Query Optimization on Random Databases

Silviu Laurențiu VASILE

University of Bucharest, Faculty of Mathematics and Computer Science,
Academiei 14, RO-010014, Bucharest, Romania

Institute of Mathematical Statistics and Applied Mathematics,
Calea 13 Septembrie 13, RO-050711, Bucharest, Romania

vsl@fmi.unibuc.ro

Abstract: We analyze the cardinality of the ε-join operation into a random database for tables having column values following some representative types of probability distributions (uniform distribution, exponential distribution and normal distribution). The goal of our research is to use our analysis to make some query optimization on random databases. Some numerical results are studied and the conclusions we have drawn as a result of the practical experiments are presented.

Keywords: Random database, Join, Cardinality, Query optimization.

1. Introduction

The relational data model is one of the main database models and the basis for most existing database management systems. It is based on the first-order predicate logic which was first formulated and proposed in 1969 by Edgar F. Codd [3], [4]. The relational database model is the most popular data model since it is very simple and easily understandable by the information systems professionals and by the end users. In the relational model of a database, all users' data is represented in terms of tuples (records) which are grouped into relations (tables). A database organized in terms of the relational model is called a relational database. The rows of relations (relational matrices) 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. A large variety of algorithms used in database technology have as a goal the identification of tuples through keys. Examples in this direction are algorithms for selection, joining, constructing and maintaining tuples. These algorithms are as simple as search algorithms if key indexes are used. Therefore, keys and minimal keys are absolutely fundamental to database models.

Analysts often need to work with large datasets in which some of the data is uncertain [2]. This is the case when the data is connected to hypothetical or future events. For example, the data of interest might be the customer order sizes for some product under a hypothetical price increase of 5%. Data uncertainty is frequently modeled as a probability distribution over possible data values. These distributions are usually specified by means of a complex stochastic model. Various system characteristics of interest to the analyst can be viewed as answers to queries over the uncertain data values [9], [10]. Because the data is uncertain, there is a probability distribution over the possible results of running a given query, and the analysis of the underlying stochastic model is equivalent to studying the features (mean, variance, and so forth) of the query-result distribution. The query-result distribution is often very complex, and must be analyzed using Monte Carlo methods [5]. Such analysis is important to assessing for example the enterprise risk, as well as making decisions under uncertainty.

Random databases are databases where the value of some attributes or the presence of some records are uncertain and known only with some probability. Applications of random databases can be found in many areas such as information extraction, Radio-frequency identification (RFID) and scientific data management, data cleaning, data integration, and financial risk assessment.

At present Data models and databases for uncertain and/or probabilistic data are a hot topic in data management research.

The traditional relational databases are deterministic. Every record stored in the database is meant to be present with certainty, and fields in that record have a precise value.

The theoretical foundations of relational databases are if a logical sentence is true or false, if a record is, or is not, in a relation, or in a query result [1], but they cannot make any less precise statement. Today, however, data management needs to include new data sources, where data are uncertain [8], and which are difficult or impossible to model with traditional database systems. In this situation we have the following question: what do the traditional relational operators become in the case of random databases?

Definition: Consider $D_1, D_2, \ldots, D_n$ finite domains, not necessarily disjoint. The Cartesian product $D_1 \times D_2 \times \ldots \times D_n$ of the domains $D_1, D_2, \ldots, D_n$ is defined by the set of the tuples $(V_1, V_2, \ldots, V_n)$ where $V_1 \in D_1, V_2 \in D_2, \ldots, V_n \in D_n$.

Definition: A relation *R* on the sets $D_1, D_2, \ldots, D_n$ is a subset of the Cartesian product $D_1 \times D_2 \times \ldots \times D_n$.

We can represent a relation by a bi-dimensional table in which each line corresponds to a tuple and each column corresponds to a domain in the Cartesian product. A column corresponds to an attribute. The number of attributes defines the relation's degree, and the number of tuples in the relation defines the relation's cardinality. The relational databases are perceived by the users as a set of tables. We consider a table as a representation of a relation.

We consider a distance $d(x, y)$ for the elements in the two sets D_i and D_j, where $i, j \in \{1, 2, ..., n\}$, which are assumed to be subsets of a metric space where the distance *d* is defined [11]. Many ε-join techniques use as a metric the standard Euclidean distance. We denote by $B_\varepsilon(x)$ the ball centered in *x* and with radius ε. In this context, the operations in the relational algebra will be based upon approximations in order to deal with uncertainty. The name of the operation we intend to investigate are *ε-equi-join.*

Definition: Consider two relations *R* and *S*. Then, the ε-operation above can be described as follows: $join_\varepsilon(R,S) = \{(x,y) \in R \times S \mid d(x_A, y_B) \le \varepsilon\}$, where x_A takes all values of column *A* that belongs to relation *R* and y_B take all values of column *B* that belongs to relation *S*. We denote by $N_\varepsilon(join_\varepsilon(R,S))$ the number of lines in the result of the ε-join operation [11]. Whenever the context in which we refer these values are clear, we denote them by N_ε for simplicity.

During our studies concerning the random databases, which involved different experiments with data sets and analysis of the results, we have noticed the existence of some relations between the cardinalities of the approximate join (ε-join) operation in case of various probability distributions of the columns of tables.

Our empirical observations encouraged to find the proofs of the relations we conjectured. We were able to state them as theoretical results. The practical experiments dealt with tables containing at least 1000 records, with columns having values distributed uniformly, exponentially and, respectively, normally.

Such relations have a great impact in random database query optimization, as we have shown in [13].

2. Problem Formulation

We denote by N_U the number of rows obtained from the *ε-join* operation between two tables using columns whose values have uniform probability distribution. Analogous we denote by N_{Exp} and N_N the cardinal of the *ε-join* operation between two tables using columns whose values have exponential and normal distributions.

The random variables N_U, N_{Exp} and N_N are taking values positive integers. Our aim is to prove the following inequalities:

$$E[N_U] > E[N_{Exp}] > E[N_N]. \quad (1)$$

Consider *R(A)* and *S(B)* the projections of the attributes *A* and *B* from the relations *R,* respectively *S*. Let *X*, *Y* be the common attribute of the two relations, $X \in A$ and $Y \in B$. Let $\{X_1, X_2, ..., X_m\}$ and $\{Y_1, Y_2, ..., Y_n\}$ be the sets of distinct values of the attributes *X*, respectively *Y*. The elements of these sets are independent and identically distributed, with the probability distributions *F*, respectively *G*.

We denote by $N_{m,n}^{\varepsilon} = |\{(i,j) \,||\, X_i - Y_j| \le \varepsilon\}|$ the number of records resulting from the ε-join

operation [1]. Graphically, this is the number of points in the region A_ε in Figure 1.

In this picture, the region A_ε is the band symmetric to the first bisectrix, of width $\frac{2\varepsilon}{\sqrt{2}}$.

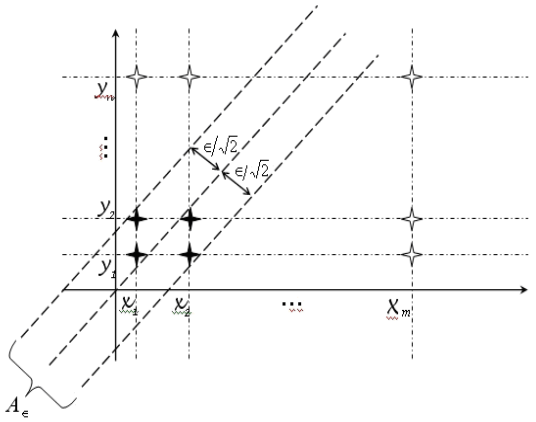

Figure 1. The ε-join of the sets of distinct values of the attributes X and Y.

Proposition 1. Let *X*, *Y* be the ε-join attributes between the relations *R* and *S*. The values of these attributes follow the probability distributions *F*, respectively *G*. Then:

$$\frac{N_{m,n}^{\varepsilon}}{m\cdot n}\longrightarrow(F\otimes G)(A_\varepsilon). \tag{2}$$

Prof. We use the Glivenko-Cantelli theorem [12]. This states that a series of empirical distribution functions F_k (k is the sample size) $F_k(x)=\frac{|\{j\le k\,|\,X_j\le x\}|}{k}$, converge almost sure to the theoretical distribution function F $\left(F_k\Rightarrow F\right)$.

A corollary of this theorem states that, for two series of distribution functions $\{F_n\}_n,\{G_n\}_n$ such that $F_n\Rightarrow F, G_n\Rightarrow G$ the following convergence holds:

$$F_n\otimes G_n\Rightarrow F\otimes G. \tag{3}$$

For the case of the empirical distribution functions corresponding to the two attributes we have that:

$$F_n\otimes G_n=\frac{|\{i\le m\,|\,X_i\le x\}|}{m}\cdot\frac{|\{j\le n\,|\,Y_j\le x\}|}{n}= \\ =\frac{N_{m,n}^{\varepsilon}}{m\cdot n}. \tag{4}$$

From the relations (3) and (4), the conclusion is immediate. □

If the ε-join attributes follow the distribution $\mathrm{F}=\mathrm{G}=\mathrm{U}(0,\,1)$, then we have to determine the limit in *proposition 1* for $\varepsilon=0.01$. Graphically this area is:

$$Area(A_\varepsilon)=1-2\cdot\frac{(1-\varepsilon)^2}{2}=1-1+2\varepsilon-\varepsilon^2=\varepsilon(2-\varepsilon). \tag{5}$$

For the value $\varepsilon=0.01$, the area of this region is $Area(A_\varepsilon)=0.01\cdot1.99=0.0199$. This means that, for $m=n=1000$, the value of the ε-join operation's cardinality $N_{m,n}^{\varepsilon}$ can be estimated by $0.0199\cdot m\cdot n=19900$ records.

2.1 Inequalities between sets that are ε-joins for the case of uniform and exponential distributions

In the following, we denote by $p_F(\varepsilon)=(F\otimes F)(A_\varepsilon)$ the limit in proposition 1.

Using (5), for *U = U(0, 1)*, we obtain:

$$p_U(\varepsilon)=2\varepsilon-\varepsilon^2. \tag{6}$$

We want to establish the relation between $p_{\mathrm{U}}(\varepsilon)$ and $p_{Exp}(\varepsilon)$. In the following proposition we calculate the value of $p_{Exp}(\varepsilon)$.

Proposition 2. $p_{\mathrm{Exp}}(\varepsilon)=1-e^{-\varepsilon}$.

Prof. From *Figure 1*, we have:

$$Area(A_\varepsilon)=\int_0^{\varepsilon}F(x+\varepsilon)\mathrm{d}F(x)+\int_{\varepsilon}^{\infty}[F(x+\varepsilon)-F(x-\varepsilon)]\mathrm{d}F(x)= \\ =\int_0^{\varepsilon}(1-e^{-x-\varepsilon})\mathrm{d}F(x)+\int_{\varepsilon}^{\infty}(1-e^{-x-\varepsilon}-1+e^{-x+\varepsilon})\mathrm{d}F(x)= \\ =\int_0^{\varepsilon}(1-e^{-x}\cdot e^{-\varepsilon})\cdot e^{-x}\mathrm{d}x+\int_{\varepsilon}^{\infty}(-e^{-x}\cdot e^{-\varepsilon}+e^{-x}\cdot e^{\varepsilon})\cdot e^{-x}\mathrm{d}x= \\ =\int_0^{\varepsilon}e^{-x}\mathrm{d}x-e^{-\varepsilon}\cdot\int_0^{\varepsilon}e^{-2x}\mathrm{d}x-e^{-\varepsilon}\cdot\int_{\varepsilon}^{\infty}e^{-2x}\mathrm{d}x+e^{\varepsilon}\cdot\int_{\varepsilon}^{\infty}e^{-2x}\mathrm{d}x= \\ =-e^{-x}\Big|_0^{\varepsilon}+e^{-\varepsilon}\cdot\frac{e^{-2x}}{2}\Big|_0^{\varepsilon}+e^{-\varepsilon}\cdot\frac{e^{-2x}}{2}\Big|_{\varepsilon}^{\infty}-e^{\varepsilon}\cdot\frac{e^{-2x}}{2}\Big|_{\varepsilon}^{\infty}= \\ =-e^{-\varepsilon}+1+e^{-\varepsilon}\cdot\frac{e^{-2\varepsilon}}{2}-\frac{e^{-\varepsilon}}{2}-\frac{e^{-\varepsilon}}{2}\cdot e^{-2\varepsilon}+\frac{e^{\varepsilon}}{2}\cdot e^{-2\varepsilon}= \\ =1-e^{-\varepsilon}+\frac{e^{-3\varepsilon}}{2}-\frac{e^{-\varepsilon}}{2}-\frac{e^{-3\varepsilon}}{2}+\frac{e^{-\varepsilon}}{2}=1-e^{-\varepsilon}.$$

□

Proposition 3. For small value of ε the following inequality holds: $p_U(\varepsilon)>p_{Exp}(\varepsilon)$.

Prof. From the relation (6) and the proposition 2 result that we want to prove that $2\varepsilon-\varepsilon^2>1-e^{-\varepsilon}$. We develop $e^{-\varepsilon}$ in series:

$$2\varepsilon-\varepsilon^2>1-\left(1-\varepsilon+\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{6}+...\right)\Leftrightarrow$$

$$\Leftrightarrow 2\varepsilon-\varepsilon^2>\varepsilon-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{6}-\frac{\varepsilon^4}{24}+...\Leftrightarrow$$

$$\Leftrightarrow \varepsilon>\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{6}-\frac{\varepsilon^4}{24}+...$$

We divide the last inequality by ε and thus we get:

$$1>\frac{1}{2}+\frac{\varepsilon^2}{6}-\frac{\varepsilon^3}{24}+....\quad (7)$$

Relation (7) is true for a small ε. For such values of ε, we get:

$$p_U(\varepsilon)>p_{Exp}(\varepsilon).\quad (8)$$

2.2 The cases of one-dimensional and bi-dimensional distributions

The practical observations suggested the following inequality: $p_U(\varepsilon)>p_{Exp}(\varepsilon)>p_N(\varepsilon)$. In order to verify the second part of this inequality, we will evaluate the following expression:

$$\int_{-\infty}^{\infty}(\Phi(x+\varepsilon)-\Phi(x-\varepsilon))\cdot\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x.\quad (9)$$

The expression (9) could be computed using numerical methods.

The number of records in the ε-join result set, denoted by $N_{m,n}^{\varepsilon}$, can be written as:

$$N_{m,n}^{\varepsilon}=\sum_{i=1,j=1}^{m,n}\Psi_{ij},\quad (10)$$

where Ψ_{ij} is the indicator function $\Psi_{ij}=1_{((X_i,Y_j)\in A_\varepsilon)}$.

One can notice that Ψ_{ij} follows a discrete distribution:

$$\Psi_{ij}\sim\begin{pmatrix}0 & 1\\ q(A_\varepsilon) & p(A_\varepsilon)\end{pmatrix},\quad (11)$$

where $p(A_\varepsilon)=(F\otimes G)(A_\varepsilon)$ and $q(A_\varepsilon)=1-p(A_\varepsilon)$.

Reference [7] shows that Ψ_{ij} is a Bernoulli random variable, whose mean is:

$$\mathrm{E}\left(\Psi_{ij}\right)=p(A_\varepsilon).\quad (12)$$

Consequently, the mean of the random variable $N_{m,n}^{\varepsilon}$ is given by the following relation:

$$\mathrm{E}\left(N_{m,n}^{\varepsilon}\right)=m\cdot n\cdot p(A_\varepsilon).\quad (13)$$

In order to compute the variance, we use the property according to which the variance of a sum of random variables equals the sum of their covariance:

$$Var(\sum_{i=1}^{n}X_i)=\sum_{i=1}^{n}\sum_{j=1}^{m}Cov(X_i,X_j).\quad (14)$$

From this property we get:

$$\sigma^2(N_{m,n}^{\varepsilon})=\sigma^2\left(\sum_{i,j}^{m,n}\Psi_{ij}\right)=\sum_{\substack{i,j\\ i',j'}}^{m,n}Cov(\Psi_{ij},\Psi_{i'j'}).\quad (15)$$

Using that $Cov(X,X)=\sigma^2(X)$ the relation (15) becomes:

$$\sigma^2(N_{m,n}^{\varepsilon})=\sum_{(i,j)=(i',j')}^{m,n}\sigma^2(\Psi_{ij})+\sum_{(i,j)\neq(i',j')}^{m,n}Cov(\Psi_{ij},\Psi_{i'j'}).\quad (16)$$

Because Ψ_{ij} is a Bernoulli random variable:

$$\sigma^2(\Psi_{ij})=p(A_\varepsilon)\cdot(1-p(A_\varepsilon))=p(A_\varepsilon)\cdot q(A_\varepsilon).\quad (17)$$

Using (17) in the formula (16) we get:

$$\sigma^2(N_{m,n}^{\varepsilon})=m\cdot n\cdot p(A_\varepsilon)\cdot q(A_\varepsilon)+\sum_{(i,j)\neq(i',j')}^{m,n}c_{iji'j'}.\quad (18)$$

The sum is a positive number, so we have:

$$\sigma^2(N_{m,n}^{\varepsilon})>m\cdot n\cdot p(A_\varepsilon)\cdot q(A_\varepsilon).\quad (19)$$

From the last inequality, we obtain:

$$\sigma^2(N_{m,n}^{\varepsilon})>\sigma^2(K_{m,n}^{\varepsilon}).\quad (20)$$

where $K_{m,n}^{\varepsilon}=\sum_{k=1}^{m\cdot n}1_{((X_k,Y_k)\in A_\varepsilon)}$ the number of points in the region is A_ε provided the $m\cdot n$ random points follow a bi-dimensional probability distribution. We will have $(z_k)_{1\le k\le m\cdot n}=((x_k,y_k))_{1\le k\le m\cdot n}$ two dimensional random variables, independent and identically distributed. Then, the mean of this random variable is:

$$\mathrm{E}K_{m,n}^{\varepsilon} = \mathrm{E}N_{m,n}^{\varepsilon}. \quad (21)$$

Since $Cov(z_i, z_j) = 0$, where $i \neq j$, it follows that z_i, z_j are uncorrelated. This implies that the variance of this variable is:

$$\sigma(K_{m,n}^{\varepsilon}) = m \cdot n \cdot p(A_{\varepsilon}) \cdot q(A_{\varepsilon}). \quad (22)$$

The last relation allows us to state that it exists an order relation $K_{m,n}^{\varepsilon} \prec N_{m,n}^{\varepsilon}$.

Following the ideas presented above, a question could raise: what are the conditions under which one can state that $p_F(\varepsilon) > p_G(\varepsilon)$, regarding the probability distributions *F* and *G*?

We denote the following inequality $(F \otimes F)(A_{\varepsilon}) \leq (G \otimes G)(A_{\varepsilon})$ by $F \lhd G$, where $\varepsilon > 0$. This inequality is equivalent to:

$$\int (F(x+\varepsilon) - F(x-\varepsilon))\mathrm{d}F(x) \leq \int (G(x+\varepsilon) - G(x-\varepsilon))\mathrm{d}G(x), \forall \varepsilon > 0. \quad (23)$$

We try to prove the following implication:

$$F \prec_{st} G \Rightarrow G \lhd F \ (\prec_{st} - \text{stochastic dominance}). \quad (24)$$

First, we will prove this implication in particular cases of probability distributions. We intend to prove this implication for the uniform, exponential and normal probability distributions cases.

Let *F = U(0, a)*, *G = Exp(λ)*. From the relations (6) and (7) we get the equivalence:

$$G \lhd F \Leftrightarrow 2 \cdot \left(\frac{\varepsilon}{a}\right) - \left(\frac{\varepsilon}{a}\right)^2 \geq 1 - e^{-\lambda\varepsilon}. \quad (25)$$

From (24) and (25) it means that we want to prove that:

$$2 \cdot \left(\frac{\varepsilon}{a}\right) - \left(\frac{\varepsilon}{a}\right)^2 \geq 1 - e^{-\lambda\varepsilon}. \quad (26)$$

Suppose that $0 < \varepsilon < a$ and denote $x = \frac{\varepsilon}{a}$. In the particular case of the uniform and normal distributions [14], we have:

$$F \prec_{st} G \Rightarrow F(x) \geq G(x) \Rightarrow \frac{x}{a} \geq 1 - e^{-\lambda x} \Rightarrow 1 - \frac{x}{a} \leq e^{-\lambda x}. \quad (27)$$

The inequality holds if the following condition is true:

$$a\lambda < 1. \quad (28)$$

From the relation (28), it results that:

$$1 - e^{-\lambda a x} < 1 - e^{-x}. \quad (29)$$

For $x < 1$ the inequality $1 - e^{-x} < 2x - x^2$ holds, so we get $1 - e^{-\lambda a x} < 2x - x^2$.

We obtain the relation we were looking for:

$$2 \cdot \left(\frac{\varepsilon}{a}\right) - \left(\frac{\varepsilon}{a}\right)^2 \geq 1 - e^{-\lambda\varepsilon}. \quad (30)$$

The stochastic dominance between the uniform and exponential distributions leads to an order relation between the two distributions, regarding the cardinality of the ε-join result set.

3. Experimental Results

In reference [6] was shown that the estimation of the intermediate cardinality of the approximate join's result sets could be a support to the database query optimization.

In our previous research [13] we studied the probability distribution of the ε-join result set cardinality. In this paper, taking into account the probability distributions of the columns, we determined the right order of the intermediate ε-join operations. This improves the query optimization in the random databases context.

In practical tests were considered queries which extract data from at least three tables, with columns that were spread uniform, normally and exponentially.

In order to generate the tables in our experiments we used the following functions written in *PL/SQL* on *Oracle o11g*:

DBMS_RANDOM.VALUE

DBMS_RANDOM.NORMAL

```
CREATE OR REPLACE FUNCTION
EXPONENTIALA (L REAL)
RETURN REAL IS
X REAL;
BEGIN
X:=(-1/L)*
 LN(DBMS_RANDOM.VALUE(0,1));
RETURN X;
END;
```

We created three tables T_1, T_2 and T_3 using the following code:

```
CREATE TABLE TI (UI NUMBER,
NI NUMBER,
EI NUMBER);  /* I ∈ {1,2,3} */
```

Each table contain three columns: U_I values have a uniform probability distribution, N_I values have a normal probability distribution and E_I values have an exponential probability distribution.

We populate the previous tables with at least 10,000 lines using the function:

```
CREATE OR REPLACE FUNCTION
P_TABLE (N INTEGER) RETURN
INTEGER IS
LAM NUMBER: = 1;
M NUMBER: = 200;
P REAL: = 0.5;
BEGIN
 FOR I IN 1 .. N LOOP
    INSERT INTO TI
    VALUES (
    DBMS_RANDOM.VALUE,
    DBMS_RANDOM.NORMAL,
    EXPONENTIALA(LAM));
 END LOOP;
RETURN 1;
END;  /* I ∈ {1,2,3} */
```

In the Table 1 are displayed some values from the table T_1:

Table 1. First ten values from the table T_1

U_1	N_1	E_1
0.6279	-2.0536	0.3298
0.7931	-1.1162	0.8181
0.0036	0.4992	0.0681
0.6177	-0.6501	1.0179
0.3855	-1.5024	0.9870
0.0057	1.5128	0.0739
0.8816	0.4102	0.7697
0.2943	0.8052	0.2419
0.4077	-0.4987	0.6351
0.1045	0.5776	0.7663

It is known that the join operation is commutative and associative.

Even if the information is extracted from two tables, with 10000 rows each, the join operation is performed on columns which have the same distribution, some differences are noticed:

Table 2. Expected rows number of $\varepsilon - join$ between 2 columns with different probability distributions

$\varepsilon - join, \varepsilon < 0.01$	Repartition	N_ε
$T_1 \bowtie_N T_2$	*normal*	**566597**
$T_1 \bowtie_E T_2$	*exponential*	**980357**
$T_1 \bowtie_U T_2$	*uniform*	**1990504**

We denote by $T_1 \bowtie_N T_2$ the $\varepsilon - join$ operation between T_1 and T_2 using the column in which values follows a normal distribution. Analogous for the similar notations. N_ε represents the number of rows in the query evaluation. The numerical results presented above show that when the number of tables, from which the information is extracted, is bigger than 2 ($T_1 \bowtie T_2 \bowtie T_3 \bowtie ...$), the time required to evaluate such requests vary widely, depending on the probability distribution of the columns used in the $\varepsilon - join$ operations:

```
SELECT COUNT(*) N_ε
FROM TI, TJ
WHERE ABS(NI,NJ)<ε ; // I,J ∈ {1,2,3}
SELECT COUNT(*) N_ε
FROM TI, TJ
WHERE ABS(EI,EJ)<ε ; // I,J ∈ {1,2,3}
SELECT COUNT(*) N_ε
FROM TI, TJ
WHERE ABS(UI,UJ)<ε ; // I,J ∈ {1,2,3}
```

Table 3. Running time for $\varepsilon - join$ operation between three tables with $5 \cdot 10^4$ rows each

$\varepsilon = 0.01, \lvert T_i \rvert_{i=1,3} = 5 \cdot 10^4$	Time (seconds)
$T_1 \bowtie_U T_2 \bowtie_E T_3$	***2156s***
$T_1 \bowtie_N T_2 \bowtie_U T_3$	***1732s***
$T_1 \bowtie_N T_2 \bowtie_E T_3$	***1226s***

The results of the experiments presented in Figure 2 show significant differences between

running time evaluations of join operations in the same tables.

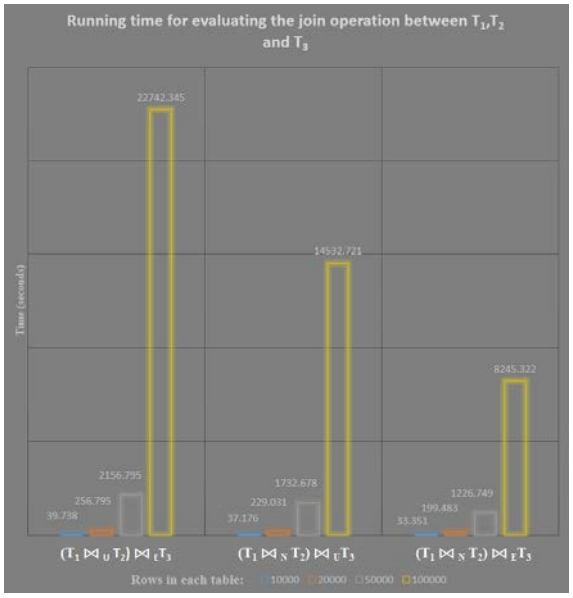

Figure 2. ε-join query optimization

In Figure 2 with blue we represented the running time for evaluating the *ε-join* operation between tables T_1, T_2 and T_3 each with 10000 tuples, with orange for 20000 tuples, with gray for 50000 tuples and with yellow for 100000 tuples. In first histogram we depicted the running time for evaluating the $T_1 \bowtie T_2 \bowtie T_3$ using in the first join the columns which follows uniform probability distribution and in the next one the columns which follows exponential distribution, in the second histogram with normal and uniform distributions, and at last with normal and exponential distributions.

In conclusion if we evaluate multiple *ε-join* between tables having columns with different probability distribution, it is recommended to choose first the columns with normal distribution (see. Table 2 – lowest number of rows in the result), afterwards the columns with exponential distribution and ultimately the columns with uniform distribution.

4. Future Work

In [8] it was proposed a new algorithm *EGO-Efficient Global Optimization* (called *Super-EGO*) for implementing the ε-join operations. The new algorithm prevails over the others through a new technique of subsets ordering which take part in join operations and through a parallel implementation that can run on devices with multiple processors.

The basic EGO-join algorithm analyzes dimensions in a sequential order from *1* to *n*. However, for higher dimensional cases, some of the dimensions might have more discriminative power than the others. The *Super-EGO* algorithm use data sampling techniques to measure this discriminative power to make a new order.

We believe that by studying the distributions of values that these dimensions could have and by using the result presented in this paper it is possible to improve the technique of reordering the dimensions enhancing the performance of this algorithm.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0908.

REFERENCES

1. ABITEBOUL, S., R. HULL, V. VIANU, **Foundations of Databases**, Addison-Wesley, 1995.
2. ANTOVA, L., T. JANSEN, C. KOCH, D. OLTEANU, **Fast and Simple Relational Processing of Uncertain Data**, ICDE, 2008, pp. 983-992.
3. CODD, E. F., **Derivability, Redundancy, and Consistency of Relations Stored in Large Data Banks**, IBM Research Report, 1969.
4. CODD, E. F., **A Relational Model of Data for Large Shared Data Banks**, Communications of the ACM, vol. 13(6), 1970, pp. 377-387.
5. DAGUM, P., R. M. KARP, M. LUBY, S. M. ROSS, **An Optimal Algorithm for Monte Carlo Estimation**, SIAM Journal on Computing, vol. 29(5), 2000, pp. 1484-1496.
6. DALVI, N., D. SUCIU, **Efficient Query Evaluation on Probabilistic Databases**, The VLDB Journal, vol. 16(4), 2007, pp. 523-544.
7. JOHNSON, N., A. KEMP, S. KOTZ, **Univariate Discrete Distributions**, 3rd edition, John Wiley & Sons, 2005.

8. KALASHNIKOV, D., **Super-EGO: Fast Multi-dimensional Similarity Join**, The VLDB Journal, vol. 22(4), 2013, pp. 561-585.

9. OSORIO, R., S. GARCIA, M. PENA, I. LOPEZ-JUAREZ, G. LEFRANC, **Movement and Color Detection of a Dynamic Object: An application to a Mobile Robot**, Studies in Informatics and Control, ISSN 1220-1766, vol. 21(1), 2012, pp. 33-40.

10. LUNGU, R., C. ROTARU, M. LUNGU, **New Systems for Identification, Estimation and Adaptive Control of the Aircrafts Movement**, Studies in Informatics and Control, ISSN 1220-1766, vol. 20(3), 2011, pp. 273-284.

11. SELEZNJEV, O., B. THALHEIM, **Random Databases with Approximate Record Matching**, Methodology and Computing in Applied Probability, Springer Verlag, vol. 12(1), 2008, pp. 63-89.

12. TALAGRAND, M., **The Glivenko-Cantelli Problem**, Annals of Probability, vol. 15(3), 1987, pp. 837-870.

13. VASILE, L., L. VELCESCU, **Inequalities between the Relational Result Sets Cardinalities in Random Databases**, Proc. of SACI, IEEE, 2012, pp. 209-212.

14. WHITMORE, G. A., M. CHAPMAN FINDLAY, **Stochastic Dominance: An Approach to Decision-making under Risk**, Lexington Books, 1978.