Characterization of the Distance between Subtrees of a Tree by the Associated Tight Span

Hiroshi HIRAI
Research Institute for Mathematical Sciences,
Kyoto University, Kyoto 606-8502, Japan
hirai@kurims.kyoto-u.ac.jp

June 2004

Abstract

A characterization is given to the distance between subtrees of a tree defined as the shortest path length between subtrees. This is a generalization of the four-point condition for tree metrics. For this, we use the theory of the tight span and obtain an extension of the famous result by A. Dress that a metric is a tree metric if and only if its tight span is a tree.

1 Introduction

Recently, mathematical treatments of phylogenetics have come to be increasingly important; see [2],[11]. The central problem in phylogenetics is reconstructing phylogenetic trees from given experimental data, e.g., DNA sequences. If the data is given as a distance matrix expressing dissimilarity between species, the problem is to search for a *tree metric* that "fits" the given distance matrix.

For a finite set X and a distance $d : X \times X \to \mathbf{R}$ with $d(x,x) = 0$ and $d(x,y) = d(y,x) \geq 0$ for $x, y \in X$, d is said to be a metric if it satisfies the triangle inequality, and a tree metric if there exists some weighted tree such that d can be expressed by the path metric between vertices of the tree. One of the most fundamental theorems in phylogenetics is the characterization of tree metrics.

Theorem 1.1 ([14][12][3][4]). *A metric d is a tree metric if and only if it satisfies the four-point condition*

$$\begin{aligned} &\forall x, y, z, w \in X,\ |\{x, y, z, w\}| = 4, \\ &d(x,y) + d(z,w) \leq \max\{d(x,z) + d(y,w), d(x,w) + d(y,z)\}. \end{aligned} \tag{1.1}$$

In this paper, we generalize this characterization for the distance between subtrees of a tree. We define the distance on subtrees of a tree by the shortest path length between subtrees (see Figure 1).

Our main result is as follows:

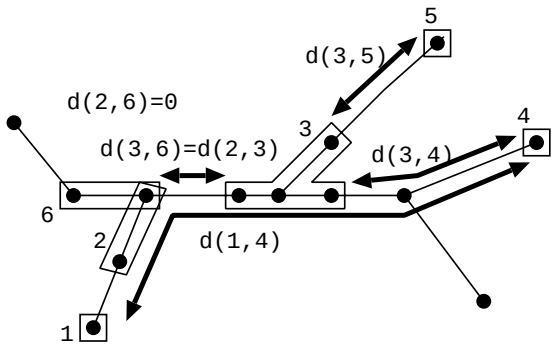

Figure 1: Shortest path lengths between six subtrees of a tree

Theorem 1.2. *A distance d can be expressed as the distance between subtrees of some tree if and only if it satisfies*

$$\begin{aligned}
&\forall x, y, z, w \in X,\ |\{x, y, z, w\}| = 4,\\
&d(x, y) + d(z, w) \leq\\
&\max\left\{\begin{array}{ll} d(x, z) + d(y, w),\ d(x, w) + d(y, z),\ d(x, y),\ d(z, w), & \\ \dfrac{d(x, y) + d(y, z) + d(z, x)}{2}, & \dfrac{d(x, y) + d(y, w) + d(w, x)}{2}, \\ \dfrac{d(x, z) + d(z, w) + d(w, x)}{2}, & \dfrac{d(y, z) + d(z, w) + d(w, y)}{2} \end{array}\right\} \quad (1.2)
\end{aligned}$$

If d satisfies the triangle inequality, then it can be verified that (1.2) coincides the four-point condition (1.1) (see Remark 2.5). Hence (1.2) is a generalization of the four-point condition.

For the proof of Theorem 1.2, we use the theory of *tight span*, which was discovered independently by J. R. Isbell [10], A. Dress [6] and M. Chrobak and L.L. Larmore [5] and developed by A. Dress and coworkers [8]. Whereas tight span has so far been considered essentially for a metric, we need to consider the tight span for a distance that may violate the triangle inequality.

This paper is organized as follows. In Section 2, we prepare definitions and notation, and present a more general version of Theorem 1.2. In Section 3, we give the proof of the theorems.

2 Definitions, Notation and Results

2.1 Distances and partial splits

Let X be a finite set. A function $d : X \times X \to \mathbf{R}$ is said to be a *distance* on X if d satisfies $d(x, x) = 0$ and $d(x, y) = d(y, x) \geq 0$ for $x, y \in X$. A distance d is said to be a *metric* if, in addition, d satisfies $d(x, y) \leq d(x, z) + d(y, z)$ for $x, y, z \in X$. For $A, B \subseteq X$ with $A \cap B = \emptyset$ and $A, B \neq \emptyset$, an unordered pair $\{A, B\}$ is called a *partial split* on X. If a partial split $\{A, B\}$ satisfies $A \cup B = X$, then $\{A, B\}$ is called a *split* on X. For a partial split $\{A, B\}$ on X, we define a *partial split*

distance $\zeta_{\{A,B\}} : X \times X \to \mathbf{R}$ by

$$\zeta_{\{A,B\}}(x,y) = \begin{cases} 1 & \text{if } x \in A,\ y \in B \text{ or } y \in A,\ x \in B \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Note that $\zeta_{\{A,B\}}$ is not a metric if $A \cup B \neq X$ and is a metric, called a *split metric*, if $A \cup B = X$. A pair of partial splits $\{A,B\}$ and $\{C,D\}$ on X is said to be *compatible* if it satisfies one of the following four conditions:

$$A \subseteq C \text{ and } B \supseteq D, \tag{2.2}$$
$$A \subseteq D \text{ and } B \supseteq C, \tag{2.3}$$
$$A \supseteq C \text{ and } B \subseteq D, \tag{2.4}$$
$$A \supseteq D \text{ and } B \subseteq C. \tag{2.5}$$

A collection of partial splits $\mathcal{S}$ is said to be *compatible* if any pair of partial splits in $\mathcal{S}$ is compatible. Note that if $\mathcal{S}$ consists of splits, then compatibility in our sense coincides with compatibility of splits in the standard definition; see [3], [2], [11].

2.2 Graphs

For a weighted graph $G = (V, E, w)$ with a vertex set V, an edge set E, and a positive weight $w : E \to \mathbf{R}$ representing edge lengths, $D_G : V \times V \to \mathbf{R}$ denotes the path metric on G defined by the shortest length of a path. We also denote vertices of G by $V(G)$ and edges of G by $E(G)$.

2.3 Tight span of distances

Next we introduce tight span and related concepts. For a distance $d : X \times X \to \mathbf{R}$, a polyhedron $P(X,d) \subseteq \mathbf{R}^X$ associated with d is defined as

$$P(X,d) = \{f \in \mathbf{R}^X \mid f(x) + f(y) \geq d(x,y)\ (x,y \in X)\}. \tag{2.6}$$

The tight span $T(X,d)$ is defined to be the union of bounded faces of $P(X,d)$, or equivalently,

$$T(X,d) = \{f \in \mathbf{R}^X \mid \forall x \in X,\ f(x) = \max_{y \in X}\{d(x,y) - f(y)\}\}. \tag{2.7}$$

The dimension of $T(X,d)$ is defined to be the maximum dimension of bounded faces of $P(X,d)$. As indicated by [6, Remark 5.4, p.370], $\dim T(X,d)$ can be characterized as follows, whether d is a metric or not.

Theorem 2.1 ([6]). *For a distance $d : X \times X \to \mathbf{R}$ and a positive interger n, the following two conditions are equivalent.*

(*a*) $\dim T(X,d) \geq n$.

(*b*) *There exists $2n$ set $\{x_1, x_{-1}, x_2, x_{-2}, \ldots, x_n, x_{-n}\} \subseteq X$ such that*

$$\sum_{i \in I} d(x_i, x_{-i}) > \sum_{i \in I} d(x_i, x_{\sigma(i)}) \tag{2.8}$$

holds for any permutation σ of $I = \{\pm 1, \pm 2, \ldots, \pm n\}$ not satisfying $\sigma(i) = -i$ for all $i \in I$.

In the appendix, we give a simple proof of Theorem 2.1 based on the standard arguments in linear programming.

Let $t^d : X \to 2^{T(X,d)}$ be defined as

$$t^d(x) = T(X,d) \cap \{f \in \mathbf{R}^X \mid f(x) = 0\} \quad (x \in X), \tag{2.9}$$

which is also the union of the bounded faces of

$$\{f \in \mathbf{R}^X \mid f(y) + f(z) \geq d(y,z) \ (y,z \in X),\ f(x) = 0\}, \tag{2.10}$$

Note that both $T(X,d)$ and $t^d(x)$ are contractible.

We define a weighted graph $G(d)$ by the 1-skeleton of $T(X,d)$ endowed with $\|\cdot\|_\infty$ norm of $\mathbf{R}^X$. For $x \in X$, let $g^d(x)$ be defined by the graph corresponding the 1-skeleton of $t^d(x)$, which is a connected subgraph of $G(d)$.

The following shows that in the case that d is a metric, $t^d(x)$ is a single point of $T(X,d)$ that coincides with the canonical map $X \to T(X,d)$.

Lemma 2.2. *If d is a metric, then we have $t^d(x) = \{h_x\}$ for $x \in X$, where $h_x \in \mathbf{R}^X$ is defined as*

$$h_x(y) = d(x,y) \quad (y \in X). \tag{2.11}$$

Proof. Let $f \in t^d(x)$. Then we have $f(z) \geq d(x,z)$ for $z \in X$ since $f(x) = 0$. For $y \in X$, by $f \in T(X,d)$, there exists $w \in X$ such that $f(y) + f(w) = d(y,w)$. By the triangle inequality, we have $d(y,x) + d(w,x) \leq f(y) + f(w) = d(y,w) \leq d(x,y) + d(x,w)$. Hence we obtain $f(y) = d(x,y)$. □

2.4 Results

We present a more general version of Theorem 1.2 below, which is also an extension of (a finite dimensional version of) the result of A. Dress [6] that a metric is a tree metric if and only if its tight span is a tree.

Theorem 2.3. *For a distance $d : X \times X \to \mathbf{R}$, the following conditions are equivalent.*

(*a*) *There exist some weighted tree T and a family of its subtrees T_x $(x \in X)$ such that*

$$d(x,y) = \min\{D_T(u,v) \mid u \in V(T_x), v \in V(T_y)\} \quad (x,y \in X). \tag{2.12}$$

(*b*) *There exist some compatible collection of partial splits $\mathcal{S}$ on X and a positive weight $\alpha : \mathcal{S} \to \mathbf{R}$ such that*

$$d = \sum_{S \in \mathcal{S}} \alpha_S \zeta_S. \tag{2.13}$$

(*c*) *$G(d)$ is a tree.*

(*d*) *$T(X,d)$ is a tree.*

(*e*) $\dim T(X,d) = 1$.

(*f*) *d satisfies the condition* (1.2).

The essential part of the proof of Theorem 2.3 relies on the following, which is an extension of the fact that finite metric space (X, d) can be isometrically embedded into $(T(X,d), \|\cdot\|_\infty)$ and realized by the 1-skeleton of $T(X,d)$ [6].

Theorem 2.4. *For a distance $d : X \times X \to \mathbf{R}$, the following holds.*

(1) $d(x,y) = \inf\{\|f-g\|_\infty \mid f \in t^d(x), g \in t^d(y)\} \quad (x,y \in X)$.

(2) $d(x,y) = \min\{D_{G(d)}(u,v) \mid u \in V(g^d(x)), v \in V(g^d(y))\} \quad (x,y \in X)$.

Remark 2.5. We show that the condition (1.2) reduces to the four-point condition (1.2) for a metric d. From the triangle inequality, we have

$$d(x,y) \leq \frac{1}{2}\{d(x,z)+d(z,y)\} + \frac{1}{2}\{d(x,w)+d(w,y)\}. \tag{2.14}$$

This implies that $d(x,y) \leq \max\{d(x,z)+d(y,w), d(x,w)+d(z,y)\}$. Similarly,

$$\{d(x,y)+d(y,z)+d(z,x)\}/2 \leq \{d(x,w)+d(w,y)+d(y,z)+d(z,x)\}/2 \tag{2.15}$$

implies

$$\{d(x,y)+d(y,z)+d(z,x)\}/2 \leq \max\{d(x,z)+d(y,w), d(x,w)+d(y,z)\}.$$

Remark 2.6. Every 3-point distance can be expressed as Theorem 2.3 (a). Let $d : \{1,2,3\} \times \{1,2,3\} \to \mathbf{R}$ be a distance on $\{1,2,3\}$. If d is a metric, then it is well known that d is a tree metric. Suppose that d does not satisfy the triangle inequality, say $d(1,2) > d(1,3) + d(2,3)$. Consider a weighted tree $T = (\{i,j,k,l\}, \{ij, jk, kl\}, w)$ with edge length $w_{ij} = d(1,3)$, $w_{jk} = d(1,2) - d(1,3) - d(2,3)$ and $w_{kl} = d(2,3)$, and a family of its subtrees $\{T_1 = (\{i\}, \emptyset), T_2 = (\{j,k\}, \{jk\}), T_3 = (\{l\}, \emptyset)\}$. Then they satisfy (2.12).

Remark 2.7. The split decomposition, due to Bandelt and Dress [1], has been extended in [9] for distances using partial split distances. A distance between subtrees of a tree, considered in this paper, is one of the examples of *totally split decomposable* distances in the sense of [9].

3 Proofs

In the following, let X be a finite set and $d : X \times X \to \mathbf{R}$ be a distance on X. For a set S, we denote by χ_S the characteristic vector of S defined as: $\chi_S(x) = 1$ if $x \in S$ and 0 otherwise. In particular we write simply χ_x instead of $\chi_{\{x\}}$ for a singleton $\{x\}$.

3.1 Preliminaries

For $f \in P(X,d)$, we define an undirected graph $K(f) = (X, E(f))$ by

$$xy \in E(f) \overset{\text{def}}{\Longleftrightarrow} f(x) + f(y) = d(x,y) \quad (x,y \in X), \tag{3.1}$$

where for $x, y \in X$, xy denotes an unordered pair, which means that xy and yx are not distinguished from each other. Note that $E(f)$ may contain loop edges,

like xx for $x \in X$. Let $F(f)$ be the face of $P(X,d)$ that contains f in its relative interior, which is also the set of solutions to the linear inequalities

$$\begin{cases} p(x)+p(y) = d(x,y) & (xy \in E(f)), \\ p(x)+p(y) \geq d(x,y) & (xy \notin E(f)). \end{cases} \tag{3.2}$$

By the same argument in the case that d is a metric [7], it is easy to observe that

$$\begin{aligned} f \in T(X,d) &\Leftrightarrow F(f) \text{ is bounded} && (3.3)\\ &\Leftrightarrow K(f) \text{ does not have isolated vertices} && (3.4)\\ &\Leftrightarrow \forall x \in X,\ f(x) = \max_{y \in X}\{f(y) - d(x,y)\}. && (3.5) \end{aligned}$$

For a connected graph (X,E), we observe

$$\operatorname{rank}\{\chi_x + \chi_y \mid xy \in E\} = \begin{cases} |X|-1 & \text{if } (X,E) \text{ is bipartite,} \\ |X| & \text{if } (X,E) \text{ is nonbipartite,} \end{cases} \tag{3.6}$$

where loops are regarded as odd cycles. Therefore, if $f \in T(f)$, we have

$$\begin{aligned} \dim F(f) &= |X| - \operatorname{rank}\{\chi_x + \chi_y \mid xy \in E(f)\} && (3.7)\\ &= \text{the number of bipartite components of } K(f). && (3.8) \end{aligned}$$

In particular, we have

$$\begin{aligned} F(f) \text{ is an edge} &\Leftrightarrow K(f) \text{ has only one bipartite component,} && (3.9)\\ F(f) \text{ is a vertex} &\Leftrightarrow K(f) \text{ is nonbipartite.} && (3.10) \end{aligned}$$

The dimension of $T(X,d)$ is given by

$$\dim T(X,d) = \max_{f \in T(X,d)} \{\text{the number of bipartite components of } K(f)\}. \tag{3.11}$$

3.2 Proof of Theorem 2.4

Let $D_1, D_2 : X \times X \to \mathbf{R}$ be defined as

$$\begin{aligned} D_1(x,y) &= \inf\{\|f-g\|_\infty \mid f \in t^d(x), g \in t^d(y)\} \quad (x,y \in X),\\ D_2(x,y) &= \min\{D_{G(d)}(u,v) \mid u \in V(g^d(x)), v \in V(g^d(y))\} \quad (x,y \in X). \end{aligned}$$

Lemma 3.1. *$d(x,y) \leq D_1(x,y) \leq D_2(x,y)$ holds for $x,y \in X$.*

Proof. For any $f \in t^d(x), g \in t^d(y)$, we have

$$f(x) = 0,\ f(y) \geq d(x,y),\ g(x) \geq d(x,y),\ g(y) = 0. \tag{3.12}$$

Hence we have $\|f-g\|_\infty \geq d(x,y)$. We may identify the graph $G(d)$ and the 1-skeleton of $T(X,d)$. Let $(f_0, f_1, \ldots, f_m)$ be a path of $G(d)$ with $f_0 \in V(g^d(x))$ and $f_m \in V(g^d(y))$. Hence the length of the path $(f_0, f_1, \ldots, f_m)$ is $\sum_{i=0}^{m-1} \|f_i - f_{i+1}\|_\infty \geq \|f_0 - f_m\|_\infty \geq D_1(x,y)$. □

In the following, we construct the path in $G(d)$ from $V(g^d(x))$ to $V(g^d(y))$ with its path length $d(x,y)$. That implies Theorem 2.4.

First, we take a vertex of $t^d(x)$. Let $X = \{x_1 = x, x_2 = y, x_3, \ldots, x_m\}$. Then, by a simple inductive argument, $f \in \mathbf{R}^X$ defined as

$$\begin{aligned} f(x_1) &= 0, \\ f(x_i) &= \max(0, \max_{k=1,\ldots,i-1}(d(x_i,x_k) - f(x_k)) \quad (i = 2, \ldots, m) \end{aligned}$$

is a vertex of $t^d(y)$. In particular, we have $xx, xy \in E(f)$, $f(y) = d(x,y)$, and $f(x) = 0$.

Next we try to move f toward $t^d(y)$ through edges of $T(X,d)$. If $yy \in E(f)$, then we have $f \in t^d(y)$ and $D_2(x,y) = D_1(x,y) = 0 = d(x,y)$. Hence we suppose $yy \notin E(f)$, i.e., $f(y) > 0$.

Let $S_y \subseteq X$ be a stable set of $K(f)$ constructed according to the following process, where for $S \subseteq X$, $\Gamma(S) = \{z \in X \setminus S \mid \exists w \in S,\ zw \in E(f)\}$:

(S0) $S_y = \{y\}$.

(S1) If there is no loopless vertex in $\Gamma(S_y \cup \Gamma(S_y))$, then output S_y and stop.

(S2) Take a loopless vertex $z \in \Gamma(S_y \cup \Gamma(S_y))$.

(S3) $S_y \leftarrow S_y \cup \{z\}$ and go to (S1).

By this construction, we see that the graph

$$G_{S_y} = (S_y \cup \Gamma(S_y), \{zw \mid z \in S_y, w \in \Gamma(S_y)\}) \tag{3.13}$$

is a connected bipartite graph. Furthermore, let $\epsilon_0 > 0$ be defined as

$$\epsilon_0 = \min \left\{ \begin{array}{c} \min_{z,w \in S_y} (f(z) + f(w) - d(z,w))/2, \\ \min_{z \in S_y, w \notin S_y \cup \Gamma(S_y)} f(z) + f(w) - d(z,w) \end{array} \right\} \tag{3.14}$$

and for $\epsilon \geq 0$, let $f^\epsilon \in \mathbf{R}^X$ be defined as

$$f^\epsilon = f + \epsilon(\chi_{\Gamma(S_y)} - \chi_{S_y}). \tag{3.15}$$

Then it is easily seen that

(1) $f^\epsilon \in T(X,d)$ for $0 \leq \epsilon \leq \epsilon_0$,

(2) $K(f^\epsilon)$ has one bipartite component G_{S_y} for $0 < \epsilon < \epsilon_0$, and

(3) $K(f^{\epsilon_0})$ is nonbipartite.

Hence the move $f \to f^{\epsilon_0}$ is on the edge of $T(X,d)$, f^{ϵ_0} is a vertex of $T(X,d)$, and we have

$$\|f^{\epsilon_0} - f\|_\infty = f^{\epsilon_0}(x) - f(x) = f(y) - f^{\epsilon_0}(y) = \epsilon_0. \tag{3.16}$$

Put $f_1 = f^{\epsilon_0}$ and repeat this process for f_1. Then we have the path $(f = f_0, f_1, f_2, \ldots)$ of $G(d)$. By (3.16), we have $f_0(y) > f_1(y) > \cdots$. After finitely many steps, we have $f_l(y) = 0$, $f_l(x) = d(x,y)$, and $f_l \in t^d(y)$. Therefore the path length of $(f = f_0, f_1, f_2, \ldots, f_l = g)$ is $\sum_{i=0}^{l-1} \|f_{i+1} - f_i\|_\infty = f(y) - g(y) = g(x) - f(x) = d(x,y)$.

3.3 Proof of Theorem 2.3

We prove Theorem 2.3 by showing the following:

$$\begin{array}{ccccc} (a) & \Leftarrow & (c) & \Leftarrow & (d) \\ \Downarrow & & & & \Updownarrow \\ (b) & \Rightarrow & (f) & \Leftrightarrow & (e) \end{array} \tag{3.17}$$

$(c) \Leftarrow (d)$ is obvious. $(a) \Leftarrow (c)$ follows from Theorem 2.4. $(d) \Leftrightarrow (e)$ follows from the contractibility of $T(X,d)$. $(f) \Leftrightarrow (e)$ follows from Theorem 2.1. Therefore, we need to prove $(a) \Rightarrow (b)$ and $(b) \Rightarrow (f)$.

$(a) \Rightarrow (b)$. Deletion of each edge e of T separates T into two trees T_e^A and T_e^B. From this, we have a disjoint pair $\{A_e, B_e\}$ defined as

$$A_e = \{x \in X \mid T_x \text{ is a subtree of } T_e^A\}, \tag{3.18}$$

$$B_e = \{x \in X \mid T_x \text{ is a subtree of } T_e^B\}. \tag{3.19}$$

For two edges $e, f \in E(T)$, we may assume that T_e^A is a subtree of T_f^A and T_f^B is a subtree of T_e^B. This implies the compatibility of $\{A_e, B_e\}$ and $\{A_f, B_f\}$. Hence we define the compatible collection of partial splits $\mathcal{S}$ on X and its positive weight $\alpha : \mathcal{S} \to \mathbf{R}$ by

$$\mathcal{S} = \{\{A_e, B_e\} \mid e \in E(T),\ \{A_e, B_e\} \text{ is a partial split}\}, \tag{3.20}$$

$$\alpha_{\{A_e, B_e\}} = \text{the length of edge } e\ . \tag{3.21}$$

Let $d' = \sum_{S \in \mathcal{S}} \alpha_S \zeta_S$. We show $d = d'$. Let $e \in E(T)$ be an edge with $\{A_e, B_e\} \in \mathcal{S}$. For $x \in A_e$ and $y \in B_e$, any path between T_x and T_y must contain e. This implies $d \geq d'$. Next we show $d \leq d'$. For $x, y \in X$, if T_x and T_y have a common vertex, i.e., $d(x,y) = 0$, then there is no edge in T that separates T_x and T_y. Hence we have $d(x,y) = d'(x,y) = 0$. Suppose that $d > 0$. Let $e \in E(T)$ be an edge of the shortest path between T_x and T_y. Neither T_x or T_y contains the edge e. Since both T_x and T_y are trees, it must be $x \in A_e,\ y \in B_e$ or $y \in A_e,\ x \in B_e$. Hence we have $\{A_e, B_e\} \in \mathcal{S}$. This implies $d \leq d'$.

$(b) \Rightarrow (f)$. It is sufficient to show this in the case that d is a distance on 4-point set. For this, we classify maximal compatible families of partial splits on 4-point set $\{1,2,3,4\}$. All partial splits on $\{1,2,3,4\}$ are listed below, where we denote a partial split $\{\{1,2\},\{3\}\}$ simply by 12|3:

(S1): 1|234, 2|134, 3|124, 4|123,

(S2): 12|34, 13|24, 23|14,

(S3): 1|2, 1|3, 1|4, 2|3, 2|4, 3|4,

(S4): 1|23, 2|13, 3|12, 1|24, 2|14, 4|12, 1|34, 3|14, 4|13, 2|34, 3|24, 4|23.

The next proposition shows that maximal compatible families of partial splits on $\{1,2,3,4\}$ are classified into six types. We illustrates this six types and their tree representations in Figure 2, where the line corresponding to a partial split $\{A,B\}$ separates points of A and B and meets points of $\{1,2,3,4\} \setminus A \cup B$.

Two families of partial splits $\mathcal{S}_1$ and $\mathcal{S}_2$ on X are said to be *isomorphic* if there exsits some bijection $\sigma : X \to X$ such that $\mathcal{S}_2 = \{\{\sigma(A), \sigma(B)\} \mid \{A,B\} \in \mathcal{S}_1\}$.

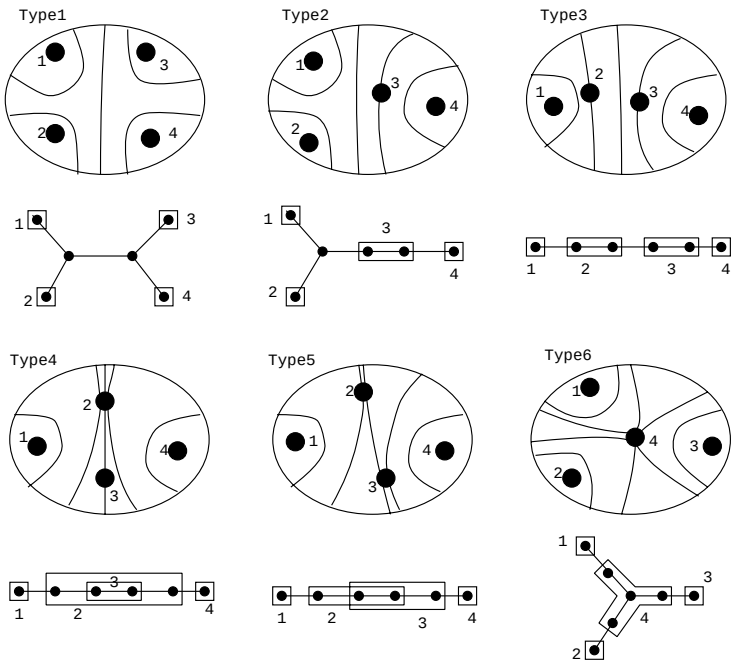

Figure 2: All types of maximal compatible families of $\{1,2,3,4\}$ and their tree representations

Proposition 3.2. *Any maximal compatible family of partial split on* $\{1,2,3,4\}$ *is isomorphic to one of the following:*

Type 1**:** $\{1|234, 2|134, 12|34, 3|124, 4|123\}$,

Type 2**:** $\{1|234, 2|134, 12|34, 12|4, 4|123\}$,

Type 3**:** $\{1|234, 1|34, 12|34, 12|4, 4|123\}$,

Type 4**:** $\{1|234, 1|34, 1|4, 13|4, 4|123\}$,

Type 5**:** $\{1|234, 1|34, 1|4, 12|4, 4|123\}$,

Type 6**:** $\{1|23, 2|13, 3|12, 1|234, 2|134, 3|124\}$.

Proof. For a family of partial splits $\mathcal{S}'$, the *incompatibility graph* of $\mathcal{S}'$ is defined to be a graph whose vertex set is $\mathcal{S}'$ and edge set is

$$\{ST \mid \ S \in \mathcal{S}' \text{ and } T \in \mathcal{S}' \text{ are not compatible}\}. \tag{3.22}$$

Then ${\mathcal{S}'}_0 \subseteq \mathcal{S}'$ is compatible if and only if ${\mathcal{S}'}_0$ is a stable set of the incompatibility graph of $\mathcal{S}'$.

Let $\mathcal{S}$ be a maximal compatible family of partial splits on $\{1,2,3,4\}$. Suppose that $\mathcal{S}$ has a partial split of (S2), say $12|34$. The set of all partial splits compatible to $12|34$ is given by

$$\mathcal{S}_1 = \{12|34, 1|234, 2|134, 3|124, 4|123, 1|34, 2|14, 12|4, 12|3\}. \tag{3.23}$$

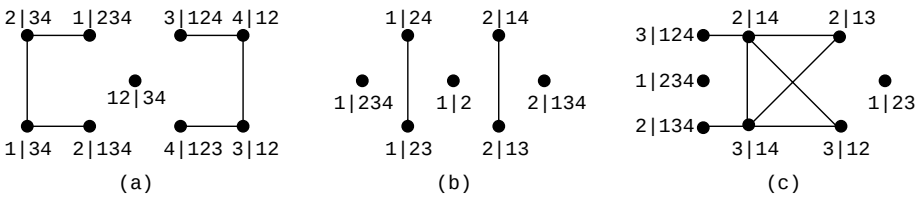

Figure 3: Incompatibility graphs

Then the incompatibility graph of $\mathcal{S}_1$ is (a) of Figure 3. From maximal stable sets of this graph, we see that $\mathcal{S}$ is of Type 1, Type 2, or Type 3.

Suppose that $\mathcal{S}$ has a partial split of (S3), say $1|2$. The set of all partial splits compatible to $1|2$ is given by

$$\mathcal{S}_2 = \{1|2, 1|234, 2|134, 1|24, 1|23, 2|34, 2|13\}. \tag{3.24}$$

Then the incompatibility graph of $\mathcal{S}_2$ is (b) of Figure 3. From maximal stable sets of this graph, we see that $\mathcal{S}$ is of Type 4 or Type 5.

Suppose that $\mathcal{S}$ has no partial splits of (S2) and (S3). If $\mathcal{S}$ consists of partial splits of (S1), $\mathcal{S}$ is not maximal compatible. Suppose that $\mathcal{S}$ has a partial split of (S4), say $1|23$. The set of all partial splits of (S1) and (S4) compatible to $1|23$ is given by

$$\mathcal{S}_3 = \{1|23, 2|13, 3|12, 1|234, 2|134, 3|124, 2|14, 3|14\}. \tag{3.25}$$

Then the incompatibility graph of $\mathcal{S}_3$ is (c) of Figure 3. Hence all maximal stable sets of this graph are

(1) $\{1|23, 2|13, 3|12, 1|234, 2|134, 3|124\}$,

(2) $\{1|23, 2|14, 1|234, 2|134\}$, and

(3) $\{1|23, 3|14, 1|234, 3|124\}$.

Neither (2) nor (3) is maximal compatible. Hence $\mathcal{S}$ must be (1) and is of Type 6.

□

Finally, we can easily confirm the condition (1.2) for each type in Proposition 3.2.

Acknowledgment

The author thanks Satoru Fujishige, Kazuo Murota, and Akihisa Tamura for helpful comments.

A Appendix

Proof of Theorem 2.1

Our proof is based on the fundamental duality principle in the theory of linear programming; see [13] for example for linear programming.

Lemma A.1. *Let $A = (a_1\ a_2\ \ldots\ a_m)$ be an $n \times m$ matrix with n-dimensional column vectors $\{a_i \mid i = 1, 2, \ldots, m\} \subseteq \mathbf{R}^n$. For $b \in \mathbf{R}^n$, consider the polyhedron*

$$Q = \{u \in \mathbf{R}^m \mid Au = b, u \geq 0\}. \tag{A.1}$$

Then $u \in Q$ is a vertex of Q if and only if vectors $\{a_i \mid u_i > 0\}$ are linearly independent.

Let E_X denote the set of unordered pairs defined as

$$E_X = \{xy \mid x \in X, y \in X\}. \tag{A.2}$$

The following is an easy consequence of the previous lemma.

Lemma A.2. *Let $Q(X)$ be a polyhedron defined as*

$$Q(X) = \{\lambda \in \mathbf{R}^{E_X} \mid \sum_{xy \in E_X} (\chi_x + \chi_y)\lambda_{xy} = 2\chi_X, \lambda_{xy} \geq 0\ (xy \in E_X)\}. \tag{A.3}$$

Then $\lambda \in Q(X)$ is a vertex of $Q(X)$ if and only if there exists some edge cover E of (X, E_X) consisting of (vertex) disjoint matching and odd cycles such that

$$\lambda_{xy} = \begin{cases} 2 & \text{if } xy \text{ is an edge of matching of } E, \\ 1 & \text{if } xy \text{ is an edge of some odd cycle of } E, \\ 0 & \text{otherwise,} \end{cases} \qquad (xy \in E_X). \tag{A.4}$$

Considering the facts that a permutation of X can be decomposed as disjoint cyclic permutations, that a cyclic permutation can be regarded as a cycle of graph (X, E_X) and that an even cycle is the union two edge-disjoint matchings, the optimal value of linear program

$$\max. \sum_{xy \in E_X} \lambda_{xy} d(x, y) \quad \text{s.t.} \quad \lambda \in Q(X) \tag{A.5}$$

is given by

$$\max\{\sum_{x \in X} d(x, \sigma(x)) \mid \sigma \text{ is a permutation of } X\}. \tag{A.6}$$

Lemma A.3. *The following holds, where $d^Y : Y \times Y \to \mathbf{R}$ denotes the restriction of d to Y.*

(1) $\dim T(Y, d^Y) \leq \dim T(X, d)$ *for* $Y \subseteq X$.

(2) *If* $\dim T(X, d) \geq n$, *there exists* $Y \subseteq X$ *with* $|Y| = 2n$ *such that* $\dim T(Y, d^Y) = n$.

Proof. For $f \in \mathbf{R}^X$ and $Y \subseteq X$, let $f^Y : Y \to \mathbf{R}$ denote the restriction of f to Y.

(1). It is sufficient to show the case $Y = X \setminus \{z\}$ for some $z \in X$. Suppose that $\dim T(Y, d^Y) = n$. Then there exists $f \in T(Y, d^Y)$ such that a graph $(Y, E(f^Y))$ has n bipartite components $(A_1 \cup B_1, E_1), \ldots, (A_n \cup B_n, E_n)$ with $A_i \cap B_i = \emptyset$ and $E_i \subseteq \{xy \mid x \in A_i, y \in B_i\}$ for $i = 1, \ldots, n$. Let $f' \in \mathbf{R}^X$ be defined as

$$f'(x) = \begin{cases} \max\{0, \max_{y \in Y}(d(z, y) - f(y))\} & \text{if } x = z, \\ f(x) & \text{otherwise.} \end{cases} \tag{A.7}$$

Then some edges connecting z appear in $(X, E(f'))$ and we have $f' \in T(X,d)$. If $(X, E(f'))$ has no edges connecting $\{z\}$ and $A_1 \cup B_1 \cup \cdots \cup A_n \cup B_n$, then $(X, E(f'))$ also has n bipartite components.

We suppose that there exists $y \in A_1$ with $zy \in E(f')$. Let S and S' be stable sets of $(X, E(f'))$ defined as $S = A_1 \cup A_2 \cup \cdots \cup A_n$ and $S' = A_1 \cup B_2 \cup \cdots \cup B_n$. Let $g \in \mathbf{R}^X$ be defined as

$$g = f' + \epsilon(\chi_{\Gamma(S)} - \chi_S) + \epsilon'(\chi_{\Gamma(S')} - \chi_{S'}) \tag{A.8}$$

for sufficiently small $\epsilon, \epsilon' > 0$. Then we have $g \in T(X,d)$. Furthermore all edges in $(X, E(f'))$ connecting $\{z\}$ and $X \setminus A_1$ vanish in $(X, E(g))$. This implies that $(X, E(g))$ has n bipartite components.

(2). Since $\dim T(X,d) \geq n$, there exists $f \in T(X,d)$ such that $(X, E(f))$ has n bipartite components. Take n edges from each bipartite component, say $\{x_1y_1, x_2y_2, \ldots, x_ny_n\}$ and put $Y = \{x_1, x_2, \ldots, x_n, y_1, \ldots, y_n\}$. Then it is easy to check that f^Y is in $T(Y, d^Y)$ and $(Y, E(f^Y))$ has n bipartite components. □

Hence, it is sufficient to show the following.

Theorem A.4. *Suppose $|X| = 2n$. The following conditions are equivalent.*

(*a*) $\dim T(X,d) = n$.

(*b*) *There exists some perfect matching M of (X, E_X) such that $\lambda^* = 2\chi_M \in \mathbf{R}^{E_X}$ is the unique optimal solution to linear program (A.5).*

Proof. (*a*) $\Rightarrow$ (*b*). There exists $f^* \in P(X,d)$ such that $K(f^*)$ has n bipartite components. Hence $E(f^*)$ must be a perfect matching of (X, E_X). Consider the dual program of (A.5):

$$\min. \sum_{x \in X} f(x) \quad \text{s.t.} \quad f \in P(X,d). \tag{A.9}$$

Then $\lambda^* = \chi_{E(f^*)}$ and f^* satisfies the (strict) complementary slackness condition

$$\lambda^*_{xy} > 0 \Leftrightarrow f^*(x) + f^*(y) = d(x,y) \quad (xy \in E_X). \tag{A.10}$$

Hence λ^* and f^* are optimal solutions to (A.5) and (A.9), respectively. Conversely, any optimal solution $\tilde{\lambda}$ of (A.5) satisfies

$$\tilde{\lambda}_{xy} = 0 \quad (xy \notin E(f^*)). \tag{A.11}$$

Since $\{\chi_x + \chi_y \mid xy \in E(f^*)\}$ is linearly independent, we have $\tilde{\lambda} = \lambda^*$. Hence λ^* is the unique optimal solution of linear program (A.5).

(*b*) $\Rightarrow$ (*a*). By the strict complementary slackness theorem, there exist optimal solutions $\tilde{\lambda}$ and f^* of (A.5) and (A.9) such that

$$\tilde{\lambda}_{xy} > 0 \Leftrightarrow f^*(x) + f^*(y) = d(x,y) \quad (xy \in E_X). \tag{A.12}$$

By the condition (*b*), we have $\tilde{\lambda} = \lambda^*$. Hence it must be that $E(f^*) = M$. This implies $\dim T(X,d) = n$. □

References

[1] H.-J. Bandelt and A. W. M. Dress: A canonical decomposition theory for metrics on a finite set, *Advances in Mathematics* **92** (1992), no. 1, 47–105.

[2] J.-P. Barthélémy and A. Guénoche: *Trees and Proximity Representations*, Translated from the French by Gregor Lawden, Wiley, Chichester, 1991.

[3] P. Buneman: The recovery of trees from measures of dissimilarity, *Mathematics in the Archaeological and Historical Sciences* (F. R. Hodson, D. G. Kendall, P. Tautu, eds.) Edinburgh University Press, Edinburgh, 1971, 387–395.

[4] P. Buneman: A note on metric properties of trees, *Journal of Combinatorial Theory, Series B* **17** (1974) 48–50.

[5] M. Chrobak and L. L. Larmore: Generosity helps or an 11-competitive algorithm for three servers, *Journal of Algorithms* **16** (1994), no. 2, 234–263

[6] A. W. M. Dress: Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: a note on combinatorial properties of metric spaces, *Advances in Mathematics* **53** (1984), no. 3, 321–402.

[7] A. W. M. Dress: Towards a classification of transitive group actions on finite metric spaces, *Advances in Mathematics* **74** (1989), no. 2, 163–189.

[8] A. Dress, V. Moulton and W. Terhalle: T-theory: an overview, *European Journal of Combinatorics* **17** (1996), no. 2-3, 161–175.

[9] H. Hirai: Geometric study on the split decomposition of finite metrics, RIMS preprint 1459, Kyoto University, May 2004.

[10] J. R. Isbell: Six theorems about injective metric spaces, *Commentarii Mathematici Helvetici* **39** (1964), 65–76.

[11] C. Semple and M. Steel: *Phylogenetics*, Oxford University Press, Oxford, 2003.

[12] J. M. S. Simões-Pereira: A note on the tree realizability of a distance matrix, *Journal of Combinatorial Theory* **6** (1969), 303–310.

[13] R. J. Vanderbei: *Linear programming*, Second edition, Kluwer, Boston, 2001.

[14] K. A. Zareckiĭ: Constructing a tree on the basis of a set of distances between the hanging vertices, *Uspehi Matematičeskih Nauk* **20** (1965), no. 6, 90–92.