Computing Phylogenetic Roots with Bounded Degrees and Errors is Hard (Evolutionary Advancement in Fundamental Theories of Computer Science)

Tsukiji, Tatsuie; Chen, Zhi-Zhong

数理解析研究所講究録 (2004), 1375: 42-48

2004-05

http://hdl.handle.net/2433/25574

Departmental Bulletin Paper

Kyoto University
Computing Phylogenetic Roots with Bounded Degrees and Errors is Hard

Tsukiji Tatsuie\textsuperscript{1} and Zhi-Zhong Chen\textsuperscript{2}

\textsuperscript{1}篠地 立家
Department of Information Science, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan.
tsukiji@.dendai.ac.jp

\textsuperscript{2}陳 賢中
Department of Mathematical Sciences, Tokyo Denki University, Hatoyama, Saitama 350-0394, Japan. chen@.dendai.ac.jp

Abstract. The Degree-$\Delta$ Closest Phylogenetic $k$th Root Problem ($\Delta$CPR$k$) is the problem of finding a (phylogenetic) tree $T$ from a given graph $G = (V,E)$ such that (1) the degree of each internal node of $T$ is at least 3 and at most $\Delta$, (2) the external nodes (i.e., leaves) of $T$ are exactly the elements of $V$, (3) The number of disagreements, $|E \oplus \{(u,v): u,v \text{ are leaves of } T \text{ and } d_T(u,v) \leq k\}|$ does not exceed a given number, where $d_T(u,v)$ denotes the distance between $u$ and $v$ in tree $T$. We show that this problem is NP-hard for every fixed constants $\Delta, k \geq 3$.

Our major technical contribution is the determination of all phylogenetic roots that approximate the almost largest cliques. In more precise, let $f_\Delta(k)$ be the size of a largest clique having a $k$th phylogenetic root with maximum degree $\Delta$. We determine the all phylogenetic $k$th roots with maximum degree $\Delta$ that approximate the $(f_\Delta(k) - 1)$-clique within error 2, where we allow the internal nodes of phylogeny to have degree 2.

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. 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 vertices of the graph (i.e. species) and for any two vertices of the graph, they are adjacent in the graph 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 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 [5]:

**Phylogenetic $k$th Root Problem (PR$k$):**
Given a graph $G = (V,E)$, find a phylogeny $T$ with 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 $d_T(u,v) \leq k$. 
Such a phylogeny $T$ (if exists) is called a **phylogenetic $k$th root**, or a **$k$th root phylogeny**, of graph $G$. 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 $P_k(T)$. That is, $P_k(T)$ has the vertex set $L(T) = \{ u : u \text{ are leaves of } T \}$ and the edge set $T^k = \{ (u, v) \mid u \text{ and } v \text{ are leaves of } T \text{ and } d_T(u, v) \leq k \}$. Therefore, $PR_k$ asks for a phylogeny $T$ such that $G = P_k(T)$.

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 $V$ is called an approximate phylogeny of $G$, and the **error** of $T$ is $|T^k \cup E| = |(E - T^k) \cup (T^k - E)|$. 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 phylogeny $T$ with at most $\ell$ errors.

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

The hardness of $PR_k$ for large $k$ seems to come from the unbounded degree of an internal node in the output phylogeny. On the other hand, in the practice of phylogeny reconstruction, most phylogenies considered are trees of degree 3 [7] because speciation events are usually bifurcating events in the evolutionary process. We call these restricted versions the **Degree-\Delta PR$k$** and the **Degree-\Delta CPR$k$**, 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 [5] where linear-time algorithms for $PR_2$ and $PR_3$ were proposed. A linear-time algorithm for the special case of $PR_4$ where the input graph is required to be connected, was also presented in [5]. At present, the complexity of $PR_k$ with $k \geq 5$ is still unknown.

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

Of special interest is CPR2. CPR2 is essentially identical to the correlation clustering problem which has drawn much attention recently [1]. The proof of the NP-hardness of CPR2 given in [2] is also a valid proof of the NP-hardness of the correlation clustering problem.

### 1.2 Our Contribution

In this paper, we will show that $\Delta CPR_k$ is NP-complete, for every $k \geq 3$ and $\Delta \geq 3$. This answers an open question in [2].

In a course of the proof we first reduce Hamiltonian Path, a famous NP-complete problem, to $\Delta CPR_3$, and then $\Delta CPR_3$ to $\Delta CPR_k$. The former reduction is tedious but a routine work. On the other hand, the latter reduction seems to require new combinatorial investigation that is proper of $\Delta CPR_k$.

A $(\Delta, k, h, \ell)$-core graph is a graph $G = (V, E)$ with the following properties:
There is a tree $T$ of maximum degree $\Delta$ whose phylogenetic $k$th power is $G$ and such that $T$ has a unique unsaturated (i.e., degree $< \Delta$) internal node $\alpha$, the degree of $\alpha$ is $\Delta - 1$, $d_T(\alpha, u) = h$ holds for one leaf $u$ and $d_T(\alpha, v) \geq h + 1$ for all leaves $v$ other than $u$.

For every approximate phylogeny $T$ of $G$ with maximum degree $\Delta$ and at most $\ell$ errors, there is at most one pair $(\alpha, u)$ such that $\alpha$ is an unsaturated internal node of $T$, $u$ is a leaf of $T$, and $d_T(\alpha, u) \leq h$; moreover, if $(\alpha, u)$ exists then the degree of $\alpha$ in $T$ is $\Delta - 1$.

Then, we establish the reduction from $\triangle CPR$ to $\triangle CPR_k$ by providing a family of $(\Delta, k, [k/2] - 1, 2)$-core graphs for every fixed $\Delta \geq 3$ and $k \geq 4$. Our construction of a $(\Delta, k, [k/2] - 1, 2)$-core graph is a pile of $(\Delta, k', 1, 2)$-core graphs for $k' = 5, 7, \ldots$, $k$ if $k \geq 5$ is odd, and $k' = 4, 6, \ldots$, $k$ if $k \geq 4$ is even. So, a more basic problem is to prove that a certain graph is a $(\Delta, k, 1, 2)$-core graph.

The maximum size of a clique having a (no-error) $k$th root phylogeny with maximum degree $\Delta$ is given by the following function,

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

We prove that the clique of size $f_\Delta(k) - 1$ is a $(\Delta, k, 1, 2)$-core graph. Moreover, we determine the all $k$th root phylogenies with maximum degree $\Delta$ that approximate the clique within error 2, where we allow the internal nodes of phylogeny to have degree 2. For example, all phylogenetic roots of the $(f_3(5) - 1)$-clique are $D_5$ in Figure 1, $E_5$ in Figure 2, and the tree obtained from $D_5$ by removing the leaf $u$.

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. An induced subgraph of a graph $G$ is the subgraph $H$ induced by a subset $W$ of $V(G)$, i.e., $E(H) = \{(u, v) : u, v \in W \text{ and } (u, v) \in E(G)\}$.

Two graphs $G = (V, E)$ and $G' = (V', E')$ are isomorphic if there is a one-to-one correspondence $\phi$ between $V$ and $V'$ such that $(u, v) \in E$ if and only if $(\phi(u), \phi(v)) \in E'$, and we denote it as $G \cong_{\phi} G'$. The distance between two vertices $u$ and $v$ in $G$ is denoted by $d_G(u, v)$. The degree of a vertex $v$ in $G$ is denoted by $d_G(v)$, which is the number of vertices adjacent to $v$ in $G$. 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 some new terminologies of trees for convenience. For a tree $T$ of maximum degree $\Delta$, an internal node $\alpha$ of $T$ is unsaturated if $d_T(\alpha) \leq \Delta - 1$. 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 $T$. A tree $T$ is $h$-away if for each unsaturated internal node $x$ of $T$, the minimum distance from $x$ to a leaf is at least $h$ and further there is exactly one leaf $v_x$ such that $d_T(x, v_x) = h$. For any set $U$ of nodes of $T$, $T[U]$ denotes the minimum subtree containing $U$. Note that each leaf of $T[U]$ belongs to $U$. A phylogeny is a tree that contains no degree-2 nodes. As already mentioned, the $k$th phylogenetic power of any tree $T$ is denoted as $P_k(T) = (L(T), T'')$, $T''$ is the set of edges $(u, v)$ with $\{u, v\} \subseteq L(T)$ and $d_T(u, v) \leq k$.

3 Construction of $(\Delta, k, [k/2] - 1, 2)$-core graphs

In this section we give a construction of $(3, k, [k/2] - 1, 2)$-core graphs for every odd $k \geq 5$. It is straightforward to generalize the arguments of this section to obtain $(\Delta, k, [k/2] - 1, 2)$-core graphs for every $\Delta \geq 3$ and $k \geq 4$. 

Throughout this section, all trees and phylogenies are of maximum degree 3 or less. We abbreviate $f_3(k)$ as $f(k)$. The proofs of the most lemmas and corollaries are omitted due to lack of space.

A phylogeny whose $k$th phylogenetic power realizes the $f(k)$-clique can be constructed as follows: Start with a path $P$ of length exactly $k$. Let $u$ and $v$ be the endpoints of $P$. Then connect as many new nodes as possible so that $P$ becomes a tree of degree 3 and every node has distance at most $k$ from both $u$ and $v$. This tree is unique up to isomorphism and hence we denote it by $C_k$. Moreover, removing an arbitrary leaf from $C_k$ yields a tree, which is unique up to isomorphism. We denote this tree by $D_k$. Figure 1 depicts $D_4, D_5$, and $D_6$ where the missing sibling leaf of $u$ has been removed. By definition, the $k$th phylogenetic power of $D_k$ is an $f(k) - 1$ clique.

![Figure 1. $D_4, D_5$ and $D_6$.](image)

**Lemma 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 $C_k$.

**Corollary 1.** For any tree $T$, if $d_T(u, v) \leq k$ for all leaves $u$ and $v$, then $T$ is isomorphic to a subtree of $C_k$.

**Fact 1** For every tree $T$ with $|L(T)| = f(k) - 1$ and $|T^k| \geq \left(\frac{f(k) - 1}{2}\right) - 2$, we have $d_T(u, v) \leq k$ for all but at most two unordered pairs $(u, v)$ of leaves of $T$.

**Lemma 2.** Let $k \geq 4$. Let $T$ be an arbitrary tree such that $|L(T)| = f(k) - 1$ and $|T^k| \geq \left(\frac{f(k) - 1}{2}\right) - 2$. Then, there are leaves $u$ and $v$ of $T$ with $d_T(u, v) = k$.

**Lemma 3.** Let $k \geq 6$. Let $T$ be an arbitrary tree having $f(k) - 1$ leaves. Suppose that there are leaves $u, v, w$ of $T$ such that $d_T(u, v) = k$ and $\max(d_T(u, w), d_T(v, w)) \geq k + 1$. Then, $|T^k| \leq \left(\frac{f(k) - 1}{2}\right) - 3$.

**Lemma 4.** Let $k \geq 6$. Let $T$ be a tree having $f(k) - 1$ leaves such that $|T^k| \geq \left(\frac{f(k) - 1}{2}\right) - 2$. Then $T$ is 0-extendable or 1-extendable. Moreover, if $T$ is 1-extendable then $T \cong D_k$.

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

**Lemma 5.** Let $k \in \{4, 5\}$. Let $T$ be a tree having $f(k) - 1$ leaves such that $|T^k| \geq \left(\frac{f(k) - 1}{2}\right) - 2$. Then $T$ is 0-extendable or 1-extendable. Moreover, if $T$ is 1-extendable then $T \cong D_k$ or $T \cong E_k$. 
Theorem 1. For every $k \geq 4$, the $(f(k) - 1)$-clique is a $(3, k, 1, 2)$-core graph.

Now we are ready to construct a $(3, k, \lfloor k/2 \rfloor - 1, 2)$-core graph for every odd $k \geq 5$. We recursively construct trees $R_{k, \lfloor k/2 \rfloor - 1}$, $k=5, 7, 9, \ldots$, and define a family of $(3, k, \lfloor k/2 \rfloor - 1, 2)$-core graphs as the $k$th phylogenetic power of the trees $R_{k, \lfloor k/2 \rfloor - 1}$ (see Figure 3 for $R_{7,2}$):

- Let $h_k = \lfloor k/2 \rfloor - 1$. For each $1 \leq i \leq h_k$, let $g(i) = \prod_{j=1}^{i} (f(2j+3) - 1)$. Let $g(0) = 1$.
- $\tilde{R}_{k,h_k}$ is a leveled tree of depth $h_k$ such that at depth $i$ ($0 \leq i \leq h_k$) are placed $g(i)$ nodes and each node $v$ at depth $i < h_k$ is connected to some $f(2i + 5) - 1$ nodes at depth $i + 1$.
- $R_{k,h_k}$ is an expansion of $\tilde{R}_{k,h_k}$ such that each node $v$ of $\tilde{R}_{k,h_k}$ at depth $i$ ($0 \leq i \leq h_k - 1$) is expanded to a copy $D(v)$ of $D_{2i+5}$ in $R_{k,h_k}$, where $v$ is identified with the degree-2 node of $D(v)$ and the child nodes of $v$ in $\tilde{R}_{k,h_k}$ are identified with the leaves of $D(v)$ in an arbitrary one-to-one manner.

By construction, $R_{k,h_k}$ is 1-extensible and $h_k$-away, and has a unique degree-2 node, namely, the unique degree-2 node of $D_5$.

Lemma 6. Let $k \geq 4$. Let $T$ be a tree having $f(k) - 1$ leaves such that $|T_k| \geq \binom{f(k)-1}{2} - 2$. Suppose further that $T$ is not 0-extensible. Let $T'(w)$ be the tree obtained by connecting a new leaf to an arbitrary leaf $w$ of $T$. Then, $|T'(w)|^k \leq \binom{f(k)-1}{2} - 3$.

Lemma 7. Let $k \geq 5$ be odd. Let $T$ be a tree such that $L(T) = L(R_{k,h_k})$ and $|T^k \oplus R_{k,h_k}^k| \leq 2$. Then, $T$ is 0-extensible or 1-extensible. Moreover, if $T$ is 1-extensible then it is $h_k$-away.
Theorem 2. For every odd \( k \geq 5 \), the graph \( P_k(R_{k,\lfloor k/2 \rfloor - 1}) \) is a \((3, k, \lfloor k/2 \rfloor - 1, 2)\)-core graph.

These constructions, lemmas and theorems can be generalized to every fixed \( k \geq 4 \) and \( \Delta \geq 3 \). A phylogenetic \( k \)th root of the \( f_{\Delta}(k) \)-clique can be constructed in the same way as \( D_i \) and we denote it as \( D_{\Delta,i} \). We can construct a phylogeny \( R_{\Delta,k,h_k} \) of degree \( \Delta \) recursively in the same way as \( R_{k,h_k} \) but replacing \( f \) and \( D_i \) therein with \( f_{\Delta} \) and \( D_{\Delta,i} \), respectively; further, if \( k \) is even then the function \( g(i) \) therein should be replaced by \( \prod_{j=1}^{i} (f_{\Delta}(2j + 2) - 1) \). Lemma 7 and Theorem 2 can be generalized as follows:

Lemma 8. Let \( k \geq 4 \) and \( \Delta \geq 3 \). Let \( T \) be a tree of maximum degree \( \Delta \) such that \( L(T) = L(R_{\Delta,k,h_k}) \) and \( |T^k \oplus R_{\Delta,k,h_k}^k| \leq 2 \). Then \( T \) is 0-extensible or 1-extensible. Moreover, if \( T \) is 1-extensible then it is \( h_k \)-away.

Theorem 3. For every \( k \geq 4 \), \( P_k(R_{\Delta,k,\lfloor k/2 \rfloor - 1}) \) is a \((\Delta, k, \lfloor k/2 \rfloor - 1, 2)\)-core graph.

4 The NP-hardness of \( \Delta\text{CPR}k \)

This section proves that \( 3\text{CPR}k \) is NP-hard for each odd \( k \geq 3 \). It is straightforward to generalize the arguments of this section to prove that \( \Delta\text{CPR}k \) is NP-hard for every \( \Delta \geq 3 \) and \( k \geq 3 \).

Throughout this section, all trees and phylogenies are of maximum degree 3 or less. Proofs of most lemmas and corollaries are omitted due to lack of space.

We begin with the NP-hardness proof of \( 3\text{CPR}3 \) because the NP-hardness proofs of \( 3\text{CPR}k \) for larger odd \( k \) are reductions from it. We reduce the following version of HAMILTONIAN PATH PROBLEM (HP) to \( 3\text{CPR}3 \), whose NP-hardness proofs can be found in [3] and [6, Section 9.3].

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

- all vertices are of degree 3 or less,
- two specific vertices are of degree 1 and each of them is 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 \( G \).

Let \( G = (V, E) \) be an arbitrary instance of HP, hence the maximum degree of \( G \) is 3 and \( G \) contains no cycle of length less than 5. Let \( T = (V, E(T)) \) be an approximate phylogeny of \( G \). We define a fractional value \( \text{cost}_5(v) \) associated with each vertex \( v \in V \) as follows:

\[
\text{cost}_5(v) = \frac{1}{2} \left| \left\{ u : (u, v) \in E \text{ and } d_T(u, v) > 3 \right\} \right| + \left| \left\{ (u, w) : u \neq v, (u, v) \in E, (v, w) \in E, (u, w) \notin E \text{ and } d_T(u, w) \leq 3 \right\} \right|.
\]

Lemma 9. The sum of \( \text{cost}_5(v) \) over all vertices \( v \in V \) is no more than \( |T^3 \oplus E| \).

Lemma 10. Let \( v \) be a vertex of \( G \) having three pairwise nonadjacent neighbors \( u_1, u_2 \) and \( u_3 \). Then, \( \text{cost}_5(v) = \frac{1}{2} \) or \( \text{cost}_5(v) \geq 1 \). Moreover, if \( \text{cost}_5(v) = \frac{1}{2} \), then \( d_T(u_i, v) > 3 \) for one \( u_i \in \{u_1, u_2, u_3\} \) and \( d_T(u_j, v) = 3 \) for the other two \( u_j \in \{u_1, u_2, u_3\} \).
Theorem 5. For every odd $k \geq 3$, 3CPR$^k$ is NP-complete.

It is straightforward to generalize Theorem 5 to every $\Delta \geq 3$ and $k \geq 4$, obtaining the following theorem.

Theorem 6. For every $\Delta \geq 3$ and every $k \geq 3$, $\Delta$CPR$^k$ is NP-complete.

5 Summary and an Open Question

We have proved that $\Delta$CPR$^k$ is NP-complete for every $\Delta \geq 3$ and $k \geq 3$. 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 TR$k$ is NP-complete when $k \geq 3$ [4]. We conjecture that $\Delta$TR$k$ is NP-complete for every fixed $\Delta \geq 3$ and $k \geq 2$.

References