Studia Scientiarum Mathematicarum Hungarica 34 (1998), 127–140

FUNCTIONAL DEPENDENCIES IN RANDOM DATABASES

J. DEMETROVICS, G. O. H. KATONA, D. MIKLÓS
O. SELEZNJEV and B. THALHEIM

To the memory of Alfréd Rényi

1. Introduction

A database can be considered as a matrix, where the rows contain the data of one individual (object, etc.) and the columns contain the data of the same type: last name, first name, date of birth, etc. The types of data are called *attributes*. These data are sometimes logically dependent. Consider the following example, where the attributes are the last name (denoted by a), the first name (b), the year of the birth (c), the month of the birth (d), the day of the birth (e), the age in years (f), the age in months (g) and the age in days (h). It is obvious that c determines f. On the other hand, the pair $\{c,d\}$ determines both f and g, finally the set $\{c,d,e\}$ determines all of f, g and h.

This is formalized in the following way. Let R be an $m\times n$ matrix with different rows and Ω denote the set of its columns, that is, $|\Omega|=n$. Suppose that $A\subseteqq\Omega, b\in\Omega$. We say that b *functionally depends* on A and write $A\to b$ if R contains no two rows containing equal entries in the columns belonging to A and different entries in b.

In most of the database theory it is supposed that the *functional dependencies* $A\to b$ are a priori known by the logic of the data, as in the above example. Our way of looking at the situation is different. We suppose that we have to find the functional dependencies in a large database (both m and n are large). If nothing is known about R, it is natural to assume that the entries are independently chosen. The question is: what the typical size of the minimal sets A such that $A\to b$ is.

Thus the first mathematical question is the following. Choose the entries of the matrix R totally independently, following the probability distribution $(q_1,\dots,q_d)$. What is the minimum size l of A such that $A\to b$ holds with

---

1991 *Mathematics Subject Classification.* Primary 68P15, 68R05, 15A52.
Key words and phrases. Functional dependency, random database, random matrix, sieve method.

Supported by the Hungarian Foundation for Scientific Research (OTKA) Grant T 016524, T 016389 and the German Natural Science Research Council Contract BB-II-B1-3141-211(94).

high probability for any set $A \subset \Omega$, $|A| \geqq l$ and any column $b \in \Omega$? The answer is

$$\frac{2m}{-\log_2(q_1^2 + \cdots + q_d^2)},$$

as it is given precisely in Corollary 1. Theorem 2 generalizes this result for the case when the entries have different distributions in the different columns.

Section 2 develops a sieve method for estimating the probability of the event that all the outcomes of a many times repeated experiment are different. This result is applied for the rows of a random matrix in Section 3: Theorem 1 determines the asymptotic probability of the event that the rows of the random matrix are different. This theorem is of crucial importance in proving Theorem 2.

If A is larger than the above critical size then $A \to b$ holds with high probability for any given b. However, it will not be true for each element b of a large set Ω. Theorem 3 determines the asymptotic size of the A's satisfying $A \to \Omega$.

The method of the present paper is combinatorial. Paper [2] of the same authors contains similar (but not identical) results. The method of that paper is probabilistic, and uses the so-called Poisson approximation technique (Stein–Chen method, see [1]).

2. A sequence of experiments with different outcomes

We may obtain a counterexample for $A \to b$ if the entries of two rows in the submatrix determined by A are equal. So the critical situation is when all these rows are different. This is why this section is devoted to the probability of the event that all the outcomes of a repeated experiment are different.

Let $E_1, \dots, E_s$ be mutually exclusive events with respective probabilities $p_1, \dots, p_s$, where $\sum_{i=1}^{s} p_i = 1$. The distribution is denoted by μ. Choose independently, m times, from these events with this distribution. That is, $\mathbf{P}(\xi_i = E_j) = p_j$ is supposed for all $1 \leqq i \leqq m$ and $1 \leqq j \leqq s$. Moreover, the ξ's are totally independent. Let $\mathbf{P}(\mu, m)$ be the probability of the event that $\xi_1, \dots, \xi_m$ are all different.

Lemma 3 is the main result of the section giving good estimates on $\mathbf{P}(\mu, m)$.

For an arbitrary sequence of outcomes a trivial graph can be defined. The outcomes are the vertices and two vertices are adjacent if they have the same value. This is why we consider the following graphs. Our goal is actually to estimate the probability that this graph is empty.

The vertex-disjoint union of complete graphs with $m_1, \dots, m_r$, resp., vertices is denoted by $G(m_1, \dots, m_r)$. A graph consisting of vertex-disjoint edges is a *matching*. The vertex-disjoint union of a matching and a path consisting of two edges is called a *V-matching*. Finally, the vertex-disjoint union of a matching and a path consisting of three edges is an *N-matching*.

LEMMA 1. *Let $m_1,\dots,m_r$ $(0\leqq r)$ be non-negative integers. Then*

$$\sum_{\text{matching of } j \text{ edges}} (-1)^j + \sum_{\text{V-matching}} 1 + \sum_{\text{N-matching}} 1 \geqq 0, \tag{1}$$

where the matchings, V-matchings and N-matchings are arbitrary subgraphs of $G(m_1,m_2,\dots,m_r)$.

PROOF. $2\leqq m_i$ $(1\leqq i\leqq r)$ can be supposed. Two cases will be distinguished.

(i) $m_1=m_2=\dots=m_r=2$. The number of matchings of j edges in $G(2,\dots,2)$ is $\binom{r}{j}$ therefore the left-hand side of (1) is

$$\sum_{j=0}^{r}\binom{r}{j}(-1)^j,$$

which is 0 if $0<r$ and 1 if $r=0$.

(ii) One of the m's >2. An injection will be given from the set of all negative terms into a set of some positive terms in (1). Actually the injection will be defined on sets of subgraphs of $G(m_1,m_2,\dots,m_r)$. A negative term is generated by a matching M of j edges, where j is odd. Suppose that there are at least two edges of M in one of the components of $G(m_1,m_2,\dots,m_r)$. Join any two endpoints of these two edges by a new edge. The injection assigns this N-matching to M.

Suppose that no component of $G(m_1,m_2,\dots,m_r)$ contains at least two edges of M but there is a component with at least 3 vertices and containing exactly one edge of M. Then this edge will be replaced by a pair of adjacent edges in the same component. As the number of such pairs is $\geqq$ the number of edges in a complete graph on $\geqq 3$ vertices, this can be defined as a part of an injection. (Actually the assignment can be made in such a way that the pair contains the edge, however, this fact is not needed and its proof is somewhat more difficult.)

The only remaining case is when all components with at least three vertices are disjoint to M. Then add an edge of this component to M. This matching contains an even number of edges therefore it generates 1 in (1).

It is easy to see that the function defined above is an injection and it assigns positive terms to negative terms, proving (1). □

LEMMA 2. *Let $m_1,\dots,m_r$ $(0<r)$ be non-negative integers, at least one of them is $\geqq 2$. Then*

$$\sum_{\text{matching of } j \text{ edges}} (-1)^j + \sum_{\text{V-matching of}} (-1) + \sum_{\text{N-matching}} (-1) \leqq 0, \tag{2}$$

where the matchings, V-matchings and N-matchings are subgraphs of $G(m_1, m_2, \ldots, m_r)$.

PROOF. The proof is analogous to the previous one. The only difference is that here the injection assigns a negative term to a positive term generated by a matching of even number of edges. □

LEMMA 3.

$$1+\sum_{j=1}^{\lfloor \frac{m}{2}\rfloor} \frac{(-1)^j}{j!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2}\Bigl(\sum_{i=1}^{s} p_i^2\Bigr)^j -$$

$$-\sum_{j=0}^{\lfloor \frac{m-3}{2}\rfloor} \frac{1}{j!}\binom{m}{3}\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2}\Bigl(\sum_{i=1}^{s} p_i^3\Bigr)\Bigl(\sum_{i=1}^{s} p_i^2\Bigr)^j -$$

$$-\sum_{j=0}^{\lfloor \frac{m-4}{2}\rfloor} \frac{1}{j!}\binom{m}{4}\binom{m-4}{2}\binom{m-6}{2}\cdots\binom{m-2j-2}{2}\Bigl(\sum_{i=1}^{s} p_i^4\Bigr)\Bigl(\sum_{i=1}^{s} p_i^2\Bigr)^j \leqq$$

$$\leqq \mathbf{P}(\mu, m) \leqq \tag{3}$$

$$\leqq 1+\sum_{j=1}^{\lfloor \frac{m}{2}\rfloor} \frac{(-1)^j}{j!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2}\Bigl(\sum_{i=1}^{s} p_i^2\Bigr)^j +$$

$$+\sum_{j=0}^{\lfloor \frac{m-3}{2}\rfloor} \frac{1}{j!}\binom{m}{3}\binom{m-3}{2}\binom{m-5}{2}\cdots\binom{m-2j-1}{2}\Bigl(\sum_{i=1}^{s} p_i^3\Bigr)\Bigl(\sum_{i=1}^{s} p_i^2\Bigr)^j +$$

$$+\sum_{j=0}^{\lfloor \frac{m-4}{2}\rfloor} \frac{1}{j!}\binom{m}{4}\binom{m-4}{2}\binom{m-6}{2}\cdots\binom{m-2j-2}{2}\Bigl(\sum_{i=1}^{s} p_i^4\Bigr)\Bigl(\sum_{i=1}^{s} p_i^2\Bigr)^j.$$

PROOF. $\mathbf{P}(\mu, m)$ is the probability of the event that $\xi_1, \xi_2, \ldots, \xi_m$ are all different, that is, one minus the sum of the probabilities

$$\mathbf{P}(\xi_u = E_{v_k} \ \text{ if } u \in C_k), \tag{4}$$

where $C_1, C_2, \ldots, C_t$ is a partition of $\{1, 2, \ldots, m\}$ with at least one C having more than one element, and $v_1, v_2, \ldots, v_t$ are different elements of $\{1, 2, \ldots, s\}$. Such partitions will be called *non-elementary*.

$\mathbf{P}(\mu, m)$ contains the probabilities in (4) with zero weight, therefore if they are counted with the weight given in (1) then it leads to an upper estimate. Consider the sum

$$\sum_{\text{partition}}\Biggl(\sum_{\text{matching of } j \text{ edges}} (-1)^j + \sum_{\text{V-matching}} 1 + \sum_{\text{N-matching}} 1\Biggr) \times$$

$$\times \mathbf{P}(\xi_u = E_{v_k} \text{ if } u \in C_k). \tag{5}$$

If the partition is the elementary one, then the inner sums are empty with one exception, the empty matching. This leads to the sum of the probabilities where all ξ's are different. Therefore (5) is the sum in which the probabilities of the events, where all ξ's are different stand with weight 1, while the other probabilities stand with a non-negative weight. Consequently, (5) is an upper estimate on $\mathbf{P}(\mu, m)$.

Change the order of sums in (5).

$$\sum_{\text{matching of } j \text{ edges}} (-1)^j \sum_{\text{partition}} \mathbf{P}(\xi_u = E_{v_k} \text{ if } u \in C_k) +$$

$$+ \sum_{\text{V-matching}} \sum_{\text{partition}} \mathbf{P}(\xi_u = E_{v_k} \text{ if } u \in C_k) + \tag{6}$$

$$+ \sum_{\text{N-matching}} \sum_{\text{partition}} \mathbf{P}(\xi_u = E_{v_k} \text{ if } u \in C_k),$$

where those partitions are taken for which the given matching is a subgraph of the graph generated by the partition. Consider

$$\sum_{\text{partition}} \mathbf{P}(\xi_u = E_{v_k} \text{ if } u \in C_k)$$

for a given matching of j edges. This is nothing else but the probability of the event that the ξ's adjacent in the matching are equal:

$$\Big(\sum_{i=1}^{s} p_i^2\Big)^j.$$

The number of matchings with j edges is

$$\frac{1}{j!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2j+2}{2}.$$

This gives the fifth row of (3). The second and third rows of (6) lead, in a similar manner, to the sixth and seventh rows of (3), resp.

The lower estimate is proved in the same way. □

3. Random matrix with different rows

The Lemma 3 will be used for random matrices. Let R be a random matrix with m rows and z columns, where the entries of the jth column can have d_j different values with probabilities $q_{j1}, \dots, q_{jd_j}$, respectively. All the entries are chosen totally independently. Then the probability of the occurrence of a certain row in R is $q_{1i_1} q_{2i_2} \cdots q_{zi_z}$, where i_j is arbitrary between 1 and d_j. The probability distribution of these sequences will be denoted by π_z. The following trivial observation will be used later.

LEMMA 4. *If* $m \leqq m'$ *then* $\mathbf{P}(\pi_z, m) \geqq \mathbf{P}(\pi_z, m')$.

We want to study the probability of the event that the rows of the above matrix are different. Therefore the probabilities $q_{1i_1} q_{2i_2} \cdots q_{zi_z}$ will be taken as p's in Lemma 3. Consider $\sum_{i=1}^{s} p_i^k$ for these probabilities:

$$(7) \qquad \sum_{1 \leqq i_1 \leqq d_1, \ldots, 1 \leqq i_z \leqq d_z} (q_{1i_1} q_{2i_2} \cdots q_{zi_z})^k = \sum_{1 \leqq i_1 \leqq d_1, \ldots, 1 \leqq i_z \leqq d_z} q_{1i_1}^k q_{2i_2}^k \cdots q_{zi_z}^k = \prod_{i=1}^{z} (q_{i1}^k + \cdots + q_{id_i}^k).$$

Our investigations will be of asymptotic nature. From now on it is supposed that m tends to the infinity and the other parameters depend on m: $z = z(m), d_i = d_i(m), q_{ij} = q_{ij}(m)$. Our asymptotic assumption on them will be such that the first non-trivial term in the Lemma 3, that is,

$$(8) \qquad m^2 \sum_{i=1}^{s} p_i^2 = m^2 \prod_{i=1}^{z} (q_{i1}^2 + \cdots + q_{id_i}^2)$$

tends to a non-zero constant. It will be done in a logarithmic way, therefore the quantities $\log(q_{i1}^2 + \cdots + q_{id_i}^2)$ will play an important role.(log will always mean log of base 2.) Denote the distribution $(q_{i1}, \ldots, q_{id_i})$ by κ_i. Rényi [3] introduced the so-called entropy of order α. For $\alpha = 2$ it is $H_2(\kappa) = -\log(q_1^2 + \cdots + q_d^2)$ if $\kappa = (q_1, \ldots, q_d)$.

LEMMA 5. *If*

$$(9) \qquad 2 \log m - \sum_{i=1}^{z} H_2(\kappa_i) \to a$$

when $m \to \infty$ *then*

$$(10) \qquad 1 + \sum_{j=1}^{\lfloor \frac{m}{2} \rfloor} \frac{(-1)^j}{j!} \binom{m}{2} \binom{m-2}{2} \cdots \binom{m-2j+2}{2} \Big(\sum_{i=1}^{s} p_i^2 \Big)^j$$

tends to

$$e^{-2^{a-1}}$$

for the distribution π_z.

PROOF. Consider the limit of one term for a fixed j.

$$\binom{m}{2} \binom{m-2}{2} \cdots \binom{m-2j+2}{2}$$

can be replaced by

$$\frac{m^{2j}}{2^j}.$$

On the other hand

$$\left(\sum_{i=1}^{s} p_i^2\right)^j = 2^{-j\sum_{i=1}^{z} H_2(\kappa_i)}$$

follows by the definition of the entropy of order 2 and (7). Therefore the limit of the jth term in (10) is the same as the limit of

$$\text{(11)} \qquad \frac{(-1)^j}{j!} 2^{j(2\log m - \sum_{i=1}^{z} H_2(\kappa_i) - 1)},$$

that is,

$$\frac{(-1)^j}{j!} 2^{j(a-1)}.$$

(9) implies that the sum of (11) and therefore (10) are uniformly convergent, hence the limit of (10) is equal to the infinite sum of the limits of its terms, that is,

$$\sum_{j=0}^{\infty} \frac{(-1)^j}{j!} 2^{j(a-1)} = e^{-2^{a-1}}. \qquad \square$$

We want to show that the other terms in the lower and upper estimates of (3) tend to zero under condition (9). Before proving that some other lemmas are needed.

LEMMA 6. *If* $\kappa = (q_1, \dots, q_d)$ *is a probability distribution, where* $\varepsilon \leqq q_1, q_2$ $(0 < \varepsilon \leqq \frac{1}{2})$ *then*

$$\text{(12)} \qquad \frac{\left(\sum\limits_{i=1}^{d} q_i^3\right)^2}{\left(\sum\limits_{i=1}^{d} q_i^2\right)^3} \leqq 1 - 4\varepsilon^6.$$

PROOF. Consider the difference of the denominator and the numerator:

$$\sum_{i=1}^{d} q_i^6 + 3\sum_{i<j} q_i^4 q_j^2 + 3\sum_{i<j} q_i^2 q_j^4 + 6\sum_{i<j<k} q_i^2 q_j^2 q_k^2 - \left(\sum_{i=1}^{d} q_i^6 + 2\sum_{i<j} q_i^3 q_j^3\right) \geqq$$

$$\geqq \left(\sum_{i<j} q_i^4 q_j^2 + \sum_{i<j} q_i^2 q_j^4 - 2\sum_{i<j} q_i^3 q_j^3\right) + 2\sum_{i<j}(q_i^4 q_j^2 + q_i^2 q_j^4) =$$

$$=\sum_{i<j} q_i^2 q_j^2 (q_i^2+q_j^2-2q_iq_j)+2\sum_{i<j}(q_i^4q_j^2+q_i^2q_j^4)\geqq$$

$$\geqq 2q_1^4q_2^2+2q_1^2q_2^4)\geqq 4\varepsilon^6.$$

Using the fact that the denominator is at most 1, (12) easily follows. □

LEMMA 7. *If* $\kappa=(q_1,\ldots,q_d)$ *is a probability distribution, where* $\varepsilon\leqq q_1,q_2$ $(0<\varepsilon\leqq\frac{1}{2})$ *then*

$$\frac{\sum\limits_{i=1}^{d} q_i^4}{\left(\sum\limits_{i=1}^{d} q_i^2\right)^2}\leqq 1-2\varepsilon^4. \tag{13}$$

PROOF. The proof is similar but easier than the previous one:

$$\sum_{i=1}^{d} q_i^4+2\sum_{i<j} q_i^2q_j^2-\sum_{i=1}^{d} q_i^4=$$

$$=2\sum_{i<j} q_i^2q_j^2\geqq 2q_1^2q_2^2\geqq 2\varepsilon^4.$$

□

LEMMA 8. *If* (9) *and*

$$\varepsilon\leqq q_{i1},q_{i2} \text{ hold for all } i \text{ with a fixed } \varepsilon \quad \left(0<\varepsilon\leqq\frac{1}{2}\right) \tag{14}$$

then the second and third rows of (3) *tend to zero.*

PROOF. The jth term of the second row of (3) can be upperbounded by

$$\left(m^3\prod_{i=1}^{z}(q_{i1}^3+\cdots+q_{id_i}^3)\right)\left(m^2\prod_{i=1}^{z}(q_{i1}^2+\cdots+q_{id_i}^2)\right)^j. \tag{15}$$

The second factor tends to

$$\frac{2^{j(a-1)}}{j!}$$

as we have seen in the proof of Lemma 5. (9) implies

$$m^2\prod_{i=1}^{z}(q_{i1}^2+\cdots+q_{id_i}^2)\to 2^a \tag{16}$$

therefore the first factor of (15) can be expressed as

$$m^3 \prod_{i=1}^{z} (q_{i1}^3 + \cdots + q_{id_i}^3) = \left(m^2 \prod_{i=1}^{z} \Big(\sum_{j=1}^{d_i} q_{ij}^2 \Big) \right)^{\frac{3}{2}} \prod_{i=1}^{z} \frac{\sum_{j=1}^{d_i} q_{ij}^3}{\Big(\sum_{j=1}^{d_i} q_{ij}^2 \Big)^{\frac{3}{2}}}. \tag{17}$$

The first factor of (17) tends to $2^{\frac{3}{2}a}$ while Lemma 6 gives the upper bound $(1-4\varepsilon^6)^{\frac{z}{2}}$ for the second factor.

The conditions of the lemma imply $-\log(2\varepsilon^2) \geqq H_2(\kappa_i)$ thus (9) results in $z \to \infty$ when $m \to \infty$. $(1-4\varepsilon^6)^{\frac{z}{2}}$ and consequently (17) tend to zero. By the uniform convergence, the infinite sum of (15) and the second row of (3) also tend to zero.

The convergence of the third row can be proved in the same way, using Lemma 7. □

THEOREM 1. *Let R be a random matrix with m rows and z columns, where the entries of the jth column can have d_j different values with probabilities $q_{j1}, \ldots, q_{jd_j}$, respectively. All the entries are chosen totally independently. Suppose that* (14) *holds. Then the probability of the event that the rows of R are all different satisfies*

$$\mathbf{P}(\pi_z, m) \to \begin{cases} 0, & \text{if } 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to +\infty, \\ e^{-2^{a-1}}, & \text{if } 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to a, \\ 1, & \text{if } 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to -\infty. \end{cases}$$

PROOF. The middle row of the statement follows by Lemmas 3 and 8. The first and third rows are consequences of Lemma 4. □

In [4] Rényi proved a theorem on random matrices in connection with search theory (see also [5] and [6]). It is basically equivalent to the special case of the above theorem when κ_i's are the same. His method was different.

REMARK. The condition that each distribution contains two "large" probabilities ($\varepsilon \leqq q_{i1}, q_{i2}$) was important in the proof. This is shown by the following example. Let $\kappa_i = (\frac{1}{2}, \frac{1}{2m}, \ldots, \frac{1}{2m})$. Then the left-hand side of (12) is

$$\left(1 - \frac{2m}{(m+1)^2}\right)\left(1 - \frac{m-1}{m^2+m}\right),$$

which is not bounded from 1. Take $z = \log m$. As $H_2(\kappa_i) \to 2$, (9) holds with zero. However, the second factor of (17) does not tend to zero. (3) cannot be used.

Another example is $\kappa_i = (\frac{1}{m}, \frac{1}{m^3}, \dots, \frac{1}{m^3})$. We do not know, however, if the statement of the theorem holds for these and similar distributions. □

4. Typical sizes of functional dependencies and minimal keys

Let $\mathbf{P}(\mu, m, k)$ denote the probability of the event that exactly k pairs of $\xi_1, \dots, \xi_m$ are equal to each other and all other pairs are different. (More precisely: there are $2k$ distinct indices $i_1, \dots, i_k$ and $j_1, \dots, j_k$ such that $\xi_{i_l} = \xi_{j_l}$ for all $1 \leqq l \leqq k$, but $\xi_{i_l} \neq \xi_{j_m}$ for all $l \neq m$, $\xi_i \neq \xi_{i_l}$ and $\xi_i \neq \xi_{j_l}$ if $i \notin \{i_1, \dots, i_k, j_1, \dots, j_k\}$.)

LEMMA 9. *Suppose that k is fixed, m tends to infinity and* (14) *holds. Then*

$$\mathbf{P}(\pi_z, m, k) \to \frac{1}{k!} 2^{k(a-1)} e^{-2^{a-1}}, \quad \textit{if } 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to a.$$

PROOF. There are

$$\frac{1}{k!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2k+2}{2}$$

ways to choose the set $\{i_1, \dots, i_k, j_1, \dots, j_k\}$. Suppose that $i_1 = 1, j_1 = 2, \dots, i_k = 2k-1, j_k = 2k$ and determine the probability of the event that $\xi_1 = \xi_2, \dots, \xi_{2k-1} = \xi_{2k}$. The probabilities for the other choices of pairs will be the same. It is easy to see that

$$\mathbf{P}(\xi_1 = \xi_2) = \prod_{i=1}^{z} \sum_{j=1}^{d_i} q_{ij}^2.$$

We need the kth power of this expression. Finally, $\xi_1, \xi_3, \dots, \xi_{2k-1}, \xi_{2k+1}, \xi_{2k+2}, \dots, \xi_m$ must be all different. The probability of this event is $\mathbf{P}(\pi_z, m-k)$.

$$\begin{aligned} &\mathbf{P}(\pi_z, m, k) \\ &= \frac{1}{k!}\binom{m}{2}\binom{m-2}{2}\cdots\binom{m-2k+2}{2}\left(\prod_{i=1}^{z}\Big(\sum_{j=1}^{d_i} q_{ij}^2\Big)\right)^k \mathbf{P}(\pi_z, m-k). \end{aligned} \tag{18}$$

The last factor is asymptotically equal to $\mathbf{P}(\pi_z, m)$ since $\log m - \log(m-k) \to 0$. Therefore Theorem 1 gives its limit. The limit of the product of the other factors of (18) was determined in the proof of Lemma 5:

$$\frac{1}{k!} 2^{k(a-1)}.$$

□

LEMMA 10. *If* (9) *and* (14) *hold then the probability of the event that there are three equal* ξ*'s tends to zero.*

PROOF. The sum of the probabilities in Lemma 9 tends to 1. □

Let Ω denote the set of columns of the matrix R. Suppose $A \subset \Omega, b \in \Omega - A$. We say that b *functionally depends* on A if R contains no two rows equal in the columns belonging to A and different in b. In notation: $A \to b$. For sake of simplicity b is supposed to be the bth column.

THEOREM 2. *Let R be a random matrix with m rows and $n = n(m)$ columns with the distribution described above ((14) holds, again). Suppose that A_z is a set of $z = z(m)$ columns of R and b is a column not in A_z.*

$$\mathbf{P}(A_z \to b, m) \to \begin{cases} 0, & \textit{if } \ 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to +\infty, \\ e^{2^{a-1}(2^{-H_2(\kappa_b)}-1)}, & \textit{if } \ 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to a, \\ 1, & \textit{if } \ 2\log m - \sum_{i=1}^{z} H_2(\kappa_i) \to -\infty. \end{cases}$$

PROOF. Consider the restrictions of the rows of R within A_z. These rows of length z define a random partition $\gamma = (m_1, \ldots, m_r)$ of m, where one class consists of the equal rows. Suppose $(m_1 \geqq \ldots \geqq m_r)$. Start with the well known equation

(19)
$$\mathbf{P}(A_z \to b, m) = \sum_{m_1,\ldots,m_r} \mathbf{P}(A_z \to b|\gamma = (m_1, \ldots, m_r))\mathbf{P}(\gamma = (m_1, \ldots, m_r)).$$

The right-hand side of (19) will be divided into two parts: (i) $m_1 \leqq 2$, (ii) $m_1 \geqq 3$. For case (ii) the following trivial inequality is needed:

$$\begin{aligned} &\sum_{m_1 \geqq 3, m_2,\ldots,m_r} \mathbf{P}(A_z \to b, m|\gamma = (m_1, \ldots, m_r))\mathbf{P}(\gamma = (m_1, \ldots, m_r)) \\ &\leqq \sum_{m_1 \geqq 3, m_2,\ldots,m_r} \mathbf{P}(\gamma = (m_1, \ldots, m_r)) = \mathbf{P}(\text{there are 3 equal } \xi\text{'s}). \end{aligned} \tag{20}$$

The last quantity tends to zero under condition (9) therefore case (i) should only be considered. More precisely, if (9) holds then the limit of $\mathbf{P}(A_z \to b, m)$ is equal to the limit of

$$\sum_{m_1 \leqq 2, m_2,\ldots,m_r} \mathbf{P}(A_z \to b, m|\gamma = (m_1, \ldots, m_r))\mathbf{P}(\gamma = (m_1, \ldots, m_r)).$$

This expression can be rewritten in the form

$$\begin{aligned}
&\sum_k \mathbf{P}(A_z \to b, m | \gamma = (2,\dots,1) \text{ (the number of 2's is } k)) \\
(21) \qquad &\times \mathbf{P}(\gamma = (2,\dots,1) \text{ (the number of 2's is } k)) \\
&= \sum_k \mathbf{P}(A_z \to b, m | \gamma = (2,\dots,1) \text{ (the number of 2's is } k))\mathbf{P}(\pi_z, m, k).
\end{aligned}$$

Here

$$\mathbf{P}(A_z \to b, m | \gamma = (2,\dots,1) \text{ (the number of 2's is } k)) = \Big(\sum_{j=1}^{d_b} q_{bj}^2\Big)^k = 2^{-kH_2(\kappa_b)}.$$

On the other hand, the limit of $\mathbf{P}(\pi_z, m, k)$ is given by Lemma 9. Therefore the limit of (21) is

$$e^{-2^{a-1}} \sum_{k=0}^{\infty} \frac{1}{k!} \left(2^{(a-1)-H_2(\kappa_b)}\right)^k = e^{2^{a-1}(2^{-H_2(\kappa_b)}-1)}.$$

The middle row of the statement is proved. The first and third rows are consequences of the inequality $\mathbf{P}(A_z \to b, m) \geqq \mathbf{P}(A_z \to b, m')$ for $m \leqq m'$. □

COROLLARY 1. *Let R be a random matrix with m rows and $n = n(m)$ columns, where the entries are chosen totally independently with probabilities $q_1, \dots, q_d$. Suppose that A_z is a set of $z = z(m)$ columns of R and b is a column not in A_z. Use the notation $H_2 = -\log \sum_{i=1}^{d} q_i^2$. Then*

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

The main content of the latter statement is that if A is a set of columns of size definitely larger than $\frac{2\log m}{H_2}$, then $A \to b$ holds with high probability for any b.

We say, in general, that B *functionally depends* on A and write $A \to B(A, B \subseteqq \Omega)$ if $A \to b$ holds for each element b of B. Theorem 2 can be easily generalized for this case. We only have to imagine the set of columns in B as one column. It is worth supposing that $A \cap B = \emptyset$. Then $H_2(\kappa_b)$ can be replaced by $H_2(\kappa_B) = \sum_{b \in B} H_2(\kappa_b)$.

Let us turn back to the case when κ_i does not depend on i. If the size of B is finite, say u, then the Consequence can be generalized for $A \to B$, only $-H_2$ should be multiplied by u. However, if $|B|$ tends to infinity, then the middle probability becomes simply $e^{-2^{aH_2-1}}$.

We say that $A \subseteq \Omega$ is a *key* if $A \to \Omega$ (or equivalently $A \to \Omega - A$) holds. A is a *minimal key* if it is a key and no proper subset is a key. The above reasoning proves the following statement.

THEOREM 3. *Let R be a random matrix with m rows and $n = n(m)$ columns, where the entries are chosen totally independently following the distribution κ. Suppose that $n - \frac{2\log m}{H_2(\kappa)}$ tends to infinity and A_z is a set of columns of R. Then*

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

It can be briefly said that the sets A of size somewhat larger than $\frac{2\log m}{H_2}$ are keys with high probability.

REFERENCES

[1] BARBOUR, A. D., HOLST L. and JANSON, S., *Poisson approximation*, Oxford Studies in Probability, **2**, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1992. *MR* **93g**:60043

[2] DEMETROVICS, J., KATONA, G. O. H., MIKLÓS, D., SELEZNJEV, O. and THALHEIM, B., Asymptotic properties of keys and functional dependencies in random databases, *Theoret. Comput. Sci.* **190** (1998), 151–166.

[3] RÉNYI, A., A few fundamental problems of information theory, *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* **10** (1960), 251–282 (in Hungarian). *MR* **26** #1218

[4] RÉNYI, A., A general method for proving theorems in probability theory and some applications, *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* **11** (1961), 79–105 (in Hungarian).

[5] RÉNYI, A., Statistical laws of accumulation of information, *Bull. Inst. Internat. Statist.* **39** (1962), livraison 2, 311–316. *MR* **28** #5845

[6] RÉNYI, A., On the statistical regularities of information accumulation, *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.* **12** (1962), 15–33 (in Hungarian). *MR* **28** #1023

(Received June 29, 1998)

J. Demetrovics

MTA SZÁMÍTÁSTECHNIKAI ÉS AUTOMATIZÁLÁSI KUTATÓ INTÉZETE
KENDE U. 13–17
H–1111 BUDAPEST
HUNGARY

dj@ilab.sztaki.hu

O. Seleznjev

FACULTY OF MATHEMATICS AND MECHANICS
MOSCOW STATE UNIVERSITY
RU–119 899 MOSCOW
RUSSIA

seleznev@top.rector.msu.su

G. O. H. Katona and D. Miklós

MTA MATEMATIKAI KUTATÓINTÉZETE
POSTAFIÓK 127
H–1364 BUDAPEST
HUNGARY

ohkatona@math-inst.hu
dezso@math-inst.hu

B. Thalheim

FACHBEREICH INFORMATIK
TECHNISCHE UNIVERSITÄT COTTBUS
POSTFACH 101344
D–03013 COTTBUS
GERMANY

thalheim@informatik.tu-cottbus.de