DIFFERENTIAL PRIVACY: PUBLISHING HISTOGRAMS

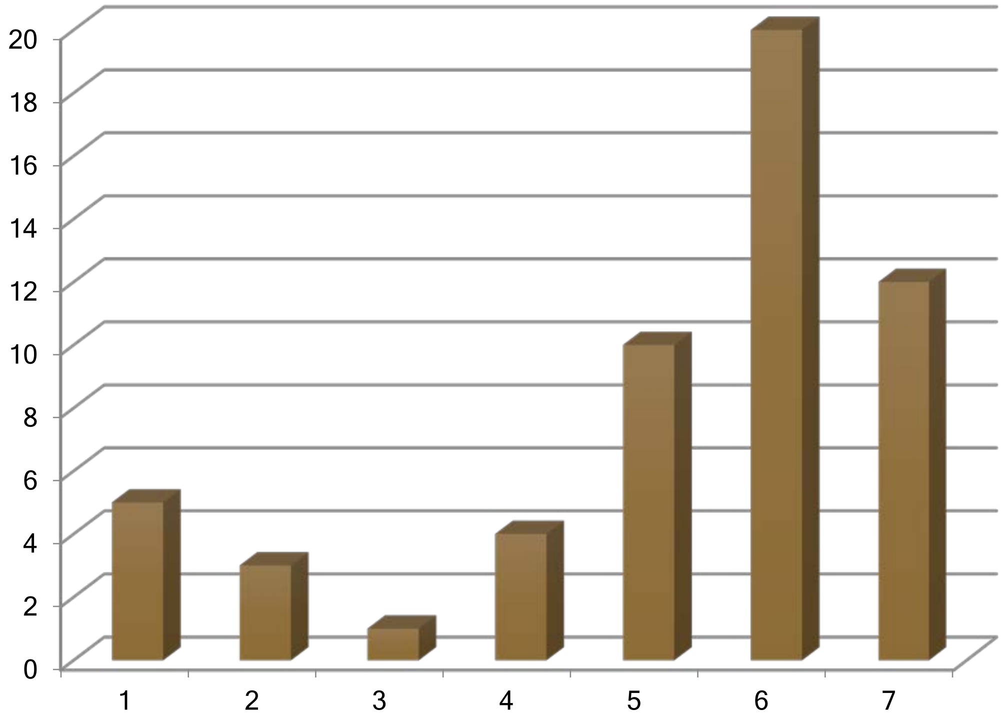
PURDUE
UNIVERSITY®

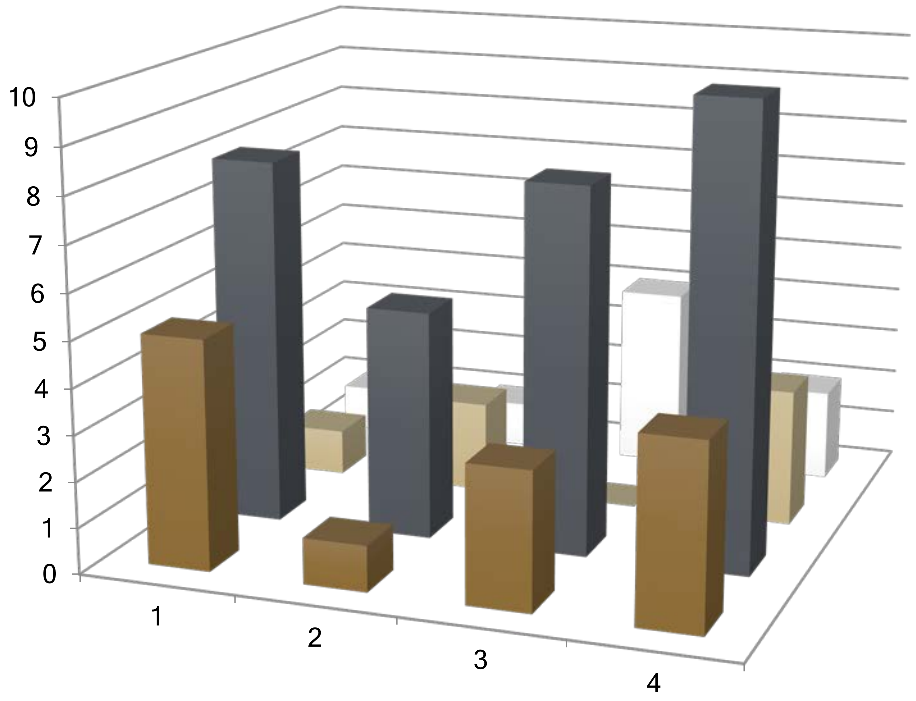
Hierarchical Methods for Histograms

- Reading
  - Wahbeh H. Qardaji, Weining Yang, Ninghui Li: Understanding Hierarchical Methods for Differentially Private Histograms. Proc. VLDB Endow. 6(14): 1954-1965 (2013)
    - M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. PVLDB, 3:1021–1032, September 2010.
    - T.-H. Hubert Chan, Elaine Shi, Dawn Song: Private and Continual Release of Statistics. ACM Trans. Inf. Syst. Secur. 14(3): 26:1-26:24 (2011)

Histogram

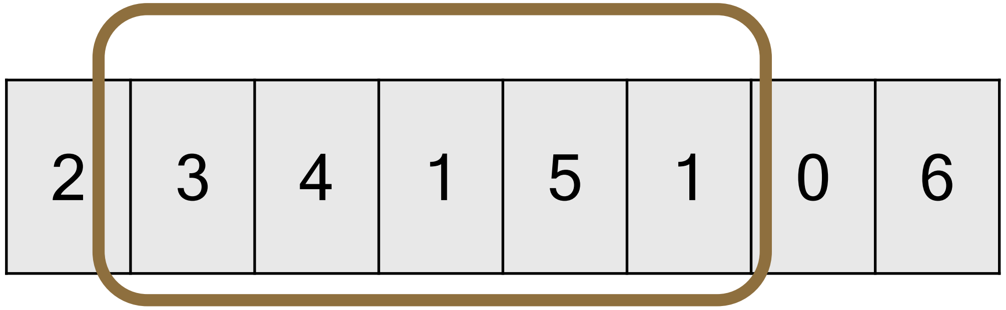
- A histogram is a graphical representation of the distribution of numerical data: a partitioning of the data domain into multiple non-overlapping bins; the number of data points in each bin

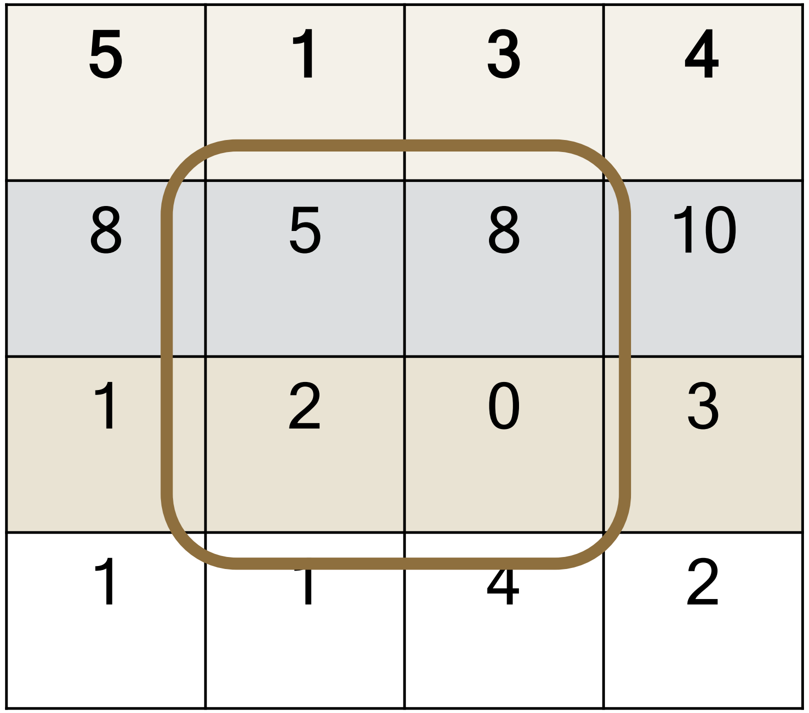




Range Query

- A range query represents a hyperrectangle in the d-dimensional domain specified by the dataset, and asks for the number of tuples that fall within the bins that are completely included in the area covered by the hyperrectangle





Histogram Settings

- Has Suitable Partitioning
  - A pre-defined partitioning
  - The average number of data points in a bin is sufficiently high
- Lacking Suitable Partitioning
  - No pre-defined partition
  - A pre-defined partitioning exists, but the average number of data points for each bin is low
  - **Determine the partitioning**

Utility Metrics (1)

- Mean Absolute Error (MAE)
  - absolute difference between the noisy answer and the true answer
- Mean Squared absolute Error (MSE)
  - often easier to compute
  - MSE is the variance of the random noise

Utility Metrics (2)

- Mean Relative Error (MRE)
  - impact of the same absolute error is different when the true answers are different
  - the true answer may be very small, or even 0
    - chooses a threshold θ to be used as the denominator

$$\text{relative error} = \frac{|\text{true anser} - \text{obtained answer}|}{\max(\theta, \text{true anser})}$$

PURDUE UNIVERSITY

Dense Pre-defined Partitioning

- The average bin count is at least $\geq \frac{5}{\varepsilon}$
- Baseline: a simple histogram
  - Add noise sampled from Lap($\frac{1}{\varepsilon}$) to each bin
    - Variance: $V_u = \frac{2}{\varepsilon^2}$

Unit Variance

  - Average length of queries:

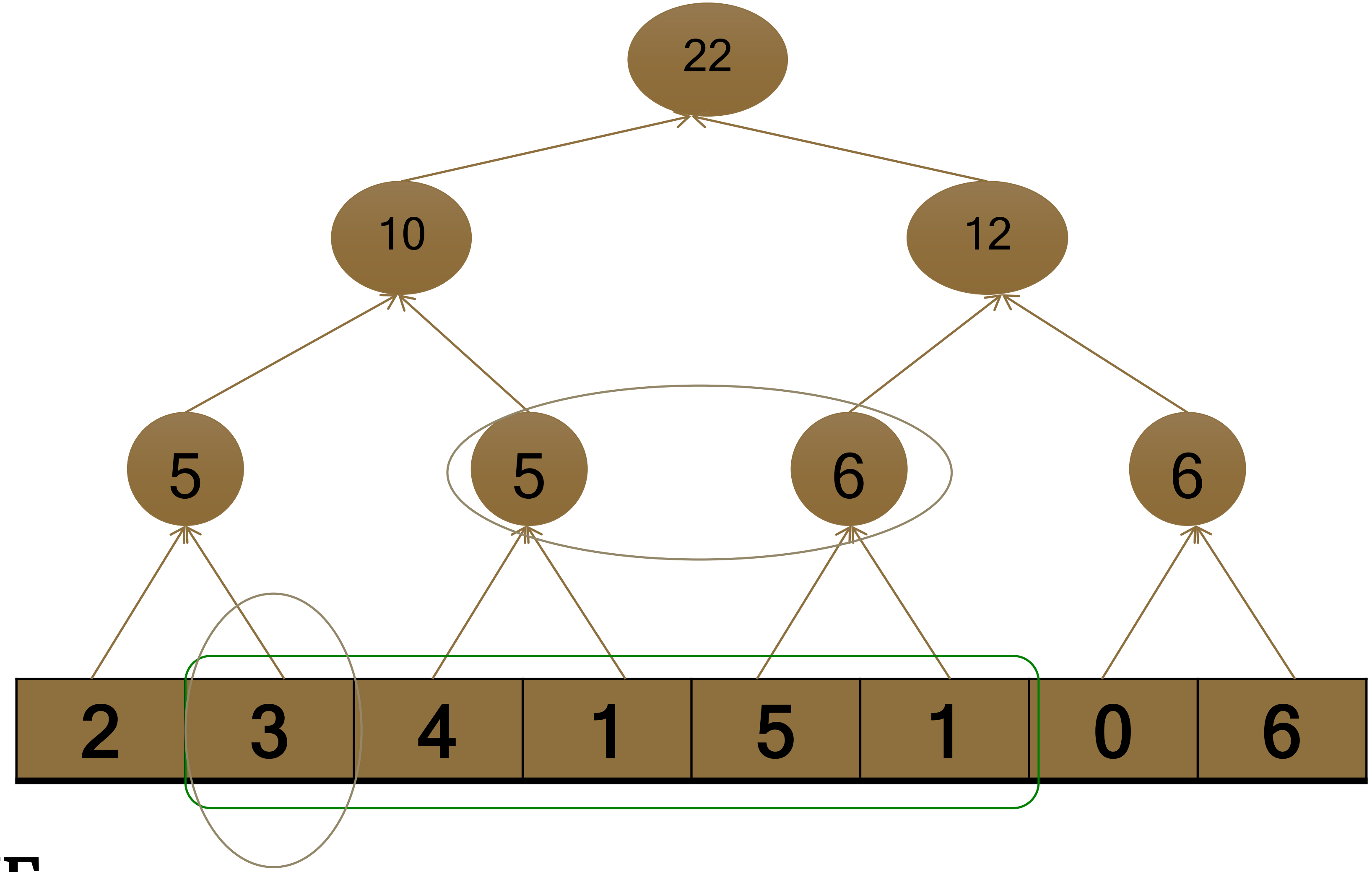
$$\frac{\sum_{j=1}^{m} j(m-j+1)}{m(m+1)/2} = \frac{(m+2)}{3}$$

  - Average-case MSE: $\frac{(m+2)}{3}$Vu

A Digression

- The hierarchy method uses a data structure called segment tree
  - Consider this problem
- We will explain what is it and why by starting with prefix sum and difference arrays
  - Let us study these slides

Hierarchical Method



Hierarchical Method

$h = \lceil \log_b N \rceil$ the tree has *h* + 1 levels

Publish h different histograms of the data

Use the least number of nodes to answer a range query, no more than *2(b−1)h*

If privacy budget is divided equally among the histogram, the noise added to each counting query has variance

$$\frac{2h^2}{\epsilon^2} = h^2 \mathsf{V_u}$$

PURDUE UNIVERSITY®

Branching Factor b

Increasing *b*

Reduce *h*, increase privacy budget to each node

More nodes to be used to answer a query on average

MSE:

Optimize *b*

$$\mathsf{V}^*_{Avg}[\mathbf{H}_b] = \left((b-1)h^3 - \frac{2(b+1)h^2}{3}\right) \cdot \mathsf{V}_u$$

$b \approx 16.8$, when m is large

Constrained Inference

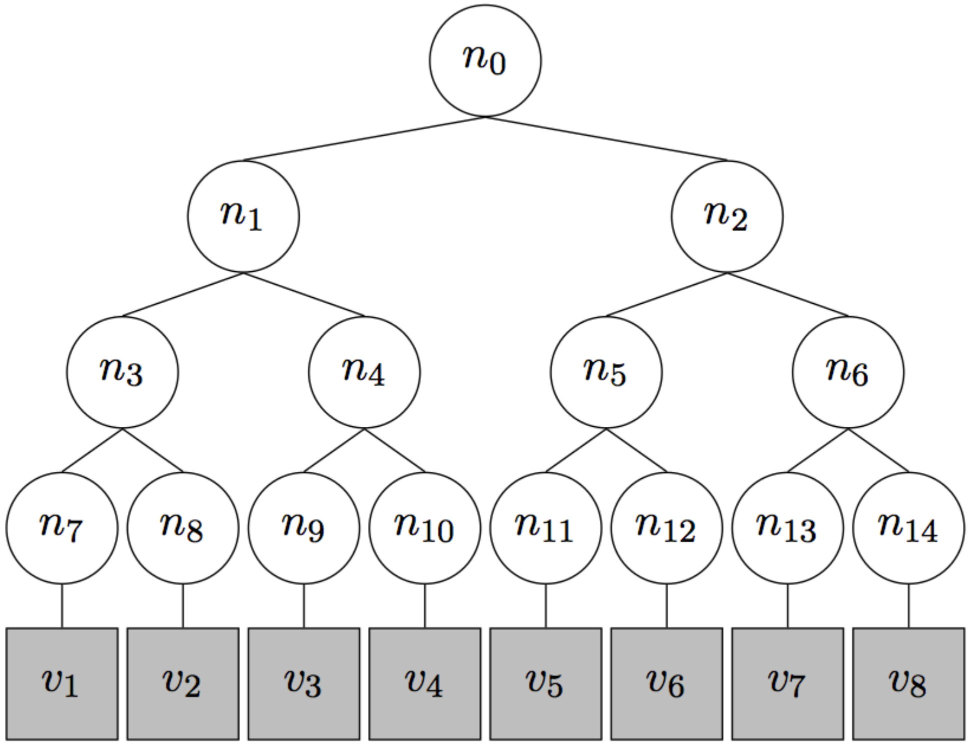
Goal: reduce variance and improve accuracy

$n_1 = n_3 + n_4$

Weighted Averaging

Mean Consistency

MSE is reduced by 3



PURDUE UNIVERSITY

Weighted Averaging

X_1 and X_2 that are both unbiased estimates of the same underlying quantity

$X = \alpha X_1 + (1 - \alpha) X_2$ is also an unbiased estimate of the quantity

To minimize variance of X,

$$\alpha = \frac{\mathrm{Var}(X_2)}{\mathrm{Var}(X_1) + \mathrm{Var}(X_2)}$$

$$\mathrm{Var}(X) = \frac{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}{\mathrm{Var}(X_1) + \mathrm{Var}(X_2)}$$

PURDUE UNIVERSITY®

Weighted Averaging

At level *i*, new count

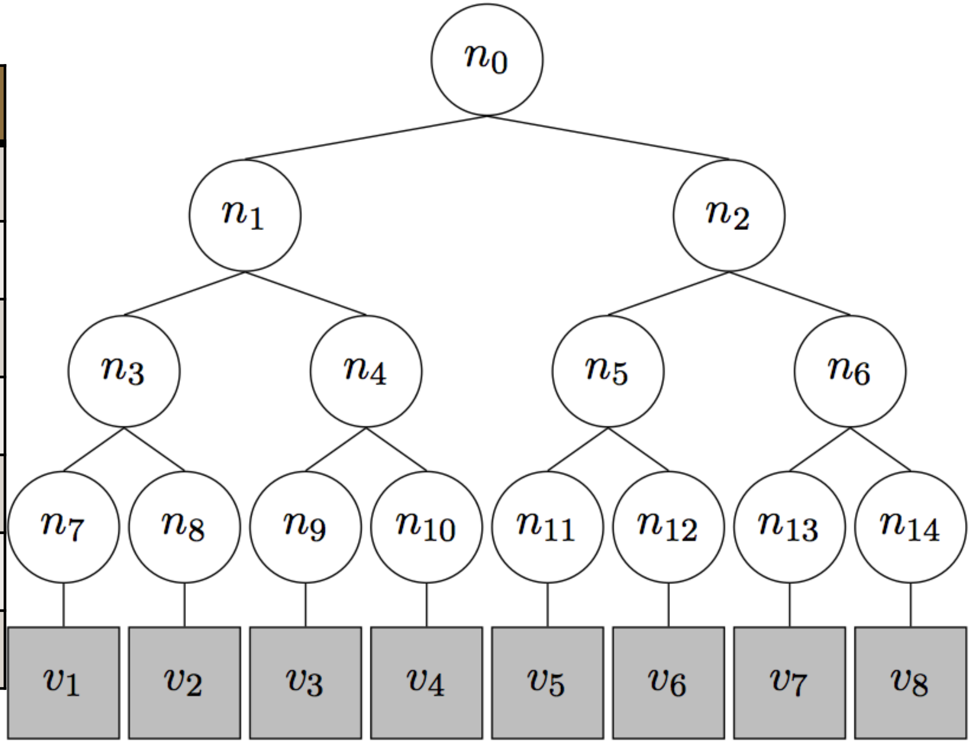
$$z_i[v] = \begin{cases} n[v], & \text{if } i = 1, \text{ i.e., } v \text{ is a leaf node} \\ \frac{b^i - b^{i-1}}{b^i - 1} n[v] + \frac{b^{i-1} - 1}{b^i - 1} \sum_{u \in child(v)} z_{i-1}[u], \text{if } i > 1 \end{cases}$$

$z_i[v]$ is that it is a weighted average of two estimates for the count at v

Weighted Averaging

$$z_i[v] = \begin{cases} n[v], & \text{if } i = 1, \text{ i.e., } v \text{ is a leaf node} \\ \frac{b^i - b^{i-1}}{b^i - 1} n[v] + \frac{b^{i-1} - 1}{b^i - 1} \sum_{u \in child(v)} z_{i-1}[u], & \text{if } i > 1 \end{cases}$$

Node	z
7	n_7
8	n_8
9	n_9
10	n_{10}
3	$(2n_3+n_7+n_8)/3$
4	$(2n_4+n_9+n_{10})/3$
1	$(4n_1+2n_3+n_7+n_8+2n_4+n_9+n_{10})/7$



PURDUE UNIVERSITY

Mean Consistency

From the root down to the leaf level, update each node count so that the sum of each node's children is the same as the node's count

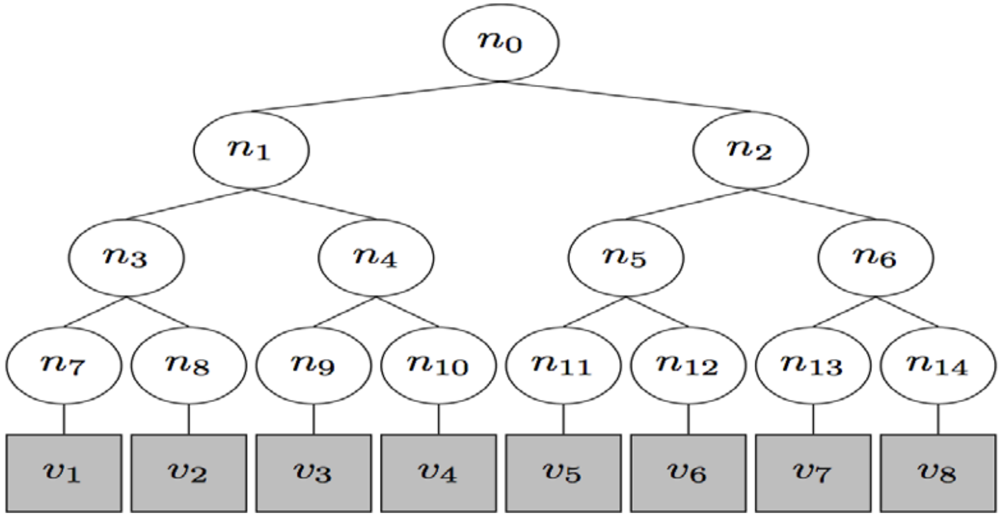
$$\bar{n}_i[v] = \begin{cases} z_i[v], & \text{if } v \text{ is root} \\ z_i[v] + \frac{1}{b}\left(\bar{n}_{i+1}[u] - \sum_{v \in child(u)} z_i[v]\right), & \text{ow} \end{cases}$$

Parent of *v*

Mean Consistency (2)

Node	Value after weighted average	Value after mean consistency
7	n_7	$(3n_1+5n_3-2n_4+13n_7-8n_8-n_9-n_{10})/21$
8	n_8	$(3n_1+5n_3-2n_4-8n_7+13n_8-n_9-n_{10})/21$
9	n_9	$(3n_1-2n_3+5n_4-n_7-n_8+13n_9-8n_{10})/21$
10	n_{10}	$(3n_1-2n_3+5n_4-n_7-n_8-8n_9+13n_{10})/21$
3	$(2n_3+n_7+n_8)/3$	$(6n_1+10n_3-4n_4+5n_7+5n_8-2n_9-2n_{10})/21$
4	$(2n_4+n_9+n_{10})/3$	$(6n_1-4n_3+10n_4+5n_7+5n_8-2n_9-2n_{10})/21$
1	$(4n_1+2n_3+n_7+n_8+2n_4+n_9+n_{10})/7$	$(4n_1+2n_3+n_7+n_8+2n_4+n_9+n_{10})/7$

$$\bar{n}_i[v] = \begin{cases} z_i[v], & \text{if } v \text{ is root} \\ z_i[v] + \frac{1}{b}\left(\bar{n}_{i+1}[u] - \sum_{v \in child(u)} z_i[v]\right), & \text{ow} \end{cases}$$



Privacy Budget Allocation

- Equally distributed
- Geometrical allocation where each level has privacy budget that is $\sqrt[3]{b}$ of its parent level
- With constrained inference
  - using the default equal privacy budget allocation performs as well as using the optimal budget allocation

Wavelet Transformation

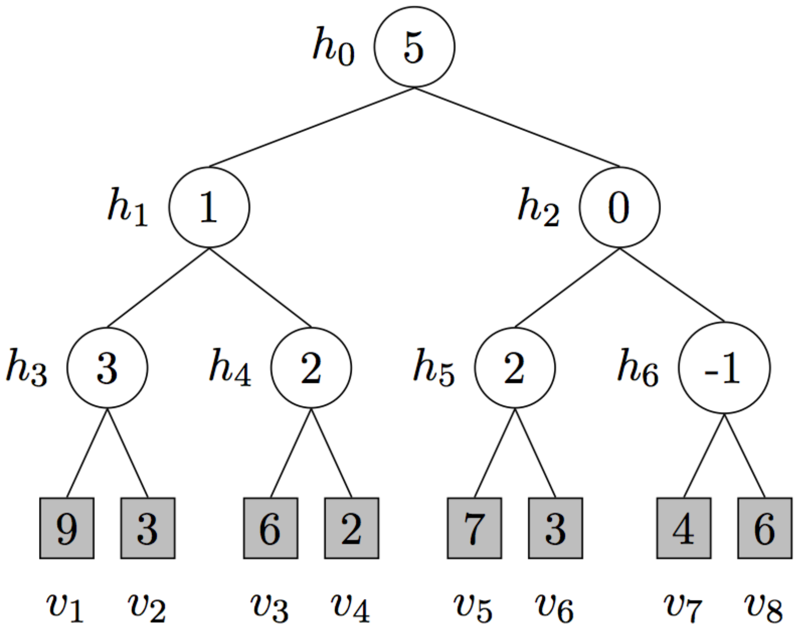
Perform a discrete Haar wavelet transform of the hierarchical histogram

h_0: average of all bins

$h_n = (a_l - a_r)/2$, a_l and a_r are the average of left and right subtree of n

Add noise to the

Haar coefficients



The Matrix Mechanism

- Optimize for a workload of counting queries
- Each counting query can be represented as a binary vector, selecting the unit bins included in the query
- A set of queries can be represented as a matrix
- Given such a matrix, the method finds an alternative set of queries, called a query strategy
- Find the best query strategy in order to answer the given workload queries with minimum error

Beyond One-Dimensional Datasets

- $d = 2$, MSE: $$b(\sqrt{m}-1)\lceil \log_{b^2} m \rceil^2$$

  Total number of bins, n^d

- MSE: $$\Theta(m^{(d-1)/d})$$

- When hierarchical method is better?
  - 1 dimensional: m > 45
  - 2 dimensional: m > 4096
  - 3 dimensional: m > 1.7E6
  - 4 dimensional: m > 2.18E9

Grid Methods

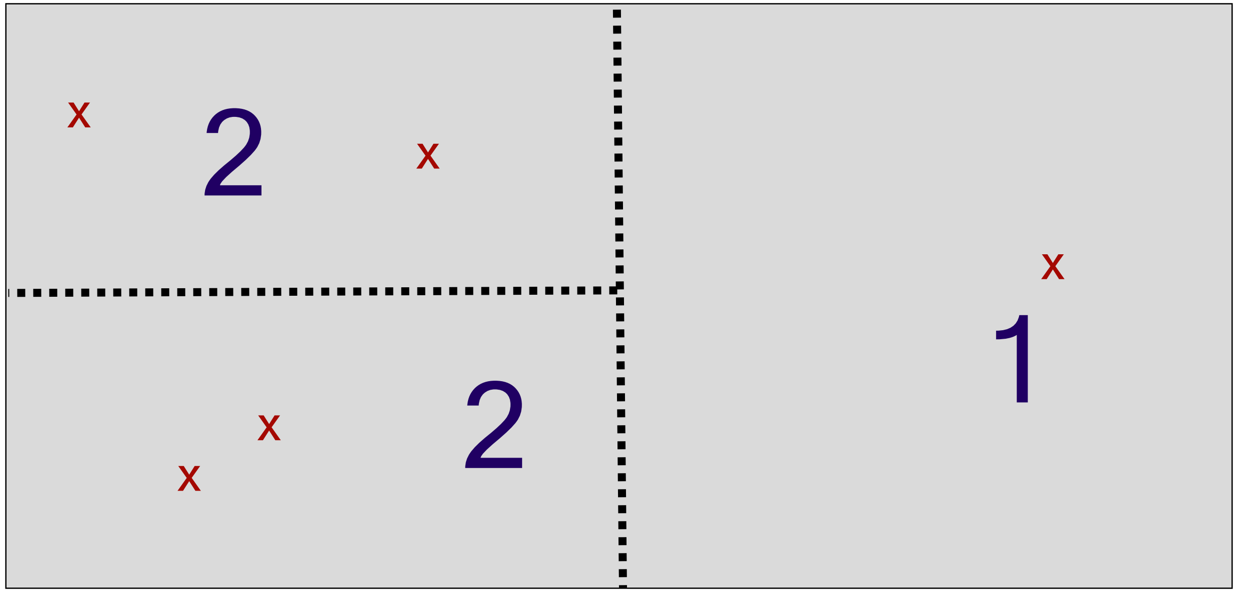
- Reading:

Lacking Suitable Partitioning

- There exists a pre-defined partitioning, but the average number of data points for each bin is low
  - Added noises likely overwhelm the true counts
- The number of natural unit bins is so large that it is infeasible to enumerate through all of them
  - Real numbers

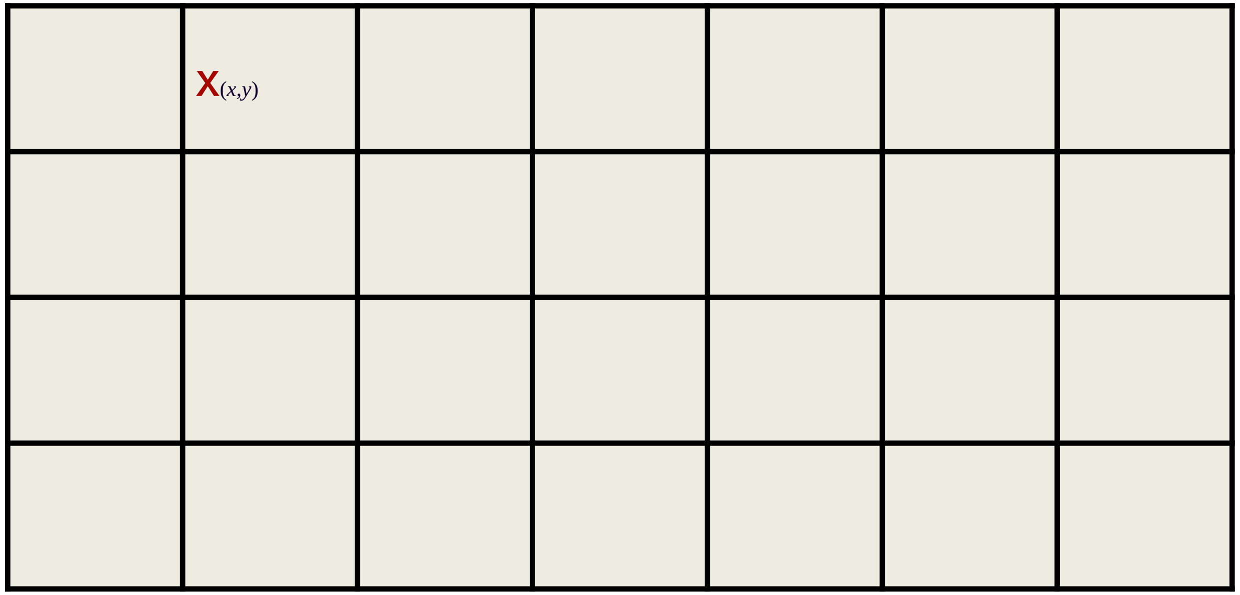
Example: Geospatial Data

Example: Geospatial Data



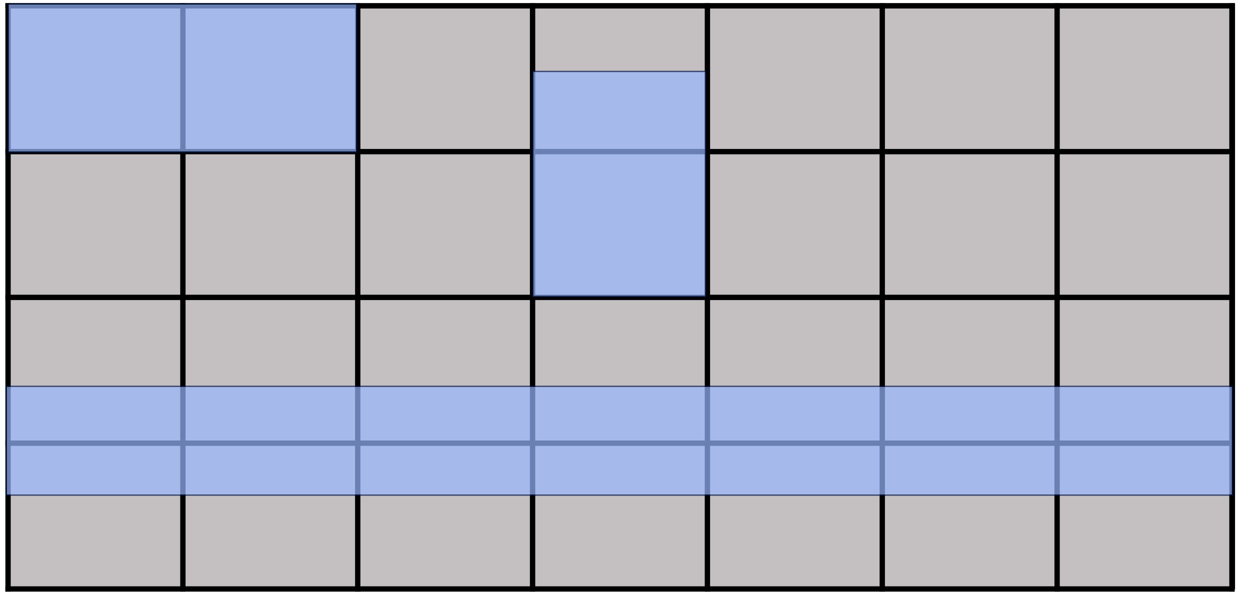
Uniform Grid

- Partition domain into m x m cells of equal size
- Add noise to counts of each cell to satisfy differential privacy



Measuring Utility

- Error from answering range queries
  - a query is a rectangle in the data domain



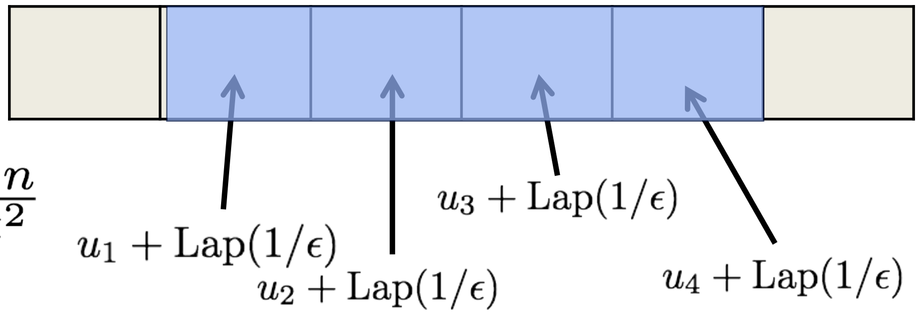
Sources of Error

1. Error from satisfying Differential Privacy (noise error)

- Adding noise from the Laplace Distribution

$$\text{Var}\left(\text{Lap}(1/\epsilon)\right) = \frac{2}{\epsilon^2}$$

$$\sum^n \text{Var}\left(\text{Lap}(1/\epsilon)\right) = \frac{2n}{\epsilon^2}$$

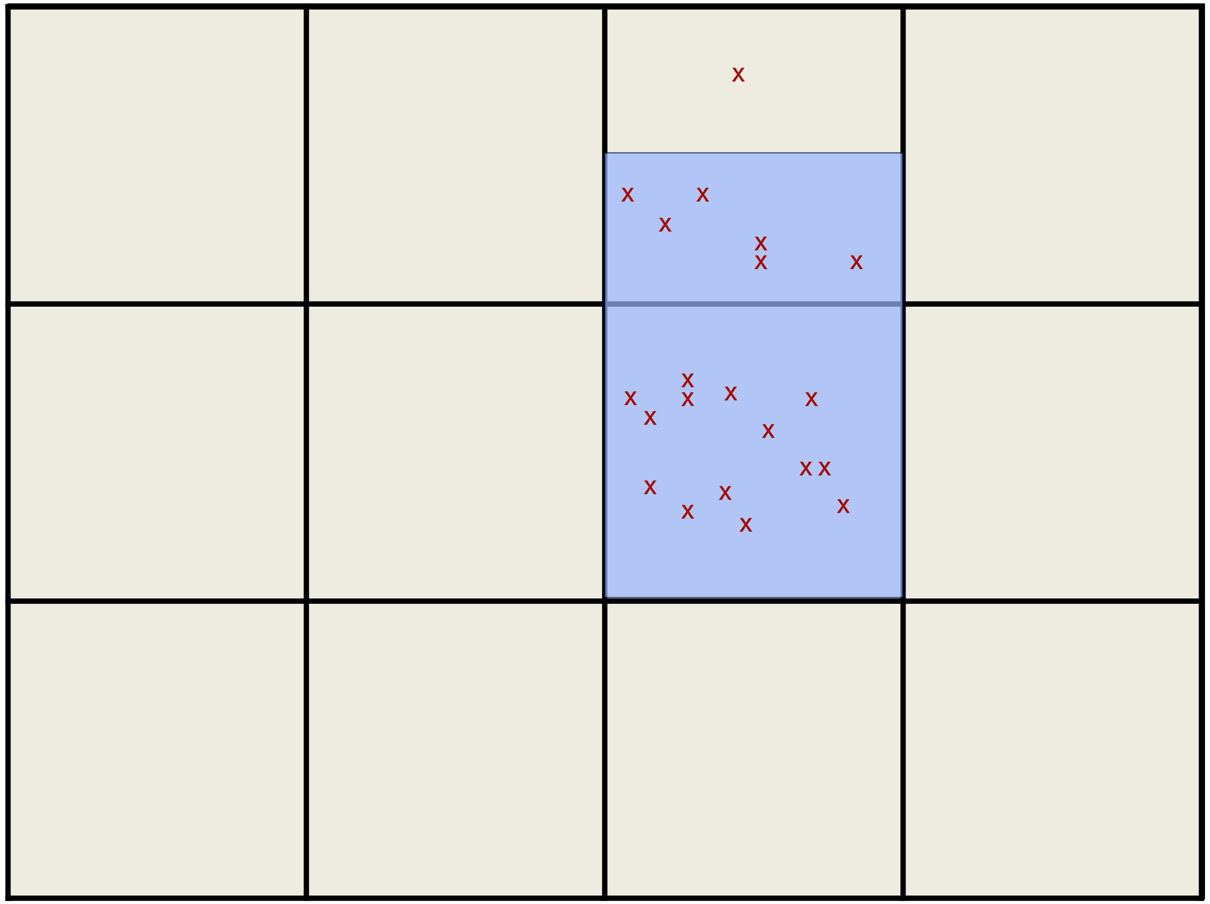


PURDUE UNIVERSITY

Sources of Error

2. Error from grid: Non-uniformity error

- Assuming the data points within each cell are uniformly distributed



Error Minimization

- Noise error: calls for coarser partitioning
- Non-uniformity error: calls for finer partitioning
- Need to choose partition granularity to minimize the sum of the two errors

Determining Grid Size

- m x m grid. Query selects a portion r of the domain.
- Standard deviation of the noise error: $\frac{\sqrt{2rm^2}}{\epsilon}$
- Standard deviation non-uniformity error: $\frac{\sqrt{r}N}{c_0 m}$
- Minimize sum of two errors

$$\arg\min_m \frac{\sqrt{2r}m}{\epsilon} + \frac{\sqrt{r}N}{mc_0}$$

$$m = \sqrt{\frac{N\epsilon}{c}}, c \approx 10$$

PURDUE UNIVERSITY®

Limitation of Uniform Grid

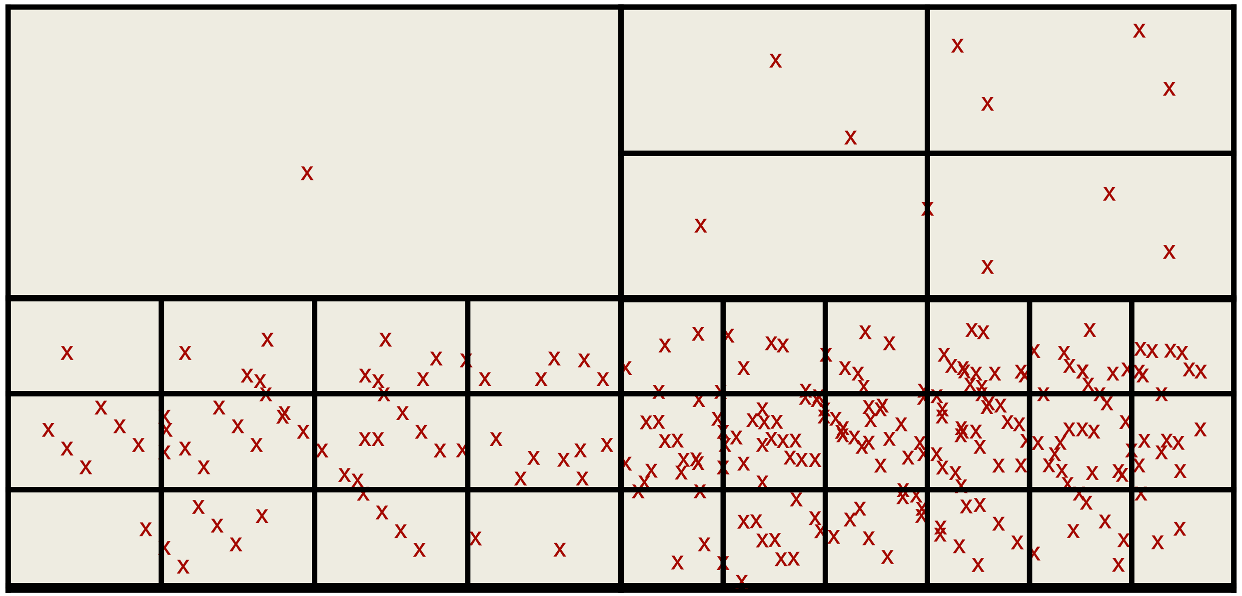
- Uniform Grid treats all regions equally
  - If a region is ***sparse***, we might ***over***-partitioning the region. This increases the noise error with little reduction in the non-uniformity error.
  - if a region is very ***dense***, this method might result in ***under***-partitioning of the region. As a result, the non-uniformity error would be quite large

PURDUE UNIVERSITY

Adaptive Grid

- Adapt the level of partitioning based on the number of data points in each region
  - If a region is dense, use finer granularity to reduce non-uniformity error
  - If a region is sparse, use a more coarse grid

Adaptive Grid



Adaptive Grid

- Two level partitioning:

$\alpha\epsilon$

1. Lay a coarse *m1* x *m1* grid over the data domain and obtain a noisy count for each cell
2. Partition each cell into an *m2* x *m2* grid, where *m2* depends on the noisy count of the cell

$(1-\alpha)\epsilon$

3. Apply constrained inference

Adaptive Grids

- Choosing Parameters (*m2*):
  - Average noise error:

$$\sqrt{\frac{(m_2)^2}{4}}\frac{\sqrt{2}}{(1-\alpha)\epsilon}$$

  - Average non-uniformity error: $\frac{N'}{c_0 m_2}$

$$m_2 = \left\lceil \sqrt{\frac{N'(1-\alpha)\epsilon}{c_2}} \right\rceil$$

Adaptive Grids

- Choosing Parameters (*m1*):
  - Parameter is less critical, since the second level adapts to the count of each cell
  - In general, we want it to be less than the choice for uniform grids.

$$m_1 = \max\left(10, \frac{1}{4}\left\lceil\sqrt{\frac{N\epsilon}{c}}\right\rceil\right).$$

Bottom-up Grouping

- NoisyFirst
  - Compute the noisy histogram by Laplace mechanism
  - Merge bins on the noisy data to reduce the error.

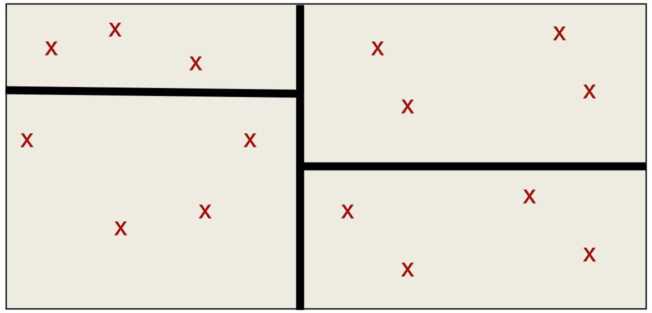
Bottom-up Grouping

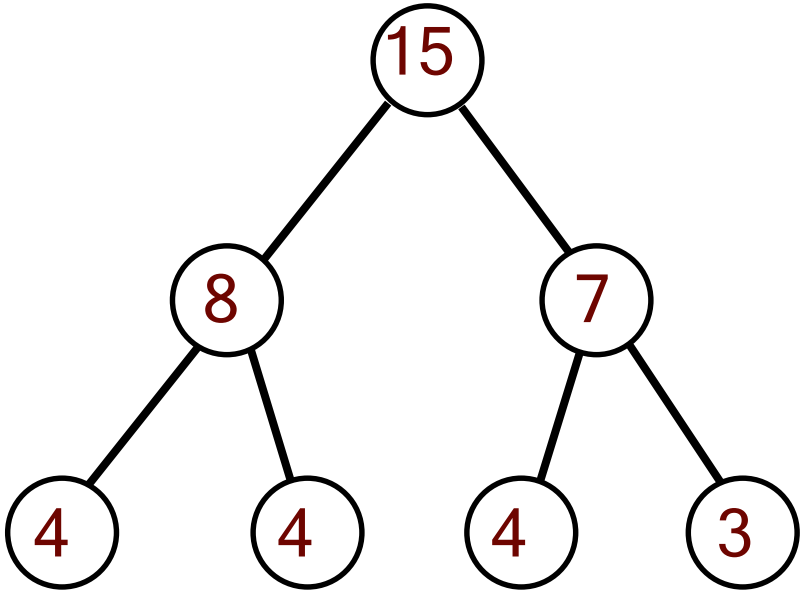
- StructureFirst
  - Use half of privacy budget to get the structure that fits the data and noise well
    - Choose right boundary of histogram bins using exponential mechanism
    - The optimal number of groups *k* is computed by running all possible *k* values in NoisyFirst
  - Spend another half of privacy budget on releasing histogram with Laplace mechanism based on the chosen structure.

Recursive Partition

KD-Trees

Recursively partition along the median
of each axis





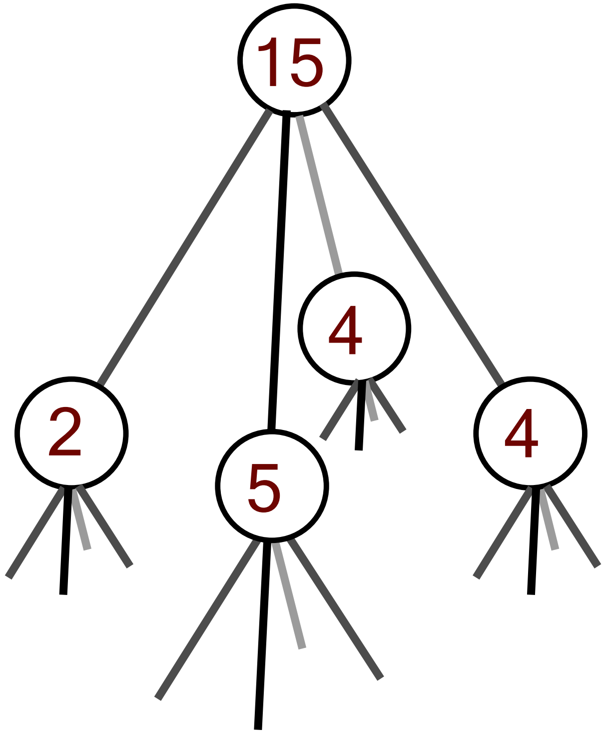
G. Cormode, M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, “Differentially private spatial decompositions,” in *ICDE*, 2012.

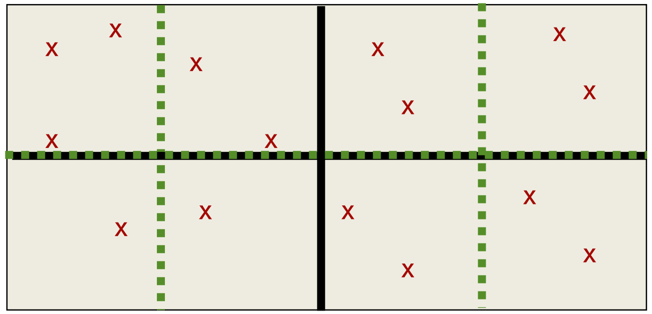
Recursive Partition

Quad-tree

Recursively partition each region into 4 quadrants

Tree of fixed depth





PURDUE UNIVERSITY

G. Cormode, M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, “Differentially private spatial decompositions,” in *ICDE*, 2012.

Recursive Partition

- KD-Hybrid
  - Quad-tree at first few levels and KD-tree for the other levels
- Quad-opt
  - Optimize division of privacy budget

G. Cormode, M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, “Differentially private spatial decompositions,” in *ICDE*, 2012.