Acta Cybernetica **13** (1998) 431-438.

Generalized Dependencies in Relational Databases [*]

Attila Sali Sr. [†] Attila Sali [‡§]

Abstract

A new type of dependencies in a relational database model introduced in [5] is investigated. If b is an attribute, A is a set of attributes then it is said that b (p,q)-depends on A, in notation $A \xrightarrow{(p,q)} b$, in a database relation r if there are no $q+1$ tuples in r such that they have at most p different values in each column of A, but $q+1$ different values in b. $(1,1)$-dependency is the classical functional dependency. Let $\mathcal{J}(A)$ denote the set $\{b \colon A \xrightarrow{(p,q)} b\}$. The set function $\mathcal{J} \colon 2^{\Omega} \longrightarrow 2^{\Omega}$ becomes a closure if $p = q$. Results on representability of closures by (p,p)-dependencies are presented.

Keywords: relational database, closure, functional dependency, branching dependency, balanced graph

1 Introduction

A relational database system of the scheme $R(A_1, A_2, \ldots, A_n)$ can be considered as a matrix, where the columns correspond to the *attributes* A_i (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*. 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 [1, 2]) 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 are different in b. We denote this by $A \longrightarrow b$.

Functional dependencies have turned out to be very useful. In the present paper we investigate a more general (weaker) dependency, than the functional dependency, which was introduced in [5].

The general concept to be studied is the (p,q)-dependency of [5] with $p = q$.

[*] AMS Subject classification Primary 68P15 68R05 Secondary 05D05

[†] Informatics Centre of Semmelweis University of Medicine Budapest, Kálvária tér 5. H-1089 HUNGARY

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

[§] Corresponding author. e-mail:sali@math-inst.hu Mathematical Institute of HAS, Budapest P.O.B. 127 H-1364 HUNGARY

Definition 1.1 *Let a relational database system of the scheme $R(A_1, A_2, \ldots, A_n)$ be given. Let $A \subseteq \Omega$ and $b \in \Omega$. We say that b* (p,q)-depends *on A if there are no $q+1$ rows (n-tuples) of r such that they contain at most p different values in each column (attribute) of A, but $q+1$ different values in b.*

For a given relation r (or its matrix M) we define a function from the family of subsets of Ω into itself, as follows.

Definition 1.2 *Let M be the matrix of the given relation r. Let us suppose, that $1 \leq p \leq q$. Then the mapping $\mathcal{J}_{Mpq}: 2^{\Omega} \to 2^{\Omega}$ is defined by*

$$\mathcal{J}_{Mpq}(A) = \left\{ b \colon A \xrightarrow{(p,q)} b \right\}.$$

We collect two important properties of the mapping $\mathcal{J}_{Mpq}$ in the following proposition, see [5].

Proposition 1.3 *Let r, Ω, M, p and q as in Definition 1.2. Furthermore, let $A, B \subseteq \Omega$. Then*

$$\begin{array}{ll} (i) & A \subseteq \mathcal{J}_{Mpq}(A) \\ (ii) & A \subseteq B \Longrightarrow \mathcal{J}_{Mpq}(A) \subseteq \mathcal{J}_{Mpq}(B). \end{array}$$

Definition 1.4 *Set functions satisfying (i) and (ii) are called* increasing-monotone functions. *We say that such an increasing-monotone function $\mathcal{N}$ is (p,q)-*representable *if there exists a matrix M such that $\mathcal{N} = \mathcal{J}_{Mpq}$.*

It was also observed in [5] that in the case $p = q$ the set function $\mathcal{J}_{Mpq}$ satisfies a third property

$$(iii) \quad \mathcal{J}_{Mpq}(\mathcal{J}_{Mpq}(A)) = \mathcal{J}_{Mpq}(A) \quad \text{for all} \quad A \subseteq \Omega.$$

Set functions satisfying (i) – (iii) are called *closures* and are widely investigated. In [6] the minimum representation of closures and increasing-monotone functions were investigated. In [7] the connection of the minimum representation and design theoretical constructions was described. Also many open problems were posed.

In the present paper the representability of closures is investigated. In [1] it was proved that functional dependencies and closures are equivalent. However, in [5] it was pointed out, that this no longer holds for general (p,p)-dependencies. It is natural to ask, which closures arise in connection with these weaker dependencies, or putting the question in another way, given a closure $\mathcal{L}$, what are those p's, for which $\mathcal{L}$ is (p,p)- representable. This motivates the following definition. Because only (p,p)-dependencies and (p,p)-representations are considered, in what follows p-dependency and p-representation are written, for the sake of simplicity.

Definition 1.5 *Let $\mathcal{L}$ be a closure on the set Ω. The spectrum* $\mathrm{SP}(\mathcal{L})$ *of $\mathcal{L}$ is defined as follows.*

$$q \in \mathrm{SP}(\mathcal{L}) \iff \mathcal{L} \textit{ is } q - \textit{representable}$$

Note that $\mathrm{SP}(\mathcal{L}) \subseteq \mathbf{N}$.

The following special type of closure plays an important role in the theory.

Definition 1.6 *Let $\mathcal{C}_n^k$ denote the following closure on Ω ($|\Omega| = n$):*

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

The following theorem was proved in [5]

Theorem 1.7

1. $\{1,2\} \subseteq \mathrm{SP}(\mathcal{L})$ *for any closure* $\mathcal{L}$.
2. $\mathrm{SP}(\mathcal{C}_n^2) = \{1,2\}$ *if* $n > 6$.
3. *If* $|\Omega| = n$ *and* $2n-3 \leq N \in \mathrm{SP}(\mathcal{L})$, *then* $\forall q \geq N$ $q \in \mathrm{SP}(\mathcal{L})$

The purpose of the present paper is to extend Theorem 1.7. The extension yields some quite surprising results about the spectra of closures. The interested reader is referred to [3, 4, 6, 7] for further investigations and open problems.

2 Spectra of closures

It was shown in [5] that for a matrix M $b \in \mathcal{J}_{Mpq}(A)$ implies $b \in \mathcal{J}_{Mp-1q-1}(A)$ provided the matrix has at least $q+1$ distinct entries in each of its columns. This may lead to the expectation that the spectrum of a closure is an interval of the integers. In this section we show the quite surprising fact that the spectrum of a closure may contain an arbitrary number of "holes", i.e., it may be far from being an interval.

Let the $m \times n$ matrix M p-represent the closure $\mathcal{L}$ on Ω. A mapping w from the edges of the complete graph K_m to the subsets of Ω can be defined, as follows. The vertices of K_m are identified with the set of rows of M. For an edge $e = \{i,j\}$ of K_m, let $w(e)$ be the set of positions where rows i and j agree. If $A \subset \Omega$ and $b \in \Omega$ such that $b \notin \mathcal{L}(A)$, then there exist $p+1$ rows $r_1, r_2, \ldots, r_{p+1}$ that contain at most p distinct values in columns of A but they are all different in column b. Equivalently, $b \notin \bigcup_{1 \leq i < j \leq p+1} w(\{r_i, r_j\}) \supset A$. The next lemma, which is an equivalent formulation of Theorem 2.12 of [5] is explained by the above observation.

Lemma 2.1 *Let $\mathcal{L}$ be a closure on Ω. $\mathcal{L}$ is p-representable if and only if there exists a mapping $w\colon E(K_m) \longrightarrow 2^{\Omega}$ of the edges of K_m for some m (where $w(e)$ is called the* weight *of edge e) that satisfies the following two properties:*

1. *For any three edges e_1, e_2, e_3 forming a triangle, $w(e_i) \cap w(e_j) \subseteq w(e_k)$ holds for any permutation (i,j,k) of $(1,2,3)$.*
2. *For any $p+1$ vertices of K_m, the union weights of edges spanned by these vertices is closed by $\mathcal{L}$, and every closed set of $\mathcal{L}$ can be obtained as intersections of sets of this type.*

Condition 1. is the necessary and sufficient condition for the existence of a matrix with prescribed edge weights, while condition 2. is that of the p-representation.

First some constructions are presented that show that certain values of p are in $\mathrm{SP}(C_n^k)$. Then we show, that these are all the elements of $\mathrm{SP}(C_n^k)$ provided n is large enough with respect to k. In what follows, edges of K_m of empty weight will be omitted for the sake of simplicity, i.e. weightings of not necessarily complete graphs will be given with the understanding that edges not mentioned have empty weight.

The following result of Rucinski and Vince [8] is needed for constructions. A graph G of $e(G)$ edges and $v(G)$ vertices is called *balanced* if $e(G)/v(G) \geq e(H)/v(H)$ holds for every subgraph H of G. G is called *strongly balanced* if $e(G)/(v(G)-1) \geq e(H)/(v(H)-1)$ holds for every subgraph H of G. A strongly balanced graph is clearly balanced.

Theorem 2.2 ([8]) *There exists a strongly balanced graph with v vertices and e edges if and only if $1 \leq v-1 \leq e \leq \binom{v}{2}$.* □

Lemma 2.3 *C_n^k is p representable if $p \leq k-2$.*

Proof of Lemma 2.3 We may assume without loss of generality that $p > 2$ by Theorem 1.7. Let $k-1 = a\binom{p+1}{2} + b$ where $0 \leq a$ and $0 \leq b < \binom{p+1}{2}$ are integers. Suppose first, that $b \geq p$. Let G be a balanced graph of $p+1$ vertices and b edges provided by Theorem 2.2. For every $k-1$-element subset of Ω we take K_{p+1} so that edges corresponding to edges of G are weighted by $a+1$-element subsets, the remaining ones by a-element subsets, such that the weights of edges are pairwise disjoint sets, and their union is the given $k-1$-element subset of Ω. We claim that the disjoint union of these weighted complete graphs satisfy the conditions of Lemma 2.1.

It is clear that Condition 1. is satisfied, because weights of adjacent edges are pairwise disjoint sets. Also clear is that every $k-1$-element subset of Ω occurs as union of weights of edges spanned by some $p+1$-element subset of vertices. The only thing to check is that larger subsets of Ω do not occur this way. Let us suppose that the $p+1$-element subset of vertices U is the union of sets U_i, $i = 1, 2, \ldots, t$, where U_i's are the intersections of U with the weighted complete graphs. Let $u_i = |U_i|$, furthermore let e_i be the number of edges of the subgraph of balanced graph G spanned by vertices corresponding to U_i. Then $e_i/u_i \leq b/(p+1)$ is satisfied. The cardinality e of the union of the weights of edges spanned by U can bounded from above, as follows:

$$\begin{aligned} e &\leq a\sum_{i=1}^{t}\binom{u_i}{2} + \sum_{i=1}^{t} e_i \\ &\leq a\binom{p+1}{2} + \sum_{i=1}^{t}\frac{e_i}{u_i}u_i \end{aligned}$$

$$\begin{aligned} &\leq a\binom{p+1}{2}+\sum_{i=1}^{t}\frac{b}{p+1}u_i \\ &= a\binom{p+1}{2}+b=k-1 \end{aligned}$$

On the other hand, if $b < p$, then $a > 0$ is satisfied. Let $k-1-p=(a-1)\binom{p+1}{2}+c$. Then $c \geq p$ holds. Let us consider two graphs, G and H, on the same $p+1$ vertices, where G is a balanced graph with c edges, and H is a path (which is clearly balanced). For every $k-1$-element subset of Ω we take K_{p+1} so that edges corresponding to edges of $G \cap H$ are weighted by $a+1$-element subsets, those corresponding to edges of $G \setminus H$ and $H \setminus G$ are weighted by a-element subsets, the remaining ones by $a-1$-element subsets, such that the weights of edges are pairwise disjoint sets, and their union is the given $k-1$-element subset of Ω. That the disjoint union of these weighted complete graphs satisfies the conditions of Lemma 2.1 can be proved by a similar argument to the one above. □

Let us recall that $\lceil x \rceil$ denotes the smallest integer not less than x.

Lemma 2.4 *If*

$$p+1-\left\lceil\frac{p+1}{s}\right\rceil=k-1 \qquad \textit{for } s>1$$

then $p \in \mathrm{SP}(C_n^k)$

Proof of Lemma 2.4 Take $\binom{n}{s-1}$ paths of s vertices whose edges have one element weights so that each $s-1$-element subset occurs as union of elements of a path. Any $p+1$ vertices span a forest that has at least $\lceil\frac{p+1}{s}\rceil$ components, so at most $k-1$ edges. □

Note, that in Lemma 2.4 $s \leq p+1$ may be assumed. Any $s \geq p+1$ gives the same $p=k-1$ case.

In the following, non-representability of closures is discussed. The general pattern is that a minimal (non-decreasable) representing matrix is assumed, then it is shown that it must contain identical rows that clearly contradicts to its minimality. The next lemma shows that the spectrum of C_n^k is finite provided n is large enough.

Lemma 2.5 *Let* $p \geq 2k-1$. *If* $n \geq k^2(k-1)$, *then* C_n^k *is not* p*-representable.*

Proof of Lemma 2.5 Let us assume indirectly that C_n^k is p-represented by the $m \times n$ matrix M, and M is minimal. Immediately follows that every column has to contain at least $p+1$ pairwise distinct entries, otherwise everything would be (p,p)-dependent on that particular column. According to Lemma 2.1 for every $k-1$-element subset A of Ω there exist $p+1$ vertices of K_m such that the union of weights of edges spanned by these vertices is A. Indeed, A is closed in C_n^k, but cannot be an intersection of other closed sets, because the only closed superset of A is Ω. In particular, for every column $a \in \Omega$ there exists and edge e_a of K_m such that $a \in w(e_a)$. Let $e_1, e_2, \ldots, e_k$ correspond to k distinct columns $\{a_1, a_2, \ldots, a_k\}$.

Suppose, that there exists a column b containing pairwise distinct entries in rows covered by edges e_i . The k edges e_i cover at most $2k \leq p+1$ points, or rows, so there exist $p+1$ points $r_1, r_2, \ldots, r_{p+1}$ such that b contains all different entries in these rows, or in other words: $\bigcup_{i=1}^{k} w(e_i) \subseteq \bigcup_{1\leq i<j\leq p+1} w(\{r_i, r_j\}) \not\ni b$. This would imply the existence of a closed set of at least k elements which is not Ω, because b is not in the closure of the set $\{a_1, a_2, \ldots, a_k\}$, a contradiction. Thus, each column b must contain at least a pair of identical entries on the at most $2k$ rows covered by $e_1, e_2, \ldots, e_k$. Now, $n \geq k^2(k-1)$ implies that there are k distinct columns $b_1, b_2, \ldots, b_k$ so that they contain identical elements on the same pair of rows, say r_1, r_2. If there exists a column c containing distinct entries on r_1, r_2, then there exist $p+1$ rows including r_1, r_2 such that c contains all different entries in them, thus a closed set $c \notin B \supseteq \{b_1, b_2, \ldots, b_k\}$ would exist, a contradiction. Consequently, every column must agree on the pair of rows r_1, r_2, i.e., these rows are identical, which contradicts the minimality of M. □

Note that in the above argument the proof of the following proposition is included.

Proposition 2.6 *If the matrix M p-represents C_n^k and minimal subject this condition, then the weight of an edge $w(e)$ is at most $k-1$-element set.*

The next proposition considers another property of a minimal representation.

Proposition 2.7 *Let $p \leq 2k-4$ and $n \geq (k-1)(2k-3)$. Let M p-represent C_n^k and let M be minimal subject to this condition. Then for any $p+1$ rows $r_1, r_2, \ldots, r_{p+1}$,*

$$\left| \bigcup_{1\leq i<j\leq p+1} w(\{r_i, r_j\}) \right| \leq k-1.$$

Proof of Proposition 2.7 According to Lemma 2.1 the union of edge weights of a $p+1$-point complete subgraph is either Ω or its size is at most $k-1$. Suppose indirectly, that there is a sub-K_{p+1} P such that the union of its edge weights is Ω. M p-represents C_n^k, so there is a sub-K_{p+1} Q such that the union of its edge weights is a $k-1$-element subset. By successively shifting vertices from $P \setminus Q$ to Q, sub-K_{p+1} P' and Q' are obtained that $|P' \setminus Q'| = 1$, but the union of edge weights of P' is still Ω, while that of Q' is still a $k-1$-element subset. Let $\{b\} = P' \setminus Q'$. Then the union of edge weights of the p edges between b and $P' \cap Q'$ is of size at least $n-k+1$, thus there exists an edge e amongst them such that $|w(e)| \geq k$, that contradicts to the minimality of M by Proposition 2.6. □

The next proposition allows considering p-representations of special type.

Proposition 2.8 *Let $2\lfloor \frac{p+1}{2} \rfloor \geq k$ and suppose that C_n^k is p-representable. Suppose furthermore that $p \leq 2k-4$ and $n \geq (k-1)(2k-3)$. Then there exists $n' \geq n-k+1$ such that $C_{n'}^k$ is p-represented so that each edge weight is at most one element set.*

Proof of Proposition 2.8 Let M be a matrix p-representing C_n^k that is minimal subject to this condition. A sequence of $\lfloor \frac{p+1}{2} \rfloor$ edges is defined . Let a_1 be the largest size of an edge weight, and let e_1 be an edge of weight of this size. Now suppose, that $e_1, e_2, \ldots, e_i$ are already defined and let a_{i+1} be the maximum of $|w(e) \setminus \cup_{j \leq i} w(e_j)|$ for any edge of K_m and define e_{i+1} to be an edge attaining this maximum. We claim, that $a_{\lfloor \frac{p+1}{2} \rfloor} = 1$. Indeed, otherwise $| \cup_{i=1}^{\lfloor \frac{p+1}{2} \rfloor} w(e_i)| \geq k$ would be, which contradicts to Proposition 2.7, because any $\lfloor \frac{p+1}{2} \rfloor$ edges can be embedded into a sub-K_{p+1}. Let $\Omega_1 = \Omega \setminus \cup_{i=1}^{\lfloor \frac{p+1}{2} \rfloor} w(e_i)$. Then $|\Omega_1| \geq n - k + 1$ and M restricted to the columns of Ω_1 p-represents $C_{|\Omega_1|}^k$ with the property, that each edge of K_m has weight of size at most one. □

The next lemma is a sort of converse of Lemma 2.4.

Lemma 2.9 *Let $n \geq (k-1)(2k-3)$ and suppose that there exists integer $s > 1$ such that*

$$p + 1 - \left\lceil \frac{p+1}{s} \right\rceil < k - 1 < p + 1 - \left\lceil \frac{p+1}{s+1} \right\rceil .$$

Then C_n^k is not *p-representable.*

Proof of Lemma 2.9 Let us suppose indirectly that C_n^k is p-represented by $m \times n$ matrix M. We may assume without loss of generality that each edge weight of K_m is at most one element set according to Proposition 2.8. In the following "number of edges" means "number of edges of pairwise different weights" for the sake of simplicity. If there are more than one edges of the same non-empty weight in a sub-K_{p+1}, then an arbitrary one of them can be picked.

Each $k-1$-element subset of Ω must occur as union of weights of edges of a sub-K_{p+1}. By the condition on k and p, the edges of non-empty but pairwise different weight of such a sub-K_{p+1} span a graph that has a non-tree component or a tree component of size at least $s + 1$. Such a component is called *big*. Let $B_1, B_2, \ldots, B_z$ be big components of different sub-K_{p+1}'s corresponding to pairwise disjoint $k-1$-element subsets. A $p+1$- vertices subgraph is constructed as follows. First, take as many non-tree components as possible, then big tree components, until the number of vertices reaches $p + 1$. Let this graph be H, and suppose the number of vertices of H covered by non-tree components is d, and let $u = p + 1 - d$. Then the number of edges $e(H)$ of H satisfies

$$e(H) \geq d + u + \left\lceil \frac{u}{s+1} \right\rceil \geq p + 1 - \left\lceil \frac{p+1}{s+1} \right\rceil > k - 1,$$

that contradicts to Proposition 2.7. □

The above results can be summarized in the following theorem.

Theorem 2.10 *Let $n \geq k^2(k-1)$. Then the spectrum* $\mathrm{SP}(C_n^k)$ *of C_n^k is determined by the following formula:*

$$\mathrm{SP}(C_n^k) = \{1, 2, \ldots, k-1\} \cup \{p \colon \exists s \in \mathrm{N} \; p + 1 - \left\lceil \frac{p+1}{s} \right\rceil = k - 1\}.$$

□

3 Open Problems

A complete characterization of $\mathrm{SP}(C_n^k)$, was given if k is small with respect to n. However, it was proved in [6] that C_n^n is p- representable for every positive integer p. Thus, the following problem arises naturally.

Open Problem 1 *Determine those k's for which* $\mathrm{SP}(C_n^k) = \mathbf{N}$ *holds!*

The constructions used in proving that certain values of p are in the spectrum of C_n^k usually result in very large matrices. Thus, the next problem is also of interest. For similar results and problems the reader is referred to [6].

Open Problem 2 *Determine the minimum number of rows of a matrix p-representing C_n^k, provided such a representation exists!*

Finally, the general question is still open.

Open Problem 3 *Determine the spectra of other closures!*

Open Problem 3 is in particular interesting for closures arising in different areas of combinatorics, for example for closures coming from matroids.

References

[1] W.W. Armstrong, Dependency Structures of database Relationships, *Information Processing 74* (North Holland, Amsterdam, 1974) 580-583.

[2] E.F. Codd, A Relational Model of Data for Large Shared Data Banks, *Comm. ACM,* **13** (1970) 377-387.

[3] J. Demetrovics, 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.

[4] J. Demetrovics, G.O.H. Katona, A survey of some combinatorial results concerning functional dependencies in database relations, *Annals of Math. and Artificial Intelligence* **7** (1993) 63-82.

[5] J. Demetrovics, G.O.H. Katona and A.Sali, The characterization of branching dependencies, *Discrete Appl. Math.*, **40** (1992), 139-153.

[6] J. Demetrovics, G.O.H. Katona and A. Sali, Minimal Representations of Branching Dependencies, *Acta Sci. Math. (Szeged)*, **60**, (1995) 213-223.

[7] J. Demetrovics, G.O.H. Katona and A.Sali, Design Type Problems Motivated by Database Theory, submitted.

[8] A. Rucinski and A. Vince, Strongly balanced graphs and random graphs, *J. Graph Theory* **10** (1986), 251-264.

Received January, 1998