Differentially Private Hierarchical Heavy Hitters

Ari Biswas*

Graham Cormode*

Yaron Kanza^

Divesh Srivastava^

Zhengyi Zhou^

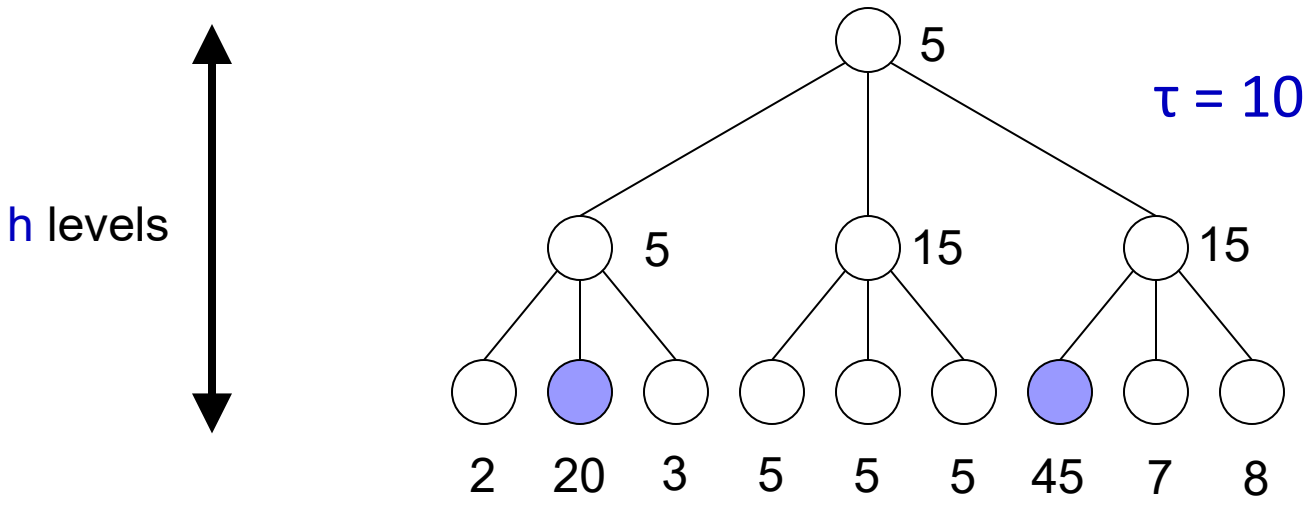
* University of Warwick ^ AT&T Chief Data Office

Private Data Analysis

- Private data analytics seeks to perform analysis on sensitive data
  - The result of the analysis must provide some useful results
  - While ensuring that they protect the privacy of the data subjects
- Much prior work has addressed core data analysis tasks
  - E.g., Histograms, clustering, classification, statistical distribution
- In this work, we study the task of hierarchical heavy hitters under the model of differential privacy
  - Variations arise for the offline and online (streaming) cases

Hierarchical Heavy Hitters (HHH)

- The hierarchical heavy hitters (HHH) identify important points in data drawn from hierarchical domains (CKMS, VLDB 2003)
- E.g., locations over street address, postal code, village, city…
- The *heavy hitters* are those points with frequency $> \tau$
- The HHHs are defined as nodes in the hierarchy that are heavy, after removing heavy contributions from lower levels



Differential Privacy (DP)

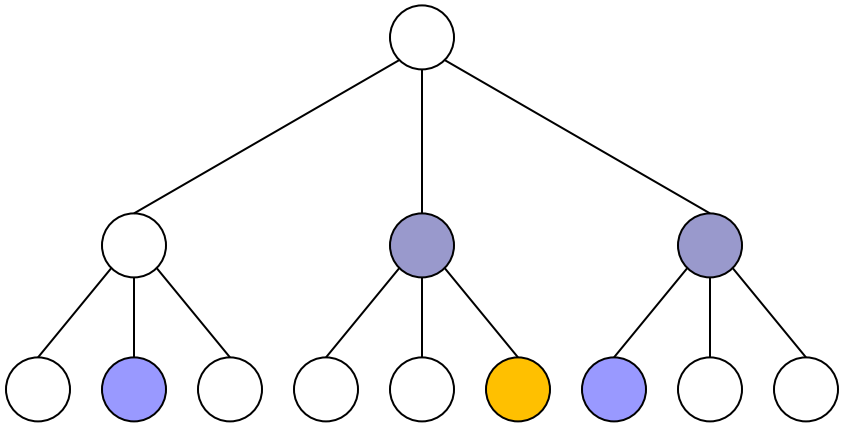
- Differential privacy is a constraint on the output distribution of a randomized algorithm A on neighboring inputs X, X'
  - Formally, ensure that $\Pr[A(X) \in O] \leq \exp(\varepsilon) \Pr[A(X') \in O] + \delta$
- Differential privacy is achieved for numeric functions by adding appropriately scaled Laplace or Gaussian noise
- (Basic) Composition: running algorithm A_1 with $(\varepsilon_1, \delta_1)$-DP then A_2 with $(\varepsilon_2, \delta_2)$-DP yields $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP
- Post-processing: any post-processing of DP output remains DP
- Sparse Vector Technique (SVT): an algorithm that outputs elements whose (noisy) count is above a threshold
  - Only pay a privacy "cost" proportional to the number of outputs

Offline DP Hierarchical Heavy Hitters

- Prior approaches can be applied to find DP-HHHs offline
  - These have costs with (explicit or implicit) dependence on h
- Laplace Histograms: materialize counts of each hierarchy level
  - Add noise proportional to h/ε to each count then find HHHs
  - The error per node scales linear to the hierarchy height, h
- DP Counting on Trees (GKKMW, ICALP 2023)
  - Proceed bottom-up, materialize node counts level-by-level
  - Only pay based on the number of heavy hitter nodes
  - We can find approximate HHHs from the noisy node counts
  - In realistic settings, the cost will still grow proportional to h

Offline DP Hierarchical Heavy Hitters

- Observation: each leaf influences at most one HHH ancestor
- Our algorithm proceeds level-by-level, bottom-up
  - Find node counts at current level without HHH descendants
  - If node count of v + Laplace noise exceeds threshold τ, output it
- Privacy proof follows by using structural properties of HHH:
  - Only leaf-to-root path matters, other nodes don't affect node v
  - Only nodes up to the first HHH ancestor of node v matter
  - Privacy proof is similar to sparse vector technique

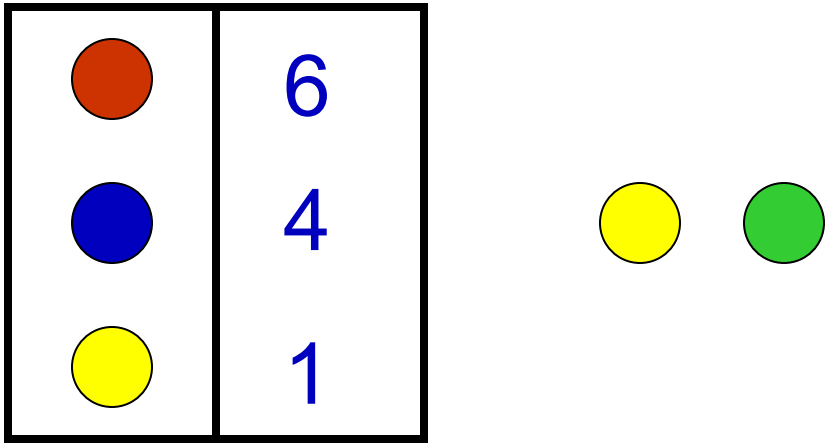


Offline DP Hierarchical Heavy Hitters

- Our DP-HHH algorithm bounds the relative error of node counts:
  - Relative error: The relative error of all HHH node counts is bounded by Δ/τ for $\Delta = O(1/\varepsilon (\log (1/\delta) + \log h))$
  - Coverage: For all nodes **not** reported by our algorithm, their HHH count is below $\tau - \Delta$
  - These bounds hold with fixed (adjustable) probability
  - Only logarithmic dependence on hierarchy height h!
- Coverage: by applying bounds on the sums of Laplace noise
- Relative error: bounds follow by summing the estimation errors in the HHHs that contribute to that node's count

Streaming DP Hierarchical Heavy Hitters

- The memory-bounded streaming HHH problem is harder
  - We see a sequence of element arrivals and must keep a sketch
  - Since the fixed-size summary is approximate, errors can amplify
- Our approach is to use a compact sketch of node frequencies
  - We use a Misra-Gries (MG) sketch for each level of the hierarchy
  - We add privacy noise to each sketch to get a noisy histogram
  - We output HHHs level-by-level based on the noisy histograms



Streaming DP Hierarchical Heavy Hitters

- Our result builds on the privacy analysis of MG (LT, PODS 2023)
  - A naïve analysis of the MG sketch suggests it has high privacy cost
  - A small change in the input can lead to a large change in the sketch
  - However, the change is correlated:
    all stored items are affected by the same amount
  - The analysis treats such changes as a single event,
    and applies the sparse vector technique
- Using the noisy MG node counts, we can extract HHH estimates
  - Conservatively reduce the count of all ancestors of an HHH node

Streaming DP Hierarchical Heavy Hitters

- Privacy by analyzing the privacy properties of each MG sketch
  - Privacy of HHH output follows by composition and post-processing
- The absolute error is a function of ε, h, δ, n, and sketch size k
  - The error of each reported HHH is $O(n/k + h/\varepsilon (\log (kh) + \log (h/\delta)))$
  - n/k is the error from (non-private) sketching
  - $h/\varepsilon \log (kh)$ is a union bound on the error from Laplace noise
  - $h/\varepsilon \log (h/\delta)$ is error from elements whose count is treated as zero

Concluding Remarks

- We can achieve accurate recovery of HHHs under DP guarantees
- Ignoring logarithmic factors in h and k, we showed:
  - In offline setting, the DP error is independent of hierarchy height h
  - In streaming setting, the DP error is independent of sketch size k
- The full paper has detailed proofs and further discussion
- Future work in this direction:
  - Handle multi-dimensional inputs without exponential blow-up
  - Use structure of problems to control the privacy noise needed