The Average Length of Keys and Functional Dependencies in (Random) Databases

J. Demetrovics[1*], G.O.H. Katona[2*], D. Miklos[2*], O. Seleznjev[3**], B. Thalheim[4]

[1] Comp. & Autom. Inst., Hungarian Academy, Kende u. 13-17, H-1111 Budapest
[2] Mathematical Inst., Hungarian Academy, POBox 127, H-1364 Budapest
[3] Moscow State University, Dept. of Mathematics and Mechanics, RU-119 899, Moscow
[4] Cottbus Technical University, Computer Science Inst., POBox 101344, D-03013 Cottbus
h935dem@ella.hu, h1164kat@ella.hu, h1162mik@ella.hu, seleznev@compnet.msu.su,
thalheim@informatik.tu-cottbus.de

Abstract. Practical database applications engender the impression that sets of constraints are rather small and that large sets are unusual and caused by bad design decisions. Theoretical investigations show, however, that minimal constraint sets are potentially very large. Their size can be estimated to be exponential in terms of the number of attributes. The gap between belief and theory causes non-acceptance of theoretical results. However, beliefs are related to average cases.

The theory known so far considered worst case complexity. This paper aims at developing a theory of *average case complexity*. Several statistic models and asymptotics of corresponding probabilities are investigated for random databases. We show that exponential complexity of independent key sets and independent sets of functional dependencies is rather unusual. Depending on the size of relations almost all minimal keys have a length which mainly depends on the size. The number of minimal keys of other length is exponentially small compared with the number of minimal keys of the derived length. Further, if a key is valid in a relation then it is probably the minimal key. The same results hold for functional dependencies.

1 Average Length of Keys

In databases keys play an important role. Records or tuples can be identified, recorded and searched in a unique way. Generally, a key is an attribute (or a combination of several attributes) that uniquely identifies a particular record. Keys are used everywhere in the database to serve as references to tuples identified by values. Keys

* Supported by the Hungarian National Foundation of Scientific Research, Grant No. 2575.

** Supported by the German Natural Science Research Council, contract BB-II-B1-3141-211(94).

are generalized to functional dependencies. Those specify the relationship between two attribute sets. In a relation the values of the first set determine the values of the second set. Functional dependencies are used for normalization of database systems. If a database designer knows the complete set of functional dependencies in a given application then unpredictable behavior during updates and update anomalies can be avoided. Therefore, the size of functional dependency and key sets is of great interest. If this size is exponential in the number of attributes then the entire approach becomes unmanageable.

In practical applications it is often the case that sets of keys and sets of functional dependencies are rather small. Based on this observation practioners believe that those sets are small in most applications. If there is an application with a large set of constraints then this application is considered to be poorly designed. This belief of engineers is opposed by theoretical results. It can be proven that key sets and sets of functional dependencies are indeed exponential. Hence the problem which case should be considered the normal one: the observation of practioners or the theory of theoreticians. The solution to this gap between beliefs of practioners and results of theoreticians can be given by developing a theory of average case complexity. There are cases in which worst case complexity really occurs. In most cases, as shown below, worst case complexity is unlikely. Thus, for average case considerations, the observation of practioners is well-founded by the approach to be developed below.

There are several reasons why complexity bounds are of interest. Firstly, most of the known algorithms, e.g. for normalization, depend on the set of all minimal keys or nonredundant sets of dependencies. Therefore, their algorithmic complexity depends on the cardinality of these sets. Secondly, the maintenance complexity of a database depends on the number of integrity constraints are under consideration. Therefore, if the cardinality of constraint sets is large then maintenance becomes infeasible. Furthermore, cardinality gives an indication whether algorithms are of interest for practical purposes since the complexity of most known algorithms is measured by the input length. For instance, algorithms for the construction of a minimal key are bounded by the maximal number of minimal keys. The decision problem whether there is a minimal key with at most k attributes is NP-complete. The decision problem whether two sets of functional dependencies are equivalent is polynomial in the size of the two sets and hence exponential.

Two different approaches can be used for specification of key set behavior:

1. The worst case size is considered.
2. The average case complexity is considered.

Although the second approach is more reliable only very few results are known (Thalheim (1991)). In almost all relations with m tuples on domains with $\mid dom(A_i) \mid = 2 \ (1 \leq i \leq n)$ the average length $av_n(m, 2)$ of minimal keys is bounded by

$$\rbrack \log_2 m \lbrack \ \leq \ av_n(m, 2) \ \leq \ 2 \rbrack \log_2 m \lbrack \ .$$

The worst case complexity has been investigated in a large number of papers (see, for example, Beeri, Dowd, Fagin, Statman (1984), Bekessy, Demetrovics, Hannak, Frankl, Katona (1980), Demetrovics, Katona (1983), Mannila, Raihä (1992), Thalheim (1992)). The number of keys and minimal keys for Bernoulli databases is investigated in Andreev (1982). Some of his results are close to those presented in sections 3 and 4. However, his techniques use rather complicated graph technique and are not directly generalizable. The number of keys of a relation is determined by the maximal number of elements in a Sperner set. More precisely, given a relational schema $R = (\{B_1, ..., B_n\}, \emptyset)$ and a relation r from $SAT(R)$. Then r has at most

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}$$

different minimal keys. This estimate is sharp, i.e a relation can be constructed with exactly this number of minimal keys.

For the solution, we now use the following approach. We consider random databases of a limited size with a constant number of attributes and restricted domains. Then we derive the likelihood of constraint validity. Next, we show how to estimate probabilities. Based on these results conditions for constraint sets can be derived. The validity of constraints for which the conditions are violated is highly unlikely in the databases under consideration. Finally, the limitations can be omitted and general conditions can be developed.

We can directly apply the results below to the solution of several database problems. Some of them are the following:

- Results of this paper can be directly applied to heuristic support for database design. If the size of relations is restricted by a certain function then the size of minimal keys can be considered in accordance to the expected length. This approach has been used in the design system RAD (Albrecht et al. (1994)).
- Database mining aims at discovering semantics in real existing databases. If any possible constraint or even any possible key is checked database mining is infeasible. However, if we consider the statistic-based approach of this paper we only have to check the validity of a very small portion of constraints provided the size of the database is limited by certain bounds.
- Often stochastic algorithms are applied to solving difficult tasks in databases. Algorithms of the Monte-Carlo or of the Las-Vegas type are more reliable if they can be applied to databases with predictable constraint sets. Thus, our approach can be used to determine in which case such algorithms are useful and in which case they should not be applied.

This short list of application areas for our results is not exhaustive. We feel that the main application of our approach should be database design. The approach restricts the set of constraints to be considered to those which are more likely.

Section 2 discusses the behavior of functional dependencies in random databases after presenting basic notions in Section 2. Sections 4 and 5 are devoted to keys in random databases. Section 6 discusses some extensions of the approach. Proofs of results are given in Seleznjev, Thalheim (1994) and Demetrovics J., Katona G.O.H., and Miklos (1994) and are omitted due to space limitations.

2 Basic notions

We use some definitions of the theory of relational databases. Given sets $D_1, \ldots, D_n$ call *domains*, an n-ary relation R defined over $D_1, \ldots, D_n$ is a subset of the cartesian product $D_1 \times \ldots \times D_n$. An *attribute* is a name assigned to a domain of a relation. Any value associated with an attribute is called attribute value. The attributes names must be distinct. The symbol U will be used to denote the set of all n attribute of R. We assume in sequel that $U = \{1, \ldots, n\}$. A set of attributes $A, A \subseteq U$, is called *key* of R, if for every n-tuple of R the values of attributes in A uniquely determine the values of the attributes in U, i.e. for any $i, j = 1, \ldots, m, i \neq j$ tuples $t_i(A) \neq t_j(A)$, where m denotes the number of tuples in a relation R. Write in sequel $M = m(m-1)/2$. A key A is called a *minimal* key if no any proper subset of A is a key. Consider a relation (or database) R as a matrix with m rows (tuples) and n columns (attributes). Note that this definition of a database implies that some of the tuples can be identical. Let $A \subset U, B \subset U \setminus A$. Following Armstrong (1974), we say that A *determines* B (or B *functionally depends* on A), if there are no tuples in R with the same data in columns A but different in columns B. Denote functional dependency B on A by $A \to B$. We say that R is a random database, if tuples $t_i(U), i = 1, \ldots, m$, are independent and identically distributed random vectors. Assume also that domains $D_i, i = 1, \ldots, m$, be a finite integer sets and distribution of $t(U)$ is defined by probabilities $P\{t(1) = k(1), \ldots, t(n) = k(n)\} = p(k(U))$, $k(U) = (k(1), \ldots, k(n)), k(i) \in D_i, i = 1, \ldots, n$. In further considerations we suppose that distribution $p(k(U))$ are given. We call a random database R *standard Bernoulli database*, if $D_k = \{1, 0\}$ and $t(i), i = 1, \ldots, n$ are standard Bernoulli random variables $P\{t(i) = 0\} = P\{t(i) = 0\} = \frac{1}{2}$, and therefore $p(k(U)) = 2^{-n}$. Say R is a *uniform random database*, if $t(i), i = 1, \ldots, n$ are independent and $P\{t(i) = k(i)\} = |D_i|^{-1}, i = 1, \ldots, n,$, i.e. $p(k(A)) = \prod_{i \in U} |D_i|^{-1}$, where for any finite set $A, |A|$ denotes its cardinality.

Based on information about distribution $p(k(A))$ we study some probability problems of database theory. At first we estimated the probability of existence of functional dependency in a random database. Some more general problems connected with functional dependency and keys are investigated as well. We consider Poisson approximation to the distribution of random number of cases N, when functional dependency fails. Analogously we investigate an asymptotic distribution for a random number of coincidences for a set of attributes X when $a(n) \to \infty$ as $n \to \infty$. Similar results for arbitrary random databases with some uniformity condition for a distribution $p(k(X)) = \prod_{i \in X} |D_i|^{-1}$ are obtained. Asymptotic distribution of a size of a minimal key and mean number of keys in standard Bernoulli database are also investigated. We consider for a set of attributes A in a standard Bernoulli database probability that A is a minimal key. In more general case of arbitrary uniform random database some asymptotic results for this probability can be obtained. Some of the problems in discrete mathematics (i.e. selection of a base for a boolean algebra (Sachkov (1982)), a random strategy of search (Ahlswede, Wegener (1979)) are close to asymptotic results for keys in random databases discussed in Section 4.

3 Functional dependencies in random databases

In some cases a set B functionally depends on a set of attributes A deterministically, e.g. if $t(i) = f_i(t(A)), i \in B$, then clearly $P\{A \to B\} = 1$. But in random databases this property connected also with joint distribution of $t(A)$ and $t(B)$. It is of interest that even for statistically independent random vectors $t(A)$ and $t(B)$ the probability of functional dependency $A \to B$ may be close to 1. We can call this case *artificial* dependency.

The notion of functional dependency can be generalized in the following way. We call that a set of attributes A *almost determines* a set of attributes $B, B \subseteq U \backslash A$, if the number of tuples with *functional independency* condition, i.e. $t(A) = t'(A), t(B) \neq t'(B)$, is a finite number $N = N(n)$, say. Let $m = m(n) \to \infty$ as $n \to \infty, n$ is a number of attributes in R.

Consider at first the case of a uniform random database. Denote by

$$a = a(n) = \sum_{i \in A} \log_2 |D_i|, b = b(n) = |B| \geq 1, \lambda(n) = M \frac{1}{2^a}(1 - \frac{1}{2^b}).$$

For a standard Bernoulli database we have $a(n) = |A|$. The following theorem allows to estimate the distribution function of number of functional independencies N.

Theorem 1 *Let R be a uniform random database, $c_1 \frac{1}{a(n)} \leq \lambda(n) \leq c_0 \frac{a(n)}{\ln a(n)}$, where $0 < c_0 < e^{-2} \frac{1}{2} \ln 2, c_1 > 0$. Then there exists $\gamma > 0$ such that*

$$P\{N = s\} = \frac{\lambda(n)^s}{s!} e^{-\lambda(n)} (1 + O(e^{-\gamma \frac{a(n)}{\ln a(n)}}))$$

$$\text{as } n \to \infty, \text{ uniformly in } 0 \leq s \leq c_0 \frac{a(n)}{\ln a(n)}.$$

We can interpret the assertion of Theorem 1 as following. The number of functional independencies has asymptotically Poisson distribution with parameter $\lambda(n)$. Clearly that the case of functional dependency simply follows by Theorem 1 since $P\{A \to B\} = P\{N = 0\}$.

Corollary 1 *Let R be a uniform random database and the conditions of Theorem 1 are valid. Then there exists $\gamma > 0$ such that*

$$P\{A \to B\} = e^{-\lambda(n)} (1 + O(e^{-\gamma \frac{a(n)}{\ln a(n)}})) \text{ as } n \to \infty.$$

To formulate the next corollary we introduce some additional denotations. Write

$$p(\alpha, \beta) = \begin{cases} 0, & \text{if } \alpha = -\infty, \\ \exp\{-2^{-(\alpha+1)} c_\beta\}, & \text{if } |\alpha| < \infty, \\ 1, & \text{if } \alpha = +\infty, \end{cases}$$

where

$$c_\beta = \begin{cases} 1, & \text{if } \beta = +\infty, \\ 1 - 2^{-\beta}, & \text{if } 1 \leq \beta < +\infty. \end{cases}$$

Denote by $\alpha(n) = \sum_{i \in A} \log_2 |D_i| - 2 \log_2 m$. And if $\lambda(n) \to \lambda_0, 0 \leq \lambda_0 \leq +\infty$, i.e. $\alpha(n) \to \alpha_0$, $|\alpha_0| \leq \infty$, as $n \to \infty$, we obtain the following corollary.

Corollary 2 *Let R be a uniform random database. Let $b(n) \to \beta \geq 1$ and $\alpha(n) \to \alpha_0$, as $n \to \infty$, $|\alpha_0| \leq \infty$. Then $P\{A \to B\} \to p(\alpha_0, \beta)$ as $n \to \infty$. Moreover if $|\alpha_0| < \infty$, then for any $s = 0, 1, \ldots$,*

$$P\{N = s\} \to \frac{\lambda_0^s}{s!} e^{-\lambda_0} \text{ as } n \to \infty, \quad \lambda_0 = 2^{-(\alpha_0+1)} c_\beta.$$

Thus the number of functional independencies N converges in distribution to a Poisson random variable with parameter λ_0.

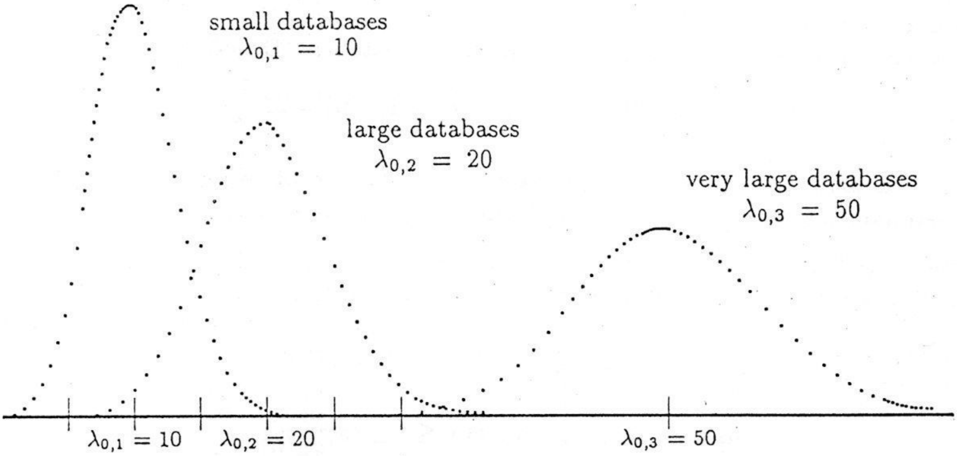

Fig. 1. Poisson distribution of Independencies in Different Databases with Different Parameters

Different distributions are shown in Figure 1. The parameter λ restricts the length of left sides in functional dependencies. Therefore, the size of the database very stongly indicates the length of possible candidates for left sides of functional dependencies. If the database is small then $\lambda_{0,1}$ is small and only sets A with a length in this small interval are likely to be left sides. If the database is large then $\lambda_{0,3}$ is large. The interval is determined by some small portion of possible candidates. Therefore, there exists a subset $\mathcal{F}_0(B)$ of all possible functional dependencies $\mathcal{F}(B)$ with right side B which is likely. This set is much smaller than $\mathcal{F}(B)$ and, for the average case, a functional dependency from $\mathcal{F}(B)$ which is valid in the random database belongs with high probability to $\mathcal{F}_0(B)$. Thus, for example, heuristic algorithms which are used to check validity of functional dependencies from $\mathcal{F}(B)$ should begin with constraints from $\mathcal{F}_0(B)$. In this case, they succeed much faster.

The high in curves depend upon the size of B. If B is smaller then the curves are zoomed up. In this case the intervall with the most probable candidates gets smaller. If B is large then the intervall gets larger. We shall return to these curves during consideration of keys.

In the case of general random database R we introduce some additional denotations. For any tuple $t(U)$ write

$$p(k(A), k(B)) = P\{t(A) = k(A), t(B) = k(B)\},$$
$$p(k(A)/k(B)) = P\{t(B) = k(B)/t(A) = k(A)\},$$

i.e. $p(k(A)/k(B))$ is a conditional probability of the event $t(B) = k(B)$ when $t(A) = k(A)$, $k(A) \in \prod_{i\in A} D_i$, $k(B) \in \prod_{i\in B} D_i$, with cartesian product of domains. Write

$$\lambda(n) = \frac{m(m-1)}{2} \sum_{k(A),k(B)} p(k(A))(1 - p(k(B)/k(A)))p(k(B), k(A)) =$$
$$ME[p(t(A))(1 - p(t(B)/t(A)))].$$

where $M = m(m-1)/2$. Introduce the following uniformity condition. There exist $K_i > 0, i = 1, 2, 3$, such that for any $k(A) \in \prod_{i\in A} D_i$, $k(B) \in \prod_{i\in B} D_i$,

$$\frac{K_1}{2^{a(n)}} \leq p(k(A)) \leq \frac{K_2}{2^{a(n)}}; \qquad p(k(B)/k(A)) \leq K_3 < 1, \tag{1}$$

where $a(n) = \sum_{i\in A} \log_2 |D_i|$.

Theorem 2 *Let R be a random database with a given distribution $p(k(A), k(B))$. Let the assumptions of Theorem 1 and (1) hold. Then the assertion of Theorem 1 for the number of functional independencies $N = N(n)$ is valid.*

Clearly the analogous corollaries to Corollaries 1 and 2 can be formulated for the case of general random database too.

Corollary 3 *Let R be a random database with given distribution $p(k(A), k(B))$. Let the assumptions of Corollary 1 and (1) hold. Then the assertion of Corollary 1 for $P\{A \to B\}$ is valid.*

Corollary 4 *Let R be a random database with a given distribution $p(k(A), k(B))$ and (1) hold. Let $\lambda(n) \to \lambda_0$ as $n \to \infty, 0 \leq \lambda_0 \leq +\infty$. Then $P\{A \to B\} \to e^{-\lambda_0}$ as $n \to \infty$. Moreover for $|\lambda_0| < \infty$, and for any $s = 0, 1, \ldots$,*

$$P\{N = s\} \to \frac{\lambda_0^s}{s!} e^{-\lambda_0} \text{ as } n \to \infty.$$

The results of Theorem 2 explain that the property of functional dependency for a random database with increasing number of tuples m is related to two main factors:

- determination of an attribute set A (probability $p(k(A))$),
- correlation between attribute values in A and B (conditional probability $p(k(B)/k(A))$).

Therefore, in the case that the values in $dom(A_i)$ are not uniformly distributed the results presented in Theorem 1 are valid.

4 Keys in random databases

The asymptotic results and corresponding proofs for probability for a set of attributes A to be a key are very close to the problem of functional dependency when $b(n) = \sum_{i\in B}|D_i| \to \infty$ as $n \to \infty$. But in the case of an uniform random database there is an exact expression for the corresponding probability $P\{R \models A\}$. Write as beforehand $a = a(n) = \sum_{i\in A}\log_2|D_i|$. For a standard Bernoulli database we have $a(n) = |A|$.

The notion of a key can be generalized in the following way. We call that a set of attributes A a key with a finite number of exceptions, if the number of tuples with $t(A) = t'(A)$ is a finite random variable $N = N(n)$, say. Then $P\{R \models A\} = P\{N = 0\}$. A random number N for a database R means a number of tuples with coincident values for a set of attributes A. Let $m = m(n) \to \infty$ as $n \to \infty, n$ is a number of attributes in R.

Consider at first the case of a uniform random database with $\lambda(n) = M2^{-a(n)}$ and $\alpha(n) = a(n) - 2\log_2 m$.

Theorem 3 *Let R be a uniform random database and $m \le 2^{a(n)}$. Then the following statements are valid:*

(i) $P\{R \models A\} = \prod_{j=1}^{m-1}(1 - j2^{-a(n)})$,

(ii) if $m < 2^{(2-\gamma)a(n)/3}, 0 < \gamma < 2$, *then*

$$P\{R \models A\} = e^{-\lambda(n)}(1 + O(2^{-\gamma a(n)})) \text{ as } n \to \infty,$$

(iii) if $\lambda(n) \to \lambda_0, 0 \le \lambda_0 \le \infty$, *i.e.* $\alpha(n) \to \alpha_0$, *as* $n \to \infty, |\alpha_0| \le \infty$, *then*

$$P\{R \models A\} \to e^{-\lambda_0} = \exp\{-2^{-(\alpha_0+1)}\} \text{ as } n \to \infty,$$

(iv) if $\lambda(n) \to \lambda_0$ *as* $n \to \infty, 0 \le \lambda_0 < \infty$, *then for any* $s = 0, 1, \ldots,$

$$P\{N = s\} \to \frac{\lambda_0^s}{s!}e^{-\lambda_0} \text{ as } n \to \infty,$$

(v) if $c_1\frac{1}{a(n)} \le \lambda(n) \le c_0\frac{a(n)}{\ln a(n)}$, *where* $0 < c_0 < \frac{1}{2}e^{-2}\ln 2, c_1 > 0$. *Then there exists* $\gamma > 0$ *such that*

$$P\{N = s\} = \frac{\lambda(n)^s}{s!}e^{-\lambda(n)}(1 + O(e^{-\gamma\frac{a(n)}{\ln a(n)}}))$$

$$\text{as } n \to \infty \text{ uniformly in } 0 \le s \le c_0\frac{a(n)}{\ln a(n)}.$$

Therefore, worst case complexity results are highly unlikely. In the remaining cases, if m, n and the size of domains fulfill the conditions *(iv)* or *(v)* then we observe the same behavior as shown in Figure 1.

Consider now a general random database R with a given distribution of a tuple

$$P\{t(A) = k(A)\} = p(k(A)), k(A) \in \prod_{i \in A} D_i.$$

Denote by

$$\lambda(n) = \frac{m(m-1)}{2} \sum_{k(A)} p(k(A))^2 = ME[p(t(A)].$$

and introduce also the following condition (cf. (1)) There exist $K_i > 0, i = 1, 2$, such that for any $k(A) \in \prod_{i \in A} D_i$,

$$\frac{K_1}{2^{a(n)}} \leq p(k(A)) \leq \frac{K_2}{2^{a(n)}}. \tag{2}$$

Theorem 4 *Let R be a random database with a given distribution $p(k(A))$ and for an attribute set A condition (2) hold. Let $\lambda(n) \to \lambda_0$ as $n \to \infty$. Then assertions (iii)-(v) of Theorem 3 are valid.*

The assertion of Theorem 4 can be interpreted in a different way. Denote by ν_n an integer random variable which equals size of a minimal subset B in a set of attributes A, when B is a key, i.e. the size of a minimal key in $A, |A| = c(n)$. Then

$$P\{\nu_n \leq c(n)\} = P\{R \models A, |A| = c(n)\}.$$

Corollary 5 *Let (2) hold and $\lambda(n) \to \lambda_0$ as $n \to \infty, 0 \leq \lambda_0 \leq \infty$. Then*

$$P\{\nu_n \leq c(n)\} \to e^{-\lambda_0} \textit{ as } n \to \infty.$$

For a standard Bernoulli database we have $c(n) = a(n) = |A| = \alpha(n) + 2\log_2 m$. Therefore, if $\alpha(n) \to \alpha_0$ as $n \to \infty$ $|\alpha_0| \leq \infty$, then

$$P\{\nu_n \leq a(n)\} = P\{\nu_n - 2\log_2 m \leq \alpha(n)\} \to \exp\{-2^{-(\alpha_0+1)}\} \text{ as } n \to \infty.$$

Thus in the case of a standard Bernoulli database shifted size of a minimal key $\mu_n = \nu_n - 2\log_2 m$ has asymptotically double exponential distribution and values of random variable ν_n concentrates near $2\log_2 m$ as displayed in Figure 1.

We observe that keys are more likely in a very small interval. Therefore, algorithms which search for keys in relations are faster if first this interval is checked.

It means that for large values of $a(n) = |A|$ the random variable $\nu_n - 2\log_2 m$ has an asymptotical double exponential distribution, i.e. $F(x) = \exp(-\exp(-cx))$, $|x| < \infty$ where $c = \ln 2$. The number $\nu_n - 2\log_2 m$ expresses the shifted minimal key length.

Consider now a standard Bernoulli database R and attributes $U, |U| = n$. Assume that we choose a set of attributes A and include an attribute in a set A with probability $p = p(n) = 1 - q(n) \leq 1$, after some random experiment. Denote a random set of attributes by $\mathbf{A}$. Then by Theorem 3(i) $P\{R \models \mathbf{A}/|\mathbf{A}| = k\} = p(n,k) = \prod_{j=1}^{m-1}(1 - j2^{-a(n)})$ and we obtain

$$P\{R \models \mathbf{A}\} = \sum_{k=0}^{n} P\{R \models \mathbf{A}/|\mathbf{A}| = k\}P\{|\mathbf{A}| = k\} = \sum_{k=0}^{n} p(n,k)\binom{n}{k}p^k q^{n-k}$$

Denote the standard Gaussian distribution function by

$$\Phi(x) = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x} e^{-y^2/2}dy.$$

Theorem 5 *Let R be a standard Bernoulli database with a set of attributes $U, n = |U|$.*

(i)Let $n = 2\log_2 m + \alpha(n)$ and $\alpha(n) \to \alpha_0$ as $n \to \infty, |\alpha_0| \leq \infty$. If $nq(n) \to \tau, 0 \leq \tau < \infty$, then $P\{R \models \mathbf{A}\} \to E[\exp^{-2^{-(\alpha_0 - \theta + 1)}}]$ as $n \to \infty$, where θ is a Poisson random variable with parameter τ.

(ii) Let $n = \frac{2}{p}\log_2 m + \beta(n)\left(\frac{nq}{p}\right)^{\frac{1}{2}}$. If $nq(n) \to \infty, q(n) \to q_0 > 0$ as $n \to \infty$, and additionally , $\beta(n) \to \beta_0$ as $n \to \infty$, $|\beta_0| \leq \infty$, then $P\{R \models \mathbf{A}\} \to \Phi(\beta_0)$ as $n \to \infty$.

We can interpret these results in the following way. Let R be a standard Bernoulli database and K, say, be a number of keys in L selected independently sets $\mathbf{A}_i, i = 1, \ldots, L$. Then every set of attributes $\mathbf{A}_i, i = 1, \ldots, L$, is a key with a probability $P\{R \models \mathbf{A}_i\}$ and we obtain

$$K = \sum_{i=1}^{L} I_i, \qquad E[K] = LP\{R \models \mathbf{A}_1\},$$

where I_i is a indicator function of the event $\{R \models \mathbf{A}_i\}$ and $E[K]$ is a mean number of keys in a sample of L sets.

When $L = 2^n$ and $p = q = \frac{1}{2}$, a random number of keys K can be represented also in the following form. For any set of attribute A_i denote by J_i the indicator function of the event $\{R \models A_i\}$, $i = 1, \ldots, 2^n$. Then $K = \sum_{i=1}^{2^n} J_i$, $P\{R \models A_i\} = p(n,k), |A| = k$, and

$$E[K] = \sum_{k=0}^{n}\binom{n}{k}p(n,k) = 2^n\sum_{k=0}^{n}\binom{n}{k}\frac{1}{2^n}p(n,k) = 2^n P\{R \models \mathbf{A}\}.$$

Applying Theorem 5 we have the following asymptotic for a mean number of keys $E[K]$.

Corollary 6 *Let R be a standard Bernoulli database with a set of attributes $U, |U| = n, n = 2\log_2 m + \alpha(n)$ and $p = q = \frac{1}{2}$, $n = 4\log_2 m + \beta(n)n^{\frac{1}{2}}, \beta(n) \to \beta_0$ as $n \to \infty, |\beta_0| \leq \infty$. Then*

$$E[K] \sim 2^n\Phi(\beta_0) \text{ as } n \to \infty.$$

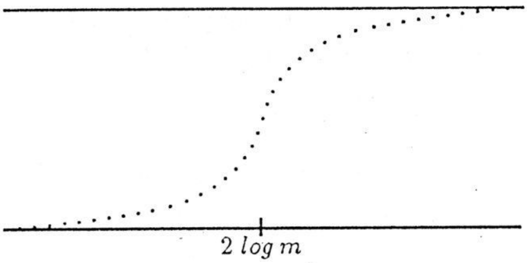

Fig. 2. Mean number of Keys Depending on Length

Further, the expectation that K is a key can be displayed as shown in Figure 2. This figure displays the mean number of key depending of the number of attributes and the size m of relations.

Based on this corollary we can show that Monte-Carlo algorithms can be used for checking keys in relations. Randomly attributes are added until the set reaches the size specified by $a(n)$ and m. Then we get keys with a probability $1-(\frac{1}{2})^k$ for k trials with the Monte-Carlo algorithm.

5 Minimal keys

The methods and results of the previous section can be applied also to the investigation of the probability of the event that a set of attributes $A, |A| = a$, is a minimal key in a random database R. Let R be a standard Bernoulli database with $D_i = \{0,1\}, i = 1,\ldots,n$, the case when $D_i = \{0,1,\ldots,d\}, i = 1,\ldots,n$, can be considered in a similar way. Denote by $K(R)$ and $K_{\min}(R)$ the sets of keys and minimal keys in R respectively. Write $A_k = A \setminus \{k\}$, $\mathcal{A}_k = \{A_k \in K(R)\}$, $k = 1,\ldots,a$, and $\mathcal{A} = \{A \in K(R)\} = \{R \models A\}$. Then it follows directly by the definition of a minimal key and Bonferroni inequality (see e.g. Feller (1968), Bender (1974)) for any $r,t = 0,\ldots,M$,

$$|P\{N = r\} - \sum_{j<t} (-1)^j \binom{r+j}{j} S_{r+j}| \leq \binom{r+t}{t} S_{r+t}. \tag{3}$$

that $P\{A \in K_{\min}(R)\}$ can be represented in the following form.

Proposition 1 *Let R be a standard Bernoulli database. Then*

$$P\{A \in K_{\min}(R)\} = P(\mathcal{A}) - \sum_{j=1}^{a(n)} (-1)^{j-1} \binom{a(n)}{j} P(\mathcal{A}_1 \ldots \mathcal{A}_j).$$

Denote by $\lambda = \lambda(n) = M2^{-a(n)}$.

Theorem 6 *Let R be a standard Bernoulli database and $\lambda(n) \geq \ln\ln a(n) + \lambda_0$. Then there exists $\gamma > 0$ such that for any sufficiently large $\lambda_0 > 0$*

$$P\{A \in K_{\min}(R)\} = e^{-\lambda(n)}(1 - e^{-\lambda(n)})^{a(n)}(1 + O(e^{-\gamma\frac{a(n)}{\ln a(n)}})) \ \textit{as}\ n \to \infty.$$

The assumption about lower bound for $\lambda(n)$ in Theorem 6 has a technical character. In the more general case $\lambda(n) \geq \lambda_0 > 0$ we obtain only the following estimate for $P\{A \in K_{\min}(R)\}$.

Proposition 2 *If $\lambda(n) \geq \lambda_0 > 0$ then there exists $\gamma > 0$ such that*

$$P\{A \in K_{\min}(R)\} \leq e^{-\lambda(n)}(1 - e^{-\lambda(n)})^{b(n)}(1 + O(e^{-\gamma\frac{a(n)}{\ln a(n)}})) \ \textit{as}\ n \to \infty,$$

where $b(n) = b_0\frac{a(n)}{\ln a(n)}, b_0 > 0$.

Theorem 6 and Proposition 2 can be used to estimate the asymptotic maximum value of $P\{A \in K_{\min}(R)\}$.

Proposition 3 *If $\lambda(n) \geq \lambda_0 > 0$ then*

$$P\{A \in K_{\min}(R)\} = P(a) \leq P_{\max}(a) = \frac{e^{-1}}{a(n)+1}(1 + o(1)) \ \textit{as}\ n \to \infty,$$

and also $P(a) = P_{\max}(a)$ if and only if $\lambda(n) = \ln(a(n)+1)(1+o(1))$ as $n \to \infty$.

The result of Proposition 3 shows that the probability for a set of attributes A to be a minimal key becomes small as $a \to \infty$ and this property does not depend on the relation between number of tuples m and the size of A as in the case of key (cf. Theorem 3). But the following corollary shows that this dependence is important when we consider the conditional probability $P(a/K) = P\{A \in K_{\min}(R)/A \in K(R)\}$.

Corollary 7 *Let $m^2 = 2^{a(n)+1}(\ln a(n) + b(n))$ and $b(n) \to \beta$ as $n \to \infty$, $|\beta| \leq \infty$, and also $b(n) \geq \ln\ln a(n) - \ln a(n) + \lambda_0$, $\lambda_0 > 0$, then*

$$P(a/K) \to \begin{cases} 0, & \textit{if}\ \beta = -\infty, \\ \exp\{-e^{-\beta}\}, & \textit{if}\ |\beta| < \infty, \\ 1, & \textit{if}\ \beta = +\infty, \end{cases}$$

as $n \to \infty$.

This result can be now compared with worst case complexity of key systems. It means that if m is small or too large then key systems can contain only a very small number of minimal keys. Therefore, worst case complexity results are highly unlikely.

As a straightforward corollary to the previous result we have that if $\lambda(n) = \ln a(n) + b(n)$ and $b(n) \to +\infty$ as $n \to \infty$, then

$$P\{A \in K_{\min}(R)\} \sim P\{A \in K(R)\} \sim e^{-\lambda(n)} \text{ as } n \to \infty. \qquad (4)$$

Therefore, if β is unbounded then it is highly unlikely that a set of this size is a key.

This corollary states now the following surprising result:
For any bounded behavior of β , if a set of length $\lambda(n)$ is a key then this set is with high probability also a minimal key.

6 Concluding Remarks

We assumed in previous sections that the initial distribution of tuples in random database R is approximately given. Often a distribution of tuples is unknown. We can generalize the above discussed results to statistical problems for statistical analysis of databases:

(i) to test of homogeneity of data in R,
(ii) to test independency of tuples and attributes in a tuple,
(iii) to fit a distribution of tuples.

For the first problem we can use for example clustering methods (see e.g. Tou, Gonsales (1974)). Then we can further investigate selected homogeneous clusters. It is possible to use for clusterization some attributes in database or its functionals. For example, a bank database keeps information about residuals in accounts for a long period. Assume that large and active in some sense accounts have different statistical characteristics than small ones. Then we can use as features for clusterization instead of full information the following functionals of attributes:

- mean residual for accounts for a period,
- mean residual for accounts with large residuals (greater than a given level),
- mean absolute values of differences (current and next days).

The independency of different tuples and attributes in R can be verified by some statistical goodness-of-fit tests. To estimate a distribution of a tuple we can use parametric and nonparametric models. The simplest model is a uniform random database. For dependent attributes it is possible to use polynomial or multidimensional Gaussian distribution or histograms, kernel density estimates etc.. For Gaussian distribution we have to modify some previous definitions, e.g., say, $t(A) = t'(A)$ if $|t(k) - t'(k)| < \delta, \delta > 0, k \in A$. To estimate mean characteristics we can use analogous empirical ones, e.g. for $s(n) = E[p(t(A))]$ the statistic

$$\hat{s}(n) = \frac{1}{|S|} \sum_{i,j \in S} I_{\{t_i(A)=t_j(A)\}},$$

where S is a sample of homogeneous tuples and I_C is an indicator function of the event C. The estimate $\hat{s}(n)$ is a standard and optimal in definite sense estimate of $s(n)$. Then we can apply some previous theoretical results with corresponding investigation of statistical errors.

References

1. Ahlswede, R., Wegener, I. (1979).*Suchprobleme*, Teubner B.G., Stuttgart.
2. Albrecht M., Altus M., Buchholz B., Düsterhöft A., Schewe K.-D., Thalheim B. (1994), Die intelligente Tool Box zum Datenbankentwurf RAD. *Datenbank-Rundbrief*, 13, FG 2.5. der GI, Kassel.
3. Andreev, A. (1982), Tests and pattern recognition. PhD thesis, Moscov State University, 1982.
4. Armstrong, W.W.(1974). Depending structures of database relationships. *Information Processing*-74, North Holland, Amsterdam, 580-583.
5. Beeri C., Dowd M., Fagin R., Statman R. (1984), On the structure of Armstrong relations for functional dependencies. *Journal of ACM*, Vol.31, No.1, 30–46.
6. Bekessy A., Demetrovics J., Hannak L., Frankl P., Katona G. (1980), On the number of maximal dependencies in a database relation of fixed order. *Discrete Math.*, 30, 83–88.
7. Bender, E.A. (1974) Asymptotic methods in enumeration. *SIAM Review*, 16, 4, 485-515.
8. Billingsley, P. (1975) *Convergence of Probability Measures*, Wiley, N.Y.
9. Codd E.F. (1970), A relational model for large shared data banks.. *Comm. ACM* 13, 6, p. 197–204.
10. Demetrovics J. (1979), On the Equivalence of Candidate Keys with Sperner sets. *Acta Cybernetica*, Vol. 4, No. 3, Szeged, 247 - 252.
11. Demetrovics J. , Katona G.O.H. (1983), Combinatorial problems of database models. Colloquia Mathematica Societatis Janos Bolyai 42, Algebra, Combinatorics and Logic in Computer Science, Gÿor (Hungary), 331–352.
12. Demetrovics J., Katona G.O.H., and Miklos (1994), Functional Dependencies in Random Relational Databases. Manuscript, Budapest.
13. Demetrovics J., Libkin L.O., and Muchnik I.B. (1989), Functional dependencies and the semilattice of closed classes. *Proc. MFDBS-89*, LNCS 364, 136–147.
14. Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, Wiley, N.Y.
15. Gottlob G. (1987), On the size of nonredundant FD-covers. *Information Processing Letters*, 24, 6, 355–360.
16. Mannila H., Räihä K.-J. (1982), On the relationship between minimum and optimum covers for a set of functional dependencies. Res. Rep. C-1982-51, University of Helsinki.
17. Mannila H., Räihä K.-J. (1992), *The design of relational databases.* Addison-Wesley, Amsterdam.
18. Sachkov, V.N. (1982). *An Introduction to Combinatorics Methods of Discrete Mathemathics*, Moscow, Nauka.
19. Seleznjev O., Thalheim B. (1988), On the number of minimal keys in relational databases over nonuniform domains. Acta Cybernetica, Szeged, 8, 3, 267–271.
20. Seleznjev O., Thalheim B. (1994), Probability Problems in Database Theory. Preprint I-3/1994, Cottbus Technical University.
21. Thalheim B. (1987), On the number of keys in relational databases. *Proc. FCT-87-Conf.*, Kazan, LNCS 1987.
22. Thalheim B. (1989), On Semantic Issues Connected with Keys in Relational Databases Permitting Null Values. *Journal Information Processing and Cybernetics, EIK*, 25, 1/2, 11–20.
23. Thalheim B. (1991), *Dependencies in Relational Databases.* Leipzig, Teubner Verlag.
24. Thalheim B. (1992), On the number of keys in relational and nested relational databases. *Discrete Applied Mathematics*, 38.
25. Tou, J., Gonsales, R. (1974). *Pattern Recognition Principles*, Add.-Wesley, London.