A survey of some combinatorial results concerning functional dependencies in database relations*

J. Demetrovics

*Computer and Automation Institute, Hungarian Academy of Sciences,
Kende u. 12–17, H-1111 Budapest, Hungary*

and

G.O.H. Katona

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

Abstract

A database R has some obvious and less obvious parameters such as the number of attributes, the size $|r|$, the maximum size of a domain, the number of some special functional dependencies (e.g. the minimal keys), and so on. The main aim of this paper is to survey some of the results giving connections and inequalities among these parameters. The methods are of a combinatorial nature. A generalization of the numerical dependency is also considered.

1. Introduction

The simplest model of a database is a matrix. The entries in one column are the data of the same kind (name, date of birth, etc.), the entries of one row are the data of one individual. Thus, in fact, we are dealing with finite sets of homogeneous finite functions which can be illustrated by matrices.

However, in the literature the names of these concepts are traditionally different from the above ones. One kind of data (e.g. name) is called an *attribute*. It can be identified with a column of the above matrix. The set of attributes will be denoted by $U = \{a_1, a_2, \ldots, a_n\}$. The set of possible entries in the ith column is the *domain* of a_i. It is denoted by $D(a_i)$. Thus, the data of one individual (row of the matrix) can be viewed as an element r of the direct product $D(a_1) \times D(a_2) \times \ldots \times D(a_n)$. Therefore, the whole database (or matrix) can be decsribed by the *relation* $R \subseteq D(a_1) \times D(a_2) \times \ldots \times D(a_n)$. If $r = (e_1, e_2, \ldots, e_n) \in R$, then $r(i)$ denotes the ith component of r, that is, e_i.

*This work was supported by the Hungarian National Foundation for Scientific Research, Grant No. 2575.

There might be some logical connections among the data. For instance, the date of birth determines the age (in a given year). Let A and B be two sets of attributes ($A, B \subseteq U$). The data in A might uniquely determine the data in B. Formally, we say that $B \subseteq U$ *functionally depends* on $A \subseteq U$ if

$$r_1(i) = r_2(i) \text{ for all such } i \text{ that } a_i \in A$$

implies

$$r_1(i) = r_2(i) \text{ for all such } i \text{ that } a_i \in B.$$

It is denoted by $A \to B$ [2,8]. Less formally, $A \to B$ if any two elements of R having the same values in the attributes belonging to A must have the same values also in B. Functional dependencies have a very important role in practical applications. In most of the present paper, we will consider these and some natural generalizations.

A special functional dependency has an even greater importance. If $K \to U$, that is, the values in K determine all other values, then K is called a *key*. If K is a key and contains no other key as a proper subset, then it is a *minimal key*.

A database R has some obvious and less obvious parameters, such as the number of attributes n, the size $m = |R|$ of the relation, the maximum size of a domain, the number of some special functional dependencies (e.g. the number of minimal keys), etc. The main aim of this paper is to survey some of the results giving connections among these parameters. An easy example is the following statement.

THEOREM 1.1 [12]

The number of minimal keys is at most

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}$$

and this estimate is sharp.

The inequality easily follows from the well-known theorem of Sperner [35], which states that the size of $|\mathcal{S}|$ of an *inclusion-free family* $\mathcal{S}$ ($A, B \in \mathcal{S}$ implies $A \not\subseteq B$) is at most $\binom{n}{\lfloor n/2 \rfloor}$. To prove that the inequality cannot be improved, one has to construct a relation R with n attributes and this many minimal keys. This can be more easily done if the possible systems of functional dependencies are characterized. For similar reasons, other equivalent descriptions are needed, too. These are collected in section 2. Section 3 contains inequalities similar to theorem 1.1. Section 4 investigates a partial ordered set whose elements are the relations on the same U, modeled with their systems of functional dependencies. Finally, section 5 tries to characterize the system of some more general, so-called (p, q)-dependencies.

2. Different characterizations of the systems of functional dependencies

It is easy to see that the following four properties hold for the functional dependencies in any relation R. Let A, B, C and D be subsets of the set U of the attributes of R.

$$A \to A, \tag{2.1}$$

$$A \to B \text{ and } B \to C \text{ imply } A \to C, \tag{2.2}$$

$$A \subseteq C, D \subseteq B \text{ and } A \to B \text{ imply } C \to D, \tag{2.3}$$

$$A \to B \text{ and } C \to D \text{ imply } A \cup C \to B \cup D. \tag{2.4}$$

A system $\mathfrak{D}$ of pairs (A, B) of subsets of U satisfying (2.1)–(2.4) is called a *determination*. (To be precise, we should repeat here these conditions with some other kind of arrows, like $A \rightsquigarrow B$ instead of $A \to B$, but we will use the same notation for functional dependencies in a relation and in a determination.)

The system of all functional dependencies in a relation R was called a *full family* by Armstrong [2]. He also found the characterization of the possible full families.

THEOREM 2.1 [2]

A system of pairs $A \to B$ of sets is a full family for some relation R iff it is a determination.

Using this theorem, it is easy to complete the proof of theorem 1.1. Define the determination $\mathfrak{D}$ to contain the pairs $A \to B$ for all $B \subseteq A \subseteq U$ and $A \to B$ for all $A, B \subseteq U$ such that $\lfloor n/2 \rfloor \le |A|$. It is easy (but somewhat tedious) to check that this is a determination, therefore there is a relation R in which the full family consists of these functional dependencies. The minimal keys in this relation are all the sets of size $\lfloor n/2 \rfloor$. This proves the sharpness of the statement of theorem 1.1.

Given a determination $\mathfrak{D}$ on U, one can define

$$\mathscr{L}(A) = \{a : A \to a\} \quad \text{for all } A \subseteq U. \tag{2.5}$$

The following properties can easily be proved for all $A, B \subseteq U$.

$$A \subseteq \mathscr{L}(A), \tag{2.6}$$

$$A \subseteq B \text{ implies } \mathscr{L}(A) \subseteq \mathscr{L}(B), \tag{2.7}$$

$$\mathscr{L}(\mathscr{L}(A)) = \mathscr{L}(A). \tag{2.8}$$

A set-function satisfying these properties is called a *closure operation* or, shortly, a *closure*.

PROPOSITION 2.2

The correspondence $\mathfrak{D} \to \mathcal{L}(\mathfrak{D})$ defined by (2.5) gives a bijection between the set of determinations and the set of closures.

Consider the following closure:

$$\mathcal{L}^n_{\lfloor n/2 \rfloor}(A) = \begin{cases} A & \text{if } |A| < \lfloor n/2 \rfloor, \\ U & \text{if } |A| \geq \lfloor n/2 \rfloor, \end{cases} \quad A \subseteq U.$$

This closure defines the determination $\mathfrak{D}^n_{\lfloor n/2 \rfloor}$ by proposition 2.2. On the other hand, there is such a relation R that its full family is $\mathfrak{D}^n_{\lfloor n/2 \rfloor}$. It is obvious that the minimal keys in this relation are the $\lfloor n/2 \rfloor$-element subsets of U. This gives an easier proof of the sharpness of theorem 1.1.

Given a closure $\mathcal{L}$ on U, define the *closed sets* by $B = \mathcal{L}(B)$. The family of closed sets is denoted by $\mathfrak{Z} = \mathfrak{Z}(\mathcal{L})$. It is easy to see that $\mathfrak{Z}$ is closed under intersection, that is $A, B \in \mathfrak{Z}$ implies $A \cap B \in \mathfrak{Z}$. Furthermore, $U \in \mathfrak{Z}$. A family $\mathfrak{Z}$ satisfying these properties is called an *intersection semi-lattice.*

PROPOSITION 2.3

The correspondence $\mathcal{L} \to \mathfrak{Z}(\mathcal{L})$ is a bijection between the set of closures and the set of intersection semi-lattices.

Denote by $\mathfrak{D}^n_k$ the determination containing the pairs $A \to B$ for all $B \subseteq A \subseteq U$ and $A \to B$ for all $A, B \subseteq U$ such that $k \leq |A|$. The corresponding closure is

$$\mathcal{L}^n_k = \mathcal{L}(\mathfrak{D}^n_k) = \begin{cases} A & \text{if } |A| < k, \\ U & \text{if } |A| \geq k, \end{cases} \quad A \subseteq U.$$

Note that $\mathfrak{Z}^n_k = \mathfrak{Z}(\mathcal{L}^n_k)$ consists of U and all sets of size at most $k-1$.

Given an intersection semi-lattice $\mathfrak{Z}$, define

$$\mathcal{M} = \mathcal{M}(\mathfrak{Z}) = \{M : M \in \mathfrak{Z} \text{ and there are no } r \geq 2 \text{ sets in } \mathfrak{Z}, \text{ all different from } M \text{ such that their intersection is } M\}. \tag{2.9}$$

It is easy to see that (i) no member M of $\mathcal{M}$ is an intersection of other members (all different from M), and (ii) $U \in \mathcal{M}$. Such families of subsets are called *intersection-free families.*

PROPOSITION 2.4

The correspondence $\mathfrak{Z} \to \mathcal{M}(\mathfrak{Z})$ is a bijection between the set of intersection semi-lattices and the set of intersection-free families.

It is easy to see that $\mathcal{M}_k^n = \mathcal{M}(\mathfrak{L}_k^n)$ consists of U and the sets of size $k-1$.

Propositions 2.2, 2.3 and 2.4 give different equivalent notions describing the same thing in different ways. The given goal determines which one of them should be used. They more or less belong to folklore, but their proofs (and the inverse mappings) can be found in [9] and [17]. In the rest of this section, we show some other equivalent notions which are/might be useful for some applications.

Let $\mathfrak{L}$ be an intersection semi-lattice on U and suppose that $H \subset U$, $H \notin \mathfrak{L}$ hold and $\mathfrak{L} \cup \{H\}$ is also closed under intersection. Consider the sets A satisfying $A \in \mathfrak{L}, H \subset A$. The intersection of all of these sets is in $\mathfrak{L}$, therefore it is different from H. Denote it by $\mathcal{L}(H)$. (If $\mathfrak{L} = \mathfrak{L}(\mathcal{L})$, then $\mathcal{L}(H)$ is the closure of H according to $\mathcal{L}$.) $H \subset \mathcal{L}(H)$ is obvious. Let $\mathcal{H}(\mathfrak{L})$ denote the set of all pairs $(H, \mathcal{L}(H))$, where $H \subset U, H \notin \mathfrak{L}$ but $\mathfrak{L} \cup \{H\}$ is closed under intersection. The following theorem characterizes the possible sets $\mathcal{H}(\mathfrak{L})$:

THEOREM 2.5 [9]

The set $\{(A_i, B_i)\}_{i=1}^m$ is equal to $\mathcal{H}(\mathfrak{L})$ for some intersection semi-lattice $\mathfrak{L}$ iff the following conditions are satisfied:

$$A_i \subset B_i \subseteq U, \quad A_i \neq B_i, \tag{2.10}$$

$$A_i \subseteq A_j \text{ implies either } B_i \subseteq A_j \text{ or } B_i \supseteq A_j, \tag{2.11}$$

$$A_i \subseteq B_j \text{ implies } B_i \subseteq B_j, \tag{2.12}$$

for any i and $C \subset U$ satisfying $A_i \subset C \subset B_i (A_i \neq C \neq B_i)$ 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. (2.13)

The set of pairs (A_i, B_i) satisfying (2.10)–(2.13) is called an *extension*. Its definition is not really beautiful, but it is needed in some applications. On the other hand, it is also an equivalent notion to the closures:

THEOREM 2.6 [9]

$\mathfrak{L} \to \mathcal{H}(\mathfrak{L})$ is a bijection between the set of intersection semi-lattices and the set of extensions.

Let $\mathcal{L}$ be a closure on U. Define $\mathcal{S}_i$ as the family of minimal sets $A \subseteq U$ such that $a_i \in \mathcal{L}(A)$. It is clear that no member of $\mathcal{S}_i$ is a subset of another member of it. Such families are called *inclusion-free* or *Sperner families*. Hence, it is obvious that

$$\mathcal{S}_i \ (1 \leq i \leq n) \text{ is a Sperner family} \tag{2.14}$$

and

$$\text{either } \mathcal{S}_i = \{\emptyset\} \text{ or } \{a_i\} \in \mathcal{S}_i. \tag{2.15}$$

One more, essential property can be proved:

$$\begin{aligned}&\text{if } A \subset U \text{ contains no subset belonging to } \mathcal{S}_i\\ &\text{then } \{\mathcal{S}_j : A \text{ contains a subset belonging to } \mathcal{S}_j\}\\ &\qquad\text{contains no subset belonging to } \mathcal{S}_i.\end{aligned} \tag{2.16}$$

PROPOSITION 2.7 [22]

The $|U|$ Sperner families satisfying (2.14)–(2.16) give an equivalent description of the closures.

A function $\mathcal{C}$ satisfying

$$\mathcal{C}(A) \subseteq A \quad (A \subseteq U) \tag{2.17}$$

is a *choice function*. Given a closure $\mathcal{L}$,

$$\mathcal{C}(A) = U - \mathcal{L}(U - A) \tag{2.18}$$

is a choice function.

THEOREM 2.8 [15]

The correspondence defined by (2.18) is a bijection between the sets of closures and the set of choice functions satisfying

$$\mathcal{C}(A) \subseteq B \subseteq A \text{ implies } \mathcal{C}(A) = \mathcal{C}(B) \text{ for all } A, B \subseteq U$$

and

$$A \subseteq B \text{ implies } \mathcal{C}(A) \subseteq \mathcal{C}(B) \qquad \text{for all } A, B \subseteq U.$$

Given a determination $\mathcal{L}$ (or a closure $\mathcal{D}$, etc.), it determines the family $\mathcal{K} = \mathcal{K}(\mathcal{L})$ (or $\mathcal{K} = \mathcal{K}(\mathcal{D})$) of minimal keys. It is a non-empty Sperner family. Conversely, if a non-empty Sperner family $\mathcal{K}$ is given, then

$$\mathcal{L}(A) = \begin{cases} A & \text{if there is no } K \in \mathcal{K} \text{ such that } K \subseteq A,\\ U & \text{if there is a } K \in \mathcal{K} \text{ such that } K \subseteq A\end{cases}$$

is a closure and the set of minimal keys in it is $\mathcal{K}$. This, theorem 2.1 and proposition 2.2 prove the following proposition.

PROPOSITION 2.9

For any non-empty Sperner family $\mathcal{K}$, there is a relation R in which the family of minimal keys is $\mathcal{K}$.

Of course, $\mathcal{K}$ does not determine $\mathcal{L}$ uniquely.

3. Inequalities for the parameters of a database

We have shown an example of these kinds of problems in the introduction. Theorem 1.1 determined the maximum number of minimal keys. Thalheim observed that this bound can be improved if the domains are bounded.

THEOREM 3.1 [37]

Suppose that $D(a_i) \le k$ $(1 \le i \le n)$, where $k^4 < 2n + 1$. Then the number of minimal keys cannot exceed

$$\binom{n}{\lfloor \frac{n}{2} \rfloor} - \left\lfloor \frac{n}{2} \right\rfloor.$$

PROBLEM 3.2

Improve this bound for small k's.

Let us show here an extension of theorem 1.1. In many practical cases, it is known that a certain set of attributes cannot uniquely determine a too large set of attributes. Formally, $|B - A| \le k$ (suppose $k \le n/2$) must hold for any functional dependency $A \to B$. It follows that the keys are of size at least $n - k$. As earlier, the minimal keys form a Sperner family. Thus, we have to find the largest Sperner family with members of size at least $n - k$. However, this is an easy task knowing the YBLM-inequality (Yamamoto [39], Bollobás [6], Lubell [29], Meshalkin [31]); it is often called the LYM-inequality:

YBLM-INEQUALITY

If the number of i-element members in a Sperner family $\mathcal{S}$ of n elements is f_i, then

$$\sum_{i=0}^{n} \frac{f_i}{\binom{n}{i}} \le 1. \tag{3.1}$$

In our case, if the Sperner family is the family of minimal keys, then $f_0 = f_1 = \ldots = f_{n-k-1} = 0$ holds. Use the inequality

$$\binom{n}{i} \le \binom{n}{n-k} \quad \text{if} \quad k \le n/2, \;\; n - k \le i \le n$$

in (3.1):

$$1 \ge \sum_{i=n-k}^{n} \frac{f_i}{\binom{n}{i}} \ge \sum_{i=n-k}^{n} \frac{f_i}{\binom{n}{n-k}} = \frac{\sum_{i=n-k}^{n} f_i}{\binom{n}{n-k}} = \frac{|\mathcal{S}|}{\binom{n}{n-k}}.$$

This proves the following statement.

THEOREM 3.3

Suppose that all functional dependencies $A \to B$ on an n-element set of attributes satisfy $|B - A| \le k$, where $k \le n/2$. Then the number of minimal keys is at most

$$\binom{n}{k}.$$

Thalheim [36,37] obtained interesting results for the same problems for the case of null-values (some data of some individuals are unknown).

The maximum number of functional dependencies is uninteresting, since the determination uniquely determined by the functional dependency $\varnothing \to U$ serves as the extremal one. (The number of functional dependencies is 2^{2n} here.) The situation is rather different if we consider only those functional dependencies which are non-trivial and non-reducible. A functional dependency $A \to B$ is called *basic* if

$$A \ne B,$$

$$\text{there is no } A' \subset A(A' \ne A) \text{ such that } A' \to B,$$

$$\text{there is no } B' \supset B(B' \ne B) \text{ such that } A \to B'.$$

Let $N(n)$ denote the maximum number of basic functional dependencies in a determination on n elements.

THEOREM 3.4 [3,28]

$$2^n\left(1 - \frac{4\log_2\log_2 n}{\log_2 e \log_2 n}\right)(1 + o(1)) \le N(n) \le 2^n\left(1 - \frac{\log_2^{3/2} n}{150\sqrt{n}}\right).$$

At first sight, it might be surprising that this number is near to the obvious upper bound 2^n. However, in this case the real question is to determine the deviation from this upper bound, that is, the second term. The above theorem gives only estimates.

A similar, but perhaps more natural parameter of a determination $\mathscr{D}$ is the following one. Let $\mathscr{E}$ be a set of functional dependencies on a set U, not necessarily satisfying the conditions (2.1)–(2.4). We say that $\mathscr{E}$ generates the determination $\mathscr{D}$ iff $\mathscr{E} \subseteq \mathscr{D}$ and $\mathscr{D}$ is the smallest such determination. The size $|\mathscr{E}|$ of the smallest $\mathscr{E}$ generating the determination $\mathscr{D}$ can be considered as the design complexity of $\mathscr{D}$. It is denoted by $C(\mathscr{D})$. Furthermore, introduce the notation $C(n) = \max\{C(\mathscr{D}) : \mathscr{D}$ is a determination on an n-element set$\}$ for the design complexity of the most complex determination in this sense. There is an obvious upper estimate by theorem

3.4 and $C(n) \leq N(n)$. (It is not known how far these two parameters can be.) The lower estimate can be obtained proving that

$$C\left(\mathfrak{D}^n_{\lfloor n/2 \rfloor}\right) = \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

(See Thalheim [38].)

THEOREM 3.5

$$c\,\frac{2^n}{\sqrt{n}} \sim \binom{n}{\lfloor \frac{n}{2} \rfloor} \leq C(n) \leq 2^n\left(1 - \frac{\log_2^{3/2} n}{150\sqrt{n}}\right).$$

A relation R in which the full family is $\mathfrak{D}^n_{\lfloor n/2 \rfloor}$ must have exponentially many rows (see lemma 3.10), that is $|R|$ must be very large. Mannila and Räihä [30] started to investigate the analogous question with bounded $|R|$. Let $C(n, m)$ denote $\max\{C(\mathfrak{D}) : \mathfrak{D}$ is a relation R on an n-element set U with size $|R|$ at most $m\}$. The following result, surprisingly, states that the minimum number of functional dependencies generating the worst determination remains exponential even in the case of linearly many rows.

THEOREM 3.6 [30]

$$C(2u + 1,\ 3u + 2) \geq 2^u.$$

PROBLEM 3.7

Find estimates for $C(n, m)$, in general.

Mannila and Räihä also investigated the algorithms finding the smallest (that is, of size $|C(\mathfrak{D})|$) $\mathscr{E}$ generating $\mathfrak{D}$. They have shown in [30] that the number of steps is at least $cm \log m$ for fixed n, where m is the number of rows and c is a constant independent of m. The brute force algorithm needs $O(n^2 2^m m \log m)$ steps. On the other hand, as a function of n the number of steps must be exponential. In the proof, they use the number of different $\mathfrak{D}$'s on an n-element set. (See section 4.)

PROBLEM 3.8

Give estimates on $N_k(n)$ and $C_k(n)$, where these numbers are defined analogously to $N(n)$ and $C(n)$ under the restriction $|B - A| \leq k$ for all functional dependencies $A \to B$.

It is known from the results surveyed in section 2 that there is a relation R for any determination $\mathfrak{D}$, closure $\mathscr{L}$, or set of minimal keys $\mathscr{K}$ such that R generates exactly this given $\mathcal{D}$, $\mathcal{L}$ or $\mathcal{K}$. It is not clear, however, what is the minimum of $|R|$ satisfying these goals. Let $s(\mathcal{D})$, $s(\mathcal{L})$ and $s(\mathcal{K})$ denote these minima.

THEOREM 3.9 [12, 14]

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

holds for any non-empty Sperner family $\mathcal{K}$ on n elements. On the other hand, there is such a $\mathcal{K}$ satisfying

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

The proof of the latter inequality is not constructive. We do not know the (nearly) worst Sperner families. A possible candidate is $\mathcal{D}^n_{\lfloor n/2 \rfloor}$. This is one of the motivations to study $s(\mathcal{K}^n_k)$, where $\mathcal{K}^n_k$ denotes the family of all k-element sets of an n-element set. The following easy lemma is surprisingly strong.

LEMMA 3.10 [13, 16]

$$\binom{s(\mathcal{K}^n_k)}{2} \geq \binom{n}{k-1} \quad (0 \leq k \leq n).$$

For $k = 1$, 2 or $n - 1$, the lower estimate obtained by this lemma is sharp. It can be shown by easy constructions. For $k = n$, this inequality is too weak, but the exact result can be obtained by a small trick.

THEOREM 3.11 [13, 16]

$$s(\mathcal{K}^n_1) = 2, \quad s(\mathcal{K}^n_2) = \left\lceil \frac{1 + \sqrt{1 + 8n}}{2} \right\rceil,$$

$$s(\mathcal{K}^n_{n-1}) = n, \quad s(\mathcal{K}^n_n) = n + 1.$$

The case $k = 3$ is very interesting from a mathematical point of view. Lemma 3.10 leads to $s(\mathcal{K}^n_3) \geq n$. In [13], we proved the equality for n's of the form $12r + 1$ and $12r + 4$ and conjectured that the equality holds for all $n \geq 7$. We also stated a conjecture for Steiner triple systems, where n is of the form $3r + 1$. This conjecture would imply the equality $s(\mathcal{K}^n_3) = n$. Rausche [32] found a counterexample for $n = 10$, but Gronau and Ganter [25] proved the second conjecture (therefore the first one, as well) for the integers $n = 3r + 1 \geq 13$. Bennett and Wu [4] independently proved the original conjecture for all $n \geq 7$ with the possible exception $n = 8$. Somewhat later, but independently, Gronau and Mullin [26] also settled the general case. (Very recently, Yeow Meng Chee [7] found a new proof for the second conjecture.)

THEOREM 3.12

$$s(\mathcal{K}_3^n) = n, \quad n \geq 7, \ n \neq 8.$$

For $k = 4, 5, \ldots, n-3, n-2$, one cannot expect a nice formula for $s(\mathcal{K}_k^n)$. However, it is asymptotically determined for fixed k and large n. In fact, lemma 3.10 gives an asymptotically correct lower estimate and the non-trivial construction given in [13] ensures the validity of

THEOREM 3.13

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

where c_1 and c_2 do not depend on n.

There is a similar result for large k's.

THEOREM 3.14 [23]

$$\tfrac{1}{12} n^2 \leq s(\mathcal{K}_{n-2}^n) \leq \tfrac{1}{2} n^2,$$

$$c_3 n^{(2k+1)/3} \leq s(\mathcal{K}_{n-k}^n) \leq c_4 n^k,$$

where c_3 and c_4 do not depend on n.

Theorem 3.9 gives some information on the worst (in the sense of minimum number of rows) key systems. It would be interesting to study smaller subclasses. We are able to offer only open problems.

PROBLEM 3.15

Determine max $s(\mathcal{K})$ for Sperner families on n elements inducing a determination containing functional dependencies $A \to B$ satisfying $|B - A| \leq k$, where k is a fixed integer.

PROBLEM 3.16

Determine max $s(\mathcal{K})$ (and min $s(\mathcal{K})$) for Sperner families on n elements, satisfying $|\mathcal{K}| = k$.

Practically nothing is known about this problem. However, it has a connection to another, perhaps easier problem. Let a subset $A \subseteq U$ be an *antikey* if it is not a key (= superset of a member of $\mathcal{K}$). The set of maximal antikeys is denoted by $\mathcal{K}^{-1}$. The following inequalities are known from [13]:

$$|\mathcal{K}^{-1}| \leq \binom{s(\mathcal{K})}{2}$$

and

$$s(\mathcal{K}) \leq 1 + |\mathcal{K}^{-1}|,$$

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

PROBLEM 3.17

Determine $\max|\mathcal{K}^{-1}|$ and $\min|\mathcal{K}^{-1}|$ for Sperner families having exactly $|\mathcal{K}| = k$ members.

We think that the minimum is attained for a family consisting of i and $(i + 1)$-element subsets, where i is determined by

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

if k is not too large relative to n.

There is a slight "philosophical" problem around $s(\mathcal{K})$. Mostly, it is supposed that the functional dependencies are given in advance, independently of the actual individuals (rows). Even in this case, it has some sense to determine the minimum number of rows generating the "full" system of functional dependencies, as was formulated by Thalheim [38]. On the other hand, our opinion is that one cannot completely exclude the view that the set of functional dependencies is (at least partially) determined by the actual data. In connection with this way of thinking, it is natural to ask what is the dependency structure of a random relation, as was suggested by Biskup [5].

Very little is known about $s(\mathcal{L})$ for closures (or, equivalently, determinations). Of course,

$$s(\mathcal{L}^n_k) = s(\mathcal{K}^n_k) \tag{3.2}$$

holds. Furthermore, there is a result on the s-function of direct products. This is not true for key systems.

Let $U = U_1 \cup U_2$ be a partition of U and let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two closures defined on U_1 and U_2, respectively. The direct product $\mathcal{L}_1 \times \mathcal{L}_2$ is defined by

$$(\mathcal{L}_1 \times \mathcal{L}_2)\,(A) = \mathcal{L}_1(A \cap U_1) \cup \mathcal{L}_2(A \cap U_2).$$

THEOREM 3.18 [13]

$$s(\mathcal{L}_1 \times \mathcal{L}_2) = s(\mathcal{L}_1) + s(\mathcal{L}_1) - 1.$$

Theorems 3.11, 3.18 and (3.2) enable us to determine $s(\mathcal{L})$ for several closures.

4. Partially ordered sets of closures

In this section, we will consider relations (databases) on a fixed attribute set U. More precisely, the closures generated by them will serve as models. That is, we forget about other properties of the databases (like other types of dependencies) – only the functional dependencies are considered. A further very natural condition is added, namely only the closures satisfying

$$\mathscr{L}(\varnothing) = \varnothing \tag{4.1}$$

are considered. (For the determinations, this means that $\varnothing \to A$ holds only for $A = \varnothing$.)

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 $A \to \{a\}$ $(A \subseteq U, a \in U)$ is true, then it remains true after the change. This implies

$$\mathscr{L}_1(A) \subseteq \mathscr{L}_2(A) \quad \text{(for all } A \subseteq U\text{)}, \tag{4.2}$$

where $\mathscr{L}_1$ and $\mathscr{L}_2$ denote the closures before and after the change. We write $\mathscr{L}_1 \geq \mathscr{L}_2$ in this case. It is easy to see that this property is transitive, consequently the closures on a fixed n-element set U satisfying (4.1) 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.

In section 2, we saw that the family of closed sets is an equivalent form of a closure. A closure satisfies (4.1) iff

$$\varnothing \in \mathfrak{X}(\mathscr{L}). \tag{4.3}$$

On the other hand, it is easy to see [9] that

$$\mathscr{L}_1 \leq \mathscr{L}_2 \quad \text{iff} \quad \mathfrak{X}(\mathscr{L}_1) \subseteq \mathfrak{X}(\mathscr{L}_2).$$

Hence, it follows that an equivalent form of P consists of the intersection semi-lattices containing $\varnothing$, ordered by inclusion as families.

It is easy to see that P has a *rank function* $r(\mathfrak{X}) = |\mathfrak{X}| - 2$, that is, r is zero for some element (namely, for $\mathfrak{X} = \{\varnothing, U\}$) and if $\mathfrak{X}_1 \subset \mathfrak{X}_2$ and there is no third element between them, then $r(\mathfrak{X}_2) = r(\mathfrak{X}_1) + 1$.

The first thing to study is the size of P. Consider the intersection semi-lattices consisting of U, some subsets of size $\lfloor n/2 \rfloor$ and all of their intersections. They are distinct and their number is

$$2^{\binom{n}{\lfloor n/2 \rfloor}}.$$

It was shown in [10] that the exponent in the upper estimate is at most

$$2\sqrt{2}\binom{n}{\lfloor n/2\rfloor}.$$

Recently, Alekseyev [1] proved that $2\sqrt{2}$ can be omitted.

THEOREM 4.1

$$2^{\binom{n}{\lfloor n/2\rfloor}} \le |P| \le 2^{\binom{n}{\lfloor n/2\rfloor}(1+o(1))}.$$

PROBLEM 4.2

Determine P asymptotically.

Since the number of Sperner families is determined by Korshunov [27] asymptotically (not only the asymptotics of the exponent!), there is some hope that the same can be done using proposition 2.7 and Korshunov's theorem.

There are some initial results concerning the sizes of the lower levels of P.

THEOREM 4.3 [11]

The number $\alpha(n, k)$ of the elements of rank k in P satisfies

$$\alpha(n, k) \sim \delta(k)\,(k+1)^n,$$

where k is fixed and n tends to infinity.

The next theorem deals with the levels near to the top. (The top rank is $2^n - 2$).

THEOREM 4.4 [11]

$$\alpha(n, 2^n - 2 - k) \sim \rho(k) n^k,$$

where k is fixed and n tends to infinity.

Comparing theorems 4.3 and 4.4, one can see that P is very asymmetric: the lower levels are much wider than the top ones.

PROBLEM 4.5

Determine approximately the widest level of P.

Theorems 4.3 and 4.4 suggest that the widest level is far below the middle.

To continue our investigations to understand the structure of P, the next task is to determine the minimum and maximum degrees at each level. Let $\deg_a(\mathcal{L})$ and

$\deg_b(\mathcal{L})$ denote the number of edges going upward and downward, respectively, from $\mathcal{L}$ in the Hasse diagram of P. The following functions are defined:

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

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

THEOREM 4.6 [9]

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

THEOREM 4.7 [9]

$$f_2(n, k) = 0 \qquad \text{iff} \ \ k = 2^n - 2,$$
$$f_2(n, k) = 1 \qquad \text{iff} \ \ k = 2^n - 2^{n-a-1} - 2,$$

for some $0 < a < n$. If $k > 2^{n-1} + 2$, then $f_2(n, k) \le$ the number of bits 1 in the binary expansion of $2^n - k - 2$. This is at most $n - 1$.

Let us mention that the proof is based on the somewhat strange notion of $\mathcal{H}(\mathfrak{L})$, see theorem 2.6.

THEOREM 4.8 [9]

$$\lceil \log_2(k+1) \rceil \le f_4(n, k) \le \lfloor \log_2(k+2) \rfloor - 1$$
$$+ \text{(the number of non-zero digits in the binary form of } k+2).$$

PROBLEM 4.9

Give estimates on $f_3(n, k)$.

5. Branching and partial dependencies

We now introduce a more general (weaker) dependency than the functional dependency. We do this first in a very particular case to show the usefulness of the concept. Let $A \subseteq U$ and $b \in U$. We say that b *(1, 2)-depends* on A if the values in

A determine the values in a "two-valued" way. That is, there exist no three rows being the same in A but having different values in b. We denote this by $A \to (1, 2) \to b$. Similarly, $A \to (1, q) \to b$ if there exist no $q + 1$ rows having the same values in each column of A, but containing $q + 1$ different values in column b.

As an example, 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 a truck goes through exactly four countries on each trip (counting the start and endpoints, too) and does not enter a country twice during one trip. Suppose furthermore that there are thirty possible countries, and one country has at most five neighbours. Let a_1, a_2, a_3, a_4 denote the countries as attributes. It is easy to see that $a_1 \to (1, 5) \to a_2$, $\{a_1, a_2\} \to (1, 4) \to a_3$ and $\{a_2, a_3\} \to (1, 4) \to a_4$. Now, we cannot decrease the size of the stored matrix, as in the case of functional (that is, (1, 1)-) dependencies, but we can decrease the range of the values in the new matrices. The domains $D(a_i)$ in the original database have thirty possible values, names of the countries or some codes of them (5 bits each, at least). Let us store a small table ($30 \times 5 \times 5 = 750$ 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 the attribute a_1 by these numbers (a_1^*), because the value in a_1 gives the starting country and the value in a_1^* determines the second country with the help of the small table. The same holds for the attribute a_3, but here the number of possible values can be even further decreased if another table is given containing the numbering of 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 value of the original relation could be encoded by 5 bits, now for the cost of two small auxiliary tables we could decrease the length of the values in the second column to 3 bits, and that of the elements in the third and fourth columns to 2 bits.

It is easy to see that the same idea can be applied in each case when the paths of a graph are stored, whose maximum degree is much less than the number of its vertices.

After this long motivation, let us give the general definition. Fix a relation R on the set of attributes U. Let $A \subseteq U$, $b \in U$ and $1 \le p \le q$ integers. We say that b *(p, q)-depends* on A if there are no $q + 1$ rows of R such that they contain at most p different values in each attribute in A, but $q + 1$ different values in b.

In [20], we called these dependencies *branching*. The referee called our attention to the paper of Grant and Minker [24], in which they introduced the numerical dependencies. These are identical to our $(1, q)$-dependencies. Their theorems are special cases of the forthcoming ones.

Define the mapping $\mathcal{J} = \mathcal{J}_{Rpq} : 2^U \to 2^U$ by

$$\mathcal{J}(A) = \{b : A \to (p, q) \to b\}.$$

PROPOSITION 5.1

$\mathcal{J}$ has the following properties:

$$A \subseteq \mathcal{J}(A), \tag{5.1}$$

$$A \subseteq B \text{ implies } \mathcal{J}(A) \subseteq \mathcal{J}(B), \tag{5.2}$$

for any subsets $A, B \subseteq U$.

The set-functions satisfying conditions (5.1) and (5.2) are called *increasing-monotone functions*. Note that (5.1) is identical with (2.6) and (5.2) is identical with (2.7). An increasing-monotone function, however, does not in general satisfy the third property (2.8) of closures.

Are these two conditions enough? We have only partial answers to this question. We say that an increasing-monotone function $\mathcal{N}$ is *(p, q)-representable* iff there is a relation R such that $\mathcal{N} = \mathcal{J}_{Rpq}$.

THEOREM 5.2 [20]

Let $\mathcal{N}$ be an increasing-monotone function satisfying $\mathcal{N}(\varnothing) = \varnothing$. Then, $\mathcal{N}$ is (p, q)-representable if one of the following conditions holds:

$$p = 1 \text{ and } 1 \le q,$$

$$p = 2 \text{ and } 3 < q,$$

$$2 < p \text{ and } p^2 - p - 1 < q.$$

PROBLEM 5.3

Is the statement of theorem 5.2 true for any $p < q$? Is it possible to drop the condition $\mathcal{N}(\varnothing) = \varnothing$?

The first undecided case is $p = 2$, $q = 3$. The situation is significantly different if $p = q$.

PROPOSITION 5.4 [20]

$\mathcal{J}_{Rpp}$ is a closure for any $1 \le p$.

Thus, it is natural to ask if all closures are (p, p)-representable for any given p. If $p < q$, then we know that $\mathcal{J}$, in general, is not a closure. However, is it at least true that all closures are (p, q)-representable? The answer, in general, is negative.

THEOREM 5.5 [20]

If $p > 2$, $n > 6$, then $\mathcal{L}_2^n$ is not (p, p)-representable.

The situation is better if $p = q = 2$ or $p < q$.

THEOREM 5.6 [20]

Every closure is (p, q)-representable if one of the following condition holds:

$$p = 1 \text{ and } 1 \le q,$$
$$p = 2 \text{ and } 2 \le q,$$
$$2 < p \text{ and } \frac{(p+1)^2}{2} < q.$$

Hence, we can see that any closure is (1, 1)-representable (we knew this a lot earlier!) and (2, 2)-representable. However, it is not true for (3, 3)-representation.

PROBLEM 5.7

Characterize the (3, 3)-representable closures.

One might think that this characterization, if found, is good for all (p, p). This is not true, as we will see using the following theorem.

THEOREM 5.8 [34]

$\mathcal{L}_k^n$ is (p, p)-representable for $p = 1, 2, 3, 4, 2k-3, 2k-2$, if $k > 2$ and is not (p, p)-representable for $\frac{3}{2}k - 1 \le p \le 2k-4$ and for $2k-1 \le p$ if $n > n_0(k)$, $k > 1$.

For instance, $\mathcal{L}_4^n$ is (p, p)-representable for large n iff $p = 1, 2, 3, 4, 5, 6$. This is a closure which is (6, 6)-representable but not (7, 7)-representable.

In the cases where a representation is found, one can define $s_{pq}(\mathcal{N})$ as the minimum of $|R|$ for relations representing $\mathcal{N}$. In this part, we do not pose open problems since they are obvious; the results are very modest.

THEOREM 5.9 [21]

$$s_{1q}(\mathcal{N}) \le 2qn2^n$$

holds for any integer $q > 1$ and increasing-monotone function $\mathcal{N}$.

Lemma 3.10 can easily be generalized for this case. This generalization helps to prove the following statement:

THEOREM 5.10 [21]

$$s_{pq}(\mathcal{L}_1^n) = q + 1,$$
$$s_{22}(\mathcal{L}_2^n) = 2n \text{ if } n > 3.$$

THEOREM 5.11 [33]

$$s_{pp}(\mathscr{L}_n^n) = \min\left\{v : \binom{v-1}{p} \geq n\right\}.$$

Finally, let us only briefly mention the *partial dependencies*. The vector $\alpha = (a_{i_1}, \ldots, a_{i_k}; r_1, \ldots, r_k)$ is called a *partial function* where the a's are elements of U and $r_h \in D(a_{i_h})$. We say that $\beta = (b_{j_1}, \ldots, b_{j_l}; s_1, \ldots, s_l)$ *depends* on α in R if each row containing r_h in the column of the attributes a_{i_h} (for all $1 \leq h \leq k$) contains s_h in the attribute b_{j_h} (for all $1 \leq h \leq l$). The paper [19] contains investigations concerning this dependency.

Acknowledgement

The authors are indebted to the referee for his valuable suggestions and for calling our attention to some papers in the literature.

References

[1] V.B. Alekseyev, Diskret. Mat. 1(1989)129–136.

[2] W.W. Armstrong, Dependency structures of data base relationship, in: *Information Processing 74* (North-Holland, Amsterdam) pp. 580–583.

[3] A. Békéssy, J. Demetrovics, L. Hannák, P. Frankl and G.O.H. Katona, On the number of maximal dependencies in a data base relation of fixed order, Discr. Math. 30(1980)83–88.

[4] F.E. Bennett and Lisheng Wu, On minimum matrix representation of closure operations, Discr. Appl. Math. 26(1990)25–40.

[5] J. Biskup, private communication.

[6] B. Bollobás, On generalized graphs, Acta Math. Hungar. 16(1965)447–452.

[7] Yeow Meng Chee, Design-theoretic problems in perfectly $(n-3)$-error-correcting databases, SIAM J. Discr. Math., submitted.

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

[9] G. Burosch, J. Demetrovics and G.O.H. Katona, The poset of closures as a model of changing databases, Order 4(1987)127–142.

[10] G. Burosch, J. Demetrovics, G.O.H. Katona, D.J. Kleitman and A.A. Sapozhenko, On the number of database closure operations, Theor. Comput. Sci. 78(1991)377–381.

[11] G. Burosch, J. Demetrovics, G.O.H. Katona, D.J. Kleitman and A.A. Sapozhenko, On the number of database closure operations, II, Discr. Appl. Math., submitted.

[12] J. Demetrovics, On the equivalence of candidate keys with Sperner systems, Acta Cybern. 4(1979) 247–252.

[13] J. Demetrovics, Z. Füredi and G.O.H. Katona, Minimum matrix representation of closure operations, Discr. Appl. Math. 11(1985)115–128.

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

[15] J. Demetrovics, G. Hencsey, L.O. Libkin and I.B. Muchnik, On the interaction between closure operations and choice functions with applications to relational databases, Acta Cybern., to appear.

[16] J. Demetrovics and G.O.H. Katona, Extremal combinatorial problems in a relational database, in: *Fundamentals of Computation Theory 81, Proc. 1981 Int. FCT-Conf.*, Szeged, Hungary, 1981, Lecture Notes in Computer Science 117 (Springer, Berlin, 1981) pp. 110–119.

[17] J. Demetrovics and G.O.H. Katona, Combinatorial problems of database models, in: *Coll. Math. Soc. János Bolyai, 42. Algebra, Combinatorics and Logic in Computer Science,* Győr, Hungary, 1983 (North-Holland, Amsterdam, 1986) pp. 331–353.

[18] J. Demetrovics and G.O.H. Katona, Extremal combinatorial problems of databases, in: *MFDBS'87, 1st Symp. on Mathematical Fundamentals of Database Systems*, Dresden, Germany, Lecture Notes in Computer Science (Springer, 1987) pp. 99–127.

[19] J. Demetrovics, G.O.H. Katona and D. Miklós, Partial dependencies in relational databases and their realization, Discr. Appl. Math., to appear.

[20] J. Demetrovics, G.O.H. Katona and A. Sali, On the characterization of branching dependencies, Discr. Appl. Math., to appear.

[21] J. Demetrovics, G.O.H. Katona and A. Sali, Branching dependencies in relational databases (in Hungarian), Alkalmaz. Mat. Lapok, to appear.

[22] J. Demetrovics and Son Hua Nam, Closures and Sperner families, *Coll. Math. Soc. János Bolyai, Extremal Problems for Families of Subsets,* Visegrád, Hungary (1991), submitted.

[23] Z. Füredi, Perfect error-correcting databases, Discr. Appl. Math. 28(1990)171–176.

[24] J. Grant and J. Minker, Normalization and axiomatization for numerical dependencies, Inf. Contol 65(1985)1–17.

[25] H.-O.O.F. Gronau and B. Ganter, On two conjectures of Demetrovics, Füredi and Katona concerning partitions, Discr. Math. 88(1991)149–155.

[26] H.-O.O.F. Gronau and R.C. Mullin, Preprint.

[27] A.D. Korshunov, On the number of monotone Boolean functions, Problemy Kibernet. 38(1981) 5–108, in Russian.

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

[29] D. Lubell, A short proof of Sperner's lemma, J. Combinat. Theory 1(1966)299.

[30] H. Mannila and K.-J. Räihä, On the complexity of inferring functional dependencies, Discr. Appl. Math., to appear.

[31] L.D. Meshalkin, A generalization of Sperner's theorem on the number of subsets of a finite set, Teor. Veroyatnost. i Primenen. 8(1963)219–220, in Russian.

[32] A. Rausche, On the existence of special block designs, Rostock Math. Kolloq. 35(1985)13–20.

[33] A. Sali, Extremal problems for finite partially ordered sets and matrices, Thesis for "kandidátus" degree, Hungarian Academy of Sciences, Budapest (1990), in Hungarian.

[34] A. Sali, private communication.

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

[36] B. Thalheim, A review of research on dependency theory in relational databases I, II, Preprint, Technische Universität Dresden, Sektion Mathematik (1986).

[37] B. Thalheim, On the number of keys in relational databases, Discr. Appl. Math., to appear.

[38] B. Thalheim, *Dependencies in Relational Databases* (Teubner, Leipzig, 1991).

[39] K. Yamamoto, Logarithmic order of free distributive lattices, J. Math. Soc. Japan 6(1954)347–357.