Minimal representations of branching dependencies

Zdeněk Dvořák, György D. H. Katona and Attila Rödl

Dedicated to Professor László Lovász for his 60th and to Professor György Szekeres for the 70th birthday.

Communicated by P. Frankl

Abstract. A new type of dependence in a relational database model is investigated. If a is an attribute, Aij is a set of values that a has and the set of dependencies that a is said that Aij is independent of A in some x, y, z, Aij is a database of then is an s, t, u, v.e. (t, u, v) such that they have a most a different values in in the i, j, different values in A, (t, u, v) is the defined functional dependency. Let F(t, u, v) denote the set of {Aij, Aij} ⊆ SFLS. Using some characteristics of the set function F(t, u, v) we give conditions for the minimum number of values in a database that is function F(t, u, v).

1. Introduction

A relational database system of the scheme \((A_1, A_2, \ldots, A_n)\) can be considered as a matrix, where the column corresponds to the attribute \(A_i\) (for example, name, date of birth, place of birth etc.) while the rows are the tuples of the relation. That is, a row contains the data of a given individual. Let \(G\) denote the set of attributes (the set of columns of the scheme). Let \(G = \{A_1, A_2, \ldots, A_n\}\).
We shall say that \( f(x, y, z) \) depends on \( x \) if \( x \) is not a constant of the function \( f \), or if it is the only variable of \( f \) for which the derivative of \( f \) with respect to \( x \) is not zero. The notation \( \frac{\partial f}{\partial x} \) denotes the derivative of \( f \) with respect to \( x \), evaluated at the point \( (x, y, z) \).

In the present paper we investigate a more general problem of the dependence of \( f \) on \( x \), which is usually called the functional equation. The functional equation is of the form:

\[
\frac{\partial f}{\partial x} = \gamma(x, y, z)
\]

where \( \gamma(x, y, z) \) is a given function of \( x, y, z \). The problem is to find all functions \( f(x, y, z) \) which satisfy the functional equation for all \( x, y, z \) in some domain \( D \).

To illustrate this, consider the function \( f(x, y, z) = x^2 + y^2 + z^2 \). The functional equation for \( f \) is:

\[
\frac{\partial f}{\partial x} = 2x
\]

which is satisfied by the function \( f(x, y, z) = x^2 + y^2 + z^2 \). However, this is not the only function that satisfies the functional equation, as we will see later.

Another example is the function \( f(x, y, z) = x^2 + y^2 - z^2 \). The functional equation for \( f \) is:

\[
\frac{\partial f}{\partial x} = 2x
\]

which is also satisfied by the function \( f(x, y, z) = x^2 + y^2 - z^2 \). However, this is not the only function that satisfies the functional equation, as we will see later.

In the present paper we investigate a more general problem of the dependence of \( f \) on \( x \), which is usually called the functional equation. The functional equation is of the form:

\[
\frac{\partial f}{\partial x} = \gamma(x, y, z)
\]

where \( \gamma(x, y, z) \) is a given function of \( x, y, z \). The problem is to find all functions \( f(x, y, z) \) which satisfy the functional equation for all \( x, y, z \) in some domain \( D \).

To illustrate this, consider the function \( f(x, y, z) = x^2 + y^2 + z^2 \). The functional equation for \( f \) is:

\[
\frac{\partial f}{\partial x} = 2x
\]

which is satisfied by the function \( f(x, y, z) = x^2 + y^2 + z^2 \). However, this is not the only function that satisfies the functional equation, as we will see later.

Another example is the function \( f(x, y, z) = x^2 + y^2 - z^2 \). The functional equation for \( f \) is:

\[
\frac{\partial f}{\partial x} = 2x
\]

which is also satisfied by the function \( f(x, y, z) = x^2 + y^2 - z^2 \). However, this is not the only function that satisfies the functional equation, as we will see later.
counties in some order of them (if this order, at least). Let us show a little table
(16) $a_i = 5 	imes 7 = 35$ for the estimates of distances of each county,
which averages to their maximum $5, 3, 3, 3$ for some order. Now we can define
attribute $A_4$ by these numbers (16), because the value of $A_4$ gives the
starting county and the value of $A_5$ estimates the second county with the help of the
little table. The same holds for the attributes $A_6$, but we can decrease the number
of possible values even further, if we give a table of numbering the possible
values for each $A_i$, where $i = 4, 5, 6$. The values for attribute $A_4$ are the
same: the same for $A_6$, too. That is, while each element of the original
matrix could be counted by $5$ lines, now for the rank of the little auxiliary table
we could decrease the rank of the element in the second column to $1$ line, and
then of the columns to the third and fourth column is $1$ line.

It is easy to see, that the rank of the matrix in each case when we create
the table of a graph, whose maximum figure is much less than the number of its
them matrix. The result is that we can write the number of possible values, which
are not less than the number of all possible values in much lesser, than the number
of possible non-existing values of the rest or in any case when there hold many (1,4)-dependent, where $k$ is small.

The general context we shall study in the (1,4)-dependency ($0 \leq r \leq 4$
commands).

Definition 1.1. Let a relational database system of the scheme $(R, A_1, A_2, \ldots, A_n)$
be given. Let $\chi \in (R)$ and $\lambda \in (R)$. We say that $\lambda$ is the $\alpha$-representation of $\chi$, if there are no
elements $\lambda$ such that there is different value in each element (attribute) of $A_i$, but $\lambda$ is a different value from $\chi$.

For a given relation $\chi$ (see matrix $M$) we define a function from the handy
of elements of $M$ (in itself) as follows:

Definition 1.2. Let, $\chi$ be the matrix of the given relation $\chi$. Let us suppose, that
$1 \leq \alpha \leq 4$. Then the mapping $\lambda_{\alpha}(\chi) = \{1, \ldots, 4\}$

We collect two important properties of the mapping $\lambda_{\alpha}(\chi)$ in the following
proposition, see [2].
Proposition 1.5. Let $A$, $M$, and $g$ be as above. Furthermore, let $A$, $B$ $\subseteq$ $\mathbb{R}$. Then

1. $A \subseteq B \Rightarrow A \cap g[M] \subseteq B \cap g[M]$.
2. $A \cap B = \emptyset \Rightarrow A \cap g[M] \cap g[B] = \emptyset$.

Definition 1.6. Set functions satisfying (1) and (2) are called increasing-monotone functions. We say that such an increasing-monotone function $f$ is a $(\mathfrak{p}, \mathfrak{q})$-representation if there exists a matrix $M$ such that $A = \text{ran}(M)$.

The aim of this paper is to generalize theorems on categorical representations valid for functional dependencies to $(\mathfrak{p}, \mathfrak{q})$-dependencies. There arise several very interesting combinatorial problems in this context.

2. Maximal representations

In this section, we investigate the maximum number of rows of a matrix $M$ that $(\mathfrak{p}, \mathfrak{q})$ represents a given increasing-monotone set function $f$, provided such representations exist. We always assume that $\mathfrak{p} \leq \mathfrak{q}$.

Definition 2.1. For an increasing-monotone function $f$, let $\text{max}(f)$ denote the maximum number of rows of a matrix that $(\mathfrak{p}, \mathfrak{q})$ represents $f$. If $f$ is not a $(\mathfrak{p}, \mathfrak{q})$-representation, then we put $\text{max}(f) = 0$.

Let us note that, in the case of $\mathfrak{p} \leq \mathfrak{q}$, we have maximality of the latter equality proving.

It is proved in [3] that an increasing-monotone function $f$ with $N(f) = \emptyset$ is $(\mathfrak{p}, \mathfrak{q})$-representable if and only if $N(f) = \emptyset$, $\mathfrak{p} = \mathfrak{q}$, and $\mathfrak{p} = \mathfrak{q}$. From the proof one can easily deduce the following general upper bounds.

Theorem 2.2. Let $f$ be an increasing-monotone function with $N(f) = \emptyset$ and let $\mathfrak{p} \leq \mathfrak{q}$ satisfy one of (10)–(12) above. Then

$$\text{max}(f) \leq \mathfrak{p} + \mathfrak{q} + 12\mathfrak{p}.$$
Definition 2.3. An increasing-relation function, \( f \) satisfying
\[
X(f(x)) = 2^n \quad \forall x \in X
\]
is called a closure.

In the rest of this section we will consider closures only. First, we prove a direct product theorem analogous to the (2.1) case considered in [2].

Definition 2.4. Let \( I \) and \( J \) be closure on ground sets \( X \) and \( Y \), respectively, with \( X \cap Y = \emptyset \). The direct product of \( I \) and \( J \) is the closure on ground set \( X \cup Y \) defined by
\[
I \times J = \{(x,y) \mid x \in X \land y \in Y\}
\]

The direct product plays an important role in the theory and practice of relational database systems.

Theorem 2.5. Let \( I \) and \( J \) be closures on ground sets \( X \) and \( Y \), respectively. Then
\[
\mu(I \times J) = \mu(I) \cdot \mu(J) = \mu(I) \cdot \mu(J)
\]

Proof. This statement is trivial because \( I \) and \( J \) are sets (p, q) representable. Then, we may assume that both \( \mu(I) \) and \( \mu(J) \) are finite. Let \( MI \), be a minimal representation matrix for \( I \) and let \( MJ \) be that for \( J \). We then the following matrices:
\[
M = \left[ \begin{array}{c|c}
Q & W \\
\hline
V & F
\end{array} \right]
\]

where \( Q \) is obtained from \( MI \), by dropping the first row. \( W \) consists of the next row of \( MI \), taken as many times as the number of rows of \( Q \). It consists of the last \( p \) rows of \( MI \), while \( V \) consists of the first \( q \) rows of \( MJ \). \( F \) is obtained from \( MJ \) by dropping the first \( m \) rows. \( V \) consists of the last \( m \) rows of \( MJ \), and \( F \) consists of \( m \) copies each of the first \( m \) rows of \( MJ \), as many copies as the number of rows of \( F \). We then that \( \Phi_{(I,J)} = 1 \times 1 \).

Let us suppose that \( r \in Q \cap F \) for some \( r \in X \cap Y \). We may assume without loss of generality that \( p < q \). This implies that \( r \) encounters \( r \) times the number of \( r \) times in union of \( I \) but not all closures in

\[
\mu(I \times J) = \mu(I) \cdot \mu(J) = \mu(I) \cdot \mu(J)
\]
Thus, they have at most $p$ different values in their extensions to $G$, which implies $y \notin \text{dom}(\pi_G)$. On the other hand, let $y \notin \delta(G)$ and let $y_1, \ldots, y_m$ be such that they contain at most $p$ different values in columns of $A$. Assume again that $y \notin \text{dom}(\pi_G)$. If there are two among $y_1, \ldots, y_m$ with that both of them are in the left row of $G$, then these two must agree on $y$ and they cannot contain a $p$ different values in $y$. Hence, if $y \in \text{dom}(\pi_G)$, then $y$ must have at most $p$ different values in all rows that have some values in $y$ and $y \in \text{dom}(\pi_G)$. Then, $y \notin \text{dom}(\pi_G)$ implies that $y$ must have at most $p$ different values in all rows.

Next, we will calculate certain $\delta_{\pi}(G)$ values for the following well-stated clauses.

**Definition 2.4.** Let $\mathcal{G}$ denote the following clauses on $D$:

$$\mathcal{G}(x) = \begin{cases} 
\gamma & \text{if } x \in \emptyset \\
\emptyset & \text{otherwise} 
\end{cases}$$

First we state a general lemma.

**Lemma 3.1.** Let us assume that $\mathcal{G}(x)$ is $\emptyset$, $x \in \emptyset$, representing $\mathcal{G}$. For any $k \in \mathcal{G}$ column, if $D$ then exists a set of all possible $y$ values such that match in $x \in \emptyset$, and only those $y$ values belong to $\emptyset$. Then $\mathcal{G}(x) \notin \emptyset$ if $\emptyset$. Indeed, if the exist $y \in \emptyset$ and $y \notin \emptyset$, then $\mathcal{G}(x) \notin \emptyset$.
Proposition 2.8.  
\[ n_{1/2}(G) = e \geq 1. \]

Proof. The inequality \( n_{1/2}(G) \geq e \geq 1 \) follows from Lemma 2.7. The inequality in the other direction and the representability proved by the following construction:

\[
\begin{array}{ccccccc}
1 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

Theorem 2.8.  If \( k \geq 3 \), then
\[ n_{1/2}(G) = 3k. \]

Proof. We construct a matrix \( M \) of size \( 3 \times 2^k \) representing \( G \) as follows. Note that \( A_{i,j} \) will contain \( 1 \) in columns \( 1 \) and \( 1 \) and \( 0 \) in other columns, respectively. If \( 0 \leq i \leq 2 \) has more than one element, then there exist at least two rows of \( A \) containing at least two different columns of \( A \). By applying \( R^1 \) for \( G \) on the other hand, for any pair of one internal subset \( [1] \) and \( [2] \) of \( A \) rows \( 0 \) and \( 1 \) shows that \( (1) \subseteq (2) \) and \( (2) \subseteq (1) \).

In order to prove that we need at least \( 3k \) rows to \( G \), we represent \( G \), let us assume that \( 0 \leq i \leq 2 \) has a representing matrix of maximum number of rows. As in the proof of Proposition 2.8., we consider distinct \( e \) columns for any subset of \( A \) and we consider \( e \) different \( e \) columns in that column. That is, for every column, there is a pair of rows that agree in that column. We claim that there are at least \( k \) different columns.

Case (i): There exist columns \( i \) and \( j \) such that the same pair of rows we chose above, say \( r \) and \( s \). Then \( (i) \subseteq (j) \) implies that the two rows \( r \) and \( s \) are connected through an arc, which contradicts the assumption of \( G \). Indeed, if \( r \) and \( s \) connected through an arc, then there exists a third different value in column \( e \) by \( (1) \subseteq (2) \). But any column contains at least \( 3k \) different elements, the entries \( (r,s) \) would show \( (i) \subseteq (j) \), a contradiction.

Case (ii): There exist columns \( i \) and \( j \) such that the same pair of rows we chose above, say \( r \) and \( s \). Then \( (i) \subseteq (j) \) implies that the two rows \( r \) and \( s \) are connected through an arc, which contradicts the assumption of \( G \). Indeed, if \( r \) and \( s \) connected through an arc, then there exists a third different value in column \( e \) by \( (1) \subseteq (2) \). But any column contains at least \( 3k \) different elements, the entries \( (r,s) \) would show \( (i) \subseteq (j) \), a contradiction.
Note three cases. Then again (3.17) implies that all other column vectors at most two different values in three columns in \( n > 3 \) we get back in Case


case 1 when we go back to Case

by previous chapter \( n > 3 \). Let us assume that \( n_{0}(2) = 4 \) and \( n_{0}(2) = 5 \). The lower bounds follow from Lemmas 2.8 either directly or by some argument. The upper bounds are given by the following constructions

\[
\begin{align*}
\sigma(n_{0}(2)) & = 4, \\
\sigma(n_{0}(2)) & = 5,
\end{align*}
\]

The next theorem is an interesting application of a theorem of Lovász.

**Theorem 3.18.**

\[
\sigma(n_{0}(2)) = \min \left( \begin{array}{c}
\frac{1}{2} \left( n + 1 \right) + 1
\end{array} \right)
\]

**Proof.** Let us first prove a notion bound by a construction. Assume that \( \left( \sigma \left( n_{0}(2) \right) \right) \geq n \). Construct a matrix \( A \) of \( n \) rows and \( n \) columns as follows. The first row consists of all \( 1 \)'s. Then assign a distinct positive integer of the remaining \( n-1 \) rows to every column and put the constant \( 1, 2, \ldots, n \) in these, respectively. The remaining rows are \( B \). We show the case \( p = 2, n = 6 \) and \( s = 5 \):

\[
\begin{array}{cccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1
\end{array}
\]

Let us now assume that \( f \leq G \). Then there is a \( p \)-distinct entries in columns \( f \) in one of and a \( p \)-distinct assigned to \( h \), while \( G \) occur at least twice in these row and columns of \( A \). This means that \( G \neq \emptyset \) and, i.e. every subset \( f \) of \( G \) is filled under \( f_{(.)} \).
On the other hand, let } \{ n \} [p] (p) \subseteq \mathbb{Q} \text{ be a set of some } \forall n \in \mathbb{N} \text{. Every } n \in \mathbb{N} \text{ but not in each } \mathbb{Z} \text{, then there exist } n \in \mathbb{N} \text{ for any column } x \in \mathbb{Z} \text{ such that there exist } n \in \mathbb{N} \text{ but not at most } n \text{ distinct rows in each of the remaining columns. Thus, } n \text{ is } 1 \text{ and the rows are all different. Let } \mathcal{A} \text{ denote the set belonging to column } x \text{. We may assume without loss of generality that for every } i \text{ the columns } \mathcal{A}_i \text{ are standing in } 1 \text{ and } \mathcal{B}_i \text{ in } 0 \text{. How to do this?} \text{ The columns of } \mathcal{A} \text{ which is not between } 1 \text{ and } x \text{ (not in column } x \text{) are } 0 \text{. It is easy to see that the following matrix will } n \text{ of } [n] \text{ of some } x \text{.}

Let us consider the case } n \geq 1 \text{. We construct a bipartite graph } \Gamma \text{ with } \mathcal{A} \text{ and } \mathcal{B} \text{ as vertices and } E \text{ as edges.}

Let } n \text{ be the number of vertices in } \mathcal{A} \text{. If } n \text{ is even, } \mathcal{A}_i \text{ is a vertex with } n \text{ edges, as Lemma } 3 \text{ holds true.}

\begin{align*}
\text{For any positive integer } n \geq 1 \text{,}
\end{align*}

\begin{align*}
r_{n}(2) = \frac{n^2 - 1}{2} < 2^{\frac{n-1}{2}} - 1.
\end{align*}

Lemma 2.1 gives a lower bound of the order of magnitude } n^{2/3} \text{, since we do not know yet the right order for } n^{2}(2).

Proof of Lemma 3.1. We give a construction that proves the upper bound. Let } \mathcal{A} \text{ be a set of } n \text{ rows and } \mathcal{B} \text{ be a set of } n \text{ columns defined as follows. For each column } x \text{, we assign } x \text{ pairs of rows that will contain the number } 1 \text{, } \ldots \text{, or, respectively, while the other columns of the rows are all distinct and different from } 1 \text{, } \ldots \text{.}

Let us take the rows in two parts, } 1 \text{, } \ldots \text{, } x \text{ for the first } x - 1 \text{ rows and } x \text{, } \ldots \text{, } n \text{ for the last } n - 1 \text{ rows. (Thus we have one } x \text{ of } 1 \text{ and one } x \text{ of } 0 \text{ and so on.)}

The pairs we assign will consist of a } (1) \text{ row and a } (0) \text{ row. (Thus, we assign...}
We consider the following set \( \{1, 2, 3, \ldots, n \} \) and a sequence of \( n \) elements \( (a_1, a_2, \ldots, a_n) \). We introduce a new relation \( R \) on the set \( \{1, 2, 3, \ldots, n\} \) defined as follows: for any two elements \( a_i \) and \( a_j \), \( (a_i, a_j) \in R \) if and only if there exists a number \( k \) such that \( a_i \leq k < a_j \). We are interested in finding the number of pairs \( (a_i, a_j) \) such that \( (a_i, a_j) \in R \) and \( a_i \neq a_j \).

We prove the following theorem:

**Theorem:** The number of pairs \( (a_i, a_j) \) such that \( (a_i, a_j) \in R \) and \( a_i \neq a_j \) is given by \( n(n-1) \).

Proof: We use the principle of inclusion-exclusion. Let \( A \) be the set of all pairs \( (a_i, a_j) \) such that \( a_i \neq a_j \) and \( (a_i, a_j) \in R \). We compute the size of \( A \) by counting the number of pairs in each case.

1. Count all possible pairs: \( \binom{n}{2} = \frac{n(n-1)}{2} \) pairs.
2. Subtract the pairs where \( a_i = a_j \): \( \frac{n(n-1)}{2} \) pairs.
3. Add back the pairs where \( a_i \) and \( a_j \) are consecutive: \( n \) pairs.

Thus, the total number of pairs is \( \frac{n(n-1)}{2} - \frac{n(n-1)}{2} + n = n(n-1) \).

We conclude that the number of pairs \( (a_i, a_j) \) such that \( (a_i, a_j) \in R \) and \( a_i \neq a_j \) is \( n(n-1) \).