Space-Efficient Private Estimation of Quantiles

M.Cafaro, A. Coluccia, I. Epicoco and M. Pulimeno

Prof. Massimo Cafaro, Ph.D.

Dept. of Engineering for Innovation
University of Salento

Outline

- The problem
- Differential Privacy (DP)
- The FRUGAL-1U algorithm
- Global sensitivity of FRUGAL-1U
- Analisys of FRUGAL-1U-L, FRUGAL-1U-G, FRUGAL-1U-ρ
- Experimental results
- (α, β)-accuracy

The problem

- Online estimation of robust statistics, namely quantiles, is of great interest in several applications where high-rate data streams must be processed as quickly as possible and discarded, being their storage usually unfeasible
- Fast and accurate estimation is challenging when considering the additional constraint of Differential Privacy (DP), which leads to the well-known privacy-utility trade-off
- Recent approaches further require the use of a minimal amount of space (even a single memory variable), so as to reduce the complexity

Our solution

- We present three differentially-private streaming algorithms for frugal estimation of a quantile, based on different modifications of the FRUGAL-1U algorithm:
  - DP-FRUGAL-1U-L
  - DP-FRUGAL-1U-G
  - DP-FRUGAL-1U-ρ
- We provide a theoretical analysis and experimental results

Preliminary definitions

Definition

(Lower and upper q-quantile) Given a set A of size n over $\mathbb{R}$, let $R(x)$ be the rank of the element x, i.e., the number of elements in A smaller than or equal to x. Then, the lower (respectively upper) q-quantile item $x_q \in A$ is the item x whose rank $R(x)$ in the sorted set A is $\lfloor 1 + q(n-1) \rfloor$ (respectively $\lceil 1 + q(n-1) \rceil$) for $0 \leq q \leq 1$.

Definition

(Rank accuracy) For all items v and a given tolerance ϵ, return an estimated rank $\tilde{R}$ such that $|\tilde{R}(v) - R(v)| \leq \epsilon n$.

Differential Privacy

Definition

(Unbounded differential privacy, also known as the add-remove model [1] [2]) Two datasets x and x' are considered neighbors if x' can be obtained from x by adding or removing one row. Under unbounded DP, the sizes of x and x' are different (by one row): $|x \backslash x'| + |x' \backslash x| = 1$.

Definition

(Bounded differential privacy, also known as the swap or the update/replace model [1] [5]) Two datasets x and x' are considered neighbors if x' can be obtained from x by changing one row. Under bounded DP, the sizes of x and x' are equal: $|x \backslash x'| = 1$ and $|x' \backslash x| = 1$.

In the sequel, we adopt bounded DP

ϵ-DP and (ϵ, δ)-DP

Definition

(ϵ-differential privacy) A function which satisfies DP is called a mechanism; we say that a mechanism $\mathcal{F}$ satisfies pure DP if for all neighboring datasets x and x' and all possible sets of outputs $\mathcal{S}$, it holds that $\frac{\Pr[\mathcal{F}(x)\in\mathcal{S}]}{\Pr[\mathcal{F}(x')\in\mathcal{S}]} \leq e^{\epsilon}$. The ϵ parameter in the definition is called the privacy parameter or privacy budget.

Definition

((ϵ, δ)-differential privacy) A mechanism $\mathcal{F}$ satisfies (ϵ, δ)-DP if for all neighboring datasets x and x' and all possible sets of outputs $\mathcal{S}$, it holds that $\Pr[\mathcal{F}(x) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{F}(x') \in \mathcal{S}] + \delta$, where the privacy parameter $0 \leq \delta < 1$ represents a failure probability.

The FRUGAL-1U algorithm

Algorithm Frugal-1U

Require: Data stream S, quantile q, one unit of memory $\tilde{m}$
Ensure: estimated quantile value $\tilde{m}$

1: $\tilde{m} = 0$
2: **for each** $s_i \in S$ **do**
3: $\quad rand = random(0, 1)$
4: $\quad$ **if** $s_i > \tilde{m}$ and $rand > 1 - q$ **then**
5: $\quad\quad \tilde{m} = \tilde{m} + 1$
6: $\quad$ **else if** $s_i < \tilde{m}$ and $rand > q$ **then**
7: $\quad\quad \tilde{m} = \tilde{m} - 1$
8: $\quad$ **end if**
9: **end for**
10: **return** $\tilde{m}$

Introduced in [3]

Global sensitivity of FRUGAL-1U

Definition

(Global sensitivity) Given a function $f : \mathcal{D} \to \mathbb{R}$ mapping a dataset in $\mathcal{D}$ to a real number, the global sensitivity of f is $GS(f) = \max_{x,x':d(x,x')\leq 1} |f(x) - f(x')|$, where $d(x, x')$ represents the distance between two datasets x, x'.

Our DP versions of the algorithm are based on the definition of bounded DP, in which two datasets x and x' are considered neighbors if x' can be obtained from x by changing one row. Owing to our choice, we need to analyze the impact of changing one incoming stream item with a different one on the quantile estimate $\tilde{m}$. The following Lemma is our fundamental result to obtain DP versions of Frugal-1U

Lemma

Under bounded DP, the global sensitivity of the FRUGAL-1U *algorithm is 2.*

Pure DP: DP-FRUGAL-1U-L

Definition

(Laplace mechanism) Given a function $f : \mathcal{D} \to \mathbb{R}$ mapping a dataset in $\mathcal{D}$ to a real number, $\mathcal{F}(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right)$ satisfies ϵ-DP. $Lap(S)$ denotes sampling from the Laplace distribution with center 0 and scale S, whilst s is the sensitivity of f.

Theorem

FRUGAL-1U *can be made ϵ-DP by adding to the quantile estimate returned by the algorithm noise sampled from a Laplace distribution with center 0 and scale S as follows: $\tilde{m} = \tilde{m} + Lap(\frac{2}{\epsilon})$.*

Approximate DP: DP-FRUGAL-1U-G

Definition

(Gaussian mechanism) Given a function $f : \mathcal{D} \to \mathbb{R}$ mapping a dataset in $\mathcal{D}$ to a real number, $\mathcal{F}(x) = f(x) + \mathcal{N}\left(\sigma^2\right)$ satisfies (ϵ, δ)-DP, where $\sigma^2 = \frac{2s^2 \ln(1.25/\delta)}{\epsilon^2}$ and s is the sensitivity of f. $\mathcal{N}\left(\sigma^2\right)$ denotes sampling from the Gaussian (normal) distribution with center 0 and variance σ^2.

Theorem

FRUGAL-1U *can be made* (ϵ, δ)*-DP by adding to the quantile estimate returned by the algorithm noise sampled from a Gaussian distribution as follows:* $\mathcal{F}(x) = f(x) + \mathcal{N}\left(\sigma^2\right)$ *where* $\sigma^2 = \frac{8 \ln(1.25/\delta)}{\epsilon^2}$.

ρ zero-concentrated differential privacy (ρ-zCDP): DP-FRUGAL-1U-ρ

Definition

(ρ-zCDP) A mechanism $\mathcal{F}$ satisfies zero-concentrated DP if for all neighboring datasets x and x' and all $\alpha \in (1, \infty)$, it holds that $D_\alpha\left(\mathcal{F}(x) \| \mathcal{F}\left(x'\right)\right) \leq \rho\alpha$, where $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \ln E_{x\sim Q}\left(\frac{P(x)}{Q(x)}\right)^\alpha$ is the Rényi divergence.

Theorem

FRUGAL-1U *can be made ρ-zCDF by adding to the quantile estimate returned by the algorithm noise sampled from a Gaussian distribution as follows:* $\mathcal{F}(x) = f(x) + \mathcal{N}\left(\sigma^2\right)$ *where* $\sigma^2 = \frac{2}{\rho}$.

Experimental results

Table: Synthetic datasets

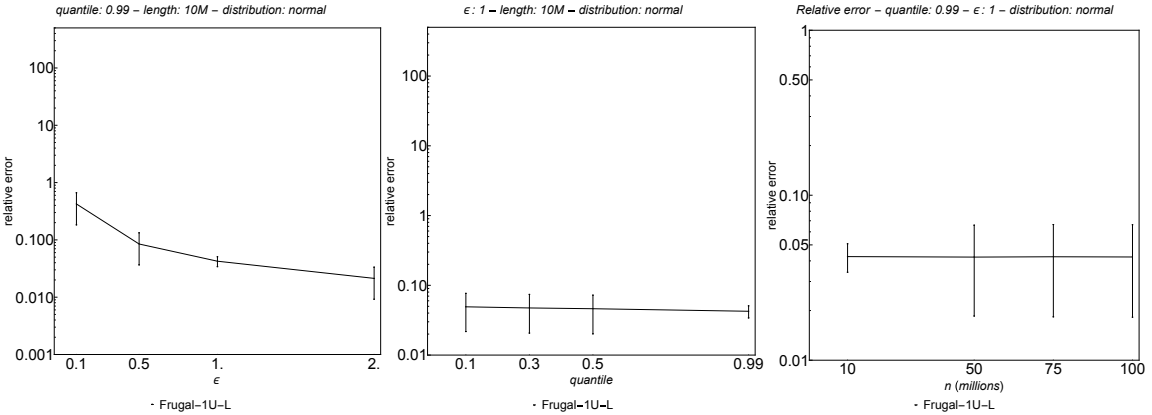
Dataset	Distribution	Parameters	PDF
D1	Uniform	$[0, 1000]$	
D2	Chi square	$\alpha = 5$	
D3	Exponential	$\alpha = 0.5$	
D4	Lognormal	$\alpha = 1, \beta = 1.5$	
D5	Normal	$\mu = 50, \sigma = 2$	
D6	Cauchy	$\alpha = 10000, \beta = 1250$	
D7	Extreme Value	$\alpha = 20, \beta = 2$	
D8	Gamma	$a = 2, b = 4$	

Experimental results

Table: Default settings of the parameters

Parameter	**Values**	**Default**
quantile	$\{0.1, 0.3, 0.5, 0.99\}$	0.99
stream length	$\{10M, 50M, 75M, 100M\}$	$10M$
ϵ	$\{0.1, 0.5, 1, 2\}$	1
δ	$\{0.01, 0.04, 0.08, 0.1\}$	0.04
ρ	$\{0.1, 0.5, 1, 5\}$	1

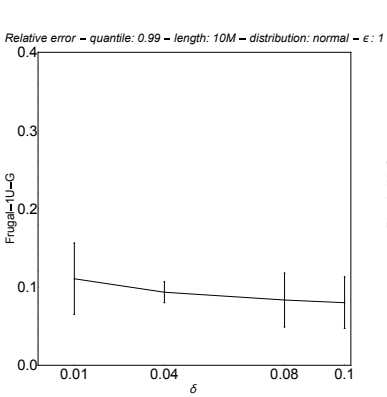
FRUGAL-1U-L: relative error



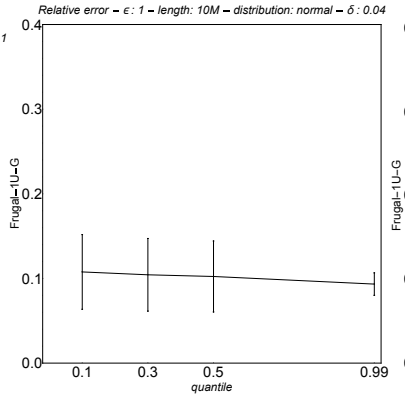
(a) privacy budget (b) quantile (c) n

Figure: FRUGAL-1U-L. Relative error varying the privacy budget ϵ, the quantile q and the stream size n

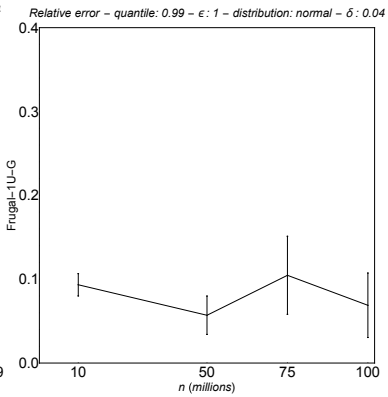
FRUGAL-1U-G: relative error



(a) δ



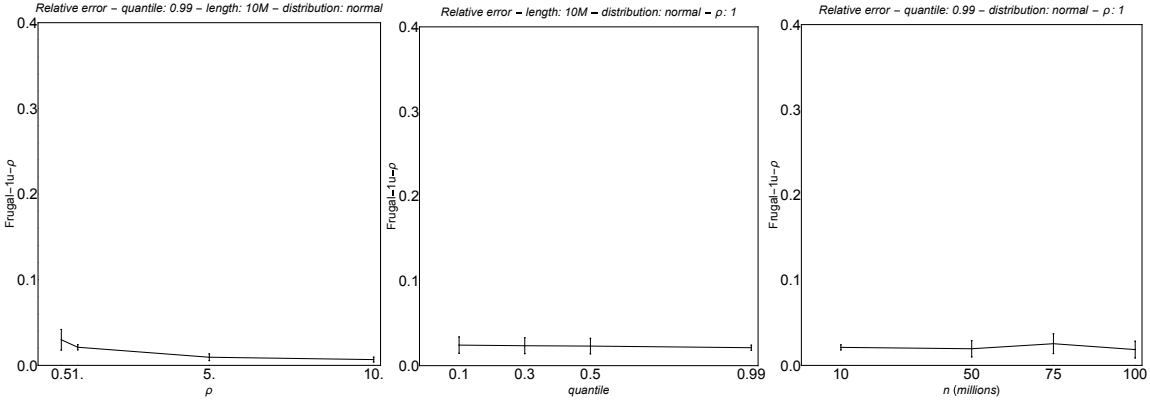
(b) Quantile



(c) n

Figure: FRUGAL-1U-G. Relative error varying the probability δ, the quantile q and the stream size n

Frugal-1U-ρ: relative error



(a) ρ (b) Quantile (c) n

Figure: FRUGAL-1U-ρ. Relative error varying the parameter ρ, the quantile q and the stream size n

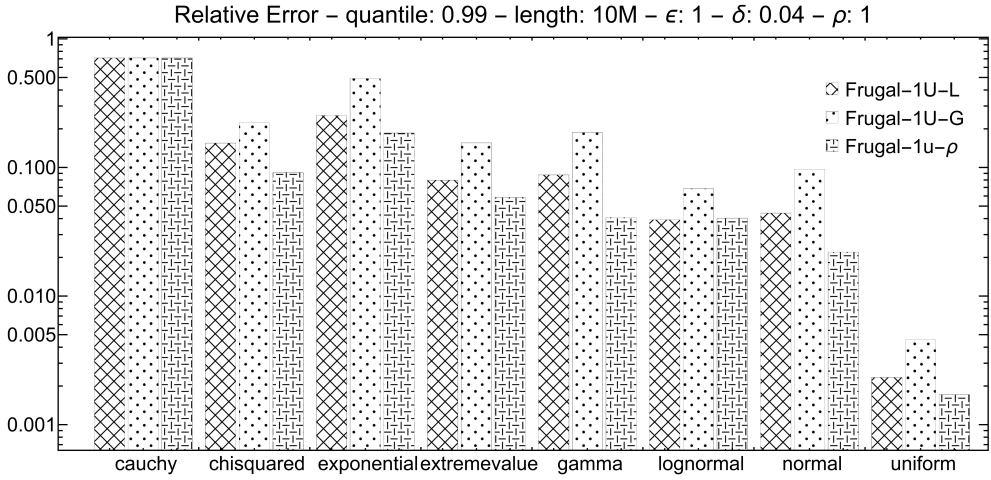


Figure: FRUGAL-1U-L: relative error varying the distributions

(α, β)-accuracy

- Utility quantifies how much a DP result is useful for a subsequent data analysis. Therefore, the analysis to be performed plays a key role here, since DP results affected by a significant error may or may not be useful to the analyst
- One way to overcame the dependence from the analysis is the use of the related concept of accuracy, which is the distance between the true value computed without DP and the DP released value
- Therefore, accuracy is often used in place of utility, because more accurate results are generally more useful for an analysis. The so-called (α, β)-accuracy framework [4] can be used to measure accuracy
- Here, α represents an upper bound on the absolute error committed, whilst β is the probability to violate this bound

Definition

((α, β)-accuracy) Given a function $f : \mathcal{D} \to \mathbb{R}$ mapping a dataset $x \in \mathcal{D}$ to a real number, and a DP mechanism $\mathcal{M}_f : \mathcal{D} \to \mathbb{R}$, $\mathcal{M}_f$ is (α, β)-accurate if $\Pr\left[\left\|f(x) - \mathcal{M}_{f(x)}\right\|_\infty \geq \alpha\right] \leq \beta$.

(α, β)-accuracy of the proposed algorithms

- FRUGAL-1U-L is (6.4, 0.04)-accurate
- FRUGAL-1U-G is (9.1, 0.04)-accurate
- FRUGAL-1U-ρ is (2.4, 0.04)-accurate

Conclusions

- We proposed DP algorithms for tracking quantiles in a streaming setting
- Our algorithms are DP variants of the well-known FRUGAL-1Ualgorithm, characterized by the property of requiring just a tiny amount of memory to process a stream while guaranteeing surprising accuracy for the estimates of a quantile
- for FRUGAL-1U we gave corresponding ϵ-DP, (ϵ, δ)-DP, and ρ-zCDF algorithms after proving that the global sensitivity of FRUGAL-1U is equal to 2
- Finally, we also showed that the proposed algorithms achieve good accuracy in the experimental results

References


Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith.
Calibrating noise to sensitivity in private data analysis.
In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.


Cynthia Dwork and Aaron Roth.
The algorithmic foundations of differential privacy.
Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014.


Qiang Ma, S. Muthukrishnan, and Mark Sandler.
Frugal Streaming for Estimating Quantiles, pages 77–96.
Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.


Joseph P. Near and Xi He.
Differential privacy for databases.
Foundations and Trends® in Databases, 11(2):109–225, 2021.


Salil Vadhan.
The Complexity of Differential Privacy, pages 347–450.
Springer International Publishing, Cham, 2017.