

## Tree structures for proximity data

H. Colonius and H. H. Schulze

---

Previous work on representing proximity data by tree structures is reviewed. The authors then present two paradigms for data collection and develop appropriate measurement structures for non-numerical data. Two theorems on representing these data by rooted or unrooted trees are given and their relationship to the more conventional representation by a dissimilarity index is discussed.

---

### 1. Introduction

Among typical paradigms for the collection of proximity data one finds various rating or sorting tasks; for example, a subject may be asked to give a rating of the subjective similarity for each pair of the set of objects under study, to judge which of two pairs of objects are more similar, or to sort the objects into subsets of equal similarity (e.g. Miller, 1969; Fillenbaum & Rapoport, 1971). The data analysis usually proceeds by computing some numerical index of subjective distances between the objects. It is assumed that the objects under study can be embedded in some geometric space in such a way that the subjective distances are represented by the metric distances between the respective points. Complementing this multidimensional scaling approach, there has been a growing interest in recent years in representing proximity data by graph-theoretical structures and in comparing both kinds of representations (Holman, 1972; Carroll, 1976; Sattath & Tversky, 1977). Here it is assumed that the objects can be represented as nodes of a graph, typically a tree, so that the subjective distances are reflected by some distance measure on the set of nodes of the graph. While for both approaches numerous efficient algorithms are available for producing representations that are best fitting in some specified sense, research on which properties of the proximity data actually guarantee the existence of the assumed representation is scarce. In this paper, we analyse the tree representation problem from a measurement-theoretic point of view. We should like to emphasize that the starting point of our analysis is not some numerical index of the subjective distances derived from the subjects' judgements but rather an empirical relation on the set of objects under study which is defined directly in terms of those judgements. For example, if the subject is asked to indicate, for each triple of objects, that pair which seems 'most similar', we are dealing with a ternary relation on the set of objects. This approach may be compared to Beals *et al.*'s (1968) foundation of multidimensional scaling where they consider conditions on a relation between pairs of objects of a set that guarantee the existence of a metric on that set. It emerges from this axiomatic point of view that the two kinds of approaches—geometric and graph-theoretical—are not different in principle since both search for conditions on proximity data that allow embeddings in a metric space. They differ, however, as to the properties one wishes the metric space to possess.

The organization of this paper is as follows: we start with some needed graph-theoretical terminology; then we review previous work on representing proximity data by tree structures. In the fourth section, we introduce two paradigms for collecting proximity data and systematically develop appropriate measurement

structures. We provide proofs for two representation theorems suggested without proof by Colonius & Schulze (1979). Finally, we report some preliminary results on the application of our method to the representation of verbal meaning.

## 2. Terminology

A set with  $m$  elements will be called an  $m$ -set. A graph  $G = (N, E)$  consists of a finite non-empty set  $N$  of nodes and a set  $E$  of unordered pairs of distinct nodes of  $N$ . Each such pair  $\{a, b\}$ ,  $a, b \in N$ , is an edge of  $G$  and is said to be *incident* with both  $a$  and  $b$ . A *subgraph* of  $G$  is a graph which has all its nodes and edges in  $G$ . The *degree* of a node  $a$  in  $G$ , denoted by  $\text{deg}(a, G)$ , is the number of edges incident with  $a$  in  $G$ . Nodes of degree one are called *terminal nodes*. A *walk* of a graph  $G$  is an alternating sequence of nodes and edges  $a, \{a, b\}, b, \{b, c\}, c, \dots$  beginning and ending with nodes, in which each edge is incident with the two nodes immediately preceding and following it. A *path* is a walk where all nodes are distinct. A *cycle* is a path where the first and the last node are identical. A graph is *connected* if every pair of nodes is joined by a path. A *tree* is a connected graph having no cycles. A *rooted tree* has one node, its root, distinguished from the others. Any node may be elected to be the root. The *graphical distance* of node  $a$  to node  $b$  is equal to the number of edges in the path from  $a$  to  $b$ . A *dissimilarity index* on a set of objects  $S$  is a function  $\delta$  from  $S \times S$  into the non-negative real numbers such that (i)  $\delta(a, a) = 0$ , (ii)  $\delta(a, b) = \delta(b, a)$  for any  $a, b \in S$ . For simplicity, we also assume that (iii)  $\delta(a, b) = 0$  implies  $a = b$ .

## 3. Previous work

The representation problem has usually been stated as follows: Given a set of objects and a dissimilarity index find a tree in which the objects are represented as terminal nodes such that a to be defined tree metric corresponds to the dissimilarity index. Two different definitions of a tree metric have been considered:

(a) *The ultrametric*. In a rooted tree representation a non-negative weight is attached to every node such that the terminal nodes have weight zero and the root has the largest weight and the sequence of weights of the nodes on a path from a terminal node to the root is strictly increasing. The distance between two nodes  $a$  and  $b$  is defined as the maximum of the weights of the nodes on the path from  $a$  to  $b$ .

A necessary and sufficient condition on the dissimilarity for this representation is well known. It is the so-called ultrametric inequality (Johnson, 1967). For all the elements  $a, b, c$  of the set the following condition must hold:

$$d(a, c) \leq \max(d(a, b), d(b, c)).$$

Numerous algorithms are available for the construction of rooted trees from fallible data (Johnson, 1967; Hartigan, 1967, 1975; Lerman, 1970; Jardine & Sibson, 1971; Hubert, 1972, 1973).

(b) *Additive tree metric*. An alternative representation has been explored more recently by several investigators. A weight is attached to each edge and the distance between two nodes is defined as the sum of the weights of the edges on the path from one node to the other.

There exist several parallel proofs in the literature that the so-called four-point condition is necessary and sufficient for an additive tree representation (Simoes-

Pereira, 1969; Patrinos & Hakimi, 1972; Dobson, 1974; Buneman, 1974). This condition is defined as follows:

$$d(a, b) + d(c, d) \leq \max \begin{cases} d(a, c) + d(b, d) \\ d(a, d) + d(b, c) \end{cases}$$

Dobson (1974) has shown that the ultrametric inequality implies the four-point condition, which in turn implies the triangle inequality

$$d(a, b) + d(b, c) \geq d(a, c).$$

The problem of finding an optimal tree that gives the minimum discrepancy between the dissimilarity index and the tree distance has not yet been completely solved.

Note that two subproblems arise: first one has to find an appropriate tree and second the optimal weights must be computed. The second problem has an analytic solution for a least squares criterion, but the properties of the algorithms for the construction of the tree are not fully understood (Cunningham, 1974, 1978; Carroll & Pruzanski, 1975; Carroll, 1976; Sattath & Tversky, 1977).

*Additive bidirectional trees.* If similarity judgements are asymmetric, the concept of an additive tree must be generalized further. An *additive bidirectional tree* is an additive tree in which each edge has been replaced by a pair of edges directed in opposite directions. Patrinos & Hakimi (1972) give necessary and sufficient conditions for the existence of an additive bidirectional tree representation and Cunningham (1978) presents a least squares solution for the tree construction problem. For psychological models incorporating asymmetric similarity judgements the reader is referred to Tversky (1977), Krumhansl (1978), and Möbus (1979).

*Betweenness.* Still another approach apparently closer to the one that will be presented in the next section was given by Sholander (1952) and, more recently, Defays (1979). They consider an (empirical) ternary relation called 'betweenness' on the set of objects under study which is then represented by a betweenness relation among the nodes of a tree ('the node  $b$  is on the path from node  $a$  to node  $c$ '). Note, however, that this representation has some peculiar features. The more empirical 'betweenness' judgements can be given by the subject the more trivial the tree structure becomes; specifically, if the subject is able to give a 'betweenness' judgement for any triple of objects, the tree becomes a single line or serial order. On the other hand, the less 'betweenness' judgements are available the more structurally different trees exist that represent the 'betweenness' relation correctly. In the latter case, only a subset of the nodes of the representing tree corresponds to the set of given objects.

#### 4. Tree structure representation

When psychologists construct a tree from dissimilarity data they hope that this tree reflects the subjective representation of the objects under study. If they are lucky and imaginative they can give a plausible interpretation of the tree. Typically, the interpretation consists of giving names to the edges of the tree that denote properties of the objects such that all the objects on one side of the edge share this property while the objects on the other side do not. A property may be a particular value of an attribute or a logical combination of values of several attributes. This suggests a different approach to the construction of the tree. Rather than asking the subjects to

give a numerical judgement (rating) of the dissimilarity of the objects one may ask them to sort the objects into different classes. Specifically, we are concerned here with two paradigms. In the first, a variation of the method of triads (Luce & Galanter, 1963), a trial consists of presenting a set of three objects and asking the subject to indicate the 'most similar' pair. Proximity data of this kind have the form of a ternary relation  $T$  on  $A$ , i.e.  $T \subset A^3$ . For example,  $abTc$  will indicate that the subject considers neither  $a, c$  nor  $b, c$  the 'most similar' pair (for  $a, b, c \in A$ ). In the second paradigm, the subject is confronted with four objects at a time and has to sort them into two classes of two objects each such that the 'most similar' objects are together. These data have the form of a quaternary relation  $H$  on  $A$ , i.e.  $H \subset A^4$ . For example,  $abHcd$  will indicate that the subject considers that sorting  $a, b$  against  $c, d$  expresses the respective similarities at least as well as the other two possible partitions. It should be noted that in both paradigms the subject is given the option of not sorting the objects at all if all similarities appear equal. It will turn out that data of the latter kind may appropriately be represented by an unrooted tree while those of the former paradigm suggest a rooted tree representation. For a further discussion of the psychological aspects of these data representations the reader is referred to Colonius & Schulze (1979). We now turn to the tree representations of the empirical relations.

Let  $a, b, c$  be elements of the set of terminal nodes of some rooted tree  $R = (N, E, r)$  with root  $r$ . Obviously, either the three nodes can be relabelled  $u, v, w$  such that the path from  $u$  to  $v$  has no nodes in common with the path from  $w$  to the root, or any path connecting two of the terminal nodes contains a node of the path from the third terminal node to the root. The possible configurations are given in Fig. 1. We

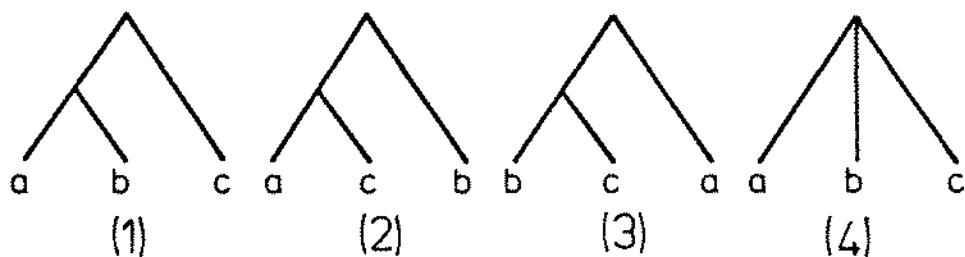


Figure 1. Four possible configurations with three terminal nodes in a rooted tree.

abbreviate case (1) by  $(ab)*c$ , (2) and (3) analogously, and (4) by  $(abc)$ . As will be seen, specifying for any triple of terminal nodes which one of the above configurations holds uniquely determines the root and the whole structure of the tree up to deletion or addition of non-root nodes of degree two. For simplicity, in what follows, we consider only rooted trees with the root not being an element of the set of terminal nodes. Here we do not want to presuppose the existence of a tree but rather investigate the following problem:

*Problem (T).* Given a set of objects  $S$ , which restrictions have to be imposed on a ternary relation  $T$  on  $S$  such that there exists a rooted tree where  $T$  can be represented in terms of the ternary relation on the terminal nodes depicted in Fig. 1 in the following sense

$$abTc \text{ iff } (ab)*c \text{ or } (abc)$$

(for all  $a, b, c \in A$ )?

Two remarks are in order here. First, it may not be obvious to the reader why a tree with a root is called for in this case. Suppose there is no root specified; then, for any triple of terminal nodes, the four configurations in Fig. 1 are undiscernible and would thus not reflect the relation  $T$ . We do not claim that there are no other ways of representing the relation  $T$  in a tree but the one we know of—the betweenness relation—does not seem appropriate for similarity representations. Second, it is not necessary to restrict attention to the set of terminal nodes. It would not be difficult conceptually to generalize the representation in such a way that objects may also be represented by non-terminal nodes. However, limits of space prevent us from doing so.

Before presenting a solution for Problem ( $T$ ) we introduce the corresponding problem for the relation  $H$  of the second paradigm.

Let  $a, b, c, d$  be elements of the set of terminal nodes  $N$ , of some tree  $T = (N, E)$ . Obviously, either the four nodes can be relabelled  $u, v, w, z$  such that the path joining  $u$  and  $v$  has no edge in common with the path joining  $w$  and  $z$ , or all paths joining any two of the nodes intersect at one and only one non-terminal node of  $T$ . The possible configurations are given in Fig. 2.

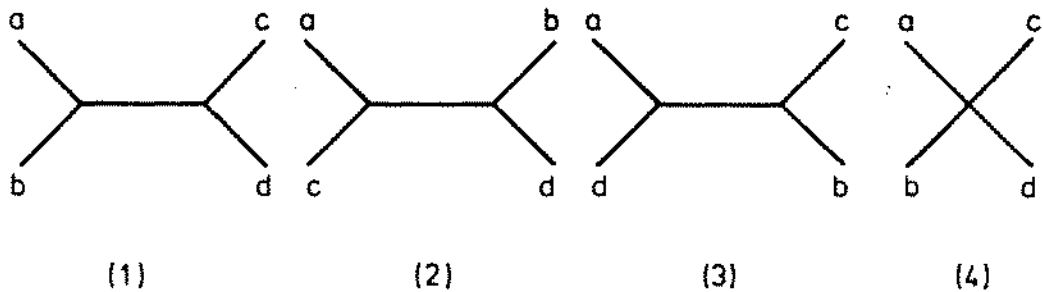


Figure 2. Four possible configurations with four terminal nodes in an unrooted tree.

We will abbreviate case (1) by  $(ab)^*(cd)$ , (2) and (3) analogously, and (4) by  $(abcd)$ . As will be seen, specifying for any quadruple of terminal nodes which one of the above configurations holds (path intersection property) uniquely determines the whole structure of the tree up to deletion or addition of nodes of degree two.

*Problem (H).* Given a set of objects  $S$ , which restrictions have to be imposed on a quaternary relation  $H$  on  $S$  such that there exists a tree where  $H$  can be represented in terms of the quaternary relation on the terminal nodes depicted in Fig. 2 in the following sense

$$abHcd \text{ iff } (ab)^*(cd) \text{ or } (abcd)$$

(for all  $a, b, c, d \in A$ )?

The remark above on confining attention to the terminal nodes applies to this case, too. The following definition captures the notion of an empirical relational structure for the first paradigm:

*Definition 1.* A pair  $(S, T)$  is a  $T$ -structure iff  $S$  is a finite non-empty set and  $T$  a ternary relation on  $S$ , such that for all  $a, b, c \in S$

(i)  $abTc$  iff  $baTc$ ;

- (ii)  $abTc$  or  $acTb$  or  $bcTa$ , and if any two of these hold so does the third (in which case we write  $abcT$ );  
 (iii)  $abTc$  implies either  $a \neq c$  and  $b \neq c$  or  $a = b = c$ .

We write  $abT^*c$  iff neither  $acTb$  nor  $bcTa$  hold.

In the above definition, (i) reflects the inherent symmetry in the task of picking a pair of objects, while (ii) asserts that every triple of objects is taken into consideration. (ii) and (iii) imply  $aaT^*b$  for any distinct  $a, b \in S$ , i.e. a pair of identical objects is always judged 'most similar'.

*Example 1.* The set of terminal nodes of a rooted tree with the ternary relation illustrated in Fig. 1 provides an obvious example for a  $T$ -structure in the sense suggested by Problem ( $T$ ). It will become apparent, however, that further restrictions on a  $T$ -structure are needed for a solution of the representation problem.

*Example 2.* Take a family of subsets of some base set  $M$  where the pairwise intersections satisfy the properties of a  $T$ -structure with respect to the following interpretation of the relation  $T$ :  $abTc$  iff  $a \cap c = b \cap c \subset a \cap b$  for  $a, b, c \subset M$ . Such a family has a natural interpretation as a family of feature sets associated with objects that can be represented by a rooted tree: each terminal node/object is assigned those edges/features that constitute the path from this terminal node to the root (see also Colonius & Schulze, 1979).

*Definition 2.* A  $T$ -structure  $(S, T)$  is called *rooted-tree-realizable* (*rt-realizable*) iff there exists a rooted tree  $R = (N, E, r)$  unique up to addition or deletion of non-root nodes of degree two and a one-to-one correspondence  $f$  between the elements of  $S$  and the terminal nodes of  $R$  such that for any  $a, b, c \in S$

$$abTc, \text{ iff } (f(a)f(b))^*f(c) \text{ or } (f(a)f(b))f(c).$$

Because of the one-to-one correspondence between  $a$  and  $f(a)$ , in what follows no distinction will be made in the notation between the elements of  $S$  and their corresponding terminal nodes.

The following lemma is tantamount to Johnson's (1967) result mentioned in Section 3a.

*Lemma 1.* Let  $(S, T)$  be a  $T$ -structure; then the following two conditions are equivalent:

- (i)  $(S, T)$  is *rt-realizable*;  
 (ii) there exists an ultrametric on  $S$  such that

$$abTc \text{ iff } d(a, c) = d(b, c) \geq d(a, b)$$

for all  $a, b, c \in S$ .

The following theorem provides a solution to Problem ( $T$ ). Its proof as well as the proof of the corresponding theorem for Problem ( $H$ ) are postponed to the next section which may be skipped without losing the main line of reasoning of this paper.

*Theorem 1.* Let  $(S, T)$  be a  $T$ -structure; the following two conditions are equivalent:

- (i)  $(S, T)$  is  $rt$ -realizable;  
 (ii) whenever  $adT^*c$  or  $bdTa$ , then  $abT^*c$  implies  $bdT^*c$  for all  $a, b, c, d \in S$ .

Some comments on this theorem are in order.

*Remark 1.* (ii) is a simple necessary and sufficient condition for  $rt$ -realizability which is, in principle, accessible to empirical testing. However, much like the representations of fundamental measurement structures (cf. Krantz *et al.*, 1971),  $rt$ -realizability could be refuted by a single violation of Condition (ii). This is due to the deterministic nature of our structures and a formulation that takes statistical variability into account is in fact desirable.

*Remark 2.* The relation between our results and the representation of a dissimilarity index by an ultrametric mentioned in Section 3a is worth some attention. The representation problem has two aspects in this case. The first aspect concerns the existence problem: Does there exist a rooted tree that represents the dissimilarity index and what is its structural form? The second aspect is to find optimal weights for the nodes. The various algorithms do not provide an answer to the first question: they always produce a rooted tree whatever the dissimilarities look like. If the dissimilarity index is not a perfect ultrametric, these trees will be different depending on the special algorithm being applied. The weights for the nodes computed by these algorithms, however, satisfy certain optimality criteria (for further details the reader is referred to the literature cited in Section 3). On the other hand, our theorem addresses the existence question while no information is given concerning the node weights. In fact, it will become clear from Lemma 1 and the proof of Theorem 1 that  $rt$ -realizability determines an ultrametric only up to monotone transformations. Given a dissimilarity index  $\delta$ , Theorem 1 may be used to check the existence problem in the following way: define relation  $T^*$  by

$$abT^*c \text{ iff } \delta(a, b) < \min \{ \delta(a, c), \delta(b, c) \}.$$

Then Condition (ii) of Theorem 1, formulated in terms of the dissimilarity index, gives a necessary and sufficient condition for  $rt$ -realizability of the index.

We now turn to a solution of Problem (H).

*Definition 3.* A pair  $(S, H)$  is an  $H$ -structure iff  $S$  is a finite set,  $|S| > 2$ ,  $H$  a quaternary relation on  $S$  such that for all  $a, b, c, d \in S$

- (i)  $abHcd$  iff  $cdHab$  iff  $dcHab$ ;  
 (ii)  $abHcd$  or  $acHbd$  or  $adHbc$ , and if any two of these hold, so does the third (in which case we write  $abcdH$ )  
 (iii)  $abHcd$  implies either  $(a \neq c \text{ and } a \neq d \text{ and } b \neq c \text{ and } b \neq d)$  or  $a = b = c = d$ .

We write  $abH^*cd$  iff neither  $acHbd$  nor  $adHbc$  hold. For  $a$  different from  $b$  and  $c$ , (i)–(iii) imply  $aaH^*bc$ , and  $aaHaa$ . There are some obvious analogies to the  $T$ -structure concept. We thus keep our discussion of  $H$ -structures somewhat cursory and invite the reader to fill in the gaps. Examples 1 and 2 for  $T$ -structures can be transformed in an obvious manner to apply to  $H$ -structures as well (see also the discussion in Colonius & Schulze, 1979).

*Definition 4.* An  $H$ -structure  $(S, H)$  is called *tree-realizable* ( $t$ -realizable) iff there exists a tree  $(N, E)$  unique up to addition or deletion of nodes of degree two and a

one-to-one correspondence  $f$  between the elements of  $S$  and the terminal nodes of  $(N, E)$  such that for any  $a, b, c, d \in S$

$$abHcd \text{ iff } (f(a)f(b))^*(f(c)f(d)) \text{ or } (f(a)f(b)f(c)f(d)).$$

*Lemma 2.* Let  $(S, H)$  be an  $H$ -structure; then the following two conditions are equivalent:

- (i)  $(S, H)$  is  $t$ -realizable;
- (ii) there exists an additive tree metric on  $S$  such that  $abHcd$  iff  $d(a, c) + d(b, d) = d(a, d) + d(b, c) \geq d(a, b) + d(c, d)$  for all  $a, b, c, d \in S$ .

This lemma essentially follows from the result by Dobson (1974) and others mentioned in Section 3*b*. The following theorem provides a solution to Problem (H).

*Theorem 2.* Let  $(S, H)$  be an  $H$ -structure; the following two conditions are equivalent:

- (i)  $(S, H)$  is  $t$ -realizable;
- (ii) whenever  $aeH^*cd$  or  $aeHbc$ , then  $abH^*cd$  implies  $beH^*cd$  for all  $a, b, c, d, e \in S$ .

*Remark 3.* It is instructive to consider the relationship between this result and the representation of a similarity index by an additive tree metric (Section 3*b*). Again the available algorithms provide weights for the edges of the tree in an optimal way but they do not address the existence problem. On the other hand, the additive tree metric following from  $t$ -realizability is unique only up to transformations that leave invariant the inequality stated in Lemma 2. Given a dissimilarity index  $\delta$ , Theorem 2 may be used to check the existence problem in the following way: define a relation  $t$  by

$$abH^*cd \text{ iff } \delta(a, b) + \delta(c, d) < \min \{ \delta(a, c) + \delta(b, d), \delta(a, d) + \delta(b, c) \}.$$

Then Condition (ii) of Theorem 2, formulated in terms of the dissimilarity index, gives a necessary and sufficient criterion for  $t$ -realizability of the index.

*Remark 4.* The relation between  $t$ - and  $rt$ -realizability is also worth some attention. It is not difficult to show that  $t$ -realizability is less restrictive than  $rt$ -realizability in the following sense: given a set of data in form of the relation  $T$  violating Condition (ii) of Theorem 1, it may nonetheless be possible to define a relation  $H$  in terms of  $T$  satisfying Condition (ii) of Theorem 2. Thus, while there is no rooted tree that could represent the  $T$  relation, there might be an unrooted tree representing an appropriately defined  $H$  relation. Rather than developing this in an abstract and general setting we present an example of a dissimilarity index  $\delta$  demonstrating this point.

*Example 3.* Suppose we have an index  $\delta$  such that

$$\delta(a, d) = 8 < \delta(a, b) = 9 < \delta(b, c) = \delta(a, c) = 10 < \delta(b, d) = 11 < \delta(c, d) = 13$$

for some  $a, b, c, d$ . The reader may check that the corresponding  $T$ -structure is not  $rt$ -realizable (where  $T$  is defined in the way indicated in Remark 2). On the other hand, we have

$$\delta(a, d) + \delta(b, c) < \min \{ \delta(a, c) + \delta(b, d), \delta(a, b) + \delta(c, d) \},$$

i.e. the corresponding  $H$ -structure is  $t$ -realizable (where  $H$  is defined in the way indicated in Remark 3).

5. Proofs of Theorems 1 and 2

Necessity of Condition (ii) in Theorems 1 and 2, respectively, can be checked easily and is left to the reader.

Theorem 1.

We show that (ii) in Theorem 1 implies (ii) in Lemma 1. Define subsets of  $A$  for  $a, b \in A, a \neq b$

$$N(a, b) = \{x \in A \mid abT^*x\}$$

and

$$d(a, b) = \begin{cases} [ |N(a, b)| + 1 ]^{-1} & a \neq b \\ 0 & a = b. \end{cases}$$

We claim that  $d$  is an ultrametric with the property stated in (ii) of Lemma 1:  $d(a, b) = 0$  iff  $a = b$  and  $d(a, b) = d(b, a)$  for any  $a, b \in S$  are obvious from the definition. For the ultrametric inequality, let us suppose without loss of generality that  $abT^*c$  holds. We show that

$$d(a, c) = d(b, c) \geq d(a, b). \tag{D}$$

This is trivial for  $a = b = c$ . If  $a = b \neq c$ , (D) follows from  $d(a, b) = 0$  and the non-negativity of  $d$ . If all three elements are distinct,  $acT^*x$  implies  $bcT^*x$  and vice versa by (ii) of Theorem 1 for any  $x$ , so that  $N(a, c) = N(b, c)$  which establishes the equality in (D); moreover,  $acT^*x$  and  $bcT^*x$  imply  $abT^*x$  by (ii), thus  $N(a, c) \subset N(a, b)$ , yielding the inequality in (D). This also settles the only-if part of Condition (ii) in Lemma 1. For the 'if' part, assume  $acT^*b$  (if  $bcT^*a$ , a symmetric argument follows). For any  $x$  with  $abT^*x$ , (ii) of Theorem 1 implies  $acT^*x$ , i.e.  $N(a, c) \supseteq N(a, b)$ . But then actually  $|N(a, c)| > |N(a, b)|$  holds because  $b \in N(a, c) \setminus N(a, b)$ . This completes the proof. It is not difficult to see from this proof that the representing ultrametric  $d$  is unique only up to strictly monotone transformations.

Theorem 2.

The same method of proof via Lemma 2 should be feasible in this case, too, but it seems to be much more cumbersome. We found it more convenient to use a method characterizing trees in terms of certain partitions of the base set  $S$ , the main tool of which is due to Buneman (1971). Suppose  $(S, H)$  is an  $H$ -structure. A non-trivial partition  $\sigma$  of a finite set  $S$  such that  $\sigma = \{S^0, S^1\}$ ; i.e.  $S^0 \cap S^1 = \emptyset, S^0 \cup S^1 = S$  and  $S^0, S^1 \neq \emptyset$ , is called a *split* of  $S$ .  $\sigma$  is called a *split* of  $(S, H)$ , or simply *H-split* if for all  $a, b \in S^0$  and all  $c, d \in S^1$   $abH^*cd$  holds.

Splits share the following property with edges of a tree: they break up their corresponding sets of elements or nodes into two subsets. Two splits of a set  $S, \sigma_1 = \{S_1^0, S_1^1\}$  and  $\sigma_2 = \{S_2^0, S_2^1\}$  are called *compatible* if at least one of the following sets is empty:  $S_1^0 \cap S_2^0, S_1^1 \cap S_2^0, S_1^0 \cap S_2^1, S_1^1 \cap S_2^1$ . In order for splits to be interpretable as the edges of a to be constructed tree, splits must obviously be pairwise compatible.

**Lemma 3.** Any two splits of an  $H$ -structure are compatible.

*Proof.* Let  $\sigma_1 = \{S_1^0, S_1^1\}$  and  $\sigma_2 = \{S_2^0, S_2^1\}$  be splits of  $(S, H)$ ; if  $\sigma_1 = \sigma_2$  they are compatible because  $S_1^0 \cap S_1^1 = \emptyset$ ; if  $\sigma_1 \neq \sigma_2$  suppose all intersections are non-empty, i.e. there exist  $a \in S_1^0 \cap S_2^0, b \in S_1^0 \cap S_2^1, c \in S_1^1 \cap S_2^0$ , and  $d \in S_1^1 \cap S_2^1$ ; then these elements are pairwise distinct and by definition of  $H$ -splits, we have  $abH^*cd$  and  $acH^*bd$  which is impossible by definition of  $H^*$ .

The following lemma by Buneman (1971) tells us that one gets a tree from any set of pairwise compatible splits.

**Lemma 4.** (Buneman, 1971) Any set  $E = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$  of distinct, pairwise compatible splits of a finite set  $S$  constitutes a tree  $T$  if  $T$  is defined in the following way:

- (i) the set of edges of  $T$  is identified with  $E$ ;
- (ii) the set of nodes of  $T$  is identified with the set

$$N(E) = \{\{S_1^{i_1}, S_2^{i_2}, \dots, S_n^{i_n}\} \mid S_k^{i_k} \in \sigma_k, i_k = 1, 0$$

$$\text{and } S_k^{i_k} \cap S_m^{i_m} \neq \emptyset \text{ for } k, m = 1, \dots, n\};$$

- (iii) and edge  $\sigma_k$  is incident with two nodes iff these are, in some order,  $\{S_1^{i_1}, \dots, S_k^0, \dots, S_n^{i_n}\}$  and  $\{S_1^{i_1}, \dots, S_k^1, \dots, S_n^{i_n}\}$ .

To establish  $t$ -realizability of an  $H$ -structure  $(S, H)$  satisfying Condition (ii) of the theorem we may thus proceed as follows: first, determine the set of all distinct  $H$ -splits of the set  $S$ ; since  $S$  is finite, we can, in principle, decide for every subset of  $S$  if, together with its set complement, it constitutes an  $H$ -split. By Lemma 3, these  $H$ -splits are pairwise compatible. Second, by Lemma 4 we know there exists a tree with nodes and edges defined as above. Third, we have to show that there is a one-to-one correspondence  $f$  between the elements of  $S$  and the terminal nodes of the tree such that all terminal nodes have the appropriate configuration. To establish this third step and thus complete the proof we need some further lemmas.

**Lemma 5.** (Buneman, 1971) For a tree  $T$  defined as in Lemma 4, the degree of a node is equal to the number of minimal elements the node contains, where minimality refers to set inclusion. Moreover, for a terminal node the intersection of its elements is non-empty.

**Lemma 6.** Let  $(S, H)$  be an  $H$ -structure,  $E$  the set of all pairwise compatible splits of  $(S, H)$ , and  $T$  the tree that exists according to Lemma 4; then there is a one-to-one correspondence  $f$  between the elements of  $S$  and the terminal nodes of  $T$ .

*Proof.* Suppose  $E = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$ ; for  $a \in S$ ,  $\{\{a\}, S \setminus \{a\}\} \in E$  by (i)–(iii) of Definition 3; we may suppose, for convenience,  $a \in S_i^0 (i = 1, \dots, n)$ ; then, obviously,  $f(a) = \{S_1^0, \dots, \{a\}, \dots, S_n^0\}$  is a node; suppose among the members of  $f(a)$  there is a minimal one different from  $\{a\}$ , say  $S_k^0$ ; then we have  $\{a\} \subset S_k^0$ , violating minimality of  $S_k^0$ ; thus, by the previous lemma,  $f(a)$  is a terminal node; for  $b \in S, a \neq b$ , if  $f(b)$  is constructed in a similar manner,  $f(a)$  and  $f(b)$  would at least differ on  $\sigma = \{\{a\}, S \setminus \{a\}\}$ , i.e.  $f(a) \neq f(b)$ , and  $f$  is injective; conversely, for  $a'$  a terminal node of

$T, a' = \{S_1^{i_1}, \dots, S_n^{i_n}\}$ , let us take some  $S_k^0 \in a'$  to be the one minimal member; suppose  $S_k^0$  contains more than one element; take any  $b \in S_k^0$  and consider the  $H$ -split  $\{\{b\}, S \setminus \{b\}\}$ .  $\{b\}$  cannot be a member of  $a'$  because minimality of  $S_k^0$  would be violated; thus  $S \setminus \{b\} \in a'$ . Then  $b$  is not element of the intersection of all members of  $a'$ , i.e.  $b \notin \bigcap_{i=1}^n S_i^{i_i}$ ; this implies  $\bigcap_{i=1}^n S_i^{i_i} = \emptyset$ , contrary to the last statement in Lemma 5. Thus  $S_k^0$  is a one-element-set,  $S_k^0 = \{a\}$  say. To summarize, for any terminal node  $a'$  we may find an element  $a$  of  $S$  such that  $f(a) = a'$ , i.e.  $f$  is surjective and the one-to-one correspondence is established.

The following lemma is essential for showing that the terminal nodes have the appropriate configuration.

*Lemma 7.* Let  $(S, H)$  be an  $H$ -structure satisfying Condition (ii) of Theorem 2; then for any  $a, b, c, d \in S$  such that  $abH^*cd$  there is an  $H$ -split  $\{S^0, S^1\}$  with  $a, b \in S^0$  and  $c, d \in S^1$ .

*Proof.* In what follows, we distinguish between the two propositions contained in Condition (ii) of Theorem 2; if the premise  $aeH^*cd$  is valid we refer to Condition (ii) as Condition (I), if  $aeHbc$  is valid we call it Condition (II). If  $a = b$ ,  $\{\{a\}, S \setminus \{a\}\}$  is an  $H$ -split of the desired form; if  $c = d$ , then  $\{\{c\}, S \setminus \{c\}\}$  as well. Otherwise, we define a split by

$$S^1 = \{e \in S \mid abH^*ce\} \text{ and } S^0 = \{e \in S \mid acH^*be \text{ or } aeH^*bc \text{ or } abceH\}.$$

Thus, if  $abH^*cd$ , any other element  $e$  is contained either in  $S^1$  or in  $S^0$  and  $a, b \in S^0$  and  $c, d \in S^1$ . For  $\{S^0, S^1\}$  to be an  $H$ -split we have to show that, for any  $u, v \in S^0$  and any  $x, y \in S^1$ ,  $uvH^*xy$ .

For  $x, y$  we have  $abH^*cx$  and  $abH^*cy$ ; for  $u, v$  we must distinguish six cases: (a)  $acH^*bu$  &  $acH^*bv$ , (b)  $acH^*bu$  &  $avH^*bc$ , (c)  $acH^*bu$  &  $abcvH$ , (d)  $abcvH$  &  $abcvH$ , (e)  $auH^*bc$  &  $avH^*bc$ , (f)  $auH^*bc$  &  $abcvH$ ;

(a) :  $caH^*bv$  &  $cyH^*ab$  imply  $ayH^*bv$  by (II);  $caH^*bu$  &  $cyH^*ab$  imply  $ayH^*bu$  by (II);  $abH^*cx$  &  $abH^*cy$  imply  $abH^*xy$  by (I);  $baH^*xy$  &  $buH^*ay$  imply  $avH^*xy$ ,  $baH^*xy$  &  $buH^*ay$  imply  $auH^*xy$ , by (II); thus  $uvH^*xy$  by (I);

(b) :  $caH^*bu$  &  $cyH^*ab$  imply  $ayH^*bu$  by (II);  $abH^*cx$  &  $avH^*bc$  imply  $bvH^*cx$  by (II), thus  $avH^*cx$  by (I);  $abH^*cy$  &  $avH^*bc$  imply  $bvH^*cy$  by (II), thus  $avH^*cy$  by (I); thus  $avH^*xy$  by (I);  $baH^*xy$  &  $buH^*ay$  imply  $auH^*xy$  by (II); then  $uvH^*xy$  &  $auH^*xy$  imply  $uvH^*xy$  by (I);

(c) :  $acH^*bu$  &  $avHbc$  imply  $cvH^*bu$  by (II);  $abH^*cx$  &  $avHbc$  imply  $bvH^*cx$ ,  $abH^*cy$  &  $avHbc$  imply  $bvH^*cy$ , by (II); then  $buH^*vc$  implies  $vuH^*cx$  &  $vuH^*cy$  by (II), thus  $uvH^*xy$  by (I);

(d) :  $abH^*cx$  &  $avHbc$  imply  $bvH^*cx$ ,  $abH^*cy$  &  $avHbc$  imply  $bvH^*cy$ , by (II); thus  $bvH^*xy$  by (I);  $abH^*cx$  &  $auHbc$  imply  $buH^*cx$ ,  $abH^*cy$  &  $auHbc$  imply  $buH^*cy$ , by (II); thus  $buH^*xy$  by (I); then  $uvH^*xy$  follows from (I);

(e) :  $uvH^*xy$  follows from (a) by interchanging  $a$  and  $b$ ;

(f) :  $uvH^*xy$  follows from (c) by interchanging  $a$  and  $b$ ; this completes the proof.

Now, suppose we have  $a, b, c, d \in S$  such that  $abH^*cd$ . According to the definition of  $t$ -realizability, we have to show that  $(f(a)f(b))^*(f(c)f(d))$  holds for the terminal nodes. By the previous lemma, there is an  $H$ -split  $\sigma = \{S^0, S^1\}$  such that  $a, b \in S^0$  and  $c, d \in S^1$ . Thus, by Lemma 4 and Lemma 6,  $\sigma$  is an edge in the tree which is incident

with two nodes  $p_1, p_2$ , let us say

$$p_1 = \{S_1^{i_1}, S_2^{i_2}, \dots, S_1^0, \dots, S_n^{i_n}\} \text{ and } p_2 = \{S_1^{i_1}, S_2^{i_2}, \dots, S_1^1, \dots, S_n^{i_n}\},$$

i.e.  $p_1$  and  $p_2$  differ only on  $\sigma$ . Since in a tree, any two nodes are joined by a path, there are paths from  $p_1$  to  $f(a)$ , from  $p_1$  to  $f(b)$ , from  $p_2$  to  $f(c)$ , and from  $p_2$  to  $f(d)$ , respectively. Obviously,  $(f(a)f(b)) * (f(c)f(d))$  holds if  $\sigma$  is not contained in any of these paths. To this end, we refer to Buneman's (1971, Lemma 1) construction of a path between any two nodes of the tree. For example, an alternating sequence of nodes and edges  $q_1, \{q_1, q_2\}, q_2, \dots, \{q_{m-1}, q_m\}, q_m$  is a path from  $p_1$  to  $f(a)$  if  $q_1 = p_1$ ,  $q_m = f(a)$  and if  $q_j$  has one element more in common with  $f(a)$  than  $q_{j-1}$ , for  $j = 2, \dots, m$ . By Lemma 6, we have  $\{a\} \in f(a)$ ; but then  $S^0 \in f(a)$ , too, since  $\{a\} \cap S^1 = \emptyset$ . Because  $S^0 \in p_1$ , in the path from  $p_1$  to  $f(a)$ ,  $S^0$  will never be replaced by  $S^1$ , i.e.  $\sigma$  is not contained in the path from  $p_1$  to  $f(a)$ . The analogous argument follows for the paths from  $p_1$  to  $f(b)$  and from  $p_2$  to  $f(c)$  and  $f(d)$ . The case  $abcdH$  can be dealt with similarly and is left to the reader. This ends the proof of Theorem 2.

## 6. Concluding remarks

While Theorems 1 and 2 give a complete answer to the representation problem for  $T$ - and  $H$ -structures, respectively, this is somewhat less than one might desire from a pragmatic point of view. In fact, if the set of objects to be judged is large, the number of triples or quadruples of objects that have to be presented to the subject may be prohibitive. On the other hand, it is obvious that given an  $rt$ -realizable  $T$ -structure (or a  $t$ -realizable  $H$ -structure) not all 3-element (or 4-element) configurations have to be known for constructing the tree: some of them can be deduced from others, for example via the axiom stated in Theorem 1 (or Theorem 2). The question then is if there is some general way to characterize a minimal set of configurations sufficient for the tree to be determined. We do not have an answer to this but our conjecture is rather to the negative.

A preliminary experiment on verbal meaning, described in full in Schulze & Colonius (1979), provides an illustration of the methods suggested in this paper. Six subjects were presented with 3- and 4-element subsets of verbs of judging taken from a set of verbs investigated in Fillenbaum & Rapoport (1971). Three lists of seven verbs each were used. For example, one list consisted of *admit, confess, defend, excuse, forgive, justify, pardon*. Subjects had to decide, for every 3-element and 4-element subset of each list, which  $T$ - and which  $H$ -relation, respectively, was appropriate. For each subject, the data for each list were then analysed by tree construction algorithms that are variants of the single-link algorithm (e.g. Jardine & Sibson, 1971) and Sattath & Tversky's (1977) algorithm for additive trees. The number of  $T$ - or  $H$ -relations correctly represented in the tree was taken as an indicator of consistency for the subjects' judgements. The data varied between perfect fit and about 66 per cent correctly represented relations.

While strictly speaking a tree representation is warranted only if the axioms are perfectly satisfied, the results suggest that, at least, the procedure is a meaningful one for the subjects. Obviously, a statistical evaluation or a simulation study is needed. In general the number of incorrectly represented relations was higher for  $T$ -structures than for  $H$ -structures and Schulze & Colonius (1979) discuss the possibility of interpreting the edges of the two sorts of tree as semantic features.

## Acknowledgements

We thank K. F. Wender for a critical reading of the manuscript. This research was in part supported by grant Co 94/1 from Deutsche Forschungsgemeinschaft to the first author.

## References

- Beals, R., Krantz, D. H. & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, **75**, 127–142.
- Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In: F. R. Hodson, D. G. Kendall & P. Tautu (eds), *Mathematics in the Archaeological and Historical Sciences*. Edinburgh: Edinburgh University Press.
- Buneman, P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory*, **17** (B), 48–50.
- Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika*, **41**, 439–463.
- Carroll, J. D. & Pruzansky, S. (1975). Fitting of hierarchical tree structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares. *Proceedings of the U.S.–Japan Seminar: Theory, Methods, and Applications of Multidimensional Scaling and Related Techniques*. University of California at San Diego.
- Colonius, H. & Schulze, H.-H. (1979). Repräsentation nicht-numerischer Ähnlichkeitsdaten durch Baumstrukturen. *Psychologische Beiträge*, **21**, 98–111.
- Cunningham, J. P. (1974). Finding an optimal tree realization of a proximity matrix. Paper presented at the Mathematical Psychology Meeting, Ann Arbor, Michigan.
- Cunningham, J. P. (1978). Free trees and bidirectional trees as representations of psychological distance. *Journal of Mathematical Psychology*, **17**, 165–188.
- Defays, D. (1979). Tree representations of ternary relations. *Journal of Mathematical Psychology*, **19**, 208–219.
- Dobson, J. (1974). Unrooted trees for numerical taxonomy. *Journal of Applied Probability*, **11**, 32–42.
- Fillenbaum, S. & Rapoport, A. (1971). *Structures in the Subjective Lexicon*. New York: Academic Press.
- Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association*, **62**, 1140–1158.
- Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.
- Holman, E. W. (1972). The relation between hierarchical and Euclidean models for psychological distances. *Psychometrika*, **37**, 417–423.
- Hubert, L. (1972). Some extensions of Johnson's hierarchical clustering algorithms. *Psychometrika*, **37**, 261–274.
- Hubert, L. (1973). Monotone invariant clustering procedures. *Psychometrika*, **38**, 47–62.
- Jardine, N. & Sibson, R. (1971). *Mathematical Taxonomy*. New York: Wiley.
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.
- Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). *Foundations of Measurement*, Vol. 1. New York: Academic Press.
- Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The relationship between similarity and spatial density. *Psychological Review*, **85**, 445–463.
- Lerman, I. C. (1970). *Les Bases de la Classification Automatique*. Paris: Gauthier-Villars.
- Luce, R. D. & Galanter, E. (1963). Psychophysical scaling. In R. D. Luce, R. R. Bush & E. Galanter (eds), *Handbook of Mathematical Psychology*, vol. 1. New York: Wiley.
- Müller, G. A. (1969). A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, **6**, 169–191.
- Möbus, C. (1979). Zur Analyse nichtsymmetrischer Ähnlichkeitsurteile: Ein dimensionales Driftmodell, eine Vergleichshypothese, Tversky's Kontrastmodell und seine Fokushypothese. *Archiv für Psychologie*, **131**, 105–136.
- Patrinos, A. N. & Hakimi, S. L. (1972). The distance matrix of a graph and its tree realization. *Quarterly Journal of Applied Mathematics*, **30**, 255–269.
- Sattath, S. & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, **42**, 319–347.
- Schulze, H.-H. & Colonius, H. (1979). Eine neue Methode zur Erforschung des Subjektiven Lexikons. In L. H. Eckensberger (ed.), *Bericht über den 31. Kongress der Deutschen Gesellschaft für Psychologie*. Göttingen: Hogrefe.
- Sholander, M. (1952). Trees, lattices, order and betweenness. *Proceedings of the American Mathematical Society*, **3**, 369–381.

- Simoes-Pereira, J. M. S. (1969). A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, **6**, 303-310.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

*Received 22 October 1980; revised version received 25 March 1981*

Requests for reprints should be addressed to Hans Colonius, Technische Universität Braunschweig, Institut für Psychologie, Spielmannstrasse 19, D-3300 Braunschweig, West Germany or to Hans-Henning Schulze, Universität Marburg/Lahn, Fachbereich Psychologie, Gutenbergstrasse 18, D-3550 Marburg, West Germany.