Sets of Double and Triple Weights of Trees

Elena Rubei
Dipartimento di Matematica “U. Dini”, viale Morgagni 67/A, 50134 Firenze, Italia
rubei@math.unifi.it

Received April 28, 2009

Mathematics Subject Classification: 05C05, 05C12, 92B05

Abstract. Let $T$ be a weighted tree with $n$ leaves numbered by the set $\{1, \ldots, n\}$. Let $D_{i,j}(T)$ be the distance between the leaves $i$ and $j$. Let $D_{i,j,k}(T) = \frac{1}{2} \left( D_{i,j}(T) + D_{j,k}(T) + D_{i,k}(T) \right)$. We will call such numbers “triple weights” of the tree. In this paper, we give a characterization, different from the previous ones, for sets indexed by 2-subsets of an $n$-set to be double weights of a tree. By using the same ideas, we find also necessary and sufficient conditions for a set of real numbers indexed by 3-subsets of an $n$-set to be the set of the triple weights of a tree with $n$ leaves. Besides we propose a slight modification of Saitou-Nei’s Neighbour-Joining algorithm to reconstruct trees from the data $D_{i,j}$.

Keywords: trees, weights of trees, neighbour-joining algorithm

1. Introduction

Consider a positive-weighted tree $T$ (that is, a tree such that every edge is endowed with a positive real number, which we call the length of the edge) with $n$ leaves numbered by the set $\{1, \ldots, n\}$. Let $D_{i,j}(T)$ be the sum of the lengths of the edges of the shortest path connecting $i$ and $j$. We call such number the “distance” between the leaves $i$ and $j$ or the “double weight” for $i$ and $j$.

In 1971, Buneman characterized the metrics on finite sets coming from a tree:

Theorem 1.1. (Buneman) A metric $(D_{i,j})$ on $\{1, \ldots, n\}$ is the metric induced by a positive-weighted tree if and only if for all $i, j, k, h \in \{1, \ldots, n\}$ the maximum of $\{D_{i,j} + D_{k,h}, D_{i,k} + D_{j,h}, D_{i,h} + D_{k,j}\}$ is attained at least twice.

The problem of reconstructing trees from data involving the distances between the leaves has several applications, such as internet tomography and phylogenetics; evolution of species can be represented by trees and, given distances between genetic sequences of some species, one can try to reconstruct the evolution tree from these distances. Some algorithms to reconstruct trees from the data $\{D_{i,j}\}$ have been proposed. Among them is neighbour-joining method, invented by Saitou and Nei in 1987 (see [9, 11, 14]).
For any weighted tree $T$, let now

$$D_{i,j,k}(T) = \frac{1}{2} (D_{i,j}(T) + D_{j,k}(T) + D_{i,k}(T)),$$

that is the sum of the lengths of the edges of the minimal subtree with $i, j, k$ as the set of leaves. We call such numbers “triple weights” of the tree. More generally, define the $k$-weights of the tree $D_{i_1,\ldots,i_k}(T)$ as the sum of the lengths of the edges of the minimal subtree connecting $i_1,\ldots,i_k$.

In 2004, Pachter and Speyer proved the following theorem (see [7]).

**Theorem 1.2.** (Pachter-Speyer) Let $k, n \in \mathbb{N}$ with $n \geq 2k - 1$ and $k \geq 3$. A positive-weighted tree $T$ with $n$ leaves $1,\ldots,n$ and no vertices of degree 2 is determined by the values $D_I$ where $I$ varies in the $k$-subsets of $\{1,\ldots,n\}$.

It can be interesting to characterize the sets of real numbers which are sets of $k$-weights of a tree, for instance triple weights of a tree; in fact (I quote Speyer and Sturmfels’s paper [13]) “it can be more reliable statistically to estimate the triple weights $D_{i,j,k}$ rather than the pairwise distances $D_{i,j}$”. We refer to [13] and above all to [7] for an analysis of this and the references.

In [1], Bocci and Cools gave a description of $k$-dissimilarity maps of a tree and generalized Buneman’s result for sets of real numbers $\{D_{i,j}\}$ indexed by 3-subsets of $\{1,\ldots,n\}$ coming from sets $\{D_{i,j}\}$.

In this paper, we give a characterization (different from Buneman’s one) for sets indexed by 2-subsets of an $n$-set to be double weights of a tree with $n$ leaves (see Theorem 2.9) (please note that in this paper a weight is not necessarily positive).

By using the same ideas, we find also necessary and sufficient conditions for a set of real numbers indexed by 3-subsets of an $n$-set to be the set of the triple weights of a tree (see Theorem 2.10).

Finally, by using the characterization of neighbours we used to deduce the above theorems, we propose a slight modification of the neighbour-joining algorithm.

### 2. The Main Theorems

**Definition 2.1.** A 2-cherry $B$ in a tree $T$ is a subtree $B$ with two leaves such that only one of the inner vertices is not bivalent; we call this vertex “stalk” of the bell and we say that the two leaves are neighbours. We call the path from a leaf of a bell to its stalk “twig” of this leaf.

A cherry is a union of 2-cherries with the same stalk.

**Notation 2.2.**
- For every $n \in \mathbb{N}$, let $[n] = \{1,\ldots,n\}$; besides, if $k \leq n$, denote the set of the $k$-subsets of $[n]$ by $[n]_k$.
- For any set $\{D_{i_1,\ldots,i_k}\}$ of real numbers indexed by the $k$-subsets $\{i_1,\ldots,i_k\}$ of $[n]$, we denote $D_{\{i_1,\ldots,i_k\}}$ by $D_{i_1,\ldots,i_k}$ for any order of $i_1,\ldots,i_k$.
- A weighted tree is a tree such that every edge is endowed with a real number called weight or length of the edge. If the weights are positive we say that the tree is positive-weighted. Please note that in other papers “weighted” means positive-weighted.