DP-tree: Indexing Multi-Dimensional Data under Differential Privacy*

Shangfu Peng, Yin Yang, Zhenjie Zhang, Xiaokui Xiao, Marianne Winslett and Yong Yu

Introduction

GOAL: count the number of records in a multi-dimensional range as accurately as possible while satisfying ε-differential privacy.

Range Count Query

A range count query returns the number of records in a multi-dimensional range.

Patient	Age	Systolic (mmHg)	Diastotic (mmHg)
Alice	45	140	95
Bob	59	120	80
Carol	52	130	90
Dave	57	135	90

select count(*) from Patients where Age≥50 and age≤60 and systolic≥120

Existing Methods

Privlet is the first index structure for answering 1D range count queries under ε-DP using Haar wavelet transform.

Universal histogram (UH) is a tree structure over a 1D domain, with a post-processing method for enforcing consistency and improving accuracy.

DP-compliant quad-tree (Quad) is a 2D structure based on the quad-tree. Quad assigns a portion of the privacy budget to each level of the tree, and computes the optimal budget assignment.

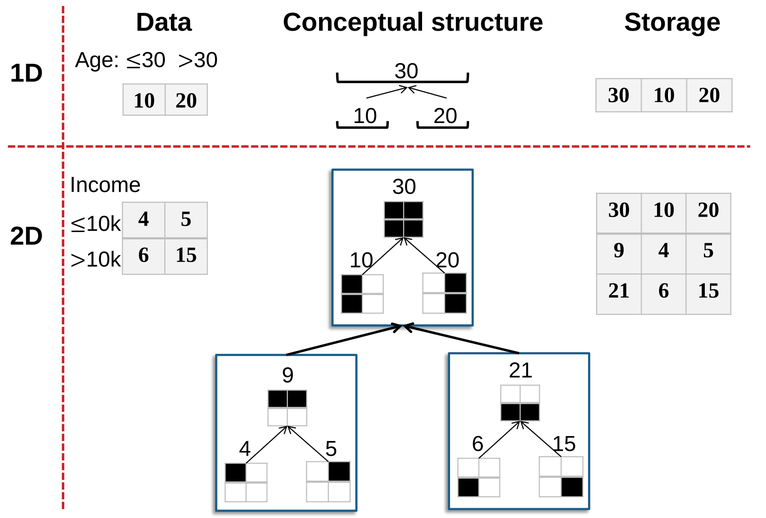
Summary of range count algorithms under ε-DP

Method	Asymptotic error bound	Practical accuracy performance
Laplace Mechanism	$O(n^d / \varepsilon^2)$	Poor
Privlet	$O(\log^{3d} n / \varepsilon^2)$	Poor
UH	$O(\log^3 n / \varepsilon^2)$ (limited to 1D data)	Good (with optimal fanout)
Quad	$O(n / \varepsilon^2)$ (limited to 2D data)	Good
DP-tree	$O(\log^{3d} n / \varepsilon^2)$	Very good

*This work is supported by TSRP Grant No. 1021580074 from Singapore's A*STAR.

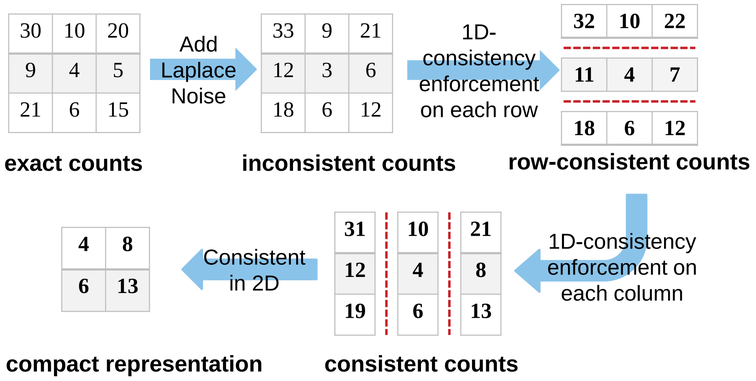
DP-tree

Nested Tree Structure



Consistency Enforcement

Example 2D consistency enforcement



Theorem. (*Multi-Dimensional Consistency Enforcement*) the optimal consistent DP-tree can be computed using the consistency enforcement method for 1D data.

Optimization

Fanout Analysis

Goal: minimize expected error for a random range query q.

Domain Size	2^6	2^{12}	2^{18}	2^{24}	2^{30}	2^{36}	2^{42}
Optimal	12	12	13	13	14	14	15

- 2 is a very poor choice for the fanout although it is used in many papers.
- In practice, one can get good accuracy with a fanout of 8 for small domains, and a fanout of16 for larger domains.

Adaptive DP-tree

Adaptive privacy budget assignment: extract from past workload statistics the visit frequency r_i for each node N_i. Then optimize the following:

$$Minimize\ E[Err(Q)] \propto \sum_{i=1}^{m} \frac{r_i}{\varepsilon_i^2}$$

$$Subject\ to:\quad \varepsilon_i > 0; \qquad \forall\, l \in leaves, \sum_{v \in ancestor(l)} \varepsilon_v = \varepsilon$$

Consistency enforcement: With node-wise budget assignment, solve the consistency enforcement problem in $O(n)$ time:

$$Minimize \sum_{v} \varepsilon_v^2 \left(\bar{h}[v] - \tilde{h}[v]\right)^2$$

Experimental Results