Design type problems motivated by database theory [1]

János Demetrovics [a], Gyula O.H. Katona [b], Attila Sali [b,*]

[a] *Computer and Automatization Institute of HAS, Budapest, Kende u. 13-17, H-1111, Hungary*
[b] *Mathematical Institute of HAS, Budapest P.O.B. 127 H-1364, Hungary*

Abstract

Let $k \leqslant n$, $p \leqslant q < m$ be positive integers. Suppose that the $m \times n$ matrix M satisfies the following two properties:

- for any choice of k distinct columns $c_1, c_2, \ldots, c_k$, there are $q+1$ rows such that the number of different entries in c_i $(1 \leqslant i \leqslant k-1)$ in these rows is at most p, while all $q+1$ entries of c_k in these rows are different;
- this is true for no choice of $k+1$ distinct columns.

We review results minimizing m, given n, p, q, k. Two of the results are new. The optimal or nearly optimal constructions can be considered as n partitions of the m-element set satisfying certain conditions. This version leads to the *orthogonal double covers*, also surveyed here. © 1998 Elsevier Science B.V. All rights reserved.

AMS classification: primary: 68P15 68R05; secondary: 05D05

Keywords: Relational Database; Design; Orthogonal double cover; Hamiltonian cycle

1. Introduction, motivation

The motivation of this study is the *relational database model*. Suppose, we want to store data of students of R.C. Bose. For each individual we create a record that contains certain fields, such as *name*, *initials*, *year of birth*, etc. We imagine this written in a table, where rows correspond to individuals and columns to types of data, as in Table 1. The types of data, the columns in our table are called *attributes*. The set of attributes is usually denoted by Ω. Some attributes determine the values of other

[*] Corresponding author.

[1] The work was supported by the Hungarian National Foundation for Scientific Research grant numbers T016389, T016524 and 4267, and European Communities (Cooperation in Science and Technology with Central and Eastern European Countries) contract number CIPACT930113.

Table 1

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**
Last name	Initials	M or F	Year of birth	Month of birth	Day of birth	Age in years	Age in months	Age in days
Srivastava	J.N.							
Shrikhande	S.S.							
Rao	C.R.							
Ray-Chaudhuri	D.K.							
Connor	W.S.							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

attributes, in our example we have the following implications (among others):

$$\{d\} \to \{g\}$$
$$\{d,e\} \to \{h\} \quad \text{(and } \{g\}\text{)}$$
$$\{d,e,f\} \to \{i\} \quad \text{(and } \{g,h\}\text{).}$$

We shall give formal definitions of the concepts discussed here in the next section. These types of implications, which are called *functional dependencies* play very important role in the implementations of relational database models. For detailed explanation, see Codd (1970), Demetrovics et al. (1992), Ullman (1989). However, there are other types of determinations, for example if we know the value of d, then there can occur at most *twelwe* different values of h. These types of dependencies are called *branching dependencies* and were introduced in Demetrovics et al. (1992).

Armstrong observed that functional dependencies give rise to *closure operations* on the set of attributes. He proved in Armstrong (1974) that in fact, all closure operations can be derived from suitable relational database models. The question naturally arises that for a given closure, which is the smallest database that yields it. We survey results about this problem and show how these questions lead to the very interesting design-theoretical concept of *orthogonal double cover* in Section 2.

Branching dependencies do not always yield closures, instead they give rise to *monotone-increasing* set functions. In Demetrovics et al. (1992), the problem that whether all monotone-increasing functions could be derived from relational database models was raised, and partial results were presented. In Demetrovics et al. (1995), the minimum representations of monotone-increasing set functions and closures by branching dependencies were studied. These problems also lead to nice design-theoretical questions that are interesting for their own sake, as well. We discuss known results on this topic and present two new theorems in Section 3. In Section 4 some intriguing open problems are listed.

Finally, let us fix some notations. Ω always denotes the n-element set of columns of a database matrix. Capital letters refer to subsets of Ω, while elements of Ω are denoted by lower case letters.

2. Functional dependency

A relational database system of the scheme $R(A_1, A_2, \ldots, A_n)$ will be considered as a matrix, where the columns correspond to the *attributes* A_i's (for example name, date of birth, place of birth, etc.), while the rows are the n-tuples of the relation r. That is, a row contains the data of a given *individual*. For the sake of convenience, it is assumed that the rows of the matrix are pairwise distinct. Let Ω denote the set of attributes (the set of the columns of the matrix). Let $A \subseteq \Omega$ and $b \in \Omega$. We say that b (*functionally*) *depends* on A (see Armstrong (1974), Codd (1970)) if the data in the columns of A determine the data of b, that is there exist no two rows which agree in A but differ in b. We denote this by $A \to b$. A set function on the subsets of Ω can be defined with the help of functional dependency.

Definition 2.1. Let M be the matrix of a relational database. The function $\mathscr{C}_M : 2^{\Omega} \to 2^{\Omega}$ is defined by

$$\mathscr{C}_M(A) = \{b \colon b \in \Omega,\ A \to b\}$$

for any $A \subseteq \Omega$. We shall write $\mathscr{C}$ instead of $\mathscr{C}_M$ if it does not cause confusion.

The function defined above has the following three properties.

Proposition 2.2.
(1) $A \subseteq \mathscr{C}(A)$,
(2) $A \subseteq B \Rightarrow \mathscr{C}(A) \subseteq \mathscr{C}(B)$,
(3) $\mathscr{C}(\mathscr{C}(A)) = \mathscr{C}(A)$.
Set functions satisfying properties (1)–(3) *are called* closure operations. *Armstrong proved that the above correspondence could be reversed.*

Theorem 2.3 (Armstrong (1974)). *For any given closure $\mathscr{C}$ there exists a matrix M such that*

$$\mathscr{C}_M = \mathscr{C}.$$

It is evident that a matrix with a small number of rows cannot yield a complicated closure. Furthermore, as closures and database matrices are equivalent by Armstrong's theorem, the following number is a measure of complexity of closures.

Definition 2.4. Let $\mathscr{C}$ be a closure on Ω. Then let

$$s(\mathscr{C}) = \min_{M : \mathscr{C}_M = \mathscr{C}} \{\text{number of rows in } M\}.$$

It is very hard to determine $s(\mathscr{C})$ for an arbitrary closure $\mathscr{C}$. However, there are nice combinatorial results for certain closures.

Definition 2.5. Let $\mathscr{C}_n^k$ denote the following closure on Ω:

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

The following lemma gives a general lower bound for $s(\mathscr{C}_n^k)$.

Lemma 2.6 (Demetrovics and Katona (1981)).

$$\binom{s(\mathscr{C}_n^k)}{2} \geqslant \binom{n}{k-1}.$$

Proof of Lemma 2.6. Suppose that M represents $\mathscr{C}_n^k$ and let $|A| = k - 1$ be a subset of Ω, furthermore let $b \notin A$. Then by the definition of $\mathscr{C}_n^k$, $A \not\to b$, i.e., there is a pair of rows i and j, such that they are identical in A, but different in b. If there is another $(k-1)$-subset B of Ω such that i and j are identical on B, as well, then $A \cup B \not\to b$ would hold, but $|A \cup B| \geqslant k$, so by the definition of $\mathscr{C}_n^k$ this cannot happen. Thus, we can assign distinct pairs of rows to distinct $(k-1)$-subsets of columns. □

The exact value of $s(\mathscr{C}_n^k)$ is determined for certain values of k.

Theorem 2.7 (Demetrovics and Katona (1981)). *The following equalities hold*:

(a) $s(\mathscr{C}_n^1) = 2$,
(b) $s(\mathscr{C}_n^2) = \lceil (1 + \sqrt{1 + 8n})/2 \rceil$,
(c) $s(\mathscr{C}_n^{n-1}) = n$,
(d) $s(\mathscr{C}_n^n) = n + 1$.

We give the proof of Case (b) as an example.

Proof of Case(b) of Theorem 2.7. Let $s = s(\mathscr{C}_n^2)$. Lemma 2.6 gives $\binom{s}{2} \geqslant n$. Note that the number of the right hand side of equality in Case (b) is the smallest s satisfying the previous inequality. If s is such, then we construct a matrix M with s rows such that $\mathscr{C}_M = \mathscr{C}_n^2$ as follows:

$$M = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 1 & 1 & 1 & \ldots & 1 \\ 0 & 2 & 2 & \ldots & 2 & 0 & 0 & 0 & \ldots & 2 \\ 3 & 0 & 3 & \ldots & 3 & 0 & 3 & 3 & \ldots & 3 \\ 4 & 4 & 0 & \ldots & 4 & 4 & 0 & 4 & \ldots & 4 \\ 5 & 5 & 5 & \ldots & 5 & 5 & 5 & 0 & \ldots & 5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s & s & s & \ldots & 0 & s & s & s & \ldots & s \end{pmatrix}.$$

There is a pair of zeros in every column of M such that for different columns the zeros are in different pairs of rows, which implies that every one-element subset of Ω

is closed. This can be done by the choice of s. On the other hand, no two rows agree in more than one column, so if $A \subseteq \Omega$ with $|A| > 1$, then $\mathscr{C}_M(A) = \Omega$. □

Let us note that in Case (d) of Theorem 2.7 Lemma 2.6 yields only $s(\mathscr{C}_n^n) > \sqrt{2n}$, hence some other tricks are needed to prove $n+1$ as a lower bound.

Let us now consider the case $k = 3$. From Lemma 2.6 we obtain that

$$\binom{s(\mathscr{C}_n^3)}{2} \geqslant \binom{n}{2},$$

hence $s = s(\mathscr{C}_n^3) \geqslant n$. Equality holds if we can construct an $n \times n$ matrix M such that:

(1) for any distinct $a, b, c \in \Omega$ there are two rows equal in columns a and b, but different in c.

(2) for any distinct $a, b, c \in \Omega$ there are no two rows equal in all of them.

Consider the dual problem. A column naturally determines a partition of the set Y of rows, by the equalities of its entries. We say that a partition *covers* the pair (α, β) $(\alpha, \beta \in Y,\ \alpha \neq \beta)$ iff α and β are in the same class of the partition. We can state the previous two properties as follows.

Find n partitions of Y $(|Y| = n)$ such that:

(1′) for any two partitions there exists a pair (α, β) covered by both,

(2′) no pair (α, β) is covered by three different partitions.

However, the number of pairs of partitions is also $\binom{n}{2}$ and different pairs of partitions cannot cover the same pair of elements by (2′). Thus, we may conclude that (1′) and (2′) (consequently (1) and (2)) are equivalent to:

(i) for any two partitions there is exactly one pair of elements, which is covered by both,

(ii) each pair of elements is covered by exactly two different partitions.

Definition 2.8. A collection of partitions satisfying (i) and (ii) is called an *orthogonal double cover.*

The following conjecture was formulated in Demetrovics et al. (1985). (It was posed in other terms, since the notion of orthogonal double cover was introduced later, in Ganter et al. (1994).)

Ex-Conjecture 1. *There exists an orthogonal double cover of the n-element set by n partitions provided $n \geqslant 7$.*

In the same paper they proved that Ex-Conjecture 1 is true for certain n's.

Theorem 2.9 (Demetrovics et al. (1985)). *Ex-Conjecture* 1 *is true if* $n = 12r + 1$ *or* $n = 12r + 4$.

In the proof they used a theorem of Hanani to construct special type of partitions, namely each partition consisted of one 1-element class and $4r$ ($4r+1$, resp.) 3-element classes. This motivated the following conjecture.

Ex-Conjecture 2 (Demetrovics et al. (1985)). *If $n=3r+1$, then there exists an orthogonal double cover of the n-element set by n partitions that have one 1-element class and r of the 3-element classes.*

Note that the two conjectures are independent in the sense that the solution of one of them does not imply the solution of the other. The first result about these conjectures was negative. In 1987 (Rausche, 1987), Rausche observed that Ex-Conjecture 2 is not true for $n=10$. However, that turned out to be the only 'bad case'. Ganter and Gronau and later Yeow Meng Chee (Preprint) independently proved the following.

Theorem 2.10 (Gronau and Ganter (1991)). *Ex-Conjecture 2 is true for $n \geqslant 13$.*

The first conjecture was decided affirmatively, as well. Bennett and Wu proved the following theorem.

Theorem 2.11 (Bennett and Lisheng Wu (1990)). *Ex-Conjecture 1 is true.*

If we have an orthogonal double cover by partitions, then we can define a graph for each partition. The vertex set is the underlying set Ω, the edges are the pairs covered by that partition. These graphs are unions of disjoint cliques. Furthermore, in the case of Ex-Conjecture 2 these graphs are pairwise isomorphic, namely they are unions of r K_3's and an isolated point. This observation motivated the following definition.

Definition 2.12. A collection of n pairwise isomorphic graphs $G_1, G_2, \ldots, G_n$ with the same vertex set V, where $|V|=n$ and $G_i=(V,E_i)$, is called an *orthogonal double cover by graphs* iff

(1) each edge of K_n is contained in exactly two of the E_i's,
(2) $|E_i \cap E_j|=1$ for $i \neq j$.

With this concept, Theorem 2.11 states that there exists a double cover by graphs where the $G_i=K_1 + r * K_3$. Gronau et al. (1995) proved a conjecture of Chung and West (1994) stating that there is an orthogonal double cover by graphs where each G_i has maximum degree at most two. A sharpening of this result was given by Ganter et al.

Theorem 2.13 (Ganter et al. (1994)). *For all $n \geqslant 4$, $n \neq 8$ there is an orthogonal double cover by graphs where each G_i consists of the isolated vertex i and a union of disjoint cycles of length 3,4 or 5 only.*

Further results are obtained by Gronau et al. (Preprint), and Leck and Leck (Preprint) for G_i consisting of cycles, Gronau et al. (Preprint) for G_i being certain trees, respectively.

The exact value of $s(\mathscr{C}_n^k)$ is not known for $k>3$. However, if k is fixed, then its asymptotic behaviour is known.

Theorem 2.14 (Demetrovics et al. (1985)). *If k is fixed and $n>n_0(k)$, then*

$$c_1(k)n^{(k-1)/2} \leqslant s(\mathscr{C}_n^k) \leqslant c_2(k)n^{(k-1)/2}.$$

The lower bound in Theorem 2.14 follows from Lemma 2.6. The upper bound is proven by a construction involving polynomials over a finite field. Füredi proved some bounds for the 'other end' of the range of k.

Theorem 2.15 (Füredi (1990)). *If k is fixed and $n>n_0(k)$, then*

$$c_3(k)n^{(2k+1)/3} \leqslant s(\mathscr{C}_n^{n-k}) \leqslant c_4(k)n^k.$$

The following concept allows us to find $s(\mathscr{C})$ for infinitely many closures.

Definition 2.16. Let $\mathscr{L}$ and $\mathscr{N}$ be closures on the ground sets U and V, respectively, with $U \cap V = \emptyset$. The *direct product* of $\mathscr{L}$ and $\mathscr{N}$ is the closure on the ground set $U \cup V$ defined by

$$(\mathscr{L} \times \mathscr{N})(A) = \mathscr{L}(A \cap U) \cup \mathscr{N}(A \cap V) \quad \text{for } A \subseteq U \cup V.$$

The size of a minimum representation of a direct product of closures can be calculated provided that the minimum representation known for the members of the product.

Theorem 2.17 (Demetrovics et al. (1985)).

$$s(\mathscr{C}_1 \times \mathscr{C}_2) = s(\mathscr{C}_1) + s(\mathscr{C}_2) - 1.$$

Theorem 2.17 provides an alternative proof for Case (d) of Theorem 2.7, one has only to observe that $\mathscr{C}_n^n = \mathscr{C}_{n-1}^{n-1} \times \mathscr{C}_1^1$.

3. Branching dependencies

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

Definition 3.1. Let M be an $m \times n$ matrix, with column set Ω. Let $A \subseteq \Omega$ and $b \in \Omega$. We say that b *(p,q)-depends* on A if there are no $q+1$ rows of M such that they contain at most p different values in each column of A, but $q+1$ different values in b.

The functional dependency discussed in the previous section is a special case, namely it is the $(1,1)$-dependency. For a given matrix M we define a function from the family of subsets of Ω into itself as follows.

Definition 3.2. Let M be the given matrix. Let us suppose, that $1 \leqslant p \leqslant q$. Then the mapping $\boldsymbol{J}_{Mpq} : 2^{\Omega} \to 2^{\Omega}$ is defined by

$$\boldsymbol{J}_{Mpq}(A) = \{b : A \xrightarrow{(p,q)} b\}.$$

We collect two important properties of the mapping $\boldsymbol{J}_{Mpq}$ in the following proposition, see Demetrovics et al. (1992).

Proposition 3.3. *Let r, Ω, M, p and q as above. Furthermore, let $A, B \subseteq \Omega$. Then*
(1) $A \subseteq \boldsymbol{J}_{Mpq}(A)$,
(2) $A \subseteq B \Rightarrow \boldsymbol{J}_{Mpq}(A) \subseteq \boldsymbol{J}_{Mpq}(B)$.

Definition 3.4. Set functions satisfying (i) and (ii) are called *increasing-monotone functions*. We say that such an increasing-monotone function $\mathscr{N}$ is (p,q)-representable if there exists a matrix M such that $\mathscr{N} = \boldsymbol{J}_{Mpq}$.

It is not known yet whether any increasing-monotone function is (p,q)-representable for arbitrary $1 \leqslant p < q$ or not. The following is known about representability of general monotone-increasing functions.

Theorem 3.5. *If*
(iii) $p=1$, $1<q$ *or*
(iv) $p=2$, $3<q$ *or*
(v) $2<p$, $p^2-p-1<q$
then any $\mathscr{N}$ is (p,q)-representable.

Let us note that it is neccessary to assume $p<q$ if we want to (p,q)-represent a general increasing-monotone function, because the following proposition was proved in Demetrovics et al. (1992).

Proposition 3.6. *For any matrix M, $\boldsymbol{J}_{Mpp}$ is a closure operation.*

A statement analogous to Theorem 2.3 holds for $(2,2)$-representation. However, if $p>2$, then this analogy cannot be continued:

Proposition 3.7 (Demetrovics et al. (1992)). *If $n>6$ and $p>2$, then $\mathscr{C}_n^2$ is* not *(p,p)-representable.*

Theorem 3.5 allows us to formulate the following definition.

Definition 3.8. For an increasing-monotone function $\mathcal{N}$ let $s_{pq}(\mathcal{N})$ denote the minimum number of rows of a matrix that (p,q)-represents $\mathcal{N}$. If $\mathcal{N}$ is not (p,q)-representable, then we put $s_{pq}(\mathcal{N})=\infty$.

The following general upper bound can be proved.

Theorem 3.9 (Demetrovics et al. (1995)). *Let $\mathcal{N}$ be an increasing-monotone function with $\mathcal{N}(\emptyset)=\emptyset$ and let (p,q) satisfy one of* (iii)–(v) *from Theorem* 3.5. *Then*

$$s_{pq}(\mathcal{N})\leqslant q(n+1)2^n.$$

The above bound is quite coarse, which is caused by the lack of knowledge about the structure of an increasing-monotone function. However, there is a nice structure theory of closures (that are special increasing-monotone functions). A lemma analogous to Lemma 2.6 can be proved in a similar way.

Lemma 3.10. *Let us assume that $\mathscr{C}_n^k$ is (p,q)-representable. Then*

$$\binom{s_{pq}(\mathscr{C}_n^k)}{q+1}\geqslant\binom{n}{k-1.}.$$

The exact value of $s_{pq}(\mathscr{C}_n^k)$ is known in a few cases only Demetrovics et al. (1995).

Theorem 3.11 (Demetrovics et al. (1995)).

$$s_{pq}(\mathscr{C}_n^1)=q+1,$$
$$s_{22}(\mathscr{C}_n^2)=2n,$$
$$s_{pp}(\mathscr{C}_n^n)=\min\left\{v\ \textit{integer:}\ \binom{v-1}{p}\geqslant n\right\}.$$

The proof of the last equality in Theorem 3.11 is based on a theorem of Lovász on k-trees (Lovasz, 1979). We present two new theorems in the rest of the section. The first one answers the Open Problem in Demetrovics et al. (1995).

Theorem 3.12.

$$3^{1/3}\,n^{2/3}+\mathrm{O}(n^{1/3})<s_{22}(\mathscr{C}_n^3)<\frac{3}{4^{1/3}}n^{2/3}+\mathrm{o}(n^{2/3}).$$

In order to prove Theorem 3.12 we need the following lemma.

Lemma 3.13. *The point-line pairs (P,l) $(P\in l)$ of the projective plane* PG$(2,q)$ *can be colored with $q+1$ colors so that pairs with the same first or second coordinates receive distinct colors.*

Proof of Lemma 3.13. Such a coloring corresponds to the complete 1-factorization of the regular bipartite point-line incidence graph of $\mathrm{PG}(2,q)$. □

Proof of Theorem 3.12. The lower bound follows from Lemma 3.10. The upper bound will be proved by a construction. We will construct a bipartite graph $\mathrm{G}(A,B,E)$ with color classes A and B ($|A|+|B|=r$), where the set of edges E is a union of matchings $T_1,T_2,\dots,T_t$. Let $V(T_j)$ denote the set of vertices covered by T_j. G will satisfy the following three properties:

(vi) $V(T_i)\cap V(T_j)\neq\emptyset$ for any i,j,
(vii) $T_i\cap T_j=\emptyset$ for $i\neq j$ (no edge is covered twice),
(viii) $\forall C\in\binom{A\cup B}{3}$ $\binom{C}{2}\not\subseteq\bigcup T_j$ (no triangle).

Suppose for a moment that $\mathrm{G}(A,B,E)$ is constructed. The $r\times t$ matrix M showing the upper bound is constructed as follows. The columns of M will be indexed by the matchings, while the rows will be indexed by the points of the bipartite graph. In a column indexed by some T_i we will have identical elements for the row pairs determined by the edges of T_i, different identical pairs for different edges, the other elements will be pairwise distinct and distinct from these pairs. In other words, columns of M correspond to partitions into two and one element classes. We claim, that this matrix $(2,2)$-represents $\mathscr{C}_t^3$. Let T_x denote the matching corresponding to column x of M. Indeed, by property (vi) there exist three rows u,v,w for any pair (a,b) of columns that contain at most two different entries in these columns. If there were a third column c also containing at most two distinct values in u,v and w, then by (vii) the equal entries in a, b and c must be in pairwise distinct pairs of rows. So, with $C=\{u,v,w\}$, $\binom{C}{2}\subseteq T_a\cup T_b\cup T_c$ would hold, that contradicts (viii). This proves that for every two-element subset $A\subset\Omega$ $\boldsymbol{J}_{M22}(A)=A$. The same argument shows that if $D\subset\Omega$ with $|D|>2$, then there exist no three rows containing at most two different entries in each column from D, hence $\boldsymbol{J}_{M22}(D)=\Omega$. $\boldsymbol{J}_{M22}(\emptyset)=\emptyset$ and $\boldsymbol{J}_{M22}(\{a\})=\{a\}$ for all $a\in\Omega$ follows from (ii) of Proposition 3.3.

Now the only thing left is to construct $\mathrm{G}(A,B,E)$ with $r\sim ct^{2/3}$. Let A be the point set of $\mathrm{PG}(2,q)$ and let $C=\{1,2,\dots,q+1\}$ be a $q+1$-element set. q^2+q+1 matchings can be constructed using Lemma 3.13, as follows. The matching T_l will correspond to line l of $\mathrm{PG}(2,q)$, namely if P is a point incident to l and the color of (P,l) is i, then T_l contains the edge (P,i), so $|T_l|=q+1$. The graph $\mathrm{G}(A,C,\bigcup T_l)$ satisfies (vii), which follows from Lemma 3.13. Finally, any bipartite graph satisfies (viii) trivially.

If B is a union of k pairwise disjoint copies of C $(C_1,C_2,\dots,C_k)$ and the above matchings from A are constructed for each copy C_i, then it is not hard to see that the graph $\mathrm{G}(A,B,\bigcup T_j)$ also satisfies (vi)–(viii). For example, $V(T_i)$ and $V(T_j)$ intersect in A, because $V(T_i)\cap A$ is a line of $\mathrm{PG}(2,q)$, for all i.

This G results in an $r\times t$ matrix M, where $r=q^2+q+1+k\,(q+1)$ and $t=k\,(q^2+q+1)$. This gives $r\sim 3q^2$ and $t\sim 2q^3$ if $k=2q$. □

Theorem 3.14.

$$s_{12}(\mathscr{C}_n^2) = \min\left\{ s \text{ integer: } \binom{s}{3} \geqslant 2n \right\},$$

provided $n > 452$.

First we prove the lower bound in Theorem 3.14. Note, that Lemma 3.10 gives only $\binom{s_{12}(\mathscr{C}_n^2)}{3} \geqslant n$. Suppose, that M is a matrix of m rows and n columns that $(1, 2)$-represents $\mathscr{C}_n^2$. Each column of M determines a partition of the row set $\{1, 2, \ldots, m\}$ according to which entries are the same. The partition corresponding to column i is denoted by Π_i. A triplet $\{i, j, k\}$ is an *indicator* for the partition (column) Π_t if there is another column u such that i, j, k are in the same class of Π_t but are in three different classes in Π_u. (That is, the triplet of rows shows that $t \overset{(1,2)}{\not\longrightarrow} u$.) The following two facts hold.

Fact 1. *A triplet can be an indicator for at most one column.*

Fact 2. *For any pair of columns t and u, there is an indicator triplet $\{i, j, k\}$ for Π_t such that i, j and k are in three different classes of Π_u.*

Partition Π_t is called of *first kind*, iff there exist at least two different indicator triplets for Π_t. Otherwise, the partition is called of *second kind.*

Proposition 3.15. *Let Π_u be a partition of second kind. Then the elements i, j, k of the indicator triple of Π_u are all in different classes in any other partition Π_t.*

Proof of Proposition 3.15. If not all three elements were in different classes of Π_t, then another triplet should show that $u \overset{(1,2)}{\not\longrightarrow} t$, so Π_u would not be of second kind. □

As a corollary, we obtain that the indicator triplets of partitions of second kind form an at most 1-intersecting system. We need the following easy lemma.

Lemma 3.16. *Let $\mathscr{T} = \{T_1, T_2, \ldots, T_k\}$ be an at most* 1*-intersecting system of triplets of M. Then there exists a collection $\mathscr{S}$ of k triplets of M such that each member of $\mathscr{S}$* 2*-intersects at least one member of $\mathscr{T}$.*

Proof of Lemma 3.16. Let $\mathscr{S}$, $|\mathscr{S}| = s$ be the system of such triplets that 2-intersect at least one member of $\mathscr{T}$. We use double counting, namely we count the number of pairs (T, S), $T \in \mathscr{T}$, $S \in \mathscr{S}$ and $|S \cap T| = 2$. On one hand, counting by the T's, it is $3k(m-3)$. On the other hand, for each S there are at most $3(m-3)$ T's that 2-intersects, so $s3(m-3)$ is at least as large as the number to count, which imply $s \geqslant k$. □

Now we are ready to prove the lower bound in Theorem 3.14.

Proof of Theorem 3.14. According to Fact 1 all indicator triplets are different. Partitions of the first kind each use at least two of them. The indicator triplets of partitions of the second kind can be matched with triplets of M so that matched pairs 2-intersect, by Lemma 3.16 and Hall's condition. These matched triplets cannot coincide with some indicator triplet by Fact 2, so we have found two 'own' triplets for each partition of second kind, as well. This proves $\binom{m}{3} \geqslant 2\,n$.

We prove the upper bound in Theorem 3.14 via construction. In fact, we consider the number of rows m to be given, and construct $n = \lfloor \binom{m}{3}/2 \rfloor$ columns so that the $(1,2)$-dependency in that matrix will be exactly $\mathscr{C}_n^2$. First the following technical theorem is needed.

Theorem 3.17. *Let $G_0 = (V, E_0)$ and $G_1 = (V, E_1)$ be simple graphs on the same vertex set $|V| = N$, such that $E_0 \cap E_1 = \emptyset$. The* 4*-tuple (x, y, z, v) is called an* alternating cycle *if (x, y) and (z, v) are in E_0 and (y, z) and (x, v) are in E_1. Let r be the minimum degree of G_0 and let s be the maximum degree of G_1. Suppose, that*

$$2r - 8s^2 - s - 1 > N;$$

then there is a Hamiltonian cycle in G_0 such that if (a, b) and (c, d) are both edges of the cycle, then (a, b, c, d) is not an alternating cycle.

The proof of Theorem 3.17 is based on Dirac's famous theorem (Dirac, 1952) on sufficient condition for existence of a Hamiltonian cycle and on Lemma 3.19.

Theorem 3.18. *If G is a simple graph on N points and all degrees of G are at least $N/2$, then G has a Hamiltonian cycle.*

Lemma 3.19. *Let G_0, G_1, r, s and N satisfy the conditions of Theorem* 3.17. *Let us assume that there is a Hamiltonian path from a to b in G_0. Then there exist c, $c \neq a$, and d, $d \neq b$, adjacent vertices along the path, such that c is between a and d on the path, $(a, d) \in E_0$, $(b, c) \in E_0, (a, d, b, c)$ is not an alternating cycle, and if (x, y) is an edge of the path, then neither (a, d, x, y) nor (b, c, x, y) are alternating cycles.*

Proof of Lemma 3.19. We call a vertex $x \in V$ *a-bad* (*b-bad*) if there exist an edge (y, z) of the Hamiltonian path, such that (a, x, y, z) ((b, x, y, z), respectively) is an alternating cycle. The statement of the lemma requires an edge (c, d) of the Hamiltonian path, such that $a \neq c, b \neq d, (a, d), (b, c) \in E_0, c$ is not b-bad, d is not a-bad and c is between a and d on the path, furthermore either $(a, c) \notin E_1$ or $(b, d) \notin E_1$. Thus, we call a vertex x *not a-good*, if it is a neighbour of a along the path, or a-bad, or its neighbour on the Hamiltonian path in the direction of a is b-bad or $(b, x) \in E_1$. Similarly, y is *not b-good* if it is a neighbour of b along the path, or b-bad or its neighbour on the path in the direction of b is a-bad.

Let t_a be the number of a-bad vertices and t_b be that of the b-bad vertices. Now, t_a is bounded from above by the number of four-tuples (a,z,y,x) such that (y,z) is an edge of the path and (a,z,y,x) is an alternating cycle. This latter number can be bounded from above by $2s^2$, because the number of possible z's is at most s, then z has 2 neighbours along the path, and the number of possible x's is at most s, again. Similarly, $t_b \leqslant 2s^2$.

The number of not a-good vertices is at most $t_a + t_b + s$, similarly the number of not b-good vertices is at most $t_b + t_a$. Now, a has at least r neighbours in G_0, so the number of candidates for the required vertex d is at least $r-2-4s^2-s$ ($d \neq b$ and d cannot be the neighbour of a along the path, which yields the term -2). Similarly, the number of possible c's is at least $r-2-4s^2$. The condition $2r-4-8s^2-s>N-3$ implies that there is a pair of candidates adjacent along the Hamiltonian path, as it is required. □

Proof of Theorem 3.17. Let us suppose indirectly, that $2r-8s^2-s-1>N$, but the required Hamiltonian cycle does not exist. Let K *contains an alternating cycle* mean that there exists an alternating cycle whose E_0 edges are edges of K, where K stands for a path or a cycle in G_0.

If $E_1=\emptyset$, then the condition of Dirac's theorem holds for G_0, thus it contains a Hamiltonian cycle, furthermore no alternating cycle could exist. So, we may assume that E_1 is non-empty. Let us drop edges one-by-one from E_1 until a required Hamiltonian cycle appears. Consider the last dropped edge (u,v). Dropping it, a Hamiltonian cycle containing no alternating cycle appears. This means, that there was a Hamiltonian cycle C in G_0 before, which contained such alternating cycles only that used edge $(u,v)\in E_1$. Let the neighbours of v along C be w and z. An alternating cycle using the edge (u,v) must use either (w,v) or (z,v). Thus, the path of length $N-1$ from w to z obtained by deleting the vertex v from C contains no alternating cycle.

Lemma 3.19 can be applied for the Hamiltonian path obtained from C by deleting the edge (z,v), taking $a=v$ and $b=z$. Replacing the edges (c,d) (provided by Lemma 3.19) and (z,v) with edges (v,d) and (z,c) a new Hamiltonian cycle C' is obtained, which can contain an alternating cycle only if that alternating cycle uses the edge (w,v). Now, a second application of Lemma 3.19 with $a=w$ and $b=v$ gives a Hamiltonian cycle C'' containing no alternating cycle, even without dropping the edge (u,v), a contradiction. □

Now we are able to prove the main tool for our construction.

Theorem 3.20. *Let $|X|=n$ and $2k>q$. The family of all q-subsets of X can be partitioned into unordered pairs (except possibly one if $\binom{n}{q}$ is odd), so that paired q-subsets are disjoint and if A_1,B_1 and A_2,B_2 are two such pairs with $|A_1\cap A_2|\geqslant k$, then $|B_1\cap B_2|<k$, provided $n>n_0(k,q)$.*

Proof of Theorem 3.20. We construct graphs $G_0=(V,E_0)$ and $G_1=(V,E_1)$ that satisfy the requirements of Theorem 3.17. The vertex set V consists of the q-subsets of X,

$|V|=\binom{n}{q}=N$. Two q-subsets are adjacent in G_0 if their intersection is empty, while two q-subsets are adjacent in G_1 if they intersect in at least k elements. The minimum degree of G_0 is $r=\binom{n-q}{q}$ and the maximum degree of G_1 is $s=\sum_{j=k}^{q}\binom{q}{j}\binom{n-q}{q-j}$. It is easy to check that $2r-8s^2-s-1>N=\binom{n}{q}$, provided $n>n_0(k,q)$.

According to Theorem 3.17, there is a Hamiltonian cycle H in G_0 that does not contain two disjoint edges that span an alternating cycle. Now the required partition of the q-subsets into disjoint pairs can be obtained by going around H, every other edge will form a good pair. □

Theorem 3.20 is used to prove the upper bound in Theorem 3.14.

Proof of the upper bound of Theorem 3.14. Let us suppose, that m is an integer that satisfies $\binom{m}{3}\geqslant 2n$. A matrix with m rows and n columns will be constructed that $(1,2)$-represents $\mathscr{C}_n^2$. Let us denote the set of rows by X. Apply Theorem 3.20 with $q=3$ and $k=2$ to obtain disjoint pairs of 3-subsets of X. There are $\lfloor\binom{m}{3}/2\rfloor$, that is, at least n such pairs. Choose n of them. We construct a column from such a pair, as follows. Put 1's in the rows indexed by the first 3-set, 2's in the rows indexed by the second one, and all different entries, that are at least 3, in the other positions.

If a and b are two distinct columns, then there are no 3 rows that agree in both a and b, because we used all distinct 3-subsets of rows, hence $\{a,b\}\xrightarrow{(1,2)}\Omega$. On the other hand, if a is constructed from the pair of 3-subsets A_1,A_2 and b is constructed from B_1,B_2, then either $|A_1\cap B_1|<2$ or $|A_2\cap B_2|<2$, so there are 3 rows which contain all identical entries in column a, but all distinct ones in column b, hence $a\overset{(1,2)}{\not\longrightarrow}b$. □

We note that for $q=3$ and $k=2$ the $n_0(k,q)$ of Theorem 3.17 can be calculated exactly, namely $n_0(2,3)=452$.

4. Open problems

There are two kinds of problems involving matrix representations of closures and monotone-increasing functions. The first one is the question of representability. This is completely solved by Armstrong (Theorem 2.3) for functional ((1,1)-) dependency. However, Theorem 3.5 leaves the following question open.

Open Problem 1. *Is every monotone-increasing function* (p,q)*-representable if* $p<q$?

We believe, but dare not put it as Conjecture that the answer is yes. The first open case is $p=2, q=3$. However, if $\mathscr{N}$ is a closure, then it was proved in Demetrovics et al. (1992) that $\mathscr{N}$ is (p,q)-representable for $p=2\leqslant q$ and for $p>2$ and $q\geqslant((p+1)/2)^2$, so for closures the first open case is $p=4, q=5$. We can state the particular case of the previous question.

Open Problem 2. *Is every closure operation $\mathscr{C}$ (p,q)-representable if $p<q$?*

For $p=q$ we have seen that the answer is negative. However, it is an interesting question for which p is a closure (p,p)-representable.

Open Problem 3. *For which p's is $\mathscr{C}_n^k$ (p,p)-representable?*

The answer is known for $n>n_0(k)$ and shows surprising facts Sali and Sali (submitted). If $k=n$ then Theorem 3.11 gives the complete answer. However, nothing is known for $2<k<n_0(k)$.

The other direction of investigations is to find the minimum representation, provided a representation exists. One intriguing question is the maximum possible value of $s(\mathscr{C})$ if $\mathscr{C}$ is a closure on an n-element set. Let us denote this number by $s(n)$. We know that $(1/n^2)\binom{n}{\lfloor n/2\rfloor}\leqslant s(n)$ (Demeterovics and Gyepesi, 1983) and that $s(n)\leqslant\binom{n}{\lfloor n/2\rfloor}$. So we propose the following:

Open Problem 4. *Find the value of $s(n)$.*

Of course, we do not know the value of $s_{pq}(\mathscr{C}_n^k)$ for other triplets p,q,k than the ones in Theorems 3.11, 3.12 and 3.14.

Finally, we refer the interested reader to Demetrovics and Katona (1993), where other problems concerning relational databases are presented.

We are indebted to the referees for their helpful suggestions.

References

Armstrong, W.W., 1974. Dependency structures of database relationships. Information Processing. 74 North-Holland, Amsterdam, pp. 580–583.

Bennett, F.E., Lisheng Wu, 1990. On mimimum matrix representation of closure operations. Discrete Appl. Math. 26, 25–40.

Chee, Y.M., Preprint. Design-theoretic problems in perfectly $(n-3)$-error-correcting databases.

Chung, M.S., West, D.B., 1994. The p-intersection number of a complete bipartite graph and orthogonal double coverings of a clique. Combinatorica 14, 453–461.

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

Demetrovics, J., Füredi, Z., Katona, G.O.H., 1985. Minimum matrix representation of closure operations. Discrete Appl. Math. 11, 115–128.

Demeterovics, J., Gyepesi, Gy., 1983. A note on minimum matrix reperesentation of closure operations. Combinatorica 3, 177–180.

Demetrovics, J., Katona, G.O.H., 1981. 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, vol. 117. Springer, Berlin, pp. 110–119.

Demetrovics, J., Katona, G.O.H., 1993. A survey of some combinatorial results concerning functional dependencies in database relations. Ann. Math. Artificial Intelligence 7, 63–82.

Demetrovics, J., Katona, G.O.H., Sali, A., 1992. The characterization of branching dependencies. Discrete Appl. Math. 40, 139–153.

Demetrovics, J., Katona, G.O.H., Sali, A., 1995. Minimal Representations of Branching Dependencies. Acta Sci. Math. (Szeged) 60, 213–223.

Dirac, G.A., 1952. Some theorems on abstract graphs. Proc. London Math. Soc. Ser. 3 (2), 69–81.

Füredi, Z., 1990. Perfect error-correcting databases. Discrete Appl. Math. 28, 171–176.

Gronau, H.-D.O.F., Ganter, B., 1991. On two conjectures of demetrovics, furedi and katona concerning partitions. Discrete Math. 88, 149–155.

Ganter, B., Gronau, H.-D.O.F., Mullin, R.C., 1994. On orthogonal double covers of K_n. Ars Combin. 37, 209–221.

Gronau, H.-D.O.F., Mullin, R.C., Schellenberg, P.J., 1995. On orthogonal double covers of K_n and a conjecture of Chung and West. J. Combin. Des. 3, 213–231.

Gronau, H.-D.O.F., Mullin, R.C., Schellenberg, P.J., Preprint. On orthogonal double covers of $\boldsymbol{K_n}$.

Gronau, H.-D.O.F., Mullin, R.C., Rosa, A. Preprint.

Leck, U., Leck, V., Preprint. There is no ODC with all pages isomorphic to $C_4 \cup C_3 \cup C_3 \cup v$.

Lovász, L., 1979. Toplogical and algebraic methods in graph theory. In: Graph Theory and Related Topics, Proc. Conf. Univ. Waterloo, Ontario 1977, 1–14, Academic Press, NY.

Rausche, A., 1987. On the existence of special block designs. Rostock Math. Kolloq. 35, 13–20.

Sali, A., Sr. Sali, A., submitted. Generalized Dependencies in Relational Databases. Discrete Appl. Math.

Ullman, J.D., 1989. Principles of Database and Knowledge-Base Systems. Computer Sci. Press, New York.