Available online at www.sciencedirect.com

ScienceDirect

Theoretical Computer Science 363 (2006) 43–59

Theoretical Computer Science

www.elsevier.com/locate/tcs

# Computing phylogenetic roots with bounded degrees and errors is NP-complete

Tatsuie Tsukiji[a,*,1], Zhi-Zhong Chen[b,2]

[a] Department of Information Science, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan
[b] Department of Mathematical Sciences, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan

**Abstract**

In this paper we study the computational complexity of the following optimization problem: given a graph $G=(V, E)$, we wish to find a tree $T$ such that (1) the degree of each internal node of $T$ is at least 3 and at most $\Delta$, (2) the leaves of $T$ are exactly the elements of $V$, and (3) the number of errors, that is, the symmetric difference between $E$ and $\{\{u, v\} : u, v$ are leaves of $T$ and $d_T(u, v) \leqslant k\}$, is as small as possible, where $d_T(u, v)$ denotes the distance between $u$ and $v$ in tree $T$. We show that this problem is NP-hard for all fixed constants $\Delta, k \geqslant 3$.

Let $s_\Delta(k)$ be the size of the largest clique for which an error-free tree $T$ exists. In the course of our proof, we will determine all trees (possibly with degree 2 nodes) that approximate the $(s_\Delta(k) - 1)$-clique by errors at most 2.

*Keywords:* Phylogeny; Phylogenetic root; Computational biology; NP-hardness

## 1. Introduction

A phylogeny is a tree where the leaves are labeled by species and each internal node represents a speciation event whereby an ancestral species gives rise to two or more child species. The internal nodes of a phylogeny have degrees (in the sense of unrooted trees, i.e. the number of incident edges) at least 3. Proximity within a phylogeny in general corresponds to similarity in evolutionary characteristics. Nishimura et al. [15] and Lin et al. [12] initiated a graph-theoretic approach of reconstructing phylogenies from similarity data via a graph-theoretic approach, and investigated its computational feasibility. Specifically, interspecies similarity is represented by a graph where the vertices are the species and the adjacency relation represents evidence of evolutionary similarity. A phylogeny is then reconstructed from the graph such that the leaves of the phylogeny are labeled by the vertices of the graph (i.e. species) and for any two vertices in the graph, they are adjacent if and only if their corresponding leaves in the phylogeny are connected by a path of length at most $k$, where $k$ is a predetermined proximity threshold. To be clear, vertices in the graph are

[*] Corresponding author.
*E-mail addresses:* tsukiji@j.dendai.ac.jp (T. Tsukiji), chen@r.dendai.ac.jp (Z.-Z. Chen).

[1] Supported in part by the Grant-in-Aid for Scientific Research on Priority Areas from The Ministry of Education, Culture, Sports, Science and Technology, under Grant no. 771.

[2] Supported in part by the Grant-in-Aid for Scientific Research of the Ministry of Education, Science, Sports and Culture of Japan, under Grant no. 17500012.

0304-3975/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.tcs.2006.06.016

called *vertices* while those in the phylogeny *nodes*. Recall that the length of the (unique) path connecting two nodes $u$ and $v$ in phylogeny $T$ is the number of edges on the path, which is denoted by $d_T(u, v)$. This approach gives rise to the following algorithmic problem [12]:

> PHYLOGENETIC *k*TH ROOT PROBLEM (PR*k*): Given a graph $G = (V, E)$, find a phylogeny $T$ with the leaves labeled by the elements of $V$ such that for each pair of vertices $u, v \in V$, $\{u, v\} \in E$ if and only if $d_T(u, v) \leqslant k$.

Such a phylogeny $T$ (if exists) is called a *phylogenetic* $k$th *root*, or a $k$th *root phylogeny*, of graph $G$; conversely, graph $G$ is called the $k$th *phylogenetic power* of $T$. For convenience, we denote the $k$th phylogenetic power of a phylogeny $T$ as $\mathcal{P}_k(T)$, that is, $\mathcal{P}_k(T)$ has the vertex set $L(T) = \{u : u \text{ are leaves of } T\}$ and the edge set $T^k = \{\{u, v\} \mid u \text{ and } v \text{ are leaves of } T \text{ and } d_T(u, v) \leqslant k\}$. Thus, PR*k* asks for a phylogeny $T$ such that $G = \mathcal{P}_k(T)$.

The input graph in PR*k* is derived from some similarity data, which is usually inexact in practice and may have erroneous (spurious or missing) edges. Such errors may result in graphs that have no phylogenetic roots and hence we are interested in finding *approximate* phylogenetic roots for such graphs. For a graph $G = (V, E)$, each tree $T$ whose leaves are exactly the elements of $V$ and each internal node has degree at least 3 is called an *approximate* phylogeny of $G$, and the *error* of $T$ is $|T^k \oplus E| = |(E - T^k) \cup (T^k - E)|$. This motivated Chen et al. to consider the following problem:

> CLOSEST PHYLOGENETIC *k*TH ROOT PROBLEM (CPR*k*): Given a graph $G = (V, E)$ and a nonnegative integer $\ell$, decide if $G$ has an approximate phylogenetic $k$th root $T$ with at most $\ell$ errors.

An approximate phylogeny of $G$ with the minimum number of errors is called a *closest* $k$th *root phylogeny* of graph $G$.

In the practice of phylogeny reconstruction, most phylogenies considered are trees of degree 3 [17] because speciation events are usually bifurcating events in the evolutionary process. More specifically, in such *fully resolved* phylogenetic trees, each internal node has three neighbors and represents a speciation event that some ancestral species splits into two child species. Nodes of degrees higher than 3 are introduced only when the input biological (similarity) data are not sufficient to separate individual speciation events and hence several such events may be collapsed into a non-bifurcating (super) speciation event in the reconstructed phylogeny. These motivated Chen et al. [3] to consider restricted versions of PR*k* and CPR*k* where the output phylogeny is assumed to have degree at most $\Delta$, for some fixed constant $\Delta \geqslant 3$. We call these restricted versions the DEGREE-$\Delta$ PR*k* and the DEGREE-$\Delta$ CPR*k* problems, and denote them for short as $\Delta$PR*k* and $\Delta$CPR*k*, respectively.

### 1.1. Previous results on phylogenetic root problems

PR*k* was first studied in [12] where linear-time algorithms for PR2 and PR3 were proposed. A linear-time algorithm for the special case of PR4 where the input graph is required to be connected was also presented in [12]. At present, the complexity of PR*k* for $k \geqslant 5$ is still unknown.

Chen et al. [3] presented a linear-time algorithm that determines, for any input *connected* graph $G$ and constant $\Delta \geqslant 3$, if $G$ has a $k$th root phylogeny with degree at most $\Delta$, and if so, demonstrates one such phylogeny. Recently, Chen and Tsukiji [4] generalized it to work for any *disconnected* graph $G$, too. On the other hand, Chen et al. [3] showed that CPR*k* is NP-complete for any $k \geqslant 2$. One of their open questions asks for the complexity of $\Delta$CPR*k*.

Of special interest is CPR2. The problem CPR2 is essentially identical to the correlation clustering problem which has drawn much attention [1]. The proof of the NP-hardness of CPR2 given in [3] is also a valid proof of the NP-hardness of the correlation clustering problem. Blum et al. [1] obtained approximation algorithms for CPR2. Recently, Dom et al. [5] showed that CPR3 is fixed-parameter tractable with respect to the number of errors $\ell$.

### 1.2. Our contribution and proof idea

In this paper, we will show that $\Delta$CPR*k* is NP-complete, for all fixed constants $k \geqslant 3$ and $\Delta \geqslant 3$. This answers an open question in [3].

We first recall some basics found in the known NP-completeness proofs of CPR*k* [3]. The proof is a reduction from the FITTING ULTRAMETRIC TREES problem, using *critical cliques* for the gadget constructions in their reduction. A critical clique of graph $G$ is a maximal subset of vertices that are adjacent to each other and have a common neighborhood in $G$. In the reduction, a gadget graph $G'$ contains an input graph $G$ and a number of critical cliques $C_i$ of $G'$. Then, for an appropriate construction of such $G'$ where each $C_i$ is larger than $\ell$, and any tree $T$ such that $L(T) = V(G')$ and $T^k$ approximates $G'$ with error at most $\ell$, the following property is shown to hold: $d_T(t, v) = \lfloor k/2 \rfloor + 1$ for all vertices $v$

of $G$, where $t$ is a node of $T$ fixed independently of $v$. More precisely, if $d_T(t, v) \neq \lfloor k/2 \rfloor + 1$ then any tree $T$ is shown to break all adjacency relations between $v$ and some $C_i$, while if $d_T(t, v) = \lfloor k/2 \rfloor + 1$ then some $T$ satisfies all these between $v$ and all $C_i$. We remark that if the degree of phylogeny were bounded by a constant, then the gap of errors that any two trees could make in the adjacency relations around any fixed $v$ would be bounded by a constant, too. Hence, the proof known for the unbounded degree case do not carry over to the bounded degree case.

Next, we give basics of our NP-completeness proof of 3CPR3. The proof is a reduction from an NP-hard special case of the HAMILTONIAN PATH problem, where all vertices of the input graph $G$ have degree at most 3 and exactly two of them are of degree 1. Let $2\ell$ be the number of degree-3 vertices in $G$. We will show that $G$ has a Hamiltonian path if and only if it has an approximate phylogeny with error $\ell$. Intuitively, this is because any Hamiltonian path can be "lifted up" to form a phylogeny with error $\ell$, and vice versa. See Fig. 5.

The NP-completeness of 3CPR$k$ for each odd $k \geqslant 5$ will be given by the generalization of that of 3CPR3. Let $G$ be an instance graph of the HP. We construct a gadget graph $G' = (V', E')$ from $G$ as follows: replace each vertex $v$ of $G$ with a component graph $H(v)$ having a specific vertex identified with $v$. These components $H(v)$ of the $G'$ are copies of the same graph $H$ having the following property: for any tree $T$ of degree 3 and whose vertices are those in $V$, if $T^k$ approximates $H$ by errors at most 2, then there exists an internal node $\alpha$ of $T$ such that $\alpha$ has degree 2, the other internal nodes of $T$ have degree 3, $d_T(\alpha, v) = \lfloor k/2 \rfloor - 1$, and $d_T(\alpha, u) > \lfloor k/2 \rfloor - 1$ for all other vertices $u$ of $H$ (we call $H$ a *k-padding graph*). For this graph $G'$, we show that $G$ has a Hamiltonian path if and only if $G'$ has an approximate phylogeny with error $\ell$. For it, we will claim that for any tree $T$ of degree 3 and whose leaves are those in $V'$, the phylogenetic power $T^k$ approximates $G'$ by errors at most $\ell$ if and only if the following two conditions hold: (i) for each $v \in V$, the subtree of $T$ induced on the vertices of $H(v)$ must have the internal node $\alpha_v$ as above, and (ii) via the identification $v \leftrightarrow \alpha_v$, the third power $T^3$ must induce a Hamiltonian path of $G$.

In order to prove NP-completeness of $\Delta$CPR$k$ for each $\Delta, k \geqslant 3$, we thus provide a construction of *$(\Delta, k, h, \ell)$-padding graphs*, which are graph $G = (V, E)$ with the following properties (where trees may have degree 2 internal nodes):

- There is a tree $T$ of maximum degree $\Delta$, whose leaves are exactly the vertices in $V$, and satisfies the following condition: $T$ has a unique unsaturated (i.e. degree $< \Delta$) internal node $\alpha$, the degree of $\alpha$ is $\Delta - 1$, $d_T(\alpha, u) = h$ for just one vertex $u$ of $G$, and $d_T(\alpha, v) > h$ for all other vertices $v$ of $T$.
- For any tree $T$ of maximum degree $\Delta$, whose leaves are the vertices in $V$, and has at least one unsaturated node, if $|E \oplus T^k| \leqslant \ell$, then $T$ must satisfy the above condition.

### 1.3. Organization of the paper

Next is a short section for notations and definitions. In Section 3 we construct a family of $(\Delta, k, \lfloor k/2 \rfloor - 1, 2)$-padding graphs for every fixed $\Delta \geqslant 3$ and $k \geqslant 4$. NP-completeness of $\Delta$CPR$k$ is established in Section 4. Section 5 concludes the paper with an open problem.

## 2. Notations and definitions

We employ standard terminologies in graph theory. In particular, for a graph $G$, $V(G)$ and $E(G)$ denote the sets of vertices and edges of $G$, respectively. Two graphs $G = (V, E)$ and $G' = (V', E')$ are *isomorphic*, which we denote by $G \cong_\varphi G'$, if there is a one-to-one correspondence $\varphi$ between $V$ and $V'$ such that $\{u, v\} \in E$ if and only if $\{\varphi(u), \varphi(v)\} \in E'$. Let $G$ be a graph and $u$ and $v$ be its two vertices. The *distance* $d_G(u, v)$ between $u$ and $v$ is the number of edges in a shortest path from $u$ to $v$ in $G$. The *neighborhood* of $v$ in $G$, which we denote by $N_G(v)$, is the set of vertices adjacent to $v$ in $G$. The *degree* of $v$ in $G$ is $|N_G(v)|$, and is denoted by $d_G(v)$. Similarly, for a tree $T$, $V(T)$, $E(T)$, and $L(T)$ denote the sets of nodes, edges and leaves of $T$, respectively.

We also introduce, for convenience, some new terminologies of trees. For a tree $T$ of maximum degree $\Delta$, an internal node $\alpha$ of $T$ is *unsaturated* if $d_T(\alpha) \leqslant \Delta - 1$. Tree $T$ is *i-extensible* if $i = \sum_v (\Delta - deg_T(v))$, where the summation is taken over all unsaturated internal nodes $v$ of $T$. Tree $T$ is *h-away* if for each unsaturated internal node $\alpha$ of $T$ there is just one leaf $u_\alpha$ of $T$ such that $d_T(\alpha, u_\alpha) = h$ and further $d_T(\alpha, v) > h$ for all other leaves $v$. We call a 1-extensible and $h$-away tree an $h$-padding tree or $(\Delta, h)$-padding tree if its maximum degree is bounded by $\Delta$, and refer to its unique unsaturated internal node $\alpha$ as the *internal port* of $T$, and the unique leaf $u_\alpha$ having distance $h$ from $\alpha$ in $T$ the *external port* of $T$, respectively. Figs. 3 and 4 give examples of padding trees; a (3, 2)-padding tree in Fig. 3 and a (4, 1)-padding tree in Fig. 4.

For any set $U$ of nodes of $T$, $T[U]$ denotes the minimum subtree containing $U$. Note that all leaves of $T[U]$ are contained in $U$. A phylogeny is a tree with no degree 2 nodes. A tree is a phylogeny that is allowed to have degree 2 internal nodes. As already mentioned, the $k$th phylogenetic power of any tree $T$ is denoted as $\mathcal{P}_k(T) = (L(T), T^k)$, where $T^k$ is the set of all edges $\{u, v\}$ with $\{u, v\} \subseteq L(T)$ and $d_T(u, v) \leqslant k$.

## 3. Construction of $(\Delta, k, \lfloor k/2 \rfloor - 1, 2)$-padding graphs

In this section we construct $(\Delta, k, \lfloor k/2 \rfloor - 1, 2)$-padding graphs for all fixed constants $\Delta \geqslant 3$ and $k \geqslant 4$; we first provide padding graphs for $\Delta = 3$, and generalize them to $(\Delta, k, \lfloor k/2 \rfloor - 1, 2)$-padding graphs for any fixed $\Delta \geqslant 4$ in the second subsection.

### 3.1. *Construction for $\Delta = 3$*

Throughout this subsection, all trees and phylogenies are of maximum degree 3 or less.

As observed in [3], the maximum size of a clique that has a $k$th root phylogeny is given by the following function:

$$s(k) = \begin{cases} 3 \cdot 2^{k/2-1} & \text{if } k \text{ is even,} \\ 2^{(k+1)/2} & \text{if } k \text{ is odd.} \end{cases}$$

Obviously, up to isomorphism, there is exactly one tree of maximum degree 3 whose $k$th phylogenetic power realizes $s(k)$-clique; we denote this tree by $C_k$. Similarly, up to isomorphism, there is exactly one tree of maximum degree 3 and having exactly one degree 2 node whose $k$th phylogenetic power realizes $(s(k) - 1)$-clique; we denote this tree by $D_k$. By definition, every $D_k$ is a 1-padding tree. Fig. 1 depicts $C_4$, $C_5$, and $C_6$. Notice that the trees $D_4$, $D_5$, and $D_6$ are obtained from $C_4$, $C_5$, and $C_6$, respectively, by removing just one leaf, say the sibling leaf of $\varphi(u)$.

**Lemma 3.1.** *For every tree $T$ (of maximum degree 3), if there are two leaves $u$ and $v$ with $d_T(u, v) = k$, and all leaves $w$ of $T$ have distance at most $k$ from both $u$ and $v$, then $T$ is isomorphic to a subtree of $C_k$.*

**Proof.** Immediate from the definition of $C_k$. □

**Corollary 3.2.** *For any tree $T$, if $d_T(u, v) \leqslant k$ for all leaves $u$ and $v$, then $T$ is isomorphic to a subtree of $C_k$.*

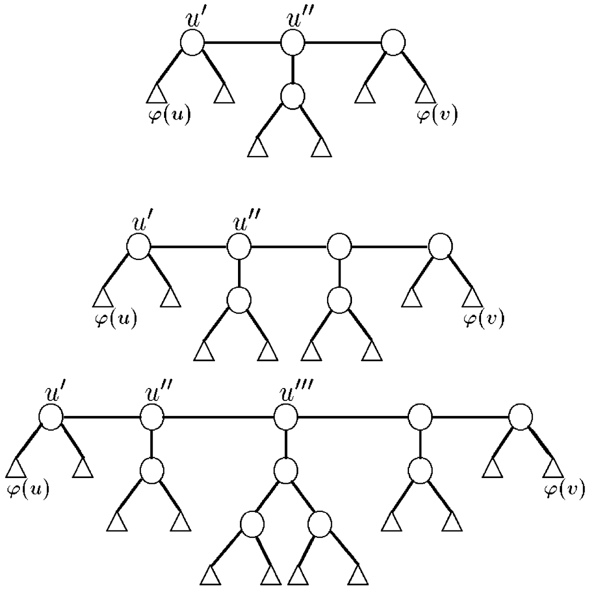

Fig. 1. $C_4$, $C_5$, and $C_6$.

**Proof.** Let $\{u, v\}$ be a pair of leaves such that $d_T(u, v)$ is maximum. If $d_T(u, v) = k$ then all leaves $w$ have distance at most $k$ from both $u$ and $v$, and Lemma 3.1 proves that $T$ is isomorphic to a subtree of $C_k$. Otherwise, $d_T(u, v) = \ell < k$ but then $T$ is isomorphic to a subtree of $C_\ell$, which is a subtree of $C_k$. So, $T$ is isomorphic to a subtree of $C_k$. □

The following trivial facts will be frequently used in the proofs of the following lemmas.

**Fact 1.** *For every tree T with* $|L(T)| = s(k) - 1$, $|T^k| \geqslant \binom{s(k)-1}{2} - 2$ *if and only if* $d_T(u, v) \leqslant k$ *for all but at most two unordered pairs* $\{u, v\}$ *of leaves of T.*

**Fact 2.** *Let T be a tree and* $u, v, w, x$ *be its nodes. If* $d_T(u, x) = d_T(w, x)$ *and both of the paths from u to w and that from v to w go through x, then* $d_T(u, w) = d_T(v, w)$.

**Fact 3.** *Let T be a tree and S be a set of leaves of T. Then any path between a node in S and that in* $V[L(T) - S]$ *contains a node x in* $V[L(T) - S]$ *such that* $deg_{T[L(T)-S]}(v) < deg_T(v)$.

**Fact 4.** *For any subtree* $C'_k$ *of* $C_k$, *if* $L(C_k) - L(C'_k)$ *consists of* $n_2$ *pairs of the two sibling leaves and the other* $n_1$ *leaves (whose sibling leaves are not in* $L(C_k)$*), then* $|L(C_k)| - |L(C'_k)| \geqslant n_1 + n_2$.

**Lemma 3.3.** *Let* $k \geqslant 4$. *For every tree T with* $|L(T)| = s(k) - 1$, *if* $|T^k| \geqslant \binom{s(k)-1}{2} - 2$, *then T has two leaves u and v of distance exactly k between them.*

**Proof.** Let $\{u, v\}$ be a pair of leaves of $T$ such that $d_T(u, v)$ is the maximum of $d_T(x, y)$ over all unordered pairs $\{x, y\}$ of leaves $x, y$ with $d_T(x, y) \leqslant k$, and let $\ell = d_T(u, v)$. By definition, $\ell \leqslant k$. To prove $\ell = k$, we assume, $\ell \leqslant k - 1$. Then, $T$ has no pair $\{x, y\}$ of leaves such that $\ell + 1 \leqslant d_T(x, y) \leqslant k$, and we can derive a contradiction in each of the following cases.

*Case* 1: $k \geqslant 6$. Since $s(\ell)$ is an upper bound on the number of leaves of $T$ whose distances from $u$ and $v$ are both bounded by $k$, at least $s(k) - 1 - s(\ell) \geqslant 3$ leaves have distance greater than $k$ from $u$ or $v$. So, $|T^k| \leqslant \binom{s(k)-1}{2} - 3$, a contradiction.

*Case* 2: $k = 4$ and $\ell = 3$. Since $s(3) = 4 = s(4) - 2$, at least one leaf $w$ of $T$ has distance greater than $k$ from $u$ or $v$. Without loss of generality, we can assume $d_T(u, w) \geqslant k + 1 = 5$. Moreover, since $d_T(u, v) = 3$ we have $|d_T(u, w) - d_T(v, w)| \leqslant 1$. So $d_T(v, w) \geqslant 5$. By Fact 1, all leaves of $T$ other than $w$ have distances at most 3 from both $u$ and $v$. By Lemma 3.1, $T[L(T) - \{w\}]$ is isomorphic to a subtree of $C_3$, and $T[L(T) - \{w\}] \ncong C_3$ by Fact 3. So, $4 = s(4) - 2 = |L(T) - \{w\}| \leqslant |L(C_3)| - 1 = 3$, a contradiction.

*Case* 3: $k = 5$ and $\ell = 4$. Since $s(4) = 6 = s(5) - 2$, at least one leaf of $T$ has distance greater than $\ell$ from $u$ or $v$. Moreover, by Fact 1, there are at most two such leaves, so by Lemma 3.1 there is a set $S \subseteq L(T)$ such that $1 \leqslant |S| \leqslant 2$ and $T[L(T) - S]$ is isomorphic to a subtree $C'_4$ of $C_4$. We further distinguish two subcases as follows.

*Subcase* 3.1: $|S| = 1$. Since $C'_4 \ncong C_4$, $6 = s(5) - 2 = |L(T) - S| \leqslant |L(C_4)| - 1 = s(4) - 1 = 5$, a contradiction.

*Subcase* 3.2: $|S| = 2$. Let $w$ be a vertex in $S$ that has distance at least 6 from $u$ or $v$, but not from both (by Fact 1). Without loss of generality, we can assume $d_T(u, w) \geqslant 6$ and $d_T(v, w) \leqslant 4$. Let $\varphi$ be an isomorphism from $T[L(T) - S]$ to $C'_4$. Tree $C_4$ has the four leaves that are not farther from $\varphi(u)$ than $\varphi(v)$, among which only $\varphi(u)$ can belong to $L(C'_4)$ by Facts 1 and 2. Thus by Fact 4, $5 = s(5) - 3 = |L(T) - S| = |L(C'_4)| \leqslant s(4) - 2 = 4$, a contradiction. □

**Lemma 3.4.** *Let* $k \geqslant 6$. *For any tree T with* $|L(T)| = s(k) - 1$, *if T has three leaves u, v, and w such that* $d_T(u, v) = k$ *and w has distance greater than k from u or v, then* $|T^k| \leqslant \binom{s(k)-1}{2} - 3$.

**Proof.** For a contradiction we assume that some tree $T$ with $|L(T)| = s(k) - 1$ has three such leaves $u, v, w$ and $|T^k| \geqslant \binom{s(k)-1}{2} - 2$. Without loss of generality, $d_T(u, w) \geqslant d_T(v, w)$. Let $S$ be the set of leaves having distance greater than $k$ from $u$ or $v$. Then $1 \leqslant |S| \leqslant 2$, $w \in S$, and by Lemma 3.1 $T[L(T) - S] \cong_\varphi C'_k$ for a subtree $C'_k$ of $C_k$. Root $T$ at $v$ and $C_k$ at $\varphi(v)$, respectively. Let $u'''$ be the great-grandparent of $\varphi(u)$ in $C_k$.

*Case* 1: $w$ is not a descendant of $\varphi^{-1}(u''')$. Thus, by Fact 3, there is at least one leaf in $L(C_k) - L(C'_k)$ that is not a descendant of $u'''$. On the other hand, in $L(C_k)$ node $u'''$ has eight leaf descendants, among which at most two

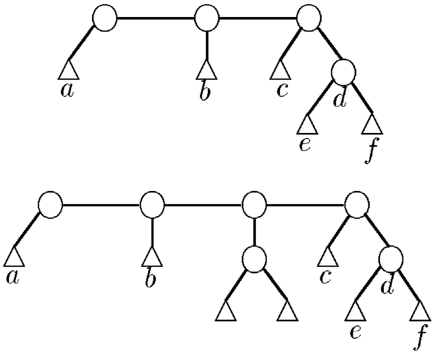

Fig. 2. $E_4$ and $E_5$.

(including $\varphi(u)$) can belong to $L(C'_k)$ by Facts 1 and 2. So by Fact 4, $s(k)-1-|S| = |L(T)-S| = |L(C'_k)| \leqslant s(k)-4 \leqslant s(k)-2-|S|$, a contradiction.

*Case* 2: $w$ is a descendant of $\varphi^{-1}(u''')$. Then, $k=6$ and $d_T(w,u) = d_T(w,v) \geqslant k+1$. Tree $C_k$ has eight leaves that have distance 4 or less from either $u$ or $v$, among which only $\varphi(u)$ and $\varphi(v)$ can belong to $L(C'_k)$ by Facts 1 and 2. So by Fact 4, $s(k)-3 \leqslant |L(T)-S| = |L(C'_k)| \leqslant s(k)-2$, a contradiction. □

**Lemma 3.5.** *Let* $k \geqslant 6$. *For any tree* $T$ *with* $|L(T)| = s(k)-1$, *if* $|T^k| \geqslant \binom{s(k)-1}{2} - 2$, *then* $T$ *is* 1-*extensible or* 0-*extensible*, *and in the former case* $T \cong D_k$.

**Proof.** Let $T$ be a tree having $s(k)-1$ leaves and $|T^k| \geqslant \binom{s(k)-1}{2} - 2$. By Lemma 3.3 $T$ has two leaves $u$ and $v$ such that $d_T(u,v) = k$, and by Lemma 3.4 all leaves $w$ of $T$ partake distance at most $k$ from both $u$ and $v$. So by Lemma 3.1 $T$ is isomorphic to a subtree $C'_k$ of $C_k$ where $|L(C'_k)| = |L(C_k)|-1$. Since all internal nodes of $C_k$ have degree 3, $C'_k$ is obtained from $C_k$ by removing one leaf or two sibling leaves. In the former case $T \cong D_k$ while in the latter case $T$ is 0-extensible. □

For $k \in \{4,5\}$, let $E_k$ be the tree in Fig. 2.

**Lemma 3.6.** *Let* $k \in \{4,5\}$. *For any tree* $T$ *with* $|L(T)| = s(k)-1$, *if* $|T^k| \geqslant \binom{s(k)-1}{2} - 2$, *then* $T$ *is* 1-*extensible or* 0-*extensible*, *and in the former case* $T \cong D_k$ *or* $T \cong E_k$.

**Proof.** Let $T$ be a tree having $s(k)-1$ leaves and $|T^k| \geqslant \binom{s(k)-1}{2} - 2$. By Lemma 3.3 $T$ has leaves $u$ and $v$ with $d_T(u,v) = k$. Let $S$ be the set of leaves having distance greater than $k$ from $u$ or $v$. By assumption $0 \leqslant |S| \leqslant 2$. By Lemma 3.1, $T[L(T)-S] \cong_\varphi C'_k$ for a subtree $C'_k$ of $C_k$. Root $T$ at $v$, $C_k$ at $\varphi(v)$, and $E_k$ at the non-leaf node adjacent to $v$, respectively, and let $u''$ be the grandparent of $\varphi(u)$ in $C_k$. The proof proceeds in three cases.

*Case* 1: $|S| = 0$, i.e. $S = \emptyset$. In this case $T \cong C'_k$, so either $T \cong D_k$ or $T$ is 0-extensible.

*Case* 2: $|S| = 1$. Let $S = \{w\}$. Without loss of generality, $d_T(u,w) \geqslant d_T(v,w)$.

*Subcase* 2.1: $w$ is not a descendant of $\varphi^{-1}(u'')$. So, by Fact 3, there is at least one leaf in $L(C_k) - L(C'_k)$ that is not a descendant of $u''$. We show that $T \cong E_k$ where the leaves $u$, $v$, and $w$ in $T$ correspond to $e$ (or $f$), $b$, and $a$ in $E_k$, respectively (see Fig. 2). In $C_k$, $u''$ has the four leaf descendants, among which at most two (including $\varphi(u)$) can belong to $L(C'_k)$ by Facts 1 and 2. Since $|L(C'_k)| \geqslant s(k)-2$, all of $\varphi(u)$, its sibling, and its uncle must belong to $C'_k$, and by Fact 1, $d_T(u,w) = k+1$. Accordingly, $T \cong E_k$.

*Subcase* 2.2: $w$ is a descendant of $\varphi^{-1}(u'')$. Since $d_T(u,w) \geqslant d_T(v,w)$, $k=4$ and $d_T(u,w) = d_T(v,w) \geqslant 5$. Among four leaves of $C_4$ having distance at most 2 from $\varphi(u)$ or $\varphi(v)$, only $\varphi(u)$ and $\varphi(v)$ belong to $L(C'_4)$. Moreover, by Fact 3, there is a leaf in $L(C_k) - L(C'_k)$ that is neither the sibling of $\varphi(u)$ nor that of $\varphi(v)$. Therefore, by Fact 4, $s(4)-2 = |L(T) - \{w\}| = |L(C'_4)| \leqslant s(4)-3$, a contradiction.

*Case* 3: $|S| = 2$. Let $S = \{w, y\}$, and without loss of generality, $d_T(u,w) \geqslant d_T(v,w)$. We show that $T \cong E_k$, where the leaves $u$, $v$, $w$, and $y$ in $T$ correspond to $a$, $c$, $e$, and $f$ (see Fig. 2). Note that $s(k)-3 = |L(T)-S| = |L(C'_k)|$.

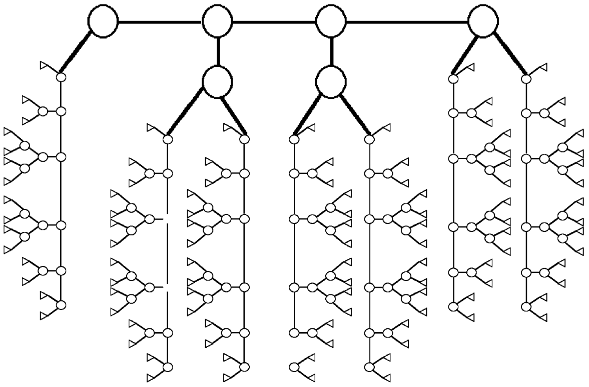

Fig. 3. $R_{7,2}$.

By Facts 1 and 2, neither $w$ nor $y$ is a descendant of $\varphi^{-1}(u'')$. Among the four leaf descendants of $u''$, only $\varphi(u)$ belongs to $L(C'_k)$. Since $|L(C'_k)| \geqslant s(k)-3$, the uncle of $\varphi(u)$ must belong to $L(C'_k)$, and by Fact 1, $d_T(u,w) = d_T(u,y) = k+1$. Accordingly, $T \cong E_k$. □

**Corollary 3.7.** *For every* $k \geqslant 4$, $(s(k)-1)$*-clique is a* $(3,k,1,2)$*-padding graph.*

**Proof.** By Lemma 3.5 for $k \geqslant 6$ and by Lemma 3.6 for $k \in \{4,5\}$. □

Let $h_k = \lfloor k/2 \rfloor - 1$. We first construct a $h_k$-padding tree and derive $(3,k,h_k,2)$-padding graph as its $k$th power. The construction of the padding trees proceeds recursively, piling up the roots of $(s(k)-1)$-cliques level by level. Formally, we define a tree $R_{k,h}$ for each odd $k \geqslant 5$ and $1 \leqslant h \leqslant h_k$ in the following way (see Fig. 3 for $R_{7,2}$):

- Let $g(i) = \prod_{j=1}^{i}(s(2j+3)-1)$ and $g(0)=1$.
- $\tilde{R}_{k,h}$ is a leveled tree of level $h$ such that $g(i)$ nodes are placed at level $i$ $(0 \leqslant i \leqslant h)$, and each node at level $i < h$ is joined to some $s(2i+5)-1$ nodes at level $i+1$.
- $R_{k,h}$ is an expansion of $\tilde{R}_{k,h}$ such that each internal node $v$ of $\tilde{R}_{k,h}$ at level $i$ $(0 \leqslant i \leqslant h-1)$ is expanded to a copy $D(v)$ of $D_{2i+5}$, where $v$ is identified with the internal port of $D(v)$ and the child nodes of $v$ in $\tilde{R}_{k,h}$ are identified with the leaves of $D(v)$.

Note that the $k$th power of $R_{k,h}$ is the following graph: it has $g(h)$ nodes $v_0, v_1, \ldots, v_{g(h)-1}$ such that for each $0 \leqslant i < h$ and each $0 \leqslant j < g(i)$, the vertices $v_{jq_i}, v_{jq_i+q_{i+1}}, v_{jq_i+2q_{i+1}}, \ldots, v_{jq_i+(s(2i+5)-2)q_{i+1}}$, where $q_i = g(h)/g(i)$, are mutually joined to form the $(s(2i+5)-1)$-clique.

By construction and the following fact, $R_{k,h}$ is an $h$-padding tree, whose internal port is the unique degree-2 node of $D_5$ and whose external port is the external port of the $D_k$ hooking to the external port of $R_{k,h-1}$ (note that $R_{k,h}$ has only one external port).

**Fact 5.** *Let $T$ be an arbitrary $k$-padding tree and $T_v$ (for each leaf $v$ of $T$) be an arbitrary $k'$-padding tree. Then, a tree obtained from these by identifying each $v$ with the internal port of $T_v$ is a $(k+k')$-padding tree. The internal port of the obtained tree is the internal port of $T$, and its external port is the external ports of $T_u$ for the external port $u$ of* $T$.

In the rest of this subsection we will demonstrate that the $k$th phylogenetic power of tree $R_{k,h_k}$ is a $(3,k,h_k,2)$-padding graph.

**Lemma 3.8.** *Let* $k \geqslant 4$. *Let $T$ be a tree such that $T$ is not* 0*-extensible*, $|L(T)| = s(k)-1$ *and* $|T^k| \geqslant \binom{s(k)-1}{2} - 2$. *Let $F$ be the tree obtained by joining a new leaf to an arbitrary leaf of $T$. Then,* $|F^k| \leqslant \binom{s(k)-1}{2} - 3$.

**Proof.** For a contradiction, suppose that $|F^k| \geqslant \binom{s(k)-1}{2} - 2$. By construction $F$ is not 0-extensible tree and has $s(k) - 1$ leaves. If $k \geqslant 6$ then by Lemma 3.5 $T \cong D_k$ and $F \cong D_k$, but obviously $T \not\cong F$, a contradiction. Similarly, if $k \in \{4, 5\}$ then by Lemma 3.6 both $T$ and $F$ are isomorphic with $D_k$ or $E_k$, but joining a new leaf to any leaf of $D_k$ or $E_k$ gives a graph having different topology with $D_k$ and $E_k$, a contradiction. □

**Lemma 3.9.** *Let $k \geqslant 5$ be odd. For any tree $T$ with $L(T) = L(R_{k,h_k})$, if $|T^k \oplus R^k_{k,h_k}| \leqslant 2$ then $T$ is* 1*-extensible or* 0*-extensible, and in the former case it is $h_k$-away as well.*

**Proof.** By induction on $k \geqslant 5$. The case $k = 5$ has been done in Lemma 3.6, because $R_{5,1} = D_5$. So fix odd $k$ to be greater than or equal to 7. For simplicity of notations let $\tilde{R} = \tilde{R}_{k,h_k}$, $R = R_{k,h_k}$, and $h = h_k$. Let $\Gamma_i$ be the set of nodes of $\tilde{R}$ at level $i$ and $L_x = L(D(x))$ for every non-leaf node $x$ of $\tilde{R}$. Then, $L(R) = \Gamma_h = \bigcup_{x \in \Gamma_{h-1}} L_x$.

Consider an arbitrary tree $T$ such that $T$ is not 0-extensible, $L(T) = L(R)$ and $|T^k \oplus R^k| \leqslant 2$. For every $x \in \Gamma_{h-1}$, $\mathcal{P}_k(R[L_x])$ is an $(s(k) - 1)$-clique and $|T^k[L_x] \oplus R^k[L_x]| \leqslant 2$. By Lemma 3.5, $T[L_x]$ is a 1-padding tree. For each $x \in \Gamma_{h-1}$ let $x'$ be its parent internal node in tree $\tilde{R}$, $\alpha_x$ be the internal port of $T[L_x]$, and $u_x$ (respectively, $v_x$) be the external port of $T[L_v]$ (respectively, $R[L_v]$).

Let $S = T[\{\alpha_x : x \in L_{x'}\}]$. To prove that $u_x = v_x$, we assume, by a contradiction, that $u_x \neq v_x$ for some $x \in \Gamma_{h-1}$. By construction $|L_{x'}| = s(k-2) - 1$, $\mathcal{P}_{k-2}(R[L_{x'}])$ is an $(s(k-2) - 1)$-clique, and so is $\mathcal{P}_k(R[\{v_x : x \in L_{x'}\}])$. Since $|T^k \oplus R^k| \leqslant 2$, $|T^k[\{v_x : x \in L_{x'}\}]| \geqslant \binom{s(k-2)-1}{2} - 2$, and $d_T(x, v_x) \geqslant 1$ for all $x \in L_{x'}$, implying $|S^{k-2}| \geqslant \binom{s(k-2)-1}{2} - 2$. Lemma 3.5 (or Lemma 3.6 for $k = 7$) thus determines the topology of $S$ (which is not 0-extensible). Let $F$ be the tree obtained from $S$ by joining a new leaf to $x$. By Lemma 3.8, $|F^{k-2}| \leqslant \binom{s(k)-1}{2} - 3$, while we also have $|F^{k-2}| \geqslant \binom{s(k-2)-1}{2} - 2$, since $1 = d_T(x, u_x) < d_T(x, v_x)$ and $|T^k[\{v_x : x \in L_{x'}\}]| \geqslant \binom{s(k-2)-1}{2} - 2$, a contradiction.

Now, $u_x = v_x$ for all $x \in \Gamma_{h-1}$. Let $T_0 = T[\{\alpha_x : x \in \Gamma_{h-1}\}]$ and $R_0 = R[\Gamma_{h-1}]$. Since $d_T(\alpha_x, u_x) = 1$ for all $x \in \Gamma_{h-1}$, $d_{T_0}(x, y) = d_T(\alpha_x, \alpha_y) - 2$ for all $x, y \in \Gamma_{h-1}$, and $d_{R_0}(x, y) = d_R(u_x, u_y) - 2$ as well, showing that $|T_0^{k-2} \oplus R_0^{k-2}| \leqslant 2$ if we identify $x$ with $\alpha_x$ for every $x \in \Gamma_{h-1}$. By the induction hypothesis $T_0$ is an $h_{k-2}$-padding tree. Further, all subtrees $T[L_x]$, $x \in \Gamma_{h-1}$, are 1-extensible and 1-away, so $T$ is 1-extensible and $h_k$-away. □

**Theorem 3.10.** *For every odd $k \geqslant 5$, $\mathcal{P}_k(R_{k,h_k})$ is a $(3, k, h_k, 2)$-padding graph.*

For every even $k \geqslant 4$, we will recursively construct trees $R_{k,\lfloor k/2 \rfloor - 1}$ in a manner parallel to the odd case, by replacing $s(2i + 5)$, $g(i)$ and $D_{2i+5}$ with $s(2i + 4)$, $\prod_{j=1}^{i}(s(2j + 2) - 1)$ and $D_{2i+4}$ here. Then, Lemma 3.9 and Theorem 3.10 hold in parallel for every even constant $k \geqslant 4$, too. In summary, we have proved the following theorems.

**Theorem 3.11.** *Let $k \geqslant 4$. For any tree $T$ such that $L(T) = L(R_{k,h_k})$, if $|T^k \oplus R^k_{k,h_k}| \leqslant 2$ then $T$ is* 1*-extensible or* 0*-extensible, and in the former case $T$ is $h_k$-away as well.*

**Theorem 3.12.** *For every $k \geqslant 4$, $\mathcal{P}_k(R_{k,h_k})$ is a $(3, k, h_k, 2)$-padding graph.*

### *3.2. Generalization to $\Delta \geqslant 4$*

This subsection sketches proofs generalizing the results in Section 3.1 to the case where $\Delta \geqslant 4$. The purpose is construction of $(\Delta, k, h_k + i, 2)$-padding graphs, $(\Delta, k, h_k - 1, 2)$-padding graphs, and $(\Delta, k, h_k + 1, 0)$-padding graphs, for all fixed constants $\Delta \geqslant 3$ and $k \geqslant 4$.

The maximum size of clique having a $k$th root phylogeny of maximum degree $\Delta$ is given by the following function:

$$s_\Delta(k) = \begin{cases} \Delta \cdot (\Delta - 1)^{k/2-1} & \text{if } k \text{ is even,} \\ 2 \cdot (\Delta - 1)^{(k-1)/2}, & \text{if } k \text{ is odd.} \end{cases}$$

We denote by $C_{\Delta,k}$ a tree of maximum degree $\Delta$ such that its $k$th phylogenetic power is an $s_\Delta(k)$-clique. Obviously, up to isomorphism, such a tree is uniquely determined. Let $E_{\Delta,k}$ be a tree of maximum degree $\Delta$ such that its $k$th

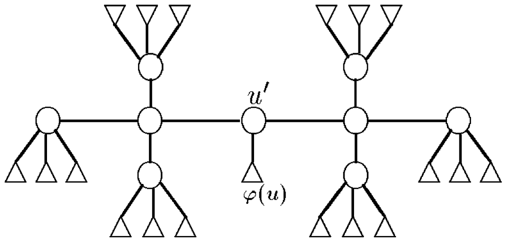

Fig. 4. $D_{4,4}$.

phylogenetic power is an $(s_\Delta(k) - \Delta + 2)$-clique and having a unique unsaturated internal node; alternatively, $E_{\Delta,k}$ is a tree of maximum degree $\Delta$ such that its $k$th phylogenetic power is a $(s_\Delta(k) - \Delta + 2)$-clique and having a leaf without any sibling leaves. Again, up to isomorphism, such a tree is uniquely determined. Let $D_{\Delta,k}$ be the tree obtained from $\Delta - 2$ copies of $E_{\Delta,k}$ by identifying their unique unsaturated internal nodes and removing all but one leaf adjacent to it. Particularly, $D_{\Delta,k}$ is a 1-padding tree of maximum degree $\Delta$. See Fig. 4 for $D_{4,4}$.

All of Lemma 3.1, Corollary 3.2, Facts 1–3, and Lemma 3.3 hold but replacing maximum degree 3 therein by maximum degree $\Delta$ here. Fact 4 is generalized in the following form.

**Fact 6.** *For any subtree $C'_{\Delta,k}$ of $C_{\Delta,k}$, if $L(C_{\Delta,k}) - L(C'_{\Delta,k})$ contains $n_i$ sets of $i$ mutually sibling leaves, for each $i \leqslant i \leqslant \Delta - 1$, then $|L(C_k)| - |L(C'_k)| \geqslant \sum_{i=1}^{\Delta} n_i \max(\Delta - i - 1, 1)$.*

**Lemma 3.13.** *Let $k \geqslant 4$ and $\Delta \geqslant 4$. For every tree $T$ of maximum degree $\Delta$ with $s_\Delta(k) - \Delta + 2$ leaves, if $T$ has a leaf with no sibling leaves and $|T^k| \geqslant \binom{s_\Delta(k) - \Delta + 2}{2} - 2$, then $T \cong E_{\Delta,k}$.*

**Proof.** Let $T$ be a tree of maximum degree $\Delta$, having $s_\Delta(k) - \Delta + 2$ leaves, one of which has no sibling leaves, and $|T^k| \geqslant \binom{s_\Delta(k) - \Delta + 2}{2} - 2$. By Lemma 3.3, $T$ has two leaves $u$ and $v$ of distance $k$ between them, and further $u$ can be assumed as a leaf with no sibling leaves. Let $S$ be the set of leaves having distance greater than $k$ from $u$ or $v$. By assumption $0 \leqslant |S| \leqslant 2$, and by Lemma 3.1, $T[L(T) - S] \cong_\varphi E'_{\Delta,k}$ for some subtree $E'_{\Delta,k}$ of $E_{\Delta,k}$.

We claim that $S = \emptyset$. For contradiction, suppose $|S| \geqslant 1$, and let $w$ be an element of $S$. We prove only the case that $d_T(u, w) \geqslant d_T(v, w)$. Root $T$ at $v$ and $E_{\Delta,k}$ at $\varphi(v)$, respectively, and let $u''$ be the grandparent of $\varphi(u)$ in $E_{\Delta,k}$.

*Case* 1: $w$ is not a descendant of $\varphi^{-1}(u'')$. So, by Fact 3, there is a leaf in $L(E_{\Delta,k}) - L(E'_{\Delta,k})$ that is not a descendant of $u''$. In $E_{\Delta,k}$, $u''$ has $(\Delta - 1)(\Delta - 2) + 1$ leaf descendants, among which at most two (including $\varphi(u)$) can belong to $L(E'_{\Delta,k})$ by Facts 1 and 2, hence by Fact 4, $s_\Delta(k) - \Delta \leqslant |L(E'_{\Delta,k})| \leqslant s_\Delta(k) - (\Delta - 2)(\Delta - 1) \leqslant s_\Delta(k) - \Delta - 1$, a contradiction.

*Case* 2: $w$ is a descendant of $\varphi^{-1}(u'')$. Then, $k = 4$, $d_T(u, w) = d_T(v, w) \geqslant 5$ and by Fact 3, there is a leaf in $L(E_{\Delta,k}) - L(E'_{\Delta,k})$ which is neither a sibling of $\varphi(u)$ nor that of $\varphi(v)$. By Facts 1 and 2, none of the siblings of $\varphi(u)$ or $\varphi(v)$ can belong to $E'_{\Delta,4}$. By Fact 4, $s_\Delta(4) - \Delta \leqslant |L(C'_{\Delta,4})| \leqslant s_\Delta(4) - 2(\Delta - 2) - 1 \leqslant s_\Delta(4) - \Delta - 1$, a contradiction.

Now, the claim $S = \emptyset$ holds, so $T \cong E'_{\Delta,k}$ and $|L(T)| = |L(E_{\Delta,k})|$, hence $T \cong E_{\Delta,k}$. □

**Lemma 3.14.** *Let $k \geqslant 4$ and $\Delta \geqslant 4$. For every tree $T$ of maximum degree $\Delta$ and $|L(T)| = (\Delta - 2)(s_\Delta(k) - \Delta + 1) + 1$, if $|T^k \oplus \mathcal{P}_k(D_{\Delta,k})| \leqslant 2$ then $T \cong E_{\Delta,k}$.*

**Proof.** By construction, graph $\mathcal{P}_k(D_{\Delta,k})$ has the vertex $v_0$ adjacent to all other vertices, while any other vertex $v \neq v_0$ is unadjacent with at least $\Delta - 1$ vertices of the graph, i.e. the co-degree of $v$ is at least $\Delta - 1$.

Let $T$ be the tree such that $L(T) = L(D_{\Delta,k})$ and $|T^k \oplus D^k_{\Delta,k}| \leqslant 2$, and let $v$ be a vertex whose degree is maximum over the vertices in $T^k$.

To prove that $v = v_0$, we suppose $v \neq v_0$ and derive a contradiction. Let $\ell$ be the co-degree of $v$ in $T^k$. Since the co-degree of $v$ in $D^k_{\Delta,k}$ is $\Delta - 1$ or greater, $|T^k - D^k_{\Delta,k}| \geqslant \Delta - 1 - \ell$. On the other hand the co-degree of $v_0$ in $D^k_{\Delta,k}$ is 0, hence $|D^k_{\Delta,k} - T^k| \geqslant \ell$ as well. Together, $|T^k \oplus D^k_{\Delta,k}| \geqslant \Delta - 1 - \ell + \ell = \Delta - 1 \geqslant 3$, a contradiction.

Now, $v = v_0$, and by the above argument $v_0$ is the only vertex having the maximum degree of the vertices in $T^k$. Particularly $v_0$ has no sibling leaf in $T$. Let $E_i$ be the $i$th copy of $E_{\Delta,k}$ constituting the tree $D_{\Delta,k}$ and let $L_i = L(E_i)$. Notice that $v_0$ belongs to every $L_i$. Since $|T^k[L_i] \oplus E^k_{\Delta,k}| \leqslant 2$ and $v_0$ has no sibling leaf in $T[L_i]$, Lemma 3.13 shows that $T[L_i] \cong E_{\Delta,k}$ for every $L_i$, hence $T \cong D_{\Delta,k}$. □

**Corollary 3.15.** *For all fixed constants $\Delta \geqslant 4$ and $k \geqslant 4$, $\mathcal{P}_k(D_{\Delta,k})$ is a $(\Delta, k, 1, 2)$-padding graph.*

We can prove the following lemma, whose proof is analogous with that of Lemma 3.14, hence omitted.

**Lemma 3.16.** *For every $\Delta \geqslant 3$, $\mathcal{P}_3(D_{\Delta,3})$ is a $(\Delta, 3, 1, 0)$-padding graph.*

We construct a phylogeny $R_{\Delta,k,h_k}$ of degree $\Delta$ recursively in the same way as $R_{k,h_k}$ but replacing $s$ and $D_i$ therein with $(\Delta - 2)(s_\Delta - \Delta + 1) + 1$ and $D_{\Delta,i}$, respectively. Then Theorems 3.11 and 3.12 are generalized as follows:

**Theorem 3.17.** *Let $k \geqslant 4$ and $\Delta \geqslant 3$. For any tree $T$ of maximum degree $\Delta$ with $L(T) = L(R_{\Delta,k,h_k})$, if $|T^k \oplus R^k_{\Delta,k,h_k}| \leqslant 2$ then $T$ is 1-extensible or 0-extensible, and in the former case $T$ is $h_k$-away.*

**Theorem 3.18.** *For every $k \geqslant 4$ and $\Delta \geqslant 3$, $\mathcal{P}_k(R_{\Delta,k,h_k})$ is a $(\Delta, k, h_k, 2)$-padding graph.*

For each odd $k \geqslant 3$, we construct $R_{\Delta,k,h_k-1}$ (respectively, $R_{\Delta,k,h_k+1}$) in the same way as $R_{\Delta,k,h_k}$ but replacing $g(i)$ therein by $\prod_{j=1}^{i}((\Delta - 2)(s_\Delta(2j + 5) - \Delta + 1) + 1)$ (respectively, $\prod_{j=1}^{i}((\Delta - 2)(s_\Delta(2j + 1) - \Delta + 1) + 1)$). For each even $k \geqslant 4$, $R_{\Delta,k,h_k-1}$ is defined similarly, replacing $g(i)$ by $\prod_{j=1}^{i}((\Delta-2)(s_\Delta(2j+4)-\Delta+1)+1)$. So, $R_{\Delta,k,1} = D_{\Delta,k}$ for each $k \in \{3, 6, 7\}$.

**Theorem 3.19.** *For all fixed constants $\Delta \geqslant 3$ and $k \geqslant 6$, $\mathcal{P}_k(R_{\Delta,k,h_k-1})$ is a $(\Delta, k, h_k - 1, 2)$-padding graph.*

**Proof.** By analogy of the recursive proof of Lemma 3.9, using Corollary 3.7 for $\Delta = 3$ and Corollary 3.15 for $\Delta \geqslant 4$. □

**Theorem 3.20.** *For every odd $k \geqslant 3$ and every $\Delta \geqslant 3$, $\mathcal{P}_k(R_{\Delta,k,h_k+1})$ is a $(\Delta, k, h_k + 1, 0)$-padding graph.*

**Proof.** By analogy of the recursive proof of Lemma 3.9, using Lemma 3.16 for $k = 3$ and Lemma 3.14 for $k \geqslant 5$. □

For each even number $k \geqslant 4$, we do not have a $(h_k + 1)$-padding tree of degree $\Delta \geqslant 3$. Instead, we will use a join of $\Delta - 1$ copies of the $h_k$-padding graph $R_{\Delta,k,h_k}$. In more precise, let $S_{\Delta,k,h_k+1}$ be a tree consisting from $\Delta - 1$ copies $R_i$ of the $R_{\Delta,k,h_k}$ (where $R_{\Delta,3,0}$ consists from the single node), and an extra internal node $\alpha$ of degree $\Delta - 1$ joined to the internal ports of $R_i$.

**Theorem 3.21.** *For every even number $k \geqslant 3$ and every $\Delta \geqslant 3$, the graph $\mathcal{P}_k(S_{\Delta,k,h_k+1})$ is a $(\Delta, k, h_k + 1, 0)$-padding graph, but having $\Delta - 1$ external ports.*

**Proof.** Let $T$ be a tree of maximum degree $\Delta$ such that $T$ is not 0-extensible, $L(T) = L(S_{\Delta,k,h_k+1})$ and $T^k = S^k_{\Delta,k,h_k+1}$. Since $T^k[L(R_i)] = R^k_i$, by Theorem 3.18, $T[L(R_i)]$ is 1-extensible and $h_k$-away. Further, $T^k = S^k_{\Delta,k,h_k+1}$ forms the clique on the external ports of $R_1, R_2, \ldots, R_{\Delta-1}$, so $T$ must induce the star graph on the internal ports of these $R_i$. Consequently, $T$ is 1-extensible and $(h_k + 1)$-away, whose external ports are those of $R_i$. □

## 4. The NP-hardness of $\Delta$CPR$k$

This section proves that $\Delta$CPR$k$ is NP-complete. The first subsection is for $\Delta = 3$ and odd $k$, and the second subsection for $\Delta = 3$ and even $k$. The last subsection generalizes these results, showing that $\Delta$CPR$k$ is NP-complete for all fixed $k \geqslant 3$ and $\Delta \geqslant 3$.

Let us briefly discuss the complexities of $\Delta$CPR$k$ for $k \leqslant 2$ or $\Delta \leqslant 2$. First of all, $\Delta$CPR1 and 3CPR2 are not interesting at all: for any sufficiently large graph $G = (V, E)$, a tree $T = (V, E(T))$, and a nonnegative integer $\ell$, $|T^1 \oplus E| \leqslant \ell$ if and only if $|E| \leqslant \ell$; suppose further that $T$ is of maximum degree 3, then $|T^2 \oplus E| \leqslant \ell$ if and only if $G$ contains a matching of size at least $|E| - \ell$. Hence, $\Delta$CPR1 and 3CPR2 are efficiently solvable. Secondly, $\Delta$CPR2 is NP-complete for all $\Delta \geqslant 4$ by a straightforward reduction from the PARTITION INTO $\Delta$CLIQUES problem; It is given a graph with $\Delta q$ vertices, and determine whether there is a partition of the vertices into $q$ disjoint sets of size $\Delta$, such that the graph induces the $\Delta$-clique on each of these sets. A proof of its NP-completeness can be found in [11].

Therefore, this section considers the complexity of $\Delta$CPR$k$ for each $k \geqslant 3$ and $\Delta \geqslant 3$. We reduce the following version of the HAMILTONIAN PATH PROBLEM (HP) to $\Delta$CPR$k$.

HAMILTONIAN PATH PROBLEM (HP): Given a graph $G = (V, E)$ such that

- all nodes are of degree 3 or less,
- exactly two (specific) vertices are of degree 1, each of which being adjacent to a vertex of degree 2, and
- there is no cycle of length less than 5,

find a Hamiltonian path of $G$, i.e. find a linear ordering of the vertices of $G$ such that each pair of consecutive vertices are adjacent in $G$.

NP-completeness proofs of the problem can be found in [16,7, Section 9.3].

### 4.1. The case where $\Delta = 3$ and $k$ is odd

Fix $\Delta = 3$. Throughout this section, all trees and phylogenies are of maximum degree 3 or less. For every fixed odd integer $k \geqslant 3$, we prove that 3CPR$k$ is NP-complete by a reduction from HP. We begin with the NP-hardness proof of 3CPR3 because those for larger odd $k$ are the generalization of it. In addition, it provides most ideas and tools for the general proofs.

For every graph $G = (V, E)$ and tree $T$ such that $L(T) = V$, and every $k \geqslant 3$, we define a function $f_{G,T,k}$ mapping each $v \in V$ to a multiple of $\frac{1}{2}$ as follows:

$$\begin{aligned} f_{G,T,k}(v) = {} & \tfrac{1}{2}\,|\{u \ : \ \{u, v\} \in E, \ \text{and}\, d_T(u, v) > k\}| \\ & + |\{\{u, w\} \ : \ u \neq w, \{u, v\} \in E, \{v, w\} \in E, \ \{u, w\} \notin E, \ \text{and}\, d_T(u, w) \leqslant k\}|. \end{aligned}$$

In words, $f$ counts the number of disagreements between $T^k$ and $E$ around each vertex $v$. Specifically, $f$ assigns weight $\frac{1}{2}$ to each disagreement contained in $E - T^k$ and adjacent to $v$ by $\frac{1}{2}$, and weight 1 to each disagreement contained in $T^k - E$ and taken between neighbors of $v$.

**Lemma 4.1.** *Let $G = (V, E)$ be a graph of maximum degree $\leqslant 3$ without cycles of length less than* 5, *and let $T$ be a tree such that $L(T) = V$. Then, $\sum_{v \in V} f_{G,T,3}(v) \leqslant |T^3 \oplus E|$.*

**Proof.** Let us figure out the contribution of each unordered pair $\{u, v\}$ of vertices in $V$ to the sum $\sum_{v \in V} f_{G,T,3}(u)$. If $\{u, v\} \notin T^3 \oplus E$ then it contributes nothing at all. Every edge $\{u, v\}$ in $E - T^3$ contributes $\frac{1}{2}$ to both $f_{G,T,3}(u)$ and $f_{G,T,3}(v)$, so contributes 1 to the sum. Every edge $\{u, w\}$ in $T^3 - E$ contributes 1 to each $f_{G,T,3}(v)$ where $v$ is a common neighbor of $u$ and $w$. Since there is at most one common neighbor of $u$ and $w$ because $G$ contains no cycle of length 4, $\{u, w\}$ contributes 1 to the sum. In total, the set of all unordered pairs of vertices adds at most $|T^3 \oplus E|$ to the sum, implying $\sum_{v \in V} f_{G,T,3}(v) \leqslant |T^3 \oplus E|$. □

**Lemma 4.2.** *Let $G$ and $T$ be as in Lemma* 4.1. *Let $v$ be a vertex of $G$ having three pairwise nonadjacent neighbors $u_1, u_2$, and $u_3$. Then $f_{G,T,3}(v) = \frac{1}{2}$ or $f_{G,T,3}(v) \geqslant 1$, and in the former case $d_T(u_i, v) > 3$ for one $u_i \in \{u_1, u_2, u_3\}$ and $d_T(u_j, v) = 3$ for the other two $u_j \in \{u_1, u_2, u_3\} - \{u_i\}$.*

**Proof.** It suffices to show that if either (i) $d_T(u_i, v) = 2$ for some $u_i$ or (ii) $d_T(u_i, v) = 3$ for all $u_i$, then $f_{G,T,3}(v) \geqslant 1$. Note that if both (i) and (ii) are false, then it must be the case that $d_T(u_i, v) \geqslant 4$ for some $u_i$, hence $f_{G,T,3}(v) \geqslant \frac{1}{2}$. Let $v'$ (respectively, $u_i'$) be the internal node adjacent to $v$ (respectively, $u_i$) in $T$.

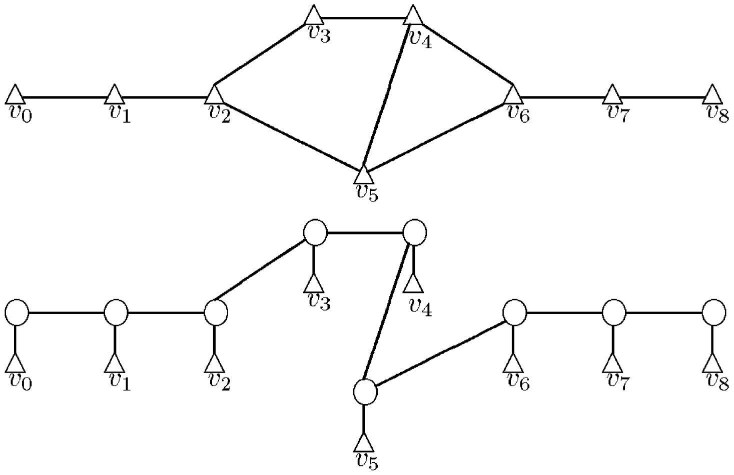

Fig. 5. A graph and the bridge along $v_0, v_1, \ldots, v_8$.

Suppose (i). Without loss of generality, $d_T(u_1, v) = 2$. If $d_T(u_2, v) \leqslant 3$ in addition then $d_T(u_1, u_2) \leqslant 3$, implying that $\{u_1, u_2\}$ contributes 1 to $f_{G,T,3}(v)$; on the other hand if both $u_2$ and $u_3$ have distance greater than 3 from $v$ then each of $u_2$ and $u_3$ contributes $\frac{1}{2}$ to $f_{G,T,3}(v)$, so they contribute 1 in total. Thus, (i) always implies $f_{G,T,3}(v) \geqslant 1$.

Suppose (ii). A simple inspection shows $|\{u'_1, u'_2, u'_3\}| \leqslant 2$, i.e. $u'_i = u'_j$ for some $u_i \neq u_j$, then $\{u_i, u_j\}$ contributes 1 to $f_{G,T,3}(v)$. □

Let $v_0, \ldots, v_{n+1}$ be distinct $n+2$ elements. The *bridge along* $v_0, \ldots, v_{n+2}$ is a tree $T$ such that $L(T) = \{v_0, \ldots, v_{n+1}\}$, $L(V) - L(T) = \{v'_0, \ldots, v'_{n+1}\}$, and $E(T) = \{\{v_i, v'_i\} : 0 \leqslant i \leqslant n+1\} \cup \{\{v'_i, v'_{i+1}\} : 0 \leqslant i \leqslant n\}$. See Fig. 5 for an example. Particularly, $T^3$ forms a simple path visiting the leaves $v_0, v_1, \ldots, v_{n+1}$ through. Note that only $v'_0$ and $v'_{n+1}$ of the internal ports in the bridge are unsaturated, which we call the *terminals* of the bridge. To extend a bridge into a phylogeny, we need to join its two terminals to other nodes outside the bridge.

**Lemma 4.3.** *Let G and T be as in Lemma* 4.1, *and further let G have exactly two vertices of degree* 1, *and exactly* $2\ell$ *vertices of degree* 3. *If* $\sum_{v \in V} f_{G,T,3}(v) \leqslant \ell$, *then T must be a bridge along a Hamiltonian path of G.*

**Proof.** Without loss of generality, $G$ has $n+2$ vertices $v_0, v_1, \ldots, v_{n+1}$, where $v_0$ and $v_{n+1}$ are of degree 1. Let $W$ be the set of degree 3 vertices in $G$.

By Lemma 4.2, $f_{G,T,3}(v) \geqslant \frac{1}{2}$ for all $v \in W$; so $\ell \leqslant \sum_{v \in W} f_{G,T,3}(v) \leqslant \sum_{v \in V} f_{G,T,3}(v) \leqslant \ell$, where all inequalities must be equality. This shows that
(i) $f_{G,T,3}(v) = \frac{1}{2}$ for all $v \in W$ and
(ii) $f_{G,T,3}(v) = 0$ for all $v \in V - W$.
Let $v'$ be the internal node in $T$ adjacent to vertex $v \in V$. We claim that $|\{v'_0, \ldots, v'_{n+1}\}| = n+2$. For a contradiction, assume $v' = w'$ for some vertices $v \neq w$ in $V$. A contradiction is derived in each of the following cases.

*Case* 1: $v \in W$. Let $u_1$, $u_2$, and $u_3$ be the three neighbors of $v$ in $G$. If further $w \in \{u_1, u_2, u_3\}$ then $f_{G,T,3}(v) \geqslant 1$; otherwise $d_T(u_i, u_j) = 2$ for some $u_i \neq u_j$, or $d_T(u_i, v) > 3$ and $d_T(u_j, v) > 3$ for some $u_i \neq u_j$. In any case, $f_{G,T,3}(v) \geqslant 1$, a contradiction against (i).

*Case* 2: $v \in V - W - \{v_0, v_{n+1}\}$. Let $u_1$ and $u_2$ be the neighbors of $v_i$. If further $w \in \{u_1, u_2\}$ then $f_{G,T,3}(v) \geqslant \frac{1}{2}$; otherwise either $u'_1 = u'_2$ or at least one $u_i$ has distance greater than 3 from $v$. In any case, $f_{G,T,3}(v) \geqslant \frac{1}{2}$, a contradiction against (ii).

*Case* 3: $v = v_0$ and $w = v_{n+1}$. Then either $d_T(v_0, v_1) > 3$ or $d_T(v_n, v_{n+1}) > 3$ must hold, and in the former case $f_{G,T,3}(v_0) \geqslant \frac{1}{2}$, while in the latter case $f_{G,T,3}(v_{n+1}) \geqslant \frac{1}{2}$, so in either case we get a contradiction against (ii).

Now, $|\{v'_0, \ldots, v'_{n+1}\}| = n+2$, i.e. $d_T(v, w) \geqslant 3$ for every distinct vertices $v$ and $w$ in $V$. Then, by (i) and Lemma 4.2, for every $v \in W$, $v'$ is adjacent to just two of the three nodes $u'_1, u'_2, u'_3$ in $T$ where $u_1, u_2, u_3$ are the neighbors of $v$

in $G$; moreover by (ii) for every $v \in V - W - \{v_0, v_{n+1}\}$ $v'$ is adjacent to both of $u_1'$ and $u_2'$ where $u_1$ and $u_2$ are the neighbors of $v$ in $G$. Consequently, tree $T$ must be a bridge along $v_0, v_1, \ldots, v_{n+1}$. □

**Lemma 4.4.** 3CPR3 *is NP-complete.*

**Proof.** Let $G = (V, E)$ be an arbitrary instance of HP. Let $V$, $W$, and $\ell$ be as in the proof of Lemma 4.3. Let and $v_1$ and $v_n$ be the unique neighbors of $v_0$ and $v_n$ in $G$, respectively. Recall that $d_G(v_1) = d_G(v_n) = 2$. Let $G' = (V', E')$ be a graph such that $V' = V \cup \{\tilde{v}_0, \tilde{v}_{n+1}\}$ and $E' = E \cup \{\{v_0, \tilde{v}\}, \{\tilde{v}_0, v_1\}, \{v_{n+1}, \tilde{v}_{n+1}\}, \{\tilde{v}_{n+1}, v_n\}\}$.

It suffices to show that $G$ has a Hamiltonian path if and only if $G'$ has an approximate phylogeny $T$ with error at most $\ell$.

Suppose that $G$ has a Hamiltonian path given by an ordering of the vertices, $v_0, v_1, v_2, \ldots, v_n, v_{n+1}$, where each pair of consecutive vertices are adjacent in $G$. Let $T'$ be the bridge along this Hamiltonian path, and let $T$ be a tree such that $L(T) = V'$, $V(T) - L(T) = V(T') - L(T')$ and $E(T) = E(T') \cup \{\{\tilde{v}_0, v_0'\}, \{\tilde{v}_{n+1}, v_{n+1}'\}\}$. Thus, we join the terminal of the bridge $T$ to $\tilde{v}_0$, and the other into $\tilde{v}_{n+1}$. Then, $T$ is a phylogeny of $G'$, and by construction, $|T^3 \oplus E'| = |\{\{v_i, v_j\} \in E \,:\, |i - j| > 1\}| = \ell$.

Conversely, suppose that $T$ is an approximate phylogeny of $G'$ with $|T^3 \oplus E'| \leqslant \ell$. By Lemma 4.1, $\sum_{v \in V} f_{G,T[V],3}(v) \leqslant |T^3 \oplus E'| \leqslant \ell$, so by Lemma 4.3, $T[V]$ must be a bridge along a Hamiltonian path of $G$. □

Fix any odd number $k \geqslant 5$. Now we turn to prove that 3CPR$k$ is NP-complete by the generalization of that of 3CPR3. Let $G = (V, E)$ be an arbitrary instance graph of HC. Let $padd_k(G) = (V_k(G), E_k(G))$ be the graph obtained from $G$ that replaces every vertex $v$ in $V$ with a copy $H(v)$ of the $(3, k, h_k, 2)$-padding graph $\mathcal{P}_k(R_{k,h_k})$ and identify $v$ with the external port of $H(v)$. Thus, $V_k(G) \supseteq V$ and $padd_k(G)$ induces the input graph $G$ as the subgraph on $V$. Further, for each $v \in V$, $V_k(G) \supseteq V(H(v))$ and $padd_k(G)$ induces $H(v)$ as the subgraph on $V(H(v))$.

For each $v \in V$ let $L_v = V(H(v))$ and $E_v = E(H(v))$. For any tree $T$ such that $L(T) = V_k(G)$, we define a function $g_{padd_k(G),T,k}$ from every $v \in V$ to a fractional value, such as to count the averaged disagreements between $T^k[L_u]$ and $E(H(u))$ over all $u \in N_G(v) \cup \{v\}$. Formally,

$$g_{padd_k(G),T,k}(v) = \sum_{u \in N_G(v) \cup \{v\}} \frac{|T^k[L_u] \oplus E_u|}{d_G(u) + 1}.$$

**Lemma 4.5.** *For any tree $T$ such that $L(T) = V_k(G)$, $\sum_{v \in V}(f_{G,T[V],k}(v) + g_{padd_k(G),T,k}(v)) \leqslant |T^k \oplus E_k(G)|$.*

**Proof.** By the definitions of $f_{G,T[V],k}$ and $g_{padd_k(G),T,k}$. □

**Lemma 4.6.** *For any tree $T$ such that $L(T) = V_k(G)$, if $\sum_{v \in V}(f_{G,T,k}(v) + g_{padd_k(G),T,k}(v)) \leqslant \ell$, then $G$ must have a Hamiltonian path.*

**Proof.** Let us abbreviate the functions $f_{G,T,k}$ and $g_{padd_k(G),T,k}$ as $f_k$ and $g_k$, respectively. We claim that $f_k(v) + g_k(v) \geqslant \frac{1}{2}$ for every degree-3 vertex $v$ of $G$. For a contradiction, assume $f_k(v) + g_k(v) < \frac{1}{2}$, so both $f_k(v)$ and $g_k(v)$ are assumed smaller than $\frac{1}{2}$. Let $u_1, u_2, u_3$ be the neighbors of $v$ in $G$, and let $L_i = V(H(u_i))$. For every $u_i$, $|T^k[L_{u_i}] \oplus E_{u_i}| \leqslant 2$, otherwise we would have $g_k(v) \geqslant \frac{3}{4}$. So by Theorem 3.10, every $T[L_{u_i}]$ is an $h_k$-padding tree, and similarly, $T[L_v]$ is an $h_k$-padding tree. These four subtrees are mutually disjoint. Let $\alpha_i$ be the internal port of $T[L_{u_i}]$ and $\alpha_i'$ be its neighbor outside $T[L_{u_i}]$; similarly let $\alpha$ be the internal port of $T[L_v]$ and $\alpha'$ be its neighbor outside $T[L_v]$. Then, $|\{\alpha_1, \alpha_2, \alpha_3, \alpha\}| = 4$, $|\{\alpha_1', \alpha_2', \alpha_3', \alpha'\}| = 4$, $d_T(u_i, \alpha_i) = h_k$ for every $u_i$, and $d_T(v, \alpha) = h_k$ (otherwise $d_T(v, u_i) \geqslant k+1$ hence $f_k(v) \geqslant \frac{1}{2}$). Therefore, $\mathcal{P}_3(T)[\{\alpha_1, \alpha_2, \alpha_3, \alpha\}] \cong \mathcal{P}_k(T)[\{u_1, u_2, u_3, v\}]$. Consequently, $f_k(v) \geqslant \frac{1}{2}$ by Lemma 4.2, a contradiction.

Now $f_k(v) + g_k(v) \geqslant \frac{1}{2}$ for every degree-3 vertex $v$ of $G$, and moreover by the above argument if $f_k(v) + g_k(v) = \frac{1}{2}$ then $f_k(v) = \frac{1}{2}$ and $g_k(v) = 0$, so together with the upper bound given in Lemma 4.5 we have

(i) $f_k(v) = \frac{1}{2}$ for all degree 3 vertices $v$ of $G$,
(ii) $f_k(v) = 0$ for all other vertices $v$ of $G'$, and
(iii) $g_k(v) = 0$ for all vertices $v$ of $G'$.

By (iii) $T[L_v]$ is an $h_k$-padding tree for every $v \in V$. Let $\alpha_v$ be the internal port of $T[L(H_v)]$. We have shown that $d_T(v, \alpha_v) = h_k$, i.e. $v$ is the external port of $T[L_v]$ for all degree-3 vertices $v$ of $G$, which in fact holds for all vertices $v$ of $G'$. Therefore $\mathcal{P}_3(T[\{\alpha_v : v \in V\}]) \cong \mathcal{P}_k(T)$, and $|T^3[\{\alpha_v : v \in V\}] \oplus \{\{\alpha_u, \alpha_v\} : \{u, v\} \in E\}| \leqslant \ell$. By Lemmas 4.1 and 4.3, $G$ must have a Hamiltonian path, say $v_0, v_1, \ldots, v_{n+1}$, and $T[\{\alpha_v : v \in V\}]$ must be a bridge along $\alpha_{v_0}, \alpha_{v_1}, \ldots, \alpha_{v_n}$. □

**Theorem 4.7.** *For every odd $k \geqslant 3$,* 3CPR$k$ *is NP-complete.*

**Proof.** We have fixed an odd $k \geqslant 5$. Let $G' = (V', E')$ be the graph constructed from $G$ as in the proof of Lemma 4.4. It suffices to show that the input graph $G$ has a Hamiltonian path if and only if $padd_k(G') = (V_k(G'), E_k(G'))$ has an approximate phylogeny with error at most $\ell$.

Suppose that the input graph $G$ has a Hamiltonian path given by an ordering of the vertices $v_0, v_1, v_2, \ldots, v_n, v_{n+1}$ of $G$. For each $v \in V'$, add a copy $R(v)$ of $h_k$-padding tree $R_{k,h_k}$ such that $L(R(v)) = L_v$, $\mathcal{P}_k(R(v)) = H(v)$ and the external port of $H(v)$ is $v$. Then, build the bridge along the internal ports of $R(v_0), R(v_1), \ldots, R(v_{n+1})$, and join one end port of the bridge to the internal port of $R(\tilde{v}_0)$, and the other to that of $R(\tilde{v}_{n+1})$. This construction gives an approximate phylogeny of $padd_k(G')$ with error $\ell$.

Conversely, suppose that $T$ is an approximate phylogeny of $padd_k(G')$ with $|T^k \oplus E_k(G')| \leqslant \ell$. By Lemma 4.5, $\sum_{v \in V'} (f_{G,T[V],k}(v) + g_{padd_k(G'),T,k}) \leqslant |T^k \oplus E_k(G')| \leqslant \ell$, so by Lemma 4.6, $G$ must have a Hamiltonian path. □

### 4.2. The case where $\Delta = 3$ and $k$ is even

This subsection assumes that $k$ is an arbitrary even number greater than or equal to 4, and proves that 3CPR$k$ is NP-complete. Throughout this section all trees and phylogenies are of maximum degree 3 or less.

Fix any even number $k \geqslant 4$. Let $G = (V, E)$ be an arbitrary instance graph of HC. Let $padd_k(G) = (V_k(G), E_k(G))$ be a graph constructed as follows:

- Replace every vertex $v \in V$, with a copy $H(v)$ of the $(3, k, h_k - 1, 2)$-padding graph $\mathcal{P}_k(R_{k,h_k-1})$ and identify $v$ with the external port in $H(v)$.
- For every edge $\{u, v\} \in E$, add an isolated copy $H(u, v)$ of the $\mathcal{P}_k(S_{3,k,h_k+1})$.
- Add four more isolated copies $H_1, \ldots, H_4$ of the $\mathcal{P}_k(S_{3,k,h_k+1})$.

Note that when $k = 4$ the first step of the above construction of $padd_4(G)$ can be skipped, since the $(3, 4, 0, 2)$-padding graph consists from a single vertex.

For each vertex $v \in V$ let $L_v = V(H(u))$, for each edge $\{u, v\} \in E$ let $L_{u,v} = V(H(u, v))$, and for each $1 \leqslant i \leqslant 4$ let $L_i = V(H_i)$.

**Lemma 4.8.** *Let $T$ be an arbitrary tree such that $L(T) = V$. Let $v$ be an arbitrary vertex of $G$ having three pairwise nonadjacent neighbors $u_1, u_2$, and $u_3$. Then $f_{G,T,4}(v) = \frac{1}{2}$ or $f_{G,T,4}(v) \geqslant 1$, and in the former case $d_T(u_i, v) > 4$ for one $u_i \in \{u_1, u_2, u_3\}$ and $d_T(u_j, v) \leqslant 4$ for the other two $u_j \in \{u_1, u_2, u_3\} - \{u_i\}$.*

**Proof.** It suffices to show that if $d_T(u_i, v) \leqslant 4$ for all $u_i$ then $f_{G,T,4}(v) \geqslant 1$. For a contradiction suppose $d_T(u_i, v) \leqslant 4$ for all $u_i$ and $f_{G,T,4}(v) < 1$. Let $v'$ (respectively, $u_i'$) be the internal node adjacent to $v$ (respectively, $u_i$) in $T$. Then $|\{u_1', u_2', u_3', v'\}| = 4$; otherwise either $u_i' = u_j'$ for some $i \neq j \in \{1, 2, 3\}$ or $u_i' = v'$ for some $i \in \{1, 2, 3\}$, but in either case $f_{G,T,4}(v) \geqslant 1$, a contradiction. Hence $d_T(u_i', v') \leqslant 2$ for all $u_i$, implying that $d_T(u_i', u_j') \leqslant 2$ for some $u_i' \neq u_j'$, so $f_{G,T,4}(v) \geqslant 1$, a contradiction. □

Both of Lemmas 4.5 and 4.6 hold as they are for the even $k$ case, too. The proofs are the same, but using Lemma 4.8 here instead of Lemma 4.2 therein. Note that when $k = 0$, the function $g$ in these lemmas becomes unnecessary (it is the zero function).

**Theorem 4.9.** *For every even $k \geqslant 4$,* 3CPR$k$ *is NP-complete.*

**Proof.** It suffices to show that the input graph $G$ has a Hamiltonian path if and only if $padd_k(G)$ has an approximate phylogeny with error at most $3\ell$.

Suppose that the input graph $G$ has a Hamiltonian path given by an ordering of the vertices $v_0, v_1, v_2, \ldots, v_n, v_{n+1}$ of $G$. We say that an edge in $G$ is *covered* (by this Hamiltonian path) if it is $\{v_i, v_{i+1}\}$ for some $0 \leqslant i \leqslant n$; otherwise it is called *uncovered*. The following construction gives an approximate phylogeny of $padd_k(G)$ with error $\ell$:

- For each $v \in V$, add a copy $R(v)$ of the $(3, k, h_k, 2)$-padding tree $\mathcal{P}_k(R_{k,h_k-1})$ such that $L(R(v)) = L_v$, $\mathcal{P}_k(R(v)) = H(v)$ and the external port of $R(v)$ is $v$. Let $\alpha_v$ denote the internal port of $R(v)$.
- For each $\{u, v\} \in E$, add a copy $R(u, v)$ of the tree $\mathcal{P}_k(S_{3,k,h_k+1})$ such that $L(R(u, v)) = L_{u,v}$ and $\mathcal{P}_k(R(u, v)) = H(u, v)$. Let $\alpha_{u,v}$ denote the internal port of $R(u, v)$.
- Build the bridge $T_{\text{cover}}$ along $\alpha_{v_0}, \alpha_{v_0,v_1}, \alpha_{v_1}, \alpha_{v_1,v_2}, \ldots, \alpha_{v_n,v_{n+1}}, \alpha_{v_{n+1}}$.
- Build another bridge $T_{\text{uncover}}$ along a sequence of $\alpha_{u,v}$ of all uncovered edges $\{u, v\}$ of $G$.
- Add four copies $R_1, \ldots, R_4$ of $S_{3,k,h_k+1}$ such that $L(R_i) = V(H_i)$ and $\mathcal{P}_k(R_i) = H_i$. Let $\alpha_i$ denote the internal port of $R_i$.
- Add two new internal nodes $\beta_1$ and $\beta_2$.
- Join $\beta_1$ to one terminal of $T_{\text{cover}}$ and $\alpha_1$. Further, if $T_{\text{uncover}}$ is empty (i.e. there is no uncovered edge) then join $\beta_1$ to $\alpha_2$; otherwise, join $\beta_1$ to one terminal of $T_{\text{uncover}}$, and join $\alpha_2$ to the other terminal of $T_{\text{uncover}}$.
- Join $\beta_2$ to the other terminal of $T_{\text{cover}}$, $\alpha_3$ and $\alpha_4$.

Conversely, suppose that $T$ is an approximate phylogeny of $padd_k(G)$ with $\left|T^k \oplus E_k(G)\right| \leqslant \ell$. By Lemma 4.5, $\sum_{v \in V}(f_{G,T[V],4}(v) + g_{padd_k(G),T,4}(v)) \leqslant |T^k \oplus E_k(G)| \leqslant \ell$, so by Lemma 4.6, $G$ must have a Hamiltonian path. □

### 4.3. Generalization to $\varDelta \geqslant 4$

In this subsection, we generalize the previous lemmas and theorems shown for $\varDelta = 3$ to $\varDelta \geqslant 4$. We show the generalization for only the odd $k$ case, because the proofs for the even $k$ case proceed in parallel. Throughout this section all trees and phylogenies are of maximum degree $\varDelta$ or less.

Fix any odd number $k \geqslant 3$. Let $G = (V, E)$ be an arbitrary instance graph of HC, where let $v_0$ and $v_{n+1}$ denote its degree 1 vertices. Let $padd_k(G) = (V_k(G), E_k(G))$ be the graph constructed from $G$ in Section 4.1. Further, a graph $padd_{\varDelta,k}(G) = (V_{\varDelta,k}(G), E_{\varDelta,k}(G))$ is constructed from $padd_k(G)$ as follows:

- For each $v \in V - \{v_0, v_{n+1}\}$, add $\varDelta - 3$ copies $H_1(v), \ldots, H_{\varDelta-3}(v)$ of the $(\varDelta, k, h_k, 2)$-padding graph $\mathcal{P}_k(R_{\varDelta,k,h_k+1})$, name the external port of $H_i(v)$ as $v(i)$, and join $v$ to every $v(i)$.
- For each $v \in \{v_0, v_{n+1}\}$, do it with $\varDelta - 2$ copies.

Note that the graph $padd_{\varDelta,k}(G)$ contains $padd_k(G) = (V_k(G), E_k(G))$ as the subgraph induced on $V_k(G) = V \cup (\bigcup_{v \in V} V(L_v))$. For every edge $v(i)$ let $L_{v(i)} = V(H_i(v))$. For any tree $T$ such that $L(T) = V_{\varDelta,k}(G)$, we define a function $h_{padd_{\varDelta,k}(G),T,k}$ from every $v \in V$ as follows:

$$h_{padd_{\varDelta,k}(G),T,k}(v) = \sum_i 1_{d_T(v,v(i))>k} + \sum_i |T^k[L_{v(i)}] \oplus E(H(v(i)))|,$$

where $i$ runs over $\{1, \ldots, \varDelta - 3\}$ if $v \in V - \{v_0, v_n\}$, and $\{1, \ldots, \varDelta - 2\}$ if $v \in \{v_0, v_n\}$; $1_{d_T(v,v(i))>k}$ is 1 if $d_T(v, v(i)) > k$ and 0 otherwise.

**Lemma 4.10.** *For any tree $T$ such that $L(T) = V_{\varDelta,k}(G)$, $\sum_{v \in V}(f_{G,T[V],k}(v) + g_{padd_k(G),T[V_k(G)],k}(v) + h_{padd_{\varDelta,k}(G),T,k}(v) \leqslant |T^k \oplus E_k(G)|$.*

**Proof.** By the definitions of $f_{G,T[V],k}$, $g_{padd_k(G),T[V_k(G)],k}$ and $h_{padd_{\varDelta,k}(G),T,k}$. □

**Lemma 4.11.** *For any tree $T$ such that $L(T) = V_{\varDelta,k}(G)$, if $\sum_{v \in V}(f_{G,T[V],k}(v) + g_{padd_k(G),T[V_k(G)],k}(v) + h_{padd_{\varDelta,k}(G),T,k}(v)) \leqslant \ell$, then $G$ must have a Hamiltonian path.*

**Proof.** Let us abbreviate $f_{G,T[V],k}(v)$, $g_{padd_k(G),T[V_k(G)],k}(v)$, and $h_{padd_k(G),T,k}$ as $f_k$, $g_k$, and $h_{\varDelta,k}$, respectively. We claim that $f_k(v) + g_k(v) + h_{\varDelta,k}(v) \geqslant \frac{1}{2}$ for every degree 3 vertex of $G$. For a contradiction suppose $f_k(v) + g_k(v) + h_{\varDelta,k}(v) < \frac{1}{2}$. Since $h_{\varDelta,k}(v) < \frac{1}{2}$, by Theorem 3.20, all $T[L_{v(i)}]$ are $(h_k + 1)$-padding trees, and further since $g_k(v) < \frac{1}{2}$, by Theorem 3.18, $T[L_v]$ is a $h_k$-padding tree. Let $\alpha_v$ (respectively, $\alpha_{v(i)}$) be the internal port of $T[L_v]$ (respectively, $T[L_{v(i)}]$) and $\alpha'_v$ (respectively, $\alpha'_{v(i)}$) be its neighbor outside $T[L_v]$ (respectively, $T[L_{v(i)}]$). Then, $\alpha'_{v(i)} = \alpha'_v$ for all

$v(i)$; otherwise $d_T(v, v(i)) > k$ so $h_{\Delta,k}(v) \geqslant 1$, a contradiction. Consequently, $\alpha'_v$ is adjacent to all $\Delta - 3$ distinct nodes $\alpha_{v(i)}$, so it can take at most three more neighbors other than these. Then by Lemma 4.2, $f_k(v) \geqslant \frac{1}{2}$, a contradiction.

Now, together with the upper bound, we have $f_k(v) + g_k(v) + h_{\Delta,k}(v) \geqslant \frac{1}{2}$ for every degree 3 vertex $v$ of $G$, and if the equality holds then $f_k(v) = \frac{1}{2}$, $g_k(v) = 0$ and $h_{\Delta,k}(v) = 0$. We thus have

(i) $f_k(v) = \frac{1}{2}$ for all degree 3 vertices $v$ of $G$,
(ii) $f_k(v) = 0$ for all other vertices $v$ of $G$, and
(iii) $g_k(v) = h_{\Delta,k}(v) = 0$ for all vertices of $G$.

By (iii) all $T[L_v]$ are $h_k$-padding trees and all $T[L_{v(i)}]$ are $(h_k + 1)$-padding trees, consequently $\alpha'_v = \alpha'_{v(i)}$ for every $v$ and $v(i)$, and the number of neighbors of $\alpha'_v$ other than $\alpha_{v(i)}$ is at most 3. Hence by Lemma 4.6 we have a Hamiltonian path of $G$. □

**Theorem 4.12.** *For every fixed constant $\Delta \geqslant 3$ and every odd $k \geqslant 3$, $\Delta$CPR$k$ is NP-complete.*

**Proof.** We have shown it for $\Delta = 3$, so fix an arbitrary $\Delta \geqslant 4$.

Suppose that the input graph $G$ has a Hamiltonian path given by an ordering of the vertices $v_0, v_1, v_2, \ldots, v_n, v_{n+1}$ of $G$. Build the tree for the graph $G$ as in the proof of Theorem 4.7, but replacing the $(3, k, h_k, 2)$-padding tree $R_{3,k,h_k}$ therein with the $(\Delta, k, h_k, 2)$-padding tree $R_{\Delta,k,h_k}$ here. Recall that there we have built the bridge along the internal ports of $R(v_0), R(v_1), \ldots, R(v_{n+1})$; let $v'$ be the node of the bridge adjacent to the internal port of $R(v)$; thus, $v'_0$ and $v'_{n+1}$ are the two terminals of the bridge. Further, for each $v \in V$ and each $1 \leqslant i \leqslant \Delta - 3$, add a copy $R(v(i))$ of the $(\Delta, k, h_k + 1, 2)$-padding tree $R_{\Delta,k,h_k+1}$ such that $L(R(v(i))) = L_{v(i)}$, $\mathcal{P}_k(R(v)) = H(v(i))$ and the external port of $H(v(i))$ is $v(i)$. Join the internal port of each $R(v(i))$ to $v'$. Further, for each $v \in \{v_0, v_{n+1}\}$, add one more copy $R(v(\Delta + 2))$ and join its internal port to $v'$. This construction gives an approximate phylogeny of $padd_{\Delta,k}(G)$ with error $\ell$.

Conversely, suppose that $T$ is an approximate phylogeny of $padd_{\Delta,k}(G)$ with error at most $\ell$. By Lemma 4.10, $\sum_{v \in V}(f_{G,T[V],k}(v) + g_{padd_k(G),T[V_k(G)],k}(v) + h_{padd_{\Delta,k}(G),T,k}(v)) \leqslant |T^k \oplus E_{\Delta,k}(G)| \leqslant \ell$, so by Lemma 4.11, $G$ must have a Hamiltonian path. □

Generalization of Theorem 4.9 to any fixed degree $\Delta \geqslant 3$ proceeds in parallel and is thus omitted. We have proved the following theorem:

**Theorem 4.13.** *For every $\Delta \geqslant 3$ and every $k \geqslant 3$, $\Delta$CPR$k$ is NP-complete.*

## 5. Summary and an open question

We have proved that $\Delta$CPR$k$ is NP-complete for all fixed constants $k \geqslant 3$ and $\Delta \geqslant 3$. A more fundamental problem is the TREE *k*TH ROOT PROBLEM (TR$k$), where the nodes (not only the leaves) of $T$ correspond to the vertices of $G$. Kearney and Corneil proved that CTR$k$ is NP-complete when $k \geqslant 3$ [10]. We conjecture that $\Delta$CTR$k$ is NP-complete for every fixed $\Delta \geqslant 3$ and $k \geqslant 2$.

## References

[1] N. Bansal, A. Blum, S. Chawla, Correlation clustering, in: Proc. 43rd Symp. on Foundations of Computer Science (FOCS 2002), 2002, pp. 238–250.

[3] Z.-Z. Chen, T. Jiang, G.-H. Lin, Computing phylogenetic roots with bounded degrees and errors, SIAM J. Comput. 32 (4) (2003) 864–879, A preliminary version appeared in Proc. of WADS2001.

[4] Z.-Z. Chen, T. Tsukiji, Computing bounded-degree phylogenetic roots of disconnected graphs, in: Proc. 30th Internat. Workshop on Graph-Theoretic Concepts in Computer Science (WG2004), Lecture Notes in Computer Science, Vol. 3353, Springer, Berlin, 2004, pp. 308–319, A revised version is going to appear in J. Algorithms..

[5] M. Dom, J. Guo, F. Hüffner, R. Niedermeier, Error compensation in leaf root problems, in: Proc. 15th Internat. Symp. on Algorithms and Computation (ISSAC2004), Lecture Notes in Computer Science, Vol. 3341, Springer, Berlin, 2004, pp. 389–401, A revised version is going to appear in Algorithmica..

[7] M.R. Garey, D.S. Johnson, R.E. Tarjan, The planar Hamiltonian circuit problem is NP-complete, SIAM J. Comput. 5 (4) (1976) 704–714.

[10] P.E. Kearney, D.G. Corneil, Tree powers, J. Algorithms 29 (1998) 111–131.

[11] D.G. Kirkpatrick, P. Hell, On the complexity of a generalized matching problem, in: The 10th Annu. ACM Symp. on Theory of Computing (STOC 1978), 1978, pp. 240–245.

[12] G.-H. Lin, P.E. Kearney, T. Jiang, Phylogenetic *k*-root and Steiner *k*-root, in: The 11th Annu. Internat. Symp. on Algorithms and Computation (ISAAC 2000), Lecture Notes in Computer Science, Vol. 1969, Springer, Berlin, 2000, pp. 539–551.

[15] N. Nishimura, P. Ragde, D.M. Thilikos, On graph powers for leaf-labeled trees, in: Proc. Seventh Scandinavian Workshop on Algorithm Theory (SWAT 2000), Lecture Notes in Computer Science, Vol. 1851, 2000, pp. 125–138.

[16] C.H. Papadimitriou, Computational Complexity, Addison-Wesley, Reading, MA, 1994.

[17] D.L. Swofford, G.J. Olsen, P.J. Waddell, D.M. Hillis, Phylogenetic inference, in: D.M. Hillis, C. Moritz, B.K. Mable (Eds.), Molecular Systematics, second ed., Sinauer Associates, Sunderland, MA, 1996, pp. 407–514.

## Further reading

[2] A. Brandstädt, V.B. Le, J.P. Spinrad, Graph classes: a survey, SIAM Monographs on Discrete Mathematics and Applications, SIAM, Philadelphia, 1999.

[6] M.R. Garey, D.S. Johnson, Computers and Intractability (A Guide to Theory of NP-Completeness), Freeman, New York, 1979.

[8] T. Jiang, G.H. Lin, J. Xu, On the closest tree *k*th root problem, Manuscript, Department of Computer Science, University of Waterloo, November 2000.

[9] R.M. Karp, Reducibility among combinatorial problems, in: R.E. Miller, J.W. Thatcher (Eds.), Complexity of Computer Computations, Plenum Press, New York, 1972, pp. 85–103.

[13] Y.-L. Lin, S.S. Skiena, Algorithms for square roots of graphs, SIAM J. Discrete Math. 8 (1995) 99–118.

[14] R. Motwani, M. Sudan, Computing roots of graphs is hard, Discrete Appl. Math. 54 (1994) 81–88.

[18] L.G. Valiant, The complexity of computing the permanent, Theoret. Comput. Sci. 8 (2) (1979) 189–201.