EXTREMAL COMBINATORIAL PROBLEMS IN RELATIONAL DATA BASE

by

J. DEMETROVICS

Computer and Automation Institute
Hungarian Academy of Sciences

G.O.H. KATONA

Mathematical Institute of the Hungarian
Academy of Sciences

1. INTRODUCTION

One of the possible models of a data base is the following "relational" model introduced by Codd [1]. In a data base there are several different data of several different individuals. E.g. a data base 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. The entries of the jth row or jth attribute can be chosen from a set D_j where these sets are not necessarily disjoint. However, for our investigations the choice of sets D_j plays no role, thus we may assume the D_j's are equal to the set of non-negative integers. That is, we shall simply consider $m\times n$ matrices with non-negative entries. As two different individuals should have different records we may assume that the rows of the matrix are different.

For a fixed $m\times n$ matrix M, a set $K\subset\{1,\ldots,n\}$ of columns is called a *key* iff the rows of the submatrix $M(K)$ determined by the given columns are different. That is, K is a key iff the attribute corresponding to K *uniquely* determine the records (individual) in the sense that given any values of these attributes there is at most one record having these values. A key K is called *minimal* iff there is no different key K' satisfying $K'\subset K$.

Let $A\subset B\subset\{1,\ldots,n\}$. We say (write) that A implies B $(A\rightarrow B)$ when two rows of $M(A)$ are equal iff the corresponding rows of $M(B)$ are also equal. In other words, A implies B when the values of the attributes in A uniquely determine the values of the attributes in B. (That is, A is a "key" in B in a certain sense.) If $A\rightarrow B$ then the pair (A,B) is called a *functional dependency* or simply a *depen-*

dency. A dependency (A,B) is called *basic* iff

1) $A \neq B$
2) there is no $A' \subset A$, $A' \neq A$ such that $A' \to B$
3) there is no $B' \supset B$, $B' \neq B$, such that $A \to B'$.

There are many natural extremal problems concerning the above concepts. The present paper gives a *survey* of the not very numerous results of this type. However, the main air is to call the reader's attention to these applicable and mathematically interesting problems.

The next section determines the maximal number of minimal keys when the number n of attributes is given. Section 3 gives lower and upper estimates on the maximal number of basic dependencies. In the last section there are some initial results in the determination of the smallest number m of rows such that there is a data base ($m \times n$ matrix) in which the family of minimal keys is given beforehand.

2. MAXIMUM NUMBER OF MINIMAL KEYS

The keys play an important role in data bases. The records can be uniquely found by them. Of course, it is worth-while to consider the minimal ones, only. It is quite natural to ask at most how many minimal keys can exist.

Theorem 1 [3]. *The maximal number of minimal keys in a data base with n attributes is*

$$\binom{n}{\left[\frac{n}{2}\right]} \tag{1}$$

Proof. The minimal keys K are subsets of $\{1,\ldots,n\}$ and do not contain each other. Sperner's well-known theorem [2] states that such a family can not contain more than $\binom{n}{\left[\frac{n}{2}\right]}$ members.

We will now construct an $m \times n$ matrix (with $m = \binom{n}{\left[\frac{n}{2}\right]-1} + 1$) having $\binom{n}{\left[\frac{n}{2}\right]}$ minimal keys.

The first row of M consists of nothing but 1's. The other rows contain $\left[\frac{n}{2}\right]-1$ 1's in all possible ways while the remaining entries of the i-th row are i's $\left(2 \leq i \leq \binom{n}{\left[\frac{n}{2}\right]-1} + 1\right)$. For $n = 4$, see the matrix below:

$$\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 3 & 1 & 3 & 3 \\ 4 & 4 & 1 & 4 \\ 5 & 5 & 5 & 1 \end{array} .$$

If we chose $\left[\frac{n}{2}\right]$ places in a row then we find there either only *1*'s or at least one number i different from *1*. Therefore the row i is uniquely determined. Any K with $|K| = \left[\frac{n}{2}\right]$ is a key. On the other hand, it is easy to see that no set K with $|K| < \left[\frac{n}{2}\right]$ can be a key, the first row coincides with another one in $M(K)$. The proof is complete.

Let us remark that a stronger statement is proved in [3]. There is a matrix M for any prescribed family of keys if it has the Sperner-property.

3. MAXIMUM NUMBER OF BASIC DEPENDENCIES

All dependencies trivially follow from the basic dependencies. Therefore their number can be considered the complexity of the data base. Thus our aim given in the title of the section is in fact equivalent to the problem of finding the most complex data base.

Let $N(n)$ denote the maximum number of basic dependencies in a data base with n attributes (or of a matrix with n columns).

It is easy to construct amatrix in which the basic dependencies are of the form $A \to A \cup \{x\}$ where x is a fixed attribute. That is, $2^{n-1} \leq N(n)$. However, the real value is greater:

Theorem 2 [4]

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

Sketch of the proof (It can be found in [4]) *Lower estimate.* Let $q_1, q_2, \ldots, q_k$ be positive integers satisfying $\sum_{i=1}^{k} q_i = n$. The $\left(\binom{q_i}{2}+1\right) \times q_i$ matrix Q_i is defined in the following way. The first row of Q_i consists of nothing but 1's. The other rows contain $q_i - 2$ 1's in all the possible ways. In the rest of the places we put i's in the ith row $(2 \le i \le \binom{q_i}{2} + 1)$. The rows of the matrix $M = Q_1 \times \ldots \times Q_k$ are all the $\prod_{i=1}^{k} (\binom{q_i}{2} + 1)$ possible combinations of the rows of $Q_1, \ldots, Q_k$. (More precisely, we put together an arbitrary row of Q_1, then an arbitrary row of Q_2 e.t.c., in this order.) It follows from $\sum_{i=1}^{k} q_i = n$ that M has n columns. We show the construction for $n = 5$, $k = 2$, $q_1 = 3$, $q_2 = 2$:

$$\begin{array}{ccc|cc}
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 2 & 2 \\
1 & 2 & 2 & 1 & 1 \\
1 & 2 & 2 & 2 & 2 \\
3 & 1 & 3 & 1 & 1 \\
3 & 1 & 3 & 2 & 2 \\
4 & 4 & 1 & 1 & 1 \\
4 & 4 & 1 & 2 & 2
\end{array} .$$

Let Q_i^* denote the set of indices of columns of Q_i in M, that is $Q_i^* = \{q_1 + \ldots + q_{i-1} + 1, \ldots, q_1 + \ldots + q_i\}$. One can see that the basic dependencies (A,B) in M are those satisfying

$$min(q_i - |A \cap Q_i^*|) = 1$$

$$B = A \cup \bigcup_{|A \cap Q_i^*| = q_i - 1} Q_i^* .$$

Their number is

$$\prod_{i=1}^{k} (2^{q_i} - 1) - \prod_{i=1}^{k} (2^{q_i} - q_i - 1). \tag{3}$$

This give a lower estimate on $N(n)$. In the above example $(n=5,\ q_1=3,\ q_2=3)$ this estimate gives $17 \le N(5)$

For large n we choose first

$$q = q(n) = \left[\log n - \log \left(\frac{1}{\log e}(\log \log n - \log \log \log n - \log \log e - 1)\right)\right]$$

where $\log$ means the logarithm of base 2. Then $k = k(n)$ and $r = r(n)$ are defined by $n = q(n)\ k(n) + r(n)$ $\quad 0 \le r(n) < q(n)$.

Finally choose

$$q_1 = q_2 = \dots = q_r = q + 1$$

$$q_{r+1} = \dots = q_k = q.$$

With this choice (3) gives the left hand side of (2) after some long but elementary calculations.

Upper estimate. Let H be the family of sets A having a pair B such that (A,B) is a basic dependency. It can be seen that the set C satisfying $A \subset C \subset B$ $|C| = |A| + 1$ is not an element of H. Such a C can be obtained from at most n different sets A only, consequently for at least $|H|/n$ sets C $C \notin H$, implying $|H| + \frac{|H|}{n} \le 2^n$. This is equivalent to the right hand side of (2). The theorem is proved.

That is, it remains an *Open question*: what is the proper second term of $N(n)$?

4. SMALLEST NUMBER OF INDIVIDUALS REALIZING A GIVEN FAMILY OF MINIMAL KEYS

Suppose a family F of subsets of n-element sets is given and it is known that the data base has exactly these sets as minimal keys. It is useful to know something about the number of rows of the data base. This section offers some results along this line.

Let F satisfy the Sperner-property, that is, $A,B \in F$ $A \neq B$ implies $A \not\subset B$. Then $s(F)$ denotes the minimum number m of the rows of an $n \times m$ matrix M in which the set of minimal keys is F.

Theorem 3. [5] *For any $n > 0$ there is an F satisfying*

$$s(F) \ge \frac{1}{n^2} \binom{n}{\left[\frac{n}{2}\right]}. \tag{4}$$

Proof. Let $s(n) = \max_F s(F)$, where F runs over the families satisfying the Sperner-property. We say that M *realizes* F if the set of minimal keys of M is F. Then any F can be realized by an $m \times n$ M where $m \leq s(n)$. Write completely new and different integers into the $(m+1)$th, $(m+2)$th,...,$s(n)$th row. The new enlarged $s(n) \times n$ matrix realizes the same F. Consequently any F can be realized by an $s(n) \times n$ matrix. Furthermore, we may suppose that the entries of these matrices are $1, 2, \dots, s(n)$, only.

Indeed, if the integers found in the first column of M are $i_1 < i_2 < \dots\ i_r$ $(1 \leq r \leq s(n))$ then let us make the changes $i_j \to j$. This change does not change the family of minimal keys. Thus, following the same procedure with all the columns independently, we finally arrive to a $s(n) \times n$ matrix containing only $1, \dots, s(n)$ as entries. The number of such matrices is $s(n)^{n\ s(n)}$, therefore we have

$$s(n)^{n\ s(n)} > \text{number of } F\text{'s}. \tag{5}$$

Any family of some $\left[\frac{n}{2}\right]$-element sets has the Sperner-property, consequently,

$$s(n)^{n\ s(n)} > 2^{\binom{n}{[\frac{n}{2}]}} \tag{6}$$

follows from (5). (6) results in (4). The theorem is proved.

Now we will try to determine $s(F)$ for some simple particular F's. Let F_k^n denote the family of k-element subsets of an n-set. First an easy lemma.

Lemma 1. *For any* $0 < k \leq n$

$$\binom{s(F_k^n)}{2} \geq \binom{n}{k-1} \tag{7}$$

holds.

Proof. Suppose that an $m \times n$ matrix M realizes F_k^n. For any $(k-1)$-element set A of columns there is a pair of rows in which these columns have identical entries. Moreover for different A's these pairs of rows must be different. Hence $\binom{m}{2} \geq \binom{n}{k-1}$ and (7) follow. The lemma is proved.

From this lemma $s(F^n_1) \geq 2$ follows. The construction

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

shows that $s(F^n_1) = 2$.

Let us determine $s(F^n_2)$. If an $m \times n$ matrix realizes F^n_2 then by the above lemma

$$n \leq \binom{m}{2} \tag{8}$$

holds. On the other hand, if (8) is satisfied we are able to construct an $m \times n$ matrix M realizing F^n_2. M contains two 0's in each column. Of course the pair of places is different for different columns. The other entries of the ith row will be i's $(1 \leq i \leq m)$. Consequently $s(F^n_2)$ is the smallest integer m satisfying (8).

Let us apply lemma 1 for $k = n-1$. We obtain $s(F^n_{n-1}) \geq n$. On the other hand $m = n$ is enough for the construction:

$$\begin{matrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 0 & 0 & \dots & 1 \end{matrix} ,$$

this leads to $s(F^n_{n-1}) = n$.

The determination of $s(F^n_n)$ needs a slightly harder consideration. If M is an $m \times n$ matrix, let $G(M)$ denote the graph whose vertices are the rows of M, two vertices are connected with an edge iff the set A of places where the two rows are equal is non-empty. The edge is *labelled* by A.

Lemma 2. *Let M be a matrix and let $A_1, \dots, A_r$ be the labels along a circuit of $G(M)$. Then.*

$$(9) \qquad \bigcap_{\substack{i=1\\ i\neq j}}^{r} A_i - A_j = \emptyset$$

Proof. Suppose that, in the contrary, (9) is non-empty, that is, there is a column, say the uth one, which is an element of all A_i but A_j. Let the vertices of the circuit be $k_1,\dots,k_r$ in such a way that the edge (k_i, k_{i-1}) is labelled by A_i $(1 \leq i < r)$ and the edge (k_r, k_1) is labelled by A_r. From $u \in A_{j+1}$ it follows that the k_{j+1}th and k_{j+2}th entries in the uth column are equal. The same holds for the k_{j+2}th and k_{j+3}th entries, etc. Consequently, the k_{j+1}th, k_{j+2}th,…,k_rth, k_1th,…,k_jth entries in the uth column are all equal. This leads to $u \in A_j$ contradicting the assumption. The proof is complete.

Now we are able to determine $s(F_n^n)$. Suppose that the $m \times n$ matrix M realizes F_n^n. An $(n-1)$-element subset A of $\{1,\dots,n\}$ is not a key therefore there must be an edge in $G(M)$ labelled with A. Consequently $G(M)$ has n different edges labelled with the $(n-1)$-element subsets of $\{1,\dots,n\}$. These edges cannot form a circuit because the $(n-1)$-element subsets cannot satisfy (9), the lemma is applicable. $G(M)$ has at least $n+1$ vertices, $m \geq n+1$. The following $(n+1) \times n$ matrix realizes F_n^n

$$\begin{matrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{matrix} .$$

We summarize our moderate results in a theorem.

Theorem 4

$$s(F_1^n) = 2, \qquad s(F_2^n) = \left\lfloor \frac{1+\sqrt{1+4n}}{2} \right\rfloor$$

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

In the case of $k = 3$ lemma 1 gives $s(F_3^n) \geq n$, again. There are problems, however, with the construction of an $n \times n$ matrix realizing F_3^n. Checking the proof of lemma 1 we can see that equality can stand in (7) iff for any pair of rows there are exactly two equal entries. In fact a column is a partition $P_i = (A_{i1}, \ldots, A_{ir_i})$ on n elements (the set of rows) $(1 \leq i \leq n)$. These partitions have to satisfy the following conditions:

1) For any pair x, y of elements of the n-element set X there are exactly two sets A_{ip} and A_{jq} containing both of them $(i \neq j)$

2) For any given $i \neq j$ there is a unique pair p, q such that $|A_{ip} \cap A_{jq}| = 2$. $|A_{ip} \cap A_{jq}| < 2$ holds for the other pairs.

Condition 1) implies

$$\sum_{j=1}^{r_i} \binom{|A_{ij}|}{2} = n - 1 \tag{10}$$

for any $1 \leq i \leq n$. If $n = 3$ or 6 (10) has no solution, therefore $s(F_3^n) > n$ for these values. However, for the values of form $n = 3k + 1$ there is always a solution $|A_{i1}| = \ldots = |A_{ik}| = 3$, $|A_{i,k+1}| = 1$. This suggests the following conjecture.

<u>Conjecture 1.</u> There is a system of 3-element subsets of a $3k + 1$ -element set 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 subfamilies where a subfamily consists of k disjoint 3-sets. 3) Exactly one pair of members of two different subclasses meet in 2 elements.

The problem of *resolvable Steiner systems* is very closely related. This problem is solved in [6]. We were able to construct them for $n = 4$ and 7 (using the Fano-geometry):

$$\begin{array}{cccc} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{array} \qquad \begin{array}{ccccccc} 2 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 2 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 2 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 2 \end{array}$$

Conjecture 1 implies $s(F_3^n) = n$ for infinitely many n. However we state

Conjecture 2 $s(F_3^n) = n \quad (n \geq 7)$.

We state the next lemma without proof.

Lemma 3.

$$s(F_k^{n+1}) \leq s(F_k^n) + s(F_{k-1}^n) .$$

This implies $s(F_3^n) \leq c\, n^{3/2}$, a fairly weak upper bound.

REFERENCES

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

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

[3] DEMETROVICS, J., On the equivalence of candidate keys with Sperner Systems, Acta Cybernetica 4 (1979) 247-252.

[4] BÉKÉSSY, J., DEMETROVICS, J., HANNÁK, K., FRANKL, P. and KATONA, G.O.H., On the number of maximal dependencies in a data base relation of fixed order, Disc. Math. 30(1980) 83-88.

[5] DEMETROVICS, J. and GYEPESI, GY., On the functional dependency and some generalization of it (to appear in Acta Cybernetica).

[6] RAY-CHAUDHURI, D.K., and WILSON, R.M., Solution of Kirkman's school girl problem, Proc. Symp. in Pure Math., Combinatorics, Am. Math. Soc. 19(1971) 187-204.