof of the theorem, we exploit the fact that each internal edge of a tree is a bridge connecting two non-overlapping subtrees rooted at its endnodes and constituted by the remaining edges of the tree.

Theorem 1 *An integer matrix τ of order $n \geq 3$ encodes the PLM of a UBT $T \in \mathcal{U}_n$ if and only if it satisfies*

$$\begin{aligned}
\tau_{ij} &\in [2, n-1] && \forall\ i, j \in [n] && \text{(integrality conditions)}\\
\tau_{ji} &= \tau_{ij} && \forall\ i, j \in [n] && \text{(symmetry conditions)}\\
\tau_{ij} + \tau_{jk} - \tau_{ik} &\geq 2 && \forall\ i, j, k \in [n] && \text{(strong triangle inequalities)}\\
\sum_{\substack{j \in [n]\\ j \neq i}} 2^{-\tau_{ij}} &= \frac{1}{2} && \forall\ i \in [n] && \text{(Kraft's equalities)}
\end{aligned}$$

and, for all $i, j, p, q \in [n]$, exactly one of the following

$$\begin{aligned}
&\tau_{ij} + \tau_{pq} + 2 \leq \tau_{ip} + \tau_{jq} = \tau_{iq} + \tau_{jp}\\
&\tau_{ip} + \tau_{jq} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{iq} + \tau_{jp} \qquad \text{(strong four-point conditions)}\\
&\tau_{iq} + \tau_{jp} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{ip} + \tau_{jq}.
\end{aligned}$$

Proof Necessity: see Section 2.2 of [19].

Sufficiency. Since τ satisfies the strong triangle inequalities and the strengthened four-point conditions, by Proposition 1 there exists a tree realization whose tree graph has $[n]$ as set of leaves and such that $\tau_{ij} = \sum_{e \in P_{ij}^T} w_e$ for $i, j \in [n]$. Now suppose that some internal node v of T has degree four or more. Pick four distinct leaves of T, say i, j, p, and q, which are located in four of the subtrees that join in v. It is easy to see that these four leaves violate the strengthened four-point condition (3). We can then assume that every internal node of T has degree less than or equal to 3. Suppose that T includes some internal nodes of degree 2. Then, we can always construct a new tree T' from T with the same τ matrix of T. Specifically, consider any path Q_{ij} in T whose endnodes i and j have degrees 1 or 3, whereas all intermediate nodes have degree 2. Replace Q_{ij} by an edge $e = (i, j)$ with weight $w_e = \sum_{k \in Q_{ij}} w_k$ and iterate this replacement until the tree has no more internal nodes of degree 2. Let T' be the resulting tree. It is easy to see that T' is a UBT that satisfies $\tau_{ij} = \sum_{e \in P_{ij}^T} w_e$ for $i, j \in [n]$. Then, we can always assume that T is a UBT.

Now, consider an internal edge $e = (r, s)$ of T and the four rooted subtrees of T whose respective roots are adjacent to r and s. Pick one leaf in each of these subtrees, and denote these leaves as i, j, p, and q. Without loss of generality, assume that i and j are in subtrees rooted at the two neighbors of r distinct from s. Then, by (3a), it holds that $\tau_{ip} + \tau_{jq} - \tau_{ij} - \tau_{pq} = 2w_e \geq 2$. A similar case occurs when considering an external edge $e = (r, s)$ of T, with r being a leaf. Indeed, by choosing any two distinct leaves of T, say $p \neq r$ and $q \neq r$, the tightened triangle inequality (2) gives $\tau_{pr} + \tau_{rq} - \tau_{pq} = 2w_e \geq 2$. Hence, we can conclude that $w_e \geq 1$ for all $e \in E(T)$. Moreover, as T fits τ, we have that

$$\tau_{ij} = \sum_{e \in P_{ij}^T} w_e \geq \sum_{e \in P_{ij}^T} 1 = |P_{ij}^T| \quad \forall\ i, j \in [n].$$

Finally, as $T \in \mathcal{U}_n$, its induced PLM has entries $t_{ij} = |P_{ij}^T|$ and must also satisfy Kraft's equality. As $\tau_{ij} \geq |P_{ij}^T|$ and

$$\sum_{\substack{j \in [n] \\ j \neq i}} 2^{-\tau_{ij}} = \sum_{\substack{j \in [n] \\ j \neq i}} 2^{-t_{ij}} = \frac{1}{2},$$

it holds that $\tau_{ij} = |P_{ij}^T|$ for all leaves $i, j \in [n]$. Thus, the statement follows. □

Note that the integrality conditions in Theorem 1 can be slightly weakened, as they are implied by Kraft's equalities and the more general condition $\tau_{ij} \in \mathbb{N}$, for all $i, j \in [n]$.

4 A refined characterization of Θ_n

We show here that the $O(n^4)$ strong four-point conditions involved in Theorem 1 are implied by a subset of them of cardinality $O(n^3)$. We start by introducing some notation and preliminary results.

Consider a UBT $T \in \mathcal{U}_n$ and any quartet of leaves i, j, p, q of T. The strong four-point conditions applied to these leaves are three disjunctive conditions. We denote the set of PLMs $\tau \in \Theta_n$ that satisfy the first of these as

$$F^{ij|pq} = \{\tau \in \Theta_n : \tau_{ij} + \tau_{pq} + 2 \leq \tau_{ip} + \tau_{jq} = \tau_{iq} + \tau_{jp}\}.$$

Similarly, we denote by $F^{ip|jq}$ and $F^{iq|jp}$ the sets of PLMs $\tau \in \Theta_n$ that satisfy the second and the third condition, respectively. Moreover, we define

$$\mathcal{F}^{ijpq} = F^{ij|pq} \cup F^{ip|jq} \cup F^{iq|jp}.$$

We will use the fact that each set $F^{ij|pq}$ can be seen as the result of the following intersection:

$$F^{ij|pq} = F_{=}^{ij|pq} \cap F_{<}^{ij|pq}$$

where

$$F_{=}^{ij|pq} = \{\tau \in \Theta_n : \tau_{ip} + \tau_{jq} = \tau_{iq} + \tau_{jp}\};$$
$$F_{<}^{ij|pq} = \{\tau \in \Theta_n : \tau_{ij} + \tau_{pq} + 2 \leq \tau_{ip} + \tau_{jq} \text{ or } \tau_{ij} + \tau_{pq} + 2 \leq \tau_{iq} + \tau_{jp}\}.$$

Note that the sets $F^{ij|pq}$, $F^{ij|pq}_{=}$, and $F^{ij|pq}_{<}$ are invariant with respect to the order of their indices before and after the vertical slash and the swapping of the indices before the vertical slash with the indices after the vertical slash, e.g., $F^{ij|pq}_{<} = F^{ji|pq}_{<} = F^{ij|qp}_{<} = F^{ji|qp}_{<} = F^{pq|ij}_{<}$, but $F^{ij|pq}_{<} \neq F^{ip|jq}_{<}$. In turn, the invariance of the above sets implies that the set $\mathcal{F}^{ijpq}$ is invariant with respect to the order of its indices, e.g., $\mathcal{F}^{ijpq} = \mathcal{F}^{qjpi}$.

Proposition 6 *For any quartet of leaves $i, j, p, q \in [2, n]$ of any UBT $T \in \mathcal{U}_n$, the following conditions hold:*

$$F^{1i|pq} \cap F^{1j|pq} \subseteq F^{ij|pq}_{=} \tag{6}$$

$$F^{1i|jp} \cap F^{1j|pq} \subseteq F^{ij|pq}_{<} \tag{7}$$

Proof Condition (6) holds because $\tau \in F^{1i|pq} \cap F^{1j|pq}$ implies $\tau_{1p} + \tau_{iq} = \tau_{1q} + \tau_{ip}$ and $\tau_{1q} + \tau_{jp} = \tau_{1p} + \tau_{jq}$, which in turn imply $\tau_{iq} + \tau_{jp} = \tau_{ip} + \tau_{jq}$ and so $\tau \in F^{ij|pq}_{=}$. Condition (7) holds because $\tau \in F^{1i|jp} \cap F^{1j|pq}$ implies $\tau_{1j} + \tau_{ip} = \tau_{1p} + \tau_{ij}$ and $\tau_{1j} + \tau_{pq} + 2 \leq \tau_{1p} + \tau_{jq}$, which in turn imply $\tau_{ij} + \tau_{pq} + 2 \leq \tau_{ip} + \tau_{jq}$ and so $\tau \in F^{ij|pq}_{<}$. □

In the light of the above proposition and definitions, we can show now that if the strong four-point conditions hold for all the quartets of indices $(1, i, j, p)$, $(1, i, j, q)$, $(1, i, p, q)$ and $(1, j, p, q)$, then the strong four-point conditions for the remaining quartets are redundant.

Proposition 7 *For any quartet of leaves $i, j, p, q \in [2, n]$ of any UBT $T \in \mathcal{U}_n$, the following condition holds*

$$\mathcal{F}^{1ijp} \cap \mathcal{F}^{1jpq} \cap \mathcal{F}^{1ipq} \cap \mathcal{F}^{1ijq} \subseteq \mathcal{F}^{ijpq}. \tag{8}$$

Proof First, rewrite (8) as

$$\overbrace{(F^{1i|jp} \cup F^{1j|ip} \cup F^{1p|ij})}^{\mathcal{F}^{1ijp}} \cap \overbrace{(F^{1j|pq} \cup F^{1p|jq} \cup F^{1q|jp})}^{\mathcal{F}^{1jpq}} \cap$$
$$\cap \underbrace{(F^{1i|pq} \cup F^{1p|iq} \cup F^{1q|ip})}_{\mathcal{F}^{1ipq}} \cap \underbrace{(F^{1i|jq} \cup F^{1j|iq} \cup F^{1q|ij})}_{\mathcal{F}^{1ijq}} \subseteq \underbrace{(F^{ij|pq} \cup F^{ip|jq} \cup F^{iq|jp})}_{\mathcal{F}^{ijpq}}. \tag{9}$$

We distribute the unions in the left-hand side of (9) over the intersection operators, by obtaining 81 possible intersections:

$$(F^{1i|jp} \cap F^{1j|pq} \cap F^{1i|pq} \cap F^{1i|jq}) \cup \ldots \cup (F^{1p|ij} \cap F^{1q|jp} \cap F^{1q|ip} \cap F^{1q|ij}) \subseteq (F^{ij|pq} \cup F^{ip|jq} \cup F^{iq|jp}).$$

Then, we prove condition (8) by showing that each of these 81 intersections is a subset, possibly empty, of one of the three subsets $F^{ij|pq}$, $F^{ip|jq}$, and $F^{iq|jp}$ of $\mathcal{F}^{ijpq}$. In the following, the presence of a dot in place of the k-th subset, $k \in [4]$, in an intersection indicates that the dot may be substituted by any of the three subsets of the k-th set on the left-hand side of (8). The proof is limited to the 27 intersections of type $(F^{1i|jp} \cap . \cap . \cap .)$, which have $F^{1i|jp}$ as a first subset. In fact, any intersection of type $(F^{1j|ip} \cap . \cap . \cap .)$ or $(F^{1p|ij} \cap . \cap . \cap .)$ can be reduced to one of type $(F^{1i|jp} \cap . \cap . \cap .)$ by renaming the leaves i, j, p, and q.

We begin by considering the nine intersections of type $(F^{1i|jp} \cap F^{1j|pq} \cap . \cap .)$. These are subsets of $F^{ij|pq}_{<}$ by condition (7). We also note that $\tau \in F^{1i|jp} \cap F^{1j|pq}$ implies $\tau_{1i} + \tau_{jp} + 2 \leq \tau_{1j} + \tau_{ip}$ and $\tau_{1j} + \tau_{pq} + 2 \leq \tau_{1q} + \tau_{jp}$, the combination of which gives

$$\tau_{1i} + \tau_{pq} + 4 \leq \tau_{1q} + \tau_{ip}. \tag{10}$$

Among the nine intersections above:

- the three intersections $(F^{1i|jp} \cap F^{1j|pq} \cap F^{1i|pq} \cap .)$ are subsets of $F^{ij|pq} \subseteq F^{ijpq}$ as they are subsets of $F^{ij|pq}_{<}$ and, as $F^{1j|pq} \cap F^{1i|pq} \subseteq F^{ij|pq}_{=}$ by (6), also subsets of $F^{ij|pq}_{=}$.
- the three intersections $(F^{1i|jp} \cap F^{1j|pq} \cap F^{1p|iq} \cap .)$ are empty sets because $\tau \in F^{1p|iq}$ implies $\tau_{1i} + \tau_{pq} = \tau_{1q} + \tau_{ip}$, which leads to an infeasibility with (10) .
- the three intersections $(F^{1i|jp} \cap F^{1j|pq} \cap F^{1q|ip} \cap .)$ are empty sets because $\tau \in F^{1q|ip}$ implies $\tau_{1q} + \tau_{ip} + 2 \leq \tau_{1i} + \tau_{pq}$, which also leads to an infeasibility with (10).

We now consider the nine intersections of type $(F^{1i|jp} \cap F^{1p|jq} \cap . \cap .)$ and in particular:

- the three intersections $(F^{1i|jp} \cap F^{1p|jq} \cap . \cap F^{1i|jq})$
- the three intersections $(F^{1i|jp} \cap F^{1p|jq} \cap . \cap F^{1j|iq})$
- the three intersections $(F^{1i|jp} \cap F^{1p|jq} \cap . \cap F^{1q|ij})$

The proof that the above intersections are subsets, possibly empty, of F^{ijpq} is identical to the one previously presented for the first nine intersections, with the exception that the leaves (i, j, p, q) are replaced by (i, p, j, q).

Finally, we consider, the nine intersections of type $(F^{1i|jp} \cap F^{1q|jp} \cap . \cap .)$. By the condition (6), these are subsets of $F^{iq|jp}_{=}$, where the leaves (i, j, p, q) are replaced by (i, q, j, p). Among these intersections:

- The three intersections $(F^{1i|jp} \cap F^{1q|jp} \cap . \cap F^{1i|jq})$ are subsets of $F^{iq|jp} \subseteq F^{ijpq}$ because they are subsets of $F^{iq|jp}_{=}$ and also subsets of $F^{iq|jp}_{<}$, as by (7) $F^{1q|jp} \cap F^{1i|jq} \subseteq F^{iq|jp}_{<}$, with the leaves (i, j, p, q) replaced by (i, q, j, p).
- The three intersections $(F^{1i|jp} \cap F^{1q|jp} \cap . \cap F^{1j|iq})$ are subsets of $F^{iq|jp} \subseteq F^{ijpq}$ because they are subsets of $F^{iq|jp}_{=}$ and also subsets of $F^{iq|jp}_{<}$, as $\tau \in F^{1i|jp} \cap F^{1j|iq}$ implies $\tau_{1i} + \tau_{jp} + 2 \leq \tau_{1j} + \tau_{ip}$ and $\tau_{1j} + \tau_{iq} + 2 \leq \tau_{1i} + \tau_{jq}$ whose combination gives $\tau_{iq} + \tau_{jp} + 4 \leq \tau_{ip} + \tau_{jq}$, which in turn means $\tau \in F^{iq|jp}_{<}$.
- The three intersections $(F^{1i|jp} \cap F^{1q|jp} \cap . \cap F^{1q|ij})$ are subsets of $F^{iq|jp} \subseteq F^{ijpq}$ because they are subsets of $F^{iq|jp}_{=}$ and also subsets of $F^{iq|jp}_{<}$, as by (7) $F^{1i|jp} \cap F^{1q|ij} \subseteq F^{iq|jp}_{<}$ with the leaves (i, j, p, q) replaced by (q, i, j, p).

□

The invariance of the sets $\mathcal{F}^{1ijp}$ with respect to the order of their indices, for $i, j, p \in [2, n]$, and Proposition 7 imply the following result.

Theorem 2 *For any UBT $T \in \mathcal{U}_n$, the strong four-point conditions restricted to all quartets of leaves $(1, i, j, p)$ imply the strong four-point conditions for the remaining quartets of leaves (i, j, p, q), with $i, j, p, q \in [2, n]$.*

Similar arguments can be used to prove the following result.

Proposition 8 *For any UBT $T \in \mathcal{U}_n$ and any quartet of leaves $(1, i, j, p)$ of T, the nine strong triangle inequalities associated with the triplets $(1, i, j)$, $(1, i, p)$ and $(1, j, p)$ and the strong four-point conditions for the quartets $(1, i, j, p)$ imply the strong triangle inequalities for the triplet of leaves (i, j, p), with $i, j, p \in [2, n]$.*

5 On the convex hull of Θ_n

In this section we study $Conv(\Theta_n)$. We first present a number of valid inequalities for $Conv(\Theta_n)$; we will then show that some of these inequalities are also facet-defining.

Proposition 9 *Consider any UBT $T \in \mathcal{U}_n$ and a tour $H = (i_1, \ldots, i_n)$ on its leaves. Then, the* circular ordering inequality

$$\sum_{j \in [n-2]} \tau_{i_j, i_{j+1}} + \tau_{i_{n-1}, i_1} \geq 4n - 8 \tag{11}$$

is valid for any $\tau \in \Theta_n$.

Proof As $T \in \mathcal{U}_n$, T has $2n-3$ edges. The subtree T' obtained from T by removing leaf i_n and its incident edge has $2n-4$ edges. Now, consider the set of paths on T': $P^{T'}_{i_1,i_2}$, $P^{T'}_{i_2,i_3}, \ldots$, $P^{T'}_{i_{n-2},i_{n-1}}$, $P^{T'}_{i_{n-1},i_1}$. If H is a circular ordering on T then, from Proposition 4 of [50], the overall number of edges involved is at least $2(2n-4) = 4n-8$. If H is not a circular ordering on T, then (11) trivially holds as a strict inequality. Thus, the statement follows. □

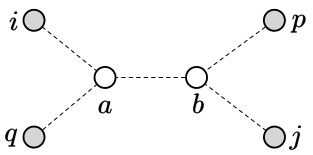


Fig. 4 An example of a UBT T such that the paths P^T_{iq} and P^T_{jp} are disjointed and connected by the path P^T_{ab}. Dashed lines indicate parts of the tree that are omitted and contain at least one edge.

Note that i_n is not included in the left-hand side of (11); if it were, the resulting inequality would be implied by the current expression of (11) and the triangle inequality $\tau_{i_{n-1},i_n} + \tau_{i_n,i_1} - \tau_{i_{n-1},i_1} \geq 2$. Note also that if H' is a circular ordering of T' then (11) holds as an equality; the converse, however, is not necessarily true. With a little abuse of notation, in the following we extend the definition of τ_{ij} also to the internal nodes of a given tree T: τ_{ij} will be intended as the length of the unique path in T between nodes i and j.

Proposition 10 *Consider any UBT $T \in \mathcal{U}_n$ and any quartet of leaves i, j, p, q of T. Then, the PLM τ induced by T satisfies the following* weak Buneman's inequality

$$2(\tau_{iq} + \tau_{jq} + \tau_{pq}) - (\tau_{ij} + \tau_{ip} + \tau_{jp}) \geq 8. \tag{12}$$

Proof As T is a tree, there exists a leaf $v \in \{i, j, p\}$ such that P^T_{qv} is disjoint from the path connecting the other two leaves. Suppose, without loss of generality, that the subtree T_{iq} of T containing leaves i and q is disjoint from T_{jp} and denote by P^T_{ab} the path on T connecting the internal node a, root of T_{iq}, to the internal node b, root of T_{jp} (see Figure 4). Then, we have

$$\begin{aligned} 2(\tau_{iq} + \tau_{jq} + \tau_{pq}) - (\tau_{ij} + \tau_{ip} + \tau_{jp}) &= 2(\tau_{qa} + \tau_{ai} + \tau_{qa} + \tau_{ab} + \tau_{bj} + \tau_{qa} + \tau_{ab} + \tau_{bp}) \\ &\quad -(\tau_{ia} + \tau_{ab} + \tau_{bj} + \tau_{ia} + \tau_{ab} + \tau_{bp} + \tau_{jb} + \tau_{bp}) \\ &= 6\tau_{qa} + 2\tau_{ab} \geq 8 \end{aligned}$$

□

We call (12) "weak Buneman's inequality" because, similarly to Buneman's conditions, it exploits the fact that there are always two disjoint paths among the paths connecting any quartet of leaves of a UBT.

Proposition 11 *For any set of $n \geq 6$ leaves and any integer constant $k \in [n-2]$, it is possible to compute in polynomial-time the largest values of α_k such that, for any two disjoint subsets of leaves S_1 and S_2 with $|S_1| = 2$ and $|S_2| = k$, the following* $(|S_1|, |S_2|)$-split inequality

$$\sum_{i \in S_1} \sum_{j \in S_2} \tau_{ij} \geq \alpha_k \tag{13}$$

holds true for all $\tau \in \Theta_n$.

Proof Assume, without loss of generality, that $S_1 = \{n-1, n\}$ and $S_2 = [k]$. We start by observing that $\alpha_k = \min_{\ell \in [2,n-1]} \alpha_{k,\ell}$ with

$$\alpha_{k,\ell} = \min_{\substack{T \in \mathcal{U}_n: \\ \tau_{n-1,n} = \ell}} \sum_{i \in [k]} (\tau_{i,n-1} + \tau_{i,n}). \tag{14}$$

Now, for all values $\ell \in [2, n-1]$, consider the path $P^T_{n-1,n}$ of a UBT $T \in \mathcal{U}_n$ with $\tau_{n,n-1} = \ell$ and let Q-subtree denote the generic subtree of T rooted in a node adjacent to one of the $\ell - 1$ internal nodes of $P^T_{n-1,n}$ but not containing any node of $P^T_{n-1,n}$. The connectivity of a tree and the uniqueness of the paths joining its leaves imply that every leaf i of T, other than $n-1$ and n, belongs to a Q-subtree. We call r_i the node of $P^T_{n-1,n}$ adjacent to the root of the Q-subtree containing i. Then, for all leaves $i \in [k]$, we have

$$\tau_{n-1i} + \tau_{ni} = \tau_{n-1r_i} + \tau_{r_i i} + \tau_{nr_i} + \tau_{r_i i} = \ell + 2\tau_{r_i i}.$$

Thus,

$$\alpha_{k,\ell} = \min_{\substack{T \in \mathcal{U}_n: \\ \tau_{n-1n} = \ell}} \sum_{i \in [k]} (2\tau_{r_i i} + \ell) = k\ell + 2 \min_{\substack{T \in \mathcal{U}_n: \\ \tau_{n-1,n} = \ell}} \sum_{i \in [k]} \tau_{r_i i}. \tag{15}$$

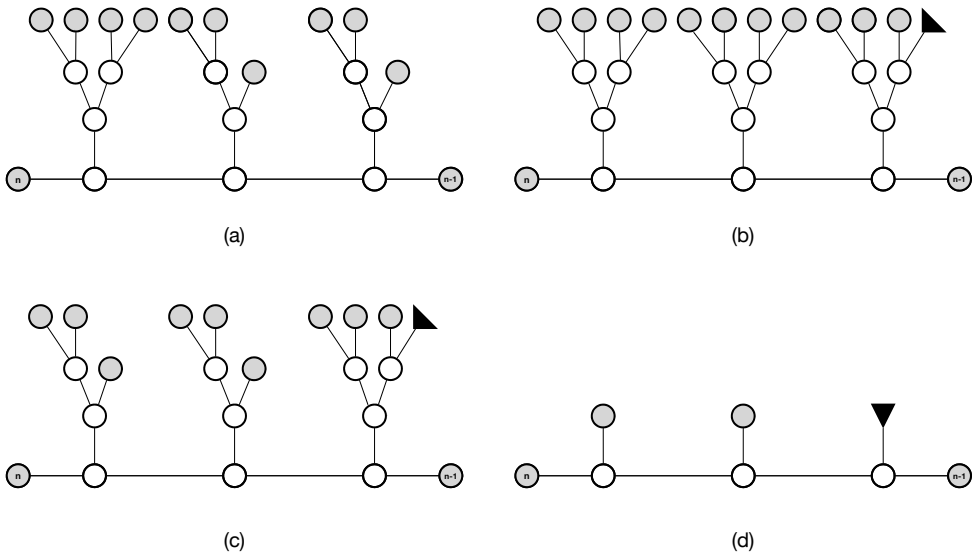


Fig. 5 The four cases that can occur for T°. The gray leaves of the four UBTs are the $[k] \cup [n-1, n]$ leaves, while the black triangle represents the subtree of the remaining $[k+1, n-2]$ leaves when $k < n-2$. The cases (a), (b), and (c) are characterized by $\ell \leq k+1$, while the case (d) is characterized by $\ell \geq k+2$. In particular, in the case (a), $k = 10, \ell = 4$, and $k = n-2$, which leads to $\alpha_{k,\ell} = 96$; in the case (b), $k = 9, \ell = 4$, $k < n-2$, and $(k+1)/(l-1)$ is a power of 2, which leads to $\alpha_{k,\ell} = 86$; in the case (c), $k = 9, \ell = 4$, $k < n-2$, which leads to $\alpha_{k,\ell} = 110$; in the case (d), $k = 2, \ell = 4$, $k < n-2$, which leads to $\alpha_{k,\ell} = 12$.

Let $T^\circ \in \mathcal{U}_n$ denote an optimal solution to (15). If $\ell \geq k+2$, there exists a T° such that each of its leaves in $[k]$ is adjacent to an internal node of the path $P^T_{n-1,n}$. In such a case, $\alpha_{k,\ell} = k(\ell+2)$ (see Figure 5.d). If $\ell \leq k+1$, some of the Q-subtrees can contain more than one leaf in $[k]$. It is easy to verify, by contradiction, that there exists a T° that satisfies the following properties. The leaves of T° in $[k]$ are uniformly distributed among the $\ell-1$ Q-subtrees, i.e., the number of leaves in $[k]$ of the Q-subtrees differs by at most of one, and the Q-subtrees are balanced with respect to the leaves in $[k]$, i.e., the lengths of the paths from the root of the Q-subtree and each of its leaves in $[k]$ differ by at most one (see Figures 5.a and 5..c). If $k < n-2$, the leaves of T° in $[k+1, n-2]$ belong to a single subtree that includes no leaves in $[k]$. This subtree is included in one of the Q-subtrees (see Figures 5.b and 5..c). In addition, $\tau_{ir_i} \geq \tau_{jr_j}$, for all $i \in [k+1, n-2]$ and all $j \in [k]$. Given the above properties, for all $\ell \leq k+1$, tedious but trivial computations lead to

$$\alpha_{k,\ell} = k\ell + 2\left[k\left(3 + \log_2\left\lfloor\frac{k+c}{\ell-1}\right\rfloor_2\right) - 2(l-1)\left\lfloor\frac{k+c}{\ell-1}\right\rfloor_2 + d\right]$$

where $\lfloor . \rfloor_2$ rounds down its argument to the nearest power of 2; c is equal to 0 if $k = n-2$ and 1 otherwise; finally, d is equal to: 0 if $k = n-2$; 2 if $(k+1)/(\ell-1)$ is a power of 2; and 1 if none of the previous conditions holds. The value of α_k can therefore be computed by comparing the values of $\alpha_{k,\ell}$, for $\ell \in [2, k+2]$, and by taking the minimum. □

We note, in particular, that for $n \geq 7$

$$\alpha_1 = 5, \quad \alpha_2 = 12, \quad \alpha_3 = 20, \text{ and } \quad \alpha_4 = 30.$$

5.1 Dimension and facets of $Conv(\Theta_n)$

In this section we determine the dimension of $Conv(\Theta_n)$ and we show that some valid inequalities are facet-defining. To start we establish when a subset of PLMs in Θ_n is linearly independent. To this end, we will look at PLMs as vectorizations of their upper triangular parts.

Given a permutation π on $[n]$, we denote $\pi(T)$ as the UBT that is obtained from T by permuting its leaves according to π, i.e., by setting $\tau(\pi(T))_{i,j} = \tau(T)_{\pi(i),\pi(j)}$, for all distinct $i, j \in [n]$. We also denote $\text{swap}^{(i,j)}(T)$ as the UBT $\pi^{i,j}(T)$, where $\pi^{i,j}$ is the identity function except for $\pi^{i,j}(i) = j$ and $\pi^{i,j}(j) = i$. This permutation reflects on the original $\tau(T)$ in that it provides a new PLM obtained from $\tau(T)$ by swapping rows i and j as well as columns i and j.

As in [19], we define a *cherry* of T as any connected subtree induced by two leaves and their common adjacent node. In particular, if i and j belong to the same cherry, then $\tau(T)_{i,j} = 2$. In Figure 1, we have that nodes 1,2 and 6 are a cherry, as well as 4, 5, and 8. We also define a *caterpillar* $T \in \mathcal{U}_n$ as a UBT whose subgraph induced by its internal nodes forms a path. If $T \in \mathcal{U}_n$ is a caterpillar, then its structure is completely determined by the sequence of its leaves. For example, if the sequence is $(i_1, i_2, i_3. \cdots, i_{n-2}, i_{n-1}, i_n)$, then we mean that i_1 and i_2 are in the same cherry, i.e. $\tau_{i_1,i_2} = 2$; moreover, for $k \in [3, n-1]$, we also understand that $\tau_{i_1,i_k} = \tau_{i_2,i_k} = k$, and finally $\tau_{i_1,i_n} = \tau_{i_2,i_n} = n-1$. We define the *fundamental caterpillar* F^n of Θ_n as the caterpillar determined by the sequence $(1, 2, 3, \cdots n-2, n-1, n)$. For example, the UBT shown in Figure 1 is the fundamental caterpillar with 5 leaves. Finally, given a subset $S \subset \mathcal{U}_n$, we define $\tau(S) = \{\tau(T) \in \Theta_n : T \in S\}$.

We now present some results that will support a number of proofs in this section. We denote $\tau^{(r,s)}$ as the path-length matrix derived from the fundamental caterpillar by means of a swap between the leaves r and s.

Proposition 12 *Consider a UBT $T' \in \mathcal{U}_n$ and any subset $\mathcal{B} \subset \mathcal{U}_n$ of UBTs such that $\tau(\mathcal{B})$ is linearly independent. If there exist two distinct linear functions $f^{=} : \Theta_n \to \mathbb{R}$ and $f^{\neq} : \Theta_n \to \mathbb{R}$ and three distinct constants, say K^0, K^1, and K^2, with $K^0 \neq 0$ and $K^1 \neq K^2$, such that $f^{=}(T) = K^0$, for all $T \in \mathcal{B} \cup \{T'\}$, $f^{\neq}(T) = K^1$, for all $T \in \mathcal{B}$, and $f^{\neq}(T') = K^2$, then $\tau(\mathcal{B} \cup \{T'\})$ is linearly independent.*

Proof Let B^1, B^2, ..., $B^{|\mathcal{B}|}$ be the UBTs in $\mathcal{B}$ and suppose, by contradiction, that $\tau(\mathcal{B} \cup \{T'\})$ is linearly dependent. Then, there exists a vector $\alpha \in \mathbb{R}^{|\mathcal{B}|}$ such that the following linear combination holds

$$\tau(T') = \sum_{k=1}^{|\mathcal{B}|} \alpha^k \tau(B^k). \tag{16}$$

In this case, the value $f^{=}(\tau(T'))$ can be written as

$$K^0 = f^{=}(\tau(T')) = f^{=}\left(\sum_{k=1}^{|\mathcal{B}|} \alpha^k \tau(B^k)\right) = \sum_{k=1}^{|\mathcal{B}|} \alpha^k f^{=}(\tau(B^k)) = \sum_{k=1}^{|\mathcal{B}|} \alpha^k K^0 \tag{17}$$

Because K^0 is by hypothesis nonzero, we can conclude that

$$\sum_{k=1}^{|\mathcal{B}|} \alpha^k = 1. \tag{18}$$

On the other hand, applying $f^{\neq}$ to $\tau(T')$ we get

$$K^2 = f^{\neq}(\tau(T')) = f^{\neq}\left(\sum_{k=1}^{|\mathcal{B}|} \alpha^k \tau(B^k)\right) = \sum_{k=1}^{|\mathcal{B}|} \alpha^k f^{\neq}(\tau(B^k)) = \sum_{k=1}^{|\mathcal{B}|} \alpha^k K^1. \tag{19}$$

From (18), we have $K^2 = K^1$, which leads to a contradiction. Thus, the statement follows. □

The following result, whose proof is omitted, states that leaf permutation preserves linear independence.

Proposition 13 *For any $n \geq 3$, let $\mathcal{B} = \{T_1, T_2, \cdots, T_{k-1}, T_k\}$ denote a subset of $\mathcal{U}_n$ such that $\tau(\mathcal{B})$ is linearly independent. Then, for any permutation π of $[n]$, the set $\mathcal{B}^\pi = \{\pi(T) \in \mathcal{U}_n : T \in \mathcal{B}\}$ is linearly independent.*

We now introduce an operation that will enable the construction of a set of $\binom{n}{2}$ linearly independent PLMs in Θ_n.

Definition 2 *Consider a UBT $T \in \mathcal{U}_n$ and a leaf i. The expansion of i in T, denoted by* $\text{expand}(T, i)$*, is the operation that consists in adding two new leaves adjacent to i, one of which is labeled i and the other is new and labeled $n+1$.*

In light of this definition, the following proposition holds:

Proposition 14 *Consider a subset $\mathcal{S} \subseteq \mathcal{U}_{n-1}$ of UBTs with $n-1$ leaves such that $\tau(\mathcal{S})$ is linearly independent. Then, the set $\mathcal{S}' = \{\text{expand}(T, n-1) \in \mathcal{U}_n : T \in \mathcal{S}\}$ is such that $\tau(S')$ is linearly independent, $|\mathcal{S}| = |\mathcal{S}'|$, and $\tau(T)_{n-1,n} = 2$ for all $T \in \mathcal{S}'$.* □

Proof Given a UBT $T \in \mathcal{U}_{n-1}$, we first observe that the UBT $T_e \in \mathcal{U}_n$, obtained from T by expanding its $(n-1)$-th leaf, induces the following PLM $\tau \in \Theta_n$:

$$\tau(T_e)_{i,j} = \begin{cases} \tau(T)_{i,j}, & \text{if } i,j \leq n-2 \\ \tau(T)_{i,n-1} + 1, & \text{if } i \leq n-2 \text{ and } j \in \{n-1, n\} \\ \tau(T)_{n-1,j} + 1, & \text{if } j \leq n-2 \text{ and } i \in \{n-1, n\} \\ 2, & \text{if } \{i,j\} = \{n-1, n\}. \end{cases} \tag{20}$$

Without loss of generality, consider a subset $\mathcal{S} = \{T^1, T^2, \cdots, T^{|\mathcal{S}|}\} \subseteq \mathcal{U}_{n-1}$ and the corresponding subset

$$\mathcal{S}' = \{\text{expand}(T^k, n-1) = T_e^k \in \mathcal{U}_n : \ k \in [|\mathcal{S}|]\} \subseteq \mathcal{U}_n.$$

By construction, $|\mathcal{S}| = |\mathcal{S}'|$ and $\tau(T_e^k)_{n-1,n} = 2$, for each $k \in [|\mathcal{S}|]$. Assume, by contradiction, that $\mathcal{S}'$ is not linearly independent. Then, there exists a non-zero vector $\alpha \in \mathbb{R}^{|\mathcal{S}|}$ such that

$$\sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T_e^k) = 0. \tag{21}$$

In particular, entrywise

$$\sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T_e^k)_{i,j} = 0$$

must hold true for all $i, j \in [n]$. Consider the entry $(n-1, n)$ and observe that the following condition holds:

$$0 = \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T_e^k)_{n-1,n} = \sum_{k=1}^{|\mathcal{S}|} 2\alpha^k. \tag{22}$$

Now, consider the entries (i, j) such that $i, j \leq n-2$. From (21), we have that

$$0 = \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T_e^k)_{i,j} = \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T^k)_{i,j}. \tag{23}$$

Finally, consider the entries $(i, n-1)$, such that $i \leq n-2$. From (21), we have that

$$\begin{aligned} 0 = \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T_e^k)_{i,n-1} &= \sum_{k=1}^{|\mathcal{S}|} \alpha^k (\tau(T^k)_{i,n-1} + 1) \\ &= \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T^k)_{i,n-1} + \sum_{k=1}^{|\mathcal{S}|} \alpha^k. \end{aligned} \tag{24}$$

From (22), we can conclude that

$$0 = \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T^k)_{i,n-1} + \sum_{k=1}^{|\mathcal{S}|} \alpha^k = \sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T^k)_{i,n-1} \tag{25}$$

and for the entries $(n-1, j)$, such that $j < n-1$, a similar result holds. Hence, for each entry (i, j) we have that $\sum_{k=1}^{|\mathcal{S}|} \alpha^k \tau(T^k)_{i,j} = 0$, which implies that also $\tau(\mathcal{S})$ is not linearly independent, a contradiction; thus, the statement follows. □

$\tau_{1,2}$	$\tau_{1,3}$	$\tau_{1,4}$	$\tau_{1,5}$	$\tau_{1,6}$	$\tau_{2,3}$	$\tau_{2,4}$	$\tau_{2,5}$	$\tau_{2,6}$	$\tau_{3,4}$	$\tau_{3,5}$	$\tau_{3,6}$	$\tau_{4,5}$	$\tau_{4,6}$	$\tau_{5,6}$
3	3	4	4	3	2	5	5	4	5	5	4	2	3	3
4	4	3	3	3	2	5	5	3	5	5	3	2	4	4
2	3	5	5	4	3	5	5	4	4	4	3	2	3	3
3	5	5	2	4	4	4	3	3	2	5	3	5	3	4
3	5	5	4	2	4	4	3	3	2	3	5	3	5	4
3	4	4	3	3	5	5	4	2	2	3	5	3	5	4
2	4	5	5	3	4	5	5	3	3	3	3	2	4	4
4	5	5	2	3	3	3	4	3	2	5	4	5	4	3
4	5	5	3	2	3	3	3	4	2	4	5	4	5	3
3	4	4	3	3	3	3	4	4	2	5	5	5	5	2
2	5	5	3	4	5	5	3	4	2	4	3	4	3	3
5	5	3	2	4	2	4	5	3	4	5	3	3	3	4
3	3	4	3	4	2	5	4	5	5	4	5	3	2	3
3	5	4	2	5	4	3	3	4	3	5	2	4	3	5
2	4	4	4	4	4	4	4	4	2	4	4	4	4	2

Table 2 The set $\tau(\mathcal{B}^6)$ described in terms of the vectorization of the upper triangular part of the PLMs of the UBTs $T \in \mathcal{B}^6$. Each row corresponds to a different UBT T. It is easy to check, by using numerical tools, that each row (i.e., each PLM) satisfies (26) at equality and that the set $\tau(\mathcal{B}^6)$ is linearly independent.

We now present a key result that will allow us both to determine the dimension of $Conv(\Theta_n)$ and to establish that the circular ordering inequalities (11) introduced in the previous section are facet-defining for $Conv(\Theta_n)$. To this end, given a UBT $T \in \mathcal{U}_n$ and an integer $k \in [4, n-1]$, we define the quantity

$$H^k(T) = \sum_{i=1}^{k-1} \tau(T)_{i,i+1} + \tau(T)_{k,1}$$

and we observe that, by Proposition 9, the following inequality holds:

$$H^{n-1}(T) \geq 4n - 8. \tag{26}$$

Moreover, for any three distinct leaves, i, j, and k of T, we define

$$\Delta_k^{i,j}(T) = \tau(T)_{i,k} + \tau(T)_{k,j} - \tau(T)_{i,j} \tag{27}$$

as the slack of the associated triangle inequality. Then, the following proposition holds:

Proposition 15 *For any integer $n \geq 6$, there exists a subset $\mathcal{B}^n \subset \mathcal{U}_n$ of $\binom{n}{2}$ UBTs, such that $\tau(\mathcal{B}^n)$ is linearly independent and $H^{n-1}(T) = 4n - 8$, for all UBTs $T \in \mathcal{B}^n$.*

Proof We prove the statement by inductively constructing $\mathcal{B}^n$. Specifically, we first show that the statement holds true for $n = 6$. Then, by assuming that the statement holds true for $n-1$, we show that we can derive from $\mathcal{B}^{n-1}$ a set $\mathcal{B}' \subset \mathcal{U}_n$ of $\binom{n-1}{2}$ UBTs whose induced PLMs are linearly independent and satisfy (26) at equality. Finally, we show that we can obtain $\mathcal{B}^n$ by joining $\mathcal{B}'$ with a specific set of $n-1$ caterpillars that satisfy (26) at equality.

Base case. Assume $n = 6$ and denote by $\mathcal{B}^6$ the set of UBTs in $\mathcal{U}_6$ associated with the PLMs listed in Table 2. We have that $\tau(\mathcal{B}^6)$ is linearly independent and each of its PLMs satisfies (26) at equality. Hence, (26) is facet-defining for $Conv(\Theta_6)$. Observe here that not all PLMs in $\tau(\mathcal{B}^6)$ are induced by caterpillars. The very last τ matrix in the Table 2 is not induced by a caterpillar (a τ matrix induced by a caterpillar must have exactly two entries equal to 2). This observation will be useful later.

Inductive step. Assume that the statement holds true for $n-1$, $n>6$. Denote by $\mathcal{B}^{n-1}$ a set of UBTs in $\mathcal{U}_{n-1}$ whose cardinality satisfies $|\mathcal{B}^{n-1}| = \binom{n-1}{2}$ and such that all PLMs in $\tau(\mathcal{B}^{n-1})$ satisfy (26) at equality. Consider the set

$$\mathcal{B}_e = \{\text{expand}(T, n-2) \in \mathcal{U}_n : \ T \in \mathcal{B}^{n-1}\}.$$

By Proposition 14, we have that $\tau(\mathcal{B}_e)$ is linearly independent. For a UBT $T \in \mathcal{B}^{n-1}$, let $T_e = \text{expand}(T, n-2)$ and note that, from the properties of the expansion operation, the following holds:

$$\begin{aligned} \tau(T_e)_{u,n-2} &= \tau(T)_{u,n-2} + 1 && \forall u \in [n-3] \\ \tau(T_e)_{u,v} &= \tau(T)_{u,v} && \forall u, v \in [n-3]. \end{aligned}$$

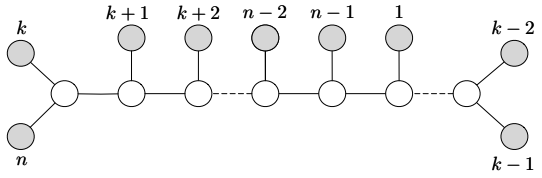


Fig. 6 The caterpillar R^k.

As $H^{n-2}(T) = 4(n-1) - 8$ holds for $T \in \mathcal{B}^{n-1}$, we have that

$$H^{n-2}(T_e) = 4(n-1) - 8 + 2 = 4n - 10 \quad \forall\, T_e \in \mathcal{B}_e \tag{28}$$

By definition of an expansion, we also have that $n-2$ and n are in the same cherry for all UBTs in $\mathcal{B}_e$. Now, consider the set

$$\mathcal{B}' = \{\text{swap}^{(n-1,n)}(T) \in \mathcal{U}_n :\ T \in \mathcal{B}_e\}.$$

By Proposition 13, $\tau(\mathcal{B}')$ is linearly independent. Moreover, due to the swap operation, $n-1$ and $n-2$ are in the same cherry of each tree in $\mathcal{B}'$. Then, we have that

$$\begin{aligned} H^{n-1}(T) &= \sum_{i=1}^{n-3} \tau_{i,i+1}(T) + \tau_{n-2,n-1}(T) + \tau_{n-1,1}(T) \\ &= H^{n-2}(T) - \tau_{n-2,1}(T) + \tau_{n-2,n-1}(T) + \tau_{n-1,1}(T). \end{aligned} \tag{29}$$

Observe that since $n-1$ and $n-2$ are in the same cherry of each $T' \in \mathcal{B}'$, $\tau(T')_{n-1,n-2} = 2$ and $\tau(T')_{n-1,1} = \tau(T')_{n-2,1}$. Together with (28), this fact implies that

$$H^{n-1}(T') = H^{n-2}(T') - \tau(T')_{n-2,1} + 2 + \tau(T')_{n-2,1} = 4n - 10 + 2 = 4n - 8. \tag{30}$$

and that

$$\Delta_{n-2}^{n-1,i}(T') = 2 \tag{31}$$

for all leaves $i \in [n] \setminus \{n-2, n-1\}$

We now construct $\mathcal{B}^n$ by complementing the $\binom{n-1}{2}$ UBTs in $\mathcal{B}'$ with a set of $n-1$ caterpillars $T \in \mathcal{U}_n$ that satisfy (26) at equality and are such that $\tau(\mathcal{B}^n)$ is linearly independent.

The first $n-2$ caterpillars are constructed as follows. For a given $k \in [3, n-1]$ consider the caterpillar $R^k \in \mathcal{U}_n$ determined by the sequence $(n, k, k+1, \cdots, n-2, n-1, 1, 2, \cdots k-2, k-1)$. For $k = 2$, consider the caterpillar $R^2 \in \mathcal{U}_n$ determined by the sequence $(n, 2, 3, \cdots, n-2, n-1, 1)$ (see, e.g., Figure 6, for $k \geq 4$). Note that, for $k \in [2, n-1]$, both n and k are in the same cherry while the opposite cherry is made of $k-1$ and $k-2$ ($n-1$ if $k = 2$). This fact implies that the leaves in $[n-1]$ are sequentially ordered such that the distance between i and $i+1$ (taken mod $n-1$) is 3, except for leaves n and k and leaves $k-1$ and $k-2$ ($n-1$ if $k = 2$), whose distance is 2 because they are in the same cherries, leaves $k-1$ and k, whose distance is $n-1$ because they are in the opposite cherries. Hence, $H^{n-1}(R^k) = 2 + (n-1) + 3(n-3) = 4n - 8$ for $k \in [2, n-1]$.
Now consider the trees R^k for $k \in [2, n-2]$ and, for each of them, a leaf $k' \in [n] \setminus \{n-2, n-1\}$ and the path between leaves k' and $n-1$. The leaf $n-2$ is adjacent to an internal node in the considered path if $k' \geq k$, otherwise it is not. Then, if $k' \geq k$ we have $\Delta_{n-2}^{n-1,k'}(R^k) = 2$, otherwise $\Delta_{n-2}^{n-1,k'}(R^k) > 2$.
We exploit this observation and conditions (30) and (31), which hold for all $T' \in \mathcal{B}'$, to show that $\tau(\mathcal{B}' \cup \bigcup_{k=2}^{n-2}\{R^k\})$ is linearly independent. Specifically, assume by contradiction that the statement is false. Let j denote the smallest index such that $\tau(\mathcal{B}' \cup \bigcup_{k=2}^{j} R^k)$ is linearly dependent and let $\mathcal{B}_{LI}$ denote the set $\mathcal{B}' \cup \bigcup_{k=2}^{j-1}\{R^k\}$. We have that $\Delta_{n-2}^{n-1,j-1}(T) = 2$ for all $T \in \mathcal{B}_{LI}$ while we have $\Delta_{n-2}^{n-1,j-1}(R^j) > 2$ for R^j since $j < n-1$. Hence, by Proposition 12, $\tau(\mathcal{B}' \cup \bigcup_{k=2}^{j}\{R^k\})$ is linearly independent, leading to a contradiction. In particular, we can use Proposition 12 by setting $f^{=}(T) = H^{n-1}(T)$, $f^{\neq}(T) = \Delta_{n-2}^{n-1,j-1}(T)$ and observing that $f^{=}(T) = K^0 = 4n-8$ for all $T \in \mathcal{B}' \cup \bigcup_{k=2}^{j}\{R^k\}$, $f^{\neq}(T) = K^1 = 2$ for all $T \in \mathcal{B}' \cup \bigcup_{k=2}^{j-1}\{R^k\}$, while $f^{\neq}(R^j) = K^2 > 2$ for R^j. Analogously, we can again use Proposition 12 to show that R^{n-1} is also linearly independent of $\tau(\mathcal{B}' \cup \bigcup_{k=2}^{n-2}\{R^k\})$. To

do this, we first observe that $\Delta_{n-2}^{n-1,n}(T) = 2$ for $T \in \mathcal{B}' \cup \bigcup_{k=2}^{n-2}\{R^k\}$, since $n-2$ and $n-1$ are in the same cherry in all these trees. Differently, $\Delta_{n-2}^{n-1,n}(R^{n-1}) = 2n-2 > 2$. Then, we note that $H^{n-1}(T) = 4n-8$ also holds when $T = R^{n-1}$. Thus we have proved that $\tau(\mathcal{B}' \cup \bigcup_{k=2}^{n-1}\{R^k\})$ is linearly independent.

To complete the construction of $\mathcal{B}^n$, consider the caterpillar T_f that is determined by the sequence $(n-2, n-3, n-4, n-5, \cdots, 5, 4, 3, n, 2, 1, n-1)$ and that is shown in Figure 7. As regards $H^{n-1}(T_f)$, observe that $\tau(T_f)_{i,i+1} = 3$, for $i \in [n-4] \setminus 2$, while

$$\begin{aligned}\tau(T_f)_{2,3} &= 4,\\ \tau(T_f)_{n-2,n-3} &= 2,\\ \tau(T_f)_{n-2,n-1} &= n-1,\\ \tau(T_f)_{n-1,1} &= 2.\end{aligned}$$

Hence, $H^{n-1}(T_f) = 3(n-5) + 4 + 2 + n - 1 + 2 = 4n-8$. As a final step, we prove that the PLMs induced by the UBTs in $\mathcal{B}^n = \mathcal{B}' \cup \bigcup_{k=2}^{n-2}\{R^k\} \cup \{T_f\}$ are linearly independent. To this end, consider the linear function

$$\begin{aligned}F_h(T) = (n-3)(\tau(T)_{1,n-1} - \tau(T)_{1,n-2}) + (n-4)\tau(T)_{n-1,n-2}\\ + (\tau(T)_{n-1,n} - \tau(T)_{n-2,n}).\end{aligned}$$

For each $T \in \mathcal{B}'$ we have that $F_H(T) = 0 + 2(n-4) + 0 = 2n-8$ as leaves $n-1$ and $n-2$ belong to the same cherry. For R^2, we have

$$F_h(R^2) = (n-3)(2-3) + (n-4)(3) + (n-1-(n-2)) = 2n-8.$$

For $k \in [3, n-2]$ we have

$$\begin{aligned}\tau(R^k)_{n-2,n-1} &= 3,\\ \tau(R^k)_{1,n-1} - \tau(R^k)_{1,n-2} &= -1,\\ \tau(R^k)_{n-1,n} - \tau(R^k)_{n-2,n} &= 1.\end{aligned}$$

Hence, $F_h(R^k) = (n-3)(-1) + (n-4)(3) + (1) = 2n-8$. Finally, for R^{n-1} we have

$$\begin{aligned}\tau(R^{n-1})_{n-2,n-1} &= n-1,\\ \tau(R^{n-1})_{1,n-1} - \tau(R^{n-1})_{1,n-2} &= (3) - (n-2) = -n+5,\\ \tau(R^{n-1})_{n-1,n} - \tau(R^{n-1})_{n-2,n} &= (2) - (n-1) = -n+3.\end{aligned}$$

Hence, $F_h(R^{n-1}) = (n-3)(5-n) + (n-4)(n-1) + (3-n) = 2n-8$. On the other hand, for T_f, we have

$$\begin{aligned}\tau(T_f)_{n-2,n-1} &= n-1,\\ \tau(T_f)_{1,n-1} - \tau(T_f)_{1,n-2} &= (2) - (n-1) = -n+3,\\ \tau(T_f)_{n-1,n} - \tau(T_f)_{n-2,n} &= (4) - (n-3) = -n+7.\end{aligned}$$

Hence, $F_h(T_f) = (n-3)(3-n) + (n-4)(n-1) + (7-n) = 2$. Because $n \geq 6$, it holds that $2n-8 \geq 12-8 = 4$, which implies that $F_h(T) = 2n-8$, for each $T \in \mathcal{B}' \cup \bigcup_{k=1}^{n-2}\{R^k\}$, while $F_h(T_f) \neq 2n-8$. Since $H^{n-1}(T)$ holds true for each $T \in \mathcal{B}^n$, by Proposition 12 we have that $\tau(\mathcal{B}^n)$ is linearly independent. Thus, the statement follows. □

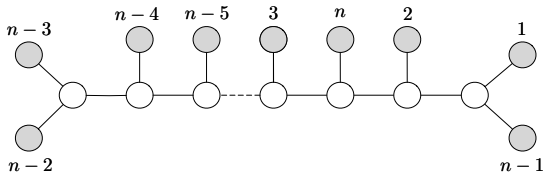


Fig. 7 The caterpillar T_f.

As an immediate consequence of Proposition 15 we have that

Theorem 3 *$Conv(\Theta_n)$ is full-dimensional, i.e.,* $\dim(Conv(\Theta_n)) = \binom{n}{2}$.

Proof We first observe that, as a PLM is completely determined by its upper triangle entries,

$$dim(Conv(\Theta_n)) \leq \binom{n}{2}.$$

Now, from Proposition 15, there are $\binom{n}{2}$ linearly independent PLMs in $Conv(\Theta_n)$ that satisfy (26) at equality. Hence, $dim(Conv(\Theta_n))$ is either $\binom{n}{2}$ or $\binom{n}{2}-1$. To conclude that $dim(Conv(\Theta_n)) = \binom{n}{2}$, it suffices to show a UBT T such that $H^{n-1}(T) > 4n-8$. To this end, consider the caterpillar $T = \text{swap}^{(3,4)}(F^n)$, which is determined by the sequence $(1, 2, 4, 3, 5, \cdots, n-2, n-1, n)$. Observe that $\tau(T)_{1,2} = 2$, $\tau(T)_{n-1,1} = n-1$ and $\tau_{i,i+1}(T) = 3$ for $i \in [2, n-2] \setminus \{4\}$. Due to the swap, we have $\tau_{4,5} = 4$. Thus,

$$H^{n-1}(T) = 2 + 3(n-3) + 4 + n - 1 = 4n - 3 > 4n - 8$$

and the statement follows. □

Now, let N' denote a subset of $[n]$ such that $|N'| = n-1$ and let $\mathcal{H}(N')$ denote the set of Hamiltonian tours of $G(N')$. Then, the following result holds:

Theorem 4 *The* circular ordering inequalities

$$\sum_{(i,j)\in H} \tau_{ij} \geq 4n - 8 \qquad \forall\ N' \subset [n] : |N'| = n-1,\ \forall\ H \in \mathcal{H}(N') \tag{32}$$

are facet-defining for $Conv(\Theta_n)$.

Proof Consider a tour $H = (i_1, i_2, \cdots, i_{n-1}) \in \mathcal{H}(N')$ and denote by v the only element in $[n] \setminus N'$. Then, $\pi = (i_1, i_2, \cdots, i_{n-1}, v)$ is a permutation of $[n]$. By Proposition 13 and Proposition 15

$$\sum_{i=1}^{n-2} \tau_{\pi(i),\pi(i+1)} + \tau_{\pi(n-1),\pi(1)} \geq 4n - 8$$

is facet-defining for $Conv(\Theta)$. Because v is not in the tour H in (32), an edge (i, j), with $i, j \in [n]$, $i < j$, belongs to H if and only if either $\pi(1) = i$ and $\pi(n-1) = j$, or there exists a $k \in [n-2]$ such that $\pi(k) = i$ and $\pi(k+1) = j$. Thus, the statement follows. □

The following proposition (whose proof in Appendix A is similar to that of Proposition 15) also holds.

Proposition 16 *The strong triangle inequalities* (2) *are facet-defining for $Conv(\Theta_n)$.*

6 Optimizing over PLMs of UBTs: The case study of the BMEP

The objective function of the BMEP and some of the conditions involved in the characterization of Θ_n are nonlinear in τ. In order to derive an Integer Linear Programming (ILP) formulation for the problem, we introduce the 0-1 variables x_{ij}^{ℓ}, which takes value 1 if and only if $\tau_{ij} = \ell$. By Theorem 2, the strong four-point condition involved in the characterization of Θ_n also requires a choice of one out of three alternative conditions for all distinct $j, p, q \in [2, n]$. To model this choice, we introduce the decision variables y_{pq}^{j} that take value 1 if and only if the subtree containing 1 and j is disjoint from that containing p and q. Then, in the light of Theorem 1, the following formulation is valid for the BMEP:

Formulation 1

$$\min \sum_{i\in[n]} \sum_{j\in[n]\setminus i} d_{ij} \left(\sum_{\ell\in[2,n-1]} 2^{-\ell} x_{ij}^{\ell} \right) \tag{33a}$$

$$s.t. \sum_{\ell\in[2,n-1]} x_{ij}^{\ell} = 1 \qquad \forall\ i, j \in [n] \tag{33b}$$

$$x_{ij}^{\ell} = x_{ji}^{\ell} \qquad \forall\ i,j \in [n],\ \ell \in [2,n-1] \tag{33c}$$

$$\tau_{ij} = \sum_{\ell \in [2,n-1]} \ell x_{ij}^{\ell} \qquad \forall\ i,j \in [n] \tag{33d}$$

$$\sum_{j \in [n]\setminus\{i\}} \sum_{\ell \in [2,n-1]} 2^{-\ell} x_{ij}^{\ell} = \frac{1}{2} \qquad \forall\ i \in [n] \tag{33e}$$

$$\tau_{1i} + \tau_{1j} - \tau_{ij} \geq 2 \qquad \forall\ i,j \in [2,n] \tag{33f}$$

$$\tau_{1i} + \tau_{ij} - \tau_{1j} \geq 2 \qquad \forall\ i,j \in [2,n] \tag{33g}$$

$$y_{pq}^{j} + y_{jq}^{p} + y_{jp}^{q} = 1 \qquad \forall\ j,p,q \in [2,n] \tag{33h}$$

$$\tau_{1j} + \tau_{pq} \geq 2(1 - y_{jp}^{q}) + \tau_{1p} + \tau_{jq} - (2n-2)y_{pq}^{j} \qquad \forall\ j,p,q \in [2,n] \tag{33i}$$

$$x_{ij}^{\ell} \in \{0,1\} \qquad \forall\ i,j \in [n],\ \ell \in [2,n-1] \tag{33j}$$

$$y_{pq}^{j} \in \{0,1\} \qquad \forall\ j,p,q \in [2,n] \tag{33k}$$

$$\tau_{ij} \in [2,n-1] \qquad \forall\ i,j \in [n]. \tag{33l}$$

The convexity constraints (33b) impose that for each i,j exactly one variable x_{ij}^{ℓ} takes value 1 for $\ell \in [2,n-1]$. Constraints (33c) impose the symmetry on the x_{ij}^{ℓ} variables. Constraints (33d) relate x variables to τ variables. Constraints (33e) impose Kraft's equality and constraints (33f)- (33g) impose the strong triangle inequalities. Constraints (33h)-(33i) impose the strong four-point conditions. Specifically, constraint (33h) ensures that exactly one of the reduced strong four-point conditions holds for a fixed triplet $j,p,q \in [2,n]$. Constraints (33i), instead, generate the following set of inequalities for all $j,p,q \in [2,n]$:

$$\tau_{1j} + \tau_{pq} \geq 2(1 - y_{jp}^{q}) + \tau_{1p} + \tau_{jq} - (2n-2)y_{pq}^{j} \tag{34}$$

$$\tau_{1p} + \tau_{jq} \geq 2(1 - y_{jp}^{q}) + \tau_{1j} + \tau_{pq} - (2n-2)y_{jq}^{p} \tag{35}$$

$$\tau_{1j} + \tau_{pq} \geq 2(1 - y_{jq}^{p}) + \tau_{1q} + \tau_{jp} - (2n-2)y_{pq}^{j} \tag{36}$$

$$\tau_{1q} + \tau_{jp} \geq 2(1 - y_{jq}^{p}) + \tau_{1j} + \tau_{pq} - (2n-2)y_{jp}^{q} \tag{37}$$

$$\tau_{1p} + \tau_{jq} \geq 2(1 - y_{pq}^{j}) + \tau_{1q} + \tau_{jp} - (2n-2)y_{jq}^{p} \tag{38}$$

$$\tau_{1q} + \tau_{jp} \geq 2(1 - y_{pq}^{j}) + \tau_{1p} + \tau_{jq} - (2n-2)y_{jp}^{q}. \tag{39}$$

When $y_{pq}^{j} = 1$, (38) and (39) imply

$$\tau_{1p} + \tau_{jq} \geq \tau_{1q} + \tau_{jp}$$
$$\tau_{1q} + \tau_{jp} \geq \tau_{1p} + \tau_{jq}$$

(35) and (37) imply

$$\tau_{1p} + \tau_{jq} \geq 2 + \tau_{1j} + \tau_{pq}$$
$$\tau_{1q} + \tau_{jp} \geq 2 + \tau_{1j} + \tau_{pq}$$

while (34) and (36) are trivially satisfied. Thus, combined, (34)-(39) impose

$$\tau_{1j} + \tau_{pq} + 2 \leq \ \tau_{1p} + \tau_{jq} = \tau_{1q} + \tau_{jp}.$$

With similar arguments, it is easy to recover the other two reduced strong four-point conditions when $y_{jq}^{p} = 1$ or $y_{jp}^{q} = 1$. Finally, constraints (33l), (33j) and (33k) enforce the bounds and integrality of the x, y and τ variables, respectively.

Before concluding this section, we note that the symmetry constraints (33c) make the variables x^{ℓ}_{ji} redundant for all $i < j \in [n]$. In practice these variables as well as the τ_{ij} variables are dropped due to (33d). However, we have kept these variables in the text both for clarity and to simplify the presentation of the constraints. Similarly, we have kept the constraints (33l) in the text and we will refer to them in the propositions of the following section, even if they are trivially implied by constraints (33d) and (33j).

7 A formulation for the BMEP in the (x, τ)-space.

Several attempts to work computationally with Buneman's conditions proved unsatisfactory. This led us to investigate an alternative formulation for the BMEP that does not rely on Buneman's conditions (and the associated y variables). The removal of Buneman's conditions is enabled by a simple valid inequality that is used as a last resort when a triple of Buneman's conditions is violated. To describe this formulation, consider a relaxation of the Formulation 1, hereafter referred to as $R1$, obtained by including only (33b)-(33g), (33j), and (33l). Denote by $(\tilde{x}, \tilde{\tau})$ an integer solution to R1. If $(\tilde{x}, \tilde{\tau})$ violates the reduced strong four-point condition for some $(1, j, p, q)$, then this solution can be cut off by imposing that at least one of the x variables encoding this solution must be 0. In other words, the inequality

$$x^{\tilde{\tau}_{1j}}_{1j} + x^{\tilde{\tau}_{1p}}_{1p} + x^{\tilde{\tau}_{1q}}_{1q} + x^{\tilde{\tau}_{jp}}_{jp} + x^{\tilde{\tau}_{jq}}_{jq} + x^{\tilde{\tau}_{pq}}_{pq} \leq 5 \tag{40}$$

trivially cuts off $(\tilde{x}, \tilde{\tau})$. This observation leads to the following proposition (whose proof is omitted):

Proposition 17 *The ILP formulation* **RF1** *obtained from Formulation 1 by including* (33b)-(33g), (33j), (33l), *and the* Projected Buneman's Inequality (40) *for any* $j, p, q \in [n]$ *and any* $\tilde{\tau}_{1j}, \tilde{\tau}_{1p}, \tilde{\tau}_{1q}, \tilde{\tau}_{jp}, \tilde{\tau}_{jq}, \tilde{\tau}_{pq} \in [2, n-1]$ *that do not satisfy the strong four-point conditions is valid for the BMEP.*

We can strengthen RF1 by means of the following valid inequalities.

Proposition 18 *For any three distinct* $i, j \in [2, n]$, *the* lifted strong triangle inequality

$$\tau_{1i} + \tau_{1j} - \tau_{ij} \geq 2 + 2 \sum_{u \in [n] \setminus \{i,j\}} x^2_{1u} \tag{41}$$

*is valid for RF*1.

Proof Let T denote the UBT in $\mathcal{U}_n$ that realizes the PLM $\tau \in \Theta_n$ and let a denote the unique internal node at the cross of P^T_{1i}, P^T_{1j} and P^T_{ij}. Then, as in Proposition 2, $\tau_{1i} + \tau_{1j} - \tau_{ij} = 2\tau_{1a}$. If $\sum_{u \in [n] \setminus \{i,j\}} x^2_{1u} = 1$, i.e., there exists a leaf u distinct from i and j, such that 1 and u belong to the same cherry then there is at least one internal node in the path between a and 1. In this case, we have that $2\tau_{1a} \geq 4$, so the inequality is valid. If $\sum_{u \in [n] \setminus \{i,j\}} x^2_{1u} = 0$, then (41) reduces to the strong triangle inequality and is valid again. Finally, note that the convexity constraints (33b) and the Kraft's equalities (33e) imply that all the addends in the Kraft's equalities are strictly positive, which in turn implies that $\sum_{u \in [n] \setminus \{i\}} x^2_{1u}$ cannot be greater than 1. □

Proposition 19 *For any triplet* $i, j, p \in [n]$, *the* lifted weak Buneman's inequality

$$2(\tau_{1i} + \tau_{1j} + \tau_{1p}) - \tau_{ij} - \tau_{ip} - \tau_{jp} \geq 8 + 6 \sum_{v \in [n] \setminus \{i,j,p\}} x^2_{1v} \tag{42}$$

*is valid for RF*1.

Proof Let T denote the UBT in $\mathcal{U}_n$ that realizes the PLM $\tau \in \Theta_n$. Assume, without loss of generality, that P^T_{1i} is disjointed from P^T_{jp}. Let a and b denote two internal nodes such that $a \in P^T_{1i}$ and $b \in P^T_{jp}$ and let P^T_{ab} be the path connecting the subtree containing 1 and i to the subtree connecting j and p. As in Proposition 10, we have that $2(\tau_{1i} + \tau_{1j} + \tau_{1p}) - \tau_{ij} - \tau_{ip} - \tau_{jp} = 6\tau_{1a} + 2\tau_{ab}$. In the proof of the previous proposition we have already shown that $\sum_{u \in [n] \setminus \{i\}} x^2_{1u}$ cannot be greater than 1. So, if $\sum_{v \in [n] \setminus \{i,j,p\}} x^2_{1v} = 1$, then there exists a leaf $u \notin \{i, j, p\}$ such that 1 and u belong to the same cherry. Hence, there is at least one internal node in the path between 1 and a. In this case, we have that $6\tau_{1a} \geq 4$ and therefore $6\tau_{1a} + 2\tau_{ab} \geq 14$. If $\sum_{v \in [n] \setminus \{i,j,p\}} x^2_{1v} = 0$, then (42) reduces to the weak Buneman's inequality and is valid again. □

Proposition 20 *For any two disjoint subsets S_1 and S_2 of $[n]$, such that $S_1 = \{i, j\}$ and $|S_2| = k \in [n-2]$, the* lifted $(2, k)$-split inequality

$$\sum_{p \in S_2} (\tau_{ip} + \tau_{jp}) \geq \sum_{\ell \in [2, n-1]} \alpha_{k,\ell} x_{ij}^{\ell}, \tag{43}$$

where $\alpha_{k,\ell}$ is defined by (14)*, is valid for RF*1*.*

Proof Consider a feasible solution (x, τ) to RF1 with $\tau_{ij} = \hat{\ell} \in [2, n-1]$, hence $x_{ij}^{\hat{\ell}} = 1$. A trivial consequence of the introductory arguments in the proof of Proposition 11 is that $\sum_{p \in S_2} (\tau_{ip} + \tau_{jp}) \geq \alpha_{k,\hat{\ell}}$. As constraint (33b) holds, we can write

$$\alpha_{k,\hat{\ell}} = \alpha_{k,\hat{\ell}} x_{ij}^{\hat{\ell}} = \sum_{\ell \in [2, n-1]} \alpha_{k,\ell} x_{ij}^{\ell}$$

and the statement follows. □

Let N' denote a subset of $[n]$ such that $|N'| = n - 1$ and let $\mathcal{H}(N')$ denote the set of Hamiltonian tours of $G(N')$. Then we can derive the following result from Theorem 4:

Proposition 21 *For any Hamiltonian tour H in $G(N')$, the* lifted circular ordering inequality

$$\sum_{(i,j) \in H} \tau_{ij} \geq 4n - 8 + 2 \sum_{(i,j) \notin H} x_{ij}^2 \tag{44}$$

*is valid for RF*1*.*

Proof Let T denote the UBT in $\mathcal{U}_n$ associated with the PLM $\tau \in \Theta_n$. Suppose that i and j belong to a cherry and that the pair (i, j) is not an edge of H. Let a denote the common neighbor of i and j in T and let $\hat{a}$ be the internal node adjacent to a. The edge $(a, \hat{a})$ of T is visited at least four times when traversing H, twice when visiting i and twice when visiting j. Note that each leaf can belong to at most one cherry, and that $(a, \hat{a})$ is only traversed if either i or j is visited. Because this observation holds for each cherry whose leaves do not belong to an edge of H, the statement follows. □

Another family of valid inequalities for $RF1$ can be derived from the *generalized Kraft's inequalities* introduced in [13, p. 18], which state that every PLM $\tau \in \Theta_n$ satisfies

$$\sum_{i \in S} \sum_{j \in [n] \setminus S} 2^{-\tau_{ij}} \geq \frac{1}{2} \qquad \forall\, \emptyset \subset S \subset [n].$$

In the (x, τ)-space, this family of inequalities can be linearized as follows:

$$\sum_{i \in S} \sum_{j \in [n] \setminus S} \sum_{\ell \in [2, n-1]} 2^{-\ell} x_{ij}^{\ell} \geq \frac{1}{2} \qquad \forall\, \emptyset \subset S \subset [n]. \tag{45}$$

In the reminder of the article we shall refer to **Formulation 2** as the ILP formulation derived from RF1, by including the lifted strong triangle inequalities (41), the lifted weak Buneman's inequalities (42), the lifted $(2, k)$-split inequalities (43), the lifted circular ordering inequalities (44), the generalized Kraft's inequalities (45), and the following linearization of the $\mathcal{U}_n$-manifold equality (4) in the (x, τ)-space:

$$\sum_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} \sum_{\ell \in [2, n-1]} \ell\ 2^{-\ell} x_{ij}^{\ell} = 2n - 3. \tag{46}$$

Dataset	N	Easy Set		Hard Set	
		min n	**max** n	**min** n	**max** n
Primates	12	10	12	-	-
M17	17	10	17	-	-
M18	18	10	18	-	-
SeedPlant500	500	10	20	21	26
M43	43	10	20	21	32
M62	62	10	20	21	29
RbcL55	55	10	20	21	30
Rana64	64	10	20	21	29
M82	82	10	20	21	25

Table 3 Characteristics of the benchmark instances of the BMEP used in the computational experiments.

8 Computational experiments

We now study the performance of the ILP formulations presented in the previous sections. First, we provide details about the benchmark instances, the computational settings, and the environment considered in our experiments. We then introduce the separation oracles for the families of strengthened valid inequalities discussed in Section 7 as well as on the heuristic used to generate the primal bound for the optimal solution to the BMEP. Finally, we present the computational results obtained by these formulations on the considered instances and we compare the performance of the best ILP formulation with that obtained for the massively parallel Branch-&-Bound algorithm presented in [13] (downloadable at `https://github.com/mfrohn/BMEPparallel` and currently constituting the state-of-the-art exact solution algorithm for the BMEP).

8.1 Instances

In our experiments, we considered the set of benchmark instances of the BMEP used in [2,11,13,15, 16,27], derived from the real molecular datasets described in [2] by means of the estimation models described in [18], and downloadable at `https://www.danielecatanzaro.com/wp-content/uploads/2024/04/BMEP-datasets.zip`. Each dataset is associated with a distance matrix $\mathbf{D}$, whose order N may vary between 12 and 500. The entries of these matrices are reals with no specific properties. For each dataset, we generated several instances of the BMEP by extracting the leading principal submatrices of order $n \in [10, \min\{N, 32\}]$ from $\mathbf{D}$. we grouped these instances into two sets referred to as the *Easy Set*, including 86 instances characterized by a distance matrix of order smaller than or equal to 20, and the *Hard Set*, including 51 instances characterized by a distance matrix of order strictly greater than 20. Table 3 summarizes the characteristics of these instances. We also considered a further *Third Set* of 42 instances of the BMEP already considered in [13] and available at `https://github.com/mfrohn/BMEPparallel`. The characteristics of this set are reported in Table 4.

8.2 Implementation details

We implemented Formulations 1 and 2 in C++17, by relying on Gurobi libraries version 10.0.5 [33] for linear programming, boost libraries [4] version 1.82.0 for some routines involved in the separation algorithms described in Section 8.3, and FASTME 2 [39] for the local searches involved in the computation of the primal bound described in Section 8.4. The source code, publicly available at `https://www.danielecatanzaro.com/wp-content/uploads/2024/04/bmep_ilp_model-main.zip`, has been compiled by using Apple clang compiler version 14.0.3 with optimization flag -O3, and run on a Mac Pro equipped with a 28-core Intel Xeon W-3275M 2.50GHz CPU, equipped with 192GB of RAM and MacOS 13.5. We assumed one hour as maximum runtime limit for each tested algorithm/formulation. In the implementation of both formulations, we exploited constraints (33d) to remove τ variables and constraints (33c) to remove variables x^{ℓ}_{ij} such that $i<j$. We preferred to keep them in the previous sections because their presence greatly simplified the notation and the discussion of several implementation aspects. Table 5 reports on the only standard parameters of Gurobi that have been modified in our implementations.

8.3 Separation oracles

Some families of strengthened valid inequalities considered in Section 7 require separation oracles to be imposed in Formulation 2 at runtime. Here, we provide implementation details on these oracles.

Initially, Formulation 2 loads only the constraints (33b)-(33e), (33l), (33j) (41), and (46). When the resolution of the linear programming relaxation of Formulation 2 at a generic node of the search tree provides a fractional solution $(\tilde{x}, \tilde{\tau})$, the following separation oracles are sequentially called (in this specific order) in the attempt to separate $(\tilde{x}, \tilde{\tau})$ from $Conv(\Theta_n)$:

- **Separation oracle for the weak Buneman's inequalities (WB)** This routine checks by inspection for a violation of the lifted weak Buneman's inequalities (42) by using the following approach: it first fixes a leaf $q \in [n]$ and then identifies the triplet of indices (i, j, p) for which the inequality (42) (with indices (i, j, p, q)) is the most violated. This routine has time complexity $O(n^4)$ and generates at most $O(n)$ cuts per separation step.
- **Separation oracle for the $(2, k)$-split inequalities (2K)** This routine checks for violation of the lifted $(2, k)$-split inequalities (43) by means of a simple greedy algorithm. Specifically, given a fractional solution $(\tilde{x}, \tilde{\tau})$ to Formulation 2, the algorithm considers each possible pair of distinct leaves $i, j \in [n]$ and sorts, in ascending order, the remaining leaves according to the quantity $\tilde{h}(p) = \tilde{\tau}_{i,p} + \tilde{\tau}_{j,p}$, $p \in [n] \setminus \{i, j\}$. Then, for $k \in [4]$, the algorithm compares $\sum_{\ell \in [2,n-1]} \alpha_{k,\ell} \tilde{x}_{ij}^{\ell}$ with the sum of $\tilde{h}_p$ for the first k leaves in the sorted list and only adds an inequality for the smallest value of k that leads to a violation. The routine has time complexity $O(n^3)$ and generates at most $O(n^2)$ cuts per separation step.
- **Separation oracle for the circular ordering inequalities (CO)** This routine checks for violation of the lifted circular ordering inequalities (44) by means of the following slightly modified version of the *Nearest Neighbor Heuristic* for the *Traveling Salesman Problem* (TSP) [1]. Specifically, consider a fractional solution $(\tilde{x}, \tilde{\tau})$ to Formulation 2 and denote by $G([n])$ the complete weighted graph with n nodes, one per leaf. Assign the weight $w_{ij} = \tilde{\tau}_{ij} - \tilde{x}_{ij}^2$ to each edge (i, j) of $G([n])$. Then, for each node $i \in [n]$, we construct in a greedy way a tour in $G([n])$ that starts and ends in i and that visits precisely $n - 1$ nodes. Let S denote the set of nodes of $G([n])$ that have been not visited yet and let j denote the last visited node. If $|S| < 2$, then we visit the node v of $G([n])$ such that $v = \arg\min_{k \in S}(w_{jk})$. This process is iterated until $|S| < n - 2$. When $|S| = n - 2$, we visit the node v such that $v = \arg\min_{k \in S}(w_{jk} + w_{ki})$. This algorithm construct $O(n)$ candidate tours at each separation step. The tours then undergo to a 2-OPT local search for possible improvement [1]. The routine has time complexity $O(n^2)$ and generates at most n cuts per separation step.
- **Separation oracle for the generalized Kraft's inequalities (GKI)** This routine checks for violation of the generalized Kraft's inequalities (45) in a way similar to the subtour elimination constraints for the Symmetric Traveling Salesman Problem, i.e., by reducing the separation problem to a min-cut problem. Specifically, we construct the complete weighted graph $G([n])$ and we assign to each edge (i, j) of $G([n])$ a weight

$$w_{ij} = \sum_{\ell \in [2,n-1]} 2^{-\ell} x_{ij}^{\ell}.$$

  Then, we call Stoer-Wagner algorithm implemented in the boost libraries [4] to solve the relative min-cut problem, by obtaining as an output a bipartition of $[n]$ into the subsets S_1 and S_2. The relative

Dataset	Instance Sizes	Dataset	Instance Sizes
B-HA-573-585-BMGE	20, 25, 26	nucleic_M2839_470x829_2006	20, 21, 22
B-HA-986-1908-BMGE	20, 21	nucleic_M3862_362x1207_2008	20
B-NA-633-1751-BMGE	20, 22, 23	proteic_M2577_40x12260_2005	20,23, 24
B-NA-684-472-BMGE	20, 24, 25	proteic_M2624_139x348_2006-BMGE	20, 25, 26
B-NS-629-1111-BMGE	20, 26, 27	proteic_M2883_91x7386_2007-BMGE	20
B-NS1-284-344-BMGE	20, 29, 30	proteic_M3068_50x1000_2006	20
eudicots-BMGE	20	proteic_M3755_77x9918_2008-BMGE	20, 22, 23
euros1	20	proteic_M3756_77x11234_2008	20
euros2	20, 22, 23	rosids	20
nucleic_M2573_346x897_2006	20, 27, 28		

Table 4 Characteristics of the benchmark instances of the BMEP in the Third Set used in the computational experiments.

Parameters	Value	Description
MIPGap	1e-8	Relative tolerance
TimeLimit	3600	Time limit of 1h
LazyConstraints	1	Forces validation of integer solutions
PreCrush	1	Disable duality-based preprocessing
MIPFocus	3	Focus on improving bounds via cutting planes
Heuristics	0	Disables the use of proprietary MIP heuristics

Table 5 Parameter settings of Gurobi that have been changed in the implementations of Formulations 1 and 2.

inequality is then checked for violation and added to the formulation in the positive case. The routine has time complexity $O(n^3)$ and generates at most one cut per separation step. In order to reduce the time spent in solving the linear programming relaxations carried out at a generic node of the search tree, we only add a cut if the violation exceeds a threshold, which depends on the inequality family and whether at the root node or not. Specific values can be found in the source-code.

– **Separation oracle for the projected Buneman's inequalities** This routine checks the validity of an integer solution $(\hat{x}, \hat{\tau})$ obtained from the resolution of the linear programming relaxation of Formulation 2. The check is carried out by inspecting the existence of at least a triplet (j, p, q) such that $(1, j, p, q)$ violates the strong four-point condition. If no violation is found, $(\hat{x}, \hat{\tau})$ is retained as a valid solution to the BMEP. Otherwise, the violated inequalities are added to the formulation. This routine is only called when an integer solution is found in the search tree; it has time complexity $O(n^3)$ and may generate up to $O(n^3)$ cuts per separation step.

8.4 Primal heuristic

We generated the primal bound for the optimal solution to an instance of the BMEP by means of a novel matheuristic inspired by the *Neighbor-Joining* (NJ) algorithm, possibly one of the most common heuristics used to approximate the optimal solutions to the BMEP [32,48]. The NJ algorithm generates a solution to a given instance of the BMEP by using a clustering approach that, starting from a star having one internal node and n leaves, iteratively merges two leaves into a new node until obtaining a UBT in $\mathcal{U}_n$. In the classical version of the NJ algorithm, the choice of the pair of leaves to merge at the first iteration step is carried out by finding the pair

$$(\hat{i}, \hat{j}) = \arg\min_{i<j\in[n]} \sum_{k\in[n]} (d_{ik} + d_{kj} - d_{ij}). \tag{47}$$

Once the pair is merged into a new node $\hat{v}$, the $\hat{i}$-th and $\hat{j}$-th rows and columns of the matrix $\mathbf{D}$ are removed from $\mathbf{D}$ and a new row (and column) is added to $\mathbf{D}$, encoding the distances of all the other leaves from $\hat{v}$. The selection criterion of the pair of leaves to join at each iteration step, as well as the way in which the new entries of $\mathbf{D}$ are computed, determines the version of the NJ algorithm. The literature on molecular phylogenetics proposes several versions of the NJ algorithm; we refer the interested reader to [5,9,10,29,30] for further details.

The primal heuristic that we used in our experiments replaces the entries of matrix $\mathbf{D}$ with those of of the matrix τ obtained by computing the linear programming relaxation of Formulation 2 at a generic node of the search tree. We refer to this heuristic as *Linear Programming-based Neighbor-Joining* (LPNJ) algorithm. The rationale at the core of this heuristic is that if matrix τ is a PLM, then the LPNJ algorithm provides a weighted UBT (T, w) that realizes τ. The LPNJ algorithm generates different primal solutions in function of the linear programming relaxation obtained at a generic node of the search tree. In order to include the information about the objective function into the LPNJ algorithm, we selected the best k pairs of leaves according to (47) by using τ instead of D and by subsequently selecting among the k candidate pairs the one that minimizes the original criterion (47). We refer to this version as the *k-LPNJ algorithm*. In preliminary tests, we observed that setting $k = 3$ provided on average solutions of better quality than just using the original NJ Algorithm. Consequently, we decided to fix k to this value.

In preliminary experiments, we compared the performance of the 3-LPNJ algorithm on the considered datasets with that of FASTME 2. We observed that on 113 instances of the three datasets considered,

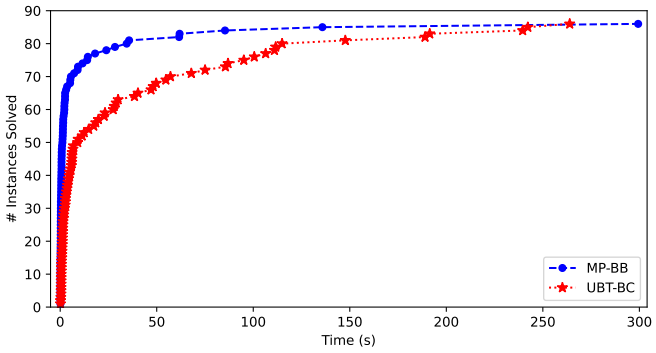


Fig. 8 Number of instances in the Easy Set over seconds for both MP-BB and UBT-BC.

the two heuristics identified the same primal solution (which, incidentally, was optimal but in two cases). In the remaining 58 instances, the 3-LPNJ algorithm found a better solution than FASTME 2 in 50 instances, with an average improvement of 0.5%. In the remaining eight instances, FASTME 2 found a better primal bound, with an average improvement of 0.13%. In these 58 instances, the best of the two primal bounds generated by the two heuristics were always optimal but in 4 cases. Given the rapidity of both heuristics (less than 0.01 seconds), we decided to run both of them at any generic node of the search tree.

After having computed the first primal bound to the problem, we try to improve it by calling classical local search procedures, such as the *Nearest Neighbor Interchange* (NNI) and the *Subtree Pruning and Regrafting* (SPR), included in FASTME 2 [39]. To reduce the computational overload caused by calling FASTME 2, the local search is just called for the first three solutions generated by the 3-LPNJ algorithm and whenever an improved primal bound is found during the tree search.

8.5 Computational results

Initially, the performances of (the methods used to solve) Formulations 1 and 2 were tested on the Easy Set, as Formulation 1 proved unable to solve the instances in both the Hard Set and Third Set within the reference time limit. The total runtimes on the Easy Set for Formulation 1 without and with the lifted constraints were of 8142 seconds and 4315 seconds, respectively. In contrast, Formulation 2 demonstrated significantly faster performance, requiring only 4193 seconds and 2953 seconds, respectively. In the light of these results, we retained Formulation 2 for further computational tests and comparisons. Hereafter, we will refer to the implementation of Formulation 2 as *UBT-BC*.

We compared the performance of UBT-BC with that of the current state-of-the-art exact solution algorithm for the BMEP [13], hereafter denoted as *MP-BB*, on all the three sets of instances of the BMEP described in Section 8.1. The results of this comparison are reported in Tables 8, 9, and 10. The columns of Table 8 report on the solution time expressed in seconds, taken by an algorithm to solve an specific instance of the BMEP; the nodes explored in the search tree; and the gap at the root node of the search tree, expressed in percentage (i.e., the ratio $(z^* - z_{LP})/z^* \cdot 100$, where z^* is the optimal solution to the instance of the BMEP, and z_{LP} is the linear programming relaxation at the root node of the search tree). The columns of Table 9 also report on the number of cuts added during the search and the final gap obtained by UBT-BC within the time limit of one hour. The symbol > 3600 in the tables indicates that the resolution of a specific instance took more than 3600 seconds. When this situation occurs consecutively (see, e.g., the case for UBT-BC in Table 9 when addressing the instance SeedPlant500 for $n = 23$ and $n = 24$, then the execution of a solution algorithm is aborted for instances of higher orders and the numerical values in the table are replaced by the symbol '-'). Bold entries indicate the best performance.

Results on the Easy Set

Figure 8 shows the number of solved instances in the Easy Set over time (expressed in seconds) for both MP-BB and UBT-BC. Although UBT-BC solved most of these instances within this set without

Instance		Time (s)			Nodes		Gap (%)		
Dataset	n	**UBT-BC**	**MP-BB**	$\Delta(\%)$	**UBT-BC**	**MP-BB**	**UBT-BC**	**MP-BB**	$\Delta(\%)$
RbcL55	20	188.97	61.60	127.37	149	4152495	2.17	2.81	**-0.64**
SeedPlants500	20	191.41	61.85	129.56	372	2894260	5.12	9.11	**-3.99**
M82	18	114.94	85.37	29.57	442	6873457	3.90	5.53	**-1.63**
	19	111.61	135.82	**-24.21**	313	9393814	4.03	6.36	**-2.33**
	20	263.91	299.37	**-35.46**	830	19360753	4.58	8.21	**-3.63**

Table 6 Comparison of the performance of both UBT-BC and MP-BB on specific instances of the BMEP belonging to the Easy Set and for which the solution times took more than 60 seconds. The i-th value in a Δ-column indicates the difference (value for UBT-BC - value for MP-BB). A negative sign means that for that specific value and instance, UBT-BC performed better than MP-BB; vice versa for positive values.

branching, the figure clearly shows that MP-BB is almost always faster than UBT-BC. MP-BB achieves these results by combining an efficient exploration strategy of the solution space of the problem based on Hendy and Penny's *Stepwise Addition Strategy* (SAS) [36] with a lower bound obtained by considering a particular relaxation of Formulation 2 consisting only of constraints (33b)-(33e) and (46). Hendy and Penny' SAS starts from a star with three leaves and then recursively extends this star until it generates potentially all the possible (2n-5)!! UBTs in $\mathcal{U}_n$ (see Figure 5 of [13]). MP-BB encodes a UBT by means of a classical arc-list encoding scheme [8] and then transforms this scheme into upper and lower bounds for the path-length values τ_{ij} of the aforementioned relaxation of Formulation 2 which is then solved at each node of the search tree. The size of this relaxation is considerably smaller than that of Formulation 2 and its resolution does not necessitate the execution of specific separation oracles. Consequently, while the resulting lower bound is poorer than the one provided by Formulation 2 (as evidenced by the gaps columns in Tables 8, 9, and 10), the overall runtime spent at each node of the search tree is considerably less than that of Formulation 2. This strategy is particularly effective when the order of the input distance matrix is smaller than or equal to 18 and the difference between the maximum and the minimum values of its entries d_{ij} is relatively small. When the lower bound implemented in MP-BB proves too poor, the number of nodes visited in the search tree becomes considerable, thereby altering the situation. Specifically, in the five instances of the Easy Set reported in Table 6, MP-BB took more than 60 seconds to reach optimality and in two of these five instances UBT-BC proved to be even faster than MP-BB. In these instances, the considerably weaker lower bounds employed by MP-BB compelled the algorithm to explore a vast number of nodes within the respective search trees, thereby slowing down the solution times of the algorithm. This phenomenon is particularly pronounced in the Hard and Third Sets. Table 8 shows that the lower bound provided by Formulation 2 at the root node of the search tree is up to 7% better than that provided by the ILP formulation used by MP-BB. However, while MP-BB solves the entire set in 919 seconds, UBT-BC requires 2953 seconds. It is noteworthy that the five hard instances displayed in Table 6 take 644 of the 919 seconds spent by the MP-BB on the set. It is also noteworthy that in the vast majority of the considered instances, UBT-BC identified the optimal solution at the root node of the search tree. This fact provides evidence of the benefit of including the valid inequalities discussed in the previous sections.

Dataset	**Instances Solved**	
	MP-BB	**UBT-BC**
SeedPlant500	**4**	2
M43	1	**10**
M62	1	**7**
RbcL55	3	**8**
Rana64	**7**	5
M82	2	**3**
Total	18	35

Table 7 Number of instances in the Hard Set solve by UBT-BC and MP-BB and grouped by dataset.

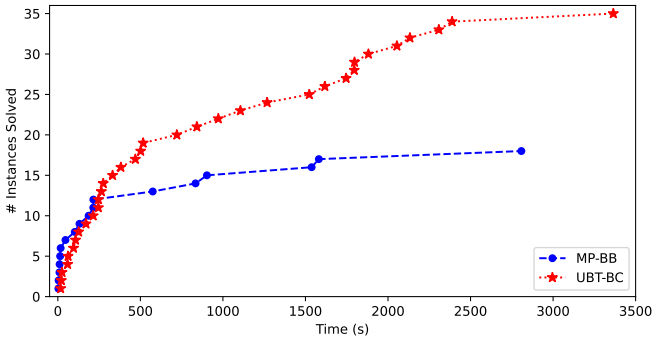


Fig. 9 Number of instances in the Hard Set over seconds for both MP-BB and UBT-BC.

Results for the Hard and Third Sets

Table 7 reports on the number of instances in the Hard Set solved within one hour for each dataset and by each algorithm, while Figure 9 shows the number of instances in the Hard Set that each algorithm can solve over elapsed time. The values reported in Table 7 show that UBT-BC outperforms MP-BB on most datasets. This behavior is particularly remarkable in M43, M62, and RbcL55, where UBT-BC was able to solve within one hour all instances, seven instances, and eight instances, respectively, versus one instance, one instance, and three instances, respectively for MP-BB. This phenomenon persists to a lesser extent in the case of M82: UBT-BC solved just three instances within one hour versus two instances for MP-BB. In the datasets SeedPlant500 and Rana64, instead we observed an inversion of this behavior: MP-BB succeeded to solve, within one hour, two instances more than UBT-BC in both datasets. This phenomenon lacks a clear justification. Basic statistical analysis of both datasets indicates that the difference between the largest and smallest entries of their distance matrices is comparable to that observed in other datasets, as is the average distance and the relative standard deviation. These instances exhibit a markedly different number of violations of triangle inequalities or Buneman's four-point conditions compared to other instances. The only thing that we remarked is that addressing the exploration of their search trees in a depth-first fashion may prove faster than other exploration strategies.

The general trend reported on Table 7 is notable also in Figure 9, where it is possible to observe how MP-BB is less effective at solving harder instances when significantly more time is allowed, in contrast to UBT-BC. Table 9 provides further insight into the performance of both algorithms on the Hard Set. The most striking information in this table is the rapid increase in the number of nodes explored by MP-BB as the size of the considered instances grows. This probably explains its relative poor performance on larger instances. In contrast, thanks to the valid inequalities, both the number of nodes explored and the number of cuts added grow much more slowly for UBT-BC.

Figure 10 shows the number of instances in the Third Set that each algorithm can solve over elapsed time. The figure highlights the different behavior of the two algorithms when tackling larger instances: while both algorithms have a comparable performance with up to 250 seconds of runtime, MP-BB is not able to solve any more instances when the run time increases from 600 to 3600 seconds. In contrast increasing the run time for the UBT-BC yields consistently more solved instances. Table 10 gives details on how the algorithms perform on the Third Set. Data continues to show the trend observed in the Hard Set where, even if the MP-BB is faster on easier instances, its solving time increases very fast with the size of the instances, thereby solving only 28 of the 42 instances in the set. The UBT-BC, instead, solved 40 instances within the reference time of one hour.

9 Open questions

Theorem 1 shows which set of conditions outlined in [19] prove both necessary and sufficient to characterize Θ_n. A question that naturally arises is whether this set could be replaced by a smaller set of (possibly linear) conditions on path-lengths. This question is nontrivial and still open. For example, Kraft's equalities seem to play a central role in the characterization of PLMs; without them, one could easily construct integer matrices that satisfy all the remaining conditions of Theorem 1 yet fail to be

PLMs, as the following one

$$\begin{pmatrix} 0 & 3 & 5 & 5 & 5 & 5 \\ 3 & 0 & 5 & 5 & 5 & 5 \\ 5 & 5 & 0 & 3 & 5 & 5 \\ 5 & 5 & 3 & 0 & 5 & 5 \\ 5 & 5 & 5 & 5 & 0 & 3 \\ 5 & 5 & 5 & 5 & 3 & 0 \end{pmatrix}.$$

Interestingly, in Theorem 1 the $\mathcal{U}_n$-manifold (4) is somewhat sidelined. This highly nonlinear equation tightens (in some cases significantly) the lower bound for the optimal solution to the BMEP, by enforcing the circular ordering inequalities associated with a UBT to be satisfied at equality. Although the role of the $\mathcal{U}_n$-manifold differs from that of the conditions involved in Theorem 1, in certain cases, its effects overlap with those of the strong four-point conditions. One such case arises when considering the following symmetric integer matrix:

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

which does not map the PLM of any UBT in $\mathcal{U}_5$. This matrix satisfies all the conditions listed in Theorem 1, but the strong four-point conditions (3). At the same time, this matrix is also cut off by the $\mathcal{U}_n$-manifold (4). By enumerating all of the PLMs in Θ_5, it is possible to show that the PLMs induced by the UBTs in $\mathcal{U}_5$ can be characterized by just the integrality constraints, the symmetry constraints, the strong triangle inequalities, Kraft's equalities, and the $\mathcal{U}_n$-manifold. An ILP formulation based on these conditions would have the benefit of involving $O(n^3)$ less variables and constraints with respect to an ILP formulation based on Theorem 1. Note, however, that the $\mathcal{U}_n$-manifold (4) cannot always replace the strong four-point conditions. In fact, the following matrix

$$\begin{pmatrix} 0 & 5 & 6 & 5 & 5 & 6 & 5 & 5 & 6 & 5 & 6 & 2 \\ 5 & 0 & 4 & 2 & 6 & 7 & 7 & 7 & 7 & 6 & 4 & 5 \\ 6 & 4 & 0 & 4 & 5 & 7 & 7 & 7 & 7 & 5 & 2 & 6 \\ 5 & 2 & 4 & 0 & 6 & 7 & 7 & 7 & 7 & 6 & 4 & 5 \\ 5 & 6 & 5 & 6 & 0 & 5 & 6 & 6 & 5 & 2 & 5 & 5 \\ 6 & 7 & 7 & 7 & 5 & 0 & 4 & 4 & 2 & 5 & 7 & 6 \\ 5 & 7 & 7 & 7 & 6 & 4 & 0 & 2 & 4 & 6 & 7 & 5 \\ 5 & 7 & 7 & 7 & 6 & 4 & 2 & 0 & 4 & 6 & 7 & 5 \\ 6 & 7 & 7 & 7 & 5 & 2 & 4 & 4 & 0 & 5 & 7 & 6 \\ 5 & 6 & 5 & 6 & 2 & 5 & 6 & 6 & 5 & 0 & 5 & 5 \\ 6 & 4 & 2 & 4 & 5 & 7 & 7 & 7 & 7 & 5 & 0 & 6 \\ 2 & 5 & 6 & 5 & 5 & 6 & 5 & 5 & 6 & 5 & 6 & 0 \end{pmatrix}.$$

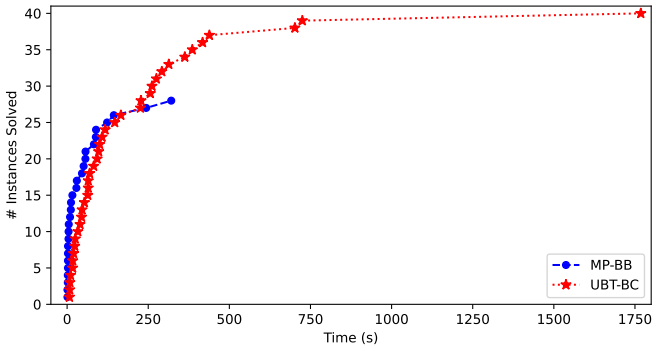


Fig. 10 Number of instances in the Third Set over seconds for both MP-BB and UBT-BC.

satisfies (4) as well as all the conditions in Theorem 1 except the strong four-point ones; however, it is not a PLM.

Identifying alternative characterizations of Θ_n not relying on the strong four-point conditions could suggest better ILP formulations for the BMEP as well as assist in the study of the BMEP polytope in Forcey et al.'s $\mathcal{X}$ space, where the expression of the strong four-point conditions becomes highly nonlinear. For this reason, it is worth noting that, in some circumstances, the presence of possible structural properties of the entries of the distance matrix might allow us to eliminate some disjunctive strong four-point conditions, e.g., because we might already know that the optimal PLM must lie within a specific subset of Θ_n. Note, indeed, that the strong four-point conditions impose three alternative ways to partition four taxa into pairs. For instance, the condition

$$\tau_{iq} + \tau_{jp} + 2 \leq \tau_{ij} + \tau_{pq} = \tau_{ip} + \tau_{jq}$$

imposes a UBT structure (shown in Figure 4), where taxa i and q share a closer common ancestor a than the ancestor b of j and p. If the relative distances (or additional information such as geographic origin of taxa) could guarantee the presence of these common ancestors, we could rule out the other two possible four-point conditions for i, j, p, and q, reducing so the size of the formulation.

A second research direction concerns the use of Formulation 2 for developing better heuristics for the BMEP. We have seen in the previous sections that the linear programming relaxation of Formulation 2 can be exploited to obtain tight primal bounds for the optimal solution to the problem. When tackling instances of the BMEP involving hundreds of sequences, Formulation 2 could be used to construct parts of the solution that are subsequently merged together to obtain a final UBT.

A third research direction concerns the study of optimization problems over Θ_n characterized by objective functions linear in τ. One problem of practical interest is the minimization of the cross-entropies associated to n sources which find applications in the area of classification, and more in general, machine learning [41].

Acknowledgments

We are very grateful to the Reviewers for their useful comments and suggestions. The first and the third authors acknowledge support from the Belgian National Fund for Scientific Research (FNRS) via the grant FNRS PDR 40007831. This first author also acknowledges support from the Fondation Louvain, via the grant "COALESCENS". Computational resources involved in this work have been provided by the super-computing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI), funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

Compliance with the ethical standard of the journal

This manuscript fully complies with the Ethical Standards of Mathematical Programming. The authors declare no conflict of interest/competing interest of any sort and fully adhere to the authorship principles and practices. All the authors equally concurred to develop the theoretical and methodological study as well as to writing the article.

References

1. D. L. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook, *The traveling salesman problem: A compuational study*, Princeton University Press, NJ, 2006.
2. R. Aringhieri, D. Catanzaro, and M. Di Summa, *Optimal solutions for the balanced minimum evolution problem*, Computers and Operations Research, 38 (2011), pp. 1845–1854.
3. L. J. Billera, S. P. Holmes, and K. Vogtmann, *Geometry of the space of phylogenetic trees*, Advances in Applied Mathematics, 27 (2001), pp. 733–767.
4. Boost, *Boost C++ libraries 1.82.0*, 2023.
5. W. J. Bruno, M. D. Socci, and A. L. Halpern, *Weighted neighbor-joining: A likelihood-based approach to distance-based phylogeny reconstruction*, Molecular Biology and Evolution, 17 (2000), pp. 189–197.

6. P. Buneman, *The recovery of trees from measure of dissimilarities*, in Archaeological and Historical Science, F. R. Hodson, D. G. Kendall, and P. Tautu, eds., Edinburgh University Press, Edinburgh, UK, 1971, pp. 387–395.
7. P. Buneman, *A note on the metric properties of trees*, Journal of combinatorial theory, 17 (1974), pp. 48–50.
8. S. Caminiti, I. Finocchi, and R. Petreschi, *On coding labeled trees*, Theoretical Computer Science, 382 (2007), pp. 97–108.
9. D. Catanzaro, *The minimum evolution problem: Overview and classification*, Networks, 53 (2009), pp. 112–125.
10. ———, *Estimating phylogenies from molecular data*, in Mathematical approaches to polymer sequence analysis and related problems, R. Bruni, ed., Springer, NY, 2011, pp. 149–176.
11. D. Catanzaro, R. Aringhieri, M. di Summa, and R. Pesenti, *A branch-price-and-cut algorithm for the minimum evolution problem*, European Journal of Operational Research, 244 (2015), pp. 753–765.
12. D. Catanzaro, M. Frohn, O. Gascuel, and R. Pesenti, *A tutorial on the balanced minimum evolution problem*, European Journal of Operational Research, 300 (2022), pp. 1–19.
13. ———, *A massively parallel branch-&-bound algorithm for the balanced minimum evolution problem*, Computers and Operations Research, (2023, to appear).
14. D. Catanzaro, L. Gatto, and M. Milinkovitch, *Assessing the applicability of the GTR nucleotide substitution model through simulations*, Evolutionary Bioinformatics, 2 (2006), pp. 145–155.
15. D. Catanzaro, M. Labbé, and R. Pesenti, *The balanced minimum evolution problem under uncertain data*, Discrete Applied Mathematics, 161 (2013), pp. 1789–1804.
16. D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-Gonzáles, *The balanced minimum evolution problem*, INFORMS Journal on Computing, 24 (2012), pp. 276–294.
17. D. Catanzaro and R. Pesenti, *Enumerating vertices of the balanced minimum evolution polytope*, Computers and Operations Research, 109 (2019), pp. 209–217.
18. D. Catanzaro, R. Pesenti, and M. Milinkovitch, *A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model*, Bioinformatics, 22 (2006), pp. 708–715.
19. D. Catanzaro, R. Pesenti, and L. A. Wolsey, *On the Balanced Minimum Evolution polytope*, Discrete Optimization, 36 (2020), pp. 1–33.
20. M. A. Cueto and F. A. Matsen, *Polyhedral geometry of phylogenetic rogue taxa*, Bulletin of Mathematical Biology, 73 (2011), pp. 1202–1226.
21. R. Desper and O. Gascuel, *Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting*, Molecular Biology and Evolution, 21 (2004), pp. 587–598.
22. K. Eickmeyer, P. Huggins, L. Pachter, and R. Yoshida, *On the optimality of the neighbor-joining algorithm*, Algorithms for Molecular Biology, 3 (2008), pp. 1–11.
23. J. Felsenstein, *Inferring phylogenies*, Sinauer Associates, Sunderland, MA, 2004.
24. S. Fiorini and G. Joret, *Approximating the balanced minimum evolution problem*, Operations Research Letters, 40 (2012), pp. 31–35.
25. S. Forcey, L. Keefe, and W. Sands, *Facets of the balanced minimal evolution polytope*, Journal of Mathematical Biology, 73 (2015), pp. 447–468.
26. S. Forcey, L. Keefe, and W. Sands, *Split-facets for balanced minimal evolution polytopes and the permutoassociahedron*, Bulletin of Mathematical Biology, in press, 79 (2017), pp. 975–994.
27. B. Fortz, O. Oliveira, and C. Requejo, *Compact mixed integer linear programming models to the minimum weighted tree reconstruction problem*, European Journal of Operational Research, 256 (2017), pp. 242–251.
28. M. Frohn, *On the approximability of the fixed-tree balanced minimum evolution problem*, Optimization Letters, 15 (2021), pp. 2321–2329.
29. O. Gascuel, *BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data*, Molecular Biology and Evolution, 14 (1997), pp. 685–695.
30. ———, *Concerning the NJ algorithm and its unweighted version, UNJ*, in Mathematical Hierarchies and Biology, B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky, eds., American Mathematical Society, Providence, RI, July 1997, pp. 149–170.
31. ———, *Mathematics of evolution and phylogeny*, Oxford University Press, New York, NY, 2005.
32. O. Gascuel and M. A. Steel, *Neighbor-joining revealed*, Molecular Biology and Evolution, 23 (2006), pp. 1997–2000.
33. Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, 2023.
34. S. L. Hakimi and S. S. Yau, *Distance matrix of a graph and its realizability*, Quarterly of Applied Mathematics, 22 (1964), pp. 305–317.
35. D. C. Haws, T. L. Hodge, and R. Yoshida, *Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope*, Bulletin of Mathematical Biology, 73 (2011), pp. 2627–2648.
36. M. D. Hendy and D. Penny, *Branch and bound algorithms to determine minimal evolutionary trees*, Mathematical Biosciences, (1982).
37. M. M. Kapranov, *The permutoassociahedron, Mac Lane's coherence theorem and asymptotic zones for the KZ equation*, Journal of Pure and Applied Algebra, 85 (1993), pp. 119–142.
38. L. G. Kraft, *A device for quantizing, grouping, and coding amplitude modulated pulses*, Master's thesis, Electrical Engineering Department, Massachussetts Institute of Technology, 1949.
39. V. Lefort, R. Desper, and O. Gascuel, *FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program*, Molecular Biology and Evolution, 32 (2015), pp. 2798–2800.
40. B. McMillan, *Two inequalities implied by unique decipherability*, IEEE Transactions on Information Theory, 2 (1956), pp. 115–116.
41. K. P. Murphy, *Probabilistic Machine Learning*, The MIT Press, London, UK, 2022.
42. R. D. M. Page and E. C. Holmes, *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science, Oxford, UK, 1998.
43. F. Pardi, *Algorithms on Phylogenetic Trees*, PhD thesis, University of Cambridge, UK, 2009.

44. A. N. Patrinos and S. L. Hakimi, *The distance matrix of a graph and its tree realization*, Quaterly of Applied Mathematics, 30 (1972), pp. 255–269.
45. Y. Pauplin, *Direct calculation of a tree length using a distance matrix*, Journal of Molecular Evolution, 51 (2000), pp. 41–47.
46. J. M. S. S. Pereira, *A note on the tree realizability of a distance matrix*, Journal of Combinatorial Theory, 6 (1969), pp. 303–310.
47. V. Reiner and G. M. Ziegler, *Coxeter-associahedra*, Zuse Institute Berlin (ZIB), Berlin, Takustrasse 7, 14195, Germany, 41 (1994), pp. 364–393.
48. N. Saitou and M. Nei, *The neighbor-joining method: A new method for reconstructing phylogenetic trees*, Molecular Biology and Evolution, 4 (1987), pp. 406–425.
49. C. Semple and M. A. Steel, *Phylogenetics*, Oxford University Press, New York, NY, 2003.
50. C. Semple and M. A. Steel, *Cyclic permutations and evolutionary trees*, Advances in Applied Mathematics, 32 (2004), pp. 669–680.
51. M. Steel, *Phylogeny: Discrete and random processes in evolution*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016.
52. M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer, *Additive evolutionary trees*, Journal of Theoretical Biology, 64 (1977), pp. 199–213.

Dataset	n	UBT-BC				MP-BB			Dataset	n	UBT-BC				MP-BB		
		Time (s)	Nodes	Root Gap (%)	Cuts	Time (s)	Nodes	Root Gap (%)			Time (s)	Nodes	Root Gap (%)	Cuts	Time (s)	Nodes	Root Gap (%)
	10	**0.15**	**1**	**1.79**	18	0.16	216	2.57		10	0.18	**1**	**0.91**	23	**0.13**	347	1.79
Primates12	11	0.51	**1**	**1.84**	21	**0.18**	344	2.59		11	0.45	**1**	**1.30**	17	**0.18**	740	1.61
	12	0.49	**1**	**1.97**	33	**0.27**	579	2.47		12	0.85	**1**	**1.75**	30	**0.22**	1360	2.2
			1							13	1.12	**1**	**1.94**	34	**0.29**	3089	2.33
	10	0.6	**1**	**0.36**	14	**0.22**	1717	1.05		14	1.83	**1**	**2.09**	38	**0.34**	3692	2.27
	11	0.87	**1**	**0.51**	24	**0.26**	2779	1.18	M62	15	4.34	**1**	**2.41**	85	**0.54**	10618	2.46
	12	0.64	**1**	**0.49**	10	**0.3**	4562	1.21		16	5.82	**1**	**2.34**	92	**0.78**	21728	2.5
	13	1.79	**1**	**0.95**	30	**0.69**	21356	1.65		17	11.35	**1**	**2.30**	119	**1.38**	56359	2.94
M17	14	2.73	**1**	**0.99**	30	**0.45**	4239	1.65		18	28.97	**1**	**1.89**	164	**2.14**	104840	2.9
	15	4	**1**	**1.02**	53	**1.41**	44108	1.65		19	86.93	**98**	**1.92**	128	**5.56**	284194	2.97
	16	4.72	**1**	**1.08**	39	**2.02**	56993	1.7		20	110.97	**116**	**1.90**	143	**11.58**	672960	4.63
	17	6.13	**1**	**1.08**	70	**2.55**	56012	1.64									
			1							10	2.2	**1**	**0.84**	43	**0.14**	1401	1.07
	10	0.49	**1**	**0.41**	5	**0.2**	1003	2.11		11	1.23	**1**	**0.70**	30	**0.19**	1975	1.5
	11	1.17	**1**	**1.16**	13	**0.21**	1799	2.57		12	3.24	**1**	**1.13**	54	**0.29**	7348	2.06
	12	1.45	**1**	**1.35**	22	**0.31**	4196	2.71		13	18.41	**24**	**1.53**	87	**0.54**	20086	2.21
	13	6.19	**1**	**1.82**	54	**0.62**	17600	3.27		14	23.22	**58**	**1.38**	69	**0.79**	39870	2.35
M18	14	22.77	**8**	**2.20**	69	**1.22**	44537	3.62	RbcL55	15	40.29	**238**	**1.71**	104	**1.45**	67940	2.64
	15	27.46	**1**	**2.22**	108	**2.49**	98974	3.47		16	57.12	**497**	**1.77**	101	**2.15**	102957	2.54
	16	47.03	**1**	**2.29**	130	**3.07**	126893	3.51		17	147.69	**1008**	**2.15**	164	**9.01**	616828	2.5
	17	54.73	**24**	**2.40**	125	**5.23**	214950	3.53		18	239.54	**1096**	**2.09**	182	**18.07**	1243608	2.14
	18	95.11	**115**	**2.52**	163	**14.32**	618454	3.55		19	242.21	**754**	**2.14**	174	**23.89**	1508113	2.19
										20	188.97	**149**	**2.17**	187	**61.6**	4152495	2.81
	10	0.6	**1**	**1.96**	18	**0.19**	825	5.51									
	11	1.12	**1**	**2.16**	39	**0.21**	1070	5.1		10	0.43	**1**	**1.65**	9	**0.16**	2049	5.74
	12	1.47	**1**	**2.18**	44	**0.24**	595	4.87		11	0.43	**1**	**4.41**	8	**0.21**	1829	4.94
	13	2.58	**1**	**2.51**	32	**0.35**	2145	5.3		12	0.98	**1**	**4.35**	40	**0.23**	1868	4.64
	14	3.57	**1**	**2.46**	45	**0.48**	4638	5.06		13	2	**1**	**5.17**	56	**0.29**	2686	5.39
SeedPlant500	15	6.86	**1**	**2.41**	34	**0.69**	9794	4.66		14	2.73	**1**	**5.43**	47	**0.37**	3030	5.6
	16	30	**12**	**3.35**	100	**0.85**	12160	5.76	Rana64	15	5.38	**1**	**6.30**	62	**0.5**	4092	6.47
	17	67.92	**47**	**3.93**	79	**1.56**	54398	6.21		16	14.74	**1**	**7.62**	96	**0.6**	8126	7.78
	18	100.53	**380**	**4.59**	136	**5.12**	243902	6.92		17	17.38	**1**	**7.83**	84	**0.71**	9516	8.01
	19	106.42	**179**	**4.79**	208	**13.96**	648335	6.97		18	49.74	**17**	**4.45**	137	**0.95**	18914	5.07
	20	191.41	**372**	**5.12**	236	**61.85**	2894260	9.11		19	48.13	**14**	**3.95**	96	**1.33**	32195	4.48
										20	85.58	**23**	**4.07**	224	**1.76**	42552	4.74
	10	0.34	**1**	**0.35**	5	**0.13**	532	1.29									
	11	0.52	**1**	**0.89**	8	**0.21**	4884	2.39		10	1.3	**1**	**1.99**	20	**0.16**	4005	4.39
	12	0.71	**1**	**0.74**	11	**0.25**	4883	1.79		11	6.65	**1**	**2.31**	21	**0.37**	43651	4.52
	13	0.97	**1**	**0.79**	12	**0.32**	8303	1.74		12	9.27	**1**	**2.42**	49	**0.52**	46440	4.43
	14	1.34	**1**	**0.87**	18	**0.43**	12731	1.75		13	5.97	**1**	**2.08**	41	**1.54**	179717	3.92
M43	15	3.04	**1**	**0.77**	37	**0.68**	22028	1.53		14	19.78	**47**	**2.82**	62	**9.22**	1272759	4.51
	16	3.77	**1**	**0.73**	27	**1.15**	51773	1.51	M82	15	27.97	**46**	**4.45**	56	**34.45**	4539819	6.26
	17	4.77	**1**	**1.04**	39	**2.52**	121001	1.82		16	38.48	**113**	**3.67**	106	**28.37**	2700408	5.13
	18	6.08	**1**	**0.74**	29	**2.27**	86731	1.44		17	75.04	**227**	**3.40**	120	**35.69**	3087231	4.76
	19	9.07	**1**	**0.78**	31	**3.6**	157449	1.43		18	114.94	**442**	**3.90**	177	**85.37**	6873457	5.53
	20	12.43	**1**	**0.87**	67	**7.21**	336111	1.48		19	**111.61**	**313**	**4.03**	154	135.82	9393814	6.36
										20	**263.91**	**830**	**4.58**	267	299.37	19360753	8.21

Table 8 Comparison of the performance obtained by UBT-BC and MP-BB on the Easy Set.

Dataset	n	Optimum	UBT-BC						MP-BB			
			Best Solution	Root Gap (%)	Final Gap (%)	Time (s)	Nodes	Cuts	Best Solution	Root Gap (%)	Time (s)	Nodes
SeedPlant500	21	332056.369	332056.369	**5.258**	0.000	381.72	**943**	296	332056.369	7.296	**130.88**	4711486
	22	671163.813	671163.813	**5.693**	0.000	1523.26	**3512**	536	671163.813	7.659	**215.15**	7349038
	23	1372759.962	1372759.962	**6.288**	0.414	> 3600	**2154**	1253	1372759.962	8.237	**903.17**	31488228
	24	2796739.184	2796739.184	**6.469**	0.821	> 3600	**898**	620	2796739.184	8.259	**1581.07**	67844153
	25	unknown	**5939815.36158**	**7.178**	-	-	-	-	6171362.306	-	> 3600	498553098
	26	unknown	**12440715.6569**	**7.806**	-	-	-	-	13010792.442	-	> 3600	468648444
M43	21	325438.135	325438.135	**0.852**	0.000	16.92	**1**	77	325438.135	1.448	**13.99**	703824
	22	656592.407	656592.407	**0.874**	0.000	**21.11**	**1**	82	656592.407	4.451	> 3600	191231641
	23	1318516.260	1318516.260	**0.894**	0.000	**25.16**	**1**	105	1318516.260	4.460	> 3600	516803942
	24	2789865.680	**2789865.680**	**1.055**	0.000	**58.99**	**1**	123	-	4.608	-	-
	25	5688555.528	**5688555.528**	**1.088**	0.000	**62.11**	**1**	152	-	4.497	-	-
	26	12008676.691	**12008676.691**	**1.353**	0.000	**275.17**	**1**	396	-	4.801	-	-
	27	25409724.737	**25409724.737**	**1.751**	0.000	**1795.69**	**1263**	308	-	4.306	-	-
	28	54171504.267	**54171504.267**	**1.902**	0.000	**2132.72**	**745**	288	-	4.030	-	-
	29	111211685.285	**111211685.285**	**1.868**	0.000	**2388.34**	**733**	342	-	3.950	-	-
	30	242403729.428	**242403729.428**	**1.755**	0.000	**2308.00**	**566**	368	-	3.549	-	-
	31	unknown	**491730541.178**	**1.914**	0.093	> 3600	**591**	450	-	-	-	-
	32	unknown	**1016257186.290**	**2.374**	0.366	> 3600	**220**	585	-	-	-	-
M62	21	606171.435	606171.435	**0.789**	0.000	94.93	**87**	149	606171.435	4.868	**9.71**	448109
	22	1220818.790	1220818.790	**0.820**	0.000	**170.42**	**123**	200	1228200.387	4.803	> 3600	529375621
	23	2482880.895	2482880.895	**1.012**	0.000	**245.81**	**223**	205	2498066.183	4.703	> 3600	502405960
	24	5132331.040	**5132331.040**	**0.981**	0.000	**517.00**	**725**	267	-	4.463	-	-
	25	10408638.046	**10408638.046**	**1.056**	0.000	**1267.88**	**959**	294	-	4.623	-	-
	26	21428870.595	**21428870.595**	**1.479**	0.000	**1797.26**	**636**	323	-	4.065	-	-
	27	43823281.741	**43823281.741**	**1.301**	0.000	**1881.49**	**1045**	335	-	4.007	-	-
	28	90114762.830	**90114762.830**	**1.401**	0.081	> 3600	**785**	389	-	4.052	-	-
	29	184186545.195	**184186545.195**	**1.408**	0.115	> 3600	**686**	461	-	4.067	-	-
RbcL55	21	493899.415	493899.415	**0.143**	0.000	263.72	**307**	193	493899.415	2.789	**102.63**	6238233
	22	1023571.265	1023571.265	**0.129**	0.000	245.52	**185**	211	1023571.265	2.733	**214.69**	12132748
	23	2106749.004	2106749.004	**0.147**	0.000	501.94	**300**	236	2106749.004	2.739	**574.82**	30025787
	24	4281426.961	4281426.961	**0.149**	0.000	**720.93**	**391**	208	4281426.961	3.165	> 3600	501709276
	25	8866707.383	8866707.383	**0.158**	0.000	**842.17**	**336**	262	8866707.383	3.053	> 3600	454075005
	26	17848947.626	**17848947.626**	**0.153**	0.000	**1106.10**	**319**	263	-	3.030	-	-
	27	36912429.716	**36912429.716**	**0.167**	0.000	**3365.02**	**823**	346	-	2.832	-	-
	28	75236032.385	**75236032.385**	**0.160**	0.000	**2054.03**	**459**	321	-	2.769	-	-
	29	unknown	**156595052.507**	**0.171**	0.062	> 3600	**359**	383	-	-	-	-
	30	unknown	**321071032.462**	**0.160**	0.093	> 3600	**272**	400	-	-	-	-
Rana64	21	186092.562	186092.562	**0.259**	0.000	107.11	**46**	149	186092.562	3.476	**3.17**	101136
	22	385972.325	385972.325	**0.327**	0.000	126.08	**48**	179	385972.325	3.528	**4.36**	155438
	23	805992.557	805992.557	**0.420**	0.000	214.25	**167**	282	805992.557	5.224	**10.69**	302983
	24	1643729.224	1643729.224	**0.432**	0.018	> 3600	**6056**	781	1643729.224	5.184	**17.16**	457205
	25	3579365.445	3579365.445	**0.389**	0.000	469.49	**196**	284	3579365.445	4.645	**46.14**	1519166
	26	7634137.020	7634137.020	**0.533**	0.000	1617.32	**1013**	368	7634137.020	4.333	**186.42**	6417881
	27	16430070.905	16430070.905	**0.826**	0.306	> 3600	**489**	523	16430070.905	4.288	**1537.49**	49425353
	28	unknown	35712637.346	**0.789**	0.509	> 3600	**301**	598	35712637.346	-	> 3600	173867803
	29	unknown	72530903.6739	**2.651**	-	-	-	-	72530903.674	-	> 3600	286501741
M82	21	178851.725	178851.725	**0.420**	0.000	**331.53**	**456**	201	178851.725	5.432	834.25	49432404
	22	392087.422	392087.422	**0.566**	0.000	**972.23**	**1502**	337	392087.422	5.900	2807.91	146436375
	23	798955.380	798955.380	**0.582**	0.000	**1746.20**	**2887**	426	798955.380	5.819	> 3600	245346036
	24	unknown	**1645768.732**	**0.551**	0.113	> 3600	**2541**	593	1645768.732	-	> 3600	339327327
	25	unknown	**3411137.819**	**0.694**	0.117	> 3600	**1715**	542	3461590.253	-	-	-

Table 9 Comparison of the performance obtained by UBT-BC and MP-BB on the Hard Set.

Dataset	n	Optimum	UBT-BC					MP-BB			
			Best Solution	Time (s)	Nodes	Final Gap (%)	Cuts	Best Solution	Time (s)	Nodes	Final Gap (%)
B-HA-573-585-BMGE	20	530072.565	530072.565	17.18	**1**	**0.000**	120	530072.565	**1.11**	8568	2.566
	25	17880701.513	17880701.513	165.72	**1**	**0.000**	265	17880701.513	**2.83**	30004	2.440
	26	35631582.718	35631582.718	225.92	**13**	**0.000**	257	35631582.718	**5.08**	79308	2.488
B-HA-986-1908-BMGE	20	432217.764	432217.764	69.17	**54**	**0.000**	163	432217.764	**56.41**	3654676	8.421
	21	869072.877	869072.877	146.75	**239**	**0.000**	233	869072.877	**143.93**	7748550	9.450
B-NA-633-1751-BMGE	20	288685.631	288685.631	17.32	**1**	**0.000**	245	288685.631	**2.64**	73818	11.981
	22	1270085.105	1270085.105	65.04	**1**	**0.000**	399	1270085.105	**11.99**	438241	12.800
	23	2660926.474	2660926.474	97.31	**8**	**0.000**	450	2660926.474	**50.96**	1658515	13.114
B-NA-684-472-BMGE	20	554220.083	554220.083	**7.18**	**1**	**0.000**	112	554220.083	9.01	325389	3.452
	24	9467537.210	9467537.210	**46.09**	**1**	**0.000**	214	9467537.210	243.67	5775880	4.252
	25	18911659.051	18911659.051	**53.23**	**1**	**0.000**	225	18911659.051	320.88	6525723	4.233
B-NS-629-1111-BMGE	20	284088.517	284088.517	23.58	**1**	**0.000**	234	284088.517	**2.63**	70783	8.608
	26	19225291.111	19225291.111	227.73	**62**	**0.000**	774	19225291.111	**16.26**	436825	10.313
	27	39704665.778	39704665.778	313.39	**221**	**0.000**	614	39704665.778	**88.56**	2520022	10.069
B-NS1-284-344-BMGE	20	600436.354	600436.354	**8.54**	**1**	**0.000**	118	600436.354	29.96	1652823	4.686
	29	341239418.934	341239418.934	**1767.97**	**539**	**0.000**	412	342683792.853	> 3600	62068495	7.083
	30	unknown	689965539.914	> 3600	**955**	**0.056**	490	692918790.535	> 3600	57964421	7.733
eudicots-BMGE	20	1294680.314	1294680.314	92.57	**1**	**0.000**	298	1294680.314	**28.58**	2170084	2.386
euros1	20	964061.058	964061.058	81.79	**34**	**0.000**	186	964061.058	**4.28**	243086	3.753
euros2	20	716214.753	716214.753	8.06	**1**	**0.000**	60	716214.753	**1.13**	8946	1.083
	22	2949200.758	2949200.758	20.04	**1**	**0.000**	117	2949200.758	**2.23**	39618	1.100
	23	5959679.541	5959679.541	33.02	**1**	**0.000**	166	5959679.541	**4.53**	105950	1.760
nucleic-M2573-346x897-2006	20	1324915.878	1324915.878	39.74	**20**	**0.000**	230	1324915.878	**2.15**	87532	2.246
	27	171652350.870	171652350.870	417.45	**74**	**0.000**	469	171652350.870	**45.79**	1487350	2.916
	28	345731096.684	345731096.684	724.91	**392**	**0.000**	424	345731096.684	**56.80**	1720310	2.881
nucleic-M2839-470x829-2006	20	3599570.958	3599570.958	**256.40**	**642**	**0.000**	214	3782376.821	> 3600	304313225	8.540
	21	7507514.791	7507514.791	**275.05**	**321**	**0.000**	151	7892763.278	> 3600	235795333	8.821
	22	15439953.018	15439953.018	**362.73**	**455**	**0.000**	180	16254955.369	> 3600	250393665	9.069
nucleic-M3862-362x1207-2008	20	3102271.026	3102271.026	**107.73**	**29**	**0.000**	153	3121879.050	> 3600	257302144	3.208
proteic-M2577-40x12260-2005	20	4793364.627	4793364.627	**44.55**	**1**	**0.000**	206	4854143.148	> 3600	198043972	7.531
	23	41973145.938	41973145.938	**261.53**	**242**	**0.000**	226	42504306.609	> 3600	180256775	8.460
	24	86565511.835	86565511.835	**438.50**	**370**	**0.000**	255	87092956.578	> 3600	166971241	8.296
proteic-M2624-139x338-2006-BMGE	20	3569327.875	3569327.875	10.60	**1**	**0.000**	61	3569327.875	**2.69**	73409	1.432
	25	125549247.540	125549247.540	**291.79**	**21**	**0.000**	296	125893577.142	> 3600	138432308	3.161
	26	257545396.454	257545396.454	**701.99**	**800**	**0.000**	267	258277857.349	> 3600	122852779	3.222
proteic-M2883-91x7386-2007-BMGE	20	unknown	2183208.518	> 3600	**5943**	**0.124**	5429	2188398.985	> 3600	358856805	6.032
proteic-M3068-50x1000-2006	20	24638785.283	24638785.283	**385.83**	**1557**	**0.000**	365	24981920.418	> 3600	197997696	3.816
proteic-M3756-77x9918-2008-BMGE	20	5063227.006	5063227.006	25.10	**1**	**0.000**	127	5063227.006	**11.45**	634147	1.004
	22	21113170.305	21113170.305	**65.18**	**1**	**0.000**	175	21113170.305	89.13	4145337	1.058
	23	43181421.785	43181421.785	**63.64**	**1**	**0.000**	143	43190975.673	123.28	5235313	1.082
proteic-M3756-77x11234-2008	20	6639947.471	6639947.471	**117.26**	**116**	**0.000**	231	6756521.943	> 3600	286309546	4.168
rosids	20	1036298.579	1036298.579	100.27	**1**	**0.000**	289	1036298.579	**82.60**	6305076	3.738

Table 10 Comparison of the performance of UBT-BC and MP-BB on the Third Set of instances of the BMEP.

$\tau_{1,2}$	$\tau_{1,3}$	$\tau_{1,4}$	$\tau_{1,5}$	$\tau_{1,6}$	$\tau_{2,3}$	$\tau_{2,4}$	$\tau_{2,5}$	$\tau_{2,6}$	$\tau_{3,4}$	$\tau_{3,5}$	$\tau_{3,6}$	$\tau_{4,5}$	$\tau_{4,6}$	$\tau_{5,6}$
2	3	4	5	5	3	4	5	5	3	4	4	3	3	2
3	3	3	4	4	2	4	5	5	4	5	5	3	3	2
4	3	4	3	3	3	2	5	5	3	4	4	5	5	2
3	2	4	5	5	3	3	4	4	4	5	5	3	3	2
4	3	2	5	5	3	4	3	3	3	4	4	5	5	2
2	4	3	5	5	4	3	5	5	3	3	3	4	4	2
3	4	3	3	4	3	4	4	3	5	5	2	2	5	5
3	4	4	3	3	5	5	4	2	2	3	5	3	5	4
5	5	4	3	2	2	3	4	5	3	4	5	3	4	3
2	4	4	4	4	4	4	4	4	2	4	4	4	4	2
4	2	4	4	4	4	2	4	4	4	4	4	4	4	2
4	4	2	4	4	2	4	4	4	4	4	4	4	4	2
3	5	5	4	2	4	4	3	3	2	3	5	3	5	4
4	3	3	3	4	5	5	3	2	2	4	5	4	5	3
5	5	2	3	4	2	5	4	3	5	4	3	3	4	3

Table 11 The set $\tau(\mathcal{B}^6)$. Each row refers to the upper triangular part of a PLM whose associated UBT $T \in \mathcal{B}^6$. It is easy to check, by using numerical tools, that each row (i.e., each PLM) satisfies (48) at equality and that the set $\tau(\mathcal{B}^6)$ is linearly independent.

A – On triangle inequalities

The strong triangle inequalities (2) are facet-defining for $Conv(\Theta_n)$. The proof of this claim, being very similar to that of Proposition 15, was omitted from the main body of the paper and is given in this appendix for completeness.

Proposition 22 *For $n \geq 6$, the tightened triangle inequality*

$$\Delta_{n-1}^{n-2,n} = \tau_{n-2,n-1} + \tau_{n-1,n} - \tau_{n-2,n} \geq 2 \tag{48}$$

is facet-defining for $Conv(\Theta)$.

Proof We prove the statement by inductively constructing a set $\mathcal{B}^n$ of $\binom{n}{2}$ UBTs such that $\tau(\mathcal{B})$ is linearly independent. Specifically, we first show that the statement holds true for $n = 6$. Then, by assuming that the statement holds true for $n-1$, we show that we can derive from $\mathcal{B}^{n-1}$ a set $\mathcal{B}' \subset \mathcal{U}_n$ of $\binom{n-1}{2}$ UBTs whose induced PLMs are linearly independent and satisfy (48) at equality. Finally, we show that we can construct the set $\mathcal{B}^n$ by joining $\mathcal{B}'$ with a specific set of $n-1$ caterpillars that satisfy (48) at equality.

Base case. Assume that $n = 6$ and denote by $\mathcal{B}^6$ the set of UBTs in $\mathcal{U}_6$ associated with the PLMs reported in Table 11. Using numerical tools (e.g., Mathematica or Matlab), it is easy to check that $\tau(\mathcal{B}^6)$ is linearly independent and each of its PLMs satisfies (48) at equality. Hence, (48) is facet-defining for $Conv(\Theta_6)$.

Inductive step. Now, assume that $n > 6$ and that the statement holds true for $n-1$. Denote $\mathcal{B}^{n-1}$ as a set of $\binom{n-1}{2}$ UBTs in $\mathcal{U}_{n-1}$ whose induced PLMs $\tau \in \tau(\mathcal{B}^{n-1})$ satisfy

$$\tau_{n-3,n-2} + \tau_{n-2,n-1} - \tau_{n-3,n-1} = 2. \tag{49}$$

Consider the set

$$\mathcal{B}_e = \{\text{expand}(T, n-1) \in \mathcal{U}_n : \ T \in \mathcal{B}^{n-1}\}$$

and observe that, by Proposition 14, $\tau(\mathcal{B}_e)$ is linearly independent. Consider a UBT $T \in \mathcal{B}^{n-1}$ and define

$$T_e = \text{expand}(T, n-1).$$

Then, by the properties of the expansion operation, the following conditions hold:

$$\begin{aligned}
\tau(T_e)_{u,n} = \tau(T_e)_{u,n-1} &= \tau(T)_{u,n-1} + 1 && \forall u \in [n-2] \\
\tau(T_e)_{u,v} &= \tau(T)_{u,v} && \forall u, v \in [n-2] \\
\tau(T_e)_{n,n-1} &= 2 && \\
\Delta_{n-1}^{k,n}(T_e) &= 2 && \forall k \in [n-2].
\end{aligned}$$

Note, in particular, that $\Delta_{n-1}^{n-2,n}(T_e) = 2$ for all $T_e \in \mathcal{B}_e$, so each PLMs induced by these UBTs satisfies (49). We now construct $\mathcal{B}^n$ by complementing the $\binom{n-1}{2}$ UBTs in $\mathcal{B}_e$ with a set $n-1$ of caterpillars $T \in \mathcal{U}_n$ that satisfy (48) at equality and that are such that $\tau(\mathcal{B}^n)$ is linearly independent. First, consider the caterpillar $Y = \text{swap}^{(n-2,n-3)}(F^n)$, which is determined by the sequence $(1, 2, \cdots, n-4, n-2, n-3, n-1, n)$. As $n-1$ and n are in the same cherry, we have that $\tau(Y)_{n,n-1} = 2$. This property also holds for each element in $\mathcal{B}_e$. However, while $\Delta_{n-2}^{n,n-3}(T_e) = 2$ for each $T_e \in \mathcal{B}_e$, we have that $\Delta_{n-2}^{n,n-3}(Y) > 2$ because, due to the swap, leaf $n-2$ is not adjacent to an internal node in the path $P_{n-3,n}^Y$. Hence, by Proposition 12, $\tau(\mathcal{B}' \cup \{Y\})$ is linearly independent. Now, for each $k \in [n-3]$, consider the

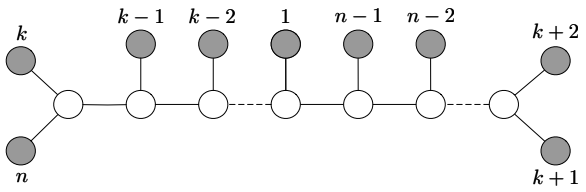


Fig. 11 The caterpillar C_k^n

caterpillar C_k^n determined by sequence $(n, k, k-1, k-2, \cdots, 2, 1, n-1, n-2, \cdots, k+2, k+1)$, shown in Figure 11. Observe that if $k' \in [k]$ then k' is adjacent to an internal node in the path $P_{n,n-1}^{C_k^n}$; alternatively, if $k' \in [k+1, n-2]$, then $n-1$ is adjacent to an internal node in the path $P_{n,k'}^{C_k^n}$. We can then conclude that

$$\Delta_{n-1}^{n,k'}(C_k^n) = 2 \text{ for } k' \in [k+1, n-2] \tag{50}$$

and

$$\Delta_{n-1}^{n,k'}(T) > 2 \text{ for } k' \in [k]. \tag{51}$$

Observe that, for each $k \in [n-3]$, $\Delta_{n-1}^{n,k}(T) = 2$ for each $T \in \mathcal{B}' \cup \{Y\} \cup_{k'=1}^{k-1} \{C_{k'}^n\}$, while $\Delta_{n-1}^{n,k}(C_k^n) > 2$. In particular, by Proposition 12, we have that $\tau(\mathcal{B} \cup \{Y, C_1^n, C_2^n\})$ is linearly independent and for $T \in \mathcal{B} \cup \{Y, C_1^n, C_2^n\}$ and $k \in [3, n-2]$ we have $\Delta_{n-1}^{n,k}(T) = 2$. Now, consider the linear function

$$F(T) = 2(\tau(T)_{1,n} - \tau(T)_{1,n-1}) + (\tau(T)_{2,n} - \tau(T)_{2,n-1}) + \tau(T)_{n-1,n}.$$

For $T \in \mathcal{B}' \cup \{Y\}$, $n-1$ and n are in the same cherry, hence

$$F(T) = 0 + 0 + 2 = 2.$$

We also have that

$$F(C_1^n) = 2(2-3) + (n-1-(n-2)) + 3 = -2 + 1 + 3 = 2.$$

Similarly, we have that

$$F(C_2^n) = 2(3-3) + (2-4) + 4 = 0 - 2 + 4 = 2.$$

Hence, $F(T) = 2$ for each $T \in \mathcal{B} \cup \{Y, C_1^n, C_2^n\}$. Now, consider the caterpillar $W = \text{swap}^{(1,2)}(C_2^n)$, determined by the sequence $(n, 1, 2, n-1, \cdots, 5, 4, 3)$ and shown in Figure 12. For W we have that

$$F(W) = 2(2-4) + (3-3) + 4 = 0$$

and $\Delta_{n-1}^{n,k}(W) = 2$, for $k \in [3, n-2]$. Hence, by Proposition 12, $\tau(\mathcal{B} \cup \{Y, C_1^n, C_2^n, W\})$ is linearly independent.

Let $\mathcal{B}^n$ denote the set

$$\mathcal{B}^n = \mathcal{B}' \cup \{Y, W\} \cup_{k=1}^{n-3} \{C_k^n\}.$$

We will show now that the set $\tau(\mathcal{B}^n)$ is linearly independent. To this end, suppose, by contradiction, that $\tau(\mathcal{B}^n)$ is linearly dependent and denote by j be the smallest index such that the PLMs induced by the set $\mathcal{B}_{LD} = \mathcal{B} \cup \{Y, W\} \cup_{k=1}^{j} C_k^n$ are linearly dependent. Because (50) and (51) hold and $j > 2$, we have that $\Delta_{n-1}^{n,k'}(T) = 2$ for all $T \in \mathcal{B}_{LD}$ and $k' \in [k, n-2]$. On the other hand, note that $\Delta_{n-1}^{n,k}(C_k^n) > 2$. In the light of Proposition 12, these facts imply that $\tau(\mathcal{B}_{LD})$ is linearly independent, which leads to a contradiction. Therefore, $\tau(\mathcal{B}^n) = \tau(B_e \cup \{Y, W\} \cup_{k=1}^{n-3} \{C_k^n\})$ is linearly independent, $|\mathcal{B}^n| = \binom{n}{2}$, $\Delta_{n-1}^{n,n-2}(T) = 2$, for each $T \in \mathcal{B}^n$, and the statement follows. □

From Proposition 13, we can conclude that

Corollary 1 *For any distinct $i, j, k \in [n]$, the inequality $\Delta_k^{i,j}(T) \geq 2$ is facet-defining for $Conv(\Theta_n)$.*

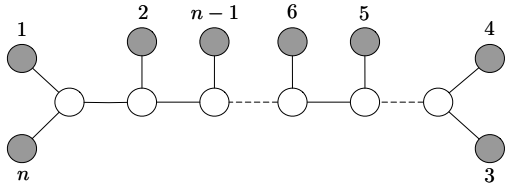


Fig. 12 The caterpillar $W = \text{swap}^{1,2}(C_2^n)$.