A Probabilistic Framework for Building Privacy-preserving Synopses of Multi-dimensional Data

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

[1]University of Calabria, Rende (CS) 87036, Italy

[2]ICAR-CNR, Rende (CS) 87036, Italy

SSDBM 2008 – July 9-11, 2008, Hong Kong

Outline

- Multi-dimensional data summarization
- Histograms and sensitive information disclosure
- A probabilistic framework for evaluating privacy preservation of histograms
- Construction of privacy-preserving histograms
- Conclusions and future works

Outline

- Multi-dimensional data summarization
- Histograms and sensitive information disclosure
- A probabilistic framework for evaluating privacy preservation of histograms
- Construction of privacy-preserving histograms
- Conclusions and future works

Multi-dimensional data summarization

- Application contexts: selectivity estimation, OLAP range queries (for preliminary explorations), etc.
- Goal: providing approximate but fast answers to range queries, which can be adopted for useful *statistical analysis*
- Dozens of existing techniques: sampling, wavelet, histograms

Multi-dimensional data summarization

- Application contexts: selectivity estimation, OLAP range queries (for preliminary explorations), etc.
- Goal: providing approximate but fast answers to range queries, which can be adopted for useful *statistical analysis*
- Dozens of existing techniques: sampling, wavelet, histograms

Example of histogram

Let ***D*** be a two-dimensional data set (discrete dimension domains and nonnegative real measure)

5	2	1	0	0	0	0	0
3	4	0	0	0	0	0	0
2	0	0	5	0	8	4	1
0	3	1	0	0	9	3	4
0	0	6	9	1	0	0	2
0	0	7	8	2	0	3	3
1	0	0	4	6	1	3	3
0	1	0	2	2	2	0	0

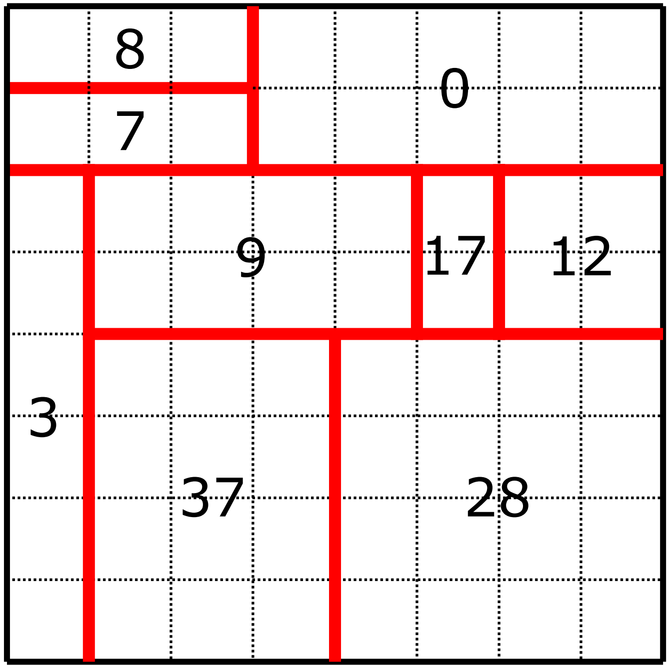
Example of histogram

Data domain is partitioned into *buckets*…

5	2	1	0	0	0	0	0
3	4	0	0	0	0	0	0
2	0	0	5	0	8	4	1
0	3	1	0	0	9	3	4
0	0	6	9	1	0	0	2
0	0	7	8	2	0	3	3
1	0	0	4	6	1	3	3
0	1	0	2	2	2	0	0

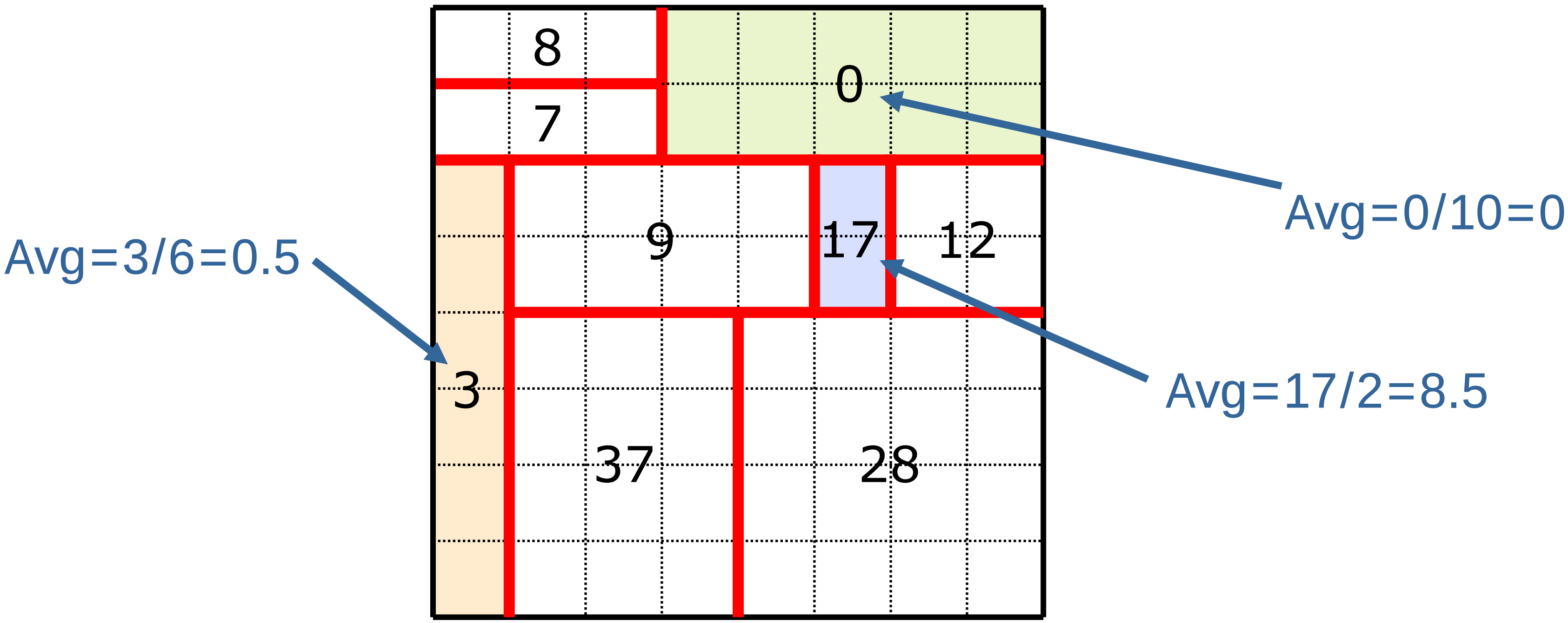
Example of histogram

…for each bucket its boundaries and the sum of its elements are stored



Example of histogram

Queries are evaluated by assuming that each point inside a bucket is associated with the same value (i.e, the bucket average, sum/volume)



Histogram construction algorithms

- The goal of algorithms for constructing histograms is to define the “best” partition of the data domain within a storage space bound
- Constructing the histogram which minimize the overall error of a query workload is a NP-Hard problem [Muthukrishnan et al., 1999]
- Several greedy approaches have been proposed in the last three decades
- Few work dealing with the disclosure of sensitive information from data summarized by means of histograms

Outline

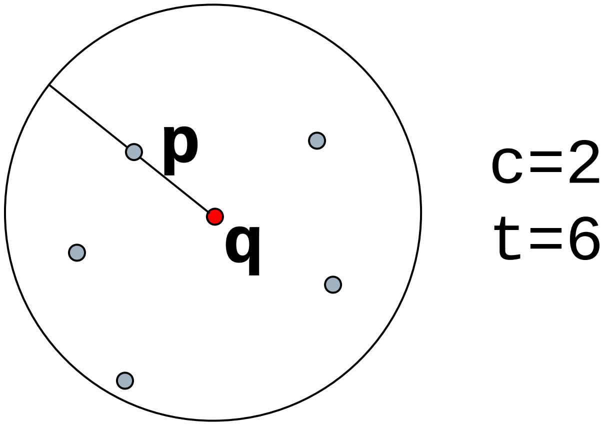
- Multi-dimensional data summarization
- Histograms and sensitive information disclosure
- A probabilistic framework for evaluating privacy preservation of histograms
- Construction of privacy-preserving histograms
- Conclusions and future works

Histograms and sensitive information disclosure

- One of the main works on histograms and privacy considers privacy as *protection from being brought to the attention of others* [Chawla et al., 2005]
- The work focuses on unlabelled points, representing individuals, whose identity (i.e., the point coordinates) must be protected
- The summarization must prevent individuals from being isolated

Histograms and sensitive information disclosure

- A point **p** is isolated by a point **q** if the ball of radius $c \cdot |\boldsymbol{q}-\boldsymbol{p}|$ centered at **q** contains less than t points

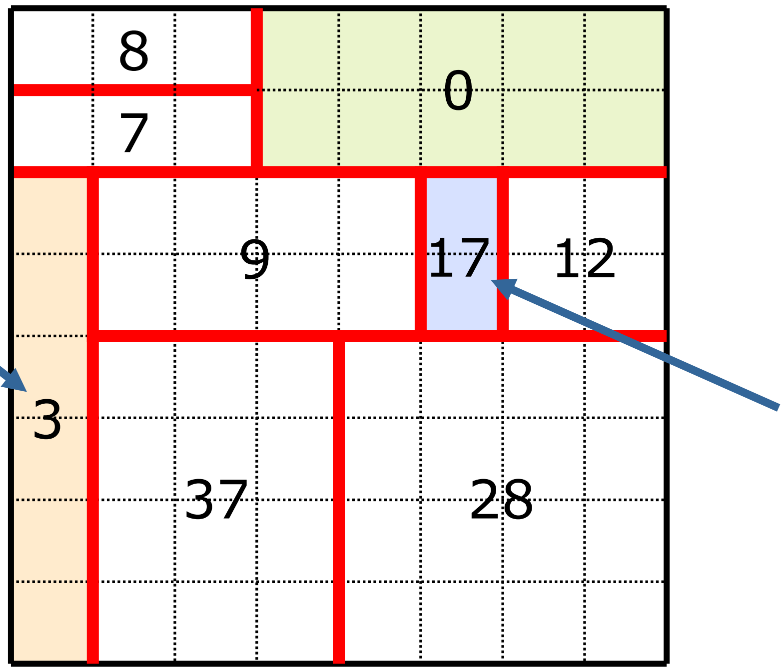


- We study a different problem: individuals, identified by their coordinates, are associated with values which must be kept confidential

Histograms and sensitive information disclosure

What can we say about individual values summarized by the histogram?

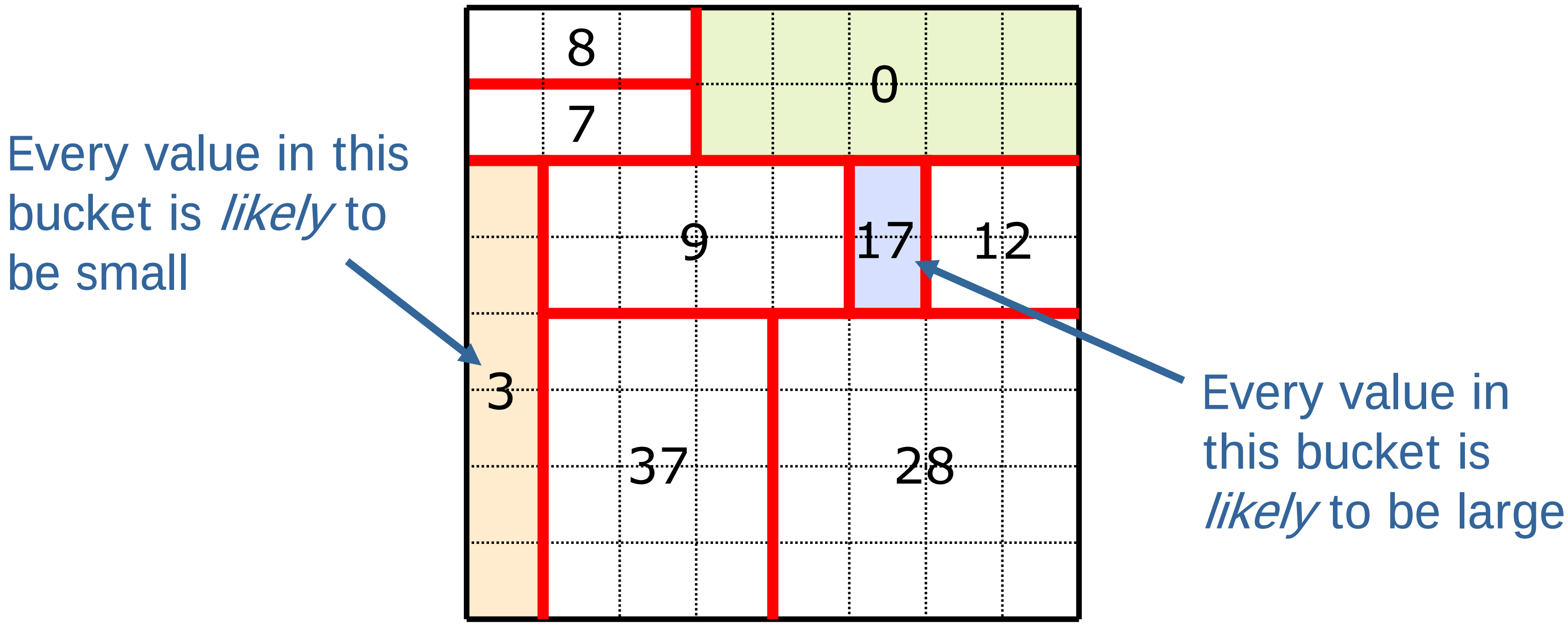
Every value in this bucket is *likely* to be small



Every value in this bucket is *likely* to be large

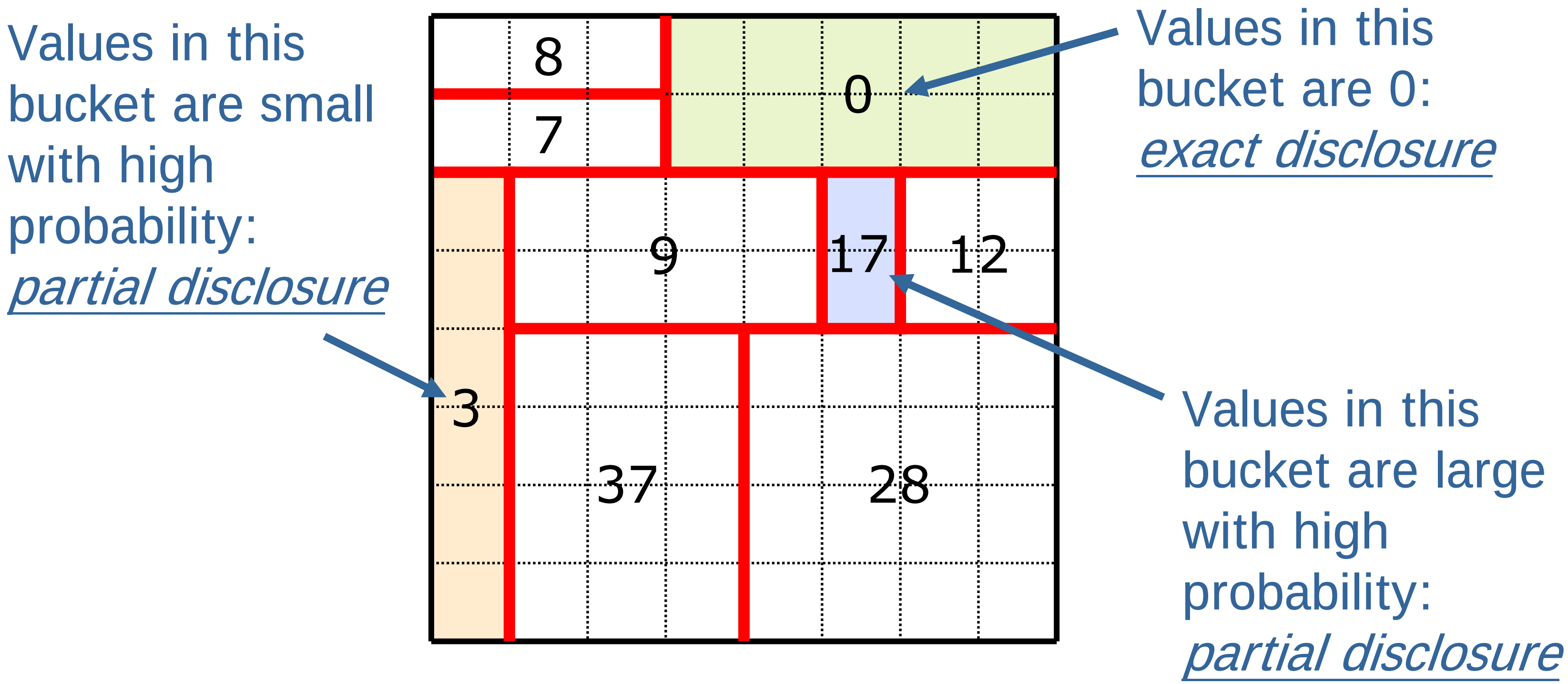
Histograms and sensitive information disclosure

How much “likely” to be small/large?
If “too likely” privacy of individuals could be compromised!



Histograms and sensitive information disclosure

How much "likely" to be small/large?
If "too likely" privacy of individuals could be compromised!



Histograms and sensitive information disclosure

- Intuitively, privacy is compromised when
  it is possible to infer with high probability that an individual value belongs to a range revealing some reserved information
- Examples: the annual income of an employee is
  - in [24,000..26,000] with probability 90%
  - greater than 1,000,000 with probability 80%
  - less than 10,000 with probability 85%

Histograms and sensitive information disclosure

- We focus our attentions on the possibility to estimate with high confidence the actual values associated with individuals (i.e., points)
- Privacy of an individual value X, estimated to be E(X), is compromised if its actual value is inside [(1-ε)·E(X), (1+ε)·E(X)] (ε-confidence-interval) with probability higher than P
- E.g., the privacy of an employee, whose income is estimated to be 25,000, is compromised if its actual income is in [24,000, 26,000] (ε=0.04) with probability higher than 90% (P=0.9)

Privacy-preserving histograms

- A pair ‹ε, P› will represent a privacy constraint
- A bucket β is said to be privacy-preserving w.r.t. a privacy constraint ‹ε, P› if for each individual value X inside β the confidence interval [(1-ε)·E(X), (1+ε)·E(X)] has confidence level (probability) less than P
- A histogram is said to be privacy-preserving w.r.t. a privacy constraint ‹ε, P› if it consists of only privacy-preserving buckets w.r.t. the same privacy constraint ‹ε, P›

Outline

- Multi-dimensional data summarization
- Histograms and sensitive information disclosure
- A probabilistic framework for evaluating privacy preservation of histograms
- Construction of privacy-preserving histograms
- Conclusions and future works

Probabilistic Framework

- By modeling individual values as random variables, their probability distribution enables to evaluate if privacy is compromised
- In order to model individual values as random variables we assume that
  - the summarized data are known (for each bucket, its sum and its volume is published)
  - all the values are nonnegative real numbers
  - there is no correlation among values inside different buckets
  - no additional information is known

Probabilistic Framework

- Since
  - each individual value belongs to exactly one bucket (β), and
  - values in buckets are not correlated,

  the probability distribution of each value depends only on the sum (*s*) and volume (*b*) of the bucket β containing it
- $\tilde{q}_{s,b}$ will denote the random variable representing an individual value inside β
- The sample space of $\tilde{q}_{s,b}$ is [*0..s*], since values in the bucket are assumed to be nonnegative and their sum is *s*

Probabilistic Framework

- If s>0, b>1 and 0≤x≤s (the other cases are straightforward)

$$Pr(\tilde{q}_{s,b} < x) = F(x) = 1 - \left(1 - \frac{x}{s}\right)^{b-1}$$

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

- By means of the cumulative probability distribution it is simple to compute the probability that an individual value is within a range [a, b] (i.e., by computing F(b)-F(a))

Probabilistic framework

- We are interested in the ε-confidence-interval [(1-ε)·E, (1+ε)·E]

$$Pr\left((1-\epsilon)\cdot\frac{s}{b} < \tilde{q}_{s,b} < (1+\epsilon)\cdot\frac{s}{b}\right) =$$

$$= F\left((1+\epsilon)\cdot\frac{s}{b}\right) - F\left((1-\epsilon)\cdot\frac{s}{b}\right) =$$

$$= \left(1 - \frac{1-\epsilon}{b}\right)^{b-1} - \left(1 - \frac{1+\epsilon}{b}\right)^{b-1}$$

Probabilistic framework

- A bucket with sum s (s>0) and volume b (b>1) is privacy-preserving w.r.t. a privacy constraint $\langle \varepsilon, P \rangle$ if

$$\left(1-\frac{1-\epsilon}{b}\right)^{b-1}-\left(1-\frac{1+\epsilon}{b}\right)^{b-1}<P$$

- The condition does not depend on the bucket sum!
- It is possible to compute the value b^* for a pair $\langle \varepsilon, P \rangle$ such that buckets are privacy-preserving iff $b \geq b^*$ (and $s>0$)

Outline

- Multi-dimensional data summarization
- Histograms and sensitive information disclosure
- A probabilistic framework for evaluating privacy preservation of histograms
- Construction of privacy-preserving histograms
- Conclusions and future works

Construction of privacy-preserving histograms

- Classical histogram-construction techniques progressively refine the partition of multi-dimensional domain according to some heuristic
- The partitioning ends when there is no more available space for storing more buckets
- We focus our attention on constructing a ‹ε, P›-privacy-preserving histogram minimizing the error on a query workload W

$$SSE(W) = \sum_{w \in W} \left(ex(w) - ap(w)\right)^2$$

A greedy strategy

1. Init a set S of (refinable) buckets with a bucket summarizing the whole data set
2. Extract a bucket β from S
3. Choose the *best safe split* of β into $\langle \beta', \beta'' \rangle$
4. If $\langle \beta', \beta'' \rangle$ exists add β' and β'' to S else mark β as *final*
5. If S is not empty go to 2
6. Return the set of final buckets

A greedy strategy

- The *best safe split* is a split yielding two privacy-preserving buckets which maximize the SSE(W) reduction
- If a bucket β admits no safe splits (i.e., every split yields at least one non-privacy-preserving bucket) then it is marked as *final*
- It easy to show that any split sequence of a non-privacy-preserving bucket yields privacy non-preserving-buckets, thus it can be correctly marked as final

Algorithm complexity

- The complexity of the algorithm is $O(N^2 \cdot |W| \cdot t \cdot d)$
  - N is the number of points in the data set
  - $O(t^d)$ is the size of the data set domain
  - |W| is the number of queries in the workload
- The upper bound is actually "quite large"
  - O(N) iterations (at most N/b^*)
  - For each iteration, $O(t \cdot d)$ splits are tried (much less as the number of iterations increases)
  - For each possible split, |W| queries are evaluated (much less if only the queries of W overlapping the bucket to be split are considered)
  - Each query evaluation has cost O(N): more precisely O(i) at the i-th iteration

Outline

- Multi-dimensional data summarization
- Histograms and sensitive information disclosure
- A probabilistic framework for evaluating privacy preservation of histograms
- Construction of privacy-preserving histograms
- Conclusions and future works

Conclusions

- The problem of sensitive information disclosure caused by publication of summarized multi-dimensional data was studied
- Privacy on values associated with individuals was considered, differently from existing work on privacy preserving histograms which focuses on anonymity of unlabelled points
- A framework for evaluating the risk for privacy of individual values was introduced
- A greedy algorithm for constructing privacy-preserving histograms was proposed

Future works

- Considering other kinds of privacy constraints
  - Preventing small/large values to be inferred
  - Considering “absolute” confidence intervals and mixed (relative/absolute) confidence intervals
- Considering the possibility that further information about values inside buckets is known (e.g., count of non-null values, max, min)
- Considering the possibility that different histograms summarizing the same data set are published

Thanks for your attention!

Questions?

Probabilistic Framework

- The probability that q=x is given by the ratio between
  - the number of ok-configurations

  and
  - the number of all the possible configurations

  (each configuration is equiprobable), that is
  - the number of configurations of (b-1) values such that their sum is s-x

  and
  - the numer of configurations of b values such that their sum is

Probabilistic Framework

- If values are cardinals, there are

$$\binom{s+b-1}{s}$$

ways to distribute sum *s* among *b* cells (s cells must be chosen, enabling repetitions, among b)

Probabilistic Framework

- If values are cardinals, the probability that a single value is v in [0..s] is given by

$$Prob(\tilde{q}_{s,b} = x) = \frac{\begin{pmatrix} s - x + b - 2 \\ s - x \end{pmatrix}}{\begin{pmatrix} s + b - 1 \\ s \end{pmatrix}}$$

Probabilistic Framework

- In the case of real values, the probability that q=x is 0 (the sample space [0,s] contains infinite values)
- The cumulative probability distribution in the continuous can be obtained by the discrete one, in which s/γ objects each of value γ are considered to be distributed among b cells, and computing the limit for $\gamma \rightarrow 0$