Recent Combinatorial Results in the Theory of Relational Databases

J. DEMETROVICS
Computer and Automation Institute
Hungarian Academy of Sciences
Kende u. 13-17, H-1111 Budapest, Hungary
dj@ilab.sztaki.hu

GY. O. H. KATONA AND D. MIKLÓS
Alfréd Rényi Institute of Mathematics
Hungarian Academy of Sciences
Reáltanoda utca 13-15, H-1053 Budapest, Hungary
<ohkatona><dezso>@renyi.hu

Abstract—Recent results of the authors—some of which are joint with Thalheim and Seleznjev—in the area of combinatorial investigations of the relational database model are presented here. In relational databases keys—combinations of attributes uniquely identifying the records—play an important role. The structure and size of keys have been widely investigated. Here, after a short review of the earlier results, we discuss two generalizations: the (average) structure and size of keys in a random database and the concept of error-correcting keys in case of unreliable data collection. © 2003 Elsevier Ltd. All rights reserved.

Keywords—Relational database, Functional dependency, Keys, Minimal keys, Random database.

1. INTRODUCTION

A *database* M (or R) is an $m \times n$ matrix where the columns are the n *attributes* of the database—the set of whose is denoted usually by Ω—and the rows correspond to the m *records* or *individuals*. It will be supposed that the data of two distinct individuals are different; that is, the rows of the matrix are different.

If $A, B \subset \Omega$, $b \in \Omega$ we say that B *(functionally) depends* on A (denoted by $A \to B$) or b (functionally) depends on A (denoted by $A \to b$) iff there are no rows (records) of M equal in A but different in B (or b).

A subset K of Ω is called a *key* if the data in K determine the individual (row) uniquely. 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. Usually, we denote the family of all minimal keys by $\mathcal{K}$ while the family of all maximal subsets of attributes which are not keys is denoted by $\mathcal{K}^{-1}$.

The work of the first author was supported by the Hungarian National Foundation for Scientific Research Grant Number T042706 and European Community's Centre of Excellence Grant Number ICA1-CT-2000-70025 while the work of the second and third authors was supported by the Hungarian National Foundation for Scientific Research Grant Number T029255 and European Community's Centre of Excellence Grant Number ICA1-CT-2000-70009.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

REMARK 1.1. $\mathcal{K}$ is a Sperner family; that is, for every $A, B \in \mathcal{K}$ we have $A \not\subset B$. Therefore, by the well-known theorem of Sperner [1], $|\mathcal{K}| \leq \binom{n}{\lfloor n/2 \rfloor}$.

The first classic question of this area was whether every Sperner family can be obtained in this way.

THEOREM 1.2. *(See [2,3].) $\mathcal{K}$ is the set of minimal keys of a certain database iff $\mathcal{K}$ is a nonempty Sperner family.*

The next question is whether we can draw a conclusion from the number of records of the database to the system of minimal keys or, more precisely, what is the minimal size (number of records, that is number of rows of the matrix) of a database realizing a given Sperner family.

DEFINITION 1.3. $s(\mathcal{K})$ = *the minimal number of records of a database where the system of minimal keys is* $\mathcal{K}$.

Regarding this question the following results were found first for the size of $\mathcal{K}$ and the size of $\mathcal{K}^{-1}$.

THEOREM 1.4.

$$|\mathcal{K}^{-1}| \leq \binom{s(\mathcal{K})}{2} \qquad \text{(Demetrovics and Katona [4]),}$$

$$\sqrt{2\,|\mathcal{K}^{-1}|} < s(\mathcal{K}) \leq 1 + |\mathcal{K}^{-1}| \qquad \text{(Demetrovics and Katona [4]),}$$

$$\forall\,\mathcal{K},\ s(\mathcal{K}) \leq \binom{n}{\left\lfloor \frac{n}{2} \right\rfloor} + 1 \qquad \text{(Demetrovics and Gyepesi [5]),}$$

$$\exists\,\mathcal{K},\ s(\mathcal{K}) \geq \frac{1}{n^2}\binom{n}{\left\lfloor \frac{n}{2} \right\rfloor} \qquad \text{(Demetrovics and Gyepesi [5]).}$$

The proof of the lower bound for $s(\mathcal{K})$ above is not constructive. Nothing is known about the (nearly) worst Sperner families, that is about the $\max_{\mathcal{K}} s(\mathcal{K})$, and similarly, nothing is known about $\max_{|\mathcal{K}|=k} s(\mathcal{K})$ or $\min_{|\mathcal{K}|=k} s(\mathcal{K})$. The similar questions for $\mathcal{K}^{-1}$ can be asked as well. Though neither $\max_{|\mathcal{K}|=k} \mathcal{K}^{-1}$ nor $\min_{|\mathcal{K}|=k} \mathcal{K}^{-1}$ are known, here at least we conjecture that for ks relatively small compared to n the minimum is attained by a family consisting of i and $i+1$ element subsets where i is determined by $\binom{n}{i} \leq k \leq \binom{n}{i+1}$.

Some results are known for the special Sperner families consisting of sets of uniform size. Let $\mathcal{K}_k^n$ denote the family of all k-element subsets of an n-element set; that is,

$$\mathcal{K}_k^n = \binom{\Omega}{k} \Rightarrow \mathcal{K}^{-1} = \binom{\Omega}{k-1}.$$

We have then the following simple, but as later described surprisingly strong lemma.

LEMMA 1.5. *(See [4].)* $\binom{n}{k-1} \leq \binom{s(\mathcal{K}_k^n)}{2}$.

The lemma implies that for $k = 2$

$$s\left(\mathcal{K}_2^n\right) = \min\left\{ s \colon n \leq \binom{s}{2} \right\},$$

and therefore, we have the equality

$$s\left(\mathcal{K}_2^n\right) = \left\lceil \frac{1 + \sqrt{1 + 8n}}{2} \right\rceil.$$

Another easy consequence of the lemma is that in case of $k = 3$ for $s = s(\mathcal{K}_3^n)$ we obtain $\binom{n}{2} \leq \binom{s}{2}$ which implies that $n \leq s$. It turned out that this upper bound is almost always the exact answer.

THEOREM 1.6. *(See [6].) If $n = 12r + 1$ or $n = 12r + 4$, then*

$$s\left(\mathcal{K}_3^n\right) = n.$$

THEOREM 1.7. *(See [7].)*

$$s\left(\mathcal{K}_3^n\right) = n, \quad \text{if } n = 7, \quad n \geq 9.$$

For higher values of ks we do not have such nice results, but at least the asymptotic of $s(\mathcal{K}_k^n)$ is determined.

THEOREM 1.8. *(See [6].)*

$$c_k\, n^{(k-1)/2} \leq s\left(\mathcal{K}_k^n\right) \leq d_k\, n^{(k-1)/2}.$$

In this section, finally we present a recent result for another type of minimal keys.

THEOREM 1.9. *(See [8].) Let*

$$\mathcal{K} = \{\{1,2\},\{2,3\},\dots,\{n-1,n\},\{n,1\}\}.$$

Then,

$$\left|\mathcal{K}^{-1}\right|^{1-\varepsilon} < s(\mathcal{K}) \leq 1 + \left|\mathcal{K}^{-1}\right|$$

where

$$\left|\mathcal{K}^{-1}\right| = 2^{\alpha n + o(n)}, \qquad \alpha = 0,4056\dots.$$

The interested reader can have a more detailed (but not so up-to-date) overview of these results in, e.g., [9].

2. ERROR CORRECTING KEYS

Suppose that the data are collected in a nonreliable way; e.g., the transfer of data is done via a noisy channel or the sources of the data are not completely reliable. We may then assume that at most 1, or more general, at most e of the data of each individual may be incorrect. Under these circumstances a collection of attributes C will be called an *e-error-correcting key* iff the data in the columns belonging to C (some of which may be false) uniquely determine the individual (row) of the database.

The first remark we make here is that in the contrary of the ordinary key systems, where a database always has keys (in the worst case only the set of all of the attributes will be a key, but it definitely will be) there may not exist error-correcting keys at all in a database. To understand the forthcoming example, we need a few preliminary remarks, definitions, and propositions. The number of different entries in two rows is called the *Hamming distance* of these two rows. The $m \times |C|$ submatrix of M determined by the set C of its columns is denoted by $M(C)$.

The following proposition can be then easily justified.

PROPOSITION 2.1. *$C \subset \Omega$ is an e-error-correcting key iff the pairwise Hamming distances of the rows of $M(C)$ are at least $2e + 1$.*

This suggests the definition.

DEFINITION 2.2. *A subset of the attributes $C \subset \Omega$ is called a d-distance key iff the pairwise Hamming distance of the rows of $M(C)$ is at least d. Remark that the one-distance keys are just the ordinary keys.*

The following proposition will make it easier to understand the reason why a set of attributes will be a d-distance key of a database with system of minimal keys $\mathcal{K}$.

PROPOSITION 2.3. *$C \subset \Omega$ is a d-distance key of M iff for any choice of $a_1, \dots, a_{d-1} \in C$ one can find a $K \in \mathcal{K}$ such that $K \subset C - \{a_1, \dots, a_{d-1}\}$.*

PROOF. Assume that for a $C \subset \Omega$ there exist $a_1, \dots, a_{d-1} \in \Omega$ such that $C - \{a_1, \dots, a_{d-1}\}$ contains no member of $\mathcal{K}$, implying that $C - \{a_1, \dots, a_{d-1}\}$ is not a key. Therefore, there must be two distinct rows of M which are equal in $M(C - \{a_1, \dots, a_{d-1}\})$ and so the Hamming distance of these two rows in $M(C)$ is less than d, C is not is d-distance key by definition.

Assume now that a $C \subset \Omega$ is not a d-distance key, i.e., that $M(C)$ contains two distinct rows with Hamming distance $< d$. Denote the columns (plus probably a few arbitrarily chosen columns) where these rows are different by $\{a_1, \dots, a_{d-1}\}$. Then, $M(C - \{a_1, \dots, a_{d-1}\})$ will contain two distinct rows that are equal everywhere, and therefore, $C - \{a_1, \dots, a_{d-1}\}$ will not be a key in M; it will not contain a member of $\mathcal{K}$. ∎

It is again easy to see that any superset of a d-distance key will be a d-distance key as well, and therefore, it is enough to focus on the minimal d-distance keys of a database, which will be denoted by $\mathcal{C}_d(M)$. Observe that by Proposition 2.3, $\mathcal{K}$ and d determine $\mathcal{C}_d$; i.e., it is enough to know the structure of the set of minimal keys $\mathcal{K}$ of a database to be able to construct $\mathcal{C}$. The notation $\mathcal{C}_d(\mathcal{K})$ will be used, if it is necessary to emphasize that $\mathcal{C}_d$ is generated by $\mathcal{K}$.

Now, we are able to give the example of a database (set of minimal keys) for which there is no d-distance key at all.

EXAMPLE 2.4. Fix an element $a \in \Omega$ (that is, a column) and an integer $2 \leq k$. Define $\mathcal{K}$ as the family of all k-element sets ($\subset \Omega$) containing a.

Then, $C - \{a\}$ cannot contain any key, so the condition of Proposition 2.3 does not hold for any C if $2 \leq d$; there is no d-distance key in this database for $2 \leq d$.

EXAMPLE 2.5. Let $\mathcal{K}$ consist of all k-element subsets of Ω. In this case, as it can be easily seen, all subsets C of Ω of size at least $k + d - 1$ are d-distance keys.

In case of ordinary keys Theorem 1.2 completely describes the structure of them: a system of sets is isomorphic to the set of minimal keys for a certain database iff it is a Sperner family. The set of minimal d-distance keys is a Sperner family as well; however, the following example shows that an arbitrary Sperner family may not be the set of minimal d-distance keys of a database (for $d \geq 2$).

EXAMPLE 2.6. Let $\mathcal{C}$ be the family of all but one $k + d - 1$ element subsets of a large enough set. Then $\mathcal{C}$ is a Sperner family, but using Proposition 2.3 it can be easily seen that if all $C \in \mathcal{C}$ are d-distance keys of a database, the only missing $k + d - 1$ element subset of the underlying set must be a d-distance key as well. Therefore, $\mathcal{C}$ may not be the set of all minimal d-distance keys of any database.

The previous example, Example 2.5, suggests that the sizes of the members of $\mathcal{C}_d$, if they exist, do not exceed the sizes of the members of $\mathcal{K}$ by too much. We will show that this is not really true.

Let $\binom{\Omega}{\leq k}$ denote, as usual, the family of all subsets of Ω with size not exceeding k. For a Sperner family $\mathcal{K}$ (system of minimal keys of a database) and $d \geq 2$ such that $\mathcal{C}_d(\mathcal{K})$ is not empty, let

$$f_1(\mathcal{K}, d) = \min\{|C| : C \in \mathcal{C}_d(\mathcal{K})\},$$
$$f_2(\mathcal{K}, d) = \max\{|C| : C \in \mathcal{C}_d(\mathcal{K})\}.$$

Using the above definitions, we define

$$f_i(n, k, d) = \max\left\{ f_i(\mathcal{K}, d) : \mathcal{K} \subset \binom{\Omega}{\leq k},\ \mathcal{C}_d(\mathcal{K}) \neq \emptyset \right\}.$$

Now, we are ready to state the main theorem of this section.

THEOREM 2.7. *(See [10].)*

$$c_1k^d \leq f_1(n,k,d) \leq f_2(n,k,d) \leq c_2k^d$$

holds for $n_0(k,d) \leq n$ where c_1 and c_2 depend only on d.

PROOF. We will need a characterization of the *minimal* d-distance keys similar to Proposition 2.3. We say that the elements $a_1, \ldots, a_{d-1} \in \Omega$ *represent* $\mathcal{K}$—a family of subsets of Ω—if each $K \in \mathcal{K}$ contains one of the as. Then, Proposition 2.3 is equivalent to the following statement: a $\mathcal{C} \subset \Omega$ is a d-distance key of the database M with set minimal keys $\mathcal{K}$ iff no $d-1$ elements can represent the family $\{K : K \in \mathcal{K},\, K \subset C\}$. If C is minimal with respect to this property then no proper subset of C has the above property; that is, for all $a \in C$ the family $\{K : K \in \mathcal{K},\, K \subset C - \{a\}\}$ can be represented by $d-1$ elements.

PROPOSITION 2.8. *$C \in \mathcal{C}_d(\mathcal{K})$ iff $\{K : K \in \mathcal{K},\, K \subset C\}$ cannot be represented by $d-1$ elements, but for every $a \in C$ and properly chosen $a_1, \ldots, a_{d-1}$ the d elements $a, a_1, \ldots, a_{d-1}$ represent C.*

The *lower estimate* of the theorem will be given by a nonempty, inclusion-free family $\mathcal{K}$ consisting of some k-element sets which generates a $\mathcal{C}_d$ consisting of one member having size at least ck^d.

For a given integer $1 \leq i$ pick a subset A of the underlying set Ω of size $i+d-1$. Let $A_1, A_2, \ldots$ be all the $\binom{i+d-1}{i}$ i-element subsets of A and

$$\mathcal{K}(i) = \{A_1 \cup B_1, A_2 \cup B_2, \ldots\},$$

where $A, B_1, B_2, \ldots$ are pairwise disjoint subsets of Ω and $|B_1| = |B_2| = \cdots = k - i$. This can be carried out if $|\Omega|$ is big enough. We will show that the only member of $\mathcal{C}_d(\mathcal{K}(i))$ is $C = A \cup (\bigcup_i B_i)$. It is easy to see that $\mathcal{K}(i)$ cannot be represented by $d-1$ elements. On the other hand, if $a \in B_j$ for some j then the d-element set $\{a\} \cup (A - A_j)$ represents $\mathcal{K}$. If, however, $a \in A$, then any d-element set $D \subset A$ containing a represents $\mathcal{K}$, and therefore, C is really a member of $\mathcal{C}_d(\mathcal{K}(i))$. It is easy to see that there is no other member.

Choose $i = \lfloor k(1-1/d) \rfloor$. Then, the size of C becomes

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

For the *upper estimate* let us consider a $C \in \mathcal{C}_d(\mathcal{K})$ where $\mathcal{K} \subset \binom{\Omega}{\leq k}$. We will prove that $|C| \leq dk^d$. By Proposition 2.8, $C \in \mathcal{C}_d(\mathcal{K})$ iff $C \in \mathcal{C}_d(\mathcal{K}')$ where $\mathcal{K}' = \{K \in \mathcal{K} : K \subset C\}$ and here $\mathcal{K}' \subset \mathcal{K} \subset \binom{\Omega}{\leq k}$ is a Sperner family as well, and thus, we have that $C \in \mathcal{C}_d(\mathcal{K}')$. Therefore, it can be, and from now on will be, supposed that all members of $\mathcal{K}$ are subsets of C.

We may assume that the d-element sets $\{a, a_1, \ldots, a_{d-1}\}$ representing $\mathcal{K}$ defined in Proposition 2.8 are all subsets of C, and therefore, they will have union C. Denote the family of them by $\mathcal{D} = \{D = \{a, a_1, \ldots, a_{d-1}\} : a \in C, a_i \in C, D \text{ represents } \mathcal{K}\}$.

We now know that

$$\bigcup_{K \in \mathcal{K}} = \bigcup_{D \in \mathcal{D}} = C, \tag{2.1}$$

$$D \cap K \neq \emptyset, \qquad \text{for all } D \in \mathcal{D}, \quad K \in \mathcal{K}, \tag{2.2}$$

and $\mathcal{K}$ cannot be represented by a set with less than d element.

For a $I \subset C$ define the I-degree of $\mathcal{D}$ as the number of members of $\mathcal{D}$ containing I; that is,

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

We will prove that

$$\deg_I(\mathcal{D}) \leq k^{d-|I|}, \tag{2.3}$$

by induction on $j = d - |I|$.

The base case is $j = d - |I| = 1$; that is, $|I| = d - 1$. If all members of $\mathcal{K}$ meet I, then $\mathcal{K}$ can be represented by $d - 1$ elements, a contradiction. Therefore, there is a $K \in \mathcal{K}$ which is disjoint to I. By (2.2), all the sets D satisfying $I \subset D$ must intersect this K, and therefore, their number is $\leq |K| \leq k$. The base case is settled.

Now suppose that the statement is true for every $I \subset C$ such that $j \geq d - |I| \geq 1$ that is $d - 1 \geq |I| \geq d - j > 1$. Let $I^* \subset C$ such that $|I^*| = d - j - 1$; that is, $j + 1 = d - |I^*|$. There must exist a $K \in \mathcal{K}$, $K \cap I^* = \emptyset$; otherwise $\mathcal{K}$ is represented by less than d elements, a contradiction. Let $K = \{x_1, \dots, x_l\}$ where $l \leq k$. By (2.2), we have

$$\{D \in \mathcal{D} : I^* \subset D\} = \bigcup_{i=1}^{l} \{D \in \mathcal{D} : (I^* \cup \{x_i\}) \subset D\}. \tag{2.4}$$

The sizes of the sets on the right-hand side are $\deg_{I^* \cup \{x_i\}}(\mathcal{D})$ which are at most k^{d-j} by the induction hypothesis. Using (2.4),

$$\deg_{I^*}(\mathcal{D}) \leq l k^{d-j} \leq k^{d-j+1}$$

is obtained, finishing the induction proof of (2.3).

Finally, consider any $K = \{y_1, \dots, y_r\} \in \mathcal{K}$ where $r \leq k$. By (2.2), the families $\{D \in \mathcal{D} : y_i \in D\}$ cover $\mathcal{D}$. Apply (2.3) for $I = \{y_i\}$

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

This implies $|\mathcal{D}| \leq k^d$ and

$$\left| \bigcup_{D \in \mathcal{D}} D \right| \leq |\mathcal{D}| \, d \leq d k^d.$$

Application of (2.1) completes the proof: $|C| \leq d k^d$. ∎

We conclude this section with a few remarks.

REMARK 2.9. Consider the simplest case, when the probability of an incorrect data is so small that practically at most one data of an individual can be incorrect. In this case, $e = 1$, $d = 3$, and therefore, if the minimal keys have at most k elements, then the minimal one-error-correcting keys have at most $3k^3$ elements by the upper estimate of Theorem 2.7 and there exists a database with minimal keys of size k and only a single minimal one-error-correcting key of size roughly $(4/27)\, k^3$ by the lower estimate. So, even in this simple case, the error-correcting keys may be much larger than the keys.

REMARK 2.10. Although Theorem 2.7 determines the order of magnitude of $f_1(n, k, d)$, it does not give the exact value. We believe that the lower estimate is sharp; that is,

$$f_1(n, k, d) = \max_i \left\{ i + d - 1 + \binom{i + d - 1}{i} (k - i) \right\}$$

holds for $n_0(k, d) \leq n$.

REMARK 2.11. The following variation of the original problem sounds similar to the problem treated here, but it is actually very different. Suppose again that the data go through a noisy channel, where each data can be distorted with a small probability or due to any other reason we might have wrong data with a small probability. Try to define new attributes to make the effective keys for the erroneous database small.

3. FUNCTIONAL DEPENDENCIES AND KEYS IN RANDOM DATABASES

In most of the cases, investigating databases the system of functional dependencies is already known at setting up the structure, before collecting the data. In some situations, however, it might not be the case. Then, the database exists in the reality (usually not coded in the computer, yet) and the need is to find the rules in it, first of all, the functional dependencies.

In most cases, there is some *a priori* information on the relations of statistical nature. More precisely, a probability distribution is given, determining the probabilities of the possible databases M. Based on this information, we want to estimate the sizes of the functional dependencies. It is useful to have functional dependencies $A \to b$ with small As and it may be easier to find them if their approximate sizes are known in advance. The statistical rules could be quite complex, but it seems to be hard to obtain any theoretical result unless the choices of entries of the database are independent. Under this assumption, we determine the asymptotic probability of the event $A \to b$, depending on the size of A, for large databases. The main questions here are: what is the typical size of the minimal sets A such that $A \to b$ for a given attribute b and what is the typical size of the minimal keys.

The results of these investigations may be mostly found in [11–13]: the results in [11,12] are of rather probabilistic flavor; the proofs use the so-called Poisson approximation technique. The results and methods of [13] use the language of this paper and the proofs there—though rather technical—are more combinatorial. In this section, we will first give the simplest versions of these type results with their proof and then list the more sophisticated ones. The interested reader can find the proofs of them in [13].

We will need some additional notations. The elements of the set of attributes will be denoted by $\Omega = \{a_1, a_2, \ldots, a_n\}$. The set of the possible entries of the i^{th} column—the *domain* of a_i—will be denoted by $D(a_i)$. Thus, the data of one individual (row of the matrix) can be viewed as an element r of the direct product $D(a_1) \times D(a_2) \times \cdots \times D(a_n)$. Therefore, the whole database (or matrix) can be described by the *relation* $M \subseteq D(a_1) \times D(a_2) \times \cdots \times D(a_n)$. This definition coincides with the property of the database we assumed in Section 1: the data of two distinct individuals cannot be identical. However, in this section, to avoid further technical difficulties, we allow identical records. If $r = (e_1, e_2, \ldots, e_n) \in M$, then $r(i)$ will denote the i component of r; that is, $e_i \in D(a_i)$.

For the simplest case of a random database, suppose that all the domains contain exactly two elements; that is, $|D(a_i)| = 2$ holds for all $1 \le i \le n$. We may also suppose that $D(a_i) = \{0, 1\}$. Further, these values are chosen with equal $(1/2, 1/2)$ probabilities. All the entries (data) are chosen totally independently. Therefore, the probability of the choice of a given $0,1$-sequence of length n as a row $r \in M$ is $1/2^n$.

All of our results will be of asymptotic nature. We will suppose that the number n of attributes (columns) tends to infinity and the number of individuals (rows) is a function of $n : m(n)$ where $m(n)$ tends to infinity with n. The investigated quantity $|A|$ is also expressed as a function of n.

THEOREM 3.1. *Suppose that entries of the random database of $m(n)$ rows and n columns are chosen randomly from $\{0,1\}$: independently and with equal probabilities. Let $A_z \subset \Omega$ be of size $z = z(n)$, and suppose that $b \in \Omega - A_z$. Then, the probability of the event that b functionally depends on A satisfies (here and later on* log *means* $\log_2$ *always)*

$$P(A \to b) \to \begin{cases} 0, & \text{if } z(n) - 2\log m(n) \to -\infty, \\ \exp\left(-\dfrac{1}{2^{d+2}}\right), & \text{if } z(n) - 2\log m(n) \to d, \\ 1, & \text{if } z(n) - 2\log m(n) \to \infty. \end{cases} \tag{3.1}$$

PROOF. Let $z(n) = 2\log m(n) + d(n)$ and $s = 2^z$ denote the number of all the possible sequences on A. Then, $s = 2^z = 2^{2\log_2 m} \cdot 2^d = m^2 \cdot 2^d$ where 2^d will be denoted by c and so $s = cm^2$. We will calculate P (we have $m - i$ different rows). This can be calculated by adding up for all the possible partitions of the m rows into $m - i$ classes the number of ways one can assign a distribution of the actual values to the given partition.

Let us suppose that we partition the rows into classes of sizes $e_1 \geq e_2 \geq \cdots \geq e_{m-i}$, where the first k classes have cardinality at least 2, and the others have cardinality 1. It is easy to see that $k \leq i$ and $\sum_{j=1}^{k} e_j \leq 2i$ with equality only if $e_1 = e_2 = \cdots = e_k = 2$.

With this we have

$$\text{P (we have } m-i \text{ different rows and the partition they generate}$$
$$\text{has classes of sizes } e_1, e_2, \ldots, e_{m-i})$$
$$= \binom{m}{i+k} \frac{(i+k)!}{e_1! e_2! \cdots e_k!} \frac{1}{f_1!} \frac{1}{f_2!} \cdots \frac{1}{f_v!} \frac{s(s-1)\cdots(s-m+1+1)}{s^m},$$

where the first term is the number of ways to choose the complete set of rows belonging to a class of the partition with cardinality at least 2, the second term is the number of ways partitioning these rows into the actual parts, the f_is denote the number of classes of equal sizes, and the last term is the actual distribution of the values among the given positions and the whole divided by the number of possible cases.

All of the above terms tend to 0, except the case when $i = k$ and so $e_1 = e_2 = \cdots e_i = 2$, $e_{i+1} = \cdots = e_{m-i} = 1$, when the first term (with a factor $1/s^i$ "borrowed" from the last term) is

$$\frac{m^{i+k}}{(i+k)!\, s^i} \to \frac{1}{(2i)!\, c^i},$$

the second and third terms together are

$$\frac{(2i)!}{2^i\, i!},$$

and the last term (less a divisor factor s^i already considered in the first term) tends to $e^{-1/2c}$.

Putting all these together we get that

$$\text{P (we have } m-i \text{ different rows)} = \frac{1}{2^i\, i!\, c^i}\, e^{-1/2c}.$$

Now, let $\mathcal{D}_i$ denote the event that there are $m - i$ different partition classes. With this we have

$$\mathrm{P}(A \to b) = \sum_i \mathrm{P}(A \to b \mid \mathcal{D}_i) \cdot \mathrm{P}(\mathcal{D}_i) = \sum_i \mathrm{P}(\mathcal{D}_i) \cdot \frac{1}{2^i}, \tag{3.2}$$

and so

$$\mathrm{P}(a \to b) \to \sum_{i=0}^{\infty} \frac{1}{2^{2i}\, i!\, c^i}\, e^{-1/2c} = e^{1/4c} \,.\, e^{-1/2c} = e^{-1/4c} \quad \blacksquare$$

COROLLARY 3.2. *If the number of rows is a polynomial of* n, *that is,* $m(n) = n^h$, *then* (3.1) *holds for* $z(n) = 2h\log_2 n + d(n)$. *On the other hand, if* $m(n) = 2^{n/2 + \log_2 n}$, *then the probability of the event that there is any nontrivial functional dependency tends to* 0.

Theorem 3.1 can be generalized in two different ways (we will consider the straightforward most general form later).

(1) The number of the values of the attributes is not necessarily two and these numbers may be different from each other, but for every attribute each value is taken with the same equal probability.

(2) Every attribute takes the same values with the same probability distribution, but the number of the values may be bigger then two and the probability distribution does not have to be even.

THEOREM 3.3. *Suppose that entries of the random database of $m = m(n)$ rows and n columns are chosen randomly, that is independently, and the entries of the i^{th} column are the elements of $D(a_i)$ with equal probabilities. For an $A \subset U$ let $\sum_{a_i \in A} \log_2 |D(a_i)| - 2\log_2 m(n)$ be denoted by $d(n)$ and suppose that $b \in U - A$. Then, the probability of the event that b functionally depends on A satisfies*

$$\mathrm{P}(A \to b) \to \begin{cases} 0, & \textit{if } d(n) \to -\infty, \\ \exp\left(-\dfrac{|D(b)| - 1}{2^{d+1}|D(b)|}\right), & \textit{if } d(n) \to d, \\ 1, & \textit{if } d(n) \to \infty. \end{cases}$$

PROOF. It is the same as the proof of Theorem 3.1, only in (3.2) the term $1/2^i$ should be replaced by $1/|D(b)|$. ∎

For the second case, assume that every element of the database is chosen mutually independently from the same set $\{1, 2, \ldots, d\}$ with the same distribution $\{q_1, q_2, \ldots, q_d\}$. Let us denote the entropy of the distribution by

$$H_2 = -\log \sum_{i=1}^{d} q_i^2.$$

THEOREM 3.4. *Assume that the random database M has $m = m(n)$ rows and n columns. Let A_z denote a $z(n)$-element subset of Ω and b an element from Ω not in A_z. Then,*

$$\mathrm{P}(A_z \to b) \to \begin{cases} 0, & \textit{if } \dfrac{2\log m}{H_2} - z \to +\infty, \\ e^{2^{aH_2-1}(2^{-H_2}-1)}, & \textit{if } \dfrac{2\log m}{H_2} - z \to a, \\ 1, & \textit{if } \dfrac{2\log m}{H_2} - z \to -\infty, \end{cases}$$

as $n \to \infty$.

In short, in case we have a subset A of the attributes of size a bit larger than $(2\log m)/H_2 - z$, then for every attribute b (not in A) of the database we have: $A \to b$. In case B is another (finite) set of attributes of the database we have a rather similar result for the dependency $A \to B$ (only the $-H_2$ factor in the middle row, in the power of 2 should be multiplied by the cardinality of B). However, to assure that a set A is a key, we need the dependency $A \to \Omega$ where the size of Ω goes to infinity.

THEOREM 3.5. *Assume that the random database M has $m = m(n)$ rows and n columns. Let A_z denote a $z(n)$-element subset of Ω. Then,*

$$\mathrm{P}(A_z \textit{ is a key}) \to \begin{cases} 0, & \textit{if } \dfrac{2\log m}{H_2} - z \to +\infty, \\ e^{2^{aH_2-1}}, & \textit{if } \dfrac{2\log m}{H_2} - z \to a, \\ 1, & \textit{if } \dfrac{2\log m}{H_2} - z \to -\infty. \end{cases}$$

That is, it can be briefly said that the sets A of size somewhat larger than $2\log m/H_2$ are the keys with high probability.

Finally, let us state the most general case, where the attributes of the database take values from different sets, and the distribution of these values are not even (but, as mentioned at the beginning of this section, the values of the database are still mutually independent).

Assume that the elements of the a_i attribute of the database are chosen mutually independently from the set $D(a_i) = \{i_1, i_2, \ldots, i_{d_i}\}$ with the distribution $\kappa_i = \{q_{i1}, q_{i2}, \ldots, q_{id_i}\}$. Let us denote the entropy of the distribution $\kappa = \{q_1, q_2, \cdots, q_d\}$ by $H_2(\kappa) = -\log \sum_{i=1}^{d} q_i^2$. For the following (last) theorem, we need an additional assumption about the above distributions, namely:

$$\varepsilon \le q_{i1}, q_{i2}, \text{ hold for all } i \text{ with a fixed } \varepsilon, \quad \left(0 < \varepsilon \le \frac{1}{2}\right). \tag{3.3}$$

THEOREM 3.6. *Assume that the random database M has $m = m(n)$ rows and n columns where the entries of the j^{th} column can have d_j different values with probabilities $q_{j1}, \ldots, q_{jd_j}$, respectively, and all the entries are chosen totally independently. Assume that (3.3) holds as well. Let A_z denote a $z(n)$-element subset of Ω and b an element from Ω not in A_z. Then,*

$$\mathrm{P}(A_z \to b) \to \begin{cases} 0, & \text{if } 2\log m - \sum\limits_{i=1}^{z} H_2(\kappa_i) \to +\infty, \\ e^{-2^{a-1}\left(2^{-H_2(\kappa_b)}-1\right)}, & \text{if } 2\log m - \sum\limits_{i=1}^{z} H_2(\kappa_i) \to a, \\ 1, & \text{if } 2\log m - \sum\limits_{i=1}^{z} H_2(\kappa_i) \to -\infty, \end{cases}$$

as $n \to \infty$.

REFERENCES

1. E. Sperner, Ein Satz über Untermengen einer endlichen Menge, *Math. Z.* **27**, 544–548, (1928).
2. W.W. Armstrong, Dependency structures of data base relationship, In *Information Processing 74*, pp. 580–583, North-Holland, Amsterdam, (1974).
3. J. Demetrovics, On the equivalence of candidate keys with Sperner systems, *Acta Cybernet.* **4**, 247–252, (1979).
4. J. Demetrovics and G.O.H. Katona, Extremal combinatorial problems in relational data base, In *Fundamentals of Computation Theory '81*, LNCS 117, pp. 110–119, Springer, Berlin, (1981).
5. J. Demetrovics and Gy. Gyepesi, A note on minimal matrix representation of closure operations, *Combinatorica* **3**, 177–180, (1983).
6. J. Demetrovics, Z. Füredi and G.O.H. Katona, Minimum matrix representation of closure operations, *Discrete Applied Mathematics* **11**, 115–128, (1985).
7. F.E. Bennett and L. Wu, On minimum matrix representation of closure operations, *Discrete Applied Math.* **26**, 25–40, (1990).
8. K. Tichler, Minimum matrix representation of some key systems, In *Foundations of Information and Knowledge Systems*, LNCS 1762, (Edited by K.-D. Schewe and B. Thalheim), pp. 275–287, Springer, Berlin, (2000).
9. G.O.H. Katona, Combinatorial and algebraic results for database relations, In *Database Theory—ICDT '92*, LNCS 646, (Edited by J. Biskup and R. Hull), pp. 1–20, Springer, Berlin, (1992).
10. J. Demetrovics, G.O.H. Katona and D. Miklós, Error-correcting keys in relational databases, In *Foundations of Information and Knowledge Systems*, LNCS 1762, (Edited by K.-D. Schewe and B. Thalheim), pp. 88–93, Springer, Berlin, (2000).
11. J. Demetrovics, G.O.H. Katona, D. Miklós, O. Seleznjev and B. Thalheim, The average length of keys and functional dependencies in (random) databases, In *Database Theory—ICDT'95*, LNCS 893, (Edited by G. Gottlob and M.Y. Vardi), pp. 266–279, Springer, Berlin, (1995).
12. J. Demetrovics, G.O.H. Katona, D. Miklós, O. Seleznjev and B. Thalheim, Asymptotic properties of keys and functional dependencies in random databases, *Theoretical Computer Science* **190**, 151–166, (1998).
13. J. Demetrovics, G.O.H. Katona, D. Miklós, O. Seleznjev and B. Thalheim, Functional dependencies in random databases, *Studia Sci. Math. Hung.* **34**, 127–140, (1998).