Differentially Private Hierarchical Count-of-Counts Histograms

Yu-Hsuan Kuo, Cho-Chun Chiu, Daniel Kifer†, Michael Hay‡, Ashwin Machanavajjhala⋆

Penn State University, † Penn State University and U.S. Census Bureau, ‡ Colgate University, ⋆ Duke University

Introduction

Consider the table **Persons**(person_name, group_id, location) and a hierarchy Γ on location associated with each group. A hierarchical count-of-counts histogram queries on this table: for each geographic region (e.g. the United States/New York), how many groups (e.g. households) in that region have j people (i.e. of size j).

Table 1: **Persons**

name	g_id	loc.
Alice	1	a
Bob	1	a
Carol	1	a
Dave	1	a
Eve	2	b
Frank	2	b
Judy	3	a
Nick	4	b

- A=SELECT groupid, COUNT(*) AS size FROM Persons GROUPBY groupid

Table 2: A

g_id	size	loc.
1	4	a
2	2	b
3	1	a
4	1	b

The count-of-counts histograms can be obtained by `H=SELECT size, COUNT(*) FROM A GROUPBY size`

- **count-of-counts histogram (coco)** H is
  $H^{\text{root}} = [2, 1, 0, 1]$
  $H^{a} = [1, 0, 0, 1]$
- **unattributed histogram** [1] H_g is
  $H_g^{\text{root}} = [1, 1, 2, 4]$
  $H_g^{a} = [1, 4]$
- **cumulative count-of-counts histogram** H_c
  $H_c^{\text{root}} = [2, 3, 3, 4]$
  $H_c^{a} = [1, 1, 1, 2]$

To protect privacy, the ϵ-differential privacy is applied at the person level. We used the geometric mechanism.

Definition (Sensitivity)

Given a query q (which outputs a vector), the global sensitivity of q, denoted by $\Delta(q)$ is defined as:

$$\Delta(q) = \max_{D_1, D_2} ||q(D_1) - q(D_2)||_1,$$

where databases D_1, D_2 contain the public Hierarchy and Groups tables, and differ by the presence or absence of one record in the Persons table.

Definition (Geometric Mechanism)

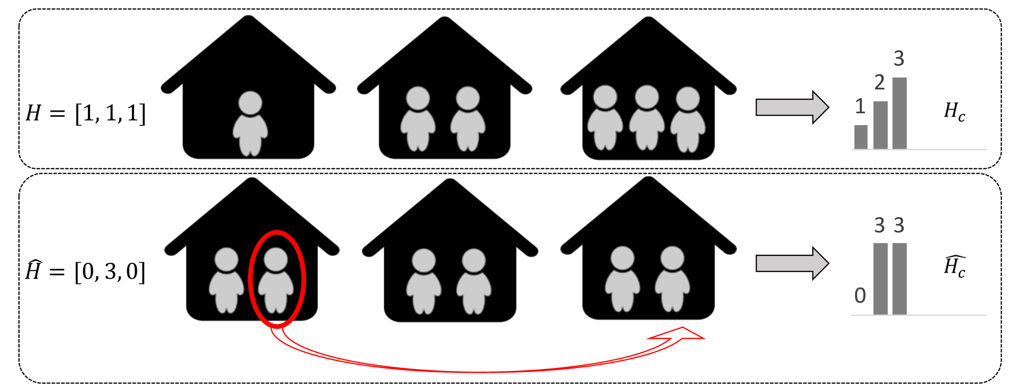
[2] Given a database D, a query q that outputs a vector, a privacy loss budget ϵ, the global sensitivity $\Delta(q)$, the geometric mechanism adds independent noise to each component of $q(D)$ using distribution: $P(X = k) = \frac{1-e^{-\epsilon}}{1+e^{-\epsilon}} e^{-\epsilon|k|/\Delta(q)}$ (for $k = 0, \pm 1, \pm 2$, etc.). This distribution is known as the double-geometric with scale $\Delta(q)/\epsilon$.

Problem Definition

For each node τ in hierarchy, create differentially private estimate $\tau.\widehat{H}$ of coco histogram H such that

- The entries are nonnegative integers
- The counts are accurate ($\tau.\widehat{H}$ and $\tau.H$ are close)
- $\tau.\widehat{H}$ matches publicly known total # of groups in τ
- Consistency: children histograms sum up to the parent.

Error Measure

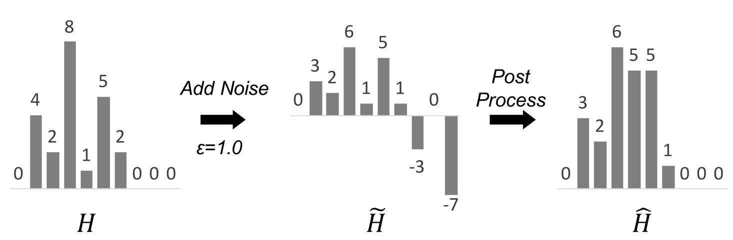


emd = $|H_c - \widehat{H}_c|_1 = |H_g - \widehat{H}_g|_1 = 2$

Non-hierarchical Count-of-counts Histograms Publishing

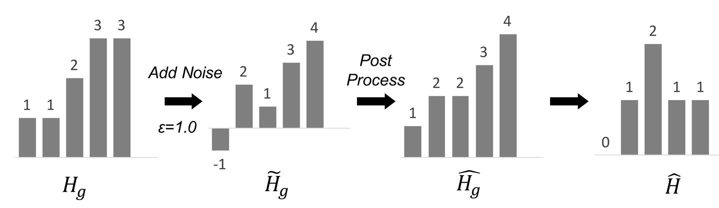
Naive Strategy

$$\widehat{H} = \arg\min_{\widehat{H}} ||\widetilde{H} - \widehat{H}||_p^p$$
$$\text{s.t. } \widehat{H}[i] \geq 0 \text{ for all i}$$
$$\text{and } \sum_i \widehat{H}[i] = G$$



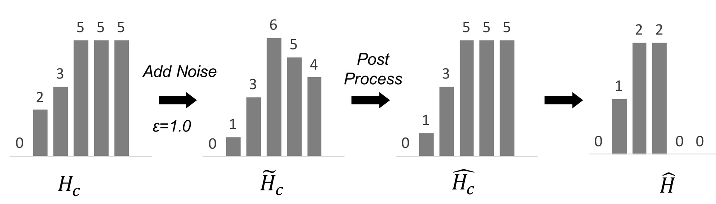
Unattributed Histogram [1] $\mathbf{H_g}$

$$\widehat{H}_g = \arg\min_{\widehat{H}_g} ||\widetilde{H}_g - \widehat{H}_g||_p^p$$
$$\text{s.t. } 0 \leq \widehat{H}_g[i] \leq \widehat{H}_g[i+1]$$
$$\text{for } i = 0, \ldots, G-1$$



Cumulative Sum Histograms $\mathbf{H_c}$

$$\widehat{H}_c = \arg\min_{\widehat{H}_c} ||\widetilde{H}_c - \widehat{H}_c||_p^p$$
$$\text{s.t. } 0 \leq \widehat{H}_c[i] \leq \widehat{H}_c[i+1]$$
$$\text{for } i = 0, \ldots, K \text{ and } \widehat{H}[K] = G$$



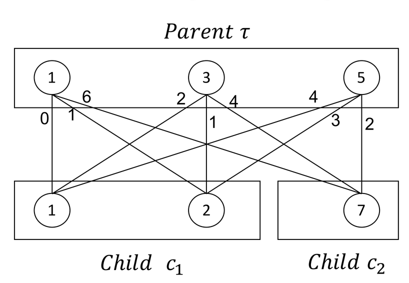
Solver: min-max algorithm [3], pool-adjacent violators (PAV), Gurobi

Hierarchical Count-of-counts Histograms Publishing

- Our proposed solution:
  1. Estimated coco $\tau.\widehat{H} \Rightarrow$ the unattributed histogram $\tau.\widehat{H}_g$
  2. Find a 1-to-1 optimal matching between groups at child nodes and groups at parent
  3. Merge those two estimates

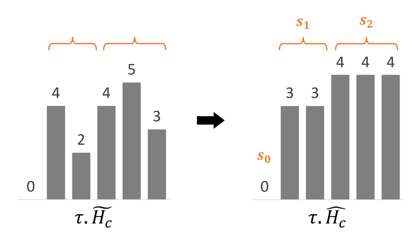
Optimal Matching

- For each node τ and its children, set up a bipartite weighted graph



- **Least cost weighted matching** on this bipartite graph.
- Optimal algorithm: match the smallest unmatched group in τ to the smallest unmatched group among any of its children.

Initial Variance Estimation



- Let ϵ be the privacy budget used in node τ in level ℓ of Γ
- Variance estimate for the i^{th} largest group $\tau.V_g[i]$

$$\begin{cases} \frac{2}{|S_i|\epsilon^2} & \text{if } \mathbf{H_g} \text{ method} \\ 4/(\epsilon^2 \times \text{number of estimated groups of size } \tau.\widehat{H}_g[i] & \text{if } \mathbf{H_c} \text{ method} \end{cases}$$

Merge Estimates

Given size estimates: $\tau.\widehat{H}_g[i]$, $c.\widehat{H}_g[j]$ & variance estimates $\tau.V_g[i]$, $c.V_g[j]$.

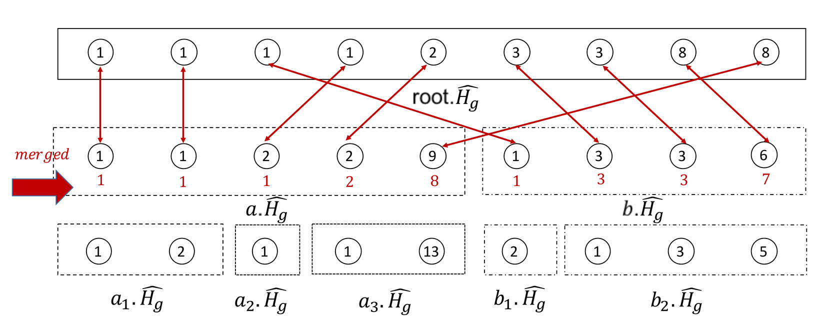
- Optimal linear combination of the estimates [1]: weighted average

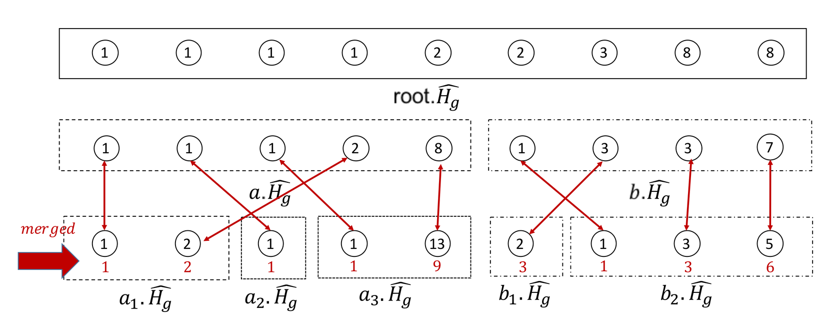
$$\left(\frac{\tau.\widehat{H}_g[i]}{\tau.V_g[i]} + \frac{c.\widehat{H}_g[j]}{c.V_g[j]}\right) \Big/ \left(\frac{1}{\tau.V_g[i]} + \frac{1}{c.V_g[j]}\right)$$

and the variance of this estimator is

$$\left(\frac{1}{\tau.V_g[i]} + \frac{1}{c.V_g[j]}\right)^{-1}$$

Top-down Consistency





Results

Table 3: Average error with $\epsilon = 1.0$ at top level

Method	Synthetic	White	Hawaiian	Taxi
Naive	4,462,728,374	4,809,679,734	4,027,891,692	208,977,518
$\mathbf{H_c}$	3,742.0	1,838.9	254.0	2,819.8
$\mathbf{H_g}$	2,219.6	6,115.3	516.2	11,227.6

1. Naive strategy is usually worse and not used in the hierarchical estimates.
2. Data dependent performance: $\mathbf{H_c}$ performs better in dense region while $\mathbf{H_g}$ performs better in sparse region

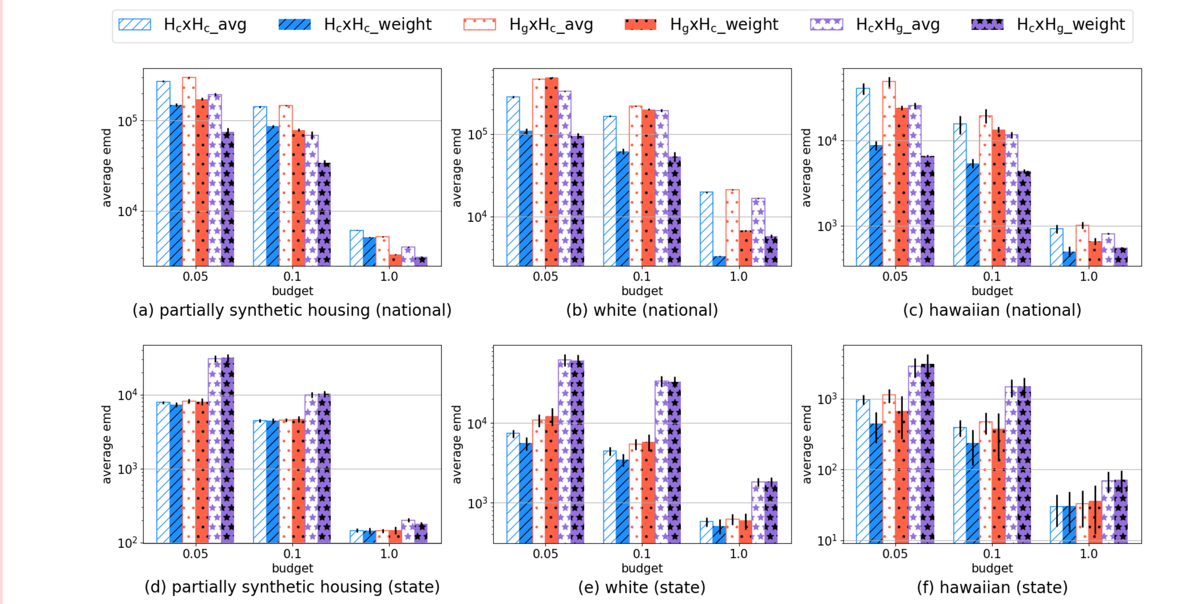


Figure 1: Merging estimates using weighted average vs. normal average. x-axis: privacy budget per level.

3. Weighted average method consistently produces large reductions in error at the top level

Table 4: Comparison to Bottom-up Aggregation

	Part. Synth.	White	Hawaiian	Taxi
	Level 0			
BU	78,459.0	448,909.0	13,968.0	20,731.0
H_c	**32,480.0**	**17,000.0**	**1,381.0**	**10,547.0**
	Level 1			
BU	1,512.2	8,722.0	270.1	10,405.5
H_c	**1,000.3**	**1,511.8**	**117.7**	**5,431.5**
	Level 2			
BU	**24.9**	**152.3**	**4.3**	**772.8**
H_c	80.1	363.8	21.6	1,601.8

4. BU has very low error at the leaves but higher error everywhere else

References

[1] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *PVLDB*, 2010.

[2] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *STOC*, 2009.

[3] RE Barlow and HD Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 1972.