Matrix representations of participation constraints

Sven Hartmann*‡ and Uwe Leck*

March 2003

Abstract

We discuss the existence of matrix representations for generalized and minimum participation constraints which are frequently used in database design and conceptual modelling. Matrix representations, also known as Armstrong relations, have been studied in literature e.g. for functional dependencies and play an important role in example-based design and for the implication problem of database constraints. The major tool to achieve the results in this paper is a theorem of Hajnal and Szemerédi on the occurrence of clique graphs in a given graph.

Keywords: participation constraints, matrix representation, Armstrong database, axiomatization, graph packing, clique graph

*Fachbereich Mathematik, Universität Rostock, Rostock, Germany
‡Information Science Research Centre, Massey University, Palmerston North, New Zealand

1 Introduction

Informally, a database relation may be considered as a matrix, where every column contains the data of the same sort and every row contains the data of some object. This approach is very similar to the two-dimensional tables that humans have used to keep track of information for centuries. As an example consider a relation schema (Teacher, Course, Weekday) and the database relation in Figure 1 containing information on classes taught at a university.

Teacher	Course	Weekday
Mary	Java	Mo
John	C++	Tu
John	Delphi	Tu
Mary	Java	We
Mary	Java	Fr

Figure 1: A database relation containing information on classes to be taught.

Often the data stored in a database relation are not independent from each other. In the example above any two classes on the same course are given by the same teacher. Let Ω denote the set of columns, and let X, Y be non-empty subsets of Ω. Then Y functionally depends on X if any two rows coinciding in the columns of X are also equal in the columns of Y. Further, entries do not occur arbitrarily often. In the example above every teacher gives between 2 and 3 classes, and for every course there is at most one class per weekday. Given some entry a in the matrix, its degree $\deg(A, a)$ counts how often this entry occurs in the column $A \in \Omega$. Analyzing these degrees provides lower and upper bounds on the number rows that coincide in column A.

Some of the dependencies discussed above may hold by accident. When the database relation is updated they could well be violated. Others, however, we wish to hold forever, no matter of how the database relation is modified. They reflect the semantics of the real world situation captured by the database. The notion of a database relation itself provides only syntax but does not carry any semantics of the data. Therefore, semantic integrity constraints are used to specify the rules which data have to satisfy in order to reflect the properties of the represented objects in the modelled real world situation. When designing a database system, integrity constraints have been proven useful in ensuring databases with semantically desirable properties, in preventing update anomalies, and in allowing the application of efficient methods for storing, accessing and querying data. Consequently, various classes of integrity constraints have been defined and studied for databases with functional dependencies, multi-valued dependencies and inclusion dependencies being the most prominent examples. In addition, properties such as axiomatization and implication have been studied for these constraints. For details we refer e.g. to [11, 13].

In the present paper we study participation constraints which gained much atten-

tion in the database design community, but also in conceptual modelling and in knowledge representation.

2 Preliminaries

Let R be a matrix with n columns and s rows and such that no two rows are identical. Let $\Omega = \{C_1, \ldots, C_n\}$ be the n-element set of columns. Further, let $range_i$ contain all entries of R in column C_i. In the context of the relational database model (RDM), the columns are called *attributes*, the elements in $range_i$ are called the values of attribute C_i, the sequence $(C_1, \ldots, C_n)$ is called a *relation schema*, and the matrix R is called a *database relation* over Ω. The rows of R are *tuples* from the cartesian product $range_1 \times \cdots \times range_n$, and each tuple contains the data of one object.

2.1 Participation constraints

Within this paper, we are mainly concerned with participation constraints. A *participation constraint* is an expression $card^{part}(C_i) = b$ with $b \in \mathbb{N}^\infty$ and $C_i \in \Omega$. This constraint holds in the database relation R every $C_i \in range_i$ appears at most b times in column C_i. For example, the participation constraint $card^{part}(Teacher) = 3$ tells us that every teacher gives at most three classes. We call a participation constraint *finite* if b is finite.

Participation constraints may be easily extended to sets of columns. A *generalized participation constraint* is an expression $card^{part}(X) = b$ with $b \in \mathbb{N}^\infty$ and $\emptyset \neq X \subseteq \Omega$. This constraint holds in a database relation R if there are at most b rows which coincide in each of the columns $C_i \in X$. For example, the generalized participation constraint $card^{part}(\{Course, Weekday\}) = 1$ tells us that every course is taught at most once per weekday. Clearly, a participation constraint $card^{part}(C_i) = b$ corresponds to a generalized participation constraint $card^{part}(R, X) = b$ with $X = \{C_i\}$. Generalized participation constraints with X containing all but one of the columns are better known as look-across constraints or Chen-style cardinality constraints, and have been widely used in database design, e.g. in entity-relationship modelling [13].

In many applications, one is not only interested in upper bounds on the number of occurrences of values but also in lower bounds. A *minimum participation constraint* is an expression $card^{min}(C_i) = a$ with $a \in \mathbb{N}^\infty$ and $C_i \in \Omega$. This constraint holds in the database relation R every $C_i \in range_i$ appears at least a times in column C_i. For example, the participation constraint $card^{min}(Teacher) = 2$ tells us that every teacher gives at least two classes.

2.2 The implication problem and closed constraint sets

The constraints satisfied by a database relation are usually not independent. A single constraint σ *follows* from a constraint set Σ if σ holds in every database relation R which satisfies Σ. We also say that Σ *implies* σ. Two constraint sets Σ and Σ' are *equivalent* if every constraint in Σ' follows from Σ and vice versa.

For a fixed class $\mathcal{Z}$ of constraints, the *implication problem* for this class $\mathcal{Z}$ reads as follows: Given a constraint set $\Sigma \subseteq \mathcal{Z}$ and a single constraint $\sigma \in \mathcal{Z}$, we want to know whether σ follows from Σ. The emergence of the implication problem in database theory is discussed e.g. in [11, 12]. A constraint set Σ is *$\mathcal{Z}$-closed* if it contains every constraint $\sigma \in \mathcal{Z}$ which follows from Σ. Special attention is devoted to the determination of closed constraint sets. Clearly, Σ implies $\sigma \in \mathcal{Z}$ if and only if σ is in the $\mathcal{Z}$-closure of Σ. Thus the characterization of closed sets in a constraint class $\mathcal{Z}$ completely solves the implication problem for this class.

In the present paper, we are interested in the joint class $\mathcal{P}$ of generalized participation constraints and minimum participation constraints.

3 Matrix representations

Given a database relation R it is often a straightforward task to extract the set $\Sigma(R) \subseteq \mathcal{Z}$ of all constraints from $\mathcal{Z}$ satisfied by R. Clearly, Σ_R must be $\mathcal{Z}$-closed. Conversely, given a constraint set $\Sigma \subseteq \mathcal{Z}$ it is natural to ask whether there is a database relation R such that $\Sigma(R)$ is just the $\mathcal{Z}$-closure of Σ. In this case, R is said to *represent* the constraint set Σ under consideration or to be a *$\mathcal{Z}$-Armstrong relation* for Σ. In this case, R satisfies exactly the logical consequences of Σ among all the constraints in $\mathcal{Z}$.

In view of this property, matrix representations are excellent tools in example-based database design. Armstrong relations satisfy exactly the conditions specified by the database designer. This makes them good examples to represent the real world situation captured by the database. Further, they help the designer to recognize omissions and mistakes in the design. Actually, a major problem that has been noted with the use of automated design tools is to get all necessary design information from the designer into the tool.

Matrix representations have been first studied for functional dependencies. A *functional dependency* is a statement $X \to Y$ where both X and Y are non-empty subsets of Ω. This constraint holds in the database relation R if any two rows coinciding in the columns of X also coincide in the columns of Y. Armstrong [1] observed that closed sets of functional dependencies correspond to closure operations on the set Ω. He proved that every closed set of functional dependencies admits a matrix representation. In [7] Demetrovics and Gyepesi proved that in the worst case the minimum size s of an Armstrong relation for a set of functional dependencies satisfies

the inequality

$$\frac{1}{n^2}\binom{n}{\lfloor n/2 \rfloor} < s \leq (1 + \tfrac{c}{\sqrt{n}})\binom{n}{\lfloor n/2 \rfloor},$$

for some suitable constant c. A functional dependency $X \rightarrow \Omega$ is, in particular, called a *key dependency* and X is said to be a *key*. Note that key dependencies are special kinds of generalized participation constraints, namely those ones with $b = 1$. Demetrovics observed that the set of minimal keys is always a Sperner family over the set Ω, that is, minimal keys are mutually inclusion-free. Again, every closed set of key dependencies admits a matrix representation. Demetrovics and Gyepesi [7] proved that in the worst case the minimum size s of an Armstrong relation for a set of key dependencies satisfies the inequality

$$\frac{1}{n^2}\binom{n}{\lfloor n/2 \rfloor} < s \leq 1 + \binom{n}{\lfloor n/2 \rfloor}.$$

Since then, matrix representations for functional dependencies have been widely studied in the literature [2, 3, 4, 8]. For a survey on similar results for other constraints such as multi-valued dependencies, inclusion dependencies or branching dependencies, see e.g. [11, 12, 13].

Unfortunately, matrix representations are not always possible. Let $n \geq 2$ and consider the empty constraint set Σ which is clearly satisfied by every database relation R over $\Omega = \{C_1, \ldots, C_n\}$. Hence, Σ does not imply any participation constraint $card^{part}(C_1) = b$ with finite b. Conversely, however, each database relation R of size s satisfies the participation constraint $card^{part}(C_1) = s$, which is not a consequence of Σ. In order to be represented by some database relation, Σ must at least imply some finite participation constraint for every $C_i \in \Omega$.

4 Inference rules

The latter observation again rises the implication problem. Of course, one will not inspect all possible database relations in order to decide the implications of a given constraint set. Rather, we are interested in inference rules which help to decide this question. An *inference rule* is an expression $\frac{\Sigma'}{\sigma}\gamma$ where Σ' is a subset of Σ, and γ states some condition on Σ' which has to be satisfied if we want to apply this rule. If Σ contains a subset Σ' satisfying the condition γ, then σ may be *derived* from Σ due to that inference rule. An inference rule is *sound* if Σ implies every constraint σ which may be derived from Σ due to that rule.

We are interested in inference rules which completely describe all the implications of a given constraint set Σ. A *rule system* $\mathcal{R}$ is a set of inference rules. The most prominent example of such a rule system is the Armstrong system for functional dependencies [1]. A set Σ is *syntactically closed* with respect to $\mathcal{R}$ if it contains every constraint σ which may be derived from Σ due to some rule in $\mathcal{R}$. The general

problem is to find a rule system $\mathcal{R}$ for the constraint class $\mathcal{Z}$ such that a given set $\Sigma \subseteq \mathcal{Z}$ is $\mathcal{Z}$-closed if and only if it is syntactically closed w.r.t. $\mathcal{R}$. Such a rule system is said to be *sound and complete* for the implication of $\mathcal{Z}$. The Armstrong system for functional dependencies is the most prominent example of a sound and complete rule system.

Let $C_i, C_j \in \Omega$, let X, Y be non-empty subsets of Ω, and let $a, a', b, b' \in \mathbb{N}^\infty$. For the class $\mathcal{P}$ of generalized and minimum participation constraints the following inference rules are clearly sound:

$$\overline{card^{part}(X) = \infty}, \quad \overline{card^{part}(\Omega) = 1}, \quad \overline{card^{min}(C_i) = 1},$$

$$\frac{card^{part}(X) = b}{card^{part}(Y) = b}\ X \subset Y, \quad \frac{card^{part}(X) = b}{card^{part}(X) = b'}\ b < b', \quad \frac{card^{min}(C_i) = a}{card^{min}(C_i) = a'}\ a > a',$$

$$\frac{card^{part}(C_i) = b, card^{min}(C_i) = a}{card^{part}(C_i) = 0}\ a > b,$$

$$\frac{card^{part}(C_i) = 0}{card^{part}(C_j) = 0}, \quad \frac{card^{part}(C_i) = 0}{card^{min}(C_i) = \infty}.$$

Note that the last three rules describe situations where Σ is only satisfied by the empty database relation. We call such a constraint set *conflicting*. Matrix representations will help us to verify that the rule system above is in fact complete, that is, provides a characterization of closed sets of generalized and minimum participation constraints.

5 Representation graphs

In the sequel we make use of a nice graph-theoretic analogue of matrix representations. For every column C_i, we introduce its *representation graph* $\mathcal{G}_i$ whose vertices are the rows of R, and where two vertices r and r' are connected by an edge just when the rows r and r' coincide in column C_i.

By $\mathcal{K}_k$ we denote the *complete graph* on k vertices. A *clique* of size k in a graph $\mathcal{G}$ is a maximal complete subgraph with k vertices in $\mathcal{G}$. A *clique graph* is a graph where every connected component is a complete graph. Obviously the representation graph $\mathcal{G}_i$ is a clique graph where each clique corresponds to exactly one value of the attribute C_i. Conversely, suppose we are given a collection $\mathcal{O}$ of subgraphs $\mathcal{G}_i$ of the complete graph $\mathcal{K}_s$ such that each of them is a clique graph. Then it is easy to construct a database relation R of size s whose representation graphs are just the given graphs $\mathcal{G}_i$.

For any non-empty subset $X \subseteq \Omega$, let $\mathcal{G}_X$ denote the intersection of the representation graphs $\mathcal{G}_i$ with $C_i \in X$. This intersection is again a clique graph. The following observation is straightforward.

Proposition 1. *A database relation R satisfies the generalized participation constraint $card^{part}(X) = b$ if and only if the intersection graph $\mathcal{G}_X$ has maximum clique size at most b. A database relation R satisfies the minimum participation constraint $card^{min}(C_i) = a$ if and only if the representation graph $\mathcal{G}_i$ has minimum clique size at least a.*

This explains our interest in collections of clique graphs whose intersections have prescribed clique sizes. In the remainder of this section we assemble a number of lemmata ensuring the existence of such collections. The final lemma in this series will then turn out to be the major tool to establish matrix representations for generalized and minimum participation constraints. In order to prove this final lemma we are going to apply a theorem of Hajnal and Szemerédi [10]. By $\mu\mathcal{K}_k$ we denote the clique graph consisting of μ vertex-disjoint copies of $\mathcal{K}_k$.

Theorem 2 (Hajnal and Szemerédi). *Let $\mathcal{H}$ be a graph with $m = \mu k$ vertices and minimum valency $\delta(\mathcal{H}) \geq m - \mu$. Then $\mathcal{H}$ has a subgraph isomorphic to the clique graph $\mu\mathcal{K}_k$.*

This deep result was first conjectured by Erdős [9] and gives a necessary condition on the occurrence of clique graphs as subgraphs in a given graph $\mathcal{H}$. For a detailed discussion, we refer to Bollobás [5].

Throughout, suppose we are given positive integers k_X for every non-empty subset $X \subseteq \Omega$ such that $k_X \geq k_Y$ whenever $X \subseteq Y$. For simplicity, we write k_j instead of k_{C_j} for every $C_j \in \Omega$.

Lemma 3. *Let $s = \sum_{\emptyset \neq X \subseteq \Omega} k_X$. Then there is a collection of spanning subgraphs $\mathcal{G}_1, \ldots, \mathcal{G}_n$ of $\mathcal{K}_s$ satisfying the following conditions:*

(i) For every j with $1 \leq j \leq n$, the subgraph G_j is a clique graph.

(ii) For every non-empty subset $X \subseteq \Omega$, the intersection graph $\mathcal{G}_X$ has maximum clique size k_X.

Proof. To begin with, we partition the vertex set of $\mathcal{K}_s$ into subsets V_Z where V_Z consists of k_Z vertices and Z runs through all non-empty subsets $Z \subseteq \Omega$. Then, for every $j = 1, \ldots, n$, we choose $\mathcal{G}_j$ to be the clique graph whose components are the complete graphs on the sets V_Z with $j \in Z$ together with the isolated vertices contained in the sets V_Z with $j \notin Z$. Each G_j satisfies the first condition as $k_Z \leq k_j$ holds whenever $j \in Z \subseteq \Omega$. Given some non-empty subset $X \subseteq \Omega$, the intersection graph $\mathcal{G}_X$ is just the clique graph whose non-singleton components are complete graphs on the sets V_Z with $X \subseteq Z$. The inequality $k_Z \leq k_X$ for $X \subseteq Z$ proves $\mathcal{G}_X$ to be of maximum clique size k_X as claimed. □

Lemma 4. *Let $s = \sum_{\emptyset \neq X \subseteq \Omega} \left(k_X - |X| k_X + \sum_{j \in X} k_j \right)$. Then there is a collection of spanning subgraphs $\mathcal{G}_1, \ldots, \mathcal{G}_n$ of $\mathcal{K}_s$ satisfying the following conditions:*

(i) For every j with $1 \leq j \leq n$, the subgraph G_j is a clique graph such that each of its cliques is of size 1 or k_j.

(ii) For every non-empty subset $X \subseteq \Omega$, the intersection graph $\mathcal{G}_X$ has maximum clique size k_X.

Proof. First, we select a subset V' of size $s' = \sum_{\emptyset \neq X \subseteq \Omega} k_X$ among the vertices of $\mathcal{K}_s$. For these vertices we proceed as in the preceding lemma which gives us a collection $\mathcal{O}'$ of graphs $\mathcal{G}'_j$ with vertex set V' satisfying the conditions in the preceding lemma. The remaining vertices not in V' should be partitioned into subsets $V_{j,Z}$ where $V_{j,Z}$ consists of $k_j - k_Z$ vertices, and j, Z runs through all pairs j, Z with $1 \leq j \leq n$ and $j \in Z \subseteq \Omega$. Next, for every $j = 1, \ldots, n$, we have to extend the subgraph $\mathcal{G}'_j$ on vertex set V' to a spanning subgraph $\mathcal{G}_j$ containing all vertices of $\mathcal{K}_s$. For that, we extend the component with vertex set V'_Z in $\mathcal{G}'_j$ to a complete graph on the vertex set $V'_Z \cup V_{j,Z}$ where Z runs through all subsets $Z \subseteq \Omega$ containing j. Due to our choice of the vertex sets V'_Z and $V_{j,Z}$, all the cliques in the resulting clique graph $\mathcal{G}_j$ are of size 1 or k_j as desired. The second condition immediately follows from our construction and the preceding lemma. Note that the intersection graph $\mathcal{G}_X$ is just the intersection graph $\mathcal{G}'_X$ on the vertex set V' augmented by a number of isolated vertices. □

Choose λ to be a positive integer such that $\lambda \prod_{j=1}^{n} k_j \geq \sum_{\emptyset \neq X \subseteq \Omega} k_X$.

Lemma 5. *Let $s = (\lambda+1) \prod_{j=1}^{n} k_j$. Then there is a collection of spanning subgraphs $\mathcal{G}_1, \ldots, \mathcal{G}_n$ of $\mathcal{K}_s$ satisfying the following conditions:*

(i) For every j with $1 \leq j \leq n$, the subgraph $\mathcal{G}_j$ is isomorphic to the clique graph $k_j \mathcal{K}_{s/k_j}$.

(ii) For every non-empty subset $X \subseteq \Omega$, the intersection graph $\mathcal{G}_X$ has maximum clique size k_X.

Proof. Let V denote the vertex set of $\mathcal{K}_s$. First, we select a subset $V' \subseteq V$ of size $s' = \sum_{\emptyset \neq X \subseteq \Omega} \left(k_X - |X| k_X + \sum_{j \in X} k_j \right)$. For these vertices we proceed as in the preceding lemma which gives us a collection $\mathcal{O}'$ of graphs $\mathcal{G}'_j$ with vertex set V' satisfying the conditions in the preceding lemma. Now, for every $j = 1, \ldots, n$, we have to extend the subgraph $\mathcal{G}'_j$ on vertex set V' to a spanning subgraph $\mathcal{G}_j$ on vertex set V.

Assume we have already constructed suitable subgraphs $\mathcal{G}_i$ for $i < j$, and are now going to construct $\mathcal{G}_j$. Let V'' consist of all the isolated vertices in $\mathcal{G}'_j$ and all the vertices in $V - V'$. Put

$$\mu = (\lambda + 1) \prod_{i \neq j} k_i - |\{Z \subseteq \Omega : j \in Z\}|.$$

It is an easy calculation to see that V'' is just of size μk_j. The subgraph of $\mathcal{G}'_j$ induced by $V - V''$ is clearly isomorphic to the clique graph $((\lambda+1)\prod_{i\neq j} k_i - \mu)\mathcal{K}_{k_j}$. Hence, to ensure condition (i), it essentially remains to arrange the vertices in V'' to cliques of size k_j each. For that, however, we may use neither the edges in the subgraphs $\mathcal{G}_i$, $i < j$, nor the edges in the subgraphs $\mathcal{G}'_i$, $i \geq j$. Let $\mathcal{H}$ be the graph on vertex set V containing all the remaining, i.e. permitted edges for $\mathcal{G}_j$. Further, let $\mathcal{H}''$ be the subgraph of $\mathcal{H}$ induced by the vertex set V''. Every vertex in $\mathcal{H}''$ has valency at least

$$\delta(\mathcal{H}'') \geq |V''| - 1 - \sum_{i\neq j}(k_i - 1) = \mu k_j - 1 - \sum_{i\neq j}(k_i - 1).$$

This allows us to apply the Theorem of Hajnal and Szemerédi which verifies that $\mathcal{H}''$ contains a subgraph with vertex set V'' which is isomorphic to $\mu\mathcal{K}_{k_j}$. Together with the copy of $((\lambda+1)\prod_{i\neq j} k_i - \mu)\mathcal{K}_{k_j}$ with vertex set $V - V''$ this gives us the subgraph $\mathcal{G}_j$ satisfying condition (i) as desired. Again, condition (ii) immediately follows from our construction and the preceding lemma. Note that the intersection graph $\mathcal{G}_X$ is just the intersection graph $\mathcal{G}'_X$ on the vertex set V' augmented by a number of isolated vertices. □

6 Main results

We are now ready to state our results on matrix representations of generalized and minimum participation constraints. As a consequence we also obtain a characterization of closed sets of these constraints.

Theorem 6. *Let Σ be a set of generalized and minimum participation constraints containing some finite participation constraint for every $C_i \in \Omega$. Then Σ may be represented by a database relation R.*

Proof. Let Σ^+ contain Σ and all the consequences of Σ derived by applying the rules in Section 4. If Σ^+ contains a constraint $card^{part}(C_i) = 0$ for some (and thus for all) $C_i \in \Omega$, the empty database relation represents Σ. Otherwise, put $b_X = \min\{b : card^{part}(X) = b \text{ is in } \Sigma^+\}$ for every non-empty $X \subseteq \Omega$, and $a_i = \max\{a : card^{min}(C_i) = a \text{ is in } \Sigma^+\}$. For short, we again write b_i instead of b_{C_i}. By hypothesis, all these values are finite. Two applications of the final lemma in the preceding section will provide representation graphs $\mathcal{G}_1, \ldots, \mathcal{G}_n$ which yield the claimed database relation R. First, we choose $k_X = b_X$ for every non-empty $X \subseteq \Omega$. This gives us a collection $\mathcal{O}^1$ of clique graphs $\mathcal{G}^1_1, \ldots, \mathcal{G}^1_n$. Next, we choose $k_i = a_i$ for every $C_i \in \Omega$ and $k_X = 1$ for every subset $X \subset \Omega$ of size at least 2. This gives us a collection $\mathcal{O}^2$ of clique graphs $\mathcal{G}^2_1, \ldots, \mathcal{G}^2_n$. Afterwards, for every $C_i \in \Omega$, we take $\mathcal{G}_i$ as the vertex-disjoint union of $\mathcal{G}^1_i$ and $\mathcal{G}^2_i$. Due to Proposition 1, it is an easy exercise to check that the database relation R corresponding to the chosen representation graphs in fact represents Σ^+ and, thus, Σ. □

Corollary 7. *Let $n \geq 2$. A set Σ of generalized and minimum participation constraints admits a $\mathcal{P}$-Armstrong relation if and only if Σ is conflicting or contains some finite participation constraint for every $C_i \in \Omega$.*

Proof. By virtue of the discussion at the end of Section 3, it suffices to show that Σ does not imply a finite participation constraint for a fixed $C_j \in \Omega$ unless it is conflicting or contains such a constraint. Suppose Σ is not conflicting and contains no finite participation constraint for C_j, but assume Σ implies some constraint $card^{part}(C_j) = b$. Adjoin new participation constraints $card^{part}(C_j) = b + 1$ and $card^{part}(C_i) = a_i$ for every C_i, $i \neq j$, without a finite participation constraint in Σ where a_i is defined as in the proof of the preceding theorem. By this theorem, the augmented constraint set may be represented by a database relation R. Hence, R satisfies Σ, but violates $card^{part}(C_j) = b$. □

Corollary 8. *The rule system presented in Section 4 is sound and complete for generalized and minimum participation constraints, that is, a set Σ of generalized and minimum participation constraints is $\mathcal{P}$-closed if and only if Σ is syntactically closed w.r.t. these rules.*

7 Final remarks

Before closing this paper, there are two remarks called for. The interested reader might wonder why we did not extend the concept of minimum participation constraints to subsets of Ω. This, of course, would be a natural idea similar to the investigation of generalized participation constraints. The reason for this is twofold. First, lower bounds for the occurrence in pairs, triples, etc. of entries are rarely used in database design. This, however, is surely not a satisfactory answer from the scientific point of view. Rather, in case we could construct matrix representations for these extended set of constraints, we would have written a completely different paper. Let $n \geq 3$ and Σ contain the constraints $card^{part}(C_i) = k$ and $card^{min}(C_i) = k$ for every $C_i \in \Omega$, and $card^{part}(X) = 1$ and $card^{min}(X) = 1$ for every two-element subset $X \subseteq \Omega$. A database relation representing Σ is a transversal design $TD(n, k)$ with block size n and group size k. For relevant notions from combinatorial design theory, we refer to [6]. It is well-known that a transversal design $TD(n, k)$ corresponds to a set of $n - 2$ mutually orthogonal Latin squares. The question whether there exists a $TD(5, 10)$, that is, a set of 3 mutually orthogonal squares of order 10, however, is one of the most famous open problems in design theory. This, we hope, explains our difficulties with extending the concept of minimum participation constraints to subsets of Ω.

Further, one may ask for matrix representations of minimum size. For functional dependencies this problem has been widely studied and partial results have been obtained. For the constraints studied in this paper, we already started to study this

question for special constraint sets. This time, let $n \geq 3$ and Σ contain the constraints $card^{part}(C_i) = k$ and $card^{min}(C_i) = k$ for every $C_i \in \Omega$, and $card^{part}(X) = 1$ for every two-element subset $X \subseteq \Omega$. If there exists a transversal design $TD(n, k)$, this would be a matrix representation of minimum size for Σ. Hence, again, it seems to be hard to achieve general results on the minimum size of a database relation representing a rather simple set of generalized and minimum participation constraints. In either case, the known results on key dependencies show that in the worst case the minimum size will be exponential in n.

References

[1] W. W. Armstrong. Dependency structures of database relationship. *Inform. Process.*, 74:580–583, 1974.

[2] C. Beeri, M. Dowd, R. Fagin, and R. Statman. On the structure of Armstrong relations for functional dependencies. *J. ACM*, 31:30–46, 1984.

[3] F. E. Bennett and L. Wu. On minimum matrix representation of closure operations. *Discrete Appl. Math.*, 26:25–40, 1990.

[4] F. E. Bennett and L. Wu. On minimum matrix representation of Sperner systems. *Discrete Appl. Math.*, 81:9–17, 1998.

[5] B. Bollobás. *Extremal graph theory.* Academic Press, London, 1978.

[6] C. J. Colbourn and J. H. Dinitz, editors. *The CRC handbook of combinatorial designs.* CRC press, Boca Raton, 1996.

[7] J. Demetrovics and G. Gyepesi. On the functional dependency and some generalizations of it. *Acta Cybernet.*, 5:295–305, 1981.

[8] J. Demetrovics, G. O. H. Katona, and A. Sali. Design type problems motivated by database theory. *J. Statist. Plann. Inference*, 72:149–164, 1998.

[9] P. Erdős. Extremal problems in graph theory. In F. Harary, editor, *A seminar in graph theory*, pages 54–64. Holt, Rinehart and Winston, 1967.

[10] A. Hajnal and E. Szemerédi. Proof of a conjecture of Erdős. In P. Erdős, A. Renyi, and V. T. Sós, editors, *Combinatorial theory and its applications*, volume 4 of *Colloq. J. Bolyai Math. Soc.*, pages 601–623. North-Holland, Amsterdam, 1970.

[11] H. Mannila and K. Räihä. *The design of relational databases.* Addison-Wesley, Reading, 1992.

[12] B. Thalheim. *Dependencies in relational databases.* Teubner, Stuttgart, 1991.

[13] B. Thalheim. *Entity-relationship modeling.* Springer, Berlin, 2000.