Random Databases and Keys

O. Seleznjev,
Faculty of Mathematics and Mechanics, Moscow State University, RU-119 899, Moscow
B. Thalheim
Cottbus Technical University, Computer Science Inst., POBox 101344, D-03013 Cottbus

Abstract

Practical database applications give the impression that sets of constraints are rather small and that large sets are unusual and caused by bad design decisions. Theoretical investigations however show 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 observation in practice and theory results in rejection of theoretical results. However, practice is related to average cases and is not related to worst cases.

The theory used until now considered the worst case complexity. This paper aims to develop a theory for the *average case complexity*. Several probabilistic models and asymptotics of corresponding probabilities are investigated for random databases formed by independent random tuples with a common discrete distribution. Poisson approximations are studied for the distributions of some characteristics for such databases where the number of tuples is sufficiently large. We intend to prove that the exponential complexity of key sets and sets of functional dependencies is rather unusual and almost all minimal keys in a relation have a length which depends mainly on the size of the relation.

Keywords: FUNCTIONAL DEPENDENCY; KEYS; MINIMAL KEYS; RANDOM DATABASE; BONFERRONI INEQUALITY; POISSON APPROXIMATION

1 Introduction

The relational data model is one of the main database models and the basis for most existing database management systems. In this model the user's data are expressed by relations (relational matrices) whereby rows represent records and columns represent the domains or attributes, respectively. Records or tuples can be identified, recorded and searched by sets of attributes, so-called *keys*, in a unique way. Generally, a key is an attribute (or a combination of several attributes) that uniquely identifies a particular record. A given set of attributes is a *minimal* key if its proper subsets are not keys. The basis for a large variety of algorithms used in database technology is the identification of tuples through keys, e.g. algorithms for selection, joining, constructing and maintaining tuples are as simple as search algorithms if key indexes are used. Therefore, keys and minimal keys are absolutely fundamental to database models. However if this approach is used, at least the combinatorial behavior of key sets should be known. For relations, keys are generalized to functional dependencies which 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 the 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 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, minimal keys and sets of functional dependencies are rather small. Based on this observation practitioners 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 observation of engineers is in contrast with theoretical results. It can be proven that in the worst case key sets and sets of functional dependencies are indeed exponential. Hence the problem is deciding which case should be considered the normal one: the observation of practitioners or the theory of theoreticians. The solution to this gap between the observation of practitioners and the results of theoreticians can be given by developing a theory of *average case complexity* (cf. the average complexity setting in theory of approximation and scientific computing, see e.g., Traub, Wasilkowski and Woźniakowski (1988)). We use the notion *complexity* for the cardinality of a minimal key set. There are cases in which the worst case complexity does occur. In most cases, as shown below, the worst case complexity is unlikely. Thus, for average case considerations, the observation of practitioners is well-founded by the approach to be developed below.

The worst case complexity has been investigated in a large number of papers (see, e.g., Békéssy, Demetrovics, Hannák, Frankl and Katona (1980), Andreev (1984), Beeri, Dowd, Fagin and Statman (1984), Thalheim (1987), Seleznjev and Thalheim (1988), Mannila and Räihä (1992)). The number of minimal keys in a relation is determined by the maximum number of elements in a Sperner set. More precisely, for a given relational schema $\mathcal{S} = (\{D_1, ..., D_n\}, \emptyset)$ with domains $D_1, \ldots, D_n$, a relation $\mathcal{R}$ from $SAT(\mathcal{S})$ has at most

$$\binom{n}{\lfloor \frac{n}{2} \rfloor} = \mathrm{O}(2^n n^{-\frac{1}{2}})$$

different minimal keys. This estimate is precise, i.e. a relation can be constructed with exactly this number of minimal keys. These considerations can be extended to sets of functional dependencies (Gottlob (1987), Thalheim (1991), Mannila and Räihä (1992)).

Considering the worst case behavior we observe that large key sets occur only when the minimal keys are $\lfloor \frac{n}{2} \rfloor$ or $\rfloor \frac{n}{2} \lfloor$ in length. In our models a minimal key set is rather small whenever the length of minimal keys is much larger or smaller and also depends on the number of tuples in a relation. For instance, if minimal keys are not higher than 3 in length, then the cardinality of the set of all minimal keys can be bounded by $\mathrm{o}(n^3)$, i.e. the bound is polynomial and not exponential. Thus we now consider a minimal key probability, i.e. the probability for a given set of attributes to be a minimal key, and a length of minimal, and in particular shortest, key instead of the cardinality of key sets. In this case we can analyze the behavior of key sets as well. In the paper it is found that the distribution of the length of a shortest key in our models has relatively little support. As the first and necessary step to estimate the mean characteristics of the cardinality of key sets it needs to evaluate minimal key probabilities for different sets of attributes.

Some problems of discrete mathematics (e.g., nonredundant tests Andreev (1984); the height of binary digital trees, Devroye (1984), Szpankowski (1991)) are similar to the corresponding problems for keys in random databases in Section 3. We can directly apply our results to the following database problems:

- Results of this paper can be directly applied to the heuristic support within database design. If the size of relations is restricted by a certain function, then the size of minimal keys can be considered in accordance with the expected length. This approach has been used in the design system RADD (Albrecht, Altus, Buchholz, Düsterhöft and Thalheim (1994)).
- Database mining aims to discover 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 probabilistic approach of this paper, we only have to check the validity of a very small section of constraints, provided that the size of the database is limited by certain bounds.
- Stochastic algorithms are often applied to the solving of difficult tasks in databases. Algorithms of the Monte-Carlo type are more reliable if they can be applied to databases with predictable constraint sets. Thus, our approach can be used to determine when such algorithms are useful and when they should not be applied.

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

1.1 Basic notation

Let R be a matrix with m rows (tuples) and n columns (attributes), and $U = \{1, \ldots, n\}$ be the set of all attributes. For any set of attributes $A \subseteq U$, the corresponding part of the jth tuple will be termed $t_j(A), j = 1, \ldots, m$. Suppose all tuples take values in the Cartesian product $\prod_{i \in U} D_i = D_1 \times \cdots \times D_n$, where the domain D_i is a finite integer set for $i = 1, \ldots, n$. Denote by $|S|$ the cardinality of a finite set S.

We consider some properties of R in a probabilistic framework. We say that R is a *random database* if tuples $t_j(U)$, $j = 1, \ldots, m$, are independent and identically distributed random vectors with a given discrete distribution $P\{t_j(U) = k(U)\} = p(k(U))$ for $k(U) \in \prod_{i \in U} D_i$. Let $h(A) := \sum_{i \in A} \log_2 |D_i|$ be the *information* function of A for a given set of domains $\{D_i, i \in A \subseteq U\}$. We say that a random database R is a *standard Bernoulli* database if R is a binary matrix and $t_j(i), j = 1, \ldots, m, i = 1, \ldots, n$, are independent standard Bernoulli random variables with the equal probability outcomes 0 and 1, and whereby $p(k(U)) = 2^{-n}$. The random database R is a *uniform* random database if all tuples have the discrete uniform distribution, i.e. $p(k(U)) = 2^{-h(U)}$. For sequences $x(m)$ and $y(m)$, $m = 1, 2, \ldots$, we write $x(m) \asymp y(m)$, if there exist positive constants c_1 and c_2 such that $c_1 y(m) \leq x(m) \leq c_2 y(m)$ for all m. We say that R is *uniform-type* if $p(k(U)) \asymp 2^{-h(U)}$.

Further, a set of attributes A is said to be a *key in a random database* R if all $t_j(A)$ are different. We say a key A is *minimal in* R if the key property fails for any proper subset $B \subset A$. Write $R \models A$ if A is a key, and $R \models_{\min} A$ if A is a minimal key. Let $A \subset U$ and $B \subseteq U \setminus A$. We say that 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 this property by $A \to B$. Henceforth we consider the asymptotic behavior of some characteristics of random databases when the number of tuples m and the number of attributes n tend to infinity. Let m be the main limit index. We drop the argument m for parameters, when doing so causes no confusion.

The notion of 'random database' extends the usual notion. In database theory, *relations* are considered, i.e. *sets* of tuples. In a random database, tuples can be identical. We treat a random database as a sequence of tuples (also referred to as *table*). Hence the number of different tuples can be less than m. Note that the notions of key and minimal key for random database correspond to those of *test* and *nonredundant test*, respectively, widely used in discrete mathematics (e.g., in pattern recognition, see Andreev (1984)). We show that the main asymptotic results for keys and minimal keys are extended in certain conditions to the case of a *random relation*, i.e. a database with different values of tuples. On occasion we adopt the convention that the random relation is termed $\mathcal{R}$ if it is the counterpart of R. Write $\mathcal{R} \models A$ if A is a key, and $\mathcal{R} \models_{\min} A$ if A is a minimal key in $\mathcal{R}$. Clearly, U is always a key in $\mathcal{R}$.

First we evaluate the functional dependency probability $P\{A \to B\}$ for a uniform-type random database. We derive asymptotic properties of keys and our main problem of interest, namely, minimal keys, in Sections 3 and 4, respectively. An asymptotic distribution of the length of a shortest key (i.e., a key with minimal length) is treated in Section 3. These results and the similar arguments based on the Poisson approximation technique (the Stein-Chen method) allow to investigate the *minimal key probability* $P\{\mathcal{R} \models_{\min} A\}$ for random databases. As the first but an important step we consider standard Bernoulli databases. For uniform random databases with different domains, we investigate the case when the minimal key probability is asymptotically equivalent to the corresponding key probability, i.e. almost all keys are minimal. We treat the extensions of the results for keys and minimal keys in random relations in Sections 3 and 4, respectively. Section 5 gives evidence to support the statements made in the previous sections.

2 Functional dependencies in random databases

Let A and $B \subseteq U \setminus A$ be sets of attributes in R. Denote by $N = N(A, B)$ the random number of pairs of tuples in a database R when the functional dependency property fails, i.e. $t(A) = t'(A)$ and $t(B) \neq t'(B)$. The distribution of $N(A, B)$ characterizes the *degree* of dependency between sets of attributes A and B. Clearly, $P\{A \to B\} = P\{N(A, B) = 0\}$. Henceforth in this section write $a = h(A)$, $b = h(B)$, and $M = m(m-1)/2$. In particular, if R is a standard Bernoulli database, then $a = |A|$ and $b = |B|$.

Further, for any tuple $t(U)$, denote by

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

i.e. $p(k(A)/k(B))$ is the conditional probability of the event $t(B) = k(B)$ when $t(A) = k(A)$. For such model, we denote the mean of $N(A,B)$ by

$$\lambda = \lambda(A,B) = \mathrm{E}[N(A,B)] = M \sum_{k(A),k(B)} p(k(A))\Big(1 - p(k(B)/k(A))\Big)p(k(B),k(A)) =$$
$$M\mathrm{E}[p(t(A))\Big(1 - p(t(B)/t(A))\Big)],$$

where $\mathrm{E}[\cdot]$ is the mathematical expectation for the joint distribution $\{p(k(A),k(B))\}$. As an example, for a uniform database, $\lambda = M2^{-a}(1-2^{-b})$.

Theorem 1 *Let R be a uniform-type database and $p(k(B)/k(A)) \le \delta < 1$. Suppose that $0 < \lambda \le ca$; then*

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

where $0 < c < 1/2\ln 2$ and $\gamma = 1/2 - c/\ln 2 > 0$.

Corollary 1 *Let R be a uniform random database. If $a - 2\log_2 m \to \alpha$ and $b \to \beta$, where $|\alpha| \le +\infty$ and $1 \le \beta \le +\infty$, then*

$$P\{A \to B\} \to \exp\{-2^{-(\alpha+1)}(1-2^{-\beta})\} \text{ as } m \to \infty.$$

We can interpret this asymptotic result in the following way. The size of a database very strongly indicates the length of possible candidates for left sides of functional dependencies. Therefore, for a given database with m tuples there is a subset $\mathcal{F}_0(B)$ of all possible functional dependencies $\mathcal{F}(B)$ with right side B which are likely. First of all the set $\mathcal{F}_0(B)$ is determined by the value of the parameter λ, since the probability $P\{A \to B\} \sim e^{-\lambda}$ as $m \to \infty$. This set is much smaller than $\mathcal{F}(B)$ and, with regard to the average case, there is a high probability that a functional dependency from $\mathcal{F}(B)$ which is valid in the random database, belongs to $\mathcal{F}_0(B)$. Thus heuristic algorithms, for example, which are used to check the validity of functional dependencies from $\mathcal{F}(B)$, should begin with constraints from $\mathcal{F}_0(B)$. In this case, they succeed much faster.

Theorem 1 shows that a functional dependency for a random database with sufficiently large m relates both with stochastic properties of an attribute set A and with stochastic relationship between attribute values in A and B. Also, the distribution of $t(A)$ defines the first factor, and the conditional distribution of $t(B)$ with a given $t(A)$ characterizes the latter.

3 Keys in random databases and relations

3.1 Random databases

The asymptotic results for the key probability $P\{R \models A\}$ are very similar to those for the functional dependency probability when $b = h(B) \to \infty$ as $m \to \infty$. However, in the case of a uniform random database there is also an explicit formula for $P\{R \models A\}$. Let $a = h(A)$ and $M = m(m-1)/2$ as above. We also introduce the random number of key condition violations (cf. the functional dependency case), $N = N(A) = |\{t_i(A) = t_j(A),\ i,j = 1,\dots,m, i < j\}|$. The distribution of $N(A)$ characterizes the capability of A to distinguish tuples in R. Clearly, $P\{R \models A\} = P\{N(A) = 0\}$ for a set of attributes $A \subseteq U$. Denote also the mean number of key condition violations by $\lambda =$

$\lambda(A) = \mathrm{E}[N(A)] = M\sum_{k(A)} p(k(A))^2 = M\mathrm{E}[p(t(A)]$, where $\mathrm{E}[\cdot]$ is the mathematical expectation for the distribution $\{p(k(A))\}$.

Theorem 2 *Let R be a uniform-type database.*

(i) If $0 < \lambda \le ca$, then

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

where $0 < c < 1/2\ln 2$ and $\gamma = 1/2 - c/\ln 2 > 0$;

(ii) If R is uniform then $P\{R \models A\} = \prod_{j=1}^{m-1}(1 - j2^{-a}) \le e^{-\lambda}$. If additionally, $0 < \lambda \le 2^{a(2-\gamma)/3}$, where $0 < \gamma < 2$, then

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

In the uniform case, $\lambda = M2^{-a}$, we also formulate the following

Corollary 2 *Let R be a uniform random database. If $a - 2\log_2 m \to \alpha$, where $|\alpha| \le +\infty$, then*

$$P\{R \models A\} \to \exp\{-2^{-(\alpha+1)}\} \text{ as } m \to \infty.$$

Remark Let $\nu = \nu(A)$ be the length of a shortest key in A, i.e. an integer random variable, which equals the length of a minimal key subset B in A. Therefore, $\{\nu(A) \le r\} = \{R \models A\}$ if $|A| = r$. For a standard Bernoulli database, we have $|A| = a$. Write $\alpha = a - 2\log_2 m$. It is a straightforward consequence of Theorem 2, that if $\alpha \to x/\ln 2 - 1$, then

$$P\{\nu(A) \le a\} = P\{\nu(A) - 2\log_2 m \le \alpha\} \to \exp\{-e^{-x}\} \text{ as } m \to \infty. \tag{2}$$

Thus, the normalized size of a shortest key in A has asymptotically the Gumbel (double exponential) distribution and the values of the random variable ν_m concentrate near $2\log_2 m$ (cf. Devroye (1984), Szpankowski (1991) for the height of a digital trees).

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

3.2 Random relations

Let R be a random database and $\mathcal{R}$ be the corresponding relation. By definition the only difference between R and $\mathcal{R}$ is that the latter *may* have identical tuples. Write $u = h(U)$. Clearly, $N(U) \le N(A)$ and therefore,

$$P\{\mathcal{R} \models A\} = P\{N(A) = 0/N(U) = 0\} = P\{N(A) = 0\}/P\{N(U) = 0\}. \tag{3}$$

It follows directly by definition that for uniform-type databases $\lambda(U)/\lambda(A) \asymp 2^{a-u}$. Now the main results about asymptotic behavior of key probability can be formulated also for random relations. Applying Theorem 2 and (3), and not striving for generality, we obtain the following

Theorem 3 *Let R be a uniform-type and $\mathcal{R}$ be the corresponding random relation.*

(i) If $0 < \lambda(A) \le ca$ and $0 < \lambda(U) \le cu$, then

$$P\{\mathcal{R} \models A\} = \exp\{-\lambda(A)(1 - \lambda(U)/\lambda(A))\}(1 + O(2^{-\gamma a})) \text{ as } m \to \infty,$$

where $0 < c < 1/2\ln 2$ *and* $\gamma = 1/2 - c/\ln 2 > 0$;

(ii) If $\lambda(A) \to \lambda_0$ *and* $u - a \to \infty$, *then* $P\{\mathcal{R} \models A\} \to e^{-\lambda_0}$ *as* $m \to \infty$, *where* $0 \le \lambda_0 \le +\infty$;

(iii) If R *is uniform, then* $P\{\mathcal{R} \models A\} = \prod_{j=1}^{m-1}(1 - j2^{-a})/(1 - j2^{-u})$.

Applying the same arguments as in Theorem 3, clearly, (2) is also valid for random relations if $u - a \to \infty$ as $m \to \infty$.

4 Minimal keys

In this section we derive our main asymptotic results for the minimal key probability $P\{R \models_{\min} A\}$, where A is a given set of attributes. For this *extreme* case in a sense, first we consider a standard Bernoulli model. If $D_i = \{0, 1, \ldots, d\}, i = 1, \ldots, n$, and $d > 2$, then the arguments are similar. By using the notation of the previous section we have the following

Theorem 4 *Let* R *be a standard Bernoulli database and* $0 < \lambda_0 < \lambda < ca$, *where* $0 < c < 1/2\ln 2$. *Then there exists* $\gamma > 0$ *such that for any sufficiently large* λ_0,

$$P\{R \models_{\min} A\} = e^{-\lambda}(1 - e^{-\lambda})^a(1 + O(2^{-\gamma a})) \ \textit{as}\ m \to \infty.$$

Now Theorem 4 allows to evaluate the asymptotic maximum value of the minimal key probability.

Proposition 1 *If* $0 < \lambda_0 \le \lambda$, *then*

$$P\{R \models_{\min} A\} = P(a) \le P_{\max}(a) \sim e^{-1}/(a+1) \ \textit{as}\ m \to \infty,$$

and also $P(a) = P_{\max}(a)$ *iff*

$$2\log_2 m - \log_2 \ln(2\log_2 m) - 1 + o(1) \le a \le 2\log_2 m - 1,$$

i.e. $\lambda = \ln(a+1)(1 + o(1))$ *as* $m \to \infty$.

This result can also now be compared with worst case complexity of minimal key systems. Thus if m is small or large enough then minimal key systems can contain only a very small number of minimal keys in average. Therefore, worst case complexity results are highly unlikely for these cases.

The result of Proposition 1 shows that the probability for a set of attributes A to be a minimal key tends to be zero as $a \to \infty$ and this property does not depend on relationship between the number of tuples m and the size of A as in the case of key (cf. Theorem 2). However, the following corollary shows that this relationship is important when we consider the conditional minimal key probability $P(a/K) = P\{R \models_{\min} A/R \models A\}$.

Corollary 3 *Let* $m^2 = 2^{a+1}(\ln a + d)$, *i.e.* $\lambda \sim \ln a + d$, *and* $b \to x$, *where* $|x| \le +\infty$, *and also* $b \ge \lambda_0 - \ln a+$, $\lambda_0 > 0$. *Then*

$$P(a/K) \to \exp\{-e^{-x}\}, \ \textit{as}\ m \to \infty.$$

Now it follows directly that if $\lambda = \ln a + d$ and $d \to +\infty$, then

$$P\{R \models_{\min} A\} \sim P\{R \models A\} \sim e^{-\lambda} \text{ as } m \to \infty. \tag{4}$$

Furthermore, the main results about asymptotic behavior of minimal key probability can be formulated for *random relations* too. It is evident, $\{R \models_{\min} A\} \subseteq \{N(U) = 0\}$ and therefore,

$$P\{\mathcal{R} \models_{\min} A\} = P\{R \models_{\min} A/N(U) = 0\} = P\{R \models_{\min} A\}/P\{N(U) = 0\}. \tag{5}$$

Applying Theorems 2, 4, and (5) we obtain the following

Theorem 5 *Let $\mathcal{R}$ be a random relation and the conditions of Theorem 4 be valid. Suppose $u = n \geq \delta \log_2 m$, where $3/2 < \delta < \infty$; then there exists $\gamma > 0$ such that for any sufficiently large $\lambda_0 > 0$,*

$$P\{\mathcal{R} \models_{\min} A\} = \exp\{-\lambda(1 - 2^{a-n})\}(1 - e^{-\lambda})^a(1 + O(e^{-\gamma a})) \text{ as } m \to \infty.$$

Remark The asymptotic results of Proposition 1 are also valid for random relations if $n - a \to \infty$ as $m \to \infty$.

Finally, let R now be a uniform random database and $|D_i| \geq 2$, for $i \in A$. Previous asymptotic result (4) for $P\{R \models_{\min} A\}$ can be generalized for this case as well. Write $d_i = |D_i| - 2 \geq 0$ for $i \in A$. Denote by $\pi_k = |A|^{-1}|\{i : d_i = k, i \in A\}|$ for $k = 0, \ldots, L$, where $L = \max_{i \in A} d_i \geq 0$. Note that $\sum_{k=0}^{L} \pi_k = 1$ and $\{\pi_k\}_0^L$ is a distribution of values d_i for $i \in A$. Denote by $g(z)$ the *generating function* for the distribution $\{\pi_k\}_0^L$ and whereby $g(z) := \sum_{k=0}^{L} z^k \pi_k$, $|z| \leq 1$. In particular, for a standard Bernoulli database, we have $L = 0, \pi_0 = 1$, and $g(z) = 1$. Let $\lambda = \ln a + d$, where $a = h(A)$. We introduce the following condition,

$$a e^{-\lambda} g(e^{-\lambda}) = e^{-d} g(e^{-d}/a) \to 0 \text{ as } m \to \infty. \tag{6}$$

Write $\pi_{\max} = \max_{1 \leq k \leq L} \pi_k$. Evidently, to ensure (6) it is sufficient, if $d \to +\infty$, since $|g(z)| \leq 1$ for $|z| \leq 1$; or if $e^{-d} \max(\pi_0, e^{-d}\pi_{\max}/a) \to 0$ as $m \to \infty$.

Proposition 2 *Let R be a uniform and (6) hold. If $\log_2(L + 2) = o(a)$ as $m \to \infty$ and $\lambda \leq 2^{\epsilon a}$, where $0 < \epsilon < 1/3$, then (4) is valid.*

In the following simple examples of non-trivial distributions we have $L \geq 1$.

Example 1 (Two-steps distribution) Let $\pi_k = 0, k = 1, \ldots, L - 1$ and $\pi_0 + \pi_L = 1, \pi_L > 0$. Then $g(z) = \pi_0 + z^L \pi_L$ and (6) applies iff $e^{-d} \max(\pi_0, \pi_L e^{-d - L \ln a}) \to 0$ as $m \to \infty$.

Example 2 (Binomial distribution) Let for any $k = 0, \ldots, L$, $\pi_k = \binom{L}{k} p^k q^{L-k}$, $p(m) = 1 - q(m) > 0$. Then $g(z) = (q + zp)^L$ and (6) applies iff $L \ln(q + e^{-d}p/a) - d \to -\infty$ as $m \to \infty$.

5 Proofs

Let $(\Omega, \mathcal{F}, P)$ be a standard probability space. For every $B \in \mathcal{F}$, denote by $\bar{B} := \Omega \setminus B$. Let $N = \sum_{\alpha \in \Gamma} I_\alpha$, where I_α is the indicator of the event C_α and $P\{I_\alpha = 1\} = P(C_\alpha)$, $\alpha \in \Gamma$. Write

$M = |\Gamma|$. There are classic limit theorems for independent indicator random variables. In general, C_α, $\alpha \in \Gamma$, are dependent and it needs a different technique. The Stein-Chen method has been developed for establishing Poisson approximation for sums of dependent indicator variables. We use one of the result of this approach following Barbour, Holst and Janson (1992).

Denote by $\Gamma_\alpha = \Gamma \setminus \{\alpha\}$, and $\lambda = \mathrm{E}[N] = Mq$. Suppose that Γ_α is partitioned into Γ_α^o, and Γ_α^i such that I_α and $\{I_\beta; \beta \in \Gamma_\alpha^i\}$ are independent. Write $m_o = |\Gamma_\alpha^o|$. The next proposition follows directly from Corollary 2.C.5., Barbour et al. (1992),

Proposition 3 *Let N be the sum of indicators, $q = P\{I_\alpha = 1\}$, and $E\ [I_\alpha I_\beta] \leq s$ for $\alpha \in \Gamma$ and $\beta \in \Gamma_\alpha^o$. Then*

$$|P\{N=0\} - e^{-\lambda}| \leq \lambda/M(m_o + 1 + m_o s/q^2).$$

We use also the following elementary inequality

$$\ln(1+x) \leq x, \quad |x| < 1. \tag{7}$$

1. Functional dependencies in random databases

Proof of Theorem 1. To apply the Stein-Chen method, note that in our case $\Gamma = \{(i,j) : i,j = 1,\ldots,m, i < j\}$, where $|\Gamma| = M = m(m-1)/2$. Further, write $N(A,B) = \sum_{\alpha\in\Gamma} I_\alpha$, where I_α is the indicator of the event $C_\alpha = C_{ij} = \{t_i(A) = t_j(A), t_i(B) \neq t_j(B)\}$. The random variables $I_\alpha, \alpha \in \Gamma$, are identically distributed, but dependent. In addition, $\Gamma_\alpha^o = \{(i,k),(l,j); k \neq j,\ l \neq i\}$, and $|\Gamma_\alpha^o| = 2(m-1)$ for $\alpha = (i,j)$. Moreover, $P\{I_\alpha = 1\} = q$ and $\mathrm{E}[I_\alpha I_\beta] = s$ for all $\alpha \in \Gamma$ and $\beta \in \Gamma_\alpha^o$. Now, Proposition 3 yields

$$|P\{N=0\} - e^{-\lambda}| \leq \lambda/M\Big(2(m-1) + 1 + 2(m-1)s/q^2\Big) = 4\lambda/m(1/2 + s/q^2).$$

Since R is uniform-type, we have $s \asymp q^2$ and therefore,

$$|P\{N=0\} - e^{-\lambda}| \leq C_1\lambda/m \leq C_2 e^{-\lambda}\exp\{\lambda + \ln\lambda - 1/2\ln 2a\}, \tag{8}$$

where $0 < C_1, C_2 < \infty$, and the assertion follows. □

Proof of Corollary 1. For a finite value of λ_0 the assertion follows directly by Theorems 1 and (8). If $\lambda_0 = 0$ we can use the following estimate

$$P\{R \models A\} = 1 - P\{N \geq 1\} = 1 - P\{\bigcup_{i,j} C_{ij}\} \geq 1 - \lambda \to 1 \text{ as } m \to \infty.$$

If $\lambda_0 = \infty$, then for every $\epsilon > 0$ and $\lambda_\epsilon > 0$, we find m_ϵ such that $m_\epsilon(m_\epsilon - 1)/2P\{C_{12}\} = \lambda_\epsilon$. Then for every $n > n_\epsilon, \lambda > \lambda_\epsilon$, and $m > m_\epsilon$, we get

$$P\{A \to B\} = P\{N = 0\} = p(m) \leq p(m_\epsilon).$$

Applying now the assertion of the theorem, we obtain

$$P\{A \to B\} \leq p(m_\epsilon) \leq e^{-\lambda_\epsilon} + \epsilon/2 < \epsilon$$

for every $m > m_\epsilon$. This completes the proof. □

2. Keys in random databases and relations

The proofs in this section are similar to those we have for functional dependencies. Applying the same notation as above we have $N(A) = \sum_{\alpha\in\Gamma} I_\alpha$, where I_α is the indicator of the event $C_\alpha = C_{ij} = \{t_i(A) = t_j(A)\}$. Now the proofs of Theorem 2 (i) and Corollary 2 repeat those of Theorem 1 and Corollary 1, respectively.

In the uniform case we use the following combinatorial argument. One can find $2^a = \prod_{i\in A} |D_i|$ variants for the first row $t(A)$ in R. For the second raw there are $2^a - 1$ variants and so on. There exist 2^{ma} different matrices $m \times |A|$, and whence,

$$p(m) = P\{R \models A\} = \frac{1}{2^{ma}} \prod_{j=0}^{m-1} (2^a - j) = \prod_{j=1}^{m-1} (1 - j2^{-a}) \le e^{-\lambda}.$$

Applying now (7) yields $e^{-\rho_m} \le e^\lambda p(m) \le 1$, where

$$\rho_m = \sum_{j=1}^{m-1} \frac{j^2 2^{-2a}}{1 - j2^{-a}} \le \frac{2^{-2a}}{1 - m2^{-a}} \sum_{j=1}^{m-1} j^2 \le C_0 \frac{\lambda^2}{m},$$

for some $0 < C_0 < \infty$, and (ii) follows, since $p(m) = e^{-\lambda}(1 + \delta_m)$, with $|\delta_m| \le 1 - e^{-\rho_m} \le \rho_m$. □

3. Minimal keys

Proof of Theorem 4. R be a standard Bernoulli database and whence $|A| = a$. Denote by $A_k = A\backslash\{k\}$, $\mathcal{A}_k = \{R \models A_k\}$, $k = 1, \dots, a$, and $\mathcal{A} = \{R \models A\}$. Then it follows directly by the definition of a minimal key that $P\{R \models_{\min} A\}$ can be represented in the following form,

$$P\{R \models_{\min} A\} = P(\mathcal{A}) - \sum_{j=1}^{a} (-1)^{j-1} \binom{a}{j} P(\mathcal{A}_1 \dots \mathcal{A}_j). \tag{9}$$

In fact, $\{R \models_{\min} A\} = \{R \models A\} \bigcap_{j=1}^a \overline{\{R \models A_j\}} = \mathcal{A} \bigcap_{j=1}^a \bar{\mathcal{A}}_j$. Therefore, $P\overline{\{R \models_{\min} A\}} = P(\bar{\mathcal{A}}) + P(\bigcup_{j=1}^a \mathcal{A}_j)$, since $\bar{\mathcal{A}} \cap \mathcal{A}_j = \emptyset$ for $j = 1, \dots, a$. Then (9) follows by using the inclusion-exclusion formula.

For the proof of the theorem, first we apply the Poisson approximation technique to evaluate the probability $P(\mathcal{A}_1 \dots \mathcal{A}_j)$; then the Bonferroni inequality yields the assertion.

The events $\mathcal{A}_k, k = 1, \dots, a$, are strongly dependent. Hence, we represent the event $E_p = \mathcal{A}_1 \dots \mathcal{A}_p$ as follows, $E_p = \bigcap_{i<j} \bar{D}_{ij}$, where $D_{ij} = \bigcup_{l=1}^p D_{ij}^l$ and $D_{ij}^l = \{t_i(A_l) = t_j(A_l)\}$.

Lemma 1 *Let* $1 \le p \le T$ *and* $0 < \lambda < ca$, *where* $1 \le T \le \max(ca/\lambda - 1, a)$ *and* $0 < c < 1/2\ln 2$. *Then there exists* $\gamma > 0$ *such that*

$$P(\mathcal{A}_1 \dots \mathcal{A}_p) = e^{-(p+1)\lambda}(1 + O(e^{-\gamma a})) \text{ as } m \to \infty.$$

Proof of Lemma 1. Denote by I_α the indicator of the event $D_\alpha = D_{ij}, \alpha \in \Gamma$, and $N(A) = \sum_{\alpha\in\Gamma} I_\alpha$. First consider $P(D_{ij}) = P(D_{12}), i, j = 1, \ldots, m, i \neq j$. We find $P(D_{12})$ by using again the inclusion-exclusion formula. Namely, $|A_1| = a - 1$,

$$P(D_{12}^l) = P(D_{12}^1) = P\{t_1(A_1) = t_2(A_1)\} = 2^{-(a-1)}.$$

If $l_1 \neq l_2$ and $k \geq 2$, then $D_{12}^{l_1} D_{12}^{l_2} = \{t_1(A) = t_2(A)\}$ and therefore,

$$P(D_{12}^{l_1} \ldots D_{12}^{l_k}) = P(D_{12}^{l_1} D_{12}^{l_2}) = P\{t_1(A) = t_2(A)\} = 2^{-a}. \tag{10}$$

Thus, the inclusion-exclusion formula yields

$$P(D_{12}) = \sum_{l_1} P(D_{12}^{l_1}) - \sum_{l_1<l_2} P(D_{12}^{l_1} D_{12}^{l_2}) + \cdots =$$

$$\frac{p}{2^{a-1}} - \binom{p}{2}\frac{1}{2^a} + \binom{p}{3}\frac{1}{2^a} + \cdots + (-1)^{p-1}\binom{p}{p}\frac{1}{2^a} = \frac{2p+1-p}{2^a} - (1-1)^p \frac{1}{2^a} = \frac{p+1}{2^a},$$

and therefore,

$$\lambda_p = E[N(A)] = M(p+1)2^{-a} = (p+1)\lambda.$$

To estimate the probability $P(D_\alpha D_\beta)$ when $D_\beta \in \Gamma_\alpha^o$, we use again the inclusion-exclusion formula for the events $C_{kl} = D_{12}^k D_{13}^l$ for $k, l = 1, \ldots, a$. Then

$$P(D_\alpha D_\beta) = P(D_{12} D_{13}) = P(\bigcup_{k,l} D_{12}^k D_{12}^l) = P(\bigcup_{k,l} C_{kl}).$$

If $k = l$, then

$$P(D_{12}^1 D_{12}^1) = P\{t_1(A_1) = t_2(A_1), t_1(A_1) = t_3(A_1)\} = s_1,$$

where $s_1 = 2^{-2(a-1)}$. If $k \neq l$, then

$$\begin{aligned} P(D_{12}^1 D_{12}^2) = P\{t_1(A_1) = t_2(A_1), t_1(A_2) = t_3(A_2)\} &= \\ P\{t_1(A_{12}) = t_2(A_{12}), t_1(A_{12}) = t_3(A_{12})\} P\{t_1(2) = t_2(2), t_1(1) = t_3(1)\} &= s_1, \end{aligned}$$

where $A_{kl} = A \setminus \{k, l\}$, and therefore,

$$\sum_{k,l} P(C_{k,l}) = (p + 2\binom{p}{2})s_1 = p^2 s_1.$$

Applying again (10) yields that if $k_1 = k_2$ or $l_1 = l_2$, then $P(C_{k_1,l_1} C_{k_2,l_2}) = 2^{-2(a-1)-1} = s_1/2$, otherwise $P(C_{k_1,l_1} C_{k_2,l_2}) = 2^{-2(a-2)} = s_1/4$. Let order the set $\{(k_i, l_i) : (k_i, l_i) \neq (k_j, l_j),\ i \neq j, \quad i, j = 1, \ldots, \binom{p^2}{2}\}$, and $T_i = C_{(k_i,l_i)}$, $i = 1, \ldots, \binom{p^2}{2}$. Hence

$$\sum_{i<j} P(T_i T_j) = \sum_{i<j} P(C_{k_i,l_i} C_{k_j,l_j}) = 2p\binom{p}{2} s_1/2 + \left(\binom{p^2}{2} - 2p\binom{p}{2}\right) s_1/4.$$

Finally, we obtain

$$P(D_{12} D_{13}) = P(\bigcup_{k,l} C_{kl}) = (p-1)^2 s_1/4 + p^2 s_1 - p(p-1)s_1 \leq C p^2 s_1 \tag{11}$$

for some $0 < C < \infty$. Thus, from Proposition 3 and (11) we get

$$|P(E_p) - e^{-\lambda_p}| \le C_1 \lambda_p / m,$$

where $\lambda_p = (p+1)\lambda \le (T+1)\lambda \le ca$, and $0 < c < 1/2\ln 2$. The assertion follows now as in Theorem 2(ii). □

Further, we represent the asymptotic function in the following inclusion-exclusion form

$$e^{-\lambda}(1-e^{-\lambda})^a = \sum_{j=0}^{a} (-1)^j \binom{a}{j} e^{-(j+1)\lambda}.$$

Hence, the Bonferroni inequality (see e.g., Bender (1974)) implies

$$|P\{R \models_{\min} A\} - e^{-\lambda}(1-e^{-\lambda})^a| \le \binom{a}{T}(e^{-(T+1)\lambda} + P(E_T)) +$$
$$\sum_{j<T} \binom{a}{j} |e^{-(j+1)\lambda} - P(E_T)| \le 2\binom{a}{T} e^{-(T+1)\lambda} + \sum_{j\le T} \binom{a}{j} R_j(m) = e^{-\lambda}(1-e^{-\lambda})^a (I_1 + I_2)$$

for every $T = 1, \ldots, a$, First consider I_1. Let T be as in Lemma 1 and choose $c_1 > 0$ such that $T = c_1 a/\lambda \le \lfloor (ca/\lambda - 1) \rfloor$, and $\lambda \ge \lambda_0$. Then for a sufficiently large λ_0 we have $T \le a$. Applying (7) and Stirling's formula we obtain

$$\begin{aligned} I_1 \le a^T/T! e^{-\lambda T}(1-e^{-\lambda})^{-a} &\le C_1 \exp\{T\ln(a/T) + T - \lambda T + ae^{-\lambda}\} \le \\ &C_2 \exp\{ac_1\lambda^{-1}\ln(\lambda/c_1) - ac_1 + ac_1/\lambda + ae^{-\lambda}\} \le \\ &C_3 \exp\{ac_1{\lambda_0}^{-1}\ln\lambda_0 - c_1 a + ac_1/\lambda_0 + ae^{-\lambda_0}\} \le e^{-\gamma_1 a}, \end{aligned} \tag{12}$$

for some γ_1 and $C_i > 0$, $i = 1, \ldots, 4$, and sufficiently large $\lambda_0 > 0$.

In the case of I_2 combining Lemma 1 and (7), we get

$$\begin{aligned} I_2 \le (1-e^{-\lambda})^{-a} e^{-\gamma a} \sum_{j=0}^{a} e^{-j\lambda} \binom{a}{j} &\le \exp\{-\gamma a - a\ln(1-e^{-\lambda}) + ae^{-\lambda}\} \le \\ &\exp\{-\gamma a + 2ae^{-\lambda_0}\} \le e^{-\gamma_2 a}, \end{aligned} \tag{13}$$

for some $\gamma_2 > 0$ and sufficiently large λ_0. Now the assertion follows by (12) and (13). □

Proof of Proposition 1. First, let $\lambda > \lambda_0 > 0$, as in Theorem 4, then we have

$$P(a) \sim e^{-\lambda}(1-e^{-\lambda})^a \le e^{-1}/(a+1) = P_{\max}(a),$$

since the function $f(x) = x(1-x)^a, x > 0$, has the unique maximum point $x = 1/(a+1)$, i.e. $\lambda = \ln(a+1) + \mathrm{o}(1)$ as $m \to \infty$. Namely,

$$f_{\max} = f(1/(a+1)) = 1/(a+1)(1 - 1/(a+1))^a \sim e^{-1}/(a+1) \text{ as } a \to \infty.$$

If $\lambda > ca$, then $P(a) \le e^{-\lambda} \le e^{-ca}$. Finally, the asymptotic estimates in the assertion can be easily checked. □

Proof of Corollary 3. If $m^2 = 2^{a+1}(\ln a + d)$, then $m2^{-a/2} \to \infty$ and

$$\lambda = M2^{-a} = (\ln a + d)(1 + \mathrm{O}(1/m)) = \ln a + c \geq \lambda_1,$$

where $c = d(1 + \mathrm{O}(e^{-\gamma a}))$ as $m \to \infty, \lambda_1 > 0$. The assertion follows now by applying Theorems 2 and 4, since

$$P(a/K) = P\{R \models_{\min} A, R \models A\}/P\{R \models A\} = P\{R \models_{\min} A\}/P\{R \models A\}.$$

□

Proof of Proposition 2. Denote by $p_i(a) = P\{R \models A_i\}$ and $p(a) = P\{R \models A\}$ for $i \in A$. By the definition of a minimal key as in the proof of (9) we obtain

$$p(a) \geq P(a) \geq p(a) - \sum_{i \in A} p_i(a) = p(a)(1 + \delta_m).$$

To prove the assertion it is sufficient to verify that $\delta_m \to 0$ as $m \to \infty$. Write $a_i = a - |D_i|$ and $\lambda_i = M2^{-a_i} = \lambda|D_i|$ for $i \in A$. By assumption

$$m \sim \sqrt{\lambda} 2^{\frac{1}{2}a} \leq 2^{\frac{a}{2}(1+\epsilon)} \leq 2^{(\frac{2}{3}-\gamma)a}, \gamma > 0,$$

where $a_i \sim a$ as $m \to \infty$. Thus $\lambda_i \leq 2^{\epsilon_1 a_i}$, $0 < \epsilon_1 < 2/3$, uniformly in $i \in A$. Hence, the claims of Theorem 2 (ii) hold uniformly in $i \in A$. We observe that there exists $\gamma_0 > 0$ such that

$$p_i(a) = e^{-\lambda_i}(1 + \mathrm{O}(e^{-\gamma_0 a_i})) = e^{-(d_i+2)\lambda}(1 + \mathrm{O}(e^{-\gamma_0 a})),\ p(a) = e^{-\lambda}(1 + \mathrm{O}(e^{-\gamma_0 a})),$$

as $m \to \infty$ uniformly in $i \in A$. Finally,

$$|\delta_m| = 1/p(a) \sum_{i \in A} p_i(a) = \Big(1 + \mathrm{O}(e^{-\gamma_0 a})\Big) e^{-\lambda} \sum_{i \in A} e^{-d_i \lambda} \sim$$

$$a e^{-\lambda} \sum_{k=0}^{L} e^{-k\lambda} \pi_k = a e^{-\lambda} g(e^{-\lambda}) = e^{-d} g(e^{-d}/a) \to 0 \text{ as } m \to \infty,$$

and the assertion follows. □

6 Conclusion

Worst case results on sets of minimal keys state that theirs sizes are exponential according to the number of attributes. For instance, the worst case number for sets of minimal keys is $O(2^n n^{-\frac{1}{2}})$, where n is the number of attributes. It is of interest whether these results are valid also for the average case. The proofs of worst case results are based on the length of minimal keys or on the length of left sides of functional dependencies. Thus, if we can prove for the average case that the length is rather small compared with the number of attributes then we can assume that the numbers are not exponential. In this paper, tables are considered. Tables are sequences of tuples. We could develop a theory of key length for tables. Relations are a special case of tables where tuples are different. Our main results apply on tables as well as on relations. We have shown that depending on the size of relations almost all minimal keys have a length which mainly depends on the size of the relation. Otherwise, the minimal key length probability is exponentially small compared with the number of minimal keys of the derived length. Thus, we have shown that for a large variety of relations the exponential complexity of sets of minimal keys is rather unusual. Further, if we have found a key in a relation and this key has the length derived from the size of the relation then this key is probably a minimal key. The similar results can be shown for functional dependencies. Furthermore, our results show that the size of the domains does not change the general picture. The presented research is only the first step. There are several problems left open in this paper. One of the main problems is the evaluation of the mean number of minimal keys. Relations on n attributes can be grouped depending on their size. In this case, different scenarios can be discussed according to the number of tuples in a relation and compared with the number of attributes: keys in rather small relations, keys in rather large relations and keys in relations in the small spectrum of an exponential size of the number of attributes. The first two cases are covered by our results. The last case leads to large sets of minimal keys. Furthermore, we can group databases into such which size does not change too much and such which size is changing to a larger extend. The behavior of key systems in the case of the last group is another open problem. We consider the uniform and uniform-type models, but our conjecture is that the main results are also valid for more general probabilistic models. Statistical estimation of database parameters for the developed models is the next stage of investigation.

Acknowledgments

The authors would like to thank the referees for the valuable remarks and comments on an earlier version of the paper. In particular due to the remarks of the reviewers the difference of our random database model (based on tables) and the usual relational database model (based relations) has been made transparent. Finally, we are very thankful to G. Gottlob and M. Vardi for their efforts and their help.

References

1. Albrecht, A., Altus, M., Buchholz, E., Düsterhöft, A. and Thalheim, B. (1995). The rapid application and database development (RADD) workbench - A comfortable database design tool. In: *Proc. CAiSE 95* (eds. Iivari, J., Lyytinen, K. and Rossi, M.), Lecture Notes in Comput. Science **932**, Springer, 327-340.

2. Andreev, A. (1984). On asymptotic behavior of the number of nonredundant tests for almost all tables, *Problemyi Kibernetiki* **41**, 117-142. (in Russian)

3. Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation.* Clarendon Press, Oxford.

4. Beeri, C., Dowd, M., Fagin, R. and Statman, R. (1984). On the structure of Armstrong relations for functional dependencies. *Journ. of ACM* **31**, 30-46.

5. Békéssy, A., Demetrovics, J., Hannák, L., Frankl, P. and Katona, G. (1980). On the number of maximal dependencies in a database relation of fixed order. *Discrete Math.* **30**, 83-88.

6. Bender, E. A. (1974). Asymptotic methods in enumeration. *SIAM Review* **16**, 485-515.

7. Devroye, L. (1984). A probabilistic analysis of the height of tries and of the complexity of trie sort. *Acta Informatica* **21**, 229-237.

8. Gottlob, G. (1987). On the size of nonredundant FD-covers. *Information Processing Letters* **24**, 355-360.

9. Mannila, H. and Räihä, K.-J. (1992). *The design of relational databases*, Addison-Wesley, Amsterdam.

10. Seleznjev, O. and Thalheim, B. (1988). On the number of minimal keys in relational databases over nonuniform domains. *Acta Cybernetica* **8**, 267-271.

11. Szpankowski, W. (1991). On the height of digital trees and related problems. *Algorithmica* **6**, 256-277.

12. Thalheim, B. (1987). On the number of keys in relational databases. In: *Proc. FCT-87-Conf.*, Kazan, Lecture Notes in Comput. Science **278**, Springer, 448 - 455.

13. Thalheim, B. (1989). On semantic issues connected with keys in relational databases permitting null values. *Journ. Information Processing and Cybernetics, EIK* **25**, 11-20.

14. Thalheim, B. (1991). *Dependencies in Relational Databases.* Teubner Verlag, Leipzig.

15. Traub, J. F., Wasilkowski, G. W. and Woźniakowski, H. (1988). *Information-based Complexity.* Academic Press, San Diego.