Better Differentially Private Approximate Histograms and Heavy Hitters using the Misra-Gries Sketch

Christian Janos Lebeda*† and Jakub Tětek*‡

†IT University of Copenhagen, ‡University of Copenhagen
*Basic Algorithms Research Copenhagen

IT UNIVERSITY OF COPENHAGEN

Problem formulation

Given a stream $S = \{S_1, \ldots, S_n\}$ of n elements $S_i \in \{1, \ldots, d\}$, design an (ε, δ)-differentially private mechanism $\mathscr{M}$ to approximate the histogram $c(S)$ where:

$$c(S)_x = |\{i : S_i = x\}|$$

Memory constraint

We cannot store $c(S)$. We can use only $O(k)$ memory with $k \ll n$ and $k \ll d$.

Error measure

$$\|\mathscr{M}(S) - c(S)\|_\infty := \max_{x \in [d]} |\mathscr{M}(S)_x - c(S)_x|$$

Misra-Gries sketch

Store k counters.

Perform one of three simple updates for each element.

Algorithm: Misra-Gries Sketch (MG)
Input: Positive integer k and stream $\mathcal{S}$
$T \leftarrow \emptyset$ (Set of non-zero counters)
foreach $x \in \mathcal{S}$ **do**
  if $x \in T$ **then**
    $c_x \leftarrow c_x + 1$
  else if $|T| < k$ **then**
    $T \leftarrow T \cup \{x\}$
    $c_x \leftarrow 1$
  else
    $c_y \leftarrow c_y - 1$ for all $y \in T$
    $T \leftarrow T \setminus \{y \in T : c_y = 0\}$
  end
end
return T, c

Optimal error for worst-case input:

$$\|\mathrm{MG}(S) - c(S)\|_\infty \leq \frac{n}{k+1}$$

Technique

Step 1:

- Compute $T, c \leftarrow \mathrm{MG}(S)$.
- Adds error at most $n/(k+1)$.

Step 2:

- Add noise from Laplace$(1/\varepsilon)$ independently to the counter for each element in T.
- Adds error $O(\log(k)/\varepsilon)$.

Step 3:

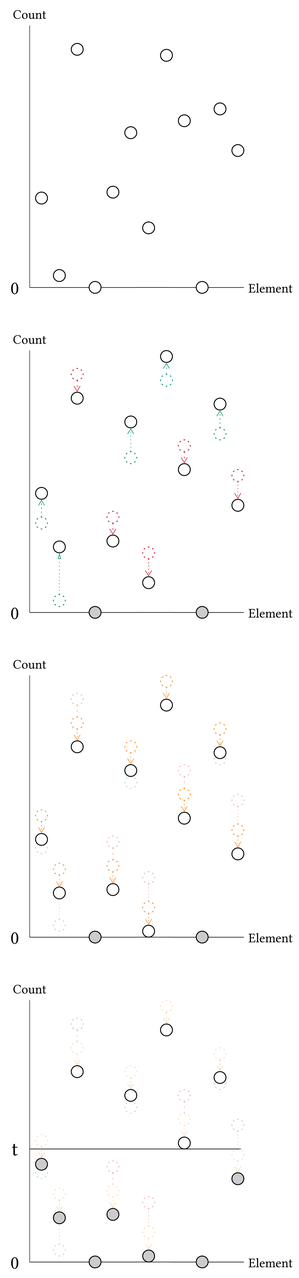
- Sample $\eta \sim$ Laplace$(1/\varepsilon)$ and add η to every element in T.
- Adds error $O(1/\varepsilon)$.

Step 4:

- Remove noisy counts below $t = 1 + 2\ln(3/\delta)/\varepsilon$.
- Adds error at most $O(\log(1/\delta)/\varepsilon)$.

Main result

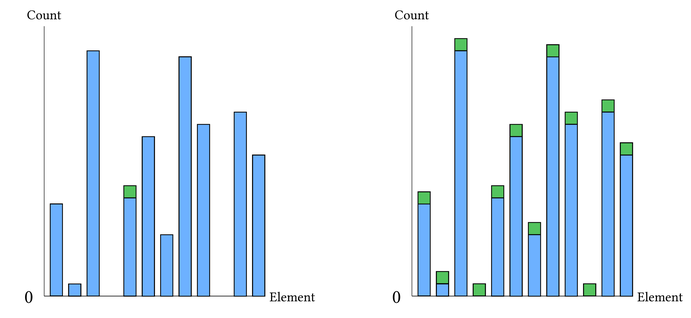
$$E[\|\mathscr{M}(S) - c(S)\|_\infty] \leq \frac{n}{k+1} + O(\log(1/\delta)/\varepsilon)$$



Sensitivity

We utilise the structure of the Misra-Gries sketch.

Between neighboring streams the sketch differs either in 1 or in all k counters.



Case 1 Case 2

Potential Future Work

- The noise scales with m if an individual can submit m distinct elements. Can this be improved to $\sqrt{m}$?

Looking for a Postdoc?

I am searching for a job starting 2024.

References

[MG82] Misra and Gries. Finding repeated elements.
[CLSX12] Chan, Li, Shi, and Xu. Differentially Private Continual Monitoring of Heavy Hitters from Distributed Streams.

Research supported by the VILLUM foundation grant number 16582.

chle@itu.dk