Europ. J. Combinatorics (1996) **17**, 191–208

Tree–Tree Matrices and Other Combinatorial Problems from Taxonomy

MICHIEL HAZEWINKEL

Let A be a bipartite graph between two sets D and T. Then A defines, via Hamming distance, metrics on both T and D. The question is studied which pairs of metric spaces can arise this way. If both spaces are trivial, the matrix A comes from a Hadamard matrix or is a BIBD. The second question studied is how A can be used to transfer (classification) information from one of the two sets to the other. These problems find their origin in mathematical taxonomy.

1. INTRODUCTION

A great deal of the literature in mathematical taxonomy focuses on clustering; i.e. summarizing the information present in a metric or dissimilarity on a set X by means of a classification tree or something similar.

Here, we focus directly on the situation that one finds in the taxonomic problems of scientific disciplines. Often, the data are in the form of a collection of documents and a collection of key words and key phrases that is supposed to be sufficiently rich to describe (up to a point) the scientific field in question. Here, I am not concerned with how such a control list or thesaurus is generated.

The data are thus in the form of a bipartite graph A (or, equivalently, a relation) between two sets, a set D (of documents) and a set T (of terms). The bipartite graph A tells us which terms occur in which documents.

These data can be used to define a metric space structure on both T and D by means of Hamming distance—the distance between two terms is the number of documents in which one term occurs and the other not. A first question that arises is what pairs of discrete metric spaces can arise this way. For trivial metric space structures on both T and D it turns out that A must be very regular (a Hadamard matrix, a Hadamard matrix minus one row or column, or a symmetric BIBD). Section 2 below is devoted to some results in this direction.

It arises frequently in practice that on one of the spaces T or D there is available metric information coming from other sources. For instance, in the case of a body of scientific literature, co-citation analysis can be used to define 'research clusters' or 'research fronts' of strongly linked clusters of documents. The question then arises how to transfer such information from one of the sets, in this case D, to the other by means of the bipartite graph between them. This matter is discussed in Section 3.

Finally, in Section 4 some recent ideas and results concerning metrics on the space of all metrics on a given finite set are summarized. These things are fundamental for addressing the question of finding, for instance, the best approximative ultrametric to a given metric or dissimilarity.

0195-6698/96/020191 + 18 $18.00/0

2. The Tree–Tree Problem

2.1. Definition of the problem. As indicated above, we shall take as the basic available data a bipartite graph A between terms and documents. Or, equivalently, A is a 0–1 matrix with the set of terms as column indices and the set of documents as row indices. A 1 at spot (i, j) means that the term j occurs in the document i. These data define two metric spaces as follows:
(i) The column space of A, $cs(A) = T(A)$. As a set, this is the set of terms. The distance between two terms t, t' is the Hamming distance between the corresponding columns, i.e. the number of row indices with different entries at spots t and t'.
(ii) The row space $rs(A) = D(A)$ of A. As a set, this is the set of documents. The distance between two documents d, d' is the Hamming distance between the corresponding rows, i.e. the number of column indices with different entries in rows d and d'.

This leads immediately to a number of natural basic questions, such as:
(i) Which metric spaces can arise as a $T(A)$ or a $D(A)$?
(ii) To what extent is A determined by $D(A)$ and $T(A)$?
(iii) Which pairs of metric spaces D, T can arise from a 0–1 matrix A?

In this paper I concentrate on the last question. Trees and classification schemes (which are special kinds of treess) are ubiquitous in (mathematical taxonomy). Thus it is important and natural to start with the question when both the column and row spaces of a 0–1 matrix are trees or related to trees.

2.1.1. Definitions. A *tree* is an unoriented connected graph such that there is a unique path between any two given vertices. A *leaf* of a tree is a vertex with just one edge incident with it. An *edge weighted tree* is a tree with each edge labelled with a real number > 0. An example is shown in Figure 1. The *distance* between two vertices of an edge weighted tree is the sum of the weights of the edges occurring in the unique path between those vertices. This defines a metric on the set of vertices (and on any subset, particularly the set of leafs). A *rooted tree* is a tree with a special, selected vertex called the root. An *hierarchical tree* is a rooted edge weighted tree such that each leaf has the same distance to the root.

Figure 1 is not a hierarchical tree but Figures 2 and 3 are. In these figures and those below an unlabelled edge is supposed to have weight 1. An hierarchical tree defines an ultrametric on its set of leaves: and, inversely, [6, 11], every finite ultrametric space arises that way. By inserting, if necessary, extra vertices of valency two (as was done in Figure 3), each ultrametric space arises as the space of leafs of some 'hierarchically organized' tree like the one in Figure 3 in which, for each vertex, all the edges pointing towards the leafs have the same weight.

It is rather easy to see that each edge weighted tree with integer weights can be realized as a $T(A)$ (or a $D(A)$). Things are rather different if both $T(A)$ and $D(A)$ are

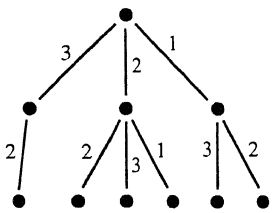

Figure 1

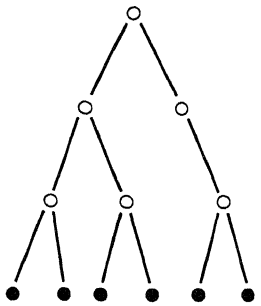

FIGURE 2

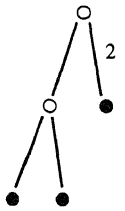

FIGURE 3

required to be trees or tree-like (definition below). This appears to be quite difficult to realize. In particular, it seems difficult to realize a pair of spaces that are not (nearly) isomorphic. This is, roughly, what I like to call the *tree–tree problem.* To make the problem more precise, let us make the following definition.

2.1.2. DEFINITION. A finite metric space (X, m) is *tree-like* if it is isometric to a subspace of the vertex metric space defined by an edge weighted tree.

2.1.3. TREE–TREE PROBLEM. Which pairs of tree-like spaces can be realized by a 0–1 matrix?

I view these 0–1 matrices as some sort of generalized hierarchical block designs. The reason for that is Theorem 2.2.5 below.

Related to the tree–tree problem is the problem of finding a good characterization of those matrices for which both the column metric space and the row metric space are tree-like.

Of course, tree-like metric spaces are characterized by the so-called four-point condition.

2.1.4. FOUR-POINT CONDITION. A finite metric space (X, m) is tree-like iff, for all not necessarily distinct four points $a_1, a_2, b_1, b_2 \in X$,

$$m(a_1, a_2) + m(b_1, b_2) \leq \max\{m(a_1, b_1) + m(a_2, b_2), m(a_1, b_2) + m(a_2, b_1)\}. \quad (1)$$

This gives a necessary and sufficient condition for a 0–1 matrix A to yield a pair of tree-like spaces—but certainly a very inelegant and unsatisfying one.

2.1.5. ULTRAMETRIC TREE–TREE PROBLEM. Which pairs of ultrametric spaces can be realized by a 0–1 matrix?

2.1.6. COMPLETE TREE–TREE PROBLEM. Which (complete) pairs of edge weighted trees can be realized by a 0–1 matrix?

2.2. Trivial tree–trivial tree matrices and BIBDs. Let us start with some very simple

examples in which the column and row metric spaces are 'trivial' in the sense of the definition below.

2.2.1. DEFINITION. A *trivial discrete metric space* (X, m) is a metric space such that there is a positive number a such that

$$m(x, y) = a \qquad \text{for all } x \neq y \text{ in } X$$

(and of course $m(x, x) = 0$ for all $x \in X$).

2.2.2. EXAMPLE: HADAMARD MATRICES. A *Hadamard matrix* is an $n \times n$ matrix H with entries 1, -1 such that

$$HH^{\mathrm{T}} = nI_n.$$

It follows that also $HH^{\mathrm{T}} = nI_n$ (and that n is even, $n = 2k$). It is immediate from these two properties that for each two rows there are precisely k entries that are equal and k entries that are unequal—and similarly for the columns. Let A be the matrix obtained from H by replacing each -1 with 0. Then both the column and the row space of A are the trivial metric space of $n = 2k$ points with distance k.

2.2.3. EXAMPLE: HADAMARD MATRICES WITH ONE ROW OR COLUMN DELETED. Now let H be a Hadamard matrix for which one row or column consists entirely of $+1$'s or entirely of -1's. Delete that row or column. Again replace -1 with 0 everywhere. The result is a 0–1 matrix with trivial column and trivial row space of sizes n and $n - 1$ and distance $n/2$.

Not every Hadamard matrix has such a column or row. However, if D is diagonal with each diagonal element equal to 1 or -1, and if H is an Hadamard matrix, then so are HD and DH. So it is easy to modify a Hadamard matrix so as to obtain one with such a column or row.

2.2.4. EXAMPLE: SYMMETRIC BIBDs. A balanced incomplete block design (BIBD) is a zero–one matrix A such that each row has the same number, r, of 1's each column has the same number, s, of 1's, and further, for each pair of column indices $i \neq j$ there are precisely λ rows which have a 1 at both locations i and j. This last condition is the same as saying that each two different columns have λ common 1's.

A BIBD is symmetric if A is square. It then follows that $r = s$ and that each two distinct rows also have λ common 1's (see, e.g., [3]).

It follows immediately that the row space and the column space of a symmetric BIBD are trivial metric spaces with n points and distance $2(r - \lambda)$.

2.2.5. THEOREM. *Let A be an $m \times n$ zero–one matrix such that both the column space and the row space are trivial. Then A is one of the Examples* 2.2.2–2.2.4; *i.e. A 'is' a Hadamard matrix, a Hadamard matrix with one constant row or column deleted, or it is a symmetric BIBD.*

Let B be the matrix obtained from A by replacing each 0 with -1. Then the trivial column and row space condition on A translates for B into the statement that the rows

of B form a system of m length n vectors, all of whom make the same angle with one another, and the columns form a system of n vectors of length m that also all make the same angle with one another.

2.2.6. PROOF OF THEOREM 2.2.5. Let B be the $m \times n$ matrix obtained from A by replacing each 0 with -1. Let d be the distance between each two distinct rows of B (or A) and e the distance between each two distinct columns. Then

$$BB^{\mathrm{T}} = \begin{pmatrix} n & p & \dots & p \\ p & n & \ddots & \vdots \\ \vdots & \ddots & \ddots & p \\ p & \dots & p & n \end{pmatrix}, \qquad p = n - 2d, \tag{2}$$

$$B^{\mathrm{T}}B = \begin{pmatrix} m & q & \dots & q \\ q & m & \ddots & \vdots \\ \vdots & \ddots & \ddots & q \\ q & \dots & q & m \end{pmatrix}, \qquad q = m - 2e. \tag{3}$$

Interchanging rows and columns if necessary, we can assume that $m \geqslant n$. By the lemma below, the $m \times m$ matrix BB^{T} is non-singular except when $p = n$ or $n = -(m-1)p$. The first case cannot happen because $d > 0$. The second case can happen. Then, because $m \geqslant n$, $n = m - 1$ and $p = -1$. Now, add one column of 1's (or -1's) to B to obtain an $m \times m$ matrix $\bar{B}$. It follows that $\bar{B}$ is a Hadamard matrix. Therefore, in this case, we are dealing with an instance of Example 2.2.3.

Continuing, we can assume that BB^{T} is non-singular and hence that

$$n = m. \tag{4}$$

Let $c_1, c_2, \dots, c_n$ be the column sums of B, and let $r_1, r_2, \dots, r_n$ be the row sums of B. Multiply (2) with B on the right, to obtain

$$BB^{\mathrm{T}}B = (n-p)B + p\begin{pmatrix} c_1 & \dots & c_n \\ \vdots & & \vdots \\ c_1 & \dots & c_n \end{pmatrix}, \tag{5}$$

and, using $n = m$, multiply (3) on the left with B, to obtain

$$BB^{\mathrm{T}}B = (n-q)B + q\begin{pmatrix} r_1 & \dots & r_1 \\ \vdots & & \vdots \\ r_n & \dots & r_n \end{pmatrix}. \tag{6}$$

Subtracting (6) from (5), we see that the matrix $(q-p)B$ is equal to a matrix of rank $\leqslant 2$. If $n = m \geqslant 3$ this is only possible if $p = q$ and hence $e = d$, because B is invertible.

Now, there are two cases:

(i) Case 1; $p = q = 0$. Then, B is a Hadamard matrix by (2).

(ii) Case 2; $p = q \neq 0$. Then, it follows from (5) and (6) that

$$c_1 = \dots = c_n = r_1 = \dots = r_n,$$

so that A is a symmetric BIBD with $r = (n + c_1)/2$ entries 1 in each column and row and $\lambda = (n + c_1 - d)/2$.

This proves the theorem for $n, m \geq 3$; it is trivial to deal with the remaining cases. □

2.2.7. LEMMA. *The determinant of the $m \times m$ matrix in* (2) *is equal to*

$$\det\begin{pmatrix} n & p & \dots & p \\ p & n & \ddots & \vdots \\ \vdots & \ddots & \ddots & p \\ p & \dots & p & n \end{pmatrix} = (n-p)^{m-1}(n+(m-1)p).$$

PROOF. The proof is straightforward. □

Using similar but more complicated arguments, one can show that if A is an $m \times n$ zero–one matrix such that each two distinct rows have exactly μ ones in common and each two distinct columns have exactly λ ones in common, then A is a symmetric BIBD. Interpreting the column indices of A as points and the row indices of A as lines, this gives the following [8].

2.2.8. THEOREM. *Let X be a finite set (of points), with a system of subsets called lines. Let there be n points and m lines. Suppose that lines distinguish points (i.e. no two distinct points have the same set of lines through them) and points distinguish lines, and that:*
(i) *each two distinct lines meet in μ points; and*
(ii) *through each pair of distinct points there pass λ lines.*
Then $n = m$ and $\lambda = \mu$, each line has r points and through each point there pass r lines (where $r(r-1) = \lambda(n-1)$).

This is a special case of a more general result of Röhmel [16]; see also [3, p. 102ff.].

2.3. More examples. Using the various symmetric BIBDs as main building blocks, a variety of examples of tree–tree matrices can be constructed. Here is a small selection. In the illustrations below (and above), the black nodes in a tree make up the tree-like space that is being realized.

2.3.1. EXAMPLE.

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad cs(A) = rs(A) =$$

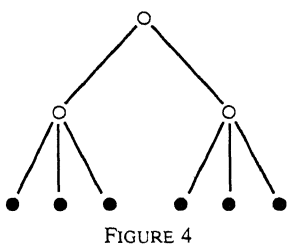

FIGURE 4

2.3.2. EXAMPLE.

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad cs(A) = rs(A) =$$

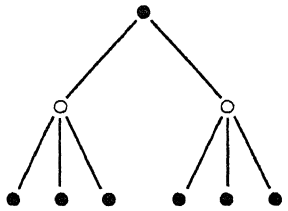

FIGURE 5

2.3.3. EXAMPLE. Let A' be the matrix obtained from that of Example 2.3.2 by deleting the top row. Then the row space of A' is equal to the space of Figure 4, while the column space is that of Figure 5.

2.3.4. EXAMPLE. Let E_n denote the $n \times n$ matrix with every entry equal to 1, let I_n denote the $n \times n$ unit matrix, and let 0 denote whatever size matrix of zeros is appropriate. Then:

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & E_3 & I_3 \\ 0 & I_3 & I_3 \end{pmatrix}, \quad cs(A) = rs(A) =$$

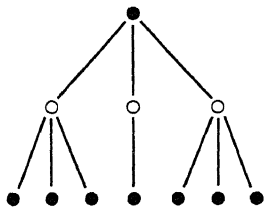

FIGURE 6

2.3.5. EXAMPLE.

$$A = \left(\begin{array}{c|cc} 0 & E_3 & I_3 \\ & I_3 & E_3 \\ \hline E_3 & I_3 & \\ I_3 & E_3 & 0 \end{array}\right), \quad cs(A) = rs(A) =$$

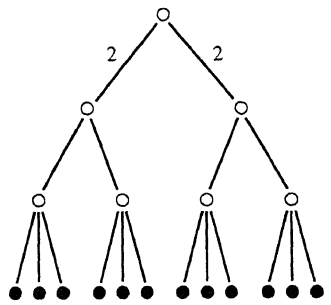

FIGURE 7

2.3.6. EXAMPLE.

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad cs(A) =$$

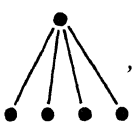

FIGURE 8

$$rs(A) =$$

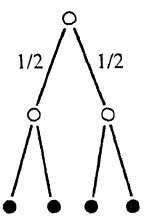

FIGURE 9

2.3.7. REMARK. It is not possible to realize the tree-like space depicted in Figure 10 with a 4×4 matrix. Here, as always, unlabelled edges have weight 1.

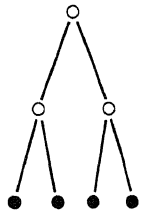

FIGURE 10

2.3.8. EXAMPLE.

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad cs(A) =$$

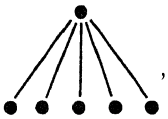

FIGURE 11

$$, \quad rs(A) =$$

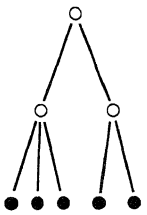

FIGURE 12

2.3.9. EXAMPLE.

$$A = \left(\begin{array}{cc|c} I_4 & E_4 & 0 \\ E_4 & I_4 & \\ \hline & 0 & \begin{matrix} I_3 & E_3 \\ E_3 & I_3 \end{matrix} \end{array}\right), \quad cs(A) = rs(A) =$$

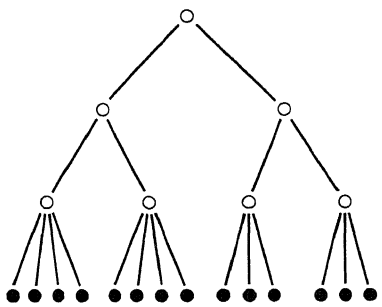

FIGURE 13

2.3.10. REMARK. Call a rooted tree for which the number of edges towards any of its leafs is equal to a, a tree of a levels. Using similar techniques as in the proof of Theorem 2.2.5, there is a great deal that one can say about the zero–one matrices that produce tree-like spaces of level ≤ 2 for their row and column spaces. I intend to return to this in a future paper.

2.4. Tree-like spaces of unbounded height. There is a systematic iterative construction that yields trees and tree-like spaces of any number of levels.

2.4.1. THE ZERO CONSTRUCTION. Let A be a zero–one matrix of size $m \times n$, and suppose that:
(i) all columns have distance $\leq d_c$ to one another;
(ii) all rows have distance $\leq d_r$ to one another;
(iii) the rows of A all have precisely w_r ones;
(iv) the columns of A have precisely w_c ones;
(v) $2w_r > d_r$, $n > w_r > 0$, and $2w_c > d_c$, $0 < w_c < m$; and
(vi) the row space of A and the column space of A are both tree-like.
Now consider the $k \times k$ block matrices

$$A_k^0 = \begin{pmatrix} A & 0 & \cdots & 0 \\ 0 & A & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A \end{pmatrix}. \tag{7}$$

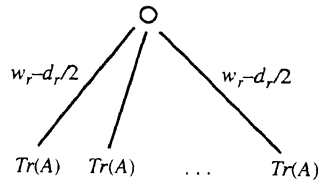

FIGURE 14

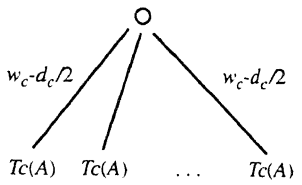

FIGURE 15

Then, if $Tr(A)$ denotes the row tree-like space of A, and $Tc(A)$ is the column tree-like space, then the row space and column space of A_k^0 look like Figures 14 and 15.

Note that

$$d_r(A_k^0) = 2w_r, \qquad d_c(A_k^0) = 2w_c, \qquad w_r(A_k^0) = w_r, \qquad w_c(A_k^0) = w_c. \tag{8}$$

As a rule, if A is just an arbitrary 0–1 matrix with tree-like column and row spaces this construction gives a 0–1 matrix for which neither the row space, nor the column space is tree-like.

2.4.2. THE ONE CONSTRUCTION. A very similar construction can be carried out with ones instead of zeros in (7). Let A be as before in Section 2.4.1, except that the conditions (v) are replaced by

(v') $2(n - w_r) > d_r$, $n > w_r > 0$, and $2(n - w_c) > d_c$, $0 < w_c < m$.

In this case, consider the $k \times k$ block matrices

$$A_k^1 = \begin{pmatrix} A & E & \cdots & E \\ E & A & \ddots & \vdots \\ \vdots & \ddots & \ddots & E \\ E & \cdots & E & A \end{pmatrix},$$

where E is the $m \times n$ matrix consisting completely of ones. Then, the row space and column space of A_k^1 look like Figures 14 and 15, except that $w_r - d_r/2$ and $w_c - d_c/2$ are replaced by $n - w_r - d_r/2$ and $n - w_c - d_c/2$, respectively.

Furthermore,

$$d_r(A_k^1) = 2(n - w_r), \qquad d_c(A_k^1) = 2(n - w_c), \tag{9}$$

$$w_r(A_k^1) = (k-1)n + w_r, \qquad w_c(A_k^1) = (k-1)m + w_c. \tag{10}$$

2.4.3. ITERATING THE CONSTRUCTIONS. It is now easy to check that if A satisfies the conditions for the zero construction, then A_k^0 satisfies the conditions for the one construction, and that if A satisfies the conditions for the one construction, then A_k^1 satisfies the conditions for the zero construction.

Indeed A_k^0 is an $km \times kn$ matrix ($k \geq 2$). So,

$$2(kn - w_r(A_k^0)) = 2kn - 2w_r > 2w_r = d_r(A_k^0),$$

because $k \geq 2$ and $n > w_r$. Also, $0 < w_r = w(A_k^0) < n < kn$. The column conditions are checked similarly, and it follows that the conditions for the one construction are satisfied for A_k^0.

Analogously, A_k^1 is also a $km \times kn$ matrix, and

$$2w_r(A_k^1) = 2(k-1)n + 2w_r > 2(n - w_r) = d_r(A_k^1)$$

because $k \geq 2k$ and $n > w_r$. Also, $0 < (k-1)n + w_r = w_r(A_k^1) < kn$. The column conditions are checked similarly and it follows indeed that A_k^1 satisfies the conditions for the zero construction.

Thus, provided that a starting A can be found, the two constructions can be applied alternatively to yield tree-like spaces with an arbitrary number of levels.

There are many possible starting matrices: e.g. the unit matrix of size 3 or more

satisfies the conditions for the one construction; the matrix $E_n - I_n$, $n \geqslant 3$, satisfies the conditions for the zero construction, and the incidence matrix M of the projective space $\mathbf{P}^2(\mathbf{F}_2)$, i.e.

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

satisfies the conditions for both the zero construction and the one construction.

2.5. Complete trees. To conclude this selection of examples, here are some in which both the row and column space are not just tree-like (i.e. isometric to a subspace of the vertex space of an edge labelled tree) but isometric to the full vertex space of an edge labelled tree.

Let T_k be the following $k \times k$ matrix

$$T_k = \begin{pmatrix} 1 & \cdots & 1 & 0 \\ \vdots & \ddots & \iddots & \vdots \\ 1 & \iddots & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}$$

and let E denote matrices consisting entirely of 1's of the appropriate sizes. Consider the block zero–one matrix

$$A = \begin{pmatrix} 1 & E & E & \cdots & E \\ E & T_{k_1} & E & \cdots & E \\ E & E & T_{k_2} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & E \\ E & E & \cdots & E & T_{k_m} \end{pmatrix}.$$

The column and row spaces of A are both complete trees with just one node of valence >2, as depicted in Figure 16. They consist of one central node of valencey m, from

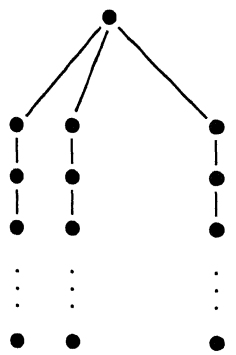

FIGURE 16

which issue m branches with k_i nodes, $i = 1, \ldots, m$. These are the only kind of examples I know for which both the row and column space are complete trees. Modifying the example a bit, the edges can be given arbitrary positive integer weights.

3. Transfer of Metrics

As noted in the Introduction, a bipartite graph connecting terms and documents should also permit the transfer of information about one of the two sets to the other. This section is devoted to aspects of that problem.

3.1. The transfer problem.. Loosely stated, the transfer problem is concerned with the following situation. Let $\Gamma \subset D \times T$ be a bipartite graph (or, equivalently, a relation) between a set D of documents and a set T of terms. Let there be given a metric on D (resp. T). What is the 'best' corresponding metric on T (resp. D).

This sort of situation frequently arises in practice. In the case of the taxonomy of a scientific field for instance, the technique of cocitation analysis (cf. e.g. [5, 20] gives clustering type information on the set D of documents, and the question arises how to transfer this information optimally to classification information on the set of terms.

3.2. The canonical embedding in function space. To discuss various aspects of the transfer problem we first need to describe a canonical embedding of a (discrete) metric space into the space of functions on it.

3.2.1. Definition. Let (X, m) be a (discrete) metric space. let $F(X)$ be the space of all real valued functions on X. Give $F(X)$ the max (or sup) norm metric:

$$m_F(f, g) = \max_{x \in X} |f(x) - g(x)|. \tag{11}$$

The canonical embedding of X into $F(X)$ is given by

$$\alpha_X : X \to F(X), \qquad x \mapsto g_x, \quad g_x(y) = m(x, y) \tag{12}$$

3.2.2. Lemma. *The canonical embedding α_X is an isometry.*

The proof of this lemma is a straightforward application of the triangle inequality.

3.3. The Hausdorff metric. Below, the Hausdorff metric is defined only for finite metric spaces. The definitions extend to more general cases. To do this, replace 'max' by 'sup' and 'min' by 'inf'.

3.3.1. Definition. Let (X, m) be a finite metric space, and let A and B be subsets of X. Then, the Hausdorff distance between the sets A and B is defined as

$$m_{Hd}(A, B) = \max\left\{\max_{a \in A} \min_{b \in B} m(a, b), \max_{b \in B} \min_{a \in A} m(a, b)\right\}. \tag{13}$$

It is well known that the Hausdorff metric is a metric on the set of all subsets of X, i.e.

it satisfies $m_{Hd}(A, B) \geqslant 0$, $m_{Hd}(A, B) = 0 \Leftrightarrow A = B$, and the triangle inequality $m_{Hd}(A, B) \leqslant m_{Hd}(A, C) + m_{Hd}(C, B)$, cf., e.g., [2, 17].

3.3.2. DEFINITION (extension of the canonical embedding). For a subset A of X define

$$g_A: X \to \mathbf{R}, \qquad g_A(x) = \min_{a \in A} m(a, x). \tag{14}$$

3.3.3. PROPOSITION. *For all subsets A and B of X:*

$$m_F(g_A, g_B) = m_{Hd}(A, B). \tag{15}$$

PROOF. Take $x \in X$. Let $a_1 \in A$ be such that $g_A(x) = m(a_1, x)$. Let $b_1 \in B$ be such that $m(a_1, b_1) \leqslant m(a_1, b)$ for all $b \in B$. We have

$$m_{Hd}(A, B) = \max\left\{\max_{a \in A} \min_{b \in B} m(a, b), \max_{b \in B} \min_{a \in A} m(a, b)\right\}$$

$$\geqslant \max_{a \in A} \min_{b \in B} m(a, b)$$

$$\geqslant \min_{b \in B} m(a_1, b) = m(a_1, b_1).$$

Now,

$$g_B(x) \leqslant m(x, b_1) \leqslant m(x, a_1) + m(a_1, b_1) \leqslant m(x, a_1) + m_{Hd}(A, B).$$

Hence

$$g_B(x) - g_A(x) \leqslant m_{Hd}(A, B)$$

and, similarly $g_A(x) - g_B(x) \leqslant m_{Hd}(A, B)$, showing that

$$\forall x \in X \; |g_A(x) - g_B(x)| \leqslant m_{Hd}(A, B).$$

On the other hand, switching A and B is necessary, we can assume that

$$m_{Hd}(A, B) = \max_{b \in B} \min_{a \in A} m(a, b).$$

Let this maximum be assumed at $b_2 \in B$. Then $g_A(b_2) = m_{Hd}(A, B)$ and $g_B(b_2) = 0$. Hence also

$$m_F(g_A, g_B) \geqslant |g_A(b_2) - g_B(b_2)| = m_{Hd}(A, B)$$

and the proposition is proved. □

3.3.4. REMARK. In the literature, one also frequently encounters the following different definition of Hausdorff distance:

$$\bar{m}_{Hd}(A, B) = \max_{a \in A} \min_{b \in B} m(a, b) + \max_{b \in B} \min_{a \in A} m(a, b). \tag{16}$$

Proposition 3.3.3 fails for this alternative definition. Instead, one has

$$\bar{m}_{Hd}(A, B) = \max_x (g_B(x) - g_A(x)) + \max_x (g_A(x) - g_B(x)). \tag{17}$$

This is proved in practically the same way.

3.4. Five transfer procedures. Now, let us return to the basic situation in which we have a bipartite graph between two sets D and T and we want to transfer a given metric on D to one on T (or vice versa). In this subsection I describe five potential methods for doing this. They have different background philosophies and which one (if

any of these five) is appropriate in a given situation will probably depend on the particular circumstances. All need further investigation.

3.4.1. HAUSDORFF TRANSFER. Given $\Gamma \subset D \times T$, for each $t \in T$ let

$$D_t = \{d \in D: (d, t) \in \Gamma\}.$$

Now, given a metric m_D on D, a metric $m_T = \varphi_\Gamma(m_D)$ on T is defined by

$$m_T(t, t') = (m_D)_{Hd}(D_t, D_{t'}).$$

This transfer method has a number of advantages (and looks very natural). For instance, if D is a trivial metric space (no information) then so is the induced metric on T. Another nice aspect is the following.

3.4.2. PROPOSITION. *If the metric m on D is an ultrametric, then so is $m' = \varphi_\Gamma(m)$ on T.*

PROOF. This is an immediate consequence of the lemma below. □

3.4.3. LEMMA. *Let (X, u) be an ultrametric space. Let $\bar{u}$ be the Hausdorff metric on the subsets of X defined by formulas* (13). *Then $\bar{u}$ is an ultrametric.*

PROOF. By definition

$$\bar{u}(A, C) = \max\left\{\max_{a \in A} \min_{c \in C} m(a, c), \max_{c \in C} \min_{a \in A} m(a, c)\right\}.$$

Interchanging A and C if necessary, we can assume that

$$\bar{u}(A, C) = u(a_1, c_1) = \max_a \min_c u(a, c)$$

for a certain $a_1 \in A$ and $c_1 \in C$. Consider the set $\{u(a_1, b): b \in B\}$ and let the minimum be assumed at $b_1 \in B$. If $u(a_1, b_1) \geqslant u(b_1, c_1)$, then

$$u(a_1, c_1) \leqslant \max\{u(a_1, b_1), u(b_1, c_1)\} = u(a_1, b_1)$$

$$= \min_b u(a_1, b) \leqslant \max_a \min_b u(a, b) \leqslant \bar{u}(A, B)$$

and we are through. It remains to deal with the case

$$u(a_1, b_1) < u(b_1, c_1). \tag{18}$$

Consider the set $\{u(b, c_1): b \in B\}$ and let the minimum be assumed at b_2. If $u(b_2, c_1) \geqslant u(a_1, b_2)$, then we have

$$u(a_1, c_1) \leqslant \max\{u(a_1, b_2), u(b_2, c_1)\} = u(b_2, c_1)$$

$$= \min_b u(b, c_1) \leqslant \max_c \min_b u(b, c) \leqslant \bar{u}(B, C)$$

and we are through. It remains to deal with the case

$$u(b_2, c_1) < u(a_1, b_2). \tag{19}$$

Thus, in total, it remains to deal with the case in which both (18) and (19) hold. By the ultrametric inequality, we then have:

$$\begin{aligned} u(a_1, c_1) &= u(a_1, b_2) > u(b_2, c_1), \\ u(a_1, c_1) &= u(b_1, c_1) > u(a_1, b_1). \end{aligned} \tag{20}$$

Now suppose that $u(b_1, c_1) \leqslant u(b_1, c)$ for all $c \in C$. Then

$$u(a_1, c_1) = u(b_1, c_1) = \min_c u(b_1, c) \leqslant \max_b \min_c u(b, c) \leqslant \bar{u}(B, C)$$

and we are done. Thus it remains to deal with the case in which there exists a $c_2 \in C$ such that

$$u(b_1, c_2) < u(b_1, c_1). \tag{21}$$

But then, using (21) and (20),

$$\begin{aligned} u(a_1, c_2) &\leqslant \max\{u(a_1, b_1), u(b_1, c_2)\} \\ &< \max\{u(a_1, c_1), u(b_1, c_1)\} \\ &= u(a_1, c_1), \end{aligned}$$

contradicting that

$$u(a_1, c_1) = \min_c u(a_1, c)$$

This finishes the proof. □

3.4.4. Remark. Proposition 3.4.2 fails if the alternative definition (16) is taken for the Hausdorff distance.

3.4.5. Another Description of the Hausforff Metric of an Ultrametric. Let $\pi = \{Y_1, \ldots, Y_n\}$ be a partition of X. For each subset J of $\{1, 2, \ldots, n\}$, $J \neq \varnothing$, let

$$P_J = \{A \subset X : A \cap Y_j \neq \varnothing \text{ for all } j \in J \text{ and } A \cap J_j = \varnothing \text{ for all } j \notin J\}.$$

Then, as is easily checked, the P_j form a partition Π of $\mathcal{P}(X)$, the set of subsets of X. Now, an ultrametric u on X is given by a series of coarser and coarser partitions

$$\{\text{singletons}\} = \pi_0 < \pi_1 < \cdots < \pi_k = X,$$

with levels $d_0, d_1, \ldots, d_k$ attached to them. Then $u(x, y) = d_l$ if l is the index of the finest partition of these that does not separate x and y. Associated to the sequence of partitions above there is the sequence of partitions

$$\{\text{singletons}\} = \Pi_0 < \Pi_1 < \cdots < \Pi_k = \mathcal{P}(X).$$

Then the Hausdorff metric on $\mathcal{P}(X)$ is defined by this series of partitions with the same levels as above, i.e. $u(A, B) = d_l$ if l is the index of finest partition from the Π_i that does not separate A and B.

3.4.6. Averaging Transfer. The central idea here is that given two terms t, t' it is unknown which of the documents in D_t and $D_{t'}$ really represent t and t'. This leads to the idea that the dissimilarity of t and t' should be measured by the average distance of documents in D_t and $D_{t'}$, i.e.

$$\delta(D_t, D_{t'}) = \frac{1}{\#D_t} \frac{1}{\#D_{t'}} \sum_{d \in D_t, d' \in D_{t'}} m(d, d').$$

However, this expression does not define a metric. It does suggest, however, considering the *averaging transfer*. This transfer method attaches to a metric m on D the metric $\varphi_\Gamma^{av}(m)$ on T defined by:

$$m_{av}(A, B) = m_{F(D)}\Big(\frac{1}{\#A}\sum_{d\in A} g_d, \frac{1}{\#B}\sum_{d'\in B} g_{d'}\Big),$$
$$\varphi_\Gamma^{av}(m)(t, t') = m_{av}(D_t, D_{t'}) \tag{22}$$

Another way to think about this is that m_{av} somehow measures the distance between the (non-existing) centres of D_t and $D_{t'}$. (For a subspace of the line, and non-interlacing subsets of it, this is exactly the case.)

This idea is reinforced by the following observation. For a subset A of X with metric m, let

$$h_A = \frac{1}{\#A}\sum_{a\in A} g_a.$$

Then, for any $x \in X$,

$$m_F(h_A, g) = \frac{1}{\#A}\sum_{a\in A} m(a, x),$$

as is easily proved.

Note that the metric on T comes again, via Γ, from a metric on the set of all subsets of D, as defined by the first part of (22). Observe that, for all $A, B \subset X$,

$$\delta(A, B) \geqslant m_{av}(h_A, h_B)$$

and it could well be that it is the largest metric subordinate to the averaging dissimilarity δ.

Easy examples show that there is no particular relation between the Hausdorff distance, m_{Hd}, on the set of all subsets $\mathcal{P}(X)$ of a metric space (X, m) and the averaging distance, m_{av}, on $\mathcal{P}(X)$.

3.4.7. TRANSFER VIA WEIGHTS. Let $t, t' \in T$ be terms, $A = D_t$, $A' = D_{t'}$, and χ_A, $\chi_{A'}$ be the characteristic functions of these subsets. Then the Hamming distance between t and t' is equal to the sum (or integral)

$$\sum_{d\in D} |\chi_A(d) - \chi_{A'}(d)|.$$

In this formula, all $d \in D$ are given equal weight. Now let there be given a metric m on D. This can be used to assign a measure of relative importance to the elements of D in which 'central elements' acquire more weight than 'peripheral' ones. For instance, we could proceed as follows:

$$\mu(y) = \frac{S}{\sum_{x\in D} m(x, y)}, \qquad S = \sum_{x,y\in D} m(x, y).$$

Now, for $t, t' \in T$, define

$$\varphi_\Gamma^{w}(m)(t, t') = \sum_d |\chi_A(d) - \chi_{A'}(d)|\,\mu(d).$$

3.4.8. The last two transfer of metrics procedures, 3.4.9 and 3.4.10 below, require

that there is given a metric on the space of all metrics on T or D, so that it is possible to talk about a best approximating metric from a given class to a given metric. Matters pertaining to this will be discussed briefly in the next section. For the moment, let us assume that we have a suitable metric μ on the set $\mathcal{M}(X)$ of all metrics on X, where X is T or D.

3.4.9. Transfer by Approximation. The basic idea is here that the bipartite graph linking D and T perhaps embodies only part of the information linking D and T, and that some other information is hidden in the given metric m on D which comes from a similar source (perhaps another, overlapping, document collection).

Consider all possible bipartite graphs Γ' between D and T. For each Γ', we have the following numbers:
(i) the Hamming distance between Γ and Γ'; and
(ii) $\mu(m, m_D(\Gamma'))$,
where $m_D(\Gamma')$ is the Hamming distance on D defined by Γ'. Let $\nu(m, \Gamma')$ be the set of all Γ' that minimize a suitable chosen convex linear combination of these two numbers. Now define $\varphi_\Gamma^a(m)$ on T as the average of the Hamming distances on T defined by the bipartite graphs in $\nu(m, \Gamma')$.

3.4.10. Inverse Hausdorff Transfer. For each $d \in D$, let T_d be defined by

$$T_d = \{t \in T: (t, d) \in \Gamma\}.$$

Assign the number $m(d, d')$ to the pair of subsets T_d, $T_{d'}$. Now define $\varphi_\Gamma^{iHd}(m)$ as (the average of) the metric(s) m' on T for which the induced Hausdorff metric $\bar{m}'$ on the collection $\{T_d: d \in D\}$ best approximates the metric $m(T_d, T_{d'}) = m(d, d')$ on that same collection.

A important question here is what the conditions are for a metric on a collection of subsets $\mathcal{A} \subset \mathcal{P}(X)$ to be such that it is the Hausforff metric induced by a metric on X. Preliminary to this is the question what collections of subsets $\mathcal{A}$ are such that the associated functions g_A for $A \in \mathcal{A}$ (see (14)) span the linear space $F(X)$.

4. Clustering and Transfer

Much of the literature on mathematical taxonomy and clustering has focused on the question of 'abstracting' from a given dissimilarity a suitable (classification) tree. See [1, 4, 7] and the references therein for some recent results and ideas. Standard references on clustering are [10, 15, 18]. A central question is as follows: Given a metric on a space X, which is the metric of a special kind that best approximates the given one? Very often 'special kind' means ultrametric, so that there is a corresponding hierarchical classification scheme. More generally, tree-like metrics are also often considered. Still more generally, (cf. [1]), it is very interesting (and very natural in some cases) to consider metrics that are sums of splits, and best approximation by such metrics.

I will not discuss here the question of whether trees are really as appropriate for classification and information-finding purposes as one would infer from the dominance of these structures in the literature. It may well be that we have here a relic of the hard copy period: trees are just about the only classification schemes that can be more or less decently printed.

As remarked, the question of best approximating metrics is central. That in turn raises the question of finding a good metric on the set of all metrics. This section is mostly concerned with some matters pertaining to that question.

4.1. Urysohn distance. Let X and Y be metric spaces. The most natural distance between X and Y is probably the Urysohn distance, which is defined by

$$\delta_U(X, Y) = \inf_{\alpha,\beta} \mathrm{Hd}(\alpha(X), \beta(Y))$$

where the infimum is over all inbeddings α, β of X and Y into a third metric space Z, and where Hd denotes the Hausdorff distance. Because of the existence of universal metric spaces (Urysohn spaces) in which each metric space (of cardinality less than a given cardinal) can be inbedded, the space Z in the above can be taken to be a fixed Urysohn space.

As said, this is probably the most natural idea of distance between metric spaces. However, I know of no way to calculate it in concrete cases.

4.2. Function space distance. Consider a fixed set X and the set $\mathcal{M}(X)$ of all metrics on X. Each metric m in $\mathcal{M}(X)$ defines an isometric inbedding $\alpha_m: X \to F(X)$. Now define the distance between two metrics on X by

$$\delta_F(m, m') = \mathrm{Hd}(\alpha_m(X), \alpha_{m'}(X)).$$

This is quite probably related to Urysohn distance, because one of the constructions of Urysohn space uses similar function spaces and embeddings [13].

4.3. Lipshits distance. Consider a set X and two metrics (or dissimilarities), m_1, m_2, defined on it. The *distortion* of m_2 with respect to m_1 is defined by

$$\mathrm{distor}(m_2, m_1) = \sup \frac{m_2(x, y)}{m_1(x, y)},$$

where the sup is taken over all $x, y \in X$, $x \neq y$. The Lipshits distance between m_1, m_2 is now defined as

$$\delta_L(m_1, m_2) = \log(\mathrm{distor}(m_2, m_1)\, \mathrm{distor}(m_1, m_2)).$$

Note that if the two distances are proportional, their Lipshits distance is zero. This is really an advantage for classification problems, because a constant scalar factor should not matter.

It is easy to see the following.

4.3.1. PROPOSITION [12, 14]. *The Lipshits distance δ_L defines a metric on isometry-classes-up-to-a-scalar-factor of metrics (or definite dissimilarities) on a fixed set X.*

Lipshits distance is well adapted to one popular clustering technique.

4.3.2. THEOREM [9]. *The single-link clustering technique applied to a dissimilarity m on X yields an ultrametric u on X that is maximally close to m in the sense of the Lipshits distance (compared to all other ultrametrics on X).*

I know of no other metrics on $\mathcal{M}(X)$ that are linked to a well known clustering method in precisely this way. See, however, [19], in which a connection is established between local optima for the L_2 distance between metrics and the average clustering method.

4.4. Transfer and clustering. A clustering method can be seen as a mapping

$\gamma: \mathcal{M}(X) \to \mathcal{U}(X)$, where $\mathcal{U}(X)$ is a chosen subset of $\mathcal{M}(X)$. Now choose any transfer method (or two of them) to go back and forth from D to T. The combination of such a transfer with a clustering method on, say, $\mathcal{M}(D)$, yields a clustering method on $\mathcal{M}(T)$. What can be said about the resulting clustering method?

ACKNOWLEDGEMENT

I thank one of the referees for some useful and informative comments.

REFERENCES

1. H.-J. Bandelt and A. W. M. Dress, A canonical decomposition theory for metrics on a finite set, *Adv. Math.*, **92** (1992), 47–105.
2. G. Beer, *Topologies on Closed and Closed Convex Sets*, Kluwer, Dordrecht, 1993.
3. T. Beth, D. Jungnickel and H. Lenz, *Design Theory*, Cambridge University Press, 1987.
4. 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, *Adv. Math.*, **53** (1984), 231–402.
5. J. G. Gregory, Citation study of peripheral theories in an expanding research front, *J. Inform. Sci.*, **7** (1983), 71–80.
6. J. A. Hartigan, Representation of similarity matrices by trees, *J. Am. Statist. Assoc.*, **62** (1967), 1140–1158.
7. M. Hazewinkel, Classification in mathematics, discrete metric spaces, and approximation by trees, *Nieuw Archief voor Wiskunde*, to appear, 1995.
8. M. Hazewinkel, Linked balanced designs are BIBD's, Preprint, AM-R9509 CWI (1995).
9. M. Hazewinkel, Lipshitz distance and hierarchical clustering, *J. Classification*, to appear, 1995.
10. N. Jardine and R. Sibson, *Mathematical Taxonomy*, John Wiley, New York, 1971.
11. S. C. Johnson, Hierarchical clustering schemes, *Psychometrika*, **32** (1967), 241–254.
12. W. B. Johnson, J. Lindenstrauss and G. Schechtman, On Lipschitz embeddings of finite metric spaces into low dimensional normed spaces, in: *Geometric Aspects of Functional Analysis*, J. Lindenstrauss and V. D. Millman (eds), Springer, Heidelberg, 1987, pp. 177–184.
13. M. Katetov, On universal metric spaces, in: *General Topology and its Relations to Modern Algebra and Analysis VI*, Z. Frolik (ed.), Heldermann Verlag, Berlin, 1988, pp. 323–330.
14. J. Matousek, Bi-Lipschitz embeddings into low dimensional Euclidean spaces, *Communs Math. Univ. Carolinae*, **31** (1990), 589–600.
15. F. Murtagh and A. Heck, *Multivariate Data Analysis*, Reidel, Dordrecht, 1987.
16. J. Röhmel, Über die Existenz von Inzidenzstrukturen met Regularitätsbedingungen, *Math. Z.*, **133** (1973), 203–218.
17. B. Sendov, *Hausdorff Approximations*, Kluwer, Dordrecht, 1990.
18. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Francisco, 1973.
19. W. Vach and P. O. Degens, A new approach to isotonic agglomerative hierarchical clustering, *J. Classification*, **8** (1991), 217–237.
20. A. F. J. Van Raan and R. J. W. Tijssen, The neural net of neural research, *Scientometrics*, **26** (1992), 169–192.

Received 18 July 1995 and accepted 25 July 1995

MICHIEL HAZEWINKEL
CWI, P.O. Box 94079,
1090 GB Amsterdam, The Netherlands
E-mail: mich@cwi.nl