A Probabilistic Framework for Building Privacy-Preserving Synopses of Multi-dimensional Data[⋆]

Filippo Furfaro[1], Giuseppe M. Mazzeo[1,2], and Domenico Saccà[1,2]

[1] University of Calabria, Rende (CS) 87036, Italy
[2] ICAR-CNR, Rende (CS) 87036, Italy
{furfaro,mazzeo,sacca}@si.deis.unical.it

Abstract. The problem of summarizing multi-dimensional data into lossy synopses supporting the estimation of aggregate range queries has been deeply investigated in the last three decades. Several summarization techniques have been proposed, based on different approaches, such as histograms, wavelets and sampling. The aim of most of the works in this area was to devise techniques for constructing effective synopses, enabling range queries to be estimated, trading off the efficiency of query evaluation with the accuracy of query estimates. In this paper, the use of summarization is investigated in a more specific context, where privacy issues are taken into account. In particular, we study the problem of constructing *privacy-preserving synopses*, that is synopses preventing sensitive information from being extracted while supporting 'safe' analysis tasks. In this regard, we introduce a probabilistic framework enabling the evaluation of the quality of the estimates which can be obtained by a user owning the summary data. Based on this framework, we devise a technique for constructing histogram-based synopses of multi-dimensional data which provide as much accurate as possible answers for a given workload of 'safe' queries, while preventing high-quality estimates of sensitive information from being extracted.

1 Introduction

In the last three decades, a great deal of attention has been devoted to the problem of summarizing multi-dimensional data into synopses supporting the estimation of aggregate range queries. Several lossy compression techniques have been proposed, based on different approaches (such as histograms [11], wavelets [3], and sampling [7]). These techniques can be profitably applied in several application contexts (e.g., On-line Analytical Processing [7], query optimization [15], statistical and scientific databases [12]), where a high precision of query estimates is not mandatory, and fast query answers (affected by reasonable error rates) suffice to effectively support the tasks to be accomplished.

Intuitively enough, the experience acquired by the research community in designing effective lossy compression techniques could be applied in a new emerging scenario, where data should be published to support different analysis tasks,

[⋆] This work was supported by a grant from the Italian Research Project FIRB "TOCAI", funded by MUR.

B. Ludäscher and Nikos Mamoulis (Eds.): SSDBM 2008, LNCS 5069, pp. 114–130, 2008.

with no risk for privacy issues. That is, the compression process could be driven so that the loss of information is exploited to hide sensitive information, while 'safe' information is enabled to be accurately extracted from the synopses. Indeed, most of summarization techniques proposed in the previously mentioned scenarios provide no warranty on the privacy preservation of sensitive information. In fact, the compression process accomplished by these techniques aims at reducing as much as possible the loss of information resulting from summarizing data in a limited amount of storage space, paying no attention to the risk that sensitive information could be extracted from the summarized data with a high degree of accuracy. This makes the problem of refining traditional compression techniques to deal with privacy-preserving issues intriguing, also due to its practical impact in many application contexts.

In this paper we focus our attention on histogram-based summarization techniques, which are widely used in the context of data compression. A histogram is a synopsis obtained by suitably partitioning the data domain into a set of blocks and then replacing the set of individual data inside each block with some aggregate data. First, we introduce a probabilistic framework for evaluating the quality of the estimates of sensitive information which can be obtained by accessing a histogram. Specifically, the quality of estimates is measured by evaluating the probability associated with confidence intervals of individual-data estimates. This framework can be used to assign a 'safety certificate' to histograms, as it provides a measure of the privacy threat owing to the summary data published through a histogram. Thus, we exploit the proposed probabilistic framework to devise a technique for constructing *privacy-preserving histograms*. Our technique is based on a greedy strategy for constructing a partition of data which aims at two objectives: on the one hand, the resulting histogram should provide as much accurate as possible estimates for a workload of queries considered 'safe'; on the other hand, the resulting histogram should provide low-quality estimates of individual data. Finally, we address future directions towards which our work could be extended.

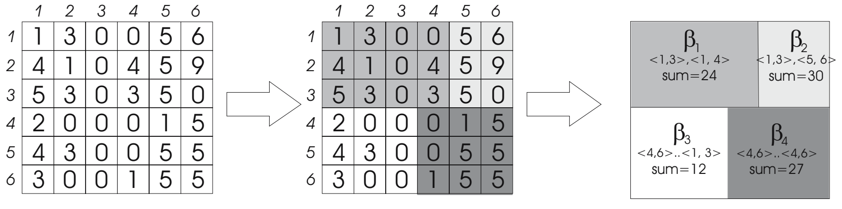
2 Preliminaries

In this work, we focus our attention on multi-dimensional data defined on a domain whose dimensions are discrete, and the values associated with the points of the domain are non-negative real numbers. Specifically, a d-dimensional data set D is a set of tuples of the form $\langle p_1, \ldots, p_d, m \rangle$, where $p_1, \ldots, p_d$ identify a point in a multi-dimensional space of size $n_1 \times \cdots \times n_d$ and m is a measure associated with the point. Thus, D can be viewed as a d-dimensional array of size $n_1 \times \cdots \times n_d$, where n_i is the cardinality of the i-th dimension. Given a point $\mathbf{p} = \langle p_1, \ldots, p_d \rangle$ of the domain of D, where $p_i \in [1..n_i]$ ($\forall i \in [1..d]$), the value m associated with $\mathbf{p}$ will be denoted as $D[\mathbf{p}]$ (if D contains no tuple associated with point $\mathbf{p}$, then $D[\mathbf{p}] = 0$). A range over D is a d-tuple $\varrho = \langle \varrho_1, \ldots, \varrho_d \rangle$, where ϱ_i ($\forall i \in [1..d]$) is a pair of the form $\langle \varrho_i^l, \varrho_i^u \rangle$ such that $1 \leq \varrho_i^l \leq \varrho_i^u \leq n_i$. Basically, a range is a hyper-rectangular subset of the domain of D. We define the volume of a range ϱ as $\Pi_{i \in [1..d]}(\varrho_i^u - \varrho_i^l + 1)$ and denote it as $vol(\varrho)$.

A histogram over D is a synopsis of aggregate values which is constructed by first partitioning the domain of D into a number of non-overlapping ranges, called *buckets*, and then storing, for each bucket, some aggregate data summarizing the data set underlying it. Histograms can be used to support the estimation of aggregate range queries, which are evaluated by exploiting the summary data stored in its buckets. Specifically, we study the case that, for each bucket, the sum of the values of the points inside it is stored (as it will be clearer in the following, this allows sum-range queries to be estimated). In this scenario, a bucket of a histogram over D can be viewed as a pair $\langle \varrho, s \rangle$, where ϱ is a range over D and $s = \sum_{\mathbf{p} \in \varrho} D[\mathbf{p}]$. Given a bucket $\beta = \langle \varrho, s \rangle$, the terms ϱ and s will be referred to as the *range* and the *sum* of β, respectively, and will be denoted as $range(\beta)$ and $sum(\beta)$. Moreover, we will denote the volume of the range of β as $vol(\beta)$ and the average value of β (i.e., $\frac{sum(\beta)}{vol(\beta)}$), as $\mu(\beta)$. This kind of histogram (where buckets are associated with sums) supports the evaluation of sum-range queries, that is, queries asking for the sum of the values of the points of D inside a specified range. A range-sum query over D is an expression of the form $q = sum(\varrho_q)$, where ϱ is a range over D. The actual answer of q is the value $\sum_{\mathbf{p} \in \varrho_q} D[\mathbf{p}]$. Given a histogram $H = \{\beta_1, \ldots, \beta_n\}$ over D, the estimated answer of q over H is $\tilde{q} = \sum_{i=1}^{n} vol\big(range(\beta_i) \cap \varrho_q\big) \cdot \frac{sum(\beta_i)}{vol(\beta_i)}$. Hence, the estimation is performed adopting linear interpolation, that is, assuming that the points inside a bucket β_i have the same value, namely the average value $\mu(\beta_i)$.

A point query on a data set D is a pair $q = \langle D, \mathbf{p} \rangle$ asking for the value $D[\mathbf{p}]$. Thus, q can be viewed as a range query where the specified range has volume 1, then the estimate of its answer obtained from histogram H is given by $\frac{sum(\beta)}{vol(\beta)} = \mu(\beta)$, where β is the bucket of H whose range contains $\mathbf{p}$. In the following, given a point query $q = \langle D, \mathbf{p} \rangle$ and a histogram H over D, we denote the bucket of H whose range contains $\mathbf{p}$ as $\beta(q)$ (observe that $\beta(q)$ is unique, as the buckets of H do not overlap).

Example 1. A two-dimensional data set D is shown in Fig. 1(a). A histogram on D is shown in Fig. 1(c). It has been obtained by first partitioning the domain of D (as in Fig. 1(b)) and then storing the boundaries and the sum of values of each block (bucket) of the partition. Consider the point query $q = \langle D, \langle 3, 4 \rangle \rangle$. In this case, the bucket involved in the query is $\beta(q) = \beta_1$, being the point $\langle 3, 4 \rangle$



	1	2	3	4	5	6
1	1	3	0	0	5	6
2	4	1	0	4	5	9
3	5	3	0	3	5	0
4	2	0	0	0	1	5
5	4	3	0	0	5	5
6	3	0	0	1	5	5

(a) (b) (c)

Fig. 1. A two-dimensional data set D (a), a partition of the domain of D (b) and a histogram H summarizing D (c)

inside the range of the bucket β_1. The estimated answer of q is $24/12 = 2$, since the sum and the volume of β_1 are 24 and 12, respectively. □

3 A Probabilistic Framework for Estimating Individual Values from a Histogram

In this section we present a probabilistic framework supporting the estimation of individual values based on the summary data stored in a histogram. Specifically, this framework provides a measure of the quality of the estimates of individual data which can be obtained by exploiting the aggregate data stored in a histogram. The quality measure is given in terms of probability that the estimation of an individual value is within a confidence interval.

Given a histogram H on a data set D and a point query $q = \langle D, \mathbf{p} \rangle$, we model the answer of q estimated on H as a random variable $\tilde{q}_{s,b}$ defined over the sample space $\Omega(q) = [0, s]$, where s and b are the sum and the volume of $\beta(q)$. Basically, $\tilde{q}_{s,b}$ can assume all the values inside the interval $[0, s]$ as the actual value associated with $\mathbf{p}$ is non-negative and cannot exceed the overall sum of the bucket of H whose range contains $\mathbf{p}$. It is worth noting that this random variable does not depend on parameters other than s and b, as histogram buckets do not overlap and we assume independence among the values summarized into different buckets, thus the sum values and the volumes of the buckets different from $\beta(q)$ do not affect the estimation of q.

We now characterize the above-introduced random variable $\tilde{q}_{s,b}$.

Theorem 1. *Let D be a data set, H a histogram over D, $q = \langle D, \boldsymbol{p} \rangle$ a point query, and s and b be the sum and the volume of bucket $\beta(q)$ of H, respectively. The probability density function of the random variable $\tilde{q}_{s,b}$ is:*

$$f(x) = \begin{cases} \delta(0) & \text{if } s = 0; \\ \delta(s) & \text{if } b = 1; \\ \frac{b-1}{s} \cdot \left(1 - \frac{x}{s}\right)^{b-2} & \text{if } b > 1,\ s > 0, \text{ and } x \in [0, s]; \\ 0 & \text{if } b > 1,\ s > 0, \text{ and } x \notin [0, s]; \end{cases} \tag{1}$$

where $\delta(x)$ denotes the Dirac function, its cumulative distribution function, is:

$$F(x) = \begin{cases} H(0) & \text{if } s = 0; \\ H(s) & \text{if } b = 1; \\ 1 - \left(1 - \frac{x}{s}\right)^{b-1} & \text{if } b > 1,\ s > 0, \text{ and } x \in [0, s]; \\ 0 & \text{if } b > 1,\ s > 0, \text{ and } x < 0; \\ 1 & \text{if } b > 1,\ s > 0, \text{ and } x > s; \end{cases} \tag{2}$$

where $H(x)$ denotes the Heaviside step function. The expected value and the variance of $\tilde{q}_{s,b}$ are

$$E(\tilde{q}_{s,b}) = \frac{s}{b} \tag{3}$$

and

$$\sigma^2(\tilde{q}_{s,b}) = \frac{b-1}{b+1}\left(\frac{s}{b}\right)^2 , \tag{4}$$

respectively.

Proof. We first focus on the expressions for $f(x)$ and $F(x)$. In the case that $s = 0$, as the elements of D are non-negative, the actual value associated with each point inside β is 0. Hence, $\tilde{q}_{s,b}$ takes value 0 with probability 1.

In the case that $b = 1$, β contains a unique element, thus the definition of $f(x)$ derives from the fact that the value associated with $\mathbf{p}$ is exactly s (the sum associated with β).

We now consider the case that $b > 1$ and $s > 0$. In this case, clearly $f(x)$ is null for $x \notin [0, s]$, as individual values are assumed to be non-negative and their sum is s (thus, no individual value can be larger than s). For the same reason, $F(x) = 0$ for $x < 0$ (it is impossible that any individual value is less than 0) and $F(x) = 1$ for $x > s$ (it is certain that any individual value is less than or equal to s). Now we derive $f(x)$ and $F(x)$ for the most interesting case, that is $b > 1$, $s > 0$, and $x \in [0, s]$. We first characterize a discrete random variable V_j different from $\tilde{q}_{s,b}$, whose probability distribution will be exploited to derive $f(x)$ and which is defined as follows. Given a real number $\gamma > 0$ and a set $S = \{k_1 \cdot \gamma, \ldots, k_b \cdot \gamma\}$ of cardinality b, where, for each $i \in [1..b]$, $k_i \in N$, and $\sum_{i=1}^{b} k_i \cdot \gamma = s$, $Pr(V_j = x)$ denotes the probability that the value of $k_j \cdot \gamma$ is equal to x. Intuitively enough, V_j can be viewed as the translation of $\tilde{q}_{s,b}$ to the case that the domain of the values of D is discrete (i.e., the points D can be assigned only multiples of γ). Thus, V_j is a discrete random variable defined over the sample space $\Omega(V_j) = \{x | 0 \leq x \leq s \text{ and } x \text{ is a multiple of } \gamma\}$. We now show that

$$Pr(V_j = x) = \frac{\binom{\frac{s-x}{\gamma} + b - 2}{\frac{s-x}{\gamma}}}{\binom{\frac{s}{\gamma} + b - 1}{\frac{s}{\gamma}}} . \tag{5}$$

This formula can be explained as follows. If a value in S is equal to x, then the sum of the remaining $b-1$ elements is $s - x$. Therefore, $Pr(V_j = x)$ is equal to the ratio between all the possible value assignments to $b-1$ elements such that their sum is $s - x$ and all the possible assignments to b elements such that their sum is s. The formula derives from the facts that each element can be assigned a multiple of γ, and that all the possible value assignments to n elements such that their sum is y is equal to number of combinations with repetitions of n objects from which y have to be chosen, that is $\binom{y+n-1}{y}$.

We denote the cumulative distribution function of V_j as $F_V(x)$, and derive a formula for $F_V(x)$:

$$F_V(x) = Pr(V_j \leq x) = \sum_{k=0}^{\lfloor \frac{x}{\gamma} \rfloor} Pr(V_j = k \cdot \gamma) =$$

$$= 1 - \sum_{k=\left\lfloor \frac{x}{\gamma} \right\rfloor + 1}^{\frac{s}{\gamma}} Pr(V_j = k \cdot \gamma) = 1 - \frac{1}{\binom{b+\frac{s}{\gamma}-1}{\frac{s}{\gamma}}} \sum_{k=\left\lfloor \frac{x}{\gamma} \right\rfloor + 1}^{\frac{s}{\gamma}} \binom{b + \frac{s}{\gamma} - k - 2}{\frac{s}{\gamma} - k}$$

Let $i = \frac{s}{\gamma} - k$. We obtain:

$$\sum_{k=\left\lfloor \frac{x}{\gamma} \right\rfloor + 1}^{\frac{s}{\gamma}} \binom{b + \frac{s}{\gamma} - k - 2}{\frac{s}{\gamma} - k} = \sum_{i=0}^{\frac{s}{\gamma} - \left\lfloor \frac{x}{\gamma} \right\rfloor - 1} \binom{b - 2 + i}{i}$$

and by adopting the identity

$$\sum_{j=0}^{k} \binom{n+j}{j} = \binom{n+k+1}{k}$$

we obtain:

$$F_V(x) = 1 - \frac{\binom{b - 2 + \frac{s}{\gamma} - \left\lfloor \frac{x}{\gamma} \right\rfloor}{\frac{s}{\gamma} - \left\lfloor \frac{x}{\gamma} \right\rfloor - 1}}{\binom{b + \frac{s}{\gamma} - 1}{\frac{s}{\gamma}}}. \tag{6}$$

The cumulative distribution function $F(x) = Pr(\tilde{q}_{s,b} < x)$ of $\tilde{q}_{s,b}$ can be obtained as $F(x) = \lim_{\gamma \to 0} F_V(x)$. In fact, as γ tends to 0, the elements of set S can be assigned any real value in $[0, s]$ (under the constraint that their sum is s), thus at the limit the distribution functions F and F_V coincide. Then, we obtain:

$$F(x) = \lim_{\gamma \to 0} F_V(x) = 1 - \lim_{\gamma \to 0} \frac{\left(b - 2 + \frac{s}{\gamma} - \left\lfloor \frac{x}{\gamma} \right\rfloor\right)! \cdot \left(\frac{s}{\gamma}\right)! \cdot (b-1)!}{\left(\frac{s}{\gamma} - \left\lfloor \frac{x}{\gamma} \right\rfloor - 1\right)! \cdot \left(b + \frac{s}{\gamma} - 1\right)! \cdot (b-1)!} =$$

$$= 1 - \lim_{\gamma \to 0} \frac{\overbrace{\frac{(s - \left\lfloor \frac{x}{\gamma} \right\rfloor \cdot \gamma) + (b-2) \cdot \gamma}{\gamma} \times \cdots \times \frac{(s - \left\lfloor \frac{x}{\gamma} \right\rfloor \cdot \gamma)}{\gamma}}^{b\text{–}1 \text{ factors}}}{\underbrace{\frac{s + (b-1) \cdot \gamma}{\gamma} \times \cdots \times \frac{s + 1 \cdot \gamma}{\gamma}}_{b\text{–}1 \text{ factors}}} =$$

$$= 1 - \lim_{\gamma \to 0} \frac{(s - \left\lfloor \frac{x}{\gamma} \right\rfloor \cdot \gamma)^{b-1} + o(\gamma)}{s^{b-1} + o(\gamma)} = 1 - \left(1 - \frac{x}{s}\right)^{b-1}$$

From definition of probability density function of a continuous random variable, we have that the probability density function $f(x)$ and the cumulative distribution function $F(x)$ are related as follows:

$$F(x) = \int_0^x f(u)du.$$

By resolving the latter and exploiting the boundary condition $F(s) = 1$, we obtain the expression for $f(x)$ reported in the statement.

We now derive the expected value of $\tilde{q}_{s,b}$. From definition of expected value, we obtain:

$$E(\tilde{q}_{s,b}) = \int_{x=0}^{s} f(x) \cdot x \cdot dx = \frac{b-1}{s} \int_{x=0}^{s} \left(1 - \frac{x}{s}\right)^{b-2} \cdot x \cdot dx =$$

$$= \frac{b-1}{s} \left[-\left(1 - \frac{x}{s}\right)^{b-1} \cdot \frac{s}{b-1} \cdot x \right]_{x=0}^{s} + \frac{b-1}{s} \int_0^s \left(1 - \frac{x}{s}\right)^{b-1} \cdot \frac{s}{b-1} dx =$$

$$= \frac{b-1}{s} \left[\frac{s}{b-1} \left(1 - \frac{x}{s}\right)^{b} \frac{s}{b} \right]_{x=0}^{s} = \frac{s}{b}.$$

Similarly, from the definition of variance, we obtain:

$$\sigma^2(\tilde{q}_{s,b}) \int_{x=0}^{s} f(x) \cdot \left(x - \frac{s}{b}\right)^2 dx = \frac{b-1}{s} \int_{x=0} s \left(1 - \frac{x}{s}\right)^{b-2} \cdot \left(x - \frac{s}{b}\right)^2 dx =$$

$$= \frac{b-1}{s} \left[-\left(1 - \frac{x}{s}\right)^{b-1} \frac{s}{b-1} \left(x - \frac{s}{b}\right)^2 + \frac{2s}{b-1} \int \left(1 - \frac{x}{s}\right)^{b-1} \left(x - \frac{s}{b}\right) dx \right]_{x=0}^{s} =$$

$$= \left(\frac{s}{b}\right)^2 + 2 \left[-\left(1 - \frac{x}{s}\right)^{b} \frac{s}{b} \left(x - \frac{s}{b}\right) + \frac{s}{b} \int \left(1 - \frac{x}{s}\right)^{b} dx \right]_{x=0}^{s} =$$

$$= \left(\frac{s}{b}\right)^2 - 2\left(\frac{s}{b}\right)^2 + 2\frac{2s}{b} \left[-\left(1 - \frac{x}{s}\right)^{b+1} \frac{s}{b+1} \right]_{x=0}^{s} =$$

$$= -\left(\frac{s}{b}\right)^2 + \frac{2s}{b} \frac{s}{b+1} = \frac{b-1}{b+1} \left(\frac{s}{b}\right)^2. \qquad \square$$

The characterization of random variable $\tilde{q}_{s,b}$ can be exploited to determine the quality of the point-query estimates which can be obtained by accessing the summary data stored in H. In fact, a user owning the histogram can estimate the answer of a point query q as the expected value of $\tilde{q}_{s,b}$ (which corresponds to performing linear interpolation), and evaluate the quality of this estimate as the probability that the actual answer of q lies inside an interval containing $E(\tilde{q}_{s,b})$ as wide as desired. For instance, consider the data set D and the histogram H shown in Fig. 1, as well as the point query $q = \langle D, \mathbf{p} \rangle$, with $\mathbf{p} = \langle 3, 4 \rangle$. If only the aggregate data stored in the histogram is available, the answer of q is estimated as $E(\tilde{q}_{24,12}) = 2$, since the sum and the volume of the bucket β_1 of H containing the point $\langle 3, 4 \rangle$ are 24 and 12, respectively, as seen in Example 1. A user owning the histogram cannot infer the actual value associated with $\mathbf{p}$, but she can evaluate the probability associated to any confidence interval. For instance, a user could be interested in evaluating the probability that the value associated with $\mathbf{p}$ is inside $[1.8, 2.2]$, that is a 'narrow' range centered at the expected value. Using the results provided in Theorem 1, the user obtains that the probability that the actual answer is in $[1.8, 2.2]$ is $F(2.2) - F(1.8) = (1 - \frac{1.8}{24})^{11} - (1 - \frac{2.2}{24})^{11}$.

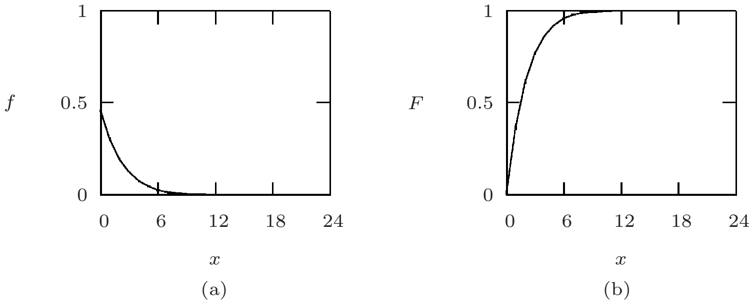


Fig. 2. Probability density (a) and distribution (b) functions of $\tilde{q}_{24,12}$

Fig. 2 depicts the probability density function (a) and the distribution function (b) of the random variable $\tilde{q}_{24,11}$.

Intuitively enough, as our framework can be used to measure the quality of estimates of queries asking for sensitive information, it can be exploited to determine whether a histogram can be considered safe or not w.r.t. a privacy standpoint. This matter is investigated in the following section.

3.1 Privacy and Histograms

Given a histogram H over a data set D, a *privacy breach* occurs if an adversary can retrieve from H "high"-quality estimates of individual data, that is she can reveal sensitive information by establishing with a high confidence level that an individual value is within a certain interval.

In the following we will devise a histogram construction technique which aims at preventing any user owning a histogram from establishing that the actual value associated with a point is "close" to its estimate with a probability higher than a certain threshold. This is tantamount to requiring that the estimated value of every individual data must be affected by a certain error with a probability at least equal to a certain threshold. For instance, a company publishing summary data about the incomes of its employees would like to impose that the estimate of the the income of a single employee evaluated by accessing the summary data is affected by at least 50% error with a probability greater than 70%.

In this example, we used the relative error to define the threshold guaranteeing the safeness of the summary data. However, different metrics could be used, such as the absolute error. In the following, we will consider the relative error as it is quite intuitive and it has been largely stressed in literature [6,8] that it represents a significant measure of the quality of the estimates. On the basis of this idea, we introduce the notion of *privacy preserving bucket* and *privacy preserving histogram.*

Definition 1. *Given a data set D, a histogram H on D, and two real numbers $\epsilon, \mathcal{P} \in (0,1)$, a bucket β of H is said to be $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving if, for every point query $q = \langle D, \boldsymbol{p} \rangle$, where* **p** *is a point laid inside the range of β, it holds that:*

$$Pr\big(|\tilde{q}_{s,b} - E(\tilde{q}_{s,b})| \leq \epsilon \cdot E(\tilde{q}_{s,b})\big) < \mathcal{P}, \tag{7}$$

where s and b are the sum and the volume of β. □

Definition 2. *Given a data set D, a histogram H on D, and two real numbers $\epsilon, \mathcal{P} \in (0,1)$, H is said to be $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving if every bucket of H is $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving.* □

According to Definition 2, a histogram H on a data set D is not privacy preserving (w.r.t. a pair $\langle \epsilon, \mathcal{P} \rangle$) if it does not protect the privacy of at least one point **p**, that is the value associated with **p** is summarized in a bucket with sum s and volume b such that

$$Pr\big(|\tilde{q}_{s,b} - E(\tilde{q}_{s,b})| \leq \epsilon \cdot E(\tilde{q}_{s,b})\big) \geq \mathcal{P},$$

where q is the point query asking for the value of **p**.

Hence, a pair $\langle \epsilon, \mathcal{P} \rangle$ defines a *privacy constraint*, and the values assigned to ϵ and $\mathcal{P}$ must be chosen according to the specific context where privacy must be guaranteed.

As $Pr\big(|\tilde{q}_{s,b} - E(\tilde{q}_{s,b})| \leq \epsilon \cdot E(\tilde{q}_{s,b})\big) = F\big(\frac{s}{b} \cdot (1+\epsilon)\big) - F\big(\frac{s}{b} \cdot (1-\epsilon)\big)$, where $F(\cdot)$ is the cumulative distribution function of $\tilde{q}_{s,b}$ derived in Theorem 1 (see Equation 2), under the assumption that $s > 0$, we find that[1]:

$$Pr\big(|\tilde{q}_{s,b} - E(\tilde{q}_{s,b})| \leq \epsilon \cdot E(\tilde{q}_{s,b})\big) = \left(1 - \frac{1-\epsilon}{b}\right)^{b-1} - \left(1 - \frac{1+\epsilon}{b}\right)^{b-1}. \tag{8}$$

Interestingly, from Equation 8, it turns out that the probability associated with a confidence interval of an estimate does not depend on the sum of the bucket summarizing the value to be estimated. Thus, in the following, we will refer to $Pr\big(|\tilde{q}_{s,b} - E(\tilde{q}_{s,b})| \leq \epsilon \cdot E(\tilde{q}_{s,b})\big)$ simply as $P(b, \epsilon)$.

In Fig. 3, the diagrams of $P(b, \epsilon)$ against ϵ and b are shown. It is worth observing that $P(b, \epsilon)$ is monotone increasing w.r.t. ϵ and monotone decreasing w.r.t. b. This means that a privacy constraint $\langle \epsilon, \mathcal{P} \rangle$ implies a lower bound on

[1] The case $s = 0$ can be disregarded, as it implies that every individual value inside the bucket is 0 with probability 1.

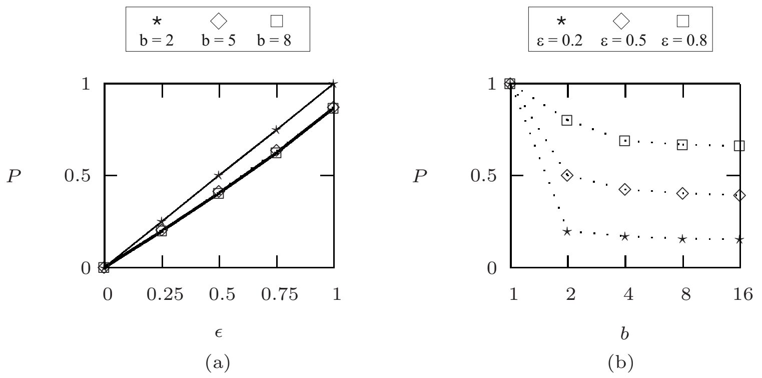


Fig. 3. $P(b, \epsilon)$ vs. ϵ, for different values of b (a), and vs. b for different values of ϵ

the volume of the buckets of a histogram: in order to satisfy the constraint $P(b, \epsilon) < \mathcal{P}$, a histogram must consist of only buckets having at least volume $b_{min} = \lfloor b^\star + 1 \rfloor$, where $b^\star$ is the solution of the equation

$$P(b^\star, \epsilon) = \mathcal{P}. \tag{9}$$

Solving Equation 9 is not possible through analytical methods, but the value of b_{min} can be efficiently computed[2] starting from $b = 1$ and iteratively incrementing b by one and computing $P(b, \epsilon)$, by adopting Equation 8, until $P(b, \epsilon) < \mathcal{P}$ holds.

The monotonicity of $P(\epsilon, b)$ w.r.t. b is at the basis of the property stated in the following proposition.

Proposition 1. *If a histogram H summarizing a data set D is not $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving, then there is no split-sequence of its buckets that can yield a $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving histogram.*

Proof. If a histogram is not $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving, then it has at least a bucket β of volume b with $P(\epsilon, b) \geq \mathcal{P}$. Any split of β yields new buckets with volume $b' < b$. It is easy to see that $\frac{\partial P}{\partial b} < 0$, thus $P(\epsilon, b') > P(\epsilon, b) \geq \mathcal{P}$ holds too. This implies that any histogram obtained from H by splitting β is not $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving. □

The result is quite intuitive after considering that a privacy constraint $\langle \epsilon, \mathcal{P} \rangle$, as discussed above, is satisfied only if each bucket of the histogram has volume not less than b_{min}, where b_{min} can be computed by solving Equation 9. Thus, if a bucket β has volume $b < b_{min}$, any sub-bucket of β will have volume less than b_{min}.

In the following section we will introduce a greedy algorithm for constructing privacy preserving histograms, which exploits the property stated in Proposition 1.

[2] The time needed to compute b_{min} for one million different combinations of ϵ and $\mathcal{P}$ is only 2.7 seconds with a Pentium IV 3.2 GHz.

4 A Greedy Algorithm for Constructing Privacy-Preserving Histograms

In many practical cases, summarized data can be effectively exploited for performing statistical analysis. Several summarization techniques have been devised to support an efficient query evaluation, aiming at providing query answers affected by the least possible error. When privacy-constraints are defined over published data, the summarization has to take them into account. That is, on the one hand, answers of queries should be accurate enough to enable statistical analysis to be performed. On the other hand, published data must prevent sensitive information from being inferred.

In this work, we focus our attention on constructing privacy-preserving histograms which can be profitably exploited for statistical data analysis. Specifically, we consider the problem of constructing a privacy preserving histogram which aims at providing as much accurate as possible estimates of sum range queries considered 'safe'. More formally, given a privacy constraint $\langle \epsilon, \mathcal{P} \rangle$ and a query workload W consisting of m sum range queries defined over a multidimensional data set D, we consider the problem of constructing a $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving histogram H summarizing D which minimizes the error over the queries in W. In order to measure the error over a workload we consider the sum of squared errors of the range queries in the workload. That is, if $W = \{sum(\rho_1), \ldots, sum(\rho_m)\}$, being q_i the exact answer of the sum range query in W over the range ρ_i and $\tilde{q}_i$ the approximate one, the overall estimation error w.r.t. W is defined as:

$$SSE(D, H, W) = \sum_{i=1}^{m} (q_i - \tilde{q}_i)^2 .$$

Constructing the optimal histogram for a query workload over a multidimensional data set has been proved to be an NP-hard problem (in [14], Muthukrishnan et al. showed that constructing the optimal histogram of a two-dimensional data set is NP-hard when the query workload consists of all the possible point queries). Thus, we show how our probabilistic framework can be exploited in a greedy algorithm for building (possibly non-optimal) histograms for a given query workload.

Our algorithm (see Fig. 4) works as follows. It takes as input a data set D, a query workload W, and a privacy constraint $\langle \epsilon, \mathcal{P} \rangle$, and returns a $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving histogram summarizing D. It starts from a histogram consisting of a unique bucket (corresponding to the whole data domain), and it iteratively refines the current histogram by taking a bucket and splitting it into two smaller buckets. Being H' and H'' the histogram at the beginning and at the end of the current iteration, respectively, the choice of the most suitable split for a bucket β of H' is accomplished by function bestSafeSplit, which returns, among all the splits yielding two privacy preserving sub-buckets of β, the split which maximizes the difference $SSE(D, H', W) - SSE(D, H'', W)$, where H'' is the histogram obtained from H' by replacing β with the pair of buckets resulting

```
Input:  A data set $D$, a query workload $W$, and a privacy constraint $\langle \epsilon, \mathcal{P} \rangle$
Output: An $\langle \epsilon, \mathcal{P} \rangle$-privacy-preserving histogram summarizing $D$

begin
  Histogram definitive=new Histogram();
  Histogram refinable=new Histogram();
  Bucket $\beta$=new Bucket(range(D), sum(D));
  refinable.add($\beta$);
  while (!refinable.isEmpty()) do begin
    $\beta$=refinable.remove();
    $\langle \beta', \beta'' \rangle$=$\beta$.bestSafeSplit(D,W,$\epsilon$,$\mathcal{P}$);
    if ($\langle \beta', \beta'' \rangle$==null) then
      definitive.add($\beta$);
    else begin
      refinable.add($\beta'$);
      refinable.add($\beta''$);
    end;
  end;
  return definitive;
end;
```

Fig. 4. A greedy algorithm for constructing privacy preserving histograms

from the split. If no split exists for β yielding two privacy preserving buckets, then β is considered as a *definitive* bucket, and will be not considered for further splits in the subsequent iterations. In fact, from Proposition 1 we have that, if a non-privacy-preserving bucket were created by splitting β, at least one non-privacy-preserving bucket would exist at every subsequent iteration, and then the final histogram would not be privacy-preserving. The algorithm ends when there is no bucket of the current histogram which can be safely split. In the pseudo-code implementation shown in Fig. 4, buckets which can be still considered for being split are maintained in the histogram refinable, while definitive buckets are put in the histogram definitive. Thus, at each iteration, the current histogram is the union between the sets of buckets stored in refinable and definitive.

Observe that any bucket of refinable can be chosen to be split at each iteration[3]. In fact, the split of a bucket at a given iteration does not influence the possibility to split the other buckets in refinable.

We now analyze the complexity of the algorithm. We assume that D contains N points distributed across a multidimensional domain of size n^d (i.e., d dimensions each of size n) and that W contains m queries. The number of iterations of

[3] Indeed, in the case that the histogram size were bounded by a maximum amount of storage space, the choice of the bucket to be split at each iteration could be performed according to some greedy criterion (e.g., the bucket giving the largest contribution to the overall error could be chosen).

the algorithm is $O(N)$, as each iteration increases by one the number of buckets, and the final number of buckets cannot be larger than N (a bucket must contain at least one non-null value). Thus, the complexity of the algorithm depends on the cost of the bestSafeSplit function, which is called $O(N)$ times. At each call of bestSafeSplit, all the $O(d \cdot n)$ possible splits must be tried, and for each split a range query of cost $O(N)$ must be performed for each query of the workload, in order to evaluate the SSE reduction provided by the split. The cost of checking if each split is safe is constant. In fact, according to Definition 2, in order to verify if a split is safe, the value b_{min} could be computed before starting the iterations, and then the function bestSafeSplit simply checks if the two buckets resulting from the split have volume greater than b_{min} and sum greater than 0. Therefore, the time complexity of the algorithm is $O(N^2 \cdot m \cdot n \cdot d)$.

Remark. Our probabilistic framework is suitable for being embedded in summarization techniques constructing histograms whose bucket do not overlap. This is due to the fact that, in this case, a point query can be estimated by accessing one bucket only. Several well-known techniques have this characteristic, such as *MHIST* [15] and *MinSkew* [1]). However, some techniques constructing histograms whose bucket overlap could exploit our probabilistic framework as well. For instance, when a bucket is nested inside another bucket, representing a 'hole', the estimate of a point query still depends on a unique bucket. Two techniques belonging to this class are *CHIST* [5] and *STHoles* [2].

5 Extending the Basic Results in Further Directions

In this section we trace further directions towards which our work could be extended:

- managing privacy when additional information is known about original data in buckets;
- managing other forms of privacy constraints;
- managing privacy when buckets overlap.

5.1 Managing Privacy When Additional Information Is Known about Original Data in Buckets

The results derived in Section 3 are based on the assumption that nothing is known about the original data inside each bucket, except that their sum is s and they are distributed in the bucket range of volume b. In many real cases, further information could be available due to the specific application context. For instance, if the measure associated with points is represented by integers, the probability distribution associated with the random variable representing the estimate of individual values would not be that derived in Theorem 1. In this case, the random variable would be discrete, thus its sample space would be $\{0, 1, 2, \ldots, s\}$ instead of $[0, s]$. However, the corresponding random variable could be characterized even easier than the continuous case previously studied.

In fact, the new random variable probability distribution would be represented by Equation 5, with $\gamma = 1$. Then, its cumulative distribution function would be represented by Equation 6, again with $\gamma = 1$. That is,

$$Pr(\tilde{q}_{s,b} \leq x) = 1 - \frac{\binom{b-2+s-x}{s-x-1}}{\binom{b+s-1}{s}}.$$

Another issue which is worth investigating is the case that other aggregate data (such as the count of non-null values, the minimum or the maximum value) inside each bucket is known. This may happen if either this summary information is explicitly represented in the histogram along with the sums of the buckets (to enhance the estimation process) or it is retrieved from different sources.

5.2 Managing Other Forms of Privacy Constraints

The definition of privacy provided in Section 3.1 can cover a large number of practical cases, in which exact individual values have to be protected. However, some other forms of privacy are worth investigating, due to their practical impact. According to our approach, guaranteeing the privacy of an individual value means limiting the confidence level associated with a confidence interval whose width is proportional to the expected value. It would be interesting to study the case that the width of the confidence interval is defined by an absolute value (rather than a relative one), that is that the confidence interval is expressed in the form $[E(\tilde{q}) - \Delta, E(\tilde{q}) + \Delta]$, where $\tilde{q}$ is the estimate of an individual value which must be protected, and Δ is a real number. In particular, it would be interesting to devise an algorithm managing mixed forms of constraints, where the width of confidence intervals can be expressed by either relative or absolute values. In fact, using an absolute value is more suitable for buckets summarizing “small” values, whereas a relative value is more suitable for buckets summarizing “large” values (where the meaning of “small” and “large” depends on the specific application context). This is due to the fact that adopting a relative value for describing intervals centered at “small” values would result in defining “narrow” intervals, for which guaranteeing low confidence levels would not suffice to preserve privacy.

5.3 Managing Privacy When Buckets Can Overlap

Even though classical histograms are based on partitions of the multi-dimensional data domain (thus, their buckets do not overlap) some of the most performing techniques, such as *GENHIST* [9], exploit bucket overlapping in order to summarize the data set more accurately. To this aim, our framework should be extended to enable taking into account the possibility that the estimation of a single point depends on the aggregate data stored in a number of buckets. For instance, in the case that a point **p** is within the ranges of two overlapping buckets, the random variable representing the value associated with $\mathbf{p}$ would depend on the random variables $\tilde{q}_{s',b'}$ and $\tilde{q}_{s'',b''}$, each representing the value of $\mathbf{p}$ given by one of the two buckets, independently. The random variable representing the expected value associated with $\mathbf{p}$ would be represented by the sum of the two random variables $\tilde{q}_{s',b'}$ and $\tilde{q}_{s'',b''}$. Thus, its probability density function could be obtained by computing the convolution of the probability density functions of $\tilde{q}_{s',b'}$ and $\tilde{q}_{s'',b''}$. Computing the convolution of many probability density functions could be practically infeasible. However, for a large number of random variables, that is, when the value associated to a point inside the intersection of a large number of overlapping buckets must be estimated, for the central limit theorem, the random variable could be very well approximated by a normal distribution that could be completely characterized by knowing the expected values and the variances of the random variables which have to be summed.

6 Related Work

The problem of managing privacy in statistical databases has received a lot of attention in the last few years, and several works dealing with data summarization and privacy issues have been proposed. However, few works providing formal frameworks for checking the privacy preservation of summarized data have been developed.

Some works provide techniques for summarizing data with quality guarantees [6,10]. However, in these works, the quality is intended as a measure of the "distance" between a synopsis and the optimal synopsis consuming the same amount of storage space. Thus, no guarantee is provided on the error rates of query estimates which could be exploited to measure the safeness of a synopsis. Our probabilistic framework, instead, does not aim at providing a technique for building optimal histograms, but provides a tool for evaluating the quality of individual value estimates, intended as confidence levels related to confidence intervals.

A work which deals with the privacy guaranteed by histograms is [4]. In this paper Chawla et al. consider points of a multidimensional space as individuals, which are not associated with any label. A privacy violation occurs when a user can isolate less than t points inside a spherical region of radius proportional to a value c (c and t are parameters which have to be chosen according to the practical context). This work, analogously to others based on the preservation of anonymity of individuals [16], is different from ours as it aims at masking the identity of individuals, that is the coordinates of the points inside the multidimensional domain (which are not associated with any measure). Our work, instead, deals with labelled points, more specifically, with points which are associated with an additive measure. Thus, our approach to privacy preservation is orthogonal w.r.t. [4]: we aim at protecting the measure associated with individuals, rather than their identity.

A thread of works where the attention is focused on the possibility to infer sensitive information by means of range queries on multidimensional data is that leaded by Malvestuto et al. [13]. They study the possibility to infer confidential information exploiting the answers of multiple range queries which, separately, could be considered safe. They design a query engine providing safe answers, which keeps track of past queries, and checks that the answer of each new query cannot be combined combined with the answers previously published in order to enable sensitive information to be inferred. Our approach is different since we assume that to release the whole data set is summarized and published. A very interesting point of contact between the issues studied in [13] and this paper could be the study of the possibility to release multiple safe histograms, each optimized for a different query workload. In fact, when different histograms are released, the fact that each of them is privacy preserving does not suffice to guarantee that confidential information cannot be disclosed, as a user owning different histograms on the same data set could exploit them jointly.

7 Conclusions

In this work we provided a novel approach for constructing effective histograms in the presence of privacy constraints. We introduced the notion of privacy-preserving histograms, that is histograms preventing a user owning them to obtain high quality estimates of individual values which must be kept confident. We defined a probabilistic framework for estimating individual values summarized in a histogram and, on the basis of our probabilistic framework, we proposed a greedy approach for constructing privacy-preserving histograms with high data utility, that is privacy-preserving histograms minimizing the estimation error for range queries belonging to a given query workload supporting statistical analysis tasks. Finally, we outlined the directions towards which our work could be extended.

To the best of our knowledge, this is the first work presenting a mechanism enabling the quality of the estimates of individual values which can be retrieved from a histogram to be measured. Our approach to the problem of preserving the privacy of data can be viewed as orthogonal to other ones, which aim at masking the identity of points belonging to a multi-dimensional domain. Our approach, in fact, aims at protecting a measure associated with the individuals, rather than protecting the identity of individuals.

References

1. Acharya, S., Poosala, V., Ramaswamy, S.: Selectivity estimation in spatial databases. In: Proc. of 1999 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 1999), Philadelphia (PA), USA, June 1-3, 1999, pp. 13–24 (1999)
2. Bruno, N., Chaudhuri, S., Gravano, L.: STHoles: a multi-dimensional workload aware histogram. In: Proc. of 2001 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2001), Santa Barbara (CA), USA, May 21-24, 2001, pp. 211–222 (2001)
3. Chakrabarti, K., Garofalakis, M.N., Rastogi, R., Shim, K.: Approximate query processing using wavelets. The VLDB Journal 10(2-3), 199–223 (2001)

4. Chawla, S., Dwork, C., McSherry, F., Smith, A., Wee, H.: Toward Privacy in Public Databases. In: Kilian, J. (ed.) TCC 2005. LNCS, vol. 3378, pp. 363–385. Springer, Heidelberg (2005)
5. Furfaro, F., Mazzeo, G.M., Sirangelo, C.: Exploiting cluster analysis for constructing multi-dimensional histograms on both static and dynamic data. In: Ioannidis, Y., Scholl, M.H., Schmidt, J.W., Matthes, F., Hatzopoulos, M., Böhm, K., Kemper, A., Grust, T., Böhm, C. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 442–459. Springer, Heidelberg (2006)
6. Garofalakis, M.N., Gibbons, P.B.: Wavelet synopses with error guarantees. In: Proc. of 2002 ACM SIGMOD Int. Conf. on Managment of Data (SIGMOD 2002), Madison (WI), USA, June 3-6, 2002, pp. 476–487 (2002)
7. Gibbons, P.B., Matias, Y.: New sampling-based summary statistics for improving approximate query answers. In: Proc. of 1998 ACM SIGMOD Int. Conf. on Managment of Data (SIGMOD 1998), Seattle (WA), USA, June 2-4, pp. 331–342 (1998)
8. Guha, S., Shim, K., Woo, J.: REHIST: Relative Error Histogram Construction Algorithms. In: Proc. of 30th Int. Conf. on Very Large Data Bases (VLDB 2004), Toronto, Canada, August 29-September 30, pp. 300–311 (2004)
9. Gunopulos, D., Kollios, G., Tsotras, V.J., Domeniconi, C.: Selectivity estimators for multidimensional range queries over real attributes. The VLDB Journal 14(2), 137–154 (2005)
10. Jagadish, H.V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K., Suel, T.: Optimal histograms with quality guarantees. In: Proc. of 24th Int. Conf. on Very Large Data Bases (VLDB 2004), New York (NY), USA, August 24-27, pp. 275–286 (2004)
11. Ioannidis, Y.E.: The History of Histograms (abridged). In: Proc. of 29th Int. Conf. on Very Large Data Bases (VLDB 2003), Berlin, Germany, September 9-12, pp. 19–30 (2003)
12. Malvestuto, F.M.: A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables. ACM Transactions on Database Systems 18(4), 678–708 (1993)
13. Malvestuto, F.M., Mezzini, M., Moscarini, M.: Auditing sum-queries to make a statistical database secure. ACM Transactions on Information and Systems Security 9(1), 31–60 (2006)
14. Muthukrishnan, S., Poosala, V., Suel, T.: On Rectangular Partitioning in Two Dimensions: Algorithms, Complexity and Applications. In: Proc. 7th Int. Conf. on Database Theory (ICDT), Jerusalem, Israel, January 10-12 (1999)
15. Poosala, V., Ioannidis, Y.E.: Selectivity estimation without the attribute value independence assumption. In: Proc. of 23rd Int. Conf. on Very Large Data Bases (VLDB 1997), Athens, Greece, August 25-29, pp. 486–495 (1997)
16. Sweeney, L.: k-Anomity: A model for protecting privacy. Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)