COLLOQUIA MATHEMATICA SOCIETATIS JÁNOS BOLYAI
42. Algebra, Combinatorics and Logic in Computer Science
Győr (Hungary), 1983.

COMBINATORIAL PROBLEMS OF DATABASE MODELS

J. DEMETROVICS - G.O.H. KATONA

1. INTRODUCTION

One possible model of a database is a matrix. E.g. a database may contain the name, the place of birth, the date of birth, and so on .... of different persons. The possible data are called *attributes* while the whole of the data of one individual is its *record*. In the above example the name, the place of birth, the date of birth are attributes. The whole of data of one person is a record. It is rather natural to describe this system by a matrix whose rows and columns correspond to the records and attributes, respectively.

It is clear, that the actual entries of the matrix are usually unimportant. We have to consider, whether they are equal or unequal, only. By further relaxations we will use other less fine models (mostly known from the literature). These models suggest several combinatorial problems of extremal nature. Namely, given some parameter(s) of the database determine the maximum or minimum of another parameter. The aim of the recent paper is to survey these extremal combinatorial results found by the authors and their coauthors.

In Section 2 we study the models and their relationships. The models are of combinatorial nature. Concepts like *closures* are borrowed from the combinatorial literature.

In Section 3 we survey the extremal combinatorial results concerning one database.

On the other hand, Section 4 considers a partial ordered set in which the databases are ordered in a very natural way. Some extremal problems concerning this partial ordered set are solved.

2. DATABASE MODELS

The concepts and results of this section are either published by other authors (see e.g. [1],[4]) or belong to the folklore. We do not regard them as our original work. However, we did not find the structure in the form presented here. Our approach expresses our personal taste as well as it corresponds to what is needed in our subsequent investigations.

The basic model of a database is a matrix M with m rows and n columns. The set of columns is denoted by Ω. Let $A,B \subseteq \Omega$. We say that B *depends on* A if M has no two rows equal in A but different in B.

The notation $A \rightarrow B$ is used for this case. In other words $A \rightarrow B$ means that the data in the columns belonging to A uniquely determine the data in B, that is, knowing A, B does not give any new information. If B is a one-element set, $B = \{b\}$, we write simply $A \rightarrow b$.

It is clear that different matrices can give the same system of dependencies, therefore this model is less fine than the matrix-model. However, in many cases it contains sufficient amount of information.

Define the function L on 2^{Ω} by

$$L(A) = \{b: A \rightarrow b\}.$$

This function possesses some simple properties:

Lemma 2.1. *Let* $A,B \subseteq \Omega$. *Then*

(2.1) $A \subseteq L(A)$;

(2.2) $A \subseteq B$ *implies* $L(A) \subseteq L(B)$;

(2.3) $L(L(A)) = L(A)$.

Proof. (2.1) is obvious. It means that $A \to b$ holds for all $b \in A$. Indeed, if two rows are equal in A, they must be equal in b, as well.

To prove (2.2) suppose that $a \in L(A)$, that is, $A \to a$. In other words, any two rows which are equal in A, coincide also in a. $A \subseteq B$ implies that A can be replaced by B in the latter statement, so $B \to a$, that is, $a \in L(B)$ as we wanted to show.

The part $L(L(A)) \supseteq L(A)$ is a consequence of *(2.1)*. We have to prove $L(L(A)) \subseteq L(A)$, only. Let $a \in L(L(A))$. Then any two rows equal in $L(A)$ are also equal in a. Consider now two rows known to be equal in A. By definition, these two rows must be equal in $L(A)$, therefore in a. $a \in L(A)$ is proved. The proof is complete.

The combinatorial literature calles a function satisfying (2.1)-(2.3) a *closure*. Lemma 2.1 makes us able to call L a closure.

Now we consider another relation between the closures and the dependencies. The next lemma can be easily proved:

Lemma 2.2. *Let* $A,B \subseteq \Omega$. $A \to B$ *iff* $B \subseteq L(A)$.

Lemmas 2.1 and 2.2 imply the following properties of the dependencies.

Lemma 2.3. *Let* $A,B,C \subseteq \Omega$.

(2.4) $A \to A$;

(2.5) $A \to B$ *and* $B \to C$ *imply* $A \to C$;

(2.6) $A \subseteq C$, $D \subseteq B$ *and* $A \to B$ *imply* $C \to D$;

(2.7) $A \to B$ *and* $C \to D$ *imply* $A \cup C \to B \cup D$.

Proof. (2.4) is a consequence of Lemma 2.2 and (2.1).

By Lemma 2.2, $A \to B$ can be written in the form $B \subseteq L(A)$. (2.2) implies $L(B) \subseteq L(L(A))$ and hence we have $L(B) \subseteq L(A)$ because of (2. 3). $B \to C$ is equivalent to $C \subseteq L(B)$, therefore $C \subseteq L(A)$ follows. This yields $A \to C$, again by Lemma 2.2. (2.5) is proved.

Prove now (2.6). $A \to B$ is equivalent to $B \subseteq L(A)$. $D \subseteq B$ implies $D \subseteq L(A)$. (2.2) and $A \subseteq C$ result in $L(A) \subseteq L(C)$, and hence we have $D \subseteq L(C)$ which is equivalent to the desired $C \to D$.

The conditions of (2.7) can be rewritten into the forms $B \subseteq L(A)$ and $D \subseteq L(C)$. Hence we obtain $B \cup D \subseteq L(A) \cup L(C)$. (2.2) yields $L(A) \subseteq L(A \cup C)$ and $L(C) \subseteq L(A \cup C)$ can be obtained similarly. These imply $B \cup D \subseteq L(A) \cup L(C) \subseteq L(A \cup C)$ which is equivalent to $A \cup C \to B \cup D$. The lemma is proved.

Suppose now, in general, that a system of pairs (A,B) of subsets of Ω is given which satisfies the conditions (2.4)-(2.7). Such a system is called a *full family*. Lemma 2.3 expresses the fact that the dependencies form a full family.

In this way we associated a full family with each matrix. It is easy to see that the same full family can be associated with several different matrices. On the other hand, as we will see later, there is at least one matrix to any full family.

Let T be a full family. It defines the closure $L(A) = \{b: (A,b) \in T\}$. This definition is a generalization of the earlier definition, where we have done it only for the full family of dependencies. Lemma 2.1 can be proved, of course, in full generality. The proof is the same, we have to use properties (2.4)-(2.7) in place of the properties of the matrices. As an example let us see the proof of (2.3). $L(L(A)) \supseteq L(A)$ is a consequence of (2.1). It remains to prove $L(L(A)) \subseteq L(A)$, only. Let $a \in L(L(A))$. By definition, this is equivalent to $(L(A),a) \in T$. Consider all the pairs $(A,b) \in T$ (A is fixed). Apply (2.7) to them. We obtain $(A, \{b: (A,b) \in T\}) = (A, L(A))$. The application of

(2.5) to $(A,L(A))$ and $(L(A),a)$ leads to $(A,a) \in T$ which is equivalent to the desired $a \in L(A)$.

Conversely, we can associate a full family with each closure L. Let $(A,B) \in T$ iff $B \subseteq L(A)$. Lemma 2.3 proves, in fact, that T will be really a full family. Observe, that in this way we found a one-to-one correspondence between the closures and the full families:

Theorem 2.4.

$$(2.8) \qquad T \to L(A) = \{b: (A,b) \in T\}$$

is a one-to-one correspondence between the set of full families and the set of closures on the same ground set. The inverse of (2.8) is determined by

$$(2.9) \qquad L \to T = \{(A,B): B \subseteq L(A)\}.$$

The structure of the full family of a given database can be very useful in handling and compressing the database. However, this structure, as we saw, could be uniquely characterized by the corresponding closure. This latter one is a simpler structure. In what follows, we will give other (even simpler) equivalent structures

The set $A \subseteq \Omega$ is called *closed* (with respect to the closure L) if $L(A) = A$. The next theorem determines the family Z $(=Z(L))$ of the closed sets with respect to the closure L.

Theorem 2.5. *Let Z be a family of different subsets of Ω. Z is the family of closed sets with respect to some closure L iff*

$$(2.10) \qquad \Omega \in Z,$$

$$(2.11) \qquad A,B \in Z \text{ implies } A \cap B \in Z.$$

(2.11) can be formulated as *"Z is closed under intersection"*.

Proof. Prove first that (2.10) and (2.11) hold for the closed sets of an L.
(2.1) implies $\Omega \subseteq L(\Omega)$, but $L(\Omega) \subseteq \Omega$ is obvious. Hence $L(\Omega) = \Omega$ follows. On the other hand, suppose $A,B \in Z$. They yield

$L(A) = A$, $L(B) = B$. $A \cap B \subseteq A$ and (2.2) imply $L(A \cap B) \subseteq L(A) = A$. $L(A \cap B) \subseteq B$ can be deduced similarly. The two inclusions result in $L(A \cap B) \subseteq A \cap B$. The converse inclusion is a consequence of (2.1). $A \cap B$ is really closed and is in Z.

On the other hand, suppose that the family Z satisfies (2.10) and (2.11). We have to construct a closure L in which the closed sets are exactly the members of Z.

Let

$$L(A) = \bigcap_{A \subseteq Z \in Z} Z.$$

It is easy to see that this is a closure.

This intersection is non-void by (2.10). On the other hand, (2.11) implies $L(A) \in Z$. That is, the sets closed with respect to L all belong to Z. We have to prove that any $B \in Z$ is closed with respect to L. Indeed, B occurs in the right hand side of $L(B) = \cap Z$ and all other terms Z contain B. Hence $L(B) = B$. The proof is complete.

The families Z satisfying (2.10) and (2.11) are called *intersection semi-lattices.*

Theorem 2.6.

$$(2.12) \qquad L \to Z = \{Z\colon L(Z) = Z\}$$

is a one-to-one correspondence between the set of closures and the set of intersection semi-lattices on the same ground set. The inverse of (2. 12) is determined by

$$(2.13) \qquad Z \to L(A) = \bigcap_{A \subseteq Z \in Z} Z.$$

Proof. We have proved in Theorem 2.5 that (2.12) leads to all intersection semi-lattices. Let us verify that two different closures give different intersection semi-lattices. Take two different closures L_1 and L_2. For some $A \subseteq \Omega$, we have $L_1(A) \neq L_2(A)$. There is an element a which is in exactly one of them, say $a \in L_2(A)$, $a \notin L_1(A)$. Consider $L_2(L_1(A))$. Apply (2.2) to $A \subseteq L_1(A)$: $L_2(A) \subseteq L_2(L_1(A))$. Hence $a \in L_2(L_1(A))$. However $a \notin L_1(A) = L_1(L_1(A))$. The closures of $L_1(A)$ with respect to L_1 and L_2, resp., are different, that is (2.12) determines two different intersection semi-lattices.

The proof that (2.13) is the inverse of (2.12) is left to the reader. The proof is finished.

Due to property (2.11) an intersection semi-lattice Z can be determined by much less of its members. Let $extr(Z)$ denote the family of those members $C \in Z$ which are not intersections of two other members of Z, that is, $A \neq C \neq B$, $A,B \in Z$ imply $A \cap B \neq C$.

Lemma 2.7. *Any member $A \in Z$ is an intersection of some (≥ 0) members of $extr(Z)-\{\Omega\}$, but no proper subfamily of $extr(Z)-\{\Omega\}$ has this property.*

Proof. Let us prove that $A \in Z$ is an intersection of some members of $extr(Z)-\Omega$ in an indirect way. If there is an $A \in Z$ which is not such an intersection, take a maximal one. Ω is the empty intersection, hence $A \neq \Omega$. $A \notin extr(Z)-\Omega$ is obvious. Therefore $A = B \cap C$ where $B \neq A \neq C$ and $B,C \in Z$ hold. These sets are larger than A, so, by the maximality of A, they are intersections of some members of $extr(Z)-\Omega$. Substituting these intersections into $B \cap C$ we obtain a contradictory intersection form for A. The first statement is proved.

Let now $A \in extr(Z)-\{\Omega\}$. We will prove that $extr(Z)-\{\Omega,A\}$ is not sufficient to generate all members of Z. Namely, A makes the difficulties. If, on the contrary A is an intersection of members of $extr(Z)-\{\Omega,A\}$, take the one containing the minimum number of terms. This number cannot be 0 or 1. Therefore we can write $A = B \cap (\cap C)$ where B and all $C \in extr(Z)-\{\Omega,A\}$. By the minimality, we have $A \neq \cap C \neq \Omega$. $\cap C \in Z$ is obvious. Therefore A is an intersection of two members of Z, different from A. This contradiction proves the lemma.

The next theorem characterizes the families $extr(Z)$.

Theorem 2.8. *A family* E *is equal to* $extr(Z)$ *for some intersection semi-lattice* Z *iff*

$$\Omega \in E \tag{2.14}$$

and

$$A = \bigcap_{i=1}^{r} A_i \quad (r \geq 1),\ A,\ A_1,\dots,A_r \in E \tag{2.15}$$

$$\text{imply } A = A_i \text{ for some } i (1 \leq i \leq r).$$

Proof. Let Z be an intersection semi-lattice. Then $\Omega \in extr(Z)$ is a consequence of (2.10). To prove (2.15) for $extr(Z)$ suppose that $A = \bigcap_{i=1}^{r} A_i$ $(r \geq 1)$, $A, A_1, \dots, A_r \in extr(Z)$. Let s be the smallest integer satisfying $A = \bigcap_{i=1}^{s} A_i$. Then $A = (A_1 \cap \dots \cap A_{s-1}) \cap A_s$, where $A_1, \dots, A_{s-1}, A_s, A_1 \cap \dots \cap A_{s-1} \in Z$, $A_1 \cap \dots \cap A_{s-1} \neq A$. Here $A \in extr(Z)$ implies $A_s = A$. (2.15) is proved.

Suppose now that E satisfies (2.14) and (2.15) and construct a Z such that $extr(Z) = E$. Let Z consist of all possible intersections formed from E. We have to show that the members A of E are not intersections of two members of Z different from A, but this can be done with the members of $Z - E$.

Let $A \in E$, $A = B \cap C$, $B, C \in Z$. As B and C are intersections of members of E we obtain such an intersection form for A. (2.15) implies that one of the terms is A. Consequently either B or C is also equal to A.

On the other hand, if $A \in Z - E$ then we can write $A = A_1 \cap, \dots \cap A_r$, $r \geq 2$, $A_1, \dots, A_r \in E$, $A_1, \dots, A_r \neq A$. Suppose that r is chosen minimally. Then $A_1 \cap \dots \cap A_{r-1} \neq A$, therefore $A = (A_1 \cap \dots \cap A_{r-1}) \cap A_r$ is an intersection of two members of Z, different from A. The proof is complete.

The families satisfying (2.14) and (2.15) are called *intersection-free families*.

Theorem 2.9.

(2.16) $Z \to extr(Z)$

is a one-to-one correspondence between the set of intersection semi-lattices and the set of intersection-free families. The inverse of (2.16) is determined by

(2.17) $E \to \{A_1 \cap \ldots \cap A_r,\ \ r \geq 1,\ \ A_1,\ldots,A_r \in E\}.$

Proof. Let Z_1 and Z_2 be two different intersection semi-lattices and suppose that $A \in Z_2$, $A \notin Z_1$. By Lemma 2.7, A is an intersection of members of $extr(Z_2)$. On the other hand, $extr(Z_1) \subseteq Z_1$ cannot contain A. This proves that $extr(Z_1) \neq extr(Z_2)$, that is, (2.16) is one-to-one.

The proof of (17) is left to the reader.

We have shown several concepts equivalent to the full families, however we have not proved yet, that the full families are exactly the dependencies.

Theorem 2.10. *Let T be a full family on Ω. There exists a matrix M with $|\Omega|$ columns in which the set of dependencies coincides with T.*

Proof. Let L and Z be the closure and the intersection semi-lattice associated with T. List the members of $extr(Z) = \{G_1,\ldots,G_r\}$. M will have $r+1$ rows: the 0-th row consists of zeros, while the i-th $(1 \leq i \leq r)$ row contains a 0 or i in the column c according to $c \in G_i$ or $c \notin G_i$.

Suppose first that $(A,B) \in T$. Then $B \subseteq L(A)$ follows. We have to prove that the equality of two rows in A implies their equality in B. By the definition of M, two rows can be equal only in zeros. Two cases are distinguished:

a) The 0-th and i-th $(1 \leq i \leq r)$ rows are equal in A. $A \subseteq G_i$ follows. G_i is closed with respect to L, so $B \subseteq L(A) \subseteq L(G_i) = G_i$, by (2.2). Hence the i-th row has 0 everywhere in B. The rows are equal in B.

b) The i-th and j-th $(1 \leq i < j \leq 1)$ rows are equal in A. $A \subseteq G_i$ and $A \subseteq G_j$ imply $B \subseteq G_i$ and $B \subseteq G_j$ in the above way. The two rows have everywhere 0 in B.

Suppose now that $(A,B) \notin T$, that is, $B \nsubseteq L(A)$. There is a column $b \in B$ such that $b \notin L(A)$. $L(A)$ is closed (by (2.3)), therefore it is an intersection of some G_i's. There must exist a G_i satisfying $b \notin G_i$. However $G_i \supseteq L(A) \supseteq A$ holds. The 0-th and i-th rows of M have zeros in A, but they are different in b. The proof is complete.

3. INEQUALITIES FOR THE PARAMETERS OF A DATABASE

Let us first indicate what kind of problems are discussed in this section. A database (or matrix) has different parameters like the number of columns (attributes), number of rows (individuals), number of possible entries of the matrix, etc. More generally, the total structure T of dependencies can also be considered as a "generalized parameter". Knowing some parameters of M we will look for the minimum or maximum value of another parameter.

The easiest example concerns the number of minimal keys of a database. A *key* is a set of attributes determining the values of all other attributes. Formally, $A \subseteq \Omega$ is a key, iff $A \rightarrow \Omega$, that is, if (A,Ω) is a dependency or equivalently, $L(A) = \Omega$. Moreover, K is a *minimal key* iff $L(K) = \Omega$, but no proper subset of it has this property. Our first problem is to determine the maximum number of minimal keys if the number n of attributes is given. For this purpose we need the following theorem.

Theorem 3.1. ([5]) *Given a family* K *of subsets of* Ω *there is a matrix in which the family of minimal keys is* K *iff* K *satisfies*

$$(3.1) \quad K_1, K_2 \in \mathsf{K}, K_1 \neq K_2 \Rightarrow K_1 \nsubseteq K_2.$$

(The families satisfying (3.1) are called *Sperner* families.)

Proof. The necessity of (3.1) follows by the minimality of the members of K.

We could deduce the sufficiency from Theorem 2.10, but the proof is easier in this case. Introduce

$K^{-1} = \{B: \nexists K \in K \text{ s.t. } K \subseteq B \text{ and } B \text{ is maximal for this property}\}$.

Let $K^{-1} = \{G_1,\dots,G_m\}$. The desired matrix consists of $n = |\Omega|$ columns and $m+1$ rows. The 0-th row consists of zeros, while the i-th $(1 \leq i \leq r)$ row contains a 0 or i in the column c according to $c \in G_i$ or $c \notin G_i$.

Let A be subset of Ω not containing any member of K as a subset. Then, by the definition of K^{-1}, there is an i such that $A \subseteq G_i$ holds. Hence the 0-th and i-th rows are equal in A. Therefore A is not a key in M.

On the other hand, if $A \supseteq K \in K$, then $A - G_i \neq \emptyset$ holds for all i $(1 \leq i \leq r)$. Therefore the i-th row has an i in at least one column belonging to A.

The family of keys in M is really equal to the family of all supersets of the members of K. This means that the family of minimal keys in M is exactly K. The proof is complete.

Our original problem is reduced to the determination of $max|K|$ under (3.1). This problem, however, has been solved many years ago by Sperner [9]. The answer is $\binom{n}{\lfloor \frac{n}{2} \rfloor}$. So, we can formulate

Theorem 3.2. *([5]) The maximum number of minimum keys in a database (matrix) with n attributes (columns) is* $\binom{n}{\lfloor \frac{n}{2} \rfloor}$.

Theorem 3.1 states that there is a matrix M, for any K, in which the family of minimal keys is K. We did not consider the number of rows. Now, let $s(K)$ denote the minimum number of rows in such a matrix M, where K is any Sperner family.

Theorem 3.3. *([5])*

$$s(K) \leq 1 + \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Proof. It is clear from the proof of Theorem 3.1 that the number of rows of M is at most $1 + |K^{-1}|$. However, K^{-1} is a Sperner-family, again. Hence $|K^{-1}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$ proves the theorem.

The next theorem states that there is a K for which $s(K)$ is close to the upper bound given in Theorem 3.3.

Theorem 3.4. *([7]) For any n, there is a Sperner family on an n-element set such that*

$$\frac{1}{n^2} \binom{n}{\lfloor \frac{n}{2} \rfloor} < s(K).$$

The proof is non-trivial, it can be found in [7].

We cannot construct a Sperner family which has such a large $s(K)$. There is, however a class of Sperner-families for which $s(K)$ is exactly determined. Let F_k^n denotes the family of all k-element sets of an n-element set. It is obvious that F_k^n is a Sperner family. Let us see first an easy lemma:

Lemma 3.5. *([6])*

$$\binom{s(F_k^n)}{2} \geq \binom{n}{k-1} \qquad (0 < k \leq n).$$

Proof. Suppose that the family of minimal keys of an $m \times n$ matrix M is K. Let A be a $(k-1)$-element set of columns of M. There is a pair of different rows which are equal in A, since A is not a key. However, to different sets A the corresponding pairs should be different. Hence we have $\binom{m}{2} \geq \binom{n}{k-1}$ proving the lemma.

$s(F_1^n) \geq 2$ follows. The next trivial construction proves the equality:

$$\begin{pmatrix} 0 & \dots & 0 \\ 1 & \dots & 1 \end{pmatrix}$$

If the family of minimal keys of an $m \times n$ matrix is F_2^n then, by Lemma 3.5,

$$n \leq \binom{m}{2}. \tag{3.2}$$

Conversely, we show that (3.2) implies the existence of the $m \times n$ matrix having F_2^n as column of the set of minimal keys. Let each M contain exactly two zeros. In different columns the zeros are placed differently. The feasibility is ensured by (3.2). All other elements of the i-th row will be i. It is easy to check that the family of minimal keys in this matrix is F_2^n. Hence we obtained that $s(F_2^n)$ is the minimum m satisfying (3.2):

$$\left\lceil \frac{1 + \sqrt{1 + 8n}}{2} \right\rceil .$$

The application of Lemma 3.5 for $k = n-1$ gives $s(F_{n-1}^n) \geq n$. The next construction proves the equality

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

If we try to use Lemma 3.5 for $k = n$ we obtain only $\binom{s(F_n^n)}{2} \geq n$. However, the truth is $s(F_n^n) = n + 1$. The construction is

$$\begin{pmatrix} 0 & 0 & \cdots\cdots & 0 \\ 1 & 0 & & \\ 0 & 1 & \cdots\cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots\cdots & 1 \end{pmatrix}$$

while the proof of $s(F_n^n) \geq n + 1$ can be found in [6].

Lemma 3.5 gives $s(F_3^n) \geq n$. We have to find "only" an $n \times n$ matrix M in which F_3^n is the family of minimal keys. We were able to do that if $n = 12r + 1$ or $n = 12r + 4$. The theory of resolvable Steiner triples is used.

Theorem 3.6. [6]

$$s(F_1^n) = 2, \qquad s(F_2^n) = \left\lceil \frac{1 + \sqrt{1 + 8n}}{2} \right\rceil ,$$

$$s(F_{n-1}^n) = n, \qquad s(F_n^n) = n + 1,$$

$$s(F_3^n) \geq n, \qquad s(F_3^n) = n \quad if \begin{cases} n = 12r + 1 \\ n = 12r + 4 \end{cases} .$$

Conjecture 3.7.

$$s(F_3^n) = n \quad if \quad n \geq 7.$$

Let us remark that $s(F_3^4) = 4$. The difficulties with $k = 3$ show that we may expect only asymptotic results for other k's. Lemma 3.5 gives

$$c_1 \; n^{\frac{k-1}{2}} \leq s(F_k^n)$$

where c_1 depends on k but not on n. It can be proved that this is asymptotically sharp:

Theorem 3.8. ([6])

$$c_1 \, n^{\frac{k-1}{2}} \leq s(F_k^n) \leq c_2 \, n^{\frac{k-1}{2}} ,$$

where c_1 and c_2 do not depend on n.

Let us consider now the analogous problem for dependencies in place of keys. Due to the results of Section 2 we may consider the closures. Let L be a closure on an n-element set Ω. According to Theorem 2.10 there is a matrix M in which the closure is exactly

L. We say that M *realizes* L. The minimum number of rows of such matrices M is denoted by $s(L)$.

In fact, we know $s(L)$ for some L. Let

$$L_k^n(A) = \begin{cases} \Omega & \text{if } |A| \geq k, \\ A & \text{if } |A| < k. \end{cases}$$

It is easy to see that a matrix M realizes L_k^n iff the family of minimal keys in M is exactly F_k^n. Hence we have

$$s(L_k^n) = s(F_k^n).$$

In what follows we determine $s(L)$ for some closures constructed from L_k^n. If M_1 and M_2 are $m_1 \times n_1$ and $m_2 \times n_2$ matrices, resp., let $M_1 \times M_2$ denote the $(m_1 \cdot m_2)\times(n_1 + n_2)$ matrix consisting of rows whose first n_1 or last n_2 entries form an arbitrary row of M_1 or M_2, resp. Let Ω_1 and Ω_2 denote the set of columns of M_1 and M_2, resp. M_1 and M_2 determine the closures L_1 and L_2 on Ω_1 and Ω_2, resp. It is easy to see that $A \to b$ $(A \subset \Omega_1 \cup \Omega_2, b \in \Omega_1)$ holds in $M_1 \times M_2$ if and only if $A \cap \Omega_1 \to b$ holds in M_1. The same could be stated for $b \in \Omega_2$. This implies that

$$L(A) = L_1(A \cap \Omega_1) \cup L_2(A \cap \Omega_2)$$

holds for the closure determined by $M_1 \times M_2$. Following this equation, we may define the *direct product* of closures in general:

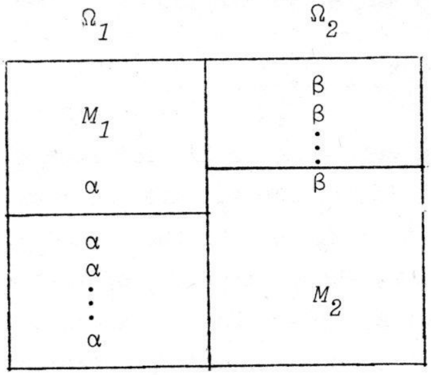
$$(L_1 \times L_2)(A) = L_1(A \cap \Omega_1) \cup L_2(A \cap \Omega_2).$$

The next theorem determines $s(L_1 \times L_2)$, given $s(L_1)$ and $s(L_2)$.

Theorem 3.9. [6]

$$s(L_1 \times L_2) = s(L_1) + s(L_2) - 1.$$

To prove $\leq$ we have to give a $(s(L_1)+s(L_2)-1)\times(n_1+n_2)$ matrix realizing $L_1 \times L_2$ using the $s(L_1)\times n_1$ and $s(L_2)\times n_2$ matrices M_1 and M_2 realizing L_1 and L_2, resp. This has much less rows than the trivial construction $M_1 \times M_2$. The better construction is given in the Figure. α and β denotes the last row of M_1 and the first row of M_2, resp. They are repeated then many times. It is easy to see that the closure determined by the matrix



Figure

is $L_1 \times L_2$. The proof of the opposite ($\geq$) inequality can be found in [6].

The next problem tries to determine the "most complex" system of dependencies in a database with n attributes. Due to the results of Section 2 we can speak about full families instead of dependencies. Let T be a full family. The pair $(A,B) \in T$ is called *basic* if

1) $A \neq B$

2) there is no $A' \subset A,\ A' \neq A,\ (A',B) \in T$

3) there is no $B' \supset B,\ B' \neq B,\ (A,B') \in T$.

Let $N(n)$ denote the maximum number of basic pairs in a full family in an n-element set.

First we show a trivial upper estimate on $N(n)$. Introduce

the notation $\mathsf{A} = \{A: (A,B)$ is a basic pair in $T\}$. Let (A,B) be a basic pair, and suppose that $A \subset C \subset B$, $|C| = |A| + 1$. It is easy to see that $C \notin \mathsf{A}$. Such a C can be obtained from at most n different sets A, consequently for at least $|\mathsf{A}|/n$ sets C holds $C \notin \mathsf{A}$. This implies $|\mathsf{A}| + \frac{|\mathsf{A}|}{n} \leq 2^n$.
Hence we have

$$N(n) \leq |\mathsf{A}| \leq 2^n(1 - \frac{1}{n+1}).$$

This estimate is considerably improved by Kostochka ([8]):

Theorem 3.10. *([2] and [8]).*

$$2^n(1 - \frac{1}{\log_2 e}\frac{\log_2 \log_2 n}{\log_2 n})(1 + o(1)) \leq N(n) \leq 2^n(1 - \frac{1}{150}\frac{\log^{3/2} n}{\sqrt{n}}).$$

It remains an *open question* what is the proper second term of $N(n)$.

4. PARTIALLY ORDERED SET OF DATABASES OF A GIVEN SET OF ATTRIBUTES

By a database we mean here the equivalent models of Section 2, namely, the full families, the closures, etc. The most convenient one for our purposes is the model of closures.

We add a natural condition, namely, no attribute is known in advance. Or by terms of matrices, the matrix has no constant column. It is easy to see that this is equivalent to the condition.

$$(4.1) \qquad L(\emptyset) = \emptyset.$$

A database is constantly changing during its life. It also changes the corresponding closure. A typical change is to delete the data of some individuals. If $A \to a$ is true then it remains true after the change.

This implies

$$(4.2) \qquad L_1(A) \subset L_2(A) \quad \text{(for all } A \subset X)$$

if L_1 and L_2 denote the closures before and after the change. We say that L_1 is *richer* than or *equal to* L_2 $(L_1 \geq L_2)$ iff they satisfy (4.2). It is easy to see that this property is transitive, consequently the closures of a fixed n-element set X form a partially ordered set (poset) for the ordering given in (4.2) The aim of the present section is to study this poset P.

Now we want to reduce the problem for the families of closed sets. The following lemma is needed:

Lemma 4.1. *$L(A)$ is equal to the smallest closed set containing A.*

We are able to introduce the equivalent poset of the intersection semi-lattices satisfying the condition $\emptyset \in Z$ which is equivalent to (4.1). $Z(L)$ denotes the family of closed sets in the closure L.

Proposition 4.2. *$L_1 \leq L_2$ iff $Z(L_1) \subset Z(L_2)$.*

Proof. Suppose that $L_1 \leq L_2$ and A is closed in L_1. By definition, $L_1(A) = A$ holds. (4.2) implies $A \supset L_2(A)$ and, by (2.1), we obtain $A = L_2(A)$.
That is, $A \in Z(L_2)$ and the first part is proved.

Conversely, suppose now $Z(L_1) \subset Z(L_2)$. By Lemma 4.1, $L_1(A)$ and $L_2(A)$ are the smallest closed sets according to L_1 and L_2, containing A, resp. . $Z(L_1) \subset Z(L_2)$ implies $L_2(A) \subset L_1(A)$ and this is the definition of $L_1 \leq L_2$. The proof is complete.

We say that L_2 *covers* L_1 and write $L_2 \bullet\!> L_1$ iff $L_2 > L_1$ and there is no L_3 satisfying $L_2 > L_3 > L_1$. The function r associating non-negative integers with the elements of a given poset satisfying the following two conditions is called a *rank-function:*

(4.3) r is zero for some element,

(4.4) if L_2 covers L_1 then $r(L_2) = r(L_1)+1$.

Proposition 4.3. $L_1 <\cdot L_2$ *iff* $Z(L_1) \subset Z(L_2)$ and $|Z(L_2)-Z(L_1)|=1$.

Proof. $L_1 <\cdot L_2$ implies $Z(L_1) \subset Z(L_2)$ by Proposition 4.2. We have to see only the second condition. Suppose, indirectly, that $|Z(L_2)-Z(L_1)| \geq 2$. Choose the member $C \in Z(L_2) - Z(L_1)$ so that C is not contained in any other member of $Z(L_2) - Z(L_1)$. We prove that $Z(L_2) - \{C\}$ is closed under intersection. If $A,B \in Z(L_2) - \{C\}$ and $A \cap B$ is not in $Z(L_2) - \{C\}$ then $A \cap B = C$ by (2.11). Neither A nor B can be in $Z(L_2) - Z(L_1)$ by the choice of C. However, if $A,B \in Z(L_1)$ then $C \in Z(L_1)$ by (2.11), again. This contradicts the choice of C.

The rest of the proof is a trivial consequence of Proposition 4.2 and Theorem 2.6. The proof is complete.

Proposition 4.4. $r(L) = |Z(L)|-2$ *is a rank-function on* P.

Proof. $L(X) = X$ and $L(\emptyset) = \emptyset$ follow by (2.1) and (4.1), resp. This implies $|Z(L)| \geq 2$ for any element of P. (4.3) is fulfilled. Proposition (4.3) implies (4.4). The proof is complete.

The rank $r(P)$ of P is the maximum value of the rank-function. It is easy to see that it is

$$r(P) = 2^n - 2$$

where $|X| = n$.

For later use we need another.

Proposition 4.5. *Let* Z *be a family closed under intersection and* $A \in Z$. $Z - \{A\}$ *is closed under intersection iff* $A \in extr(Z)$.

Proof. Suppose first that $A \in extr(Z)$. $B,C \in Z-\{A\}$ implies $B \cap C \in Z$, but $B \cap C$ can be equal to A only when one of them is equal to A. This contradiction shows that $Z - \{A\}$ is closed under intersection.

The other implication will be proved in an indirect way. Suppose that $A \in Z\text{-}extr(Z)$. Then there are B and C satisfying $A = B \cap C$, $B \neq A \neq C$, and $B,C \in Z$. Hence we obtain $B,C \in Z-\{A\}$ and $B \cap C \notin Z-\{A\}$. $Z-\{A\}$ is not closed under intersection. This contradiction completes the proof.

Let $deg_a(L)$ and $deg_b(L)$ denote the number of elements of P covering L and covered by L, respectively. We define the following functions:

$$f_1(n,k) = max\{deg_a(L): r(L) = k\}$$
$$f_2(n,k) = min\{deg_a(L): r(L) = k\}$$
$$f_3(n,k) = max\{deg_b(L): r(L) = k\}$$
$$f_4(n,k) = min\{deg_b(L): r(L) = k\}$$
$$(1 \leq n, \quad 0 \leq k \leq 2^n - 2).$$

Theorem 4.6. ([3])

$$f_1(n,k) = 2^n - k - 2.$$

Proof. If $r(L) = k$, then $|Z(L)| = k+2$ by Proposition 4.4. Proposition 4.3 implies that the closures L' covering L satisfy $Z(L') = k+3$, $Z(L') \supset Z(L)$. The number of possible choices of the member $Z(L') - Z(L)$ is at most 2^n-k-2. We construct now a $Z(L)$ allowing all these choices. Suppose that $k + 1 = \binom{n}{o} + \binom{n}{1} + \dots$ $\dots + \binom{n}{r} + a$ where $0 \leq a < \binom{n}{r+1}$. Then let $Z(L)$ consists of the empty set, all 1-element, 2-element, ..., r-element subsets, any a pieces of the $r+1$-element subsets and X. It is easy to see that this family is closed under intersection. Moreover $Z(L) \cup \{A\}$ $(A \notin Z(L))$ also has this property. We can choose A in 2^n-k-2 different ways and all these choices lead to closures covering L. The proof is complete.

We are not able to determine $f_2(n,k)$ in general, but we have shown that it can take small values for many k's.

Theorem 4.7. *([3])*

$$f_2(n,k) = 0 \quad iff \quad k = 2^n-2,$$

$$f_2(n,k) = 1 \quad iff \quad k = 2^n-2^{n-a-1}-2$$

for some $0 < a < n$;

$f_2(n,k) = 2$ *if either*

$$k = 2^n-2^{n-a-b-1}-2^{n-b-c-1}+2^{n-a-b-c-1}-2$$

for some $1 \leq a,c, \quad 0 \leq b, \quad a+b+c < n,$

or

$$k = 2^n-2^{n-a-b-1}-2^{n-b-c-1}-2$$

for some $1 \leq a,c, \quad 0 \leq b, \quad a+b+c < n,$

but $k \neq 2^n-2^{n-a-1}-2 \qquad (0 < a < n).$

We know practically nothing about $f_3(n,k)$. The only almost trivial statement is that

$$f_3(n,k) = k+2 \quad iff \quad 0 \leq k \leq 2n-1.$$

The construction of the corresponding $Z(L)$ consists of one--element sets and a chain. It is also easy to see that for $k > 2n-1$ $f_3(n,k) < k+2$, that is, there is always a member of $Z(L)$ which cannot be omitted to preserve the property that it is closed under intersection.

Theorem 4.8. *([3])*

$$\lceil log_2(k+1) \rceil \leq f_4(n,k)$$

$$\leq \lceil log_2(k+1) \rceil -2+(\text{number of non-zero digits in the binary form of } (k+2)).$$

$$f_4(n,k) = \lceil log_2(k+1) \rceil \quad if \quad n \geq (k+2).$$

We show where the expression $\lceil log_2(k+1) \rceil$ comes from. We have an intersection semi-lattice Z of $k+2$ members, containing $\emptyset$. By Proposition 4.5, only the members of $extr(Z)-\{\Omega\}$ can be omitted if we want to obtain another intersection semilattice. Hence we have

$$(4.5) \qquad f_4(n,k) = \min_{\substack{\emptyset \in Z \\ |Z| = k+2}} |extr(Z)-\{\emptyset\}| - 1.$$

By Lemma 2.7, any member of Z is an intersection of some members of $extr(Z)-\{\Omega\}$. The number of such intersections is $2^{|extr(Z)|-1}$. Hence we have

$$(4.6) \qquad k+2 = |Z| \leq 2^{|extr(Z)|-1} .$$

Using (4.5) we obtain

$$\log_2(k+2) \leq f_4(n,k),$$

providing that $\phi \notin extr(Z)$. If $\phi \notin extr(Z)$ then all the intersections containing ϕ are empty, therefore we have $k+2 = |Z| \leq 1+2^{|extr(Z)|-2}$ in place of (4.6). This leads to $\log_2(k+1) \leq f_4(n,k)$.

Added in proof. In the final version of [3] Theorem 4.7 is considerably improved, giving an upper estimate on $f_2(n,k)$ if $k>2^{n-1}+2$.

REFERENCES

[1] Armstrong, W.W., *Dependency structures of data base relationships*, Information Processing 74, North-Holland, Amsterdam, 1974, 580-583.

[2] Békéssy, J., Demetrovics, J., Hannâk, L., Frankl, P. and Katona, G.O.H., *On the number of maximal dependencies in a data base relation of fixed order*, Discrete Math., 30(1980), 83-88.

[3] Burosch, G., Demetrovics, J. and Katona, G.O.H., *The poset of closures* (Submitted to Order).

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

[5] Demetrovics, J., *On the equivalence of candidate keys with Sperner systems,* Acta Cybernetica, 4 (1979), 247-252.

[6] Demetrovics, J., Füredi, Z. and Katona, G.O.H., *Minimum matrix representation of closure operations* (to appear in Discrete Appl. Math.).

[7] Demetrovics, J. and Gyepesi, Gy., *A note on minimal matrix representation of closure operations,* Combinatorica, 3(1983), 177-180.

[8] Kostochka, A.V., *On the maximum size of a filter in the n-cube* (in Russian).

[9] Sperner, E., *Ein Satz über Untermengen einer endlichen Menge,* Math. Z., 27(1928), 544-548.

J. DEMETROVICS

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

G.O.H. KATONA

Mathematical Institute
Hungarian Academy of Sciences
1376. Budapest, P.f. 428.
Hungary