Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

On the Complexity of Degree-Constrained Network Design Problems with Linear and Nonlinear Objectives

Daniele Catanzaro[0000–0001–9427–1562, 1] | **Gwenaël Joret**[0000–0002–7157–6694, 2] | **Brieuc Pierre**[0009–0001–3935–5859, 1] | **Francesco Pisanu**[0000–0003–0799–5760, 1]

[1]Center for Operations Research and Econometrics, Université Catholique de Louvain, Voie du Roman Pays 34, Ottignies-Louvain-la-Neuve, 1348, Belgium

[2]Computer Science Department, Université libre de Bruxelles, Campus de la Plaine, CP 212, Brussels, 1050, Belgium

Correspondence

Corresponding author Francesco Pisanu, CORE, UCLouvain, Voie du Roman Pays 34, Ottignies-Louvain-la-Neuve, 1348.

Email: francesco.pisanu@uclouvain.be

Abstract

We show that the *Network Design Problem* remains $\mathcal{NP}$-hard even under strict degree constraints on its vertices. We also show that this result persists even in highly structured tree-based settings, where internal vertices have fixed degrees and the objective depends only on pairwise distances between leaves. This result applies to both minimization and maximization over a broad class of linear and nonlinear distance-based objectives depending on pairwise leaf distances.

KEYWORDS

network design, degree-constrained spanning trees, cubic trees, computational complexity.

1 | INTRODUCTION

The *Network Design Problem (NDP)* is a foundational problem in combinatorial optimization that consists of determining the optimal structure of a network while achieving specific objectives and simultaneously satisfying resources, capacity, demands, and budget constraints[1]. The NDP generalizes many classical optimization problems, such as the Steiner tree, the facility location, the multi-commodity flow, the shortest path, and the spanning tree problems, and serves as a unifying framework for modeling numerous network-based real-world applications[2,3,4]. In its most common formulation[1], the input graph of the NDP is loopless, undirected, and has no parallel edges. Throughout this article, we shall adopt this convention and restrict our attention exclusively to graphs of this type.

Consider a connected graph $G = (V, E)$, with vertex-set V and edge-set E, and a weight function $w\colon E \to \mathbb{Q}_0^+$ that associates each edge in E with a nonnegative rational number. For every spanning tree T of G and every two distinct vertices $u, v \in V$, let P_{uv}^T denote the path between u and v in T (we assume $P_{uu}^T = \emptyset$), and define the *weighted path-length* between u and v in T as

$$L_T^w(u, v) := \sum_{e \in E(P_{uv}^T)} w(e).$$

TABLE 1 A summary of known $\mathcal{NP}$-hard versions of the NDP.

Problem	Hardness & Approximability	Reference
NDP	$\mathcal{NP}$-hard	[11]
Minimum k-spanning tree	$\mathcal{NP}$-hard and approximable to within a factor 3	[5,9]
Maximum leaf spanning tree	$\mathcal{NP}$-hard and approximable to within a factor 3	[7,10]
Minimum degree spanning tree	$\mathcal{NP}$-hard and approximable to within a factor 3/2	[8]
Balanced Minimum Evolution Problem	$\mathcal{NP}$-hard to approximate within a factor c^n, for some $c > 1$ $\mathcal{NP}$-hard and approximable within a factor 2 for metric instances	[13]
Fixed-tree Balanced Minimum Evolution Problem	$\mathcal{NP}$-hard to approximate to within a factor c^n, for some $c > 1$	[12]

Then, for a given rational $B > 0$, the NDP consists of finding a spanning tree T of G with $T = (V, E')$ that solves the following optimization problem:

$$\min \sum_{u,v \in V} L_T^w(u, v)$$
$$\text{s.t.} \sum_{e \in E'} w(e) \leq B.$$

The $\mathcal{NP}$-hardness of the NDP was first established in 1978 by Johnson et al[11]. Since then, numerous $\mathcal{NP}$-hard variants of the problem have been proposed in the literature[5]. Noteworthy examples that are particularly relevant to this work include the *minimum k-spanning tree problem*[5], which seeks a minimum-length tree spanning at least k vertices of a given graph; the *maximum leaf spanning tree problem*[7], which aims to find a spanning tree with the maximum possible number of leaves; and the *minimum degree spanning tree problem*[8], which consists of finding a spanning tree that minimizes the maximum vertex degree. Table 1 outlines the key references for these problems, along with their known complexity and approximation results.

In this article, we focus on a family of network design problems that originate from the context of hierarchical clustering[13,12], where the goal is to identify a tree-like structure of nested clusters based on a given dissimilarity measure among a set of input items. These problems are closely related to the *Balanced Minimum Evolution Problem (BMEP)*, a highly nonlinear $\mathcal{NP}$-hard optimization problem[13] that has been extensively studied in molecular phylogenetics[15,16,17,18,19]. Given a set of items $\Gamma = \{1, 2, \ldots, n\}$, with $n \geq 3$, and an associated dissimilarity matrix $\mathbf{D} = \{d_{ij}\}$, the BMEP asks for a spanning tree T with leaf-set Γ, internal vertices of degree 3, that minimizes the length function

$$\sum_{i,j \in \Gamma} d_{ij} 2^{-\tau_{ij}^T},$$

where $d_{ii} = 0$, for all items $i \in \Gamma$, and $\tau_{ij}^T := |E(P_{ij}^T)|$ is the *path-length* between the leaves i and j of T.

The BMEP is particularly difficult to solve. State-of-the-art exact algorithms are currently unable to solve instances with more than 32 items within one hour of computing time on modern hardware[20]. This fact has motivated several research efforts aimed at reformulating specific aspects of the problem in order to improve tractability. Some of these efforts focus on transforming distance-based information into edge weights of suitable graphs, thereby reducing the problem to variants of the NDP under strict degree constraints. Most such approaches are heuristic in nature and often rely on learning or predicting edge weights rather than optimizing them explicitly[21,22,23].

Other lines of work currently under development address the exponential terms appearing in the BMEP objective function, which may cause severe numerical stability issues even for moderately sized instances. These approaches approximate the exponential terms by polynomial expressions (using, for instance, first or higher-order approximations) so as to preserve the hierarchical structure of the problem while improving numerical stability. Some of these approximations naturally give rise to equivalent maximization problems, whose computational complexity has not yet been fully characterized. Alternative, but related, lines of investigation examine whether reducing the diameter of the optimal solution to the problem by allowing internal vertices to have degree larger than 3 may alleviate numerical stability issues.

Contributions.
In this article, we address as broadly as possible some complexity issues connected to these attempts. To this end, we denote by Θ_n^k the set of trees with n leaves whose internal vertices all have degree exactly k, where n and k are positive integers such that $(n-2)/(k-2)$ is an integer equal to the number of internal vertices. Moreover, we denote by $\mathcal{G}_n^k$ the class of connected graphs admitting at least one spanning tree in Θ_n^k. Then, our contributions can be summarized as follows.

We first show in Section 3 that the following problem is strongly $\mathcal{NP}$-hard:

Problem 1 (The k-NDP)**.** Given a graph $G = (V, E)$ of $\mathcal{G}_n^k$ and a weight function $w : E \to \mathbb{Q}^+$, find a spanning tree T of G belonging to Θ_n^k that minimizes $\sum_{u,v\in V} L_T^w(u, v)$.

We then study the complexity of two versions of the NDP that arise when the objective function depends on the number of edges in the path between pairs of leaves. To formulate them, consider the following connected graph $G_n^k = (\Gamma \cup V_I, E)$. The sets Γ and V_I are disjoint, with $|\Gamma| = n$, $|V_I| = (n-2)(k-2)$, and the edge set is

$$E = \{uv : u, v \in V_I,\ u \neq v\} \ \cup\ \{uv : u \in \Gamma,\ v \in V_I\}.$$

We say that T is a Γ-*tree* of G_n^k, if T is a spanning tree of G_n^k belonging to Θ_n^k and has Γ as leaf-set. Figure 1 shows an example of a possible Γ-tree for a graph G_n^k with $n = 8$ and $k = 3$.

Now, let $\mathbf{D}_\Gamma$ denote a symmetric matrix of order $n = |\Gamma|$, whose entry d_{ij} belongs to $\mathbb{Q}_0^+$ is a measure of the dissimilarity associated with the vertices $i, j \in \Gamma$. Then, we consider the following two problems:

Problem 2 (The Min-(k, τ)-NDP)**.** Given G_n^k and a matrix $\mathbf{D}_\Gamma = \{d_{ij}\}$, find a Γ-tree T that minimizes $\sum_{i,j\in\Gamma} d_{ij}\tau_{ij}^T$.

Problem 3 (The Max-(k, τ)-NDP)**.** Given G_n^k and a matrix $\mathbf{D}_\Gamma = \{d_{ij}\}$, find a Γ-tree T that maximizes $\sum_{i,j\in\Gamma} d_{ij}\tau_{ij}^T$.

We show in Sections 4 and 5 that Problems 2 and 3 are $\mathcal{NP}$-hard. To simplify the relative proofs, we will first present the complexity results for the special case $k = 3$; we will then extend these results to generic $k \geq 3$. Building on these results, in Section 6 we explicitly address several reformulations that have been proposed in the literature to improve the tractability of the BMEP. We show that $\mathcal{NP}$-hardness persists when the exponential objective function is replaced by polynomial (linear or nonlinear) approximations. In particular, we prove $\mathcal{NP}$-hardness for both the minimization and maximization variants induced by these polynomial approximations. Moreover, we highlight that these results hold even when the topology is fixed. For ease of reading and cross-referencing, we summarize in Table 2 the main novel complexity results presented in this article.

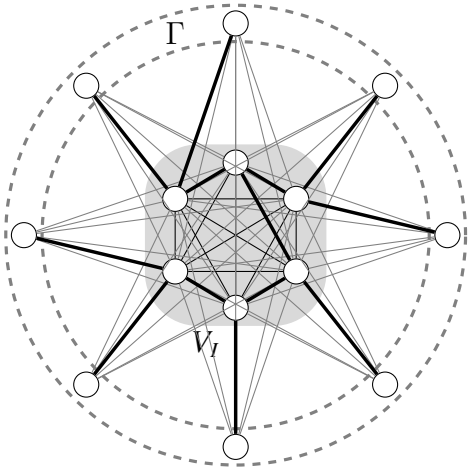

FIGURE 1 The graph $G_8^3 = (\Gamma \cup V_I, E)$. The vertices in Γ are arranged along the dashed annulus, while the vertices in V_I form a clique depicted inside the shaded region. Each vertex in Γ is adjacent to all vertices in V_I. Bold edges form a Γ-tree.

TABLE 2 New $\mathcal{NP}$-hardness results for specific versions of the NDP proved in this article.

Problem	Reference	Problem	Reference
k-NDP	Theorem 1	Leaf-restricted k-NDP	Theorem 6
Min-(k,τ)-NDP	Theorem 2	Fixed-tree Min-(k,τ)-NDP	Theorem 7
Max-(k,τ)-NDP	Theorem 5	Fixed-tree Max-(k,τ)-NDP	Theorem 8

Before proceeding, we introduce further notation and definitions that will prove instrumental in the remainder of the article.

2 | NOTATION AND DEFINITIONS

Let $G = (V, E)$ be a graph with vertex-set V and edge-set E, respectively. We denote the vertex-set and edge-set of a subgraph H of G by $V(H)$ and $E(H)$, respectively. Let $U \subseteq V$ and $E' \subseteq E$ be two nonempty subsets of vertices and edges of G, respectively, and let $E(U) = \{uv \in E\colon u, v \in U\}$ and $V(E') = \{u, v \in V\colon uv \in E'\}$. Then, we denote by $G[U] = (U, E(U))$ the subgraph of G induced by U; by $G \setminus U$ the subgraph $G[V \setminus U]$ induced by $V \setminus U$; and by $\delta(U)$ the *cut of* U, i.e., the subset of edges $uv \in E$ such that exactly one of u and v belongs to U. For the sake of notation, when $U = \{u\}$ we simply write $\delta(u)$ instead of $\delta(\{u\})$. For a fixed vertex $u \in V$, we denote by $|\delta(u)|$ the *degree* of u and by $\delta(G)$ the minimum among all degrees of the vertices of G. Finally, if the subsets $U, W \subset V$ are disjoint, then we denote by $\delta(U, W)$ the *cut of* U *and* W, defined as the subset of edges $uv \in E$ such that $u \in U$ and $v \in W$.

We say that two given graphs $G' = (V', E')$ and $G'' = (V'', E'')$ are *isomorphic* if they satisfy the standard notion of graph isomorphism provided in [24].

Let $T = (V, E)$ be a tree. The *root* $r(T)$ of T is a chosen vertex of T. A vertex of T is *internal* if it has degree at least 2, while a *leaf* of T is a vertex of degree exactly 1. We denote the set of leaves of T by $\Gamma(T)$. The *lowest common ancestor in* T of two vertices u and v in V is the unique internal vertex x of T such that

$$\tau^T_{ux} = \min\{\tau^T_{ux'}\colon \tau^T_{ux'} = \tau^T_{vx'}, \text{ for all } x' \in V\}.$$

The *height* of a tree is the maximum among the path-lengths between the root and every leaf. A *cherry* in T is a pair of distinct leaves of T that are adjacent to their lowest common ancestor.

A tree $T' = (V', E')$ is a *subtree of* T if $V' \subseteq V$ and $E' = E(V')$, while T' is a $\Gamma(T)$-*subtree* if T' is a subtree of T and $\Gamma(T') \subseteq \Gamma(T)$. Two subtrees of T are *disjoint* if their vertex sets are disjoint.

A tree $T = (V, E)$ is k-*ary*, for some integer $k \geq 3$, if each internal vertex of T has degree at most k. By definition, every tree in Θ^k_n is k-ary, and we refer to trees in Θ^3_n as *cubic*.

Throughout this paper, we consider rooted trees as subtrees of trees in Θ^k_n and therefore introduce several additional definitions. A k-ary tree T is said to be *full* if $|\delta(r(T))| = k - 1$ and $|\delta(u)| = k$ for every internal vertex $u \neq r(T)$, while T is *perfect* if it is full and the path-length between the root and any leaf equals the height of the tree. By definition, any perfect k-ary tree has $(k-1)^h$ leaves. Furthermore, perfect k-ary trees minimize $\sum_{i,j\in\Gamma(T)} \tau^T_{ij}$ [25].

We note that if T is perfect k-ary and $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ are two disjoint $\Gamma(T)$-subtrees with heights h_1 and h_2, respectively, then the set $R(T_1, T_2)$ of all lowest common ancestors of pairs of leaves (i, j) with $i \in \Gamma(T_1)$ and $j \in \Gamma(T_2)$ consists of a single vertex. We denote this vertex by $r(T_1, T_2)$. Furthermore, if $R(T_1, T_2) = \{r(T_1, T_2)\}$, we say that T_1 and T_2 are *adjacent* if there exists a perfect k-ary $\Gamma(T)$-subtree $\widehat{T}$, rooted at $r(T_1, T_2)$ and of height $\max\{h_1, h_2\} + 1$, such that $\Gamma(T_1) \cup \Gamma(T_2) \subseteq \Gamma(\widehat{T})$ (see Figure 2).

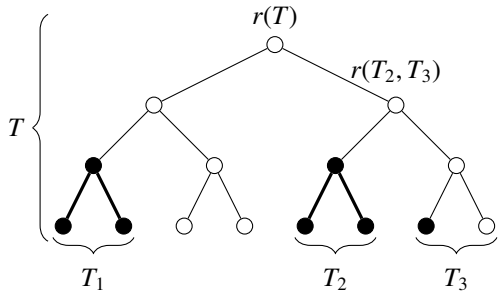

FIGURE 2 A perfect 3-ary tree T together with three disjoint perfect 3-ary $\Gamma(T)$-subtrees T_1, T_2, and T_3, highlighted by bold edges and black vertices. Since $r(T_1) = r(T_1, T_2)$ and T is the smallest perfect 3-ary tree of height 3 containing both T_1 and T_3, the subtrees T_1 and T_2 are not adjacent. For the same reason, T_1 and T_3 are not adjacent. In contrast, because $\tau^T_{r(T_2,T_3),i} = 2$, for every $i \in \Gamma(T_2) \cup \Gamma(T_3)$ and the heights of T_2 and T_3 are 1 and 0, respectively, the subtrees T_2 and T_3 are adjacent.

A tree T is a *caterpillar* if $T \setminus \Gamma(T)$ is a path. If a caterpillar belongs to Θ^k_n, for some appropriate n and k, there exist exactly two disjoint subsets of leaves $\Gamma_1, \Gamma_2 \subseteq \Gamma(T)$ of cardinality $k-1$ such that, for every $i,j \in \Gamma_1$ and $i',j' \in \Gamma_2$, (i,j) and (i',j') are cherries (see Figure 3). A folklore result shows that a tree in Θ^k_n is a caterpillar if and only if there exist two leaves whose path-length is $(n-2)/(k-2)+1$.

FIGURE 3 On the left a perfect 4-ary tree. On the right a caterpillar belonging to Θ^4_{10}.

Suppose that T is a caterpillar belonging to Θ^k_n. We say that T' is a *comb of* T if T' is a $\Gamma(T)$-subtree and every internal vertex of T' has exactly degree k but two that have exactly degree $k-1$ (see Figure 4). These last two vertices are the *extremals* of the comb. Note that a comb T' of T is such that for every $i \in \Gamma(T')$ there is $j \in \Gamma(T')$ for which $\tau^{T'}_{ij} = 3$ holds. A tree is a *comb* if it is isomorphic to a comb of some caterpillar.

Let T be a perfect k-ary tree and let $\Gamma' \subseteq \Gamma(T)$. A perfect k-ary $\Gamma(T)$-subtree T' is said to be *maximal in* Γ' if $\Gamma(T') \subseteq \Gamma'$ and there does not exist any perfect k-ary $\Gamma(T)$-subtree T'' such that $\Gamma(T') \subsetneq \Gamma(T'') \subseteq \Gamma'$. Similarly, let T be a caterpillar tree and let $\Gamma' \subseteq \Gamma(T)$. A comb T' of T is *maximal in* Γ' if $\Gamma(T') \subseteq \Gamma'$ and there does not exist any comb T'' of T such that $\Gamma(T') \subsetneq \Gamma(T'') \subseteq \Gamma'$.

A tree T is a *quasi-caterpillar* if there are two $\Gamma(T)$-subtrees T' and T'', hereafter referred to as *wings of* T, such that $T \setminus (T' \cup T'')$ is a comb (see Figure 4). When T' and T'' are both perfect k-ary trees with the same height, then we say that T is a *balanced quasi-caterpillar*. By definition, every caterpillar belonging to Θ^k_n is a balanced quasi-caterpillar.

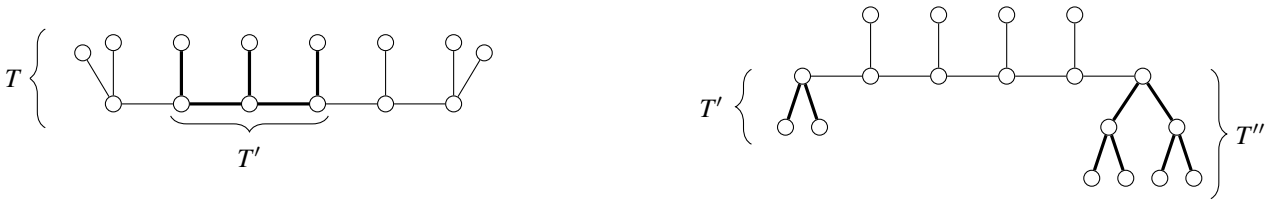

FIGURE 4 On the left a comb T' (highlighted with bold edges) of a cubic caterpillar. On the right a cubic quasi-caterpillar, with wings T' and T'' (highlighted with bold edges).

We define a *k-ary distribution* of a perfect $(k+1)$-ary tree T as a family of pairwise disjoint perfect $\Gamma(T)$-subtrees $T^1, \ldots, T^s$ of height $0 < h_1 < h_2 < \ldots < h_s$ such that T^i is adjacent to T^{i+1}, for all $i \in \{1, \ldots, s-1\}$. We say that $\Gamma' \subseteq \Gamma(T)$ *forms* a k-ary distribution of T if there exists a k-ary distribution $\mathfrak{T}$ such that $\Gamma' = \cup_{T^i \in \mathfrak{T}} \Gamma(T^i)$. Note that, if c is the k-ary expression of $|\Gamma'|$, with s non-zero digits, a k-ary distribution $T^1, \ldots, T^s$ of T associated with Γ' can be seen as associating to the leaves of T^i the i-th nonzero digit of c starting from the most significant digit to the least (see, e.g., Figure 5).

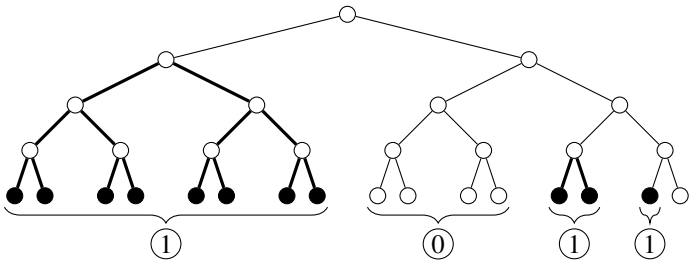

FIGURE 5 A 2-ary distribution associated with a set Γ' with 11 vertices over a perfect 3-ary tree with 16 leaves. The 2-ary (binary) expression of 11 is '1011' and it is represented by three pairwise adjacent perfect 3-ary $\Gamma(T)$-subtrees (highlighted with bold edges and black leaves) having 2^3, 2^1, 2^0 leaves.

3 | ON THE $\mathcal{NP}$-HARDNESS OF THE K-NDP

In the light of the above notation and definitions, we study the complexity of the k-NDP in this section by establishing the following result:

Theorem 1. *The k-NDP is strongly $\mathcal{NP}$-hard for every integer $k \geq 3$.*

We prove the statement by showing that the following decision version of the k-NDP is $\mathcal{NP}$-complete:

Problem 4 (Decision version of the k-NDP — (d-k-NDP))**.** Given a graph $G = (V, E)$ of $\mathcal{G}_n^k$, a weight function $w : E \to \mathbb{Q}_0^+$, and a constants $C \in \mathbb{Q}^+$, decide whether there exists a spanning tree T of G belonging to Θ_n^k such that $\sum_{u,v \in V} L_T^w(u, v) \leq C$.

We proceed in two steps: we first show that the classical $\mathcal{NP}$-complete *Exact 3-Cover Problem*[26] can be reduced in polynomial time to the following more general problem:

Problem 5 (The Exact m-Cover Problem (ECP)[26])**.** Given two positive integers $\ell \geq 1$ and $m \geq 3$, a finite set $M = \{\mu_1, \ldots, \mu_{\ell m}\}$, and a family of subsets $S \subseteq 2^M$, such that $|\sigma| = m$ for every $\sigma \in S$, decide whether there exists a subset $P \subseteq S$ such that P is a partition of M.

Subsequently, we show that the ECP can be reduced in polynomial time to the d-k-NDP, by completing so the proof. We start by showing the following preliminary result:

Proposition 1. *The ECP is $\mathcal{NP}$-complete for every integer $m \geq 3$.*

Proof. We prove that every instance of the exact m-cover can be reduced to an instance of the exact $(m+1)$-cover problem. Consequently, the result follows by the $\mathcal{NP}$-completeness of the exact 3-cover. Let $M' = \{\mu_1, \ldots, \mu_{m\ell}\}$ and define $M = \{\mu_1, \ldots, \mu_{m\ell}, \mu_{m\ell+1}, \ldots, \mu_{m\ell+\ell}\}$. Consider a collection $S' \subseteq 2^{M'}$ such that each $\sigma' \in S'$ satisfies $|\sigma'| = m$. We construct $S = \{\, \sigma' \cup \{\mu\} : \sigma' \in S',\ \mu \in \{\mu_{m\ell+1}, \ldots, \mu_{m\ell+\ell}\} \,\} \subseteq 2^M$, so that every $\sigma \in S$ has size $|\sigma| = m + 1$.

By construction, there is a partition $P \subseteq S$ of M if and only if for every $(\sigma', \mu) \in P$ the σ''s and μ's form a partition of M' and $\{\mu_{m\ell+1}, \ldots, \mu_{m\ell+\ell}\}$ respectively. Thus, solving the exact $(m+1)$-cover instance requires solving the associated exact m-cover instance. Therefore, the exact m-cover is at least as hard as the exact 3-cover for every $m > 4$, and thus $\mathcal{NP}$-complete. □

We now prove Theorem 1 by assuming that $m = (k-1)^2$. To this end, consider a positive integer r and define $s = |S|$. Consider a connected graph $G = (V, E)$ defined as follows. The vertex-set V is partitioned into the vertex-sets V_{rs}, V_R, V_S, and $M = \{\mu_1, \ldots, \mu_{\ell(k-1)^2}\}$, where the first three sets have cardinality $r+s$, $(k-2)r + (k-3)s + 2$, and ks, respectively. In particular, we consider s elements of V_S indexed by S. Then, G is such that:

- $G[V_R \cup V_{rs} \cup V_s \cup S]$ is a caterpillar belonging to $\Theta^k_{(r+s)(k-2)}$ with leaf-set $V_R \cup S$;
- $G[V_S]$ has s connected components each of which is a perfect k-ary tree of height 1;
- $G[(V_S \setminus S) \cup M]$ is a complete bipartite graph whose bipartition is given by $V_S \setminus S$ and M;
- No edges other than those described above are present in G.

Figure 6 represents G for some $r > 4$, $s = 3$, and $k = 3$. We note also that by construction G belongs to $\mathcal{G}^k_n$ for $n = |R| + |M| + (s(k-1)^2 - |M|)/(k-1)$. Finally, we complete the construction of the instance of Problem 4 by defining the edge-weight function w. For any edge e such that exactly one of its endpoints belongs to V_{rs}, we set $w(e) = 1$.

For edges of the form $e = u\mu_j$, the weight assignment is defined as follows. The edge $u\mu_j$ has assigned weight 1 if and only if all the following conditions hold:

(i) the vertex u is adjacent to some σ_i such that $\mu_j \in \sigma_i$;

(ii) at most $k-2$ edges of the form $u\mu_j$ and incident to u have weight 1;

(iii) there is no vertex $v \neq u$ adjacent to the same σ_i such that $w(v\mu_j) = 1$.

If any of the conditions (i)–(iii) is violated, we set $w(u\mu_j) = |V|^3$. Finally, all remaining edges are assigned weight 0.

In what follows, with a slight abuse of terminology, we refer to any spanning tree of G belonging to Θ^k_n as a *feasible solution*. Observe that the weight $|V|^3$ assigned to certain edges of the form $u\mu_j$ is chosen sufficiently large so that any feasible solution containing at least one such edge has strictly larger cost than any feasible solution (if any exists) that avoids all of them. This observation follows from the facts that $|V| > 4$ for every $k \geq 3$ and for any path P in G such that $w(e) \neq |V|^3$ for every $e \in E(P)$, then $\sum_{e \in E(P)} w(e) \leq 4$. Consequently, in the remainder of the proof we restrict our attention to feasible solutions of the latter type.

For a feasible solution T and any couple of vertex-sets $U, U' \subseteq V$, we define *cost of the pair* (U, U') as

$$L^w_T(U, U') = 2 \sum_{(u,v) \in U \times U'} L^w_T(u, v).$$

When $U = U'$, we simply write $L^w_T(U)$ and we refer to this quantity as the *cost of* U. In particular, when $U = V$, we refer to $L^w_T(V)$ as the *cost of* T.

By construction, the edge-set of every feasible solution contains $E(G[V_R \cup V_{rs} \cup S])$. Moreover, for every feasible solution, every $u' \in M$ and $r' \in V_R$ (respectively, $v' \in V_{rs}$) we have that $L^w_T(r', u') = 3$ (respectively, $L^w_T(r', u') = 2$). Therefore, the cost

$$C_0 = L^w_T(V_R) + L^w_T(V_R, V_{rs}) + L^w_T(V_R, S) + L^w_T(V_R, M) + L^w_T(V_{rs}) + L^w_T(V_{rs}, M) + L^w_T(V_{rs}, S) + L^w_T(S),$$

is a constant for every feasible solution, i.e., this cost does not depend on the choice of T. Finally, denote by $\mathcal{T}^k$ the set of feasible solutions whose edge-set contains $E(V_S)$. Then, in the light of the above notation, the following proposition holds:

Proposition 2. *For $r > 8k^3\ell^2s^2$, the cost of any feasible solution in $\mathcal{T}^k$ is strictly smaller than the cost of every feasible solution not in $\mathcal{T}^k$.*

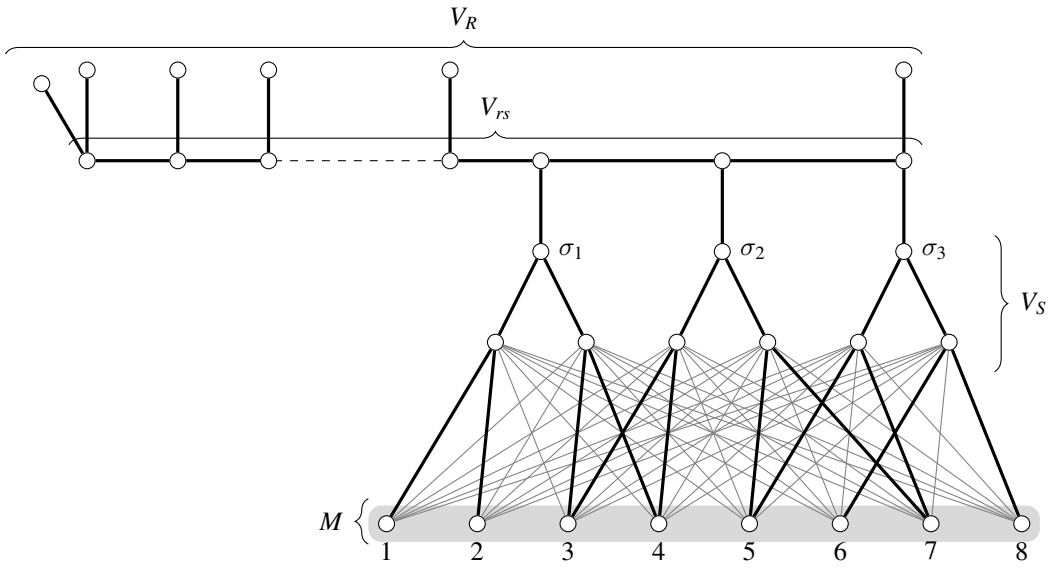

FIGURE 6 An example of an instance of Problem 4 when assuming $k = 3$, $r > 4$, $M = \{1, \ldots, 8\}$, and $S = \{\sigma_1 = (1, 2, 3, 4), \sigma_2 = (3, 4, 5, 7), \sigma_3 = (5, 6, 7, 8)\}$. The light gray edges (all of which are incident to M) are those for which the corresponding weight is $|V|^3$.

Proof. Let $T \in \mathcal{T}^k$ and T' be any feasible solution not in $\mathcal{T}^k$. First, observe that

$$L^w_T(V) = C_0 + L^w_T(V_R, V_S \setminus S) + L^w_T(V_{rs}, V_S \setminus S) + L^w_T(V_S \setminus S) + L^w_T(V_S, M) + L^w_T(M).$$

By definition of $\mathcal{T}^k$ we have that, $L^w_T(V_R, V_S) = 4(k-1)s|R|$, $L^w_T(V_{rs}, V_S \setminus S) = 2(k-1)s(r+s)$, $L^w_T(V_S \setminus S, V_S \setminus S) + L^w_T(V_S, M) \leq 4s^2(k-1) + 8\ell sk(k-1)^2$ as $L^w_T(u, v) \leq 4$ holds true for any couple of vertices (u, v) in $V_S \times M$, and $L^w_T(M) \leq 6\ell^2(k-1)^2$ as $L^w_T(u, v) \leq 3$ holds true for any couple of vertices u, v in M. Therefore, we have that

$$L^w_T(V) \leq C' = C_0 + 4(k-1)s|R| + 2(k-1)s(r+s) + 4s^2(k-1) + 8\ell sk(k-1)^2 + 6\ell^2(k-1)^2.$$

Also observe that, since T' belongs to Θ^k_n, there are $k-1$ edges of the form $u\sigma_i \in G[V_s] \setminus T'$, for some $\sigma_i \in S$. For these edges, either $L^w_{T'}(\{u\}, V_R)$ or $L^w_{T'}(\{\sigma_i\}, V_R)$ equals $8|R|$. Similarly, $L^w_{T'}(\{u\}, V_{rs})$ or $L^w_{T'}(\{\sigma_i\}, V_{rs})$ equals $6(r+s)$. Hence, the $L^w_{T'}(V) > C'' = C_0 + 4(k-1)(s+1)|R| + 2(k-1)(s+4)(r+s)$. Then, if $r > 8k^3\ell^2s^2$, $C' < C''$, i.e., $L^w_T(V) < L^w_{T'}(V)$. Thus, the statement follows. □

Now, assume that $r > 8k^3\ell^2s^2$. By Proposition 2, every feasible solution T whose cost $L^w_T(V)$ is smaller than some appropriate constant belongs to $\mathcal{T}^k$. Moreover, always by Proposition 2, the cost $L^w_T(V_R, V_S \setminus S)$ is constant for every $T \in \mathcal{T}^k$, and every vertex in M has degree 1. This fact implies that the cost $L^w_T(V_S, M)$ is constant for every $T \in \mathcal{T}^k$, since for every vertex μ_i in M there is precisely one vertex s_j in V_S such that the $L^w_T(\mu_i, s_j) = 1$, and $L^w_T(\mu_i, s_{j'}) = 3$ for all $s_{j'} \in V_S$, with $j' \neq j$. Therefore, the cost $C_1 = L^w_T(V_R, V_S \setminus S) + L^w_T(V_{rs}, V_S \setminus S) + L^w_T(V_S \setminus S) + L^w_T(V_S, M)$, is constant for any $T \in \mathcal{T}^k$.

We now observe that the following proposition holds:

Proposition 3. *There exists $T^* \in \mathcal{T}^k$ such that $L^w_{T^*}(V) \leq C_0 + C_1 + 2(k-1)^2(2(k-1)^2 - \ell - 1)$ if and only if there exists a partition $P \subseteq S$ of M.*

Proof. Let $T \in \mathcal{T}^k$. Denote by S_σ the set of vertices of V_S that are adjacent in T to $\sigma \in S$. Since $T \in \Theta^k_n$ we can define the number s_h of sets S_σ, for all $\sigma \in S$ such that the number of vertices of M adjacent in T to S_σ is exactly h, where $h \in H =$

$\{0, k-1, 2(k-1), \ldots, (k-1)^2\}$. Moreover, let $S(T)$ be the set of couples (u, v) of vertices of M such that $L_T^w(u, v) = 2$. Then,

$$L_T^w(M) = 4(k-1)^2((k-1)^2-1) - 4|S(T)| = 4(k-1)^2((k-1)^2-1) - 2\sum_{h\in H} h \cdot s_h.$$

Note that the cost $L_T^w(M)$ decreases as $2\sum_{h\in H} h \cdot s_h$ increases. In particular, since every h is a multiple of $k-1$, this sum is as small as possible when s_h is as big as possible. By definition, $s_{(k-1)^2} \leq \ell$, and $s_{(k-1)^2} = \ell$ if and only if all the corresponding S_σ's counted by $s_{(k-1)^2}$ are such that there is a partition P of M for which $\sigma \in P$. Moreover, if $s_{(k-1)^2} = \ell$, every other $s_h = 0$ for any $T \in \mathcal{T}^k$ by construction, and hence $L_T^w(M) = 2(k-1)^2(2(k-1)^2 - \ell - 1)$.

Therefore, assume that $s_{(k-1)^2} = \ell$ and set $C = C_0 + C_1 + 2(k-1)^2(2(k-1)^2 - \ell - 1)$. Then, $\sum_{u,v\in V} L_T^w(u, v) = C$ if and only if there is a partition $P \subseteq S$ of M. □

From the proof of Proposition 3 we can deduce that $L_{T^*}^w(V) \geq C$. Thus, if $r > 8k^3\ell^2 s$, answering "yes" or "no" to Problem 4 is equivalent to deciding whether S contains a partition of M or not in the ECP. As the ECP is $\mathcal{NP}$-complete, so is the d-k-NDP. □

4 | ON THE $\mathcal{NP}$-HARDNESS OF THE MIN-(K, τ)-NDP

In this section, we prove the $\mathcal{NP}$-hardness of the Min-(k, τ)-NDP. We will first show that this result holds for $k = 3$, i.e., that

Theorem 2. *The Min-$(3, \tau)$-NDP is strongly $\mathcal{NP}$-hard.*

Subsequently, we will strengthen this result by generalizing it to any arbitrary integer $k \geq 3$.

We start by considering the following problem:

Problem 6. Given a finite set Γ' and a positive integer n such that $2^n > |\Gamma'|$, find a perfect 3-ary tree T^* with 2^n leaves that minimizes $\sum_{i,j\in\Gamma'} \tau_{ij}^T$.

We first observe that the following proposition holds:

Proposition 4. *Let T^* be an optimal solution to Problem 6 with input Γ' and n. Then, no two maximal perfect 3-ary $\Gamma(T^*)$-subtrees whose leaves are indexed by Γ' have the same height.*

Proof. Suppose, by contradiction, that T^* is an optimal solution to Problem 6 containing two perfect 3-ary disjoint $\Gamma(T^*)$-subtrees T_1 and T_2 such that: (i) having the same height h, (ii) being maximal in Γ', and (iii) whose leaves are indexed by Γ'. Denote by $\hat{T}_1, \hat{T}_2$ the two connected components of $T \setminus \{r(T_1, T_2)\}$ such that that $T_1 \subset \hat{T}_1$ and $T_2 \subset \hat{T}_2$. By maximality of T_1 and T_2, there exist two families $\mathcal{T}_1$ and $\mathcal{T}_2$ of perfect 3-ary $\Gamma(T^*)$-subtrees such that every element has height h, is contained in $\hat{T}_1$ and $\hat{T}_2$, and that is disjoint from and adjacent to T_1 and T_2, respectively. Moreover, always by maximality of T_1 and T_2, both $\Gamma(\hat{T}_1) \setminus \Gamma(T_1)$ and $\Gamma(\hat{T}_2) \setminus \Gamma(T_2)$ are not subsets of Γ'.

Denote by T' the perfect 3-ary tree with 2^n leaves obtained by swapping the labels on the leaves of T_1 with those of any tree T_4 in $\mathcal{T}_2$, and T_1' and T_4' the two subtrees of T' corresponding to T_1 and T_4, respectively. Similarly, denote by T'' the perfect 3-ary tree obtained by swapping the labels on the leaves of T_2 with those of any tree T_3 in $\mathcal{T}_1$, and T_2'' and T_3'' the two subtrees of T'' corresponding to T_2 and T_3, respectively.

In what follows, we prove that the cost of the solution associated with T' or T'' is strictly smaller than the cost of T^*, contradicting the optimality of (T^*). First note that most contributions to the total costs of T^*, T', and T'' are the fixed. Specifically, for every i in $\Gamma(\hat{T}_1)$ and j in $\Gamma(\hat{T}_2)$, $\tau_{is}^{T^*} = \tau_{js}^{T^*}$ for all $s \in \Gamma(T^*) \setminus (\Gamma(\hat{T}_1) \cup \Gamma(\hat{T}_2))$. Thus, by swapping i with j, any possible change in the cost the objective function of the problem depends only on the pairwise distances of vertices in $\Gamma(\hat{T}_1) \cap \Gamma'$, those in $\Gamma(\hat{T}_2) \cap \Gamma'$, and the distances between vertices in $\Gamma(\hat{T}_1) \cap \Gamma'$ and $\Gamma(\hat{T}_2) \cap \Gamma'$. Moreover, the cost relative to the swap of the labels on $\Gamma(T_1)$ with those indexed by T_4 is fixed, since $\tau_{ij}^{T^*} = \tau_{i'j'}^{T'}$ for every $i \in \Gamma(T_1)$, $j \in \Gamma(T_4)$, $i' \in \Gamma(T_1')$, and $j' \in \Gamma(T_4')$.

Note also that all these remarks hold when swapping the labels of T_2 with those of T_3. Therefore, denote by C_0 these invariant costs. In conclusion, any variation in the cost after one of these swaps depends on the following contributions:

$$l_i^* = \sum_{p \in \Gamma(\hat{T}_1) \setminus \Gamma(T_1)} \tau_{ip}^{T^*} \quad \text{and} \quad r_i^* = \sum_{q \in \Gamma(\hat{T}_2) \setminus \Gamma(T_4)} \tau_{iq}^{T^*} \quad \text{for} \quad i \in \Gamma(T_1),$$
$$l_j^* = \sum_{p \in \Gamma(\hat{T}_1) \setminus \Gamma(T_1)} \tau_{jp}^{T^*} \quad \text{and} \quad r_j^* = \sum_{q \in \Gamma(\hat{T}_2) \setminus \Gamma(T_4)} \tau_{jq}^{T^*} \quad \text{for} \quad j \in \Gamma(T_4).$$

We also denote $l_i', l_j', l_i'', l_j'', r_i', r_j', r_i''$, and r_j'', the similar contributions relative to T' and T''. Since T_1, T_2, T_3, and T_4 are perfect, all these quantities do not depend on the choice of i and j. Thus,

$$\sum_{i \in \Gamma(T_1)} (l_i^* + r_i^*) = |\Gamma(T_1)|(l_i^* + r_i^*), \qquad \sum_{j \in \Gamma(\hat{T}_1) \setminus \Gamma(T_1)} (l_j^* + r_j^*) = |\Gamma(T_4) \cap \Gamma'|(l_j^* + r_j^*).$$

Similar equalities hold for the sums of $l_i', l_j', l_i'', l_j'', r_i', r_j', r_i''$, and r_j'', and, moreover, $|\Gamma(T_1)| = |\Gamma(T_4')|$, $|\Gamma(T_4) \cap \Gamma'| = |\Gamma(T_1') \cap \Gamma'|$, $|\Gamma(T_2)| = |\Gamma(T_3'')|$, and $|\Gamma(T_3) \cap \Gamma'| = |\Gamma(T_2'') \cap \Gamma'|$.

Now, suppose that $\sum_{i \in \Gamma(T_1)}(l_i^* + r_i^*) > \sum_{j \in \Gamma(T_4)}(l_j^* + r_j^*)$. Then, in the light of the qbove equalities, the cost of T^* corresponding to $\Gamma(T_1)$ and $\Gamma(T_4) \cap \Gamma'$ is $c^* = |\Gamma(T_1)|(l_i^* + r_i^*) + |\Gamma(T_4) \cap \Gamma'|(l_j^* + r_j^*)$, while the corresponding one of T' is $c' = |\Gamma(T_4) \cap \Gamma'|(l_i' + r_i') + (l_j' + r_j')|\Gamma(T_1)|$. Due to the maximality of T_1, we have that $|\Gamma(T_4) \cap \Gamma'| < |\Gamma(T_1)|$, which implies that c' is strictly smaller than c. With similar arguments, it is easy to see that if $\sum_{i \in \Gamma(T_1)}(l_i^* + r_i^*) < \sum_{j \in \Gamma(T_4)}(l_j^* + r_j^*)$, then the cost of T'' is smaller than the one of T^*. Finally, note that the condition $\sum_{i \in \Gamma(T_1)}(l_i^* + r_i^*) = \sum_{j \in \Gamma(T_4)}(l_j^* + r_j^*)$ cannot hold true as, by maximality of T_1 and T_2, both $|\Gamma(T_4) \cap \Gamma'|$ and $|\Gamma(T_3) \cap \Gamma'|$ are strictly smaller than $|\Gamma(T_1)|$ and $|\Gamma(T_2)|$. Thus, the cost of either T' or T'' is strictly smaller than the corresponding one of T^*, contradicting the hypothesis of optimality of T^*. □

Now, let T^* denote an optimal solution to Problem 6 with input Γ' and n. Then, the following proposition holds:

Proposition 5. *The subset of leaves Γ' forms a 2-ary distribution on T^* if and only if T^* is an optimal solution to Problem 6.*

Proof. We prove the statement by showing that every feasible solution to Problem 6 that does not form a 2-ary distribution is not optimal. On the contrary, it is easy to see that every 2-ary distribution has the same cost.

By Proposition 4, no two perfect 3-ary $\Gamma(T^*)$-subtrees that are maximal on Γ' have the same height. Thus, there is a sequence $T^1, \ldots, T^s$ of perfect k-ary $\Gamma(T^*)$-subtrees having height $h_1 < h_2 < \ldots < h_s$ that are maximal on Γ'. To conclude that Γ' forms a 2-ary distribution over T^*, it remains to prove that T^i and T^{i+1} are adjacent, for all $i = 1, \ldots, s-1$.

Let T^i be the first tree in the collection i such that T^i and T^{i+1} are not adjacent. Since T^i is maximal, there exists a (unique) perfect 3-ary $\Gamma(T^*)$-subtree $\tilde{T}$ adjacent to T^i and having height $h_i + 1$ and such that $\Gamma'(\tilde{T})$ is strictly contained in $\Gamma(\tilde{T})$. Similarly to the proof of Proposition 4, to reduce the total cost it suffices to swap the leaves of T^{i+1} with those of $\tilde{T}'$, since $|\Gamma'(\tilde{T})| < |\Gamma'(T^{i+1})|$ by assumption. □

Remark 1. As for Problem 6, 2-ary distributions corresponds to optimal solutions of the similar optimization problem of finding the minimum among all balanced quasi-caterpillars whose wing have height h and under the additional constraint of Γ' indexing only leaves of the two wings with $|\Gamma'| < 2^h$.

Let k, n and h be three positive integers such that $n > 2(k-1)^h$. Then, given a set of items Γ, with $|\Gamma| = (k-1)^h$, we denote by $\mathcal{T}_n^k(\Gamma)$the subset of quasi-caterpillars in Θ_n^k whose wings' leaf-sets are indexed by Γ. Consistently with the propositions above, we restrict ourselves to the case in which $k = 3$ in the following:

Problem 7. Given two positive integers n and h such that $n > 2^{h+1}$, and three mutually disjoint sets Γ_0, Γ_1, and Γ_2 such that $|\Gamma_0|$ is even, $|\Gamma_1| = |\Gamma_2|$ and $|\Gamma_0| + |\Gamma_1| + |\Gamma_2| = 2^{h+1}$, find a tree T belonging to $\mathcal{T}_n^3(\Gamma_0 \cup \Gamma_1 \cup \Gamma_2)$ that minimizes

$$\sum_{i,j\in\Gamma_0} \varepsilon_{ij}\tau_{ij}^T + M\left(\sum_{i,j\in\Gamma_1} \tau_{ij}^T + \sum_{i,j\in\Gamma_2} \tau_{ij}^T\right) + 2\gamma \sum_{i\in\{u,v\},j\in\Gamma'} \tau_{ij}^T,$$

where u and v are two leaves such that $\tau_{uv}^T = n-1$ and $\varepsilon_{ij} = \varepsilon_{ji} \in \{0,1\}$, $M > |\Gamma_0|$ and $\gamma > Mn|\Gamma_1|^2$.

Proposition 6. *The optimal solution of Problem 7 is a balanced quasi-caterpillar.*

Proof. We prove the statement by contradiction. Assume that the optimal solution T^* is not a balanced quasi-caterpillar, and denote by T' and T'' its wings. Without loss of generality, assume that $|\Gamma(T')| \geq |\Gamma(T'')|$ and that $\tau_{u,r(T')}^{T'} = 2$. Consequently, by assumption, $\tau_{v,r(T'')}^{T''} = 2$. Let $i \in \text{argmax}\{\tau_{ju}^{T^*} : j \in \Gamma(T')\}$. Since T^* is not a balanced quasi-caterpillar, the choice of $i \in \Gamma(T')$ satisfies $\tau_{iu}^{T^*} > \log_2|\Gamma_0 \cup \Gamma_1 \cup \Gamma_2|$, where $\tau_{iu}^T = \log_2|\Gamma_0 \cup \Gamma_1 \cup \Gamma_2|$ for every $T \in \mathcal{T}_n^3(\Gamma_0 \cup \Gamma_1 \cup \Gamma_2)$. Moreover, since T' is a cubic tree, there exists $i' \in \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$ such that (i, i') is a cherry.

Similarly, let $i'' \in \text{argmin}\{\tau_{iv}^{T''}\}$. Consider now the quasi-caterpillar $\hat{T}$ obtained from T^* by removing i and i', relabeling the vertex adjacent to i in T^* with i', relabeling i'' with a new label $\hat{i} \notin \Gamma(T^*)$, and finally adding i and i'' so that both become adjacent to $\hat{i}$. Since the longest path in a cubic tree with q leaves has length $q-1$, and since $M > |\Gamma_0| > |\Gamma_0|-1$, the contributions of i, i', and i'' to the cost difference between T^* and $\hat{T}$ are at least $\gamma - Mn|\Gamma_1|^2$, 2γ, and $-M|\Gamma_1|^2$, respectively. A direct computation then shows that the difference between the cost of T^* and the one of $\hat{T}$ is at least $\gamma - Mn|\Gamma_1|^2$, and this quantity is positive by hypothesis. Thus, the cost associated with $\hat{T}$ is strictly smaller than the one associated with T^*, contradicting its optimality. □

In the light of the above propositions, we now prove Theorem 7 by reducing the following classical $\mathcal{NP}$-hard problem to the Min-$(3,\tau)$-NDP:

Problem 8 (The Graph Bisection Problem (GBP) [27])**.** Given a simple connected graph $G = (V, E)$ having $2p$ vertices, for some $p \in \mathbb{Z}^+$, find a partition of V into two equal-sized subsets U^* and W^* that minimizes $|\delta(U^*, W^*)|$.

We will show that, given an optimal solution to an instance of the Min-$(3,\tau)$-NDP, appropriately built from an instance of the GBP, it is possible to compute in polynomial time a sub-partition of the leaf-set Γ that induces an optimal solution to the given instance of the GBP. The $\mathcal{NP}$-hardness of the Min-$(3,\tau)$-NDP, then, will immediately derive from the $\mathcal{NP}$-hardness of the GBP. We start by considering an input graph $G = (V, E)$ of the GBP, for some fixed $p \in \mathbb{Z}^+$. Denoted m and s two non-negative integers such that $p+m$ is the smallest possible power of two and s is the smallest integer strictly greater than $\log_2(p+m)(p^2-1)-1$. Then, we define three sets of items: V_1 and V_2 both having cardinality m and such that V, V_1, and V_2 are mutually disjoint; and $R = \{r_{-s}, \ldots, r_{-1}, r_0, \ldots, r_s\}$, such that $R \cap (V \cup V_1 \cup V_2) = \emptyset$. We observe that, by definition, $2m + 2p + 2s + 1$ is polynomially bounded in p. Finally, we denote by M, γ_1, γ_2 and γ_3 four positive rationals respecting the following inequalities:

- $M > 2\left(\log_2(p+m) + s + 1\right)p^2$,
- $\gamma_1 < (2s-2)\gamma_2 + (p+m)(5-4s)\gamma_3$,
- $\gamma_1 > 6\gamma_2 + (p+m)(4\log_2(p+m) + 4s + 9)\gamma_3$
- $\gamma_2 > \dfrac{2(p+m)(2s - \log_2(p+m) + 1)}{s-4}\gamma_3$,
- $\gamma_3 > 2M\left(p + m + s + \dfrac{1}{2}\right)m^2$,

(note that the fourth condition ensures the compatibility of the second and third conditions). Then, we build an instance of the Min-$(3,\tau)$-NDP by setting

- $\Gamma = V \cup V_1 \cup V_2 \cup R$;
- $d_{ij} = 1$ for all $ij \in E$;

- $d_{ij} = M$ for all $i,j \in V_1$ or $i,j \in V_2$;
- $d_{r_i r_{i+1}} = \gamma_1$ for all $i \in \{-s+1, \ldots, s-2\}$;
- $d_{r_{-s}, r_{-s+1}} = d_{r_{s-1}, r_s} = \gamma_2$;
- $d_{r_{-s} i} = d_{r_s i} = \gamma_3$ for all $i \in V \cup V_1 \cup V_2$;
- and $d_{ij} = 0$ otherwise.

Now, let T^* be the optimal solution to the Min-$(3,\tau)$-NDP and let $\mathcal{T}^{\min}_{(3,\tau)}$ denote the subset of feasible solutions to the Min-$(3,\tau)$-NDP that are balanced quasi-caterpillars and such that (i) the leaf-sets of the wings are a partition of $V \cup V_1 \cup V_2$ and (ii) the remaining leaves are indexed from the left to the right by $r_{-s}, \ldots, r_s$. Then, the following proposition holds:

Proposition 7. *The optimal solution T^* to the Min-$(3,\tau)$-NDP belongs to $\mathcal{T}^{min}_{(3,\tau)}$.*

Proof. We first observe that any cubic tree in $\mathcal{T}^{\min}_{(3,\tau)}$ satisfies $\tau_{r_i r_{i+1}} = 3$ for every $i \in \{-s, \ldots, s-1\}$. Hence, by definition of $\mathcal{T}^{\min}_{(3,\tau)}$ and by hypothesis on γ_3, the cost of any solution in $\mathcal{T}^{\min}_{(3,\tau)}$ is at most

$$2\left(3(2s-2)\gamma_1 + 6\gamma_2 + (p+m)(4\log_2(p+m) + 4s + 9)\gamma_3\right). \tag{1}$$

Assume by contradiction that T^* does not belong to $\mathcal{T}^{\min}_{(3,\tau)}$. Since T^* is a cubic tree, if $\tau^{T^*}_{r_j r_{j+1}} = 2$ for some $j \in \{-s+1 \ldots, s-2\}$, then one of the following holds:

(i) $\tau^{T^*}_{r_{j-1} r_j}, \tau^{T^*}_{r_{j+1} r_{j+2}} \geq 4$ if $-s+2 < j < s-3$;

(ii) one among $\tau^{T^*}_{r_{-s+1} r_{-s+2}}$ or $\tau_{r_{s-2} r_{s-1}}$ equals 2.

Suppose that (i) holds true. Then, the cost associated with T^* is at least $2(\gamma_1 - 6\gamma_2 - (p+m)(4\log_2(p+m) + 4s + 9)\gamma_3)$ and the difference between this quantity and (1) is $2(3(2s-2)\gamma_1 + \gamma_1)$. By the hypothesis on γ_1, this quantity is positive, hence if (i) holds T^* cannot lead to an optimal solution.

Now suppose that (ii) holds true and that $\tau^{T^*}_{r_{s-2} r_{s-1}} = 2$. Then, as for condition (i), assuming some $\tau^{T^*}_{r_{j-1} r_j}, \tau^{T^*}_{r_{j+1} r_{j+2}} \geq 4$ never leads to the optimum, and hence, we assume that for all $\tau^{T^*}_{r_i r_{i+1}} = 3$ with $i \in \{-s, .., s-2\}$, which leads $\tau^{T^*}_{r_s r_{s-1}} = 2s+1$. By Remark 1, the least expensive subtree with leaf-set $V \cup V_1 \cup V_2$ (with respect to both r_{-s} and r_s) is a perfect 3-ary tree. A direct calculation yields that its cost is at least $2((3(2s-3)+2)\gamma_1 + (2s+4)\gamma_2 + (p+m)(4\log_2(p+m) + 14)\gamma_3)$, and the difference between this last cost and (1) is $2(-\gamma_1 + (2s-2)\gamma_2 + (p+m)(5-4s)\gamma_3)$, which is positive by the hypothesis on γ_1, γ_2 and γ_3.

Therefore, the optimal solution must be a quasi-caterpillar whose wings are indexed by $V \cup V_1 \cup V_2$. Thus, by Proposition 6 applied with $\gamma = \gamma_3$ (noting that the constant M used here is significantly larger than the one used in the proof of Proposition 6), the optimal solution is a balanced quasi-caterpillar. □

Given the optimal solution T^* to the Min-$(3,\tau)$-NDP, the following proposition holds:

Proposition 8. *The vertex-sets V_1 and V_2 form a 2-ary distribution over $\hat{T}'$ and $\hat{T}''$, respectively.*

Proof. First, observe that since $M > 4p^2(p+m)^2$ and $|E| < p^2$, we have

$$\sum_{ij \in E} \tau^{T^*}_{ij} < 2\left(\log_2(p+m) + s + 1\right)|E| < 2\left(\log_2(p+m) + s + 1\right)p^2 < M.$$

This implies that increasing by one the path-length between two leaves both indexed by V_1, or between two leaves both indexed by V_2, improves the objective value of the Min-$(3,\tau)$-NDP more than any possible reassignment of the labels in V among the leaves of $\hat{T}'$ and $\hat{T}''$. Consequently, any optimal solution must minimize the partial costs $\sum_{i,j \in V_1} \tau^{T^*}_{ij}$ and $\sum_{i,j \in V_2} \tau^{T^*}_{ij}$ independently of the placement of the elements of V. By applying Remark 1 first with respect to r_{-s} and then with respect to r_s, it follows that every optimal solution induces two 2-ary distributions on $\hat{T}'$ and $\hat{T}''$, corresponding to the sets V_1 and V_2, respectively. □

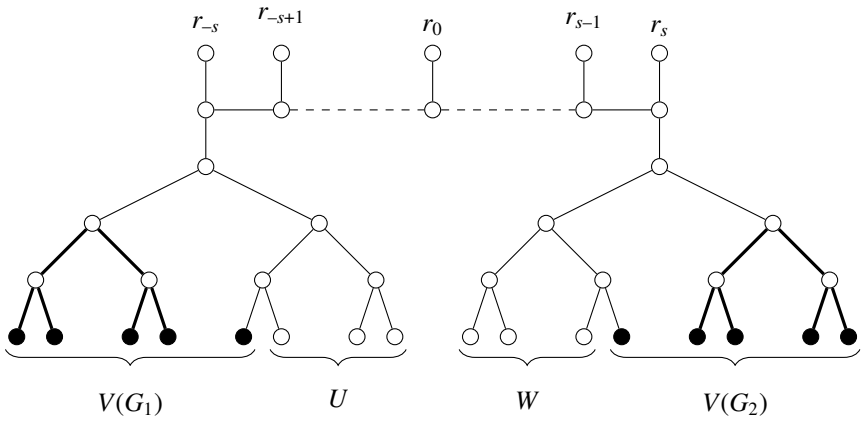

FIGURE 7 An example of the cubic tree in Proposition 8 when assuming $p = 3$ and $m = 5$. The bipartition of the vertex-set V into two equal-sized vertex-subsets U, W is a consequence of the two 2-ary distributions (highlighted with bold edges and black leaves) formed by V_1 and V_2, respectively.

By Propositions 7 and 8, $|\Gamma(\hat{T}')| = |\Gamma(\hat{T}'')| = p + m$, and V_1 and V_2 form 2-ary distributions on $\hat{T}'$ and $\hat{T}''$, respectively. Hence, there are exactly p leaves indexed by G in $\hat{T}'$ and $\hat{T}''$, respectively; therefore, there is a partition (U, W) of V such that $\Gamma(\hat{T}') = V_1 \cup U$ and $\Gamma(\hat{T}'') = V_2 \cup W$ with $|U| = |V| = p$ (see Figure 7).

Now, we show that the partition of G associated with T^* constitutes an optimal solution to the given instance of the GBP. To this end, let z^* denote the cost of any feasible solution to the Min-$(3, \tau)$-NDP respecting Propositions 7 and 8 such that the cut associated with the partition of G is optimal for the corresponding GBP instance. Let δ^* denote the cardinality of this solution. Similarly, let z' denote the cost of any feasible solution to the Min-$(3, \tau)$-NDP respecting Propositions 7 and 8 for which the cut associated with the partition of G is not optimal for the GBP instance. Let δ' denote the cardinality of this solution. Then, observe that the cost

$$C_0 = 2\left(3(2s-2)\gamma_1 + 6\gamma_2 + 2(p+m)(\log_2(p+m) + 2s + 4)\gamma_3 + \frac{1}{2}\left(\sum_{i,j\in V_1}\tau_{ij} + \sum_{i,j\in V_2}\tau_{ij}\right)M\right)$$

associated with all weights that differ from 1 is valid for both feasible solutions by construction. Then,

$$\begin{aligned} z^* &< C_0 + 4\delta^*(\log_2(p+m) + s + 1) + 4\left(\log_2(p+m)\right)(|E| - \delta^*) \\ &< \phi_1 = C_0 + 4\delta^*(\log_2(p+m) + s + 1) + 4\left(\log_2(p+m)\right)p^2, \end{aligned}$$

while $z' > \phi_2 = C_0 + 4\delta'(\log_2(p+m) + s + 1)$. By hypothesis on s, we have that $\phi_1 < \phi_2$. Thus, the cost of any optimal solution to the Min-$(3, \tau)$-NDP is at most z^* and the equal-sized partition of G provided by $(V \cap \Gamma(\hat{T}'), V \cap \Gamma(\hat{T}''))$ constitutes an optimal solution to the considered instance of the GBP. □

Generalization.
The concatenation of the propositions presented in this section allowed us to prove Theorem 2. We briefly observe here, however, that such a theorem can be easily generalized to any integer $k > 3$. Specifically, to account for generic $k > 3$ in the construction of an instance of the Min-(k, τ)-NDP, we consider the following instance:

- $\Gamma = V \cup V_1 \cup V_2 \cup R = \{r^1_{-s}, \ldots, r^{k-2}_{-s}, \ldots, r^1_{-1}, \ldots, r^{k-2}_{-1}, r^1_0, \ldots, r^{k-2}_0, \ldots, r^1_s, \ldots, r^{k-2}_s\}$, with $r^j_i \notin \Gamma \setminus R$ for all $i \in \{-s, \ldots, s\}$ and all $j \in \{1, \ldots, k-2\}$;
- $d_{ij} = 1$ for all $ij \in E$;
- $d_{ij} = M$ for all $i, j \in V_1$ and $i, j \in V_2$;
- $d_{r^j_i r^j_{i+1}} = \gamma_1$ for all $i \in \{-s+1, \ldots, s-2\}$ and $j \in \{1, \ldots, k-2\}$;

- $d_{r^j_{-s},r^1_{-s+1}} = d_{r^j_{s-1},r^1_s} = \gamma_2$ for all $j \in \{1,\dots,k-2\}$;
- $d_{r^1_{-s}i} = d_{r^1_s i} = \gamma_3$ for all $i \in V \cup V_1 \cup V_2$;
- and $d_{ij} = 0$ otherwise.

Then, it is sufficient (i) to assume that m is such that $m+p$ is the smallest possible power of $k-1$ (instead of 2); (ii) to note that, for γ_1, γ_2 and γ_3 large enough, T^* is a balanced quasi-caterpillar whose wings are two perfect k-ary subtrees of height $m+p$; (iii) to note that, for M large enough, both V_1 and V_2 form two $k-1$-ary distributions on the wings of T^*; and (iv) for s large enough, the partition of size p of G, indexing a part of the leaves of the wings of T^*, is an optimal solution to the GBP for the instance G. Thus, the following more general result holds:

Theorem 3. *The Min-(k,τ)-NDP is strongly $\mathcal{NP}$-hard for every integer $k \geq 3$.*

5 | ON THE $\mathcal{NP}$-HARDNESS OF MAX-(K,τ)-NDP

In this section, we prove the $\mathcal{NP}$-hardness of the Max-(k,τ)-NDP. As for the Min-(k,τ)-NDP, we will first show that this result holds for $k = 3$, i.e., that

Theorem 4. *The Max-$(3,\tau)$-NDP is strongly $\mathcal{NP}$-hard.*

We will then discuss how to generalize this result to arbitrary integers $k > 3$. We start by observing that the following result holds:

Proposition 9. *Let $n, m \geq 1$ be two integers, and let $\alpha \geq 1$ be a rational constant. Let $\{a_i\}_{i=1}^m$ and $\{b_i\}_{i=1}^m$ denote two sets of strictly positive integers such that $2 < a_1 < \ldots < a_n$ and $2 < b_1 < \ldots < b_m$. Then, precisely one of the following two inequalities holds:*

$$\sum_{i=1}^{n} a_i^\alpha + \sum_{i=1}^{m} b_i^\alpha < \sum_{i=1}^{n} (a_i - a')^\alpha + \sum_{i=1}^{m} (b_i + a')^\alpha, \text{ for every } a' \in \{1,\dots,a_1-3\}; \tag{2}$$

$$\sum_{i=1}^{n} a_i^\alpha + \sum_{i=1}^{m} b_i^\alpha \leq \sum_{i=1}^{n} (a_i + b')^\alpha + \sum_{i=1}^{m} (b_i - b')^\alpha, \text{ for every } b' \in \{1,\dots,b_1-3\}. \tag{3}$$

Proof. For a fixed $\varepsilon > 0$, consider the following polynomial real functions defined on the interval $[0, a_1-3+\varepsilon)$ and $[0, b_1-3+\varepsilon)$ respectively:

$$f_\alpha(x) = \sum_{i=1}^{n} (-x + a_i)^\alpha + \sum_{i=1}^{m} (x + b_i)^\alpha - A_\alpha - B_\alpha \quad \text{and} \quad g_\alpha(x) = \sum_{i=1}^{n} (x + a_i)^\alpha + \sum_{i=1}^{m} (-x + b_i)^\alpha - A_\alpha - B_\alpha,$$

where $A_\alpha = \sum_{i=1}^n a_i^\alpha$ and $B_\alpha = \sum_{i=1}^m b_i^\alpha$. We shall analyze now the behavior of these functions as α, n, and m vary, to determine when the inequalities under consideration hold. We begin by observing that when $\alpha = 1$, $f_1(x) = -nx + mx = (m-n)x$ and $g_1(x) = nx - mx = (n-m)x$. Hence, (2) holds when $m < n$, while (3) holds when $m \geq n$.

Now, assume that $\alpha > 1$ and consider the derivatives of both functions with respect to x, namely

$$f'_\alpha(x) = \alpha \left(-\sum_{i=1}^{n} (-x + a_i)^{\alpha-1} + \sum_{i=1}^{m} (x + b_i)^{\alpha-1} \right) \quad \text{and} \quad g'_\alpha(x) = \alpha \left(\sum_{i=1}^{n} (x + a_i)^{\alpha-1} - \sum_{i=1}^{m} (-x + b_i)^{\alpha-1} \right).$$

If $f'_\alpha(x_0) > 0$, for some $x_0 \in (0, a_1-3+\varepsilon) \cap (0, b_1-3+\varepsilon)$, then $\sum_{i=1}^m (x_0 + b_i)^{\alpha-1} > \sum_{i=1}^n (-x_0 + a_i)^{\alpha-1}$. Alternatively, if $f'_\alpha(x_0) \leq 0$, then we have that

$$\sum_{i=1}^{n} (a_i + x_0)^{\alpha-1} > \sum_{i=1}^{n} (a_i - x_0)^{\alpha-1} \geq \sum_{i=1}^{m} (b_i + x_0)^{\alpha-1} > \sum_{i=1}^{m} (b_i - x_0)^{\alpha-1},$$

i.e., $g'_\alpha(x_0) > 0$, for some $x_0 \in (0, b_1 - 3 + \varepsilon)$.

Moreover, as $a_i - a_1 + 3 > 0$ and $b_i - b_1 - 3 > 0$, the second derivatives of both functions, i.e.,

$$f''_\alpha(x) = \alpha(\alpha-1)(f_{\alpha-2}(x) + A_{\alpha-2} + B_{\alpha-2}) \quad \textit{and} \quad g''_\alpha(x) = \alpha(\alpha-1)(g_{\alpha-2}(x) + A_{\alpha-2} + B_{\alpha-2})$$

are always strictly positive in $(0, a_1 - 3 + \varepsilon)$ and $(0, b_1 - 3 + \varepsilon)$, respectively. Therefore, the convexity of $f_\alpha(x)$ implies that, if $f'_\alpha(x_0) > 0$ for some $x_0 \in (0, a_1 - 3 + \varepsilon)$, then $f_\alpha(x)$ is strictly increasing in the whole interval, while $g_\alpha(x)$ is weakly decreasing on $(0, b_1 - 3 + \varepsilon)$. Conversely, the convexity of $g_\alpha(x)$ implies that, if $g'_\alpha(x_0) > 0$ for some $x_0 \in (0, b_1 - 3 + \varepsilon)$, then $g_\alpha(x)$ is strictly increasing in the whole interval, while $f_\alpha(x)$ is weakly decreasing on $(0, a_1 - 3 + \varepsilon)$.

To conclude, since $f_\alpha(0) = g_\alpha(0) = 0$, and both functions are strictly convex with opposite slope sign on their respective intervals, i.e., either $f_\alpha(a') > 0$ or $g_\alpha(b') \geq 0$ when restricting them to integers. □

We prove Theorem 4 by reducing the following classical $\mathcal{NP}$-hard problem to the Max-$(3, \tau)$-NDP:

Problem 9 (The Max-Cut Problem (MCP)[27])**.** Given a graph $G = (V, E)$, with p vertices, for some $p \in \mathbb{Z}^+$, find a partition of V into two subsets U^* and W^* that maximizes $|\delta(U^*, W^*)|$.

We will show that, given an optimal solution to an instance of the Max-$(3, \tau)$-NDP, appropriately built from an instance of the MCP, it is possible to compute in polynomial time a sub-partition of the leaf-set Γ that induces an optimal solution to the given instance of the MCP. The $\mathcal{NP}$-hardness of the Max-$(3, \tau)$-NDP, then, will immediately follows from the $\mathcal{NP}$-hardness of the MCP.

We start by considering an input graph $G = (V, E)$ of the MCP, for some fixed $p \in \mathbb{Z}^+$. Denoted by ℓ any positive integer such that $\ell > 3p^3 + p^2$, consider a set of vertices L disjoint from V and having cardinality ℓ. We observe that, by definition, $p + \ell$ is polynomially bounded in p. Finally, denote by ω a positive integer such that $\omega > p^2(p + \ell + 1)$. Then, we build an instance of the Max-$(3, \tau)$-NDP as follows. Set:

- $\Gamma = V \cup L \cup \{r_1, r_2\}$, where r_1 and r_2 are two distinct vertices such that $r_1, r_2 \notin V \cup L$;
- $d_{ij} = 1$ for all $ij \in E$;
- $d_{r_1 r_2} = \omega$;
- and $d_{ij} = 0$ otherwise.

Let $\mathcal{T}^{\max}_{(3,\tau)}$ denote the set of caterpillars that are feasible solutions to the considered instance of the Max-$(3, \tau)$-NDP and such that r_1 indexes a leaf of one cherry and r_2 indexes a leaf of the (unique) other cherry. Then, the following claim holds:

Proposition 10. *The optimal solution T^* belongs to $\mathcal{T}^{max}_{(3,\tau)}$.*

Proof. Let T denote any feasible solution in $\mathcal{T}^{\max}_{(3,\tau)}$ and let T' denote any feasible solution not in $\mathcal{T}^{\max}_{(3,\tau)}$, i.e., T' is not a caterpillar, or it is a caterpillar but at least one between r_1 and r_2 does not index leaves of different cherries. Note that for a cubic tree on $|\Gamma|$ leaves, the maximal path-length between a pair of leaves is at most $|\Gamma| - 1$. Moreover, the path-length $|\Gamma| - 1$ is achieved by some pair of leaves if and only if the cubic tree is a caterpillar[25,17]. Then, $\tau^T_{r_1,r_2} = p + \ell + 1$ holds for T, while $\tau^{T'}_{r_1 r_2} < p + \ell + 1$ holds for T', but possibly $\tau^{T'}_{u,v} = p + \ell + 1$, for some $u, v \in V$. Then, if z and z' are the costs of T and T' with respect to the objective function of the Max-$(3, \tau)$-NDP, then we have that $z > 2\omega(p + \ell + 1) > \phi_1 = 2w[p + \ell + 1]$ and $z' < 2[\omega(p + \ell) + |E|(p + \ell + 1)] < \phi_2 = 2[\omega(p + \ell) + p^2(p + \ell + 1)]$. As $\omega > p^2(p + \ell + 1)$, $\phi_1 > \phi_2$, hence, $z > z'$. Thus, the cost of every feasible solution not belonging to $\mathcal{T}^{\max}_{(3,\tau)}$ is smaller than z and the statement follows. □

Now, we show that the optimal solution T^* can be partitioned with respect to G. More precisely, the following proposition holds:

Proposition 11. *There exists an optimal solution T^* to the Max-$(3, \tau)$-NDP that can be partitioned into three connected components C_L, C_U, and C_W, such that $\Gamma(C_L) = L$, $\Gamma(C_U) = U \cup \{r_1\}$, and $\Gamma(C_W) = W \cup \{r_2\}$, for some partition (U, W) of V.*

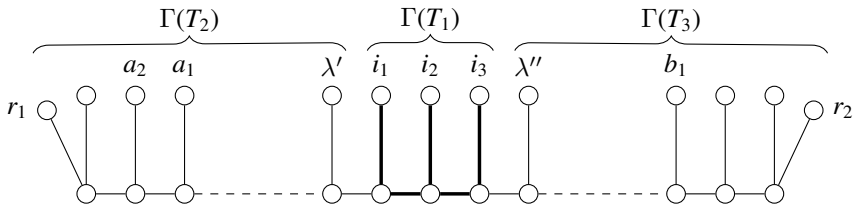

FIGURE 8 An example of a cubic tree from the family described in Proposition 11, partitioning the vertex-set V, with $|V| = 6$. In particular, $\Gamma_E(T_2, i_1) = \{a_1, a_2\}$, and $\Gamma_E(T_3, i_1) = \{b_1\}$, while bold edges represent the comb T_1 with leaf-set $\{i_1, i_2, i_3\}$.

Proof. Denote by $z(T) = \sum_{i,j} d_{ij}\tau_{ij}^T$, the cost of any feasible solution T to Max-$(3,\tau)$-NDP, and by $\Gamma_E(\tilde{T}, i) = \{j \in \Gamma(\tilde{T})\colon ij \in E\}$ for any $\Gamma(T)$-subtree $\tilde{T}$.

Suppose, by contradiction, that any optimal solution $T^* \in \mathcal{T}^{\max}_{(3,\tau)}$ does not respect the statement. Then, we will show that either T^* cannot be optimal or that it is possible to generate in polynomial time a sequence of cubic trees that starts from T^* and whose last element is an optimal solution to the Max-$(3,\tau)$-NDP that respects the statement, by leading in both cases to a contradiction.

First, assume that there is a comb T_0 of T^* such that $\Gamma(T_0) = V$. Because ℓ is strictly bigger than $2p$, then for every $i \in \Gamma\backslash\{r_1, r_2\}$, we must have either $\tau^{T^*}_{r_1,i} \geq p$, or $\tau^{T^*}_{r_2,i} \geq p$, or both. Assume the first case and denote by $\hat{i}$ the leaf of T such that r_1 and $\hat{i}$ form a cherry. Let $j_0 \in \Gamma(T_0)$ be any leaf of T^* such that $j_0 = \text{argmin}_{j\in\Gamma(T_0)}\{\tau^{T^*}_{j\hat{i}}\}$, and denote by T' the feasible solution obtained by swapping the label $\hat{i}$ with j_0. By construction, $\tau^{T'}_{j_0 r_1} \leq p$, hence $z(T') \geq z(T^*) - \sum_{j\in\Gamma_E(T_0,j_0)} \tau^{T^*}_{jj_0} + \sum_{j\in\Gamma_E(T_0,j_0)}(\tau^{T'}_{jj_0} + p) > z(T^*)$, contradicting the optimality of T^*. Thus, there are at least two disjoint combs of T^* that are maximal with respect to V.

We now proof that, if for any optimal solution T^*, there exist strictly more than two disjoint combs that are maximal with respect to the subset V, then one of the following two scenarios must hold: either T^* is not an optimal solution to the problem, or it is possible to transform in polynomial time T^* into a new feasible solution where the number of maximal combs with respect to V has been reduced by one while preserving the optimality of the solution. Specifically, let T_1 be one of the combs of T^* that are maximal in V, and suppose, without loss of generality, that both connected components T_2 and T_3 of $T^* \setminus T_1$ contain at least one leaf indexed by a vertex in V. Let $i_1, \ldots, i_{m_1}$ denote the leaves of T_1, for some integer $0 < m_1 < p-1$, and assume, without loss of generality, that $\tau^{T_1}_{i_j i_{j+1}} = 3$ for all $j = 1, \ldots, m_1 - 1$. Since i_1 and i_{m_1} are the extremals of T_1 and the comb is maximal in V, there must exist two leaves $\lambda_1, \lambda_2 \in L$ such that $\tau^{T^*}_{i_1\lambda_1} = \tau^{T^*}_{i_{m_1}\lambda_2} = 3$. Now, recall that T^* is a caterpillar by Proposition 10, hence, there exists a total order

$$2 < \tau^{T^*}_{i_1 a_1} < \ldots < \tau^{T^*}_{i_1 a_{m_2}}, \tag{4}$$

where $\{a_1, \ldots, a_{m_2}\} = \Gamma_E(T_2, i_1)$. Similarly, we have another total order

$$2 < \tau^{T^*}_{i_1 b_1} < \ldots < \tau^{T^*}_{i_1 b_{m_3}}, \tag{5}$$

where $\{b_1, \ldots, b_{m_3}\} = \Gamma_E(T_1, i_1) \cup \Gamma_E(T_3, i_1)$; see Figure 8.

Now, consider the effect of swapping i_1 with λ_1. In this case, all path-lengths $\tau^{T^*}_{i_1 a_j}$ for $j = 1, \ldots, m_2$ either increase or decrease by one unit, depending on the structure, while an opposite change occurs for all path-lengths associated with the leaves in $(\Gamma(T_1) \setminus \{i_1\}) \cup (\Gamma(T_3) \cap V)$.

Analogously, if we swap i_1 with i_2, the second leaf in the ordering of T_1, all values $\tau^{T}_{i_1 b_j}$ either increase or decrease by one, while the opposite change is observed for the leaves in $\Gamma(T_2) \cap V$ and $(\Gamma(T_1) \setminus \{i_1, i_2\}) \cup (\Gamma(T_3) \cap V)$. Then, swapping i_1 with λ_1 yields a new feasible solution T'' in which the cost changes by an amount equal to

$$\sum_{a_j\in\Gamma_E(T_2,i_1)} (\tau^{T^*}_{i_1 a_j} - 1) + \sum_{b_j\in\Gamma_E(T_1,i_1)\cup\Gamma_E(T_3,i_1)} (\tau^{T^*}_{i_1 b_j} + 1) - \sum_{a_j\in\Gamma_E(T_2,i_1)} \tau^{T^*}_{i_1 a_j} - \sum_{b_j\in\Gamma_E(T_1,i_1)\cup\Gamma_E(T_3,i_1)} \tau^{T^*}_{i_1 b_j},$$

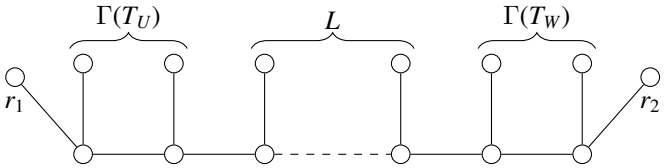

FIGURE 9 An example of a cubic tree from the family described in Proposition 11, partitioning the vertex-set V, with $|V| = 4$, into two subsets $\Gamma(T_U)$ and $\Gamma(T_W)$.

with respect to the cost of T^*. Similarly, swapping i_1 with i_2 yields a new feasible solution T''' in which the cost changes by an amount equal to

$$\sum_{a_j \in \Gamma_E(T_2, i_1)} (\tau^{T^*}_{i_1 a_j} + 1) + \sum_{b_j \in \Gamma_E(T_1, i_1) \cup \Gamma_E(T_3, i_1)} (\tau^{T^*}_{i_1 b_j} - 1) - \sum_{a_j \in \Gamma_E(T_2, i_1)} \tau^{T^*}_{i_1 a_j} - \sum_{b_j \in \Gamma_E(T_1, i_1) \cup \Gamma_E(T_3, i_1)} \tau^{T^*}_{i_1 b_j} +$$
$$\sum_{a'_j \in \Gamma_E(T_2, i_2)} (\tau^{T^*}_{i_2 a'_j} - 1) + \sum_{b'_j \in \Gamma_E(T_1, i_2) \cup \Gamma_E(T_3, i_2)} (\tau^{T^*}_{i_2 b'_j} + 1) - \sum_{a'_j \in \Gamma_E(T_2, i_2)} \tau^{T^*}_{i_2 a'_j} - \sum_{b'_j \in \Gamma_E(T_1, i_2) \cup \Gamma_E(T_3, i_2)} \tau^{T^*}_{i_2 b'_j},$$

with respect to the cost of T^*, where $\{a'_1, \dots, a'_{m'_2}\} = \Gamma_E(T_2, i_2)$ and $\{b'_1, \dots, b_{m'_3}\} = \Gamma_E(T_1, i_2) \cup \Gamma_E(T_3, i_2)$. By applying Proposition 9 first to the sequences (4) and (5), and then to the sequences $\{a'_i\}_{i=1}^{m'_2}$ and $\{b'i\}_{i=1}^{m'_3}$, one of the following three cases holds:

- $z(T'') > z(T^*)$;
- $z(T''') > z(T^*)$;
- $z(T'') \leq z(T^*)$ and $z(T''') = z(T^*)$.

The first two cases contradict the optimality of T^*. In the third case, fix i_1 in its current position and proceed by applying the same analysis to the swap between i_2 and i_3. If no improvement is found, we move to the next pair of vertices, and we iterate this swapping approach until terminating in one of the following two scenarios:

- either the final swap between i_{m_1} and λ_2 provides an increment of the total cost, contradicting the optimality of T^*;
- or, none of the swaps modifies the total cost. In this case, we can cyclically permute the positions of these leaves without affecting the objective: indeed, each i_j takes the place of i_{j+1}, for $j = 1, \dots, m_1 - 1$, i_{m_1} takes the place of λ_2, and λ_2 fills the position originally occupied by i_1, which has been vacated by the shift. Because this operation preserves feasibility and cost, we get a new feasible solution to the problem where T_1 is still a comb, but differs from T^* for the path-lengths of the extremals with respect to the cherries. Then, we can perform these cyclic swaps $\tau^{T^*}_{i_3, b_1} - 3$ times, i.e., until we obtain a unique comb $T_1 \cup T_3$ without changing the total cost or improving it, by Proposition 9.

Thus, after having performed a finite number of swaps, we can conclude that either T^* was not optimal, or that the comb T_1 is no longer maximal in V.

In conclusion, we can assume, without loss of generality, that the optimal tree T^* contains exactly two disjoint combs, denoted by T_U and T_W, both of which are maximal in V. Moreover, these two combs together cover all the vertices of V, i.e., $\Gamma(T_U) \cup \Gamma(T_W) = V$.

It is important to note that the optimality of T^* further implies that the connected components of the tree obtained by removing T_U and T_W from T^*, namely $T^* \setminus (T_U \cup T_W)$, consist precisely of the two singleton vertices $\{r_1\}$ and $\{r_2\}$, as well as an additional comb component C_L, whose leaves are indexed by the set L (see Figure 9). Therefore, to finalize the construction, it suffices to define the connected components $C_U = T_U \cup \{r_1\}$ and $C_W = T_W \cup \{r_2\}$. These components, together with C_L, yield the required partition of T^* into three connected components with the desired leaf indexing properties. Thus, setting $U = \Gamma(T_U)$ and $W = \Gamma(T_W)$ completes the proof. □

We now show that the partition of G provided by Proposition 11 constitutes an optimal solution for the given instance of the MCP. To this end, let z^* be the cost of any solution to the Max-$(3, \tau)$-NDP such that the cut induced by the partition

of Proposition 10 is optimal for the corresponding MCP instance. Let δ^* be the cardinality of such a solution. Similarly, let z' denote the cost of any feasible solution to the $(3,\tau)$-NDP for which the cut induced by the partition of Proposition 10 is not optimal for the MCP instance. Denote by δ' the cardinality of such a solution. Then, $z^* > \phi_1 = 2\tau^{T^*}_{r_1r_2} + 2\delta^*\ell$, since the contribution to the cost of r_1 and r_2 is constant for any feasible solution in $\mathcal{T}^{\max}_{(3,\tau)}$ by Proposition 10 and $\tau^{T^*}_{ij} > \ell$ for every $i \in U$ and $j \in W$ by Proposition 10. While, $z' < \phi_2 = 2\tau^{T^*}_{r_1r_2} + 4p^3 + 2\delta'(\ell + p + 1)$, where p is the maximal path length between any pair of leaves indexed by U or by W assuming that $G[U']$ and $G[W']$ are cliques such that $|U'| = |W'| = p-1$. Both assumptions give trivial upper bounds for any instance in $\mathcal{T}^{\max}_{(3,\tau)}$ not corresponding to an optimal solution to the MCP for G. By definition of the input and maximum cut, $\ell > 3p^3 + p^2$ and $\delta' \leq \delta^* - 1 < p^2$, and hence $\phi_1 < \phi_2$.

To conclude, for any optimal solution z^* there is a partition of G splitting G with the most number of edges, i.e., an optimal solution to the MCP for G. Hence, solving Problem 3 is at least as hard as solving the MCP for G. □

Generalization.
As in the previous section, we briefly observe here that Theorem 4 can be easily generalized to any integer $k > 3$. Specifically, let $G_1, G_2, \ldots, G_{k-2}$ be $k-2$ graphs isomorphic to G, then to account the hardness result of the Max-(k,τ)-NDP we consider the following instance:

- $\Gamma = V(G_1) \cup \ldots \cup V(G_{k-2}) \cup L \cup R = \{r_1^1, \ldots, r_1^{k-1}, \ldots, r_2^1, \ldots, r_2^{k-1}\}$, where L has cardinality $(k-2)\ell$ for some $\ell > 0$ and $r_1^j, r_2^j \notin V \cup L$ for all $j \in \{1, \ldots, k-1\}$;
- $d_{ij} = 1$ for all $ij \in E(G_1) \cup \ldots \cup E(G_{k-2})$;
- $d_{r_1^j r_2^j} = \omega$ for all $j \in \{1, \ldots, k-1\}$;
- and $d_{ij} = 0$ otherwise.

Then, (i) to note that for ω large enough, the optimal solution T^* is a caterpillar belonging to Θ^k_n such that $\tau^{T^*}_{r_1^j r_1^{j'}} = \tau^{T^*}_{r_2^j r_2^{j'}} = 2$ and $\tau^{T^*}_{r_1^j r_2^{j'}} = (|\Gamma| - 2)/(k-2) + 1$ for all $j \in \{1, \ldots, k-1\}$; and (ii) to note that for ℓ large enough and thank to the separability of the objective function, a similar result of 10 holds for any G_s. As a consequence, the above reduction must account for a scalar factor $k-2$ in any term involved. This term can be factored out without impacting the $\mathcal{NP}$-hardness argument, by leading us to the following general result:

Theorem 5. *The Max-(k,τ)-NDP is strongly $\mathcal{NP}$-hard for every integer $k \geq 3$.*

6 | CONSEQUENCES FOR RESTRICTED AND NONLINEAR OBJECTIVE VARIANTS

We present here additional complexity results for certain variants of the problems discussed in the previous sections. These results follow from direct adaptations of the earlier proof techniques.

On the complexity of the leaf-restricted k-NDP.
The objective function of the k-NDP measures the weighted path-length between all pairs of vertices in the input graph G. In some practical applications, however (see, e.g., [28]), this function may be restricted just to a subset of vertices of G, which usually is requested to be the set of leaves of the spanning tree of G. One can verify that the k-NDP remains $\mathcal{NP}$-hard even under this restriction.

Specifically, by reusing the same notation and input construction as in the proof for the decision version of the k-NDP, it is sufficient to observe that (i) by construction every spanning tree of G must include all edges in $E(V_R \cup V_S \cup V_{rs})$, hence the cost $L^w_T(R)$ is identical across all feasible solutions; (ii) for sufficiently large r, any optimal solution to the problem must belong to $\mathcal{T}^k$, since for any feasible solution $T' \notin \mathcal{T}^k$, there exist vertices $u \in V_S$ and $v \in R$ such that $L^w_{T'}(u,v) = 4$; and (iii) for any $T \in \mathcal{T}^k$, the quantity $L^w_T(\Gamma(T) \setminus (R \cup M))$ is constant, as $|\Gamma(T)| = |R| + |S| - \ell + |M|$. It follows that the $\mathcal{NP}$-completeness result extends to the considered restriction of the k-NDP, and the following result holds:

Theorem 6. *Given a graph $G = (V,E)$ of $\mathcal{G}^k_n$ and a weight function $w : E \to \mathbb{Q}^+$, find a spanning tree T of G belonging to Θ^k_n that minimizes $\sum_{i,j \in \Gamma(T)} L^w_T(i,j)$ is strongly $\mathcal{NP}$-hard.*

On the complexity of the fixed-tree Min-(k,τ)-NDP and Max-(k,τ)-NDP.
In the proofs of Theorems 2 and 4, respectively, we were able to fix the topology of the trees that are candidates for being optimal solutions by assuming that γ and ω are sufficiently large. Thus, without invoking Propositions 7 and 10, respectively, the reductions used for the Min-(k,τ)-NDP and the Max-(k,τ)-NDP remain valid even when we restrict feasible solutions to share the same topology, by leading to the following results:

Theorem 7. *Given G_n^k, a matrix $\mathbf{D}_\Gamma = \{d_{ij}\}$ and Γ-tree T, finding a Γ-tree isomorphic to T that minimizes $\sum_{i,j\in\Gamma} d_{ij}\tau_{ij}^T$ is strongly $\mathcal{NP}$-hard.*

Theorem 8. *Given G_n^k, a matrix $\mathbf{D}_\Gamma = \{d_{ij}\}$ and a Γ-tree T, finding a Γ-tree isomorphic to T that maximizes $\sum_{i,j\in\Gamma} d_{ij}\tau_{ij}^T$ is strongly $\mathcal{NP}$-hard.*

On the complexity of the (k,f)-NDP.
We denote (k,f)-NDP the variant of the k-NDP in which we replace every $L_T^w(u,v)$ by $f(L_T^w(u,v))$, where T is a tree of Θ_n^k, and $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing and such that $ax \leq f(x)$ for some $a > 0$. Now, as $f(x) \geq ax$, similar results to Propositions 2 and 3 hold for the (k,f)-NDP. In particular, note that the requirement on the size of $|R|$ can be considered fixed for every such f. Indeed, if $f(x) = ax$, then the objective function is nothing but $a \cdot L_T^w(V)$, while if f grows faster than any linear function, the size of R can be taken even smaller then the one suggested. Thus, the instance given remains polynomially bounded in m. Therefore the following result holds:

Theorem 9. *The (k,f)-NDP is strongly $\mathcal{NP}$-hard for every integer $k \geq 3$.*

On the complexity of the Min-$(k,f(\tau))$-NDP.
We denote Min-$(k,f(\tau))$-NDP the variant of the Min-(k,τ)-NDP in which we replace every τ_{ij}^T by $f(\tau_{ij}^T)$, where T is a Γ-tree, $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing and such that $ax \leq f(x) \leq b^x$ for some $a > 0$ and $b > 1$. Now, as $f(x) \geq ax$, similar results to Propositions 4, 5, 6, and 8 holds for the Min-$(k,f(\tau))$-NDP. Moreover, as $f(x) \leq b^x$, then the new value of the parameter s of the instance defined for proving the hardness result of the Min-(k,τ)-NDP remains polynomially bounded in p, and so is the size of the input matrix $\mathbf{D}_\Gamma$. It is easy to see that the same holds for all the other parameters M, γ_1, γ_2, and γ_3. Therefore the following result holds:

Theorem 10. *The Min-$(k,f(\tau))$-NDP is strongly $\mathcal{NP}$-hard for every integer $k \geq 3$.*

On the complexity of the Max-(k,τ^α)-NDP and the Max-(k,β^τ)-NDP.
We denote Max-(k,τ^α)-NDP and Max-(k,β^τ)-NDP the variants of the Max-(k,τ)-NDP in which we replace every τ_{ij}^T by $(\tau_{ij}^T)^\alpha$ and $\beta^{\tau_{ij}^T}$, respectively, where T is a Γ-tree, $\alpha \geq 1$ and $\beta > 1$ are real numbers. While we already proved Proposition 9 for every $\alpha \geq 1$, one can see that similar results to Propositions 10 and 11 hold for the Max-(k,τ^α)-NDP when assuming $\ell > \max\{2p, \left(2p^{\alpha+2} + p^2(p+1)^\alpha\right)^{\frac{1}{\alpha}}\}$ and $\omega > p^2(p+\ell+1)^\alpha$.

Theorem 11. *The Max-(k,τ^α)-NDP is $\mathcal{NP}$-hard for every integer $k \geq 3$ and $\alpha \geq 1$.*

The reduction described in Section 5 does not extend to the to the Max-(k,β^τ)-NDP. In fact, in this setting the input size $n = \ell + p + 2$ becomes exponential, since ℓ must grow much faster than any polynomial in order to ensure the optimality of the partition (U,W) in the corresponding instance of the reduction for the MCP with input graph G.

To the best of our knowledge, the bounds achievable in the two cases above are not sufficient to ensure that the size of the corresponding instances remains polynomial. Thus, the computational complexity of these last two problems remains an open question.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENTS

Daniele Catanzaro and Francesco Pisanu acknowledge support from the Belgian National Fund for Scientific Research (FNRS) via the grant FNRS PDR 40007831. Daniele Catanzaro also acknowledges support from the Fondation Louvain, via the grant "COALESCENS". Gwenaël Joret acknowledges support from the FNRS via the grants PDR T.0170.24, PDR T.W026.23, and CDR J.0115.26.

REFERENCES

1. Johnson DS, Lenstra JK, Kan AHGR. The complexity of the network design problem. *Networks.* 1978;8:279–285.
2. Pop PC. *Generalized network design problems: Modeling and optimization*. Berlin, Germany: De Gruyter, 2012.
3. Ahuja RK, Magnanti TL, Orlin JB. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Upper Saddle River, NJ, 1993.
4. Du DZ, Hu X. *Steiner tree problems in computer communication networks*. World Scientific Publishing Company, Singapore, 2008.
5. Crescenzi P, Kann V. A compendium of NP optimization problems. https://www.csc.kth.se/~viggo/wwwcompendium/node69.html; 2000.
6. Ravi R, Sundaram R, Marathe MV, Rosenkrantz DJ, Ravi SS. Spanning trees short or small. *SIAM Journal on Discrete Mathematics.* 1996;9(2):178–200.
7. Galbiati G, Maffioli F, Morzenti A. A short note on the approximability of the maximum leaves spanning tree problem. *Information Processing Letters.* 1994;52(1):45–49.
8. Furer M, Raghavachari B. Approximating the minimum-degree Steiner tree to within one of optimal. *Journal of Algorithms.* 1994;17(3):409-423.
9. Garg N. A 3-approximation for the minimum tree spanning k vertices. In: Proceedings of 37th Conference on Foundations of Computer Science. IEEE Computer Society, Burlington, VT 1996:302-309.
10. Lu H, Ravi R. The Power of Local Optimization: Approximation Algorithms for Maximum-Leaf Spanning Tree. In: Proceedings of the 38th Annual Allerton Conference on Communication, Control and Computing. IEEE Press, Pasadena, CA 2000.
11. Fiorini S, Joret G. Approximating the balanced minimum evolution problem. *Operations Research Letters.* 2012;40(1):31-35.
12. Frohn M. On the approximability of the fixed-tree balanced minimum evolution problem. *Optimization Letters.* 2021;15(6):2321–2329.
13. Gascuel O. *Mathematics of evolution and phylogeny*. Oxford University Press, New York, NY, 2005.
14. Gan G, Ma C, Wu J. *Data clustering: Theory, algorithms, and applications*. SIAM, Philadelphia, PA, 2007.
15. Gascuel O, Steel MA. *Reconstructing evolution*. Oxford University Press, New York, NY, 2007.
16. Catanzaro D, Labbé M, Pesenti R. The balanced minimum evolution problem under uncertain data. *Discrete Applied Mathematics.* 2013;161(13–14):1789–1804.
17. Catanzaro D, Pesenti R, Wolsey LA. On the Balanced Minimum Evolution polytope. *Discrete Optimization.* 2020;36:1-33.
18. Catanzaro D, Frohn M, Pesenti R. An information theory perspective on the balanced minimum evolution problem. *Operations Research Letters.* 2020;48(3):362-367.
19. Catanzaro D, Frohn M, Gascuel O, Pesenti R. A Tutorial on the Balanced Minimum Evolution Problem. *European Journal of Operational Research.* 2022;300(1):1–19.
20. Catanzaro D, Pesenti R, Sapucaia A, Wolsey L. Optimizing over path-length matrices of unrooted binary trees. *Mathematical Programming, in press.* 2025.
21. Qin T, Benthem vKJ, Valente L, Etienne RS. Parameter Estimation from Phylogenetic Trees Using Neural Networks and Ensemble Learning. *Systematic biology.* 2025:syaf060.
22. Azouri D, Abadi S, Mansour Y, Mayrose I, Pupko T. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nature communications.* 2021;12(1):1983.
23. Qin L, Chen Y, Pan Y, Chen L. A novel approach to phylogenetic tree construction using stochastic optimization and clustering. *Bmc Bioinformatics.* 2006;7(Suppl 4):S24.
24. Reinhard D. *Graph Theory*. Springer-Verlag, New York, NY, 2005.
25. Stott-Parker D, Ram P. The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees. *SIAM Journal on Computing.* 1996;28(5):1875–1905.
26. Garey MR, Johnson DS. *Computers and Intractability: A guide to the theory of NP-Completeness*. Freeman, New York, NY, 2003.
27. Garey MR, Johnson DS, Stockmeyer L. Some simplified NP-complete problems. In: Proceedings of the 6-th annual ACM symposium on Theory of computing. Association for Computing Machinery, Seattle, WA 1974:47–63.
28. Catanzaro D. The minimum evolution problem: Overview and classification. *Networks.* 2009;53(2):112-125.