Acta Informatica 23, 311–323 (1986)

NP-Hard Problems in Hierarchical-Tree Clustering

Mirko Křivánek[1] and Jaroslav Morávek[2]

[1] Research Institute of Mathematical Machines, Loretánské nám. 3, 11855 Praha 1, ČSSR
[2] Mathematical Institute, Czechoslovak Academy of Sciences, Žitná 25, 11567 Praha 1, ČSSR

Summary. We consider a class of optimization problems of hierarchical-tree clustering and prove that these problems are *NP*-hard. The sequence of polynomial reductions and/or transformations used in our proof is based on relatively laborious graph-theoretical constructions and starts in the *NP*-complete problem of 3-dimensional matching. Using our main result we establish the *NP*-completeness of a problem of the best approximation of a symmetric relation on a finite set by an equivalence relation, thus answering in the negative a question proposed implicitly by C.T. Zahn.

I. Introduction and Statement of the Main Result

Within the last twenty years an enormous number of strategies for cluster analysis have been proposed [3, 6, 8, 15]. Though the main emphasis in these efforts has been concentrated on the creation or application of clustering techniques, relatively little attention has been paid to the study of computational complexity of clustering algorithms. Because of wide range of applications of cluster analysis [1, 13] there are many variations of problem formulation. Generally we can consider two types of goals:

(i) *Nonhierarchical clustering* where the goal is to partition a given finite set of objects into nonempty clusters (the number of clusters can be specified beforehand), those objects in the same cluster being considered as close or similar and those in different clusters as distant or dissimilar. The "quality" of this clustering is usually expressed by a real objective function defined on the family of all partitions of the set of objects.

(ii) *Hierarchical clustering (hierarchical-tree clustering)* where the goal is to construct a sequence of nested nonhierarchical clusterings which form a so called *hierarchical tree* and which have to be optimal with respect to a criterion [7, 9].

The problems of nonhierarchical clustering were studied partially from the point of view of computational complexity, see e.g. [2, 5]. On the other hand,

to our knowledge, no comparable results concerning the complexity of hierarchical clustering have been published. Our main aim is to present a result in this respect. For the terminology concerning the computational complexity (*NP*-theory) used in this paper see [4].

Let us review and formalize the main concept of hierarchical clustering in which lies our main interest, cf. e.g. [10]. Throughout this paper, n will denote an integer, $n \geq 3$, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ an n-element set, and $D = (d_{i,j})$ a symmetric real $n \times n$-matrix such that

$$d_{i,j} > 0 \quad \text{if } i \neq j, \quad \text{and} \quad d_{i,j} = 0 \quad \text{if } i = j \qquad (i, j \in \{1, 2, \ldots, n\}).$$

Elements of Ω are called objects (these are to be clustered), and D is called a dissimilarity matrix. Within our context we interpret a "small" value of $d_{i,j}$ as a close relationship between objects ω_i and ω_j.

A hierarchical tree T over Ω is defined as a finite sequence of pairs $T = ((P_1, l_1), (P_2, l_2), \ldots, (P_q, l_q))$ where

(i) $P_1, P_2, \ldots, P_q$ are partitions[1] of Ω;

(ii) $l_1, l_2, \ldots, l_q$ are integers,

$$0 = l_1 < l_2 < \ldots < l_q;$$

(iii) P_k is proper refinement of P_{k+1} $(1 \leq k \leq q-1)$;

(iv) $P_1 = \{\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_n\}\}$ and $P_q = \{\Omega\}$.

The integer q is called the height of T and number l_k, the k-th level of partition P_k in T. It follows that $2 \leq q \leq n$.

Let $\mathfrak{A}(\Omega)$ be the set of all hierarchical trees T over Ω and $\mathfrak{A}_q(\Omega)$, where $2 \leq q \leq n$, the set of all hierarchical trees over Ω, having the height q; if Ω is evident from the context we shall write $\mathfrak{A}$ and $\mathfrak{A}_q$ instead of $\mathfrak{A}(\Omega)$ and $\mathfrak{A}_q(\Omega)$, respectively.

Further we define the function $u\langle T\rangle$: $\Omega \times \Omega \to \mathbb{N}_0$ ($\mathbb{N}_0$ is the set of all nonnegative integers), corresponding to a given hierarchical tree

$$T = ((P_1, l_1), (P_2, l_2), \ldots, (P_q, l_q)) \in \mathfrak{A}(\Omega),$$

as follows:

$$u\langle T\rangle(\omega_i, \omega_j) \overset{\text{df}}{=} \min\{l_k \mid \text{there exists } M \in P_k \; (1 \leq k \leq q) \text{ such that } \{\omega_i, \omega_j\} \subseteq M\}.$$

Remarks: 1) Function $u\langle T\rangle$ is an ultrametric on Ω (cf. [10]). From the point of view of graph theory, a hierarchical tree can be interpreted as a rooted tree. As the rooted tree is an upper semilattice we can interpret $u\langle T\rangle(\omega_i, \omega_j)$ as the level l_k assigned to the least upper bound of elements ω_i and ω_j. In Fig. 1 we give the graphical representation of the hierarchical tree

$$T' = ((P'_1, l'_1), (P'_2, l'_2), (P'_3, l'_3), (P'_4, l'_4)),$$

[1] i.e. finite disjoint decompositions into nonempty classes

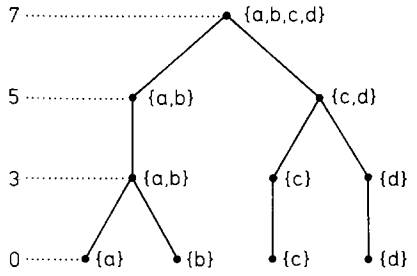

Fig. 1

where

$$(P_1', l_1') = (\{\{a\},\{b\},\{c\},\{d\}\}, 0)$$

$$(P_2', l_2') = (\{\{a, b\},\{c\},\{d\}\}, 3)$$

$$(P_3', l_3') = (\{\{a, b\},\{c, d\}\}, 5)$$

$$(P_4', l_4') = (\{\{a, b, c, d\}\}, 7).$$

The ultrametric $u\langle T'\rangle$ corresponding to T' is given in the underlying tableau:

$u\langle T'\rangle$	a	b	c	d
a	0	3	7	7
b	3	0	7	7
c	7	7	0	5
d	7	7	5	0

//

For the evaluation of hierarchical clustering we use the objective function F: $\mathfrak{A} \to \mathbb{R}_+$ (the set of all nonnegative reals), defined as follows (cf. [9]):

$$F(T) \stackrel{\text{df}}{=} \sum_{1 \leqq i < j \leqq n} |d_{i,j} - u\langle T\rangle(\omega_i, \omega_j)|.$$

Using F we introduce the following optimization problems: Problem HIC (hierarchical clustering):

INSTANCE: Set of objects Ω and dissimilarity matrix D of the size $n \times n$, where $n = \text{card}(\Omega)$;

PROBLEM: Determine a hierarchical tree $T_* \in \mathfrak{A}$ such that

$$F(T_*) = \min\{F(T) \mid T \in \mathfrak{A}\}. \quad //$$

Problem HIC_q $(q = 2, 3, \ldots)$:

INSTANCE: Set of objects Ω and dissimilarity matrix D of the size $n \times n$, where $n = \text{card}(\Omega)$;

PROBLEM: Determine a hierarchical tree $T_+ \in \mathfrak{A}_q$ such that

$$F(T_+) = \min\{F(T) \mid T \in \mathfrak{A}_q\}. \quad //$$

It is evident that problem HIC_2 has a polynomial time complexity. Our main result can be summarized as follows:

Theorem. *Problems* HIC *and* HIC_q *for* $q \geqq 3$ *are NP-hard. This result was presented at* COMPSTAT 1984 *(Prague)*, cf. [12]. □

II. Proof of the Main Result

We shall obtain the proof of our theorem by proving several lemmas.

Lemma 1. *For an arbitrary integer* $q \geqq 2$ *we have* $\mathsf{HIC}_q \propto \mathsf{HIC}_{q+1}$.

Proof. We apply the well-known method of local replacement, see e.g. [4]. To each instance (Ω, D) of HIC_q let us assign an instance (Ω', D') of HIC_{q+1} as follows:

$$\Omega' \stackrel{\text{df}}{=} \Omega \cup \{\omega_{n+1}\} = \{\omega_1, \omega_2, \ldots, \omega_n, \omega_{n+1}\},$$

where ω_{n+1} is a 'new' object (joined to Ω),

$$D' = (d'_{i,j}), \quad (1 \leqq i, j \leqq n+1),$$

where

$$d'_{i,j} \stackrel{\text{df}}{=} d_{i,j} \quad \text{for } 1 \leqq i,\ j \leqq n,$$

$$d'_{i,n+1} \stackrel{\text{df}}{=} d'_{n+1,j} \stackrel{\text{df}}{=} n^2(q + \max\{d'_{i',j'} \mid 1 \leqq i', j' \leqq n\}) + 1, \quad (1 \leqq i, j \leqq n),$$

and

$$d'_{n+1,n+1} \stackrel{\text{df}}{=} 0.$$

(Let us observe that the $d'_{i,j}$ are computable from $d_{i,j}$ in polynomial time.)

Now, the following equivalence is easily verified:

$$T' = ((P'_1, l'_1), (P'_2, l'_2), \ldots, (P'_q, l'_q), (P'_{q+1}, l'_{q+1}))$$

is a solution of HIC_{q+1} if and only if

$$T \stackrel{\text{df}}{=} ((P'_1 \setminus \{\omega_{n+1}\}, 0), (P'_2 \setminus \{\omega_{n+1}\}, l'_2), \ldots, (P'_q \setminus \{\omega_{n+1}\}, l'_q))$$

belongs to $\mathfrak{A}_q(\Omega)$ and it is a solution of HIC_q. □

By virtue of Lemma 1 it is sufficient to prove the *NP*-hardness of HIC_3. In fact, we shall obtain a slightly stronger result. A dissimilarity matrix $D = (d_{i,j})$ will be called *binary* if each off-diagonal element of D equals either 1 or 2, $(d_{i,j} \in \{1,2\}$ if $i \neq j)$. Let ${}^b\mathsf{HIC}$ and ${}^b\mathsf{HIC}_q$ denote the 'binary restrictions' of HIC and HIC_q, respectively, i.e. the computational problems defined in the precisely same way as HIC and HIC_q respectively, except that the instance D is a binary matrix. It follows immediately that

$${}^b\mathsf{HIC} \propto \mathsf{HIC} \quad \text{and} \quad {}^b\mathsf{HIC} \propto \mathsf{HIC}_q \quad \text{for } q \geqq 2. \tag{1}$$

Lemma 2. *It holds that* ${}^b\mathsf{HIC}_3 \propto {}^b\mathsf{HIC}$.

Proof. It is sufficient to prove the following assertion: If T is a solution of ${}^{b}\mathsf{HIC}$ then $T \in \mathfrak{A}_2 \cup \mathfrak{A}_3$. Let us assume on the contrary, that there exists a solution

$$T=((P_1,l_1),(P_2,l_2),\ldots,(P_r,l_r))$$

of ${}^{b}\mathsf{HIC}$ with the property $r>3$. Then it is easy to show that for the hierarchical tree

$$T^*=((P_1,0),(P_2,l_2),(P_r,l_2+1)) \in \mathfrak{A}_3$$

we have

$$u\langle T^*\rangle(\omega_i,\omega_j)=u\langle T\rangle(\omega_i,\omega_j) \quad \text{if } u\langle T\rangle(\omega_i,\omega_j)\leqq 1 \tag{2}$$

and

$$\leqq u\langle T\rangle(\omega_i,\omega_j) \quad \text{otherwise,} \tag{3}$$

$$\max(2,u\langle T^*\rangle(\omega_i,\omega_j))<u\langle T\rangle(\omega_i,\omega_j) \tag{4}$$

for some pair $(\omega_i,\omega_j)\in\Omega\times\Omega$.

Combining (2), (3) and (4) we obtain $F(T^*)<F(T)$, which completes the proof. □

In the next lemma we investigate the levels of a hierarchical tree solving ${}^{b}\mathsf{HIC}_3$.

Lemma 3. *If* $T=((P_1,0),(P_2,l_2),(P_3,l_3)) \in \mathfrak{A}_3$ *is a solution of* ${}^{b}\mathsf{HIC}_3$ *then* $l_2=1$ *and* $l_3=2$.

Proof. It is sufficient to prove that for each

$$T=((P_1,0),(P_2,l_2),(P_3,l_3)) \in \mathfrak{A}_3$$

with the property $l_3 \geqq 3$ there exists

$$T'=((P_1',0),(P_2',l_2'),(P_3',l_3')) \in \mathfrak{A}_3$$

such that

$$l_3'<l_3 \quad \text{and} \quad F(T')<F(T). \tag{5}$$

Indeed, let us consider following two cases for $(0,l_2,l_3)$:

$$(\alpha)\ l_3>3 \quad \text{or} \quad (l_2,l_3)=(1,3); \qquad (\beta)\ (l_2,l_3)=(2,3).$$

In the (α) case we can put evidently $P_j'=P_j\,(j=1,2,3)$, $l_2'=\min(2,l_2)$ and $l_3'=l_3-1$.

In the (β) case we put $l_2'=1$, $l_3'=2$ and

$$P_2'=\{\{\omega_1,\omega_2\},\{\omega_3\},\{\omega_4\},\ldots,\{\omega_f\}\}.$$

Since $l_3'<l_3$ we have to verify inequality $F(T')<F(T)$. Indeed,

$$F(T')=(|d_{1,2}-1|-|d_{1,2}-2|)+\sum_{1\leqq i<j\leqq n}|d_{i,j}-2|$$

i.e.

$$F(T')\leqq 1+\operatorname{card}\{(i,j)\,|\,1\leqq i<j\leqq n, d_{i,j}=1\}. \tag{6}$$

On the other hand, let $P_2=\{I_1, I_2, \ldots, I_r\}$, hence $2 \leqq r \leqq n-1$. Then we have

$$\begin{aligned} F(T) &= \sum_{\rho=1}^{r} \sum_{\{i,j\} \subset I_\rho} |d_{i,j}-2| + \sum_{1 \leqq \rho' < \rho'' \leqq r} \sum_{i \in I_{\rho'}} \sum_{j \in I_{\rho''}} |d_{i,j}-3| \\ &= \sum_{\rho=1}^{r} \sum_{\{i,j\} \subset I_\rho} |d_{i,j}-2| + \sum_{1 \leqq \rho' < \rho'' \leqq r} \sum_{i \in I_{\rho'}} \sum_{j \in I_{\rho''}} |d_{i,j}-2| \\ &\quad + \sum_{1 \leqq \rho' < \rho'' \leqq r} \sum_{i \in I_{\rho'}} \sum_{j \in I_{\rho''}} (|d_{i,j}-3| - |d_{i,j}-2|) \\ &= \sum_{1 \leqq i < j \leqq n} |d_{i,j}-2| + \sum_{1 \leqq \rho' < \rho'' \leqq r} \operatorname{card}(I_{\rho'}) \operatorname{card}(I_{\rho''}), \end{aligned}$$

i.e.

$$F(T) \geqq \operatorname{card}\{(i,j) \mid 1 \leqq i < j \leqq n, d_{i,j}=1\} + (n-1). \tag{7}$$

Combining (6) and (7) we obtain (5) (since $n \geqq 3$). This completes the proof. □

In the next lemma, problem ${}^b\mathsf{HIC}_3$ is restated using the following terminology. For an arbitrary partition $\{I_1, I_2, \ldots, I_r\}$ of Ω let us set $i_\rho = \operatorname{card}(I_\rho)$ and

$$j_\rho = \operatorname{card}\{\{i,j\} \subseteq I_\rho \mid d_{i,j}=1\} \qquad (\rho=1,2,\ldots,r).$$

Lemma 4. *Problem* ${}^b\mathsf{HIC}_3$ *can be stated equivalently as follows: Find a partition* $\{I_1, I_2, \ldots, I_r\}$ *of* Ω *such that*

$$\sum_{\rho=1}^{r} \left(\binom{i_\rho}{2} - 2j_\rho \right)$$

is minimum.

Proof. For every hierarchical tree

$$T = ((P_1, 0), (\{I_1, I_2, \ldots, I_r\}, 1), (\{\Omega\}, 2)) \in \mathfrak{A}_3$$

we have

$$\begin{aligned} F(T) &= \sum_{\rho=1}^{r} \sum_{\{i,j\} \subset I_\rho} |d_{i,j}-1| + \sum_{1 \leqq \rho' < \rho'' \leqq r} \sum_{i \in I_{\rho'}} \sum_{j \in I_{\rho''}} |d_{i,j}-2| \\ &= \sum_{\rho=1}^{r} \left(\binom{i_\rho}{2} - j_\rho \right) + \operatorname{card}\{\{i,j\} \mid d_{i,j}=1\} - \sum_{\rho=1}^{r} j_\rho \\ &= \operatorname{card}\{\{i,j\} \mid d_{i,j}=1\} + \sum_{\rho=1}^{r} \left(\binom{i_\rho}{2} - 2j_\rho \right), \end{aligned}$$

which concludes the proof. □

For proving the *NP*-hardness of ${}^b\mathsf{HIC}_3$ we shall use the following decision problem: Problem EC3 (exact cover by ordered 3-tuples):

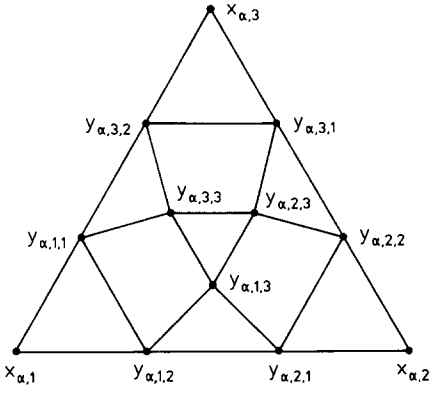

Fig. 2

INSTANCE: 1) finite set X with $\text{card}(X)=3m$ for some positive integer m;

2) finite indexed family

$$\mathscr{C}=((x_{\alpha,1},x_{\alpha,2},x_{\alpha,3}))_{\alpha=1}^{a}$$

of ordered 3-tuples of elements of X with the property that each element of X occurs at least in one 3-tuple of $\mathscr{C}$, i.e.

$$\bigcup_{\alpha=1}^{a}\{x_{\alpha,1},x_{\alpha,2},x_{\alpha,3}\}=X;$$

QUERY: Decide whether $\mathscr{C}$ contains a subfamily $\mathscr{C}'$ such that each element of X occurs in exactly one 3-tuple of $\mathscr{C}'$. //

$\mathscr{C}'$ is called an exact cover for X.

Lemma 5. *Problem* EC3 *is NP-complete*, (cf. [11]). □

Our next aim is to reduce polynomially EC3 to ${}^{b}\text{HIC}_3$. For this aim we assign to each instance $(X,\mathscr{C})$ of EC3 an instance (Ω,D) of ${}^{b}\text{HIC}_3$. First, let us put

$$n\overset{\text{df}}{=}3m+9\,\text{card}(\mathscr{C})=3m+9a \tag{8}$$

and

$$\Omega\overset{\text{df}}{=}X\cup\{y_{\alpha,\beta,\gamma}|\alpha\in\{1,2,\ldots,a\};\beta,\gamma\in\{1,2,3\}\}, \tag{9}$$

where $y_{\alpha,\beta,\gamma}$ are 9.a 'new' objects joined to X. (It is appropriate to index these objects by the triple subscripts.)

The dissimilarity matrix D will be introduced using certain graphs (we consider in this paper finite undirected graphs without loops and parallel edges). For each $\alpha\in\{1,2,\ldots,a\}$ let us consider the graph $G_\alpha=(V_\alpha,E_\alpha)$, see Fig. 2, where

(i) The vertex-set is:

$$V_\alpha\overset{\text{df}}{=}\{x_{\alpha,1},x_{\alpha,2},x_{\alpha,3}\}\cup\{y_{\alpha,\beta,\gamma}|1\leqq\beta,\gamma\leqq3\};$$

(ii) The edge-set is (edges are defined as 2-element subsets of the vertex-set):

$$\begin{aligned} E_\alpha \overset{\text{df}}{=} & \{\{x_{\alpha,\beta}, y_{\alpha,\beta,\gamma}\} \mid 1 \leqq \beta \leqq 3, \gamma = 1,2\} \\ & \cup \{\{y_{\alpha,\beta,3}, y_{\alpha,\beta',3}\} \mid 1 \leqq \beta \neq \beta' \leqq 3\} \\ & \cup \{\{y_{\alpha,\beta,2}, y_{\alpha,\beta,\gamma}\} \mid 1 \leqq \beta \leqq 3, \gamma = 1,3\} \\ & \cup \{\{y_{\alpha,\beta,1}, y_{\alpha,\beta',\gamma}\} \mid 1 \leqq \beta \leqq 3, \gamma = 2,3 \text{ and } (\beta - \beta' = 1) \vee (\beta' - \beta) = 2\}. \end{aligned}$$

Using graphs G_α we introduce the graph $\mathbf{G} \overset{\text{df}}{=} (\Omega, \mathbf{E})$, where

$$\mathbf{E} = \bigcup_{\alpha=1}^{a} E_\alpha.$$

The dissimilarity matrix $D = (d_{i,j})$ will be now defined as follows: We consider an arbitrary fixed numbering $\omega_1, \omega_2, \ldots, \omega_n$ of elements of Ω and put

$$d_{i,j} = 0 \quad \text{if } i = j, \tag{10}$$

$$d_{i,j} = 1 \quad \text{if } i \neq j \quad \text{and} \quad \{\omega_i, \omega_j\} \in \mathbf{E}, \tag{11}$$

$$d_{i,j} = 2 \quad \text{otherwise.} \tag{12}$$

In the sequel, we shall use some additional graph-theoretical definitions and notations: The term *subgraph* will denote an induced subgraph; the subgraph of $\mathbf{G}$ induced by an nonempty subset $I \subseteq \Omega$ will be denoted by $G(I)$. A subgraph of $\mathbf{G}$ with 3 vertices and 3 edges will be called a *triangle*.

Let $E = \{(W_t, H_t)\}$ be a finite set of triangles; W_t is the vertex-set and H_t is the edge-set of the triangle (W_t, H_t). E will be called a vertex-partition of the graph $\mathbf{G}$ into triangles if $\bigcup_t W_t = \Omega$ and $\bigcup_t H_t \subseteq \mathbf{E}$.

Lemma 6. *A solution of* EC3 *exists if and only if there exists a vertex-partition of* $\mathbf{G}$ *into triangles.*

Proof. Let $\mathscr{C}'$ be an exact cover for X with respect to EC3. Let E be defined as the minimum, with respect to the cardinality, set of triangles such that the following two conditions are fulfilled ($\alpha \in \{1,2,\ldots,\text{card}(\mathscr{C})\}$):

(i) If $(x_{\alpha,1}, x_{\alpha,2}, x_{\alpha,3}) \in \mathscr{C}'$ then

$$G(\{y_{\alpha,1,3}, y_{\alpha,2,3}, y_{\alpha,3,3}\} \in E$$

and

$$G(\{x_{\alpha,\beta}, y_{\alpha,\beta,1}, y_{\alpha,\beta,2}\}) \in E$$

for all $\beta = 1,2,3$.

(ii) If $(x_{\alpha,1}, x_{\alpha,2}, x_{\alpha,3}) \in \mathscr{C} - \mathscr{C}'$ then

$$G(\{y_{\alpha,\beta',1}, y_{\alpha,\beta,2}, y_{\alpha,\beta,3}\} \mid (\beta' - \beta) = 1 \vee (\beta - \beta') = 2, 1 \leqq \beta, \beta' \leqq 3\}.$$

The set E contains $4m + 3(a-m) = m + 3a$ triangles, and it is easy to see that E is a vertex-partition of $\mathbf{G}$.

Conversely, given a vertex-partition E of $\mathbf{G}$ into triangles we define $\mathscr{C}'$ as the family of all $(x_{\alpha,1}, x_{\alpha,2}, x_{\alpha,3}) \in \mathscr{C}$ such that

$$G(\{y_{\alpha,1,3}, y_{\alpha,2,3}, y_{\alpha,3,3}\}) \in E.$$

It is left to the reader to show that $\mathscr{C}'$ is an exact cover in the sense of problem EC3 with respect to the instance $(X, \mathscr{C})$. (Observe that if E contains a "central" triangle

$$G(\{y_{\alpha,1,3}, y_{\alpha,2,3}, y_{\alpha,3,3}\})$$

then it contains also 3 triangles

$$G(\{x_{\alpha,\beta}, y_{\alpha,\beta,1}, y_{\alpha,\beta,2}\}), \quad \text{where} \quad \beta \in \{1,2,3\}.$$

This completes the proof. □

For $i \in \{1,2,\ldots,\mathrm{card}(\Omega)\}$ let ${}^{m}j(i)$ denote the maximum number of edges in a subgraph $G(I)$ with $\mathrm{card}(I)=i$.

Lemma 7. *It holds that*

$$^{m}j(i) \leqq 2i-3, \quad i \geqq 1.$$

Proof. Let us consider an arbitrary $I \subset \Omega$ with $\mathrm{card}(I)=i$. Let i_α denote the number of all edges in the subgraph $G(V_\alpha \cap I)$, where $\alpha \in \{1,2,\ldots,a\}$ and let A denote the set of all α with the property $(V_\alpha \setminus X) \cap I \neq \emptyset$. Since each graph G_α is a block of graph **G** (cf. e.g. [14], p. 529) we have

$$^{m}j(i) = \max\{\sum_{\alpha \in A} i_\alpha \mid I \subset \Omega, \mathrm{card}(I)=i\}. \tag{13}$$

Examining the graph G_α (cf. Fig. 2) one can easily verify the following values of ${}^{m}j(i)$ in G_α:

$$^{m}j(1)=0, \quad ^{m}j(2)=1, \quad ^{m}j(3)=3, \quad ^{m}j(4)=4, \quad ^{m}j(5)=6,$$

$$^{m}j(6)=8, \quad ^{m}j(7)=10, \quad ^{m}j(8)=12, \quad ^{m}j(9)=15, \quad ^{m}j(10)=17,$$

$$^{m}j(11)=19, \quad ^{m}j(12)=21.$$

Hence

$$i_\alpha \leqq {}^{m}j(\mathrm{card}(V_\alpha \cap I) \leqq 2\,\mathrm{card}(V_\alpha \cap I) - 3/\mathrm{card}(A), \quad \alpha \in A. \tag{14}$$

Altogether (13) and (14) yield

$$^{m}j(i) \leqq 2 \cdot \mathrm{card}(I) - 3 = 2i-3, \quad i \geqq 1.$$

The proof is concluded. □

Lemma 8. *Let* $\emptyset \neq I \subseteq \Omega$, $i=\mathrm{card}(I)$, *and let* j *be the number of all edges in* $G(I)$. *Then*

$$\binom{i}{2} - 2j + i \geqq 0. \tag{15}$$

Moreover

$$\binom{i}{2} - 2j + i = 0$$

if and only if $i=3$ *and* $G(I)$ *is a triangle.*

Proof. Obviously it holds that

$$2i-3 \leqq \left(\binom{i}{2}+i\right)\Big/2, \quad i \geqq 1.$$

Thus by applying Lemma 7 the inequality (15) follows. Moreover if $G(I)$ is a triangle then

$$\binom{i}{2}-2j+i=3-6+3=0.$$

Conversely, if

$$\binom{i}{2}-2j+i=0$$

then it follows from the definition of **G** and from Lemma 7 that $i=j=3$ and $G(I)$ is a triangle. □

Now, let the instance (Ω, D) for ${}^{b}\mathsf{HIC}_3$ be defined by (8)–(12). For each partition $\{I_1, I_2, \ldots, I_r\}$ of Ω let us set

$$\Psi(\{I_1, I_2, \ldots, I_r\}) \stackrel{\text{df}}{=} \sum_{\rho=1}^{r} \left(\binom{i_\rho}{2}-2j_\rho\right),$$

where i_ρ, j_ρ $(\rho=1,2,\ldots,r)$ are defined as in Lemma 4. It is easily observed that j_ρ equals the number of all edges in the graph $G(I_\rho)$.

Lemma 9. *Let $\{I_1, I_2, \ldots, I_r\}$ be a partition of Ω. Then*

$$\Psi(\{I_1, I_2, \ldots, I_r\}) \geqq -3(m+3\,\mathrm{card}(\mathscr{C})). \tag{16}$$

Moreover,

$$\Psi(\{I_1, I_2, \ldots, I_r\}) = -3(m+3\,\mathrm{card}(\mathscr{C})) \tag{17}$$

if and only if $G(I_\rho)$ is a triangle for each $\rho \in \{1,2,\ldots,r\}$, i.e.

$$\{G(I_\rho) \mid \rho = 1,2,\ldots,r\}$$

is a vertex-partition of G into triangles. (Recall that $m = 1/3\,\mathrm{card}(X)$.)

Proof. Let η_s denote the number of all $\rho \in \{1,2,\ldots,r\}$ with $\mathrm{card}(I_\rho)=s$. Then

$$\Psi(\{I_1, I_2, \ldots, I_r\}) = \sum_{s=1}^{\infty} \left(\binom{s}{2}-2\cdot {}^{m}j(s)\right)\cdot \eta_s$$

On the other hand

$$\sum_{s=1}^{\infty} s\cdot\eta_s = \mathrm{card}(\Omega) = 3(m+3\,\mathrm{card}(\mathscr{C})).$$

Thus we have

$$\Psi(\{I_1, \ldots, I_r\}) = \Psi(\{I_1, \ldots, I_r\}) + \left(\sum_{s=1}^{\infty} s\cdot\eta_s - 3(m+a)\right)$$

$$= \sum_{s=1}^{\infty} \left(\binom{s}{2}-2\cdot {}^{m}j(s)+s\right)\cdot\eta_s - 3(m+3\,\mathrm{card}(\mathscr{C})).$$

The inequality (16) now follows immediately from Lemma 7. Moreover, if $\{G(I_\rho)|\rho=1,2,\ldots,r\}$ is a vertex-partition of G into triangles then (17) clearly holds.

Conversely, let us assume that (17) holds. Then, using Lemma 7 we see successively that

(j) $\eta_s=0$ for $s\neq 3$,

(jj) $\eta_3=m+3\,\mathrm{card}(\mathscr{C})$,

(jjj) for each $\rho\in\{1,2,\ldots,r\}$ the subgraph $G(I_\rho)$ is a triangle.

The proof is completed. □

Lemma 10. *It holds* $\mathsf{EC}3\propto{}^{b}\mathsf{HIC}_3$.

Proof. It is sufficient to see that given a solution of ${}^{b}\mathsf{HIC}_3$ for the instance (Ω, D), defined by (8)-(12), we obtain the answer to EC3 using a polynomially bounded algorithm. Indeed, solving ${}^{b}\mathsf{HIC}_3$ we obtain a partition $\{I_1, I_2, \ldots, I_r\}$ of Ω such that $\Psi(\{I_1, I_2, \ldots, I_r\})$ is minimum. Now, by virtue of Lemma 9 and Lemma 6 we have

$$\Psi(\{I_1, I_2, \ldots, I_r\}) = -3(m+3\,\mathrm{card}(\mathscr{C}))$$

if and only if there exists a vertex-partition of **G** into triangles i.e. if and only if there exists an exact cover by ordered 3-tuples with respect to EC3 for the instance $(X, \mathscr{C})$. This completes the proof. □

The proof of the announced result (Theorem): The *NP*-hardness of HIC_q for $q\geqq 3$ follows from (1), Lemma 1, Lemma 5 and Lemma 10. The *NP*-hardness of HIC follows from (1), Lemma 2, Lemma 5 and Lemma 10.

III. Best Approximation of Symmetric Relation by an Equivalence

In [16] the following optimization problem is proposed: Given a finite non-empty set Z and a symmetric relation $s\subseteq Z\times Z$ we are asked to determine an equivalence relation $e\subseteq Z\times Z$ minimizing the objective function

$$e\longmapsto \mathrm{card}(s\,\triangle e),$$

where $s\triangle e=(s\setminus e)\cup(e\setminus s)$ is the symmetric difference of the relations s, e (considered as subsets of $Z\times Z$). I.C. Lerman observed in [13] the importance of this problem in the hierarchical clustering.

By an immediate application of Lemma 10 we prove that the underlying decision computational problem is *NP*-complete: Problem S△E (best approximation of a symmetric relation by an equivalence relation):

INSTANCE: A finite set $Z=\{z_1, z_2, \ldots, z_m\}$, symmetric relation $s\subseteq Z\times Z$ and a positive integer k;

QUERY: Decide whether there exists an equivalence relation $e\subseteq Z\times Z$ such that

$$\mathrm{card}(s\,\triangle e)\leqq k.$$

To prove this assertion we observe that problem $\mathsf{S}\triangle\mathsf{E}$ is evidently in *NP*, and exhibit the polynomial transformation

$$ {}^{b}_{d}\mathsf{HIC}_3 \propto \mathsf{S}\triangle\mathsf{E}, \tag{18} $$

where ${}^{b}_{d}\mathsf{HIC}_3$ is the following decision version of ${}^{b}\mathsf{HIC}_3$:

INSTANCE: (Ω, D, k'), where D is binary and k' is a positive integer;

QUERY: Decide whether there exists $T \in \mathfrak{A}_3(\Omega)$ such that

$$ F(T) = \sum_{1 \leqq i < j \leqq n} |d_{i,j} - u\langle T\rangle(\omega_i, \omega_j)| \leqq k'. \quad // $$

(It follows immediately from Lemma 10 that ${}^{b}_{d}\mathsf{HIC}_3$ is *NP*-hard.)

To prove (18) we assign to an instance (Ω, D, k') of ${}^{b}_{d}\mathsf{HIC}_3$ the instance (Z, s, k) of $\mathsf{S}\triangle\mathsf{E}$, where

$$ Z \overset{\text{df}}{=} \Omega \quad \text{and} \quad z_j = \omega_j \quad (j = 1, 2, \ldots, m); $$

$$ s \overset{\text{df}}{=} \{(z_i, z_j) \in Z \times Z \mid d_{i,j} \leqq 1\}; $$

$$ k \overset{\text{df}}{=} 2k'. $$

Now observe that mapping

$$ T \mapsto e(T \in \mathfrak{A}_3(\Omega)) $$

defined by

$$ e = \{(z_i, z_j) \in Z \times Z \mid u\langle T\rangle(\omega_i, \omega_j) \leqq 1\}, $$

maps bijectively $\mathfrak{A}_3(\Omega)$ onto the set of all equivalences on Z (see the proof of Lemma 4), and preserves the equality

$$ \operatorname{card}(e \triangle s) = \operatorname{card}(e \setminus s) + \operatorname{card}(s \setminus e) = 2F(T). $$

Thus

$$ \operatorname{card}(e \triangle s) \leqq k \quad \text{if and only if } F(T) \leqq k', $$

which completes the proof.

Acknowledgements. We wish to express our thanks to Dr. I. Havel for his attention to this work and to an anonymous referee for pointing out a considerable simplification of our reduction of EC3 to ${}^{b}\mathsf{HIC}_3$.

References

1. Anderberg, M.: Cluster Analysis for Applications. New York: Academic Press 1973
2. Brucker, P.: On the Complexity of Clustering Problems. In: Optimization and Operations Research (R. Henn, B. Korte, W. Oletti eds.), pp. 45–54. Berlin, Heidelberg, New York: Springer 1977
3. Diday, E., Bochi, S., Brossier, G., Celeux, G., Charles, C., Chifflet, R., Darcos, J., Diday, E., Diebolt, J., Fevre, P., Govaert, G., Hanani, C., Jacquet, D., Lechevallier, Y., Lemaire, J., Lemoine, Y., Molliere, J.L., Morisset, G., Ok-Sakun, Y., Rousseau, P., Sankoff, D., Schroeder, A., Sidi, J., Taleng, F.: Optimisation en classification automatique. INRIA, Rocquencourt, 1979

4. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of *NP*-Completeness. San Francisco: Freeman 1979
5. Gonzales, T.: On the Computational Complexity of Clustering and Related Problems. In: System Modelling and Optimization (R. Drenick, F. Kozin eds.), pp. 174–182. Berlin, Heidelberg, New York: Springer 1982
6. Hartigan, J.A.: Clustering Algorithms. New York: John Wiley 1975
7. Hartigan, J.A.: Representation of Similarity Matrices by Trees. JASA **62**, 1140–1158 (1967)
8. Jambu, M., Lebeaux, M.-O.: Cluster Analysis and Data Analysis. Amsterdam: North-Holland 1983
9. Jardine, N., Sibson, R.: Mathematical Taxonomy. New York: John Wiley 1971
10. Johnson, S.C.: Hierarchical Clustering Schemes. Psychometrika **32**, 241–254 (1967)
11. Karp, R.M.: Reducibility among Combinatorial Problems. In: Complexity of Computer Computations (E.W. Miller, J.W. Thatcher, eds.), pp. 85–104. New York: Plenum Press 1972
12. Křivánek, M., Morávek, J.: On *NP*-Hardness in Hierarchical Clustering. In: Proceedings COMPSTAT '84, pp. 189–194. Vienna: Physica 1984
13. Lerman, I.C.: Classification et analyse ordinale des données. Paris: Dunod 1981
14. Lovász, L.: Combinatorial Problems and Exercises. Budapest: Akademiai Kiadó 1979
15. Späth, H.: Cluster Analysis Algorithms for Data Reduction and Classification of Objects. London: Ellis Horwood 1980
16. Zahn, C.T.: Approximating Symmetric Relations by Equivalence Relations. SIAM J. Appl. Math. **12**, 840–847 (1964)

Received April 25, 1985/Oktober 15, 1985