Functional dependencies distorted by errors

János Demetrovics [a] Gyula O.H. Katona [b,1,*] Dezső Miklós [c,1]

[a] *Comp. and Autom. Institute*
Hungarian Academy of Science
Kende u. 13-17, H-1111, Hungary

[b] *Alfréd Rényi Institute of Mathematics*
Hungarian Academy of Sciences
Budapest P.O.B. 127 H-1364 Hungary

[c] *Alfréd Rényi Institute of Mathematics*
Hungarian Academy of Sciences
Budapest P.O.B. 127 H-1364 Hungary

Abstract

A relational database D is given with Ω as the set of attributes. We assume that the rows (tuples, data of one individual) are transmitted through a noisy channel (or, as many times in case of the datamining applications, the observed data is distorded from the real values in a manner which we cannot know). In case of low probability of the error it may be supposed that at most one data in a row is changed by the transmission or observation. We say that $A \to b$ ($A \subset \Omega, b \in \Omega$) is an error-correcting functional dependency if the data in A uniquely determine the data in b in spite of this error. We investigate the problem how much larger a minimal error-correcting functional dependency can be than the original one. We will give upper and lower bounds showing that the can be considerably larger than the original sizes, but the growth is only polynomial.

Key words: relational database, functional dependencies

[*] Corresponding author.
Email addresses: `dj@ilab.sztaki.hu` (János Demetrovics), `ohkatona@renyi.hu` (Gyula O.H. Katona), `dezso@renyi.hu` (Dezső Miklós).

[1] The work of the second and third authors was supported by the Hungarian National Foundation for Scientific Research grant numbers T034702 and T037846

1 Introduction

Let us start with an example. Suppose that the pair of attributes (first name, last name) is a key in a database M (where the values in M are the real data). However, some of the data can be erroneous: the information is misunderstood in a phone conversation, the typist makes a mistake or the informant simply lies. We will denote by M^* the database containing the available, observed, and so sometimes erroneous data. Note that in most of the cases below we will only assume the existence of the real, error-free database M, but will be able to use only the observed version, M^*. Say, if we have both Maria Sklodowska and Mario Sklodowska in M and the first name "Mario" is replaced by "Maria" (M contains "Mario", M^* contains "Maria"), we might have two individuals with names Maria Sklodowska in M^*, hence the individual (row) cannot be determined from these two attributes. The question raised here is what other additional attributes we need to make us able to determine the real person (row).

A database can be considered as an $m \times n$ matrix M, where the rows are the data of one individual, the data of the same sort (*attributes*) are in the same column. Denote the set of attributes (equivalently, the set of columns of the matrix) by Ω, its size is $|\Omega| = n$. It will be supposed that the data of two distinct individuals are different, that is, the rows of the matrix are different. Let $A, B \subset \Omega$. We say that B *functionally depends* on A and write $A \to B$ if any two rows coinciding in the columns of A are also equal in the columns of B. Specially, if $K \to \Omega$ then K is called a *key*. In other words, there are no two distinct rows of the matrix which are equal in K. A key is a *minimal key* if no proper subset of it is a key. Denote the family of all minimal keys by $\mathcal{K}$.

Let M denote the matrix of the real data. These data are transmitted through a noisy channel. M^* ($m \times n$, again) denotes the matrix of the data obtained after the transmission. We will assume that the probability of errors is small and therefore that M and M^* can differ in at most e entries in each row. Although it is also supposed here that the real data of two distinct individuals are different, that is the rows of M are different, the same cannot be stated about M^*.

There might be several different reasons why we have to handle the erroneous database M^*. If the sender (possessing and sending M) and the receiver (receiving M^*) have both the possibility and the intention to cooperate, then the sender encodes the rows of M by an e-error-correcting code and sends the encoded row. The receiver decodes the received sequence and can recover the original row of M. That is, the receiver does not have, does not know any M^*. This is not our case, we assume, that the sender or original owner of the database does not cooperate with the receiver/user; for example, in case of datamining applications, the observed data is often distorded from the real values, which we cannot know. Or in case the data is transmitted via a channel, an error can occur during transmission. There might also be intentional or unintentional mistakes during providing the data to the user(s). That is, the receiver cannot recover the original row. Our only possible goal is to find connections between the structure of M and the structure of M^*, more precisely, between the system of functional dependencies of M and that of M^*.

Two different models will be considered in the present paper. In Model 1 it is assumed that the system of functional dependencies of M is known by the receiver, we only want to find a connection between this and the system of functional dependencies defined by M^*. On the other hand, in Model 2 nothing is known about the system of functional dependencies in M, we only know the received rows of M^* and our aim is to make conclusions based on this information. More precisely, the functional dependencies can be determined in M^* and the goal is to say something about the functional dependencies of the completely unknown M. There can be several interesting relationships between the two systems of functional dependencies, but in the present paper we investigate only the relationship between the sizes. (This will be later defined more precisely.)

First consider Model 1. Suppose for instance that $A \to a$ $(A \subset \Omega, a \in \Omega)$ holds in M. Then the data in a row in the columns of A determine the data of the same row in the column a. We know however only the corresponding rows in M^*. The data in the columns of A in M^* do not necessarily determine the data in a, since these data may be distorted. Can we enlarge A into an A' whose data in M^* already determine a? If yes, to what extent should it be enlarged?

For instance, if the number of errors in one row is at most one $(e = 1)$, and in our previous example sex is one of the attributes then either the first name or the sex is correct. Yet, the definitely erroneous record (Maria, Sklodowska, M) could be found in two different rows of M^* (one obtained from (Mario, Sklodowska, M) of M by changing the first name and the other one obtained from (Maria, Sklodowska, F of M by changing the sex). Further attributes might be needed to identify the individual. That is, (first name, last name, sex) is not a 1-error-correcting key (see the definition below).

Formalize our notions. The number of different entries of two sequences of the same length (two rows of the matrix) is the *Hamming distance*. If r, s are two such rows, this is denoted by $h(r, s)$. If $r = (r_1, \ldots, r_n)$ is a row of the matrix M, $A = \{i_1, \ldots, i_\alpha\}$ is a set of columns of M then let $r(A)$ denote the subrow (subsequence) of r determined by these columns: $r(A) = (r_{i_1}, \ldots, r_{i_\alpha})$. If r is a row of M, let r^* denote the corresponding row in M^*. The only information we know about r^* is that $h(r, r^*) \leq e$.

Let C and B be two sets of columns of M. We write $C \to \{e\}B$ if the knowledge of $r(C)$ for some row(s) r of M uniquely determines the subrow $r(B)$ even after changing at most e values in the row(s) in any arbitrary way. More formally, we have $C \to \{e\}B$ iff the following holds for every pair r, s of rows of M: for all distorted versions r^* and s^* satisfying $h(r, r^*) \leq e, h(s, s^*) \leq e$ the equation $r^*(C) = s^*(C)$ implies $r(B) = s(B)$. Observe that this is not necessarily true for the distorted versions $r^*(B)$ and $s^*(B)$, but they might differ only in at most $2e$ places. $C \to \{e\}B$ is called an *e-error-correcting functional dependency.* In case of $C \to \{e\}\Omega$ we say that C is an *e-error-correcting key.*

Note that we actually formulated the e-error-correcting functional dependencies in terms of M. The name might be misleading: the transmitted information is not (even theoretically cannot be) corrected here. Only the functional dependency is "corrected" in the sense that it is recognizable even in M^* in the following sense: if a certain functional dependency $A \to B$ holds in M, but does not hold in M^*, A is enlarged ("corrected") into C to make the functional dependency "valid" even in M^*.

The aim of the present paper is to find inequalities between the sizes of the sets occurring in the really existing functional dependencies in M and sizes of the sets occurring in the e-error-correcting ones. Our two conference papers [5], [6] contain our results in a preliminary form, [5] deals with the case of the keys.

It is worth mentioning that $\{a\} \to \{1\}\{a\}$ does not hold, since the knowledge of the data in the column a does not give any information, it can be erroneous. It does not determine the value in column a.

The $m \times |C|$ submatrix of M determined by (corresponding to) the set C of its columns is denoted by $M(C)$.

Proposition 1.1 *$C \to \{e\}B$ ($C \subset \Omega, B \subset \Omega$) is an e-error-correcting functional dependency iff the pairwise Hamming distance of the rows $r(C), s(C)$ of $M(C)$ is at least $2e+1$ whenever $r(B) \neq s(B)$.*

Proof. First suppose that the Hamming distance of every pair of rows $r(C), s(C)$ of $M(C)$ is at least $2e+1$ if the rows are different in some column $a \in B$ (that is $r(\{a\}) \neq s(\{a\})$). In other words $r(B) \neq s(B)$ implies $h(r(C), s(C)) \geq 2e+1$. Hence $r^*(C) \neq s^*(C)$ follows. Then $C \to \{e\}B$ is an e-error-correcting functional dependency.

The converse is also true. Suppose that $C \to \{e\}B$ is an e-error-correcting functional dependency, and $r(B) \neq s(B)$ holds for a pair of rows. Then $h(r(C), s(C)) \geq 2e+1$ must hold, otherwise one could find appropriate distorted versions $r^*(C) = s^*(C)$, contradicting the definition. □

Observe that the conditions are totally formulated in terms of M. Proposition 1.1 provides an alternative definition for e-error-correcting functional dependencies and hence we will use this form in the rest of the paper rather than the original definition. Model 1 is studied in Section 2.

Model 2 is motivated by data mining. Then only M^* is known, nothing is known about M. In Model 1 the keys, functional dependencies in M are known for the receiver. We only want to derive conditions for the functional dependencies of M^* from those of M. In case of data mining (Model 2) we have no prior information on M. Only M^* is known, it defines some (virtual) functional dependencies. The question here is what the relationship between (the unknown) original dependencies (in M) and the virtual ones (in M^*) is. Model 2 will be studied in Section 3.

2 Model 1: known dependency structure

It is easy to see that if the pairwise Hamming distance of the rows of $M(C)$ being different in attribute a is at least $2e$ then the knowledge of $M^*(C)$ detects the error (i.e. the presence of the error in M^*), but does not determine the data in a uniquely, i.e. there can be more than one row of M having the same values in $M^*(C)$ and different in $M(\{a\})$. This case is less interesting, but it makes worth introducing the more general definition: $C \to (d)B$ is called a *d-distance functional dependency* iff the pairwise Hamming distance of the rows r, s of $M(C)$, which are different in B (that is $r(B) \neq s(B)$), is at least d.

The main aim of the present investigations is to find connections between the functional dependencies (in M) and the d-distance functional dependencies. The next proposition is the first step along this line. Let $\mathcal{F}_B$ be the family of minimal subsets F of Ω satisfying $F \to B$ (in M!).

Proposition 2.1 *$C \to (d)B$ $(C \subset \Omega, B \subset \Omega)$ is a d-distance functional dependency iff for any choice $a_1, \ldots, a_{d-1} \in C$ one can find an $F \in \mathcal{F}_B$ such that $F \subseteq C - \{a_1, \ldots, a_{d-1}\}$.*

Proof. The necessity will be proved in an indirect way. Suppose that there exist $a_1, \ldots, a_{d-1} \in C$ such that $C - \{a_1, \ldots, a_{d-1}\}$ contains no member of $\mathcal{F}_B$, that is, $C - \{a_1, \ldots, a_{d-1}\} \to B$ does not hold. Therefore there are two rows r, s of M which are equal in $M(C - \{a_1, \ldots, a_{d-1}\})$ and satisfy $r(B) \neq s(B)$. The Hamming distance of these two rows in $M(C)$ is less than d. The contradiction with the definition of d-distance dependency completes this part of the proof.

To prove the sufficiency suppose, again in an indirect way, that $M(C)$ contains two rows r, s with Hamming distance $< d$ and the rows are different in B ($r(B) \neq s(B)$). Delete those columns from C where these rows are different. We found a set $C - \{a_1, \ldots, a_{d-1}\}$ satisfying the condition that $M(C - \{a_1, \ldots, a_{d-1}\})$ contains two rows which are equal everywhere, but the rows are different in B. Therefore $C - \{a_1, \ldots, a_{d-1}\} \to B$ is not true in M, hence $C - \{a_1, \ldots, a_{d-1}\}$ cannot contain a member of $\mathcal{F}_B$. □

The systems of functional dependencies were characterized in [1]. We prefer an equivalent description (see e.g. [4]) by the closure

$$\mathcal{L}(A) = \{a : a \in \Omega, A \to a\} \ (A \subseteq \Omega).$$

It is easy to see that this closure satisfies the following 3 conditions.

$$A \subseteq \mathcal{L}(A), \tag{i}$$

$$A \subseteq B \text{ implies } \mathcal{L}(A) \subseteq \mathcal{L}(B), \tag{ii}$$

$$\mathcal{L}(\mathcal{L}(A)) = \mathcal{L}(A). \tag{iii}$$

It is well-known ([1], [3]) that there is a database for any closure, in which the system of functional dependencies is exactly the one defined by this closure. This is why it is sufficient to give a closure rather than constructing the complete database or matrix.

It is possible to give a characterization with the families $\mathcal{F}_B$ as well. It is easy to see that $\mathcal{F}_B$ is a non-empty *inclusion-free* family of subsets of Ω. (Inclusion-free means that $F_1, F_2 \in \mathcal{F}_B, F_1 \neq F_2$ implies $F_1 \not\subset F_2$.) On the other hand, since $B \to B$ holds, $\mathcal{F}_B$ must have a member which is a subset of B. We need one more condition for the interrelation between these families. However, since we did not find the shortest form and no such characterization is needed in this paper we prove only the following lemma, which will be needed later.

Lemma 2.2 *Let $B \subseteq \Omega$. Given an inclusion-free family $\mathcal{F}$ of subsets of Ω with a B such that $F \subseteq B$ holds for an $F \in \mathcal{F}$, then there is a system of functional dependencies (and therefore, by the preceding remark, a relational database (matrix) M) such that it defines $\mathcal{F}_B = \mathcal{F}$.*

Proof. Let $\mathcal{L}(A) = A \cup B$ for $A \subseteq \Omega$ if $G \subseteq A$ holds for some member G of $\mathcal{F}$, and let $\mathcal{L}(A) = A$ otherwise. It is easy to see that this function satisfies conditions (i)-(iii), that is, it is a closure. On the other hand, $G \to B$ holds for every member G of $\mathcal{F}$ and these are minimal with this property. □

In other words, Lemma 2.2 says that for any inclusion-free family $\mathcal{F}$ with a member in B there is a database where the family of minimal sets F satisfying $F \to B$ is exactly equal to $\mathcal{F}_B = \mathcal{F}$.

Proposition 2.1 makes us able to give an abstract combinatorial definition, independent of databases. Let X be an n-element set and $\mathcal{F}$ be an inclusion-free family of its subsets. The *d-blowup* of $\mathcal{F}$ (in notation $\mathcal{F}(d)$) is defined by

$$\mathcal{F}(d) = \{G \subseteq X : \text{ for any choice of } x_1, \ldots, x_{d-1} \in G \ \ \exists F \in \mathcal{F} \text{ such that}$$

$$F \subseteq G - \{x_1, \ldots, x_{d-1}\} \text{ and } G \text{ is minimal for this property}\}.$$

Note that $\mathcal{F}(1) = \mathcal{F}$. As as we will see later, if the inclusion-free family of sets $\mathcal{F}$ consists of the minimal left hand sides of the functional dependencies $F \to B$ of a relation for a given B, $\mathcal{F}(d)$ will be the minimal left hand sides of the d-distance dependencies $C \to (d)B$.

Our first observation is that it may happen that the d-blowup of $\mathcal{F}$ is an empty family while the original $\mathcal{F}$ is not. Fix an element $a \in X$ and an integer $2 \leq k$. Define $\mathcal{F}$ as the family of all k-element sets ($\subset X$) containing a. Then for any $C \subseteq X$, the set $C - \{a\}$ cannot contain any member of $\mathcal{F}$ therefore $\mathcal{F}(d)$ is empty for $2 \leq d$.

On the other hand, if $\mathcal{F}$ consists of all k-element subsets of X then $\mathcal{F}(d)$ will consists of all sets $G \subseteq X$ with $k + d - 1$ elements. Our last example suggests that the sizes of the members of $\mathcal{F}(d)$ do not exceed the sizes of the members of $\mathcal{F}$ by too much. We will show that this is not really true.

We say that the family $\mathcal{F}$ can be *pinned* by p elements if there are $x_1, \ldots, x_p \in X$ such that no member of $\mathcal{F}$ avoids all of them, that is $F \cap \{x_1., \ldots, x_p\} \neq \emptyset \ \ \forall F \in \mathcal{F}$. It is obvious that if $\mathcal{F}$ can be pinned by $d-1$ elements then $\mathcal{F}(d)$ is empty. Otherwise $\mathcal{F}(d)$ is never empty since X always satisfies the first part of the definition of the blowup and if it is not minimal, one can reduce it until arriving to a minimal set. The following theorem, our main result in terms of subsets of a finite set, will be proved in Section 4.

Theorem 2.3 *Let $n_0(k, d) \leq n$ and let $\mathcal{F}$ be an inclusion-free family of subsets of size at most k of a given set of size n, such that $\mathcal{F}$ cannot be pinned by $d - 1$ elements. Then the sizes of the members of $\mathcal{F}(d)$ are at most $c_1 k^d$. On the other hand there is an (inclusion-free family of subsets of size at most k of a given set of size n) $\mathcal{F}$ for which all members of $\mathcal{F}(d)$ have size at least $c_2 k^d$. Here c_1 and c_2 depend only on d.*

Applying this theorem for the functional dependencies and error-correcting functional dependencies the following theorem will be easily deduced in Section 4.

Theorem 2.4 *Let M be a database (matrix) whose set of columns is Ω, where $n_0(k, d) = n'_0(k, e) \leq n = |\Omega|$ (with $d = 2e + 1$). Fix a subset $B \subseteq \Omega$. Suppose that all the members of $\mathcal{F}_B$ (minimal C's satisfying $C \to B$ in M) have* *sizes at most k. Then the minimal e-error-correcting dependencies $C \to \{e\}B$ satisfy $|C| \le c_1 k^{2e+1}$. On the other hand there is a database and $B \subseteq \Omega$ in which the members of $\mathcal{F}_B$ are of size k and every e-error-correcting dependency $C \to \{e\}B$ satisfies $c_2 k^{2e+1} \le |C|$. Here c_1 and c_2 depend on e, only.*

It is worth formulating the special case $e = 1$ with more specific constants.

Corollary 1 *Let M be a database (matrix) whose set of columns is Ω, where $n_0'(k) = n_0'(k, 3) \le n = |\Omega|$. Fix a subset $B \subseteq \Omega$. Suppose that all the members of $\mathcal{F}_B$ (minimal C's satisfying $C \to B$ in M) have sizes at most k. Then the minimal error-correcting dependencies $C \to \{1\}B$ satisfy $|C| \le 3k^3$. On the other hand there is a database and $B \subseteq \Omega$ in which the members of $\mathcal{F}_B$ are of size k and every error-correcting dependency $C \to \{1\}B$ satisfies $\frac{2}{27}k^3 \le |C|$.*

Our conclusion is that the errors can considerably increase the sizes of the minimal functional dependencies, but the growth is only polynomial.

3 Model 2: datamining after error

M^* defines a dependency structure. We want to determine some connections between this dependency structure and that of the original M. However, the dependency structure of M^* depends in large extent on the number of actual errors. It might happen that $M^* = M$ then we have the same dependencies. At the other end, the worst case is when all possible errors occur, that is all possible rows obtained from the rows of M by changing at most e values are considered. Let this matrix be denoted by $\hat{M}(e)$. (While the number of rows of M and M^* are the same, the number of rows of $\hat{M}(e)$ is much more.) The (usual) functional dependency defined by $\hat{M}(e)$ is denoted by $\overset{\hat{e}}{\to}$.

It is easy to prove the following lemma.

Lemma 3.1 *Let $a \in \Omega, C \subseteq \Omega$. Then $C \overset{\hat{e}}{\to} a$ holds iff either $a \in C$ or $C - \{a_1, \dots, a_{2e}\} \to a$ (defined in M) holds for every choice of $a_1, \dots, a_{2e} \in C$.*

Since the rows of M^* are selected from the set of rows of $\hat{M}(e)$, any functional dependency valid in $\hat{M}(e)$ must be valid in M^*, too. This results in the following lemma.

Lemma 3.2 *Let $C \subseteq \Omega, a \in \Omega - C$. If $C - \{a_1, \dots, a_{2e}\} \to a$ holds in M for every choice of $a_1, \dots, a_{2e} \in C$ then $C \to a$ holds in M^*.*

Let $\mathcal{G}_a$ denote the family of minimal sets F satisfying $a \notin F, F \to a$ in M. By Lemma 3.2 we have that $D \in \mathcal{G}_a(2e + 1)$ (the blowup of $\mathcal{G}_a$) implies $D \to a$ in M^*. Let k be the largest size in the family $\mathcal{G}_a$. Theorem 2.3 implies that the sizes of the members of $\mathcal{G}_a(2e+1)$ are at most c_1k^{2e+1}, supposing that $n_0(k, 2e+1) \leq n-1$ (not n, since a is deleted from the underlying set). This upper bound is valid for all minimal sets D such that $a \notin D, D \to a$ in M^*. Therefore, if ℓ is the size of the largest such minimal D, then $\ell \leq k^{2e+1}$ holds, hence $\ell^{\frac{1}{2e+1}} \leq k$ follows.

Theorem 3.3 *Fix $a \in \Omega$ and suppose that $n_0(\ell^{\frac{1}{2e+1}}, 2e+1) + 1 \leq n$. If the largest minimal set D satisfying $a \notin D, D \to a$ in M^* has size ℓ then there is a C of size at least $\ell^{\frac{1}{2e+1}}$ such that $a \notin C$ and $C \to a$ holds in M.*

In other words, knowing the size of the largest minimal set of the columns determining column a in the distorted M^*, its $(2e+1)$th root is a lower bound for the corresponding minimal set of columns in the original M (unknown for us). Similar estimates can be deduced for the general case when a is replaced by a set of columns.

4 Proofs

Proof of Theorem 2.3 This proof is analogous to the proof of the main theorem of [6]. Let $\mathcal{F}$ be an inclusion-free family of subsets of X. The definition of $\mathcal{F}(d)$ implies that the family $\{F : F \in \mathcal{F}, F \subseteq G\}$ cannot be pinned by $d-1$ elements for members $G \in \mathcal{F}(d)$. On the other hand, by the minimality of a member $G \in \mathcal{F}(d)$, this is not true for $G - \{a\}$ where $a \in G$ is chosen arbitrarily. This gives the following proposition.

Proposition 4.1 *$G \in \mathcal{F}(d)$ iff $\{F : F \in \mathcal{F}, F \subseteq G\}$ cannot be pinned by $d-1$ elements, but $\{F : F \in \mathcal{F}, F \subseteq G - \{a\}\}$ can be pinned by some $d-1$ elements for every $a \in G$.*

□

Lower estimate. We give an inclusion-free family $\mathcal{F}$ consisting of $2 \leq k$-element sets which generates an $\mathcal{F}(d)$ consisting of one member having size at least c_2k^d.

Fix an integer $1 \leq i < k$ and take a subset $A \subset X$ of size $i+d-1$. Let $B_1, B_2, \ldots$ be all the $\binom{i+d-1}{i}$ i-element subsets of A and

$$\mathcal{G}^i = \{B_1 \cup C_1, B_2 \cup C_2, \ldots\},$$

where $C_1, C_2, \ldots$ are disjoint subsets of $X - A$ with $|C_1| = |C_2| = \cdots = k - i$.

This can be carried out if

$$i+d-1+\binom{i+d-1}{i}(k-i)\le n. \tag{4.1}$$

Using Proposition 4.1 we next show that the only member of $\mathcal{G}^i(d)$ is $D = A\cup\bigcup_i C_i$. It is easy to see that $\mathcal{G}^i$ cannot be pinned by $d-1$ elements.

On the other hand, if $a\in C_j$ for some j then the $d-1$-element $A-B_j$ pins all members of $\mathcal{G}^i$ within $D-\{a\}$. If, however, $a\in A$ then any $d-1$-element $E\subset A$ not containing a pins the members of $\mathcal{G}^i$ within $D-\{a\}$. By Proposition 4.1 D is really a member of $\mathcal{G}^i(d)$. If there existed another member, D would not be minimal. This proves that D is the only member of $\mathcal{G}^i(d)$.

Choose $i=\left\lfloor k(1-\frac{1}{d})\right\rfloor$. The inequalities

$$k(1-\frac{1}{d})\le i+1, \qquad \frac{k}{d}\le k-i \tag{4.2}$$

are easy consequences. The size of D, given by the left hand side of (4.1) can be lower-bounded by

$$\frac{(i+1)^{d-1}}{(d-1)!}(k-i).$$

Substituting the inequalities of (4.2) the lower bound

$$c_2(d)=\frac{(d-1)^{d-1}}{d^d(d-1)!}k^d$$

is obtained.

(4.1) also gives a condition on how large n has to be. To obtain an explicit formula for $n_0(k,d)$ an upper estimate is needed for the left hand side of (4.1). □

Upper estimate. Let $G\in\mathcal{F}(d)$ where $\mathcal{F}\subset\binom{X}{\le k}$ (the latter one denotes the family of all subsets of X of size at most k). We will prove that $|G|\le dk^d$. Since we have to consider only the subsets of G, it can be supposed that all members of $\mathcal{F}$ are subsets of G.

Proposition 4.1 defines d-element subsets D of G each of them is pinning $\mathcal{F}$, namely every element a of G can be extended to such a set D. Moreover, still by Proposition 4.1, their union is G. Denote this family by $\mathcal{D}$. We know

$$\cup_{D\in\mathcal{D}}D=G, \tag{4.3}$$

$$D\cap F\ne\emptyset \text{ for all } D\in\mathcal{D}, F\in\mathcal{F} \tag{4.4}$$

and $\mathcal{F}$ cannot be pinned by a set with less than d elements.

Let $I \subseteq G$. Define the I-degree of $\mathcal{D}$ as the number of members of $\mathcal{D}$ containing I, that is,

$$\deg_I(\mathcal{D}) = |\{D \in \mathcal{D} : \ I \subset D\}|.$$

Lemma 4.2 *If $|I| < d$ then*

$$\deg_I(\mathcal{D}) \leq k^{d-|I|}.$$

Proof. We use induction on $j = d - |I|$. Suppose that $j = d - |I| = 1$, that is, $|I| = d - 1$. If all members of $\mathcal{F}$ meet I then $\mathcal{F}$ can be pinned by $d - 1$ elements, a contradiction. Therefore there is an $F \in \mathcal{F}$ which is disjoint from I. By (4.4) all the sets D satisfying $I \subset D$ must intersect this F, therefore their number is $\leq |F| \leq k$. This case is settled.

Now suppose that the statement is true for $j = d - |I| \geq 1$ and prove it for $j+1$. Let $|I^*| = d-j-1$. There must exist an $F \in \mathcal{F}, F \cap I^* = \emptyset$ otherwise $\mathcal{F}$ is pinned by less than d elements, a contradiction. Let $F = \{x_1, \ldots, x_l\}$ where $l \leq k$. By (4.4) we have

$$\{D \in \mathcal{D} : \ I^* \subset D\} = \cup_{i=1}^{l} \{D \in \mathcal{D} : \ (I^* \cup \{x_i\}) \subset D\}. \qquad (4.5)$$

The sizes of the sets on the right hand side are $\deg_{I^* \cup \{x_i\}}(\mathcal{D})$ which are at most $k^{d-|I^*|-1} = k^j$ by the induction hypothesis. Using (4.5)

$$\deg_{I^*}(\mathcal{D}) \leq lk^{d-|I^*|-1} \leq k^{d-|I^*|}$$

is obtained, proving the lemma. □

Finally, consider any $F = \{y_1, \ldots, y_r\} \in \mathcal{F}$ where $r \leq k$. By (4.3), the families $\{D \in \mathcal{D} : y_i \in D\}$ cover $\mathcal{D}$. Apply the lemma for $I = \{y_i\}$:

$$|\{D \in \mathcal{D} : y_i \in D\}| \leq k^{d-1}.$$

This implies $|\mathcal{D}| \leq rk^{d-1} \leq k^d$ and $|\cup_{D \in \mathcal{D}} D| \leq |\mathcal{D}|d \leq dk^d$. Application of (4.3) completes the proof: $|G| \leq dk^d$, therefore $c_1(d) = d$ is an appropriate choice. □

Proof of Theorem 2.4. By Proposition 2.1 and the definition of the blowup, the family of minimal e-error correcting functional dependencies $C \rightarrow \{e\}B$ is exactly $\mathcal{F}_B(2e+1)$. Apply the upper bound part of Theorem 2.3 for the family $\mathcal{F}_B$. Since its members are not larger than k, the theorem implies that all members of $\mathcal{F}_B(2e+1)$ are of size at most $c_1 k^{2e+1}$.

On the other hand, take the inclusion-free family $\mathcal{F}$ giving the optimum in the lower estimate in Theorem 2.3. Take e.g. a $B \in \mathcal{F}$. Lemma 2.2 defines a system of functional dependencies (database) in which $\mathcal{F}_B = \mathcal{F}$ holds. Here $\mathcal{F}(2e+1)$ contains only sets of size at least $c_2 k^{2e+1}$. □

5 Further problems

1. Although Theorem 2.3 determines the order of magnitude of the smallest size in the "worst" family, it does give the exact value. We believe that the lower estimate is sharp, our construction is the best possible.

Conjecture 5.1 *If $\mathcal{F} \subseteq \binom{X}{\leq k}$ where $n_0(k,d) \leq n = |X|$ then $\mathcal{F}(d)$ has a member with size at most*

$$\max_i \{i + d - 1 + \binom{i+d-1}{i}(k-i)\}.$$

2. Can the systems of e-error-correcting dependencies be characterized? Since the first version of the present paper was written Thalheim and Schewe gave an answer to this question in [7].

3. In Section 4 some connection was shown between the systems of functional dependencies in M and $\hat{M}$. Here we suggest to study another connection. A set $A \subseteq \Omega$ is said to be *closed* with respect to a certain system of functional dependencies (for instance the ones defined by a matrix) if $A \to a$ holds only for the elements of A. It was proved in [2] that the systems of functional dependencies (or equivalently, the closures) form a natural ranked poset with the rank function *number of closed sets* -1. (It was -2 in the paper, since it was supposed that the closure of the empty set is empty.) The question we pose here is the following. Given the rank of the system of functional dependencies (closure) in M, give estimates on the rank of the same in $\hat{M}$.

4. $\hat{M}$ represents the worst case. It is probable that the rank of the received M^* differs much less from that of M than the rank of $\hat{M}$. Under some reasonable probabilistic assumptions, give probabilistic estimates on the change of the rank.

We are indebted to the anonymous referees for their valuable suggestions which improved the paper to a great extent.

References

[1] Armstrong, W.W., Dependency structures of data base relationship, in: *Information Processing 74*, North-Holland, Amsterdam, pp. 580-583.

[2] Burosch, G., Demetrovics, J. and Katona G.O.H., The Poset of Closures as a Model of Changing Databases, *Order* **4**(1987) 127-142.

[3] Demetrovics, J., On the equivalence of candidate keys with Sperner systems, *Acta Cybernet.* **4**(1979) 247-252.

[4] Demetrovics J., Füredi, Z, and Katona, G.O.H.: Minimum matrix representation of closure operations, *Discrete Appl. Math.* **11**(1985) 115-128.

[5] Demetrovics J., Katona, G.O.H., Miklós, D.: Error-correcting keys in relational databases, in *Foundations of Information and Knowledge Systems, FoIKS 2000* (K.-D. Schewe and B. Thalheim eds.) Lecture Notes in Computer Science, **1762**, Springer, 2000, pp. 88-93.

[6] Demetrovics J., Katona, G.O.H., Miklós, D.: Functional dependencies in presence of errors, in *Foundations of Information and Knowledge Systems, FoIKS 2002* (Th. Eiter and K.-D. Schewe eds.) Lecture Notes in Computer Science, **2284**, Springer, 2002, pp. 85-92.

[7] Schewe, K.-D., Thalheim, B.: manuscript.