EXTREMAL COMBINATORIAL PROBLEMS OF DATABASE MODELS *

János Demetrovics
Computer and Automation Institute, Hungar.Acad. Sci.
1111, Budapest, Kende u. 13-17.,Hungary

Gyula O. H. Katona
Mathematical Institute, Hungar. Acad. Sci.
1364, Budapest, P.f. 127, Hungary

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 unimportant from the point of view of the

* Research supported by Hungarian National Foundation for Scientific Research, grant number 1.066

structure of the database. We have to consider, whether they are equal or unequal, only. A further relaxation of the model is when certain interrelations of the columns are considered, only: which columns are determined by the entries of given sets of columns. These models suggest several combinatorial problems of extremal nature. Namely, given some parameter(s) of the database (for instance the number of attributes), determine the maximum or minimum of another parameter (for instance the maximum number of minimal keys; see the definition later). The aim of the recent paper is to survey these extremal combinatorial results found by the authors and their coauthors. It is a refreshed version of [8].

In Section 2 we give the necessary definitions and study the models and their relationships. The proofs can be found in [3] and/or in [8].

In Section 3 we survey the combinatorial problems mentioned above.

In the last section we introduce a partially ordered set in which the databases (more precisely their models) are ordered in a natural way. Some combinatorial data of this partially ordered set are determined.

Both in Sections 3 and 4, one can find more open problems than final results.

2.THE MODEL

The concepts and results of this section (except Theorems 2.10,2.11) are either published by other authors (see e.g. [1],[4]) or belong to the folklore. We repeat them here, the proofs can be found in [3] and/or in [8].

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 *functionally depends* or shortly *depends on* A if M has no two rows equal in A but different in B.

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

It is clear that different matrices may give the same system of functional dependencies, therefore this model is less fine than the matrix model. However, in many cases it contains sufficient amount of information. The system of functional dependencies can be considered as a model of all matrices in which these dependen-

cies hold. So the rows of the actual matrix might be changed during the life of the database, but these dependencies are unchanged.

Define the function $\mathfrak{L}$ on 2^{Ω} by

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

This function possesses some simple properties:

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

(2.1) $A \subseteq \mathfrak{L}(A)$;

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

(2.3) $\mathfrak{L}(\mathfrak{L}(A)) = \mathfrak{L}(A)$.

The combinatorial literature calls a function satisfying (2.1)-(2.3) a *closure*. Lemma 2.1 makes us able to call $\mathfrak{L}$ a closure.

Obviously, the fuctional dependency can be defined by $\mathfrak{L}$:

Lemma 2.2. $A \to B$ *iff* $B \subseteq \mathfrak{L}(A)$.

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

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

(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* $A \cup C \to B \cup D$.

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

We defined the above closure with the functional dependencies determined by a matrix. It also can be done with an arbitrary full family $\mathfrak{I}$:

$$\mathfrak{L}(A) = \{b: (A,b) \in \mathfrak{I}\}.$$

Lemma 2.1 can be restated for the so defined function $\mathfrak{L}$, consequently it is a closure. So a full family defines a closure and the converse is also true:

Theorem 2.4.

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

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

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 said to be *closed* (with respect to the closure $\mathcal{L}$) if $\mathcal{L}(A) = A$. The next theorem determines the family $\mathcal{F}$ $(=\mathcal{F}(\mathcal{L}))$ of the closed sets with respect to the closure $\mathcal{L}$.

Theorem 2.5. *Let $\mathcal{F}$ be a family of different subsets of Ω. $\mathcal{F}$ is the family of closed sets with respect to some closure $\mathcal{L}$ iff*

(2.10) $\Omega \in \mathcal{F}$,

(2.11) $A,B \in \mathcal{F}$ *implies* $A \cap B \in \mathcal{F}$.

Property (2.11) can be formulated as "$\mathcal{F}$ *is closed under intersection*". The families $\mathcal{F}$ satisfying (2.10) and (2.11) are called *intersection semi-lattices*. The next theorem shows that a closure can be uniquely recovered from its closed sets.

Theorem 2.6.

(2.12) $$\mathfrak{L} \to \mathcal{F} = \{\mathcal{F}: \mathfrak{L}(\mathcal{F}) = \mathcal{F}\}$$

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) $$\mathcal{F} \to \mathfrak{L}(A) = \bigcap_{A \subseteq Z \in \mathcal{F}} Z \ .$$

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

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

The next theorem characterizes the families $M(\mathfrak{F})$.

Theorem 2.8. *A family* $\mathcal{E}$ *is equal to* $M(\mathfrak{F})$ *for some intersection semi-lattice* $\mathfrak{F}$ *iff*

(2.14) $\Omega \in \mathcal{E}$

and

(2.15) $A = \bigcap_{i=1}^{r} A_i$ $(r \geq 1)$, $A, A_1, \ldots, A_r \in \mathcal{E}$ *imply* $A = A_i$ *for some* i $(1 \leq i \leq r)$.

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

Theorem 2.9.

(2.16) $\mathfrak{F} \to M(\mathfrak{F})$

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

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

The following equivalent description of a

closure is rather unusual. Let $H \subset \Omega$, $H \notin \mathcal{F}$ and suppose that both $\mathcal{F}$ and $\mathcal{F} \cup \{H\}$ are closed under intersection. Consider the sets A satisfying $A \in \mathcal{F}$, $H \subset A$. The intersection of all these sets is in $\mathcal{F}$, therefore it is different from H. Denote it by $L(H)$. $H \subsetneq L(H)$ is obvious. Let $\mathcal{H}(\mathcal{F})$ denote the set of all pairs $(H,L(H))$ where $H \subset \Omega$, $H \notin \mathcal{F}$ but $\mathcal{F} \cup \{H\}$ is closed under intersection. The following theorem characterizes the possible sets $\mathcal{H}(\mathcal{F})$:

Theorem 2.10.[3] *The set* $\{(A_i,B_i)\}_{i=1}^{m}$ *is equal to* $\mathcal{H}(\mathcal{F})$ *for some intersection semi-lattice* $\mathcal{F}$ *if and only if the following conditions are satisfied:*

(2.18) $\quad \emptyset \neq A_i \subsetneq B_i \subseteq \Omega,$

(2.19) $\quad A_i \subseteq A_j$ *implies either* $B_i \subseteq A_j$ *or* $B_i \supseteq A_j$,

(2.20) $\quad A_i \subseteq B_j$ *implies* $B_i \subseteq B_j$,

(2.21) $\quad$ *for any* i *and* $C \subset \Omega$ *satisfying* $A_i \subsetneq C \subsetneq B_i$ *there is a* j *such that either* $C = A_j$ *or* $A_j \subset C$, $B_j \not\subset C$, $B_j \not\supset C$ *all hold.*

The set of pairs (A_i,B_i) satisfying (2.18) *to (2.21) is called an extension.* Its definition is not really beautiful but it is needed in some applications. On the other hand, there is, again, a one-to-one correspondence between the extensions and the intersection

semi-lattices. So the extension is an equivalent form of the closure. But we do not say that it is a simple one. Only useful.

Theorem 2.11.

(2.22) $\mathfrak{F} \to \mathcal{H}(\mathfrak{F})$

is a one-to-one correspondence between the intersection semi-lattices and the extensions. The inverse of (2.22) is given by

(2.23) $\mathcal{H} \to \mathfrak{F} =$

$\{A: A \subseteq \Omega,\ H \subseteq A \to L(H) \subseteq A \text{ for all } (H, L(H)) \in \mathcal{H}\}.$

We have seen several concepts equivalent to the full families, however we did not show yet that the full families are exactly the functional dependencies defined by matrices:

Theorem 2.12 *Let* $\mathfrak{T}$ *be a full family on* Ω. *There exists a matrix M with* $|\Omega|$ *columns in which the* family of functional dependencies coincides with $\mathfrak{T}$.

Beside the functional dependencies the literature knows many different dependencies (see e.g. [15]).

We call the attention to two, not really studied types.

As an example let $A=\{a_1,a_2\}$, $b \notin A \subset \Omega$. Suppose that we do not know the actual entry of the attribute a_1 for a given individual, but we know that it is either r_1 or r_2. We know that the entry of the attribute a_2 of the individual is either r_3 or r_4, that is, the entries belong to a fixed two-element set. Suppose that this always implies that the entries of the same individual must belong to a two-element set $\{r_5,r_6\}$ (depending on r_1,r_2,r_3,r_4). We say in this case that *b two-depends on A*. The formal definition of the two-dependency is the following. $A \to_2 B$ if and only if there are no three rows of the matrix having at most two different entries in every column in A and three different entries in one of the columns of B. In general, we say that *B i-depends on A*, in notation $A \to_i B$, if and only if the matrix contains no $i+1$ rows having at most i different entries in any column belonging to A and having $i+1$ different entries in a column belonging to B.

Obviously, $\to_1$ is the same as $\to$. $\to_i$ has a nice property:

Lemma 2.13. *$\mathfrak{L}_i(A)=\{b: A \to_i b\}$ is a closure.*

Unfortunately, $A \to_i b$ does not imply $A \to_{i+1} b$, as the following example shows it for $i=1$, $A=$

$\{a_1,a_2\}$ and $b=a_3$:

$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 2. \end{matrix}$$

The other dependency is much weaker than the usual ones. Suppose that the entries r_1 and r_2 of the attributes a_1 and a_2, respectively, determine uniquely the entry of the attribute b. Then we say that $(a_1,a_2;r_1,r_2)$ determines (b,r_3), in notation $(a_1,a_2;r_1,r_2) \rightarrow (b,r_3)$. In general, denote the set of possible entries in the column a_i by D_i. $\alpha=(a_1,\ldots,a_k;r_1,\ldots,r_k)$ is called a *partial function* if a_i are distinct elements of Ω and $r_i \in D_i$. We write $\alpha \rightarrow \beta=(b_1,\ldots,b_l;s_1,\ldots,s_l)$ and say that β *depends on* α if any row having r_i in the column a_i $(1\leq i\leq k)$ has the entry s_i in the column b_i $(1\leq i\leq l)$. Introduce $\mathfrak{L}(\alpha)$ as the largest possible β satisfying $\alpha \rightarrow \beta$ (that is, the one maximizing the size of $\{b_1,\ldots,b_l\}$). If there is no such row then $\alpha \rightarrow$ and $\mathfrak{L}(\alpha)$ are not defined.

Let $\alpha=(a_1,\ldots,a_k;r_1,\ldots,r_k)$ and $\beta=(b_1,\ldots,b_l;s_1,\ldots,s_l)$. $\alpha \subset \beta$ denotes that $\{a_1,\ldots,a_k\} \subset \{b_1,\ldots,b_l\}$ and $a_i=b_j$ implies $r_i=s_j$. We have an analogue of Lemma 2.1:

Lemma 2.14.

(2.24) $\alpha \subseteq \mathfrak{L}(\alpha)$;

(2.25) $\alpha \subseteq \beta$ *implies* $\mathfrak{L}(\alpha) \subseteq \mathfrak{L}(\beta)$;

(2.26) $\mathfrak{L}(\mathfrak{L}(\alpha)) = \mathfrak{L}(\alpha)$.

A basic difference between Lemma 2.1 and the present one is that $\mathfrak{L}(A)$ is defined for all subsets $A \subseteq \Omega$, while $\mathfrak{L}(\alpha)$ is defined only for some partial functions. However these partial functions have a structure. We say that a set $\mathcal{P}$ of partial functions is a *downset* iff

(2.27) $\alpha \in \mathcal{P}$ and $\beta \subset \alpha$ imply $\beta \in \mathcal{P}$

and

(2.28) for any $\alpha \in \mathcal{P}$ there is a $\gamma \in \mathcal{P}$ such that $\gamma = (c_1, \ldots, c_n; \ldots)$ where $\{c_1, \ldots, c_n\} = \Omega$.

If a function $\mathfrak{L}$, defined on a downset of partial functions, satisfies (2.24)-(2.26) then it is called a *function-closure*. The theory of closures presented in this section can be generalized for function-closures.

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 $\mathfrak{I}$ 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 functional dependency or equivalently, $\mathfrak{L}(A)=\Omega$. Moreover, K is a *minimal key* iff it is a key 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 $\mathcal{K}$ of subsets of Ω there is a matrix in which the family of minimal keys is $\mathcal{K}$ iff $\mathcal{K}$ satisfies*

$$(3.1) \qquad K_1, K_2 \in \mathcal{K}, \quad K_1 \neq K_2 \rightarrow K_1 \not\subseteq K_2.$$

The families satisfying (3.1) are called *Sperner families*. Our original problem is now reduced to the determination of $\max|\mathcal{K}|$ under (3.1). This problem, however, has been solved many years ago by Sperner [12].

The answer is $\binom{n}{\lfloor n/2 \rfloor}$. So we can formulate

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

Theorem 3.1 states that there is a matrix M, for any $\mathcal{K}$, in which the family of minimal keys is $\mathcal{K}$. We did not consider the number of rows. Now, let $s(\mathcal{K})$ denote the minimum number of rows in such a matrix M, where $\mathcal{K}$ is any Sperner family. $s(\mathcal{K})$ should not be understood as the number of individuals in the whole database, but as the minimum number of individuals generating $\mathcal{K}$ as the set of minimal keys. There is a general upper bound on $s(\mathcal{K})$:

Theorem 3.3. ([5])

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

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

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

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

We cannot construct a Sperner family which has such a large $s(\mathcal{K})$. There is, however a class of Sperner families for which $s(\mathcal{K})$ is exactly determined. Let F_k^n denote 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])

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

This leads to the exact solution in some easy cases:

Theorem 3.6. ([6])

$s(F_1^n)=2$, $\quad s(F_2^n)=\lceil (1/2)(1+\sqrt{1+8n})\rceil$,

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

$s(F_3^n)\geq n$, $\quad s(F_3^n)=n$ if $n=12r+1$ or $12r+4$.

Conjecture 3.7.

$$s(F_3^n) = n \text{ if } n\geq 7.$$

Let us remark that $s(F_3^4)=4$. The above exact results on $s(F_3^n)$ are based on some construction of certain triple systems. We were able to prove the following conjecture (using a theorem of Hanani [9]) for $n=12r+1$ and $n=12r+4$:

Conjecture 3.8. *There is a system of 3-element subsets of an $n=(3r+1)$-element set $\{1,2,\ldots,n\}$ satisfying the following conditions:*

(1) Any pair of elements is contained in exactly two 3-sets.

(2) The family of 3-sets can be divided into n subfamilies where the ith subfamily is a partition of $\{1,2,\ldots,n\}-\{i\}$.

(3) Exactly one pair of members of two dif-

ferent subfamilies meet in 2 elements.

Andrea Rausch (Greifswald) [11] proved that this is not true for $n=10$ $(r=3)$. But we still believe that the conjecture is valid for $n \geq 13$.

The difficulties with $k=3$ indicate that only asymptotic results might be expected for other k's. Lemma 3.5 gives

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

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

Theorem 3.9. ([6])

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

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 $\mathfrak{L}$ be a closure on an n-element set Ω. According to Theorem 2.12 there is a matrix M in which the closure is exactly $\mathfrak{L}$. We say that M *realizes* $\mathfrak{L}$. The minimum number of rows of such matrices M is denoted by $s(\mathfrak{L})$.

In fact, we know $s(\mathfrak{L})$ for some $\mathfrak{L}$. Let

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

It is easy to see that a matrix M realizes $\mathfrak{L}_k^n$ iff the family of minimal keys in M is exactly F_k^n. Hence we have

(3.2) $$s(\mathfrak{L}_k^n) = s(F_k^n).$$

Let $\Omega=\Omega_1 \cup \Omega_2$ be a partition of Ω and let $\mathfrak{L}_1$ and $\mathfrak{L}_2$ be closures defined on Ω_1 and Ω_2, respectively. The *direct product* $\mathfrak{L}_1 X \mathfrak{L}_2$ is defined by

$$(\mathfrak{L}_1 X \mathfrak{L}_2)(A) = \mathfrak{L}_1(A \cap \Omega_1) \cup \mathfrak{L}_2(A \cap \Omega_2).$$

Theorem 3.10. [6]

$$s(\mathfrak{L}_1 X \mathfrak{L}_2) = s(\mathfrak{L}_1) + s(\mathfrak{L}_2) - 1.$$

(3.2) and Theorem 3.10 make us able to determine $s(\mathfrak{L})$ for several other closures.

Another interesting question is the following. Fix the size of the set of minimal keys, what is the range of the possible values of $s(\mathfrak{K})$?

Problem 3.11. *Determine*

$$s(k) = \max_{|\mathcal{K}|=k} s(\mathcal{K})$$

and

$$r(k) = \min_{|\mathcal{K}|=k} s(\mathcal{K}).$$

We know very little about this problem. However there is a connection to another, probably easier problem. Let $A \subset \Omega$ be an *antikey* iff it is not a key. The family of *maximum antikeys* is denoted by $\mathcal{K}^{-1}$. It is known [6] that

$$|\mathcal{K}^{-1}| \leq \binom{s(\mathcal{K})}{2} \quad \text{and} \quad |s(\mathcal{K})| \leq 1 + |\mathcal{K}^{-1}|,$$

that is, there is a strong connection between $|\mathcal{K}^{-1}|$ and $s(\mathcal{K})$. This leads to another open problem:

Problem 3.12. *Determine*

$$\max_{|\mathcal{K}|=k} |\mathcal{K}^{-1}|$$

and

$$\min_{|\mathcal{K}|=k} |\mathcal{K}^{-1}|.$$

As regards the minimum, we think that it is attained for the family $\mathcal{K}$ consisting of i and $i+1$-element subsets, where i is determined by

$$\binom{n}{i} \leq k < \binom{n}{i+1}.$$

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

1) $A \neq B$,

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

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

Let $N(n)$ denote the maximum number of basic pairs in a full family in an n-element set. We know only the following estimates on $N(n)$:

Theorem 3.13. ([2] and [10])

$$2^n\left(1 - \frac{4\ \log_2\log_2 n}{\log_2 e\ \log_2 n}\ (1 + o(1))\right) \leq$$

$$N(n) \leq 2^n \left(1 - \frac{\log^{3/2} n}{150\sqrt{n}}\right).$$

The proper second term is an *open question.*

Another possible complexity of a database is defined in the forthcoming book of Thalheim [14]. Other interesting combinatorial problems of databases are considered in [13] and [14].

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 to start with the model of closures.

A natural condition is added: no attribute is known in advance. By terms of matrices, the matrix has no constant column. It is easy to see that this is equivalent to the condition

(4.1) $$\mathcal{L}(\emptyset) = \emptyset.$$

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

$$\mathfrak{L}_1(A) \subset \mathfrak{L}_2(A) \quad \text{(for all } A \subset \Omega\text{)} \tag{4.2}$$

if $\mathfrak{L}_1$ and $\mathfrak{L}_2$ denote the closures before and after the change. We write $\mathfrak{L}_1 \geq \mathfrak{L}_2$ in this case. It is easy to see that this property is transitive, consequently the closures of a fixed n-element set Ω 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 reduce the model to the families of closed sets. They form an intersection semi-lattice $\mathcal{F}$ satisfying the condition $\emptyset \in \mathcal{F}$ which is equivalent to (4.1). $\mathcal{F}(\mathfrak{L})$ denotes the family of closed sets in the closure $\mathfrak{L}$.

Lemma 4.1. *$\mathfrak{L}_1 \leq \mathfrak{L}_2$ iff $\mathcal{F}(\mathfrak{L}_1) \subset \mathcal{F}(\mathfrak{L}_2)$.*

We say that $\mathfrak{L}_2$ *covers* $\mathfrak{L}_1$ and write $\mathfrak{L}_1 \prec \mathfrak{L}_2$ iff $\mathfrak{L}_1 < \mathfrak{L}_2$ and there is no $\mathfrak{L}_3$ satisfying $\mathfrak{L}_1 < \mathfrak{L}_3 < \mathfrak{L}_2$.
The function r associating non-negative integers with the elements of a given poset satisfying the following conditions is called a *rank-function*:

(4.3) r is zero for some element,

(4.4) if $\mathfrak{L}_2$ covers $\mathfrak{L}_1$ then $r(\mathfrak{L}_2) = r(\mathfrak{L}_1) + 1$.

The following lemma shows the structure of P more clearly.

Lemma 4.2. *$\mathfrak{L}_1 \lessdot \mathfrak{L}_2$ iff $\mathfrak{F}(\mathfrak{L}_1) \subset \mathfrak{F}(\mathfrak{L}_2)$ and $|\mathfrak{F}(\mathfrak{L}_2) - \mathfrak{F}(\mathfrak{L}_1)| = 1$.*

Now it is easy to deduce

Lemma 4.3. *$r(\mathfrak{L}) = |\mathfrak{F}(\mathfrak{L})| - 2$ is a rank-function on P.*

Now we are able to pose the main problem of of this section. Consider the closures of a fixed rank in P and try to find the minimum (maximum) number of neighboring closures from above (from below). That is, lower and upper estimates of the degrees (from above and from below) in a fixed level of the Hasse-diagram of P are sought. The database motivation for this question is the following. A temporary state of the database correspond to an element P. A change in the database correspond to a move along an edge of this Hasse-diagram. The life of the database therefore corresponds to a random walk along the Hasse-diagram of P. As a first model, we might suppose

that all the edges at a given element are chosen with equal probabilities. To obtain any (probabilistic) statement in this model we obviously need information about the degrees.

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

$$f_1(n,k) = \max\{\deg_a(\mathfrak{L}): r(\mathfrak{L}) = k\}$$

$$f_2(n,k) = \min\{\deg_a(\mathfrak{L}): r(\mathfrak{L}) = k\}$$

$$f_3(n,k) = \max\{\deg_b(\mathfrak{L}): r(\mathfrak{L}) = k\}$$

$$f_4(n,k) = \min\{\deg_b(\mathfrak{L}): r(\mathfrak{L}) = k\}$$

$$(1 \leq n,\ 0 \leq k \leq 2^n - 2).$$

$f_1(n,k)$ is fully determined, there are estimates for $f_2(n,k)$ and $f_4(n,k)$. However we know practically nothing about $f_3(n,k)$.

Theorem 4.4. [3]

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

Theorem 4.5. [3]

$$f_2(n,k) = 0 \text{ iff } k = 2^n - 2,$$

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

for some $0 < a < n$.

If $k > 2^{n-1} + 2$ *then* $f_2(n,k) \leq$ *number of bits 1 in the binary expansion of* $2^n - k - 2$. *What is at most* $n-1$.

The proof is based on Theorems 2.11 and 2.12.

Theorem 4.6. [3]

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

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

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

The proof is based on the fact that only the members of $M(\mathcal{F}) - \{\Omega\}$ can be omitted from the intersection semi-lattice $\mathcal{F}$ if we want to obtain another intersection semi-lattice.

REFERENCES

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

[2] Békéssy, A., Demetrovics, J., Hannák, L., Frankl, P. and Katona, G.O.H., *On the 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 as a model of changing databases* (to appear in 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,* Discrete Appl. Math., 11(1985), 115-128.

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

[8] Demetrovics, J. and Katona, G.O.H., *Combinatorial problems of database models,* Coll. Math. Soc. János Bolyai, 42. Algebra, Combinatorics and Logic in Computer Science, Győr (Hungary), (1983), 331-353.

[9] Hanani, H., *The existence and construction of balanced incomplete block designs,* Ann. Math. Statist. 32 (1961), 361-386.

[10] Kostochka, A.V., *On the maximum size of a filter in the n-cube* (in Russian), Metodi Diskretnovo Analiza. 41(1984), 49-61.

[11] Rausch, A., personal communication.

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

[13] Thalheim, B., *On the number of keys in relational databases,* preprint of the Technische Universität Dresden.

[15] Ullman, J.D., *Principles of database systems*, Computer Science Press, Rockville, Maryland, 1982.

[14] Thalheim, B., Dependencies in Relational Databases, a book in preparation.