The characterization of branching dependencies*

János Demetrovics

Computer and Automation Institute, Hungarian Academy of Sciences, Kende u. 13–17, H-1111 Budapest, Hungary

Gyula O.H. Katona and Attila Sali

Mathematical Institute of the Hungarian Academy of Sciences, P.O. Box 127, H-1364 Budapest, Hungary

Received 22 September 1990

Abstract

Demetrovics, J., G.O.H. Katona and A. Sali, The characterization of branching dependencies, Discrete Applied Mathematics 40 (1992) 139–153.

A new type of dependencies in a relational database model is introduced. If b is an attribute, A is a set of attributes then it is said that b (p,q)-depends on A, in notation $A \overset{(p,q)}{\rightarrow} b$, in a database r if there are no $q+1$ rows in r such that they have at most p different values in A, but $q+1$ different values in b. $(1,1)$-dependency is the classical functional dependency. Let $\mathscr{I}(A)$ denote the set $\{b: A \overset{(p,q)}{\rightarrow} b\}$. The set function $\mathscr{I}(A)$ is characterized if $p=1$, $1<q$; $p=2$, $3<q$; $2<p$, $p^2-p-1<q$. Implications among (p,q)-dependencies are also determined.

1. Introduction

A relational database system of the scheme $R(A_1, A_2, \ldots, A_n)$ can be considered as a matrix, where the columns correspond to the *attributes* A_i*'s* (for example name, date of birth, place of birth, etc.), while the rows are the n-tuples of the relation r. That is, a row contains the data of a given *individual*. Let Ω denote the set of attributes (the set of the columns of the matrix). Let $A \subseteq \Omega$ and $b \in \Omega$. We say that b (*functionally*) *depends* on A (see [1,2]) if the data in the columns of A determine the data of b, that is there exist no two rows which agree in A but are different in b. We denote this by $A \rightarrow b$.

Correspondence to: Professor G.O.H. Katona, Mathematical Institute of the Hungarian Academy of Sciences, P.O. Box 127, H-1364 Budapest, Hungary.

* Research was (partially) supported by the Hungarian National Foundation for Scientific Research grant no. 2575.

Functional dependencies have turned out to be very useful. All existing database managing systems are based on this concept. Let us consider the following example. Suppose that $\Omega=\{A_1,A_2,A_3,A_4\}$ and $A_1\to A_2$ and $A_3\to A_4$ hold. If we store the whole matrix in the memory of a computer, then it requires $4N_1N_3$ registers in the worst case, where N_1 (N_3) denotes the number of possible different values of A_1 (A_3). Indeed, A_1 and A_3 can take values independently, but they determine A_2 and A_4, respectively. Thus, the number of different rows is at most N_1N_3. However, using the given functional dependencies, we can save a lot of memory. Indeed, it is enough to store the matrix consisting of the columns A_1 and A_3 ($2N_1N_3$ registers) together with two little matrices each having two columns. One contains values of A_1 and A_2 in the first and second columns, respectively. The first column contains all possible values of A_1, while the second one contains the values determined by the dependency $A_1\to A_2$. The other small matrix is built up from A_3 and A_4 in the same way. The number of stored values is at most $2N_1N_3+2(N_2+N_4)$, which is usually significantly smaller than $4N_1N_3$.

In the present paper we introduce a more general (weaker) dependency, than the functional dependency. We do it first in a very particular case, then we show the usefulness of the concept. Let $A\subseteq\Omega$ and $b\in\Omega$, we say that b $(1,2)$-depends on A if the values in A determine the values in b in a "two-valued" way. That is, there exist no three rows the same in A but having three different values in b. We denote it by $A\xrightarrow{(1,2)}b$. Similarly, $A\xrightarrow{(1,q)}b$ if there exist no $q+1$ rows each having the same values in columns of A, but containing $q+1$ different values in the column b.

Let us suppose that the database consists of the trips of an international transport truck, more precisely, the names of the countries the truck enters. For the sake of simplicity, let us suppose that the truck goes through exactly four countries in each trip (counting the start and endpoints, too) and does not enter a country twice during one trip. Suppose furthermore, that there are 30 possible countries and one country has at most five neighbours. Let A_1, A_2, A_3, A_4 denote the first, second, third and fourth country as attributes. It is easy to see that $A_1\xrightarrow{(1,5)}A_2$, $\{A_1,A_2\}\xrightarrow{(1,4)}A_3$ and $\{A_2,A_3\}\xrightarrow{(1,4)}A_4$. Now, we cannot decrease the size of the stored matrix, as in the case of functional ($(1,1)$-) dependency, but we can decrease the range of the elements of the matrix. The range of each element of the original matrix consists of 30 values, names of countries or some codes of them (five bits each, at least). Let us store a little table ($30\times5\times5=750$ bits) that contains a numbering of the neighbours of each country, which assigns to them the numbers $0,1,2,3,4$ in some order. Now we can replace attribute A_2 by these numbers (A_2^*), because the value of A_1 gives the starting country and the value of A_2^* determines the second country with the help of the little table. The same holds for the attribute A_3, but we can decrease the number of possible values even further, if we give a table of numbering the possible third countries for each pair A_1,A_2. In this case, the attribute A_3^* can take only four different values. The same holds for A_4, too. That is, while each element of the original matrix could be encoded by five bits, now for the cost of two little auxiliary tables we could decrease the length of the elements in the second column to three bits, and that of the elements in the third and fourth columns to two bits.

It is easy to see, that the same idea can be applied in each case when we store the paths of a graph, whose maximum degree is much less than the number of its vertices or when we want to store the sequence of states of a process, where the number of all possible states is much larger than the number of possible successing states of a state or in any case when there hold many $(1, q)$-dependencies, where q is small.

The general concept we shall study is the (p, q)-dependency ($1 \le p \le q$, integers).

Definition 1.1. Let r be a relational database system of the scheme $R(A_1, A_2, \ldots, A_n)$. Let $A \subseteq \Omega$ and $b \in \Omega$. We say that b *(p, q)-depends* on A if there are no $q+1$ rows (n-tuples) of r such that they contain at most p different values in each column (attribute) of A, but $q+1$ different values in b.

The aim of this paper is to generalize theorems valid for functional dependencies to (p, q)-dependencies. Several very interesting combinatorial problems arise in this context.

2. Characterization of (p,q)-dependencies

For a given relation r (or its matrix M) we define a function from the family of subsets of Ω into itself Ω as follows.

Definition 2.1. Let M be the matrix of the given relation r. Let us suppose, that $1 \le p \le q$. Then the mapping $\mathcal{I}_{Mpq}: 2^\Omega \to 2^\Omega$ is defined by

$$\mathcal{I}_{Mpq}(A) = \{b : A \xrightarrow{(p,q)} b\}. \tag{2.1}$$

We collect two important properties of the mapping $\mathcal{I}_{Mpq}$ in the following proposition.

Proposition 2.2. *Let r, Ω, M, p and q as above. Furthermore, let $A, B \subseteq \Omega$. Then*

$$\begin{aligned} &\text{(i)} \qquad A \subseteq \mathcal{I}_{Mpq}(A), \\ &\text{(ii)} \qquad A \subseteq B \Rightarrow \mathcal{I}_{Mpq}(A) \subseteq \mathcal{I}_{Mpq}(B). \end{aligned} \tag{2.2}$$

Proof. It is clear that if $b \in A$, then $A \xrightarrow{(p,q)} b$ which proves (i). On the other hand, if $A \subseteq B$ and $A \xrightarrow{(p,q)} b$, then $B \xrightarrow{(p,q)} b$ holds as well. $\square$

Definition 2.3. Set functions satisfying (2.2) are called *increasing-monotone functions*. We say that such an increasing-monotone function $\mathcal{N}$ is *(p, q)-representable* if there exists a matrix M such that $\mathcal{N} = \mathcal{I}_{Mpq}$.

Whether all increasing-monotone functions on subsets of any given Ω are (p, q)-

representable? If not, what are the restrictions on p and q or $\mathcal{N}$? The following theorem gives a partial answer.

Theorem 2.4. *Let $\mathcal{N}$ be an increasing-monotone function on subsets of Ω satisfying $\mathcal{N}(\emptyset)=\emptyset$. Then $\mathcal{N}$ is (p,q)-representable if one of the following holds*:

$$\begin{array}{lll} \text{(i)} & p=1 \quad \text{and} \quad 1<q, & \\ \text{(ii)} & p=2 \quad \text{and} \quad 3<q, & \qquad (2.3) \\ \text{(iii)} & 2<p \quad \text{and} \quad p^2-p-1<q. & \end{array}$$

Proof. Let us call a sequence of subsets $\emptyset\neq A_1\subset A_2\subset\cdots\subset A_k$ of Ω a *chain* if the following two conditions hold:

$$\begin{array}{ll} \text{(i)} & \mathcal{N}(A_i)=A_{i+1} \quad (1\leq i<k), \\ \text{(ii)} & \mathcal{N}(A_k)=A_k. \end{array} \qquad (2.4)$$

For such a chain L we construct the matrix $M(z,r,L)$ shown as follows:

$$\begin{array}{cccccc} A_1 & A_2\setminus A_1 & A_3\setminus A_2 & \dots & A_k\setminus A_{k-1} & \Omega\setminus A_k \\ \end{array}$$
$$\begin{bmatrix} z & z & z & \dots & z & z \\ z & z & z & \dots & z & z+1 \\ z & z & z & \dots & z & z+2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ z & z & z+r & \dots & z+(k-2)r & z+kr \\ z & z+1 & z+r+1 & \dots & z+(k-2)r+1 & z+kr+1 \\ z & z+2 & z+r+2 & \dots & z+(k-2)r+2 & z+kr+2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ z & z+r & z+2r & \dots & z+(k-1)r & z+(k+1)r \end{bmatrix}$$

Each column of the matrix begins with some z's, then from a certain position the natural numbers come in increasing order: $z,\dots,z,z+1,z+2,\dots$. The columns of $A_i\setminus A_{i-1}$ $(1<i\leq k)$ are all identical and the same holds for the columns of A_1 and $\Omega\setminus A_k$, respectively. The columns of the latter consist of $z,z+1,z+2,\dots$. On the other hand, columns of A_1 consist of all z's. Columns of $A_2\setminus A_1$ are shifted in comparison to columns of A_1 by r, i.e., the number of z's at the beginning is r less than that in columns of A_1, but their last element is $r+z$. In general, columns of $A_i\setminus A_{i-1}$ are shifted in comparison to those of $A_{i-1}\setminus A_{i-2}$ by r $(1<i\leq k)$. However, columns of $\Omega\setminus A_k$ are shifted by $2r$ in comparison to $A_k\setminus A_{k-1}$. According to the definition of a chain, $A_i\setminus A_{i-1}$ $(1\leq i\leq k)$ cannot be empty, but $\Omega\setminus A_k$ can be. In the latter case the matrix does not contain such columns. We shall only use the following easily checked properties of this matrix.

(i) If two positions in a column of $A_i\setminus A_{i-1}$ $(1<i\leq k)$ contain the same ele-

ment, then any column of $A_j \setminus A_{j-1}$ contains an identical element in those two positions for all $j<i$. ($A_0=\emptyset$ by assumption.)

(ii) Choosing a z in a column of $A_i \setminus A_{i-1}$ there can stand only z or $z+1$ or $z+2$ or ... or $z+r$ in the same position of a column of $A_{i+1} \setminus A_i$. However, if we choose a number s different from z in a column of $A_i \setminus A_{i-1}$, then only $s+r$ can stand in the same position of a column of $A_{i+1} \setminus A_i$.

(iii) For $k \geq j > i+1 \geq 2$ we can find $2r+1$ different numbers (namely $z, z+1, \ldots, z+2r$) in a column of $A_j \setminus A_{j-1}$ so that only z's stand in the same positions of a column of $A_i \setminus A_{i-1}$.

(iv) We can find $2r+1$ different numbers (namely $z, z+1, \ldots, z+2r$) in a column of $\Omega \setminus A_k$ so that only z's stand in the same positions of a column of $A_i \setminus A_{i-1}$ for $1 \leq i \leq k$.

Let $\mathscr{L}=\{L_1, L_2, \ldots, L_m\}$ be a set of chains which satisfies that for every pair A, b ($A \subseteq \Omega$, $b \in \Omega$) satisfying $A \neq \emptyset$, $b \notin \mathscr{N}(A)$ there is a chain L_j and a set A_i in that chain satisfying

$$A \subseteq A_i \quad \text{and} \quad b \notin \mathscr{N}(A_i). \tag{2.5}$$

We obtain such a set of chains for example, if we take all possible nonempty subsets of Ω as A_1. For every chain L_i we construct p matrices $M(z_1^i, r, L_i)$, $M(z_2^i, r, L_i), \ldots, M(z_p^i, r, L_i)$. We choose the numbers z_j^i so that a natural number can occur in at most one of these matrices. We write the matrices one under the other to obtain the matrix $\mathscr{M}(r)$. If some column contains less than $q+1$ different symbols, then we repeat $M(z_1^1, r, L_1)$ enough times with all different z's to obtain at least $q+1$ different symbols in every column. We claim that for a suitable choice of r, $\mathscr{I}_{\mathscr{M}(r)pq}=\mathscr{N}$ holds. This is true if (1) $b \notin \mathscr{N}(A)$ implies that $b \notin \mathscr{I}_{\mathscr{M}(r)pq}(A)$ and (2) $b \in \mathscr{N}(A)$ implies that $b \in \mathscr{I}_{\mathscr{M}(r)pq}(A)$.

(1) Let us suppose first that $b \notin \mathscr{N}(A)$ for some $A \subset \Omega$. If $A=\emptyset$, then $b \notin \mathscr{I}_{\mathscr{M}(r)pq}(\emptyset)$ follows from the fact that there are at least $q+1$ different symbols in any column of $\mathscr{M}(r)$. However, if $A \neq \emptyset$, then there exists a chain L_j and a set A_i of that chain satisfying (2.5). We have that $b \notin \mathscr{N}(A_i)=A_{i+1}$, so $b \in A_f \setminus A_{f-1}$, $k \geq f > i+1$ or $b \in \Omega \setminus A_k$ holds. In the first case we use (iii) and in the second case we use (iv) to choose altogether $p(2r+1)$ rows from $M(z_1^j, r, L_j), M(z_2^j, r, L_j), \ldots$, $M(z_p^j, r, L_j)$ so that they contain at most p different symbols in columns of $A \subseteq A_i$, but they contain all different symbols in the column b. Thus, if

$$p(2r+1) \geq q+1, \tag{2.6}$$

then $b \notin \mathscr{I}_{\mathscr{M}(r)pq}(A)$ holds.

(2) Let us suppose now that $b \in \mathscr{N}(A)$. $\mathscr{N}(\emptyset)=\emptyset$ implies that $A \neq \emptyset$. Let us consider an arbitrary chain L_v from $\mathscr{L}$: $A_1, A_2, \ldots, A_k$. Let $i=i(L_v)=\mathrm{k}+1$ if $A \cap (\Omega \setminus A_k) \neq \emptyset$. On the other hand, if $A \cap (\Omega \setminus A_k)=\emptyset$, then let i be the largest index such that $A \cap (A_i \setminus A_{i-1})$ is nonempty. $A \subseteq A_i$ implies that $b \in \mathscr{N}(A) \subseteq \mathscr{N}(A_i)=A_{i+1}$ for $i<k$. For $i=k$ we have that $b \in \mathscr{N}(A_i)=A_i$. Applying (ii), this implies that if there are at most t different symbols in a column of A in

the matrix $M(z_f^v, r, L_v)$, then in the column b there can stand only $t+r$ different values.

Let us choose $q+1$ rows that contain at most p different values in columns of A. These rows could be chosen from at most p different matrices $M(z_f^v, r, L_v)$. Suppose that they are chosen in fact from u ($u \le p$) different matrices. Because there are different symbols in different matrices, we have that in the columns of A there can only stand at most $p-u+1$ different symbols in one matrix, which implies that in one matrix at most $p-u+1+r$ different values are in the column b. Altogether there are at most $u(p-u+1+r)$ different symbols in column b in the u different matrices of type $M(z_f^v, r, L_v)$. If $r \ge p-2>0$, then this number is maximal for $u=p$. Thus, if $r \ge p-2>0$ and

$$p(r+1) \le q, \tag{2.7}$$

then $b \in \mathscr{I}_{\mathscr{M}(r)pq}(A)$ follows.

It is easy to check that for the pairs p, q satisfying (2.3) one can find r which simultaneously satisfies (2.6) and (2.7). □

It is natural to ask the following.

Problem 2.5. Is the statement of Theorem 2.4 true for arbitrary (p,q)-dependencies $(p<q)$? Is it possible to drop the condition $\mathscr{N}(\emptyset)=\emptyset$?

In the case $p=q$ the situation changes significantly. It is shown in [4] that $\mathscr{I}=\mathscr{I}_{Mpp}$ must satisfy an important condition together with (2.2).

Proposition 2.6.

$$\mathscr{I}_{Mpp}(\mathscr{I}_{Mpp}(A)) = \mathscr{I}_{Mpp}(A). \tag{2.8}$$

Proof. The inclusion $\mathscr{I}_{Mpp}(A) \subseteq \mathscr{I}_{Mpp}(\mathscr{I}_{Mpp}(A))$ follows from (2.2). Thus, we only have to prove that $b \in \mathscr{I}_{Mpp}(\mathscr{I}_{Mpp}(A))$ implies $b \in \mathscr{I}_{Mpp}(A)$. Let us consider such a set of rows of M that each column of A contains at most p different numbers in these rows. According to the definition of $\mathscr{I}_{Mpp}$, the same holds for each column of $\mathscr{I}_{Mpp}(A)$, too. This, together with the assumption $b \in \mathscr{I}_{Mpp}(\mathscr{I}_{Mpp}(A))$ implies that there are at most p different values in column b which proves that $b \in \mathscr{I}_{Mpp}(A)$. □

Set functions satisfying (2.2) and (2.8) are called *closures*. It is well known (see Armstrong [1] or in this form [3]), that for $p=1$ the converse of Proposition 2.6 is true, i.e., for every closure $\mathscr{L}$ there exists a matrix M such that $\mathscr{L}=\mathscr{I}_{M11}$ which means that every closure is (1, 1)-representable. We show in the following that this is true for $p=2$, as well, but not true for $p>2$ in general.

First, we recall some well-known concepts and propositions about closures, for detailed proofs see for example [3]. Let $\mathscr{L}$ be a closure on Ω, i.e., $\mathscr{L}: 2^\Omega \to 2^\Omega$ satisfying (2.2) and (2.8). $A \subseteq \Omega$ is called *closed* if $\mathscr{L}(A) = A$. The collection of closed sets is denoted by $\mathscr{Z} = \mathscr{Z}(\mathscr{L})$. The intersection of two closed sets is closed. $\mathscr{L}(A)$ is the intersection of all closed sets containing A. Furthermore, let $\mathscr{M} = \mathscr{M}(\mathscr{Z}(\mathscr{L}))$ denote the collection of those closed sets that cannot be obtained as an intersection of two other closed sets different from them. An arbitrary closed set can be obtained as intersection of some sets from $\mathscr{M}$, consequently $\mathscr{L}(A)$ is equal to the intersection of all members of $\mathscr{M}$ containing A.

Theorem 2.7. *Every closure is* $(2, q)$*-representable if* $2 \leq q$.

Proof. Let $\mathscr{L}$ be a closure on Ω and let $\mathscr{M} = \mathscr{M}(\mathscr{Z}(\mathscr{L})) = \{G_1, G_2, \ldots, G_m\}$. It is easy to see that $G = \mathscr{L}(\emptyset)$ is a subset of every closed set, in particular it is a subset of every G_i. We construct a matrix M.

q rows correspond to every G_i in M, namely the $(qi - q + 1)$th, $(qi - q + 2)$th, ... , (qi)th rows. In the columns of G, 0's are standing in every row. We put i to the positions corresponding to columns of $G_i \setminus G$. In the remaining positions $j + qm$ stands in the jth row. Note that there are at least $q+1$ different numbers in columns of $\Omega \setminus G$. We claim that for this matrix M, $\mathscr{L} = \mathscr{I}_{M2q}$.

Let $A \subset \Omega$ and suppose that $b \notin \mathscr{L}(A)$. Then by the properties of closures there exists a G_i such that

$$A \subseteq \mathscr{L}(A) \subseteq G_i \not\ni b. \tag{2.9}$$

Then in the q rows corresponding to G_i identical elements are standing in the columns of A (either 0 or i), while q different values stand in column b. Let us take a $(q+1)$st row such that it contains a $(q+1)$st different value in column b. Thus, we obtained $q+1$ rows that contain at most two different values in columns of A, but $q+1$ different ones in b, so $b \notin \mathscr{I}_{M2q}(A)$.

On the other hand, let us suppose now that $b \in \mathscr{L}(A)$. Consider $q+1$ rows that contain at most two different values in each column of A. We have to distinguish two cases.

Case 1: $A \subseteq \mathscr{L}(\emptyset)$. In this case $b \in \mathscr{L}(\emptyset)$ holds, as well, hence the column b contains only 0's, so $b \in \mathscr{I}_{M2q}(A)$.

Case 2: $A \setminus \mathscr{L}(\emptyset) \neq \emptyset$. Because A has a column not in the closure of the empty set, there can be at most two different values in each column of A iff the given $q+1$ rows are corresponding to at most two different G_i's. If all the $q+1$ rows correspond to the same G_i and $A \setminus G_i \neq \emptyset$, then the columns of A not in G_i contain $q+1$ different values in these $q+1$ rows, a contradiction. Thus,

$$A \subseteq G_i \tag{2.10}$$

consequently

$$b \in \mathscr{L}(A) \subseteq \mathscr{L}(G_i) = G_i. \tag{2.11}$$

This implies that b contains all identical elements in these $q+1$ rows. On the other hand, if the given $q+1$ rows correspond to two different G_i's, namely to G_i and G_j, then we may assume that at least two rows of the $q+1$ correspond to G_i. If A had a column not in G_i, then there would stand at least three different symbols in that column in the $q+1$ rows, a contradiction. Thus (2.10) and (2.11) again hold and b contains at most q different values in the given rows. This proves that $b \in \mathscr{I}_{M2q}(A)$. $\square$

Let us note, that the $(1,q)$-representability of a closure can be proved in a similar (but easier) way. Now, we show a closure, which is not (p,p)-representable if $p>2$.

Definition 2.8. Let $\mathscr{L}_n^k$ denote the following $2^\Omega \to 2^\Omega$ function:

$$\mathscr{L}_n^k(X) = \begin{cases} X, & \text{if } |X|<k, \\ \Omega, & \text{otherwise.} \end{cases} \tag{2.12}$$

It is easy to see that $\mathscr{L}_n^k$ is a closure.

Theorem 2.9. *If $p>2$ and $n>6$ then $\mathscr{L}_n^2$ is not (p,p)-representable.*

Proof. Let us suppose in the contrary that there exists a matrix M of n columns (p,p)-representing $\mathscr{L}_n^2$. Let us suppose that subject to this condition the number of rows of M is minimal. Because $\mathscr{L}_n^2(\emptyset)=\emptyset$, we have that there are at least $p+1$ different values in each column of M. If all symbols in a column a were different, then $b \in \mathscr{I}_{Mpp}(\{a\})$ would hold for each $b \in \Omega$ that contradicts to the assumption $\mathscr{I}_{Mpp}=\mathscr{L}_n^2$.

Now suppose that the rows r and s both contain identical elements in the columns a and b, respectively. By definition, $c \in \mathscr{L}_n^2(\{a,b\})$ holds for all $c \in \Omega$. Let us choose $p-1$ rows additionally to r and s such that they contain all different values in c and those values are different from the values of r and s. (This is possible, because there are at least $p+1$ different numbers in column c.) In these $p+1$ rows a and b take at most p different values. Thus, by $\mathscr{I}_{Mpp}=\mathscr{L}_n^2$, c takes at most p different values, too. This can only happen if r and s agree in c, hence r and s are identical rows that contradicts the minimality of M. We obtained that two rows may agree in at most one column.

Let us suppose now that rows t and u agree in the first column, while rows r and s agree in the second column ($t \neq u$, $r \neq s$). By the previous paragraph $\{t,u\} \neq \{r,s\}$, so we only have to consider the following two cases: (i) $|\{t,u,r,s\}|=3$; (ii) all the four rows are distinct.

(i) The first and second columns contain at most two different values in these three rows. Because $\mathscr{I}_{Mpp}=\mathscr{L}_n^2$ any other column contains at most two different values in these rows. If the number of columns is larger than three, then there must exist two columns that agree in the same pair of rows that contradicts the conclusion of the previous paragraph.

(ii) Using that $p>2$ one can see that every column contains at most three different values in rows r, s, t, u. There are six possibilities for a column to contain identical elements in two of these four rows, so for $n>6$ we can apply the pigeon hole principle to obtain a pair of different columns that contain identical elements in the same pair of rows, a contradiction. $\square$

In the following we give a certain characterization of (p,p)-representable closures. First we need a definition.

Definition 2.10. Let $\mathscr{B}=\{A_{i,j}\}$ be a system of subsets of an n-element set X, where $1\le i<j\le m$. We say that $\mathscr{B}$ satisfies the *triangle condition* if for all $i<j<k$ the intersection of any pair of $A_{i,j}$, $A_{j,k}$ and $A_{i,k}$ is contained in the third set.

The following lemma can be proved by an easy greedy construction.

Lemma 2.11. *Let $\mathscr{B}=\{A_{i,j}\}$ be a system of subsets of an n-element set X, where $1\le i<j\le m$. There exists an $m\times n$ matrix M such that its ith and jth rows agree exactly in the columns corresponding to $A_{i,j}$ iff $\mathscr{B}$ satisfies the triangle condition.*

Theorem 2.12. *The closure $\mathscr{L}$ is (p,p)-representable if and only if there exists a system of subsets of Ω, $\mathscr{B}=\{A_{i,j}\}$ $(1\le i<j\le m)$ such that it satisfies the triangle condition, the following sets are all closed by $\mathscr{L}$:*

$$\bigcup_{0\le r<s\le p} A_{j_r,j_s} \tag{2.13}$$

(where $1\le j_0,j_1,\dots,j_p\le m$ are arbitrarily fixed integers) and every $\mathscr{L}$-closed set can be obtained as intersection of sets of type (2.13).

In order to prove Theorem 2.12 we need the following easily checked lemma.

Lemma 2.13. *Let M be a matrix of m rows and suppose that the ith and jth rows of M agree in the column set $A_{i,j}$. Then $A\subseteq\Omega$ is closed according to $\mathscr{I}_{Mpp}$ if and only if it is an intersection of sets of type* (2.13).

Proof of Theorem 2.12. If $\mathscr{L}$ is (p,p)-representable, then the representing matrix M defines the set system $\{A_{i,j}\colon 1\le i<j\le m\}$ by that the ith and jth rows of M agree in the column set $A_{i,j}$. By Lemma 2.13, $A_{i,j}$'s satisfy the triangle condition. A set of type (2.13) is trivially an intersection of sets of type (2.13) (one-element intersection), so by Lemma 2.13 it is closed. It also follows from Lemma 2.13 that every closed set is an intersection of sets of type (2.13).

On the other hand, if there exist sets $\{A_{i,j}\}$ satisfying the condition of the theorem, then by the triangle condition we have a matrix M such that the ith and jth rows of M agree in the column set $A_{i,j}$. The $\mathscr{L}$-closed sets can be obtained as intersections of sets of type (2.13) by the conditions of the theorem. Conversely, nonclosed sets cannot be obtained because (2.13) type sets are all $\mathscr{L}$-closed and intersection of $\mathscr{L}$-closed sets is $\mathscr{L}$-closed, too. Thus, Lemma 2.13 completes the proof. $\square$

Even though the conditions of Theorem 2.12 are not algorithmically effective, it yields nice theoretical results like the following corollary.

Corollary 2.14. *Let $\mathscr{L}$ be a closure such that $\mathscr{M}=\mathscr{M}(\mathscr{Z}(\mathscr{L}))=\{G_1,G_2,\dots,G_t\}$ is closed under taking unions. Then $\mathscr{L}$ is (p,p)-representable for every p.*

Proof. Let $t\geq p$ (if $t<p$ then we repeat G_1 enough times to obtain at least p sets). We apply Theorem 2.12 with $m=2t$, $A_{2i-1,2i}=G_i$ $(1\leq i\leq t)$, while the other $A_{i,j}$'s are empty. $\square$

The next easy proposition shows that a closure is either (p,p)-representable only for finitely many p's, or (p,p)-representable for every large enough p. We omit its quite straightforward proof.

Proposition 2.15. *Let $\mathscr{L}$ be a closure on the n-element set Ω. Furthermore, let $N\geq 2n-3$ and suppose that $\mathscr{L}$ is (N,N)-representable. Then $\mathscr{L}$ is (p,p)-representable for all $p>N$.*

Summarizing, the question remained basically open:

Problem 2.16. Find an algorithmically good characterization of (p,p)-representable closures.

We have already shown that every closure is $(2,q)$-representable if $q\geq 2$. Furthermore, we can apply Theorem 2.4 for closures, too, because they are special increasing-monotone functions. However, we are able to utilize the additional properties of closures to prove the following.

Proposition 2.17. *Let $\mathscr{L}$ be a closure on Ω. If $3\leq p$ and $((p+1)/2)^2\leq q$, then $\mathscr{L}$ is (p,q)-representable.*

Proof. Let $\mathscr{M}(\mathscr{Z}(\mathscr{L}))=\{G_1,G_2,\dots,G_r\}$ and $G=\mathscr{L}(\emptyset)$. We construct a matrix M similarly to the proof of Theorem 2.7. There correspond $q+1$ rows to each G_i in M, namely the $((i-1)(q+1)+1)$st, $((i-1)(q+1)+2)$nd, $\dots,(i(q+1))$th. If $(i-1)(q+1)+1\leq j\leq i(q+1)$, i.e., row j belongs to G_i, then in this row 0 stands in the columns of G, i stands in the columns of $G_i\setminus G$ and $(q+1)r+j$ stands in the other columns. Let $A \subseteq \Omega$ be an arbitrary subset and let us suppose first that $b \notin \mathscr{L}(A)$. Then there exists an i such that $\mathscr{L}(A) \subseteq G_i \not\ni b$ holds. The $q+1$ rows corresponding to G_i contain all identical elements in columns of A (either 0 or i), but the values in b are all different. This shows that $b \notin \mathscr{I}_{Mpq}(A)$.

On the other hand, let us suppose that $b \in \mathscr{L}(A)$ and take $q+1$ rows that contain at most p different symbols in columns of A. Suppose that these rows belong to exactly u different G_i's ($u \le p$). Then the rows corresponding to the same given G_j contain at most $p-u+1$ different values in columns of A. We claim that b cannot contain more distinct numbers in rows belonging to a given G_j, than the maximum for columns of A. Indeed, if $A \not\subseteq G_j$, then there is a column of A that contains all different values in the rows belonging to G_j. On the other hand, if $A \subseteq G_j$, then by $b \in \mathscr{L}(A) \subseteq \mathscr{L}(G_j) = G_j$, all identical elements are standing in b. Thus, at most $u(p-u+1)$ different symbols stand in b in the chosen $q+1$ rows.

$$u(p-u+1) \le \left(\frac{p+1}{2}\right)^2 \le q \tag{2.14}$$

implies that $b \in \mathscr{I}_{Mpq}(A)$ holds, as well. □

It is natural to ask the following.

Problem 2.18. Is every closure (p, q)-representable if $p<q$? Or even more, is every increasing-monotone function (p, q)-representable if $p<q$?

For closures the smallest open case is $p=4$, $q=5$, while for increasing-monotone functions $p=2$, $q=3$. It is not hard to check that an argument similar to those above yields that if p divides $q+1$, then every closure is (p, q)-representable.

The next problem seems to be somewhat easier, than the previous ones. Let $\mathscr{N}$ be an increasing-monotone function on the set Ω. A set K is called a *key* if $\mathscr{N}(K) = \Omega$. K is a *minimal key* if it is a key and no proper subset of it is a key. It is easy to check that there cannot be inclusion between two minimal keys, so the system of minimal keys $\mathscr{K}$ satisfies the Sperner condition:

$$K_1, K_2 \in \mathscr{K}, \quad K_1 \ne K_2 \quad \Rightarrow \quad K_1 \not\subset K_2. \tag{2.15}$$

In this case $\mathscr{K}$ is called a *Sperner family*. We say that a Sperner family on Ω is (p, q)-representable ($p<q$) if there exists an increasing-monotone function on Ω that is (p, q)-representable and its system of minimal keys is exactly $\mathscr{K}$. The definition of (p, p)-representation of a Sperner family is analogous, we just have to look for a closure.

Problem 2.19. Is every nonempty Sperner family (p, q)-representable for any $p<q$? Which Sperner families are (p, p)-representable for all p?

3. Implications among (p, q)-dependencies

In this section we investigate the connections between (p,q)-dependencies for various p's and q's.

Definition 3.1. Let $(p,q) \to (p',q')$ denote the property that $b \in \mathscr{I}_{Mpq}(A)$ implies $b \in \mathscr{I}_{Mp'q'}(A)$ for every matrix M. Let $(p,q) \overset{m}{\Longrightarrow} (p',q')$ denote the above implication when we require only for matrices that have at least m different values in each of their columns.

The proof of the following lemma is obvious.

Lemma 3.2.

$$\begin{aligned} &(p,q) \to (p,q+1), \\ &(p,q) \to (p-1,q). \end{aligned} \tag{3.1}$$

We can say more, if we assume that the matrix M contains at least m different values in each of its columns.

Lemma 3.3. *We have that*

$$(p,q) \quad \overset{q+1}{\Longrightarrow} \quad (p-1,q-1), \tag{3.2}$$

but

$$(p,q) \quad \overset{q}{\not\Longrightarrow} \quad (p-1,q-1). \tag{3.3}$$

Proof. In order to prove (3.2) let us assume that $b \in \mathscr{I}_{Mpq}(A)$ in some matrix M. We want to prove that $b \in \mathscr{I}_{Mp-1q-1}(A)$ holds, as well. If it did not hold, then there would exist q rows of the matrix such that they contain at most $p-1$ different values in each column of A, but q different symbols in b. By assumption, there are at least $q+1$ distinct numbers in the column b, so we may choose a $(q+1)$st row that contains a $(q+1)$st different value in b. This, together with the previous q rows would form $q+1$ rows that contain at most p different values in columns of A, but all different values in b that contradicts to the assumption $b \in \mathscr{I}_{Mpq}(A)$.

On the other hand, a matrix that contains exactly q different values in column b and at most $p-1$ different symbols in columns of A proves (3.3). $\square$

Lemma 3.4.

$$(p,q) \quad \overset{q}{\not\Longrightarrow} \quad (1,q-1). \tag{3.4}$$

Proof. The following matrix is a counterexample giving the proof. The first column represents columns of A, while the second one represents column b. $\square$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & q \\ 2 & 1 \\ 2 & 2 \\ \vdots & \vdots \\ 2 & q \\ \vdots & \vdots \\ q & 1 \\ q & 2 \\ \vdots & \vdots \\ q & q \end{bmatrix}$$

Lemma 3.5. *If $p<q$ then*

$$(p,q) \quad \overset{m}{\not\Longrightarrow} \quad (1,q-p). \tag{3.5}$$

Proof. The following matrix is a counterexample giving the proof. The first column represents columns of A, while the second one represents column b. □

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & q-p+1 \\ 2 & q-p+1 \\ \vdots & \vdots \\ m & q-p+1 \end{bmatrix}$$

Lemma 3.6. *If $p<N$ then*

$$(p,q) \quad \overset{m}{\not\Longrightarrow} \quad (p+1,N). \tag{3.6}$$

Proof. First we give a construction that shows $(p,q) \not\to (p+1,N)$. The matrix has $N+1$ rows, which contain numbers $1,2,3,\ldots,N+1$ in column b, respectively. The numbers $1,2,\ldots,p+1$ may stand in columns of A. Let the columns of A be constructed in such a way that for any $p+1$ rows there exists a column that contains

$p+1$ different numbers in those rows. This can be done if A has enough columns. It is easy to see that in the so constructed matrix M, $b \in \mathcal{I}_{Mpp}(A)$, hence $b \in \mathcal{I}_{Mpq}(A)$ according to Lemma 3.2. However, $b \notin \mathcal{I}_{Mp+1N}(A)$ holds.

In order to prove (3.6) we only have to modify M so that each column would contain at least m different values. Let us write all $N+1+i$ in the $(N+1+i)$th row $(1 \leq i \leq m-N-1)$. This modification does not change the above property. □

Now we can say when a (p,q)-dependency implies an other in the sense of Definition 3.1.

Theorem 3.7. *Let $m>q$. Then*

$$(p,q) \overset{m}{\Longrightarrow} (p',q') \tag{3.7}$$

holds if and only if $1 \leq p' \leq p$ and $q-p \leq q'-p'$. On the other hand if $m \leq q$, then the necessary and sufficient condition for implication (3.7) *is $1 \leq p' \leq p$ and $q \leq q'$.*

Proof. The statement follows easily from Lemmas 3.2–3.6. □

Note, that in the proof of Lemma 3.6, A must be large. This means that for relatively small A's, $b \in \mathcal{I}_{Mpq}(A)$ implies $b \in \mathcal{I}_{Mp+1N}(A)$. Thus, if $|\Omega| = n$ is less than that bound for example, then the above implication holds for each A. This motivates the following problem.

Problem 3.8. What is the size bound for A that $b \in \mathcal{I}_{Mpq}(A)$ implies $b \in \mathcal{I}_{Mp+1N}(A)$ for all M (p, q and N are fixed)?

We give the solution for two special cases without proof.

Proposition 3.9. *If $|A| < \lceil \log(N+1) \rceil$, then $b \in \mathcal{I}_{M11}(A)$ implies $b \in \mathcal{I}_{M2N}(A)$ for all matrices M. However, if $|A| \geq \lceil \log(N+1) \rceil$, then this implication does not hold.*

Proposition 3.10. *If*

$$|A| < \left\lceil \frac{q+2}{2(q-p+1)} \right\rceil, \tag{3.8}$$

then $b \in \mathcal{I}_{Mpp}(A)$ implies $b \in \mathcal{I}_{Mp+1q+1}(A)$ for all matrices M, but if A is larger than (3.8), *then the implication is not true.*

References

[1] W.W. Armstrong, Dependency structures of database relationships, in: Information Processing 74 (North-Holland, Amsterdam, 1974) 580–583.

[2] E.F. Codd, A relational model of data for large shared data banks, Comm. ACM 13 (1970) 377–387.

[3] J. Demetrovics and G.O.H. Katona, Combinatorial problems of database models, in: Algebra, Combinatorics and Logic in Computer Science, GyHor, Colloquia Mathematica Societatis János Bolyai 42 (North-Holland, Amsterdam, 1983) 331–353.

[4] J. Demetrovics and G.O.H. Katona, Extremal combinatorial problems of database models, in: Proceedings MFDBS 87, Dresden, Lecture Notes in Computer Science 305 (Springer, Berlin, 1987) 99–127.