Proceedings of the 14th East-European Conference on Advances in Databases and Information Systems (ADBIS) 2010

Estimation of the Maximum Domination Value in Multi-Dimensional Data Sets

Eleftherios Tiakas[1], Apostolos N. Papadopoulos[1], and Yiannis Manolopoulos[1]

Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece
{tiakas,papadopo,manolopo}@csd.auth.gr

Abstract. The last years there is an increasing interest for query processing techniques that take into consideration the dominance relationship between objects to select the most promising ones, based on user preferences. Skyline and top-k dominating queries are examples of such techniques. A skyline query computes the objects that are not dominated, whereas a top-k dominating query returns the k objects with the highest domination score. To enable query optimization, it is important to estimate the expected number of skyline objects as well as the maximum domination value of an object. In this paper, we provide an estimation for the maximum domination value for data sets with statistical independence between their attributes. We provide three different methodologies for estimating and calculating the maximum domination value, and we test their performance and accuracy. Among the proposed estimation methods, our method *Estimation with Roots* outperforms all others and returns the most accurate results.

1 Introduction

Top-k and skyline queries are two alternatives to pose preferences in query processing. In a top-k query a ranking function is required to associate a score to each object. The answer to the query is the set of k objects with the best score. A skyline query does not require a ranking function, and the result is based on preferences (minimization or maximization) posed in each attribute. The result is composed of all objects that are not dominated. For the rest of the work we deal with multidimensional points, where each dimension corresponds to an attribute. Formally, a multidimensional point $p_i = (x_{i_1}, x_{i_2}, ..., x_{i_d}) \in D$ *dominates* another point $p_j = (x_{j_1}, x_{j_2}, ..., x_{j_d}) \in D$ ($p_i \prec p_j$) when:

$$\forall a \in \{1, ..., d\} : x_{i_a} \leq x_{j_a} \wedge \exists b \in \{1, ..., d\} : x_{i_b} < x_{j_b}$$

where d is the number of dimensions. A top-k dominating query may be seen as a combination of a top-k and a skyline query. More specifically, a top-k dominating query returns the k objects with the highest domination scores. The domination value of an object p, denoted as $dom(p)$, equals the number of objects that p dominates [12, 13].

	A	B	C	D	E	F	G	H	I	J
Distance to beach (m)	1000	400	600	1200	300	50	100	250	500	450
Price (euros)	60	80	90	15	65	100	50	20	40	30

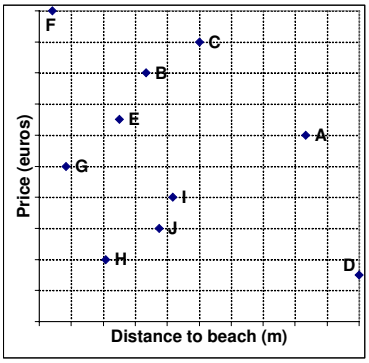

Fig. 1. The hotel data set.

The *maximum domination value* is the number of objects dominated by the top-1 (best) object. More formally, let us assign to each item t of the data set D a score, $m(t)$, which equals the number of items that t dominates:

$$m(t) = |\{q \in D : q \prec t\}|$$

Then, if p is the object with the maximum domination value we have:

$$p = \arg\max_{t}\{m(t), t \in D\}$$

An example is illustrated in Figure 1. A tourist wants to select the best hotel according to the attributes *distance to the beach* and *price per night*. The domination values of all hotels $A, B, C, D, E, F, G, H, I, J$ are 0, 1, 0, 0, 2, 0, 4, 6, 2, 3 respectively, thus the hotel with the max domination value is H. This hotel is the best possible selection, whereas the next two best choices are hotels G and J.

In this work, we focus on estimating the maximum domination value in a multi-dimensional data under the uniformity and independence assumptions. Estimating the maximum domination value contributes in: (i) optimizing top-k dominating and skyline algorithms, (ii) estimating the cost of top-k dominating and skyline queries, (iii) developing pruning strategies for these queries and algorithms. Moreover, we show that the maximum domination value is closely related to the cardinality of the skyline set.

The rest of the article is organized as follows. Section 2 briefly describes related work in the area. Section 3 studies in detail different estimation methods, whereas Section 4 contains performance evaluation results. Finally, Section 5 concludes the work.

2 Related Work

As we will show in the sequel, the maximum domination value is related to the skyline cardinality which has been studied recently. There are two different approaches for the skyline cardinality estimation problem: (i) the *parametric* methods, and (ii) the *non-parametric* methods. *Parametric* methods use only main parameters of the data set, like its cardinality N and its dimensionality d. Bentley et al. [1] established that the skyline cardinality is $O((\ln N)^{d-1})$. Buchta [3] proved another asymptotic bound of the skyline cardinality, which is: $\Theta\left(\frac{(\ln N)^{d-1}}{(d-1)!}\right)$. Bentley et al. [1] and Godfrey [6, 7], under the assumptions of attribute value independence and that all attributes in a dimension are unique and totally ordered, established that the skyline cardinality can be estimated with harmonics: $\widehat{s}_{d,N} = H_{d-1,N}$. Godfrey [6, 7] established that for sufficient large N, $\widehat{s}_{d,N} = H_{d-1,N} \approx \frac{(\ln N)^{d-1}}{(d-1)!}$. Lu et al. [10] established specific parametric formulae to estimate the skyline cardinality over uniformly and arbitrary distributed data, keeping the independence assumption between dimensions.

Non-parametric methods use a sampling process in the data set to capture its characteristics and estimate the skyline cardinality. Chaudhuri et al. [4] relax the assumptions of statistical independence and attribute value uniqueness, and they use uniform random sampling in order to address correlations in the data. They assume that the skyline cardinality follows the rule: $s = A \log^B N$ for some constants A, B (which is an even more generalized formula of $\frac{(\ln N)^{d-1}}{(d-1)!}$), and using *log sampling* they calculate the A, B values. Therefore, this method can be seen as a *hybrid* method (both parametric and non-parametric). Zhang et al. [14] use a kernel-based non-parametric approach that it does not rely on any assumptions about data set properties. Using sampling over the data set they derive the appropriate kernels to efficiently estimate the skyline cardinality in any kind of data distribution.

Both directions sometimes produce significant estimation errors. Moreover, in non-parametric methods there is a tradeoff between the estimation accuracy and the sampling preprocessing cost over the data. In this paper, we focus on estimating the maximum domination value using only parametric methods. To the best of our knowledge, this is the first work studying the estimation of the maximum domination value and its relationship with the skyline cardinality.

3 Estimation Methods

In this section we present specific methods to estimate the maximum domination value of a data set. We first explain how this maximum domination value is strongly connected with the skyline cardinality of the data set. Next, we present two estimation methods inspired from [6, 7, 10], and finally we propose a novel method that is much simpler, more efficient and more accurate than its opponents.

For each presented estimation method, the main task is to produce a formula that includes only the main data set parameters, which are: the number of items

of the data set (cardinality N), and the number of the existing attributes (dimensionality d). In this respect, several properties and results are derived for the maximum domination value and the item having this value. For the remaining part of this study we adopt the following assumptions:

- All attribute values in a single dimension are distinct (domain assumption).
- The dimensions are statistically independent, i.e., there are no pair-wise or group correlations nor anti-correlations (independence assumption).

Let p_i, $i \in \{1, ..., N\}$ be the N items of the data set, and (x_{i_1}, x_{i_2},..., x_{i_d}) their corresponding attributes in the d selected dimensions. Under our assumptions, no two items share a value over any dimension, thus the items can be totally ordered on any dimension. Therefore, it is not necessary to consider the actual attribute values of the items, but we can conceptually replace these values by their rank position along any dimension. Thus, let (r_{i_1}, r_{i_2},..., r_{i_d}) be the corresponding final distinct rank positions of item p_i in the selected dimensions (where $r_{i_j} \in \{1, ..., N\}$). Without loss of generality, we assume that over the attribute values in a dimension minimum is best. Then, the item with rank position 1 will have the smallest value on that dimension, whereas the item with rank position N will have the largest one.

3.1 Maximum Domination Value and Skyline Cardinality

Here we study how the maximum domination value is related to the skyline cardinality of the data set. Figure 2 reveals this relationship. Let p be the item of the data set with the maximum domination value. A first important property is that p is definitely a skyline point. This was first proved in [2] for any monotone ranking function over the data set, and also shown in [12, 13] for the top-1 item

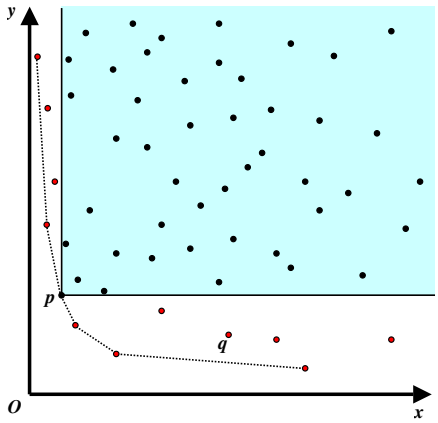

Fig. 2. Maximum domination value and skyline items.

in top-k dominating queries. Moreover, p dominates most of the items lying in the marked area. This area is called the *domination area* of p. No other point dominates more items than p does. Let dom be the exact domination value of p, which is the number of all items that lie in its domination area (i.e., the maximum domination value of the data set). On the other hand, the skyline items are the items that lie in the dotted line. Let s be the number of the skyline items (i.e., the skyline cardinality).

As p does not dominate any item contained in the skyline, its domination value satisfies the relation:

$$dom \ \leq \ N - s$$

Therefore, a simple *overestimation* of the maximum domination value is $\widehat{dom} = N - s$, and can be computed when the skyline cardinality s is already known (or it has been efficiently estimated $\widehat{dom} = N - \widehat{s}$). The error rate of this estimation depends only on the items that lie neither in the skyline nor in the domination area of p, like item q for example. These items are called *outliers*. Moreover, as the data set cardinality N increases, the number of outliers becomes significantly smaller than N, and the estimation becomes more accurate. On the contrary, as the data set dimensionality d increases, the number of outliers also increases, and the estimation becomes less accurate.

3.2 Estimation with Harmonics

Here, we present an estimation approach using harmonic numbers and their properties, inspired from [6, 7]. The analysis reveals the intrinsic similarities between the maximum domination value and the skyline cardinality, and shows that $\widehat{dom} = N - \widehat{s}$. Let $dom_{d,N}$ be the random variable which measures the number of items dominated by the top-1 item (the maximum domination value). We denote as $\widehat{dom}_{d,N}$ the expected value of $dom_{d,N}$.

Theorem 1. *In any data set under the domain and independence assumptions, the expected value $\widehat{dom}_{d,N}$ satisfies the following recurrence:*

$$\widehat{dom}_{d,N} = \frac{1}{N}\widehat{dom}_{d-1,N} + \widehat{dom}_{d,N-1}$$

for $d > 1, \ N > 0$, where $\widehat{dom}_{1,N} = N - 1$ and $\widehat{dom}_{d,1} = 0$.

Proof. If $d = 1$, then we have only one dimension and the item with rank position 1 is the top-1 item that dominates all other $N-1$ items. Thus it holds that $\widehat{dom}_{1,N} = N - 1$. If $N = 1$, then we have only one item and none item to dominate. Thus, $\widehat{dom}_{d,1} = 0$. In case that $d > 1$ and $N > 1$, there is an item with rank position 1 on dimension 1. This item has the maximum domination value as it dominates all other items on that dimension. The probability that this item will remain a top-1 item is the probability that no other item has a greater domination value in any other dimension $(2, ..., d)$, given the independence assumption. However, $\widehat{dom}_{d-1,N}$ is the maximum domination value out of these $d-1$ dimensions. Thus,

as any item has equal probability to be placed in rank position 1 on dimension 1, we have $\frac{1}{N}\widehat{dom}_{d-1,N}$ to be the probability that this item has the maximum domination value. Since, the first ranked item on dimension 1 cannot be dominated by any other item, the maximum domination value is determined by the remaining $N-1$ items which is estimated by $\widehat{dom}_{d,N-1}$. Therefore, we have:

$$\widehat{dom}_{d,N} = \frac{1}{N}\widehat{dom}_{d-1,N} + \widehat{dom}_{d,N-1}\square$$

The recurrence for $\widehat{dom}_{d,N}$ is strongly related to the harmonic numbers:

- The harmonic of a positive integer n is defined as: $H_n = \sum_{i=1}^{n} \frac{1}{i}$.
- The k-th order harmonic [11] of a positive integer n for integers $k > 0$ is defined as: $H_{k,n} = \sum_{i=1}^{n} \frac{H_{k-1,i}}{i}$, where $H_{0,n} = 1, \forall n > 0$ and $H_{k,0} = 0, \forall k > 0$. Note also that: $H_{1,n} = H_n, \forall n > 0$.

In order to retrieve the fundamental relation of $\widehat{dom}_{d,N}$ with the harmonic numbers, we compute $\widehat{dom}_{2,N}$ and using mathematical induction we derive the final formula. For the $\widehat{dom}_{2,N}$ value we have:

$$\widehat{dom}_{2,N} = \frac{1}{N}\widehat{dom}_{1,N} + \widehat{dom}_{2,N-1} = \frac{N-1}{N} + \widehat{dom}_{2,N-1} =$$

$$= \frac{N-1}{N} + \frac{1}{N-1}\widehat{dom}_{1,N-1} + \widehat{dom}_{2,N-2} = \frac{N-1}{N} + \frac{N-2}{N-1} + ... + \frac{1}{2} =$$

$$= 1 - \frac{1}{N} + 1 - \frac{1}{N-1} + ... + 1 - \frac{1}{2} = N - 1 - \sum_{i=2}^{N} \frac{1}{i} =$$

$$= N - 1 - \left(\sum_{i=1}^{N} \frac{1}{i} - 1\right) = N - H_N$$

Now, let us assume that the following equation holds for a specific k, (i.e., $\widehat{dom}_{k,N} = N - H_{k-1,N}$). We will prove that the previous equation holds also for the next natural number $k+1$. We have:

$$\widehat{dom}_{k+1,N} = \frac{1}{N}\widehat{dom}_{k,N} + \widehat{dom}_{k+1,N-1} =$$

$$= \frac{1}{N}\widehat{dom}_{k,N} + \frac{1}{N-1}\widehat{dom}_{k,N-1} + \frac{1}{N-2}\widehat{dom}_{k,N-2} + ... =$$

$$= \sum_{i=1}^{N} \frac{1}{i}\widehat{dom}_{k,i} = \sum_{i=1}^{N} \frac{1}{i}\left(i - H_{k-1,i}\right) =$$

$$= \sum_{i=1}^{N} \left(1 - \frac{H_{k-1,i}}{i}\right) = N - \sum_{i=1}^{N} \frac{H_{k-1,i}}{i} = N - H_{k,N}$$

Therefore, for any $d > 1,\ N > 0$ it holds that:

$$\widehat{dom}_{d,N} = N - H_{d-1,N} \tag{1}$$

Equation 1 generates some important properties for the maximum domination value:

- $\widehat{dom}_{d,N}$ is strongly related to the skyline cardinality of the data set. As shown in [6, 7], if $\widehat{s}_{d,N}$ is the expected value of the skyline cardinality, then it holds that:

  $$\widehat{s}_{d,N} = H_{d-1,N}$$

  Therefore, we have:

  $$\widehat{dom}_{d,N} = N - \widehat{s}_{d,N} \tag{2}$$

  In particular, $\widehat{dom}_{d,N}$ and $\widehat{s}_{d,N}$ share the same recurrence equation of Theorem 1 but with different initial conditions.
- as proved in [11], it holds that $\lim_{d\longrightarrow\infty} H_{d,N} = N$. Therefore, we have:

  $$\lim_{d\longrightarrow\infty} \widehat{dom}_{d,N} = N - \lim_{d\longrightarrow\infty} H_{d-1,N} \Leftrightarrow \lim_{d\longrightarrow\infty} \widehat{dom}_{d,N} = 0 \tag{3}$$

  Equation 3 is a validation of the fact that as the dimensionality d increases, the maximum domination value (and consequently all the following domination values) decreases until reaching zero. In particular, the dimensionality d beyond which all domination values become equal to zero, is a small number. We call that dimension the *eliminating dimension*, and denote it as d_0.

In the sequel, we focus in the computation of $\widehat{dom}_{d,N}$. Since it holds that $\widehat{dom}_{d,N} = N - H_{d-1,N}$, the main task is the efficient computation of the harmonic term $H_{d-1,N}$. There are three different methods to follow for this task:

Recursive Calculation: The calculation of $H_{d-1,N}$ can be achieved by running a recursive algorithm that follows the direct definition formula:

$$H_{k,n} = \sum_{i=1}^{n} \frac{H_{k-1,i}}{i}$$

where $H_{0,n} = 1$ and $k > 0$. We can also use a look up table at run-time, however, these recurrence computations are expensive. The algorithmic time complexity is exponential: $O(N^{d-1})$. As shown later in the experimental results section, the calculation time is not acceptable even for small dimensionality values .

Bound Approximation: This method was proposed in [6, 7] and is based on asymptotic bounds of $H_{k,N}$. Bentley et al. [1] established that: $\widehat{s}_{d,N}$ is $O((\ln N)^{d-1})$. Bentley et al. [1] and Godfrey [6, 7] established that: $\widehat{s}_{d,N} = H_{d-1,N}$, thus:

$$H_{d-1,N} \quad is \quad O((\ln N)^{d-1})$$

Buchta [3] and Godfrey [6,7] improved this asymptotic bound as follows:

$$H_{d-1,N} \approx \Theta\left(\frac{(\ln N)^{d-1}}{(d-1)!}\right)$$

Therefore, we can instantly estimate $\widehat{dom}_{d,N}$ using the following formula:

$$\widehat{dom}_{d,N} \approx N - \Theta\left(\frac{(\ln N)^{d-1}}{(d-1)!}\right)$$

or equivalently (for an appropriate real number λ):

$$\widehat{dom}_{d,N} \approx N - \lambda\left(\frac{(\ln N)^{d-1}}{(d-1)!}\right) \tag{4}$$

This is not a concrete estimation and generates a significant error rate. Moreover, by varying the dimensionality range it will be shown that this estimation is not even a monotone function and changes its monotonicity after halving the eliminating dimension (i.e., for any $d > \frac{d_0}{2}$). Therefore, it provides wrong theoretical results.

Generating Functions Approximation: This method was also proposed in [6,7] and is based on Knuth's generalization via generating functions [8,9], which established that:

$$H_{k,N} = \sum_{c_1,c_2,...,c_k} \prod_{i=1}^{k} \frac{\mathcal{H}_{i,N}^{c_i}}{i^{c_i} \cdot c_i!}, \qquad c_1, c_2, ..., c_k \geq 0 \ \ \wedge \ \ c_1 + 2c_2 + ... + kc_k = k \tag{5}$$

where $\mathcal{H}_{i,N}$ is the i-th hyper-harmonic of N and is defined as:

$$\mathcal{H}_{i,N} = \sum_{j=1}^{N} \frac{1}{j^i} \qquad (\mathcal{H}_{1,N} = H_{1,N} = H_N)$$

Note that $c_1, c_2, ..., c_k$ are positive (or zero) integer numbers, whereas the number of terms of the sum in Equation 5 stems from all possible combinations of $c_1, c_2, ..., c_k$ that satisfy the equation $c_1 + 2c_2 + ... + kc_k = k$. This number is $\wp(k)$ and expresses the number of all possible ways to partition k as a sum of positive integers. Therefore, $H_{k,N}$ can be expressed as a polynomial of $\wp(k)$ terms which contain the first k hyper-harmonics $\mathcal{H}_{i,N}$, $(i = 1, ..., k)$. For example:

$$H_{2,N} = \frac{1}{2}\mathcal{H}_{1,N}^2 + \frac{1}{2}\mathcal{H}_{2,N}$$

$$H_{3,N} = \frac{1}{6}\mathcal{H}_{1,N}^3 + \frac{1}{2}\mathcal{H}_{1,N}\mathcal{H}_{2,N} + \frac{1}{3}\mathcal{H}_{3,N}$$

This approximation of $H_{k,N}$ is remarkably accurate. In particular, with this method we reach almost exactly the theoretical values of $H_{k,N}$ when computed

with the recursive approach. This will be also evaluated in the experimental results section. For any given dimension d, the time cost to compute the d required hyper-harmonics is $O(dN)$. Then, having the previous formulae, we can immediate calculate $H_{d,N}$. The only requirement is to generate the appropriate formula for the dimension d with the $\wp(d)$ terms. Godfrey [6, 7] mentioned that this number of terms ($\wp(d)$) grows quickly, and, thus, it is not viable to compute the required formula this way, and suggests not using this approximation for large values of d. However, motivated by the accuracy of this approximation of $H_{k,N}$, we developed a dynamic-programming algorithm that efficiently produces these equations. Due to lack of space we do not elaborate further.

Therefore, we can almost instantly estimate $\widehat{dom}_{d,N}$ with hyper-harmonics using the previous approximation formula:

$$\widehat{dom}_{d,N} \approx N - \sum_{c_1,c_2,\ldots,c_{d-1}} \prod_{i=1}^{d-1} \frac{\mathcal{H}_{i,N}^{c_i}}{i^{c_i} \cdot c_i!} \tag{6}$$

by taking special care to all possible floating point overflow values, and by using the derived equations which recorded through the automation.

3.3 Estimation with Multiple Summations

In this section we present an estimation approach using a specific formula with multiple summations inspired from the study of [10]. For compatibility reasons we will keep all previous notations and variables.

Y. Lu et al. [10] introduced an estimation approach of the skyline cardinality that relaxes the domain assumption of our basic model. The statistical independence assumption still remains, but now the data can have duplicate attribute values. Their study is based in probabilistic methods, and it uses the value cardinality of each dimension. Their first main result is the following:

$$\widehat{s}_{d,N} = N \cdot \sum_{t_1=1}^{c_1} \sum_{t_2=1}^{c_2} \ldots \sum_{t_d=1}^{c_d} \left(\prod_{i=1}^{d} \frac{1}{c_i} \right) \left(1 - \prod_{j=1}^{d} \frac{t_j}{c_j} \right)^{N-1} \tag{7}$$

where $N \geq 1,\ d \geq 1$, and c_j is the value cardinality of the j-th dimension.

They also generalized this result in case of having the probability functions $f_j(x)$ of the data over each dimension, but always keeping the independence assumption:

$$\widehat{s}_{d,N} = N \cdot \sum_{t_1=1}^{c_1} \sum_{t_2=1}^{c_2} \ldots \sum_{t_d=1}^{c_d} f_1(t_1) f_2(t_2) \ldots f_d(t_d) \left(1 - \prod_{j=1}^{d} \sum_{x=1}^{t_j} f_j(x) \right)^{N-1} \tag{8}$$

However, in both cases, the computational complexity is $O(c_1 \cdot c_2 \cdot \ldots \cdot c_d)$, which is not acceptable even if in few dimensions the value cardinality is high (close to N). They also tried to relax this complexity cost by introducing high and low

cardinality criteria, but this cost remains high, and this is why their experimental results are restricted to small dimensionality and cardinality variations ($d = 1, 2, 3$ and $N \leq 1000$). We will see in our experimental results that even if we have high cardinality in 3 dimensions and up to 1000 items the estimation time is not acceptable.

Although the method of [10] works efficiently only in small cardinalities and dimensionalities, it would be very interesting to apply this method in our model and study its accuracy. Therefore, under the domain assumption of our model, all value cardinalities c_j will be equal to N and Equation 7 gives:

$$\widehat{s}_{d,N} = N \cdot \sum_{t_1=1}^{N} \sum_{t_2=1}^{N} ... \sum_{t_d=1}^{N} \left(\frac{1}{N^d}\right) \left(1 - \frac{t_1 t_2 ... t_d}{N^d}\right)^{N-1}$$

or equivalently:

$$\widehat{s}_{d,N} = \frac{1}{N^{d-1}} \cdot \sum_{t_1=1}^{N} \sum_{t_2=1}^{N} ... \sum_{t_d=1}^{N} \left(1 - \frac{t_1 t_2 ... t_d}{N^d}\right)^{N-1}$$

Thus, using the property of the estimated maximum domination value of Equation 2, the final estimation formula is:

$$\widehat{dom}_{d,N} \approx N - \frac{1}{N^{d-1}} \cdot \sum_{t_1=1}^{N} \sum_{t_2=1}^{N} ... \sum_{t_d=1}^{N} \left(1 - \frac{t_1 t_2 ... t_d}{N^d}\right)^{N-1} \quad (9)$$

which has an exponential computational complexity ($O(N^d)$).

In our experimental results we will see that Equation 9 returns values remarkably close to the harmonic $H_{d-1,N}$ values. In addition, by increasing N, the returned values converge to $H_{d-1,N}$, thus it must be related somehow with the k-th order harmonics. This strong relation remains unproven. Finally, as the two methods return almost the same estimations, their accuracy is similar.

3.4 Estimation with Roots

In this section we present a novel estimation approach using a simple formula, which provides more accurate estimation results.

Let p be the item with the maximum domination value, and $(r_{p_1}, r_{p_2}, ..., r_{p_d})$ be its corresponding final rank positions in the total ordering along any dimension. Let also a be the maximum rank position of p through all dimensions (i.e., $a = \max\{r_{p_1}, r_{p_2}, ..., r_{p_d}\}$). Then, a splits the total ordering of the items in two parts as in Figure 3: (i) the (a)-area, and (ii) the $(N - a)$-area.

Now, any item q that all its rank positions lie in the (a)-area, will be an outlier or a skyline item. Note that the opposite does not hold, thus not any skyline or outlier item lie in the (a)-area. The probability P_a that an item lies in the (a)-area is:

$$P_a = P(p_i \quad lies \quad in \quad the \quad (a) - area) =$$

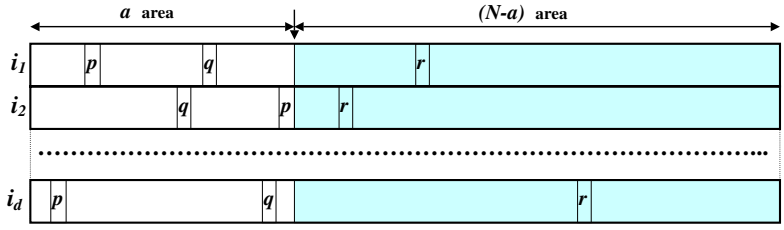

Fig. 3. Total ordering of items by rank positions

$$= P(r_{i_1} \leq a \wedge r_{i_2} \leq a \wedge ... \wedge r_{i_d} \leq a)$$

Now, due to the independence assumption we have:

$$P_a = P(r_{i_1} \leq a) \cdot P(r_{i_2} \leq a) \cdot ... \cdot P(r_{i_m} \leq a) = \frac{a}{N} \cdot \frac{a}{N} \cdot ... \cdot \frac{a}{N} = \frac{a^d}{N^d} = \left(\frac{a}{N}\right)^d$$

Moreover, any item r that all its rank positions lie in the $(N-a)$-area, will definitely be dominated by p. Thus r lies in the domination area of p. Note that the opposite does not hold, thus not any item of the domination area of p, lies also in the $(N-a)$-area). The probability P_{N-a} that an item lies in the $(N-a)$-area is:

$$P_{N-a} = P(p_i \quad lies \quad in \quad the \quad (N-a) - area) =$$
$$= P(r_{i_1} > a \wedge r_{i_2} > a \wedge ... \wedge r_{i_d} > a)$$

Due to the independence assumption we have:

$$P_{N-a} = P(r_{i_1} > a) \cdot ... \cdot P(r_{i_d} > a) = \frac{N-a}{N} \cdot ... \cdot \frac{N-a}{N} = \frac{(N-a)^d}{N^d} = \left(1 - \frac{a}{N}\right)^d$$

Therefore, the number of items lying in the (a)-area (C_a), and in the $(N-a)$-area (C_{N-a}) will be:

$$C_a = \lfloor N \cdot P_a \rfloor = \lfloor N \left(\frac{a}{N}\right)^d \rfloor \quad and \quad C_{N-a} = \lfloor N \cdot P_{N-a} \rfloor = \lfloor N \left(1 - \frac{a}{N}\right)^d \rfloor$$

respectively.

However, as it holds that all items lying in the $(N-a)$-area are dominated by p, we have $dom(p) \geq C_{N-a}$, or equivalently:

$$dom(p) \geq \lfloor N \cdot P_{N-a} \rfloor \Leftrightarrow dom(p) \geq \lfloor N \left(1 - \frac{a}{N}\right)^d \rfloor \tag{10}$$

which describes a tight lower bound for the domination value of p.

Additionally, as p definitely lies in the (a)-area, at least one item must be inside that area, thus it must hold that $C_a \geq 1$, or equivalently:

$$\lfloor N \cdot P_a \rfloor \geq 1 \Leftrightarrow \lfloor N \left(\frac{a}{N}\right)^d \rfloor \geq 1 \tag{11}$$

To efficiently estimate the maximum domination value, we have to maximize the lower bound of Inequality 10 under the a variable, respecting the condition of Inequality 11 for the a variable. Therefore, let us define the function $f(a) = N\left(1 - \frac{a}{N}\right)^d$ that expresses the lower bound values, where $a \in [0, N]$. It has $f(0) = N,\ f(N) = 0$ and the following relation derives:

$$f'(a) = -d\left(1 - \frac{a}{N}\right)^{d-1}$$

We have $f'(a) < 0, \forall a \in (0, N)$, thus f is a monotone descending function in [0,N], and returns values also in [0,N].

Moreover, the condition of Inequality 11 gives:

$$\left(\frac{a}{N}\right)^d \geq \frac{1}{N} \Leftrightarrow \frac{a}{N} \geq \sqrt[d]{\frac{1}{N}} \Leftrightarrow a \geq N\sqrt[d]{\frac{1}{N}}$$

Thus, f must be restricted in $[N\sqrt[d]{\frac{1}{N}}, N]$. Due to the descending monotonicity of f, it takes its maximum value when $a_{max} = N\sqrt[d]{\frac{1}{N}}$. Therefore, we have:

$$f(a_{max}) = N\left(1 - \frac{a_{max}}{N}\right)^d = N\left(1 - \sqrt[d]{\frac{1}{N}}\right)^d = \left(\sqrt[d]{N} - 1\right)^d$$

and the final estimation of the maximum domination value is:

$$\widehat{dom}_{d,N} \approx \left(\sqrt[d]{N} - 1\right)^d \tag{12}$$

Generalization of the Root Method To get even more accurate estimations, we can generalize the root method by allowing the a variable to take values slightly smaller than $N\sqrt[d]{\frac{1}{N}}$. This left shift of the a variable breaks the condition of Inequality 11, but allows the possibility of taking into account items that are not lying in the $(a), (N-a)$-areas and are dominated by p increasing its domination value. We further studied this estimation improvement making exhaustive experimental tests with different a values, and we conclude that the estimation is very accurate when:

$$a_{shifted} = N\sqrt[d]{\frac{1}{N\sqrt{N}}}$$

Then f takes the value:

$$f(a_{shifted}) = N\left(1 - \frac{a_{shifted}}{N}\right)^d = N\left(1 - \sqrt[d]{\frac{1}{N\sqrt{N}}}\right)^d = \frac{1}{\sqrt{N}}\left(\sqrt[d]{N\sqrt{N}} - 1\right)^d$$

This *hidden* square root factor enhances the estimation accuracy and provides the most accurate results for the maximum domination value. Thus, the final proposed estimation formula is:

$$\widehat{dom}_{d,N} \approx \frac{1}{\sqrt{N}}\left(\sqrt[d]{N\sqrt{N}} - 1\right)^d \tag{13}$$

4 Performance Evaluation

To test the estimation accuracy, we perform several experiments using independent data sets of N = 100, 1K, 10K, 100K, 1M items and varying the dimensionality d from 1 to values beyond the eliminating dimension d_0. We record the exact (average of 10 same type data sets) and the estimated maximum domination values. For brevity, we present only a small set of representative results, which depict the most significant aspects. All experiments have been conducted on a Pentium 4 with 3GHz Quad Core Extreme CPU, 4GB of RAM, using Windows XP. All methods have been implemented in C++. Table 1 summarizes the methods compared.

Notation	**Interpretation**
RealAvg	Real Averaged Values (No Estimation)
HarmRecc	Estimation with Harmonics (Recursive Calculation)
HarmBound	Estimation with Harmonics (Bound Approximation)
HarmGenF	Estimation with Harmonics (Generating Functions Approximation)
CombSums	Estimation with Multiple Summations
Roots	Estimation with Roots (Simple)
RootsGen	Estimation with Roots (Generalized with the square root)

Table 1. Summary of methods evaluated.

d	**RealAvg**	**HarmRecc**	**HarmBound**	**HarmGenF**	**CombSums**	**Roots**	**RootsGen**
1	999.0	999.000	999.000	999.000	999.418	999.000	999.968
2	955.6	992.515	993.092	992.515	993.431	937.754	988.785
3	875.2	971.162	976.141	971.166	-	729.000	908.100
4	623.1	923.542	945.064	923.556	-	456.931	732.128
5	434.4	-	905.128	842.585	-	235.430	510.298
6	294.2	-	868.930	730.456	-	102.204	308.870
7	190.3	-	849.100	598.636	-	38.198	164.043
8	110.7	-	851.089	463.172	-	12.510	77.312
9	73.2	-	871.420	338.813	-	3.642	32.674
10	33.1	-	901.311	235.082	-	0.954	12.498
11	23.8	-	931.828	155.378	-	0.227	4.362
12	18.5	-	957.189	98.316	-	0.049	1.399
13	10.1	-	975.356	59.879	-	0.010	0.415
14	6.5	-	986.905	32.465	-	0.002	0.114
15	3.6	-	993.539	21.465	-	0.000	0.029
16	2.2	-	997.025	15.645	-	0.000	0.007
17	1.5	-	998.715	9.416	-	0.000	0.002
18	0.9	-	999.478	4.196	-	0.000	0.000
19	0.8	-	999.800	2.123	-	0.000	0.000
20	0.0	-	999.927	0.170	-	0.000	0.000

Table 2. Maximum domination value estimation for $N = 1K$

Figure 4 depicts the maximum domination value estimation results for all estimation methods, varying the cardinality and the dimensionality of the data

sets. Table 2 presents the detailed estimation values of the corresponding graph for N=1K, for further inspection. We have not recorded the values where the

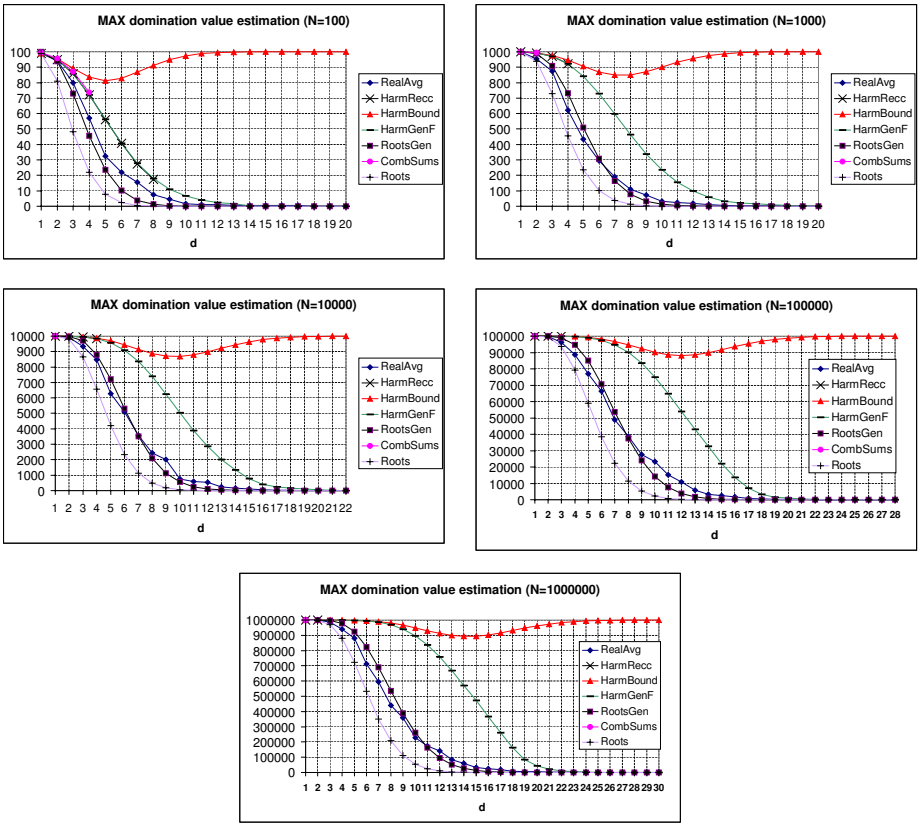

Fig. 4. Maximum domination value estimation for N=100, 1K, 10K, 100K, 1M.

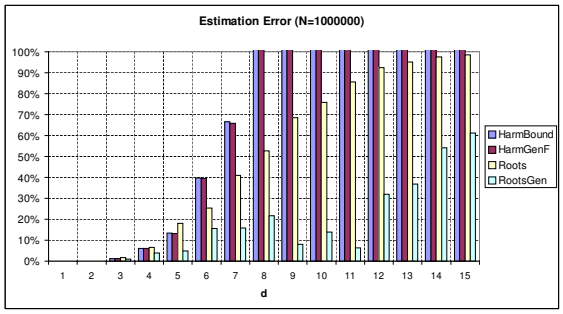

Fig. 5. Estimation error for N=1M.

computational time is more than 10 minutes. Figure 5 presents the estimation error of the 4 methods that return values into the full range for N=1M, for further inspection. Based on the previous results we observe the following:

- The HarmRecc method, due to the exponential computation complexity, returns values in short time only for small dimensionality and cardinality values. Moreover, after 3 dimensions the estimation error is significant.
- The HarmBound method returns estimation results instantly, but it is the most inaccurate method for estimation.
- The HarmGenF method computes its results very efficiently. It produces almost exactly the theoretical values of $H_{k,N}$ when computed recursively. Therefore, it returns the same estimation results with the HarmRecc method. However, we observe that as we increase the cardinality of the data set, the estimation error increases and becomes significant in almost all the dimensionality range.
- The CombSums method fails to return values even in small dimensionality and cardinality selections, due to its exponential computational complexity (we can see only 4 values when N=100 and 2 values when N=1000). In addition, the returned values are remarkably close to those of HarmGenF and HarmRecc. By increasing the cardinality, the values converge to the harmonic values of the previous methods, thus it must be related somehow with the k-th order harmonics. However, again the estimation error increases and becomes significant in the whole dimensionality range.
- The Roots method returns estimation results more efficiently than all the previous methods. By increasing the cardinality, the estimation becomes more accurate, due to the fact that the number of the outliers becomes significantly smaller than N. On the contrary, as the dimensionality increases, the number of outliers increases as well, and the estimation becomes less accurate. However, when we further move into the dimensionality range and when we approach the eliminating dimension, the estimation becomes again accurate.
- The RootsGen method is the most efficient way to get estimation results and outperforms all previous methods. It manages to approximate the maximum domination value with the smallest estimation error in the whole dimensionality and cardinality range.

5 Conclusions

This paper studies parametric methods for estimating the maximum domination value in multi-dimensional data sets, under the assumption of statistical independence between dimensions and the assumption that there are no duplicate attribute values in a dimension. The experimental results confirm that our proposed estimation method outperforms all other methods and achieves the highest estimation accuracy. Future work may include: (i) further study of the eliminating dimension d_0, providing an estimation formula for its calculation, (ii) the study the estimation of the skyline cardinality under the Roots method and its variants

and (iii)the study of the maximum domination value estimation and the eliminating dimension in arbitrary data sets, by relaxing the assumptions of uniformity, independence and distinct values that used in this work.

References

1. J.L. Bentley, H.T. Kung, M. Schkolnick and C.D. Thompson: "On the Average Number of Maxima Set of Vectors and Applications", *Journal of the ACM*, Vol.25, No.4, pp.536-543, 1978.
2. S. Borzsonyi, D. Kossmann and K. Stocker, "The Skyline Operator", *Proceedings 17th International Conference on Data Engineering (ICDE)*, pp.421-430, Heidelberg, Germany, 2001.
3. C. Buchta, "On the average number of maxima in a set of vectors.", *Information Processing Letters 33*, pp.6365, 1989.
4. S. Chaudhuri, N. Dalvi and R. Kaushik, "Robust Cardinality and Cost Estimation for the Skyline Operator", *Proceedings 22nd International Conference on Data Engineering (ICDE)*, Atlanta, GA, 2006.
5. J. Huang: "Tuning the Cardinality of Skyline", *Proceedings 10th Asia-Pacific Web Conference (APWeb) Workshops*, Shenyang, China, 2008.
6. P. Godfrey, "Cardinality Estimation of Skyline Queries: Harmonics in Data", Technical Report CS-2002-03, York University, 2002.
7. P. Godfrey, "Skyline Cardinality for Relational Processing", *Proceedings 3rd International Symposium of Foundations of Information and Knowledge Systems (FoIKS)*, pp.78-97, Wilhelminenburg Castle, Austria, 2004.
8. R.L. Graham, D.E. Knuth and O. Patashnik, "Concrete Mathematics", *Addison-Wesley*, 1989.
9. D.E. Knuth, "Fundamental Algorithms: The Art of Computer Programming.", *Addison-Wesley*, 1973.
10. Y. Lu, J. Zhao, L. Chen, B. Cui and D. Yang, "Effective Skyline Cardinality Estimation on Data Streams", *Proceedings 20th International Conference on Database and Expert Systems Applications (DEXA)*, pp.241-254, Turin, Italy, 2008.
11. S. Roman, "The harmonic logarithms and the binomial formula", *Journal of Combinatorial Theory, Series A*, Vol.63, pp.143-163, 1993.
12. M.L. Yiu and N. Mamoulis, "Efficient Processing of Top-k Dominating Queries on Multi-Dimensional Data", *Proceedings 33rd International Conference on Very Large Data Bases (VLDB)*, pp.483-494, Vienna, Austria, 2007.
13. M.L. Yiu and N. Mamoulis, "Multi-Dimensional Top-k Dominating Queries", *The VLDB Journal*, Vol.18, No.3, pp.695-718, 2009.
14. Z. Zhang, Y. Yang, R. Cai, D. Papadias and A. Tung: "Kernel-Based Skyline Cardinality Estimation", *Proceedings ACM International Conference on Management of Data (SIGMOD)*, pp.509-522, 2009.