Wavelet and Matrix Mechanism

CompSci 590.03
Instructor: Ashwin Machanavajjhala

Duke
UNIVERSITY

Announcement

- Project proposal submission deadline is **Fri, Oct 12 noon**.

Recap: Laplace Mechanism

Thm: If **sensitivity** of the query is **S**, then adding Laplace noise with parameter **λ** guarantees ε-differential privacy, when

$$\boldsymbol{\lambda = S/\varepsilon}$$

Sensitivity: Smallest number s.t. for any d, d' differing in one entry,

$$|| q(d) - q(d') || \le S(q)$$

Histogram query: Sensitivity = 2

- Variance / error on each entry = $2\lambda^2 = 2 \times 4/\varepsilon^2$

Laplace Mechanism is Suboptimal

- Query 1: Number of cancer patients
- Query 2: Number of cancer patients

- If you answer both using Laplace mechanism
  - Sensitivity = 2
  - Error in each answer: $2 \times 4/\varepsilon^2$
  - Average of two answers gives an error of $4/\varepsilon^2$

- If you just answer the first and return the same answer
  - Sensitivity = 1
  - Error in the answer: $2/\varepsilon^2$

Outline

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A

  - **Universal Histograms**
  - **Wavelet Mechanism** **[Xiao et al ICDE 09]**
  - **Matrix Mechanism** **[Li et al PODS 10]**

Note

- The following solution ideas are useful whenever
  - You want to answer a set of correlated queries.
  - Queries are based on noisy measurements.
  - Each measurement (x1 or x1+x2) has similar variance.

Range Queries

- Given a set of values {v1, v2, …, vn}
- Let xi = number of tuples with value v1.
- Range query: q(j,k) = xj + … + xk

Q: Suppose we want to answer all range queries?

Range Queries

Q: Suppose we want to answer all range queries?

Strategy 1: Answer all range queries using Laplace mechanism

- Sensitivity = $O(n^2)$
- $O(n^4/\varepsilon^2)$ total error across all range queries.
- May reduce using constrained optimization …

Duke UNIVERSITY

Range Queries

Q: Suppose we want to answer all range queries?

Strategy 2: Answer all xi queries using Laplace mechanism
Answer range queries using noisy xi values.

- $O(1/\varepsilon^2)$ error for each xi.
- $\text{Error}(q(1,n)) = O(n/\varepsilon^2)$
- Total error on all range queries : $O(n^3/\varepsilon^2)$

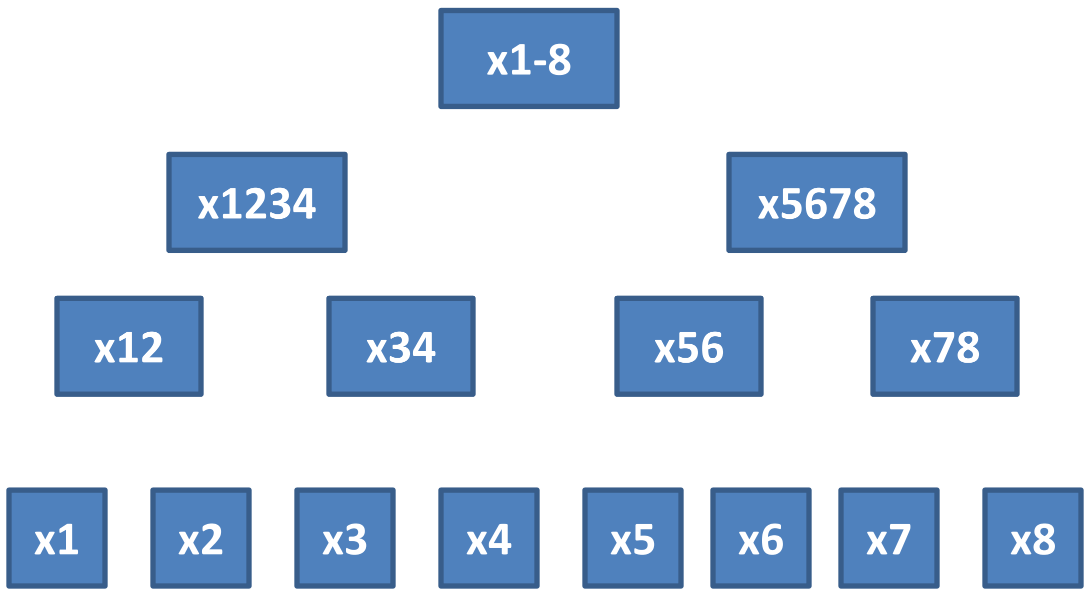
Universal Histograms for Range Queries

[Hay et al VLDB 2010]

Strategy 3:

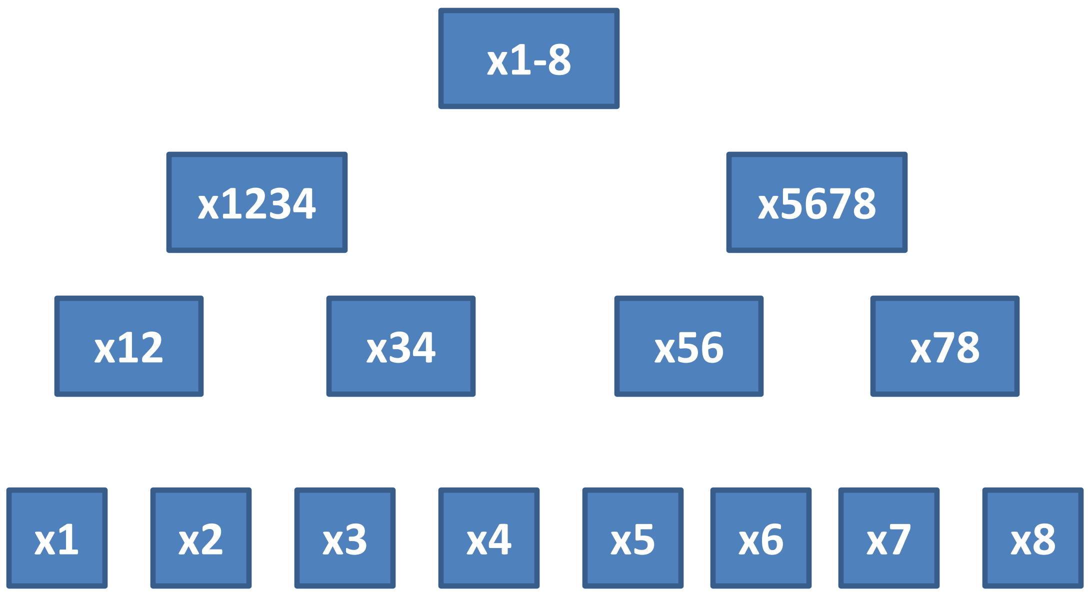
Answer *sufficient statistics* using Laplace mechanism

Answer range queries using noisy sufficient statistics.



Universal Histograms for Range Queries

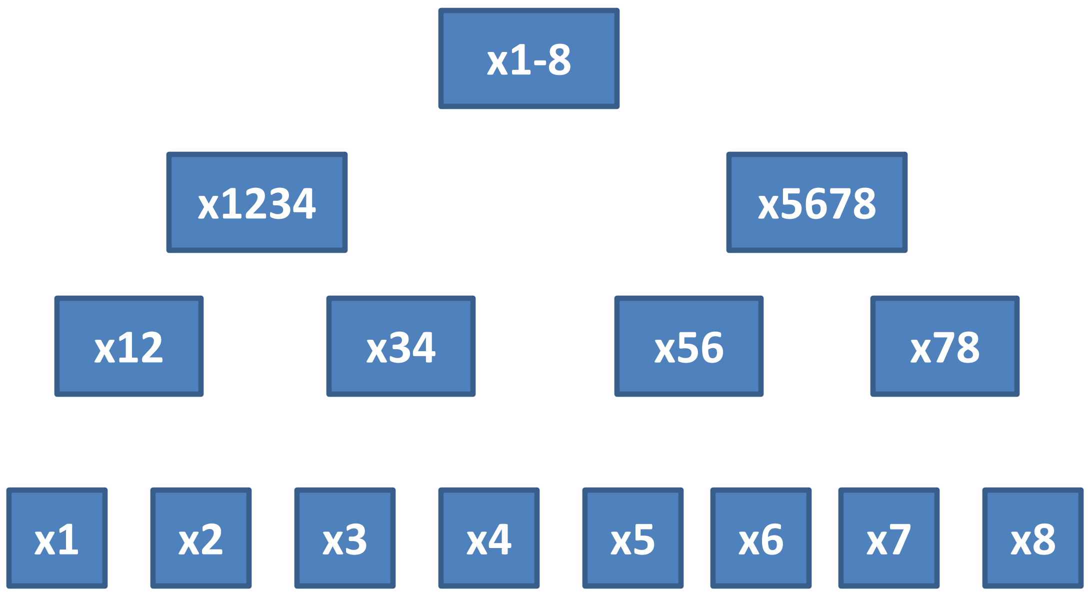
- Sensitivity: log n
- q(2,6) = x2+x3+x4+x5+x6 Error = $2 \times 5\log^2 n/\varepsilon^2$
  = x2 + x34 + x56 Error = $2 \times 3\log^2 n/\varepsilon^2$



Duke
UNIVERSITY

Universal Histograms for Range Queries

- Every range query can be answered by summing at most log n different noisy answers
- Maximum error on any range query = $O(\log^3 n / \varepsilon^2)$
- Total error on all range queries = $O(n^2 \log^3 n / \varepsilon^2)$



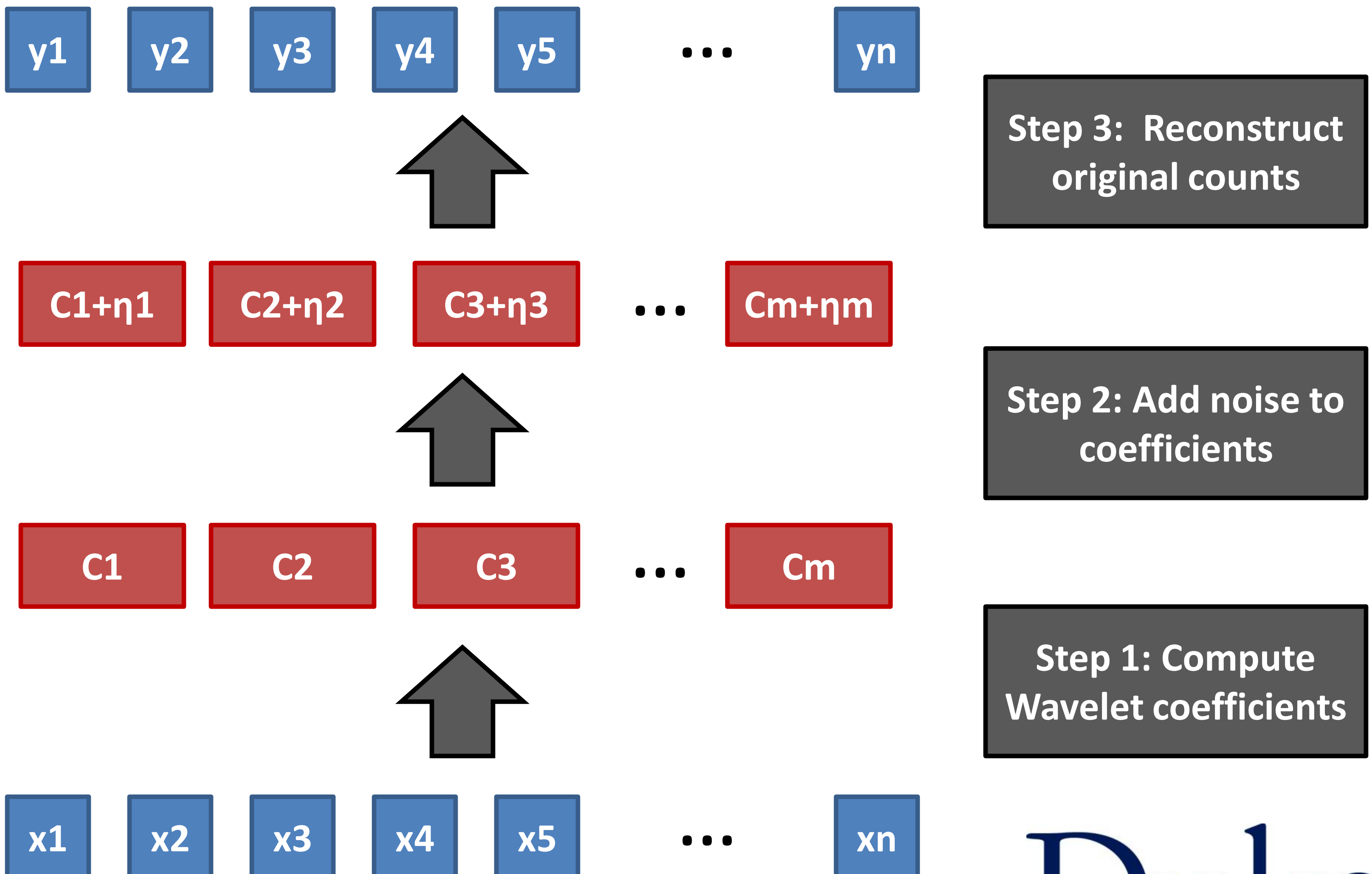
Outline

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

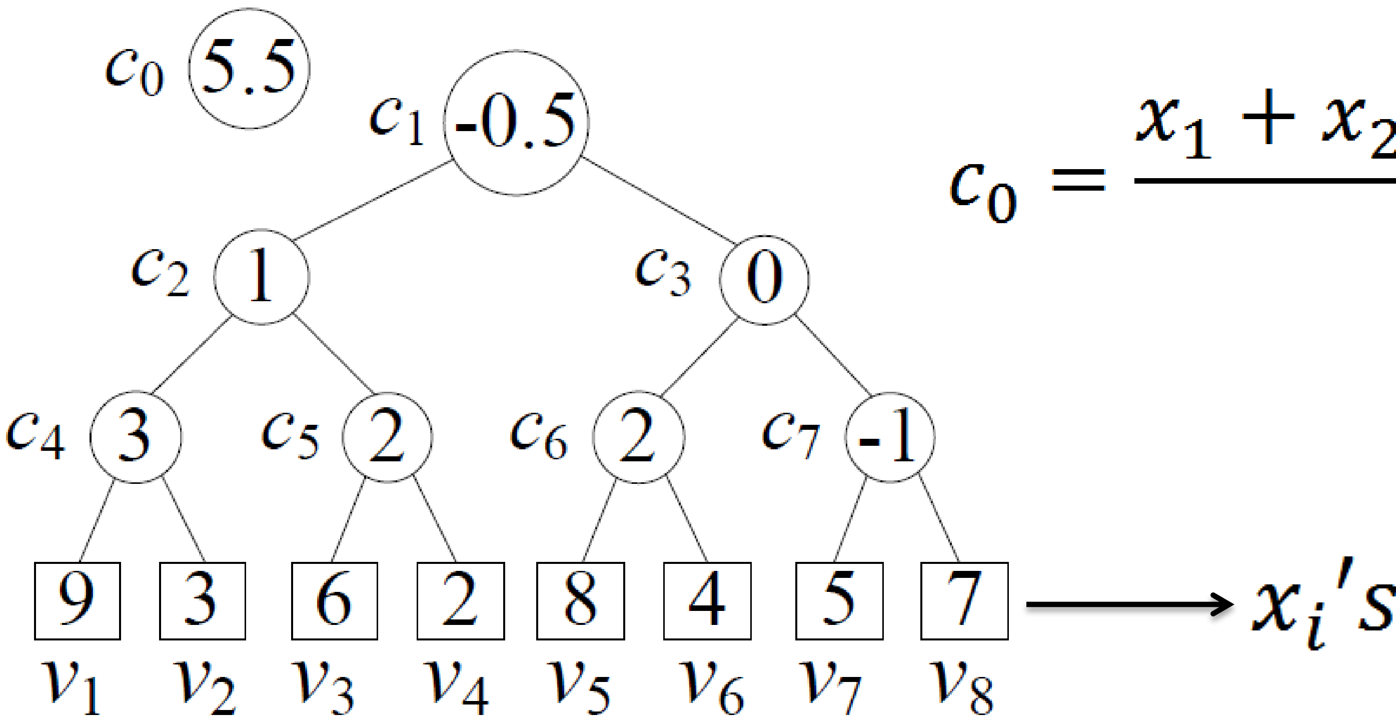
- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A

  - **Universal Histograms**
  - **Wavelet Mechanism**
  - **Matrix Mechanism**

Wavelet Mechanism



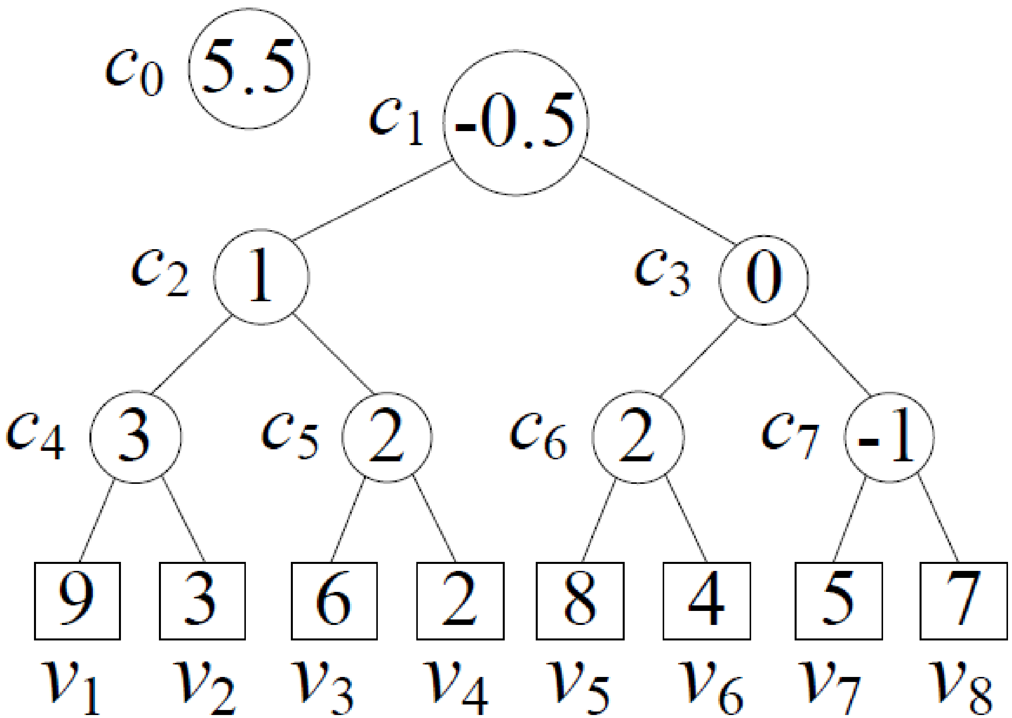
Haar Wavelet



$$c_0 = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Duke UNIVERSITY

Haar Wavelet



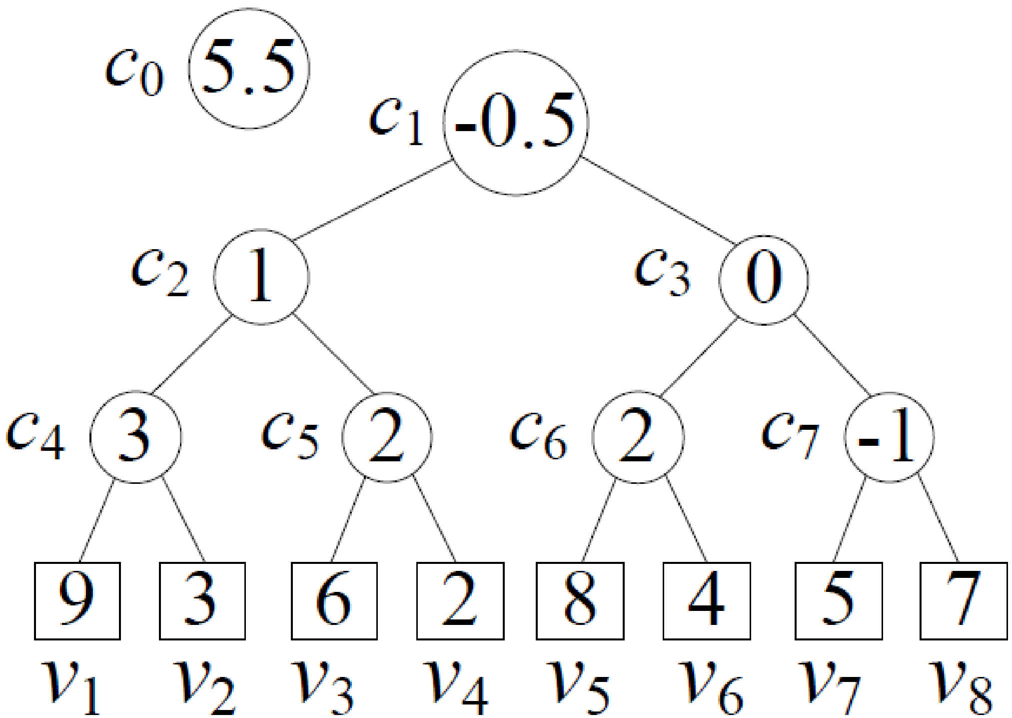
For an internal node,

Let a = average of leaves in left subtree

Let b = average of leaves in right subtree

$$c = \frac{a - b}{2}$$

Haar Wavelet Reconstruction



Sum of coefficients on root to leaf path

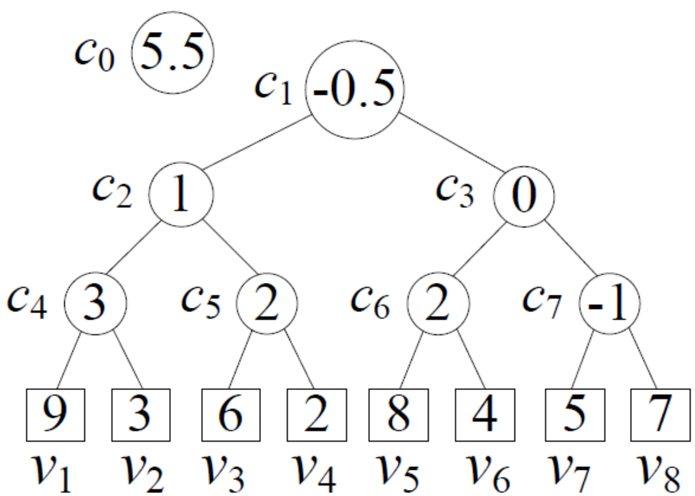
- + if x_i is in the left subtree of coefficient
- - if x_i is in right subtree

$$x_4 = c_0 + c_1 - c_2 - c_5$$

$$x_5 = c_0 - c_1 + c_3 + c_6$$

Duke UNIVERSITY

Haar Wavelet : Range Queries



Range Query: number of tuples in a range S = [a,b]

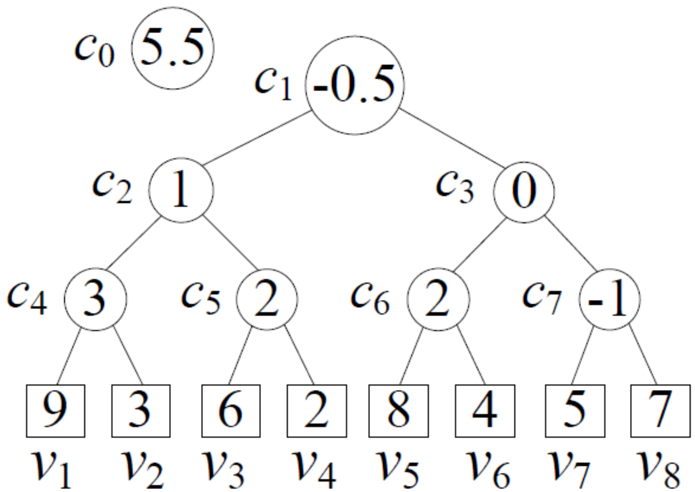
Let α(c) be the number of values in the left subtree of c that are in S

Let β(c) be the number of values in the right subtree of c that are in S

$$y = |S| \cdot c_0 + \sum_{c \neq c_0} (c \cdot (\alpha(c) - \beta(c)))$$

Duke UNIVERSITY

Haar Wavelet : Range Queries



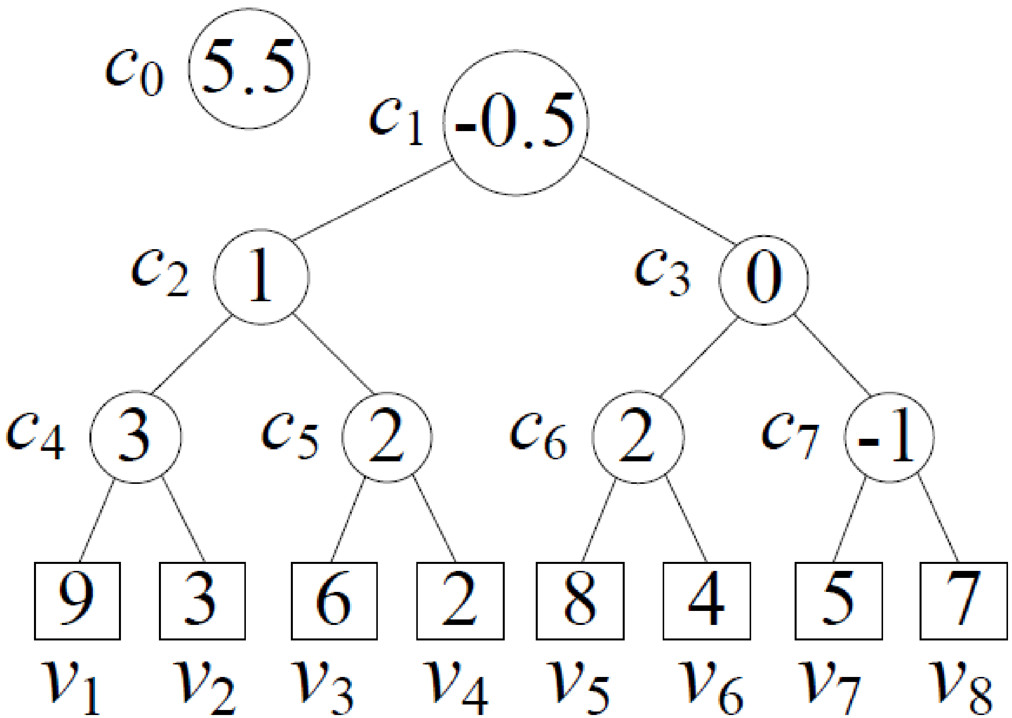
$$y = |S| \cdot c_0 + \sum_{c \neq c_0} (c \cdot (\alpha(c) - \beta(c)))$$

$\alpha(c) - \beta(c) = 0$ when no leaves under c are contained in S

$\alpha(c) - \beta(c) = 0$ when all leaves under c are contained in S

Only need to consider those coefficients with partial overlap with the range.

Haar Wavelet



For an internal node,

Let a = average of leaves in left subtree

Let b = average of leaves in right subtree

$$c = \frac{a - b}{2}$$

Adding noise to wavelet coefficients

- Associate each coefficient with a weight
- level(c) = height of c in the tree.

$$W_{Haar}(c) = 2^{h - level(c) + 1}$$

- Generalized sensitivity (ρ)

$$\sum_{c \in C} \left(W(c) \cdot |c(D) - c(D')| \right) \leq \rho \cdot ||D - D'||_1$$

Duke UNIVERSITY

Adding noise to wavelet coefficients

Theorem: Adding noise to a coefficient c from Laplace(λ/W(c)) guarantees (2ρ/λ)-differential privacy.

Proof:

$$\frac{P[M(D) = < \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_k >]}{P[M(D') = < \tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_k >]} = \frac{\prod_i e^{\left(-\frac{W(c_i)}{\lambda} \cdot |c_i(D) - \tilde{c}_i|\right)}}{\prod_i e^{\left(-\frac{W(c_i)}{\lambda} \cdot |c_i(D') - \tilde{c}_i|\right)}}$$

$$\le e^{\Sigma_i \left(\frac{W(c_i)}{\lambda} \cdot |c_i(D') - c_i(D)|\right)} \quad \le e^{\frac{2\rho}{\lambda}}$$

Duke UNIVERSITY

Generalized Sensitivity of Wavelet Mechanism

$$\rho = 1 + \log_2 n$$

Proof:

- Any coefficient changes by 1/m, where m is the number of values in its subtree.
- m = 1/W(c)
- Only c_0 and the coefficients in one root to leaf path change if some xi changes by 1.

Duke UNIVERSITY

Error in answering range queries

- Range query depends on at most O(log n) coefficients.
- Error in each coefficient is at most $O(\log^2 n/\varepsilon^2)$
- Error in a range query is $O(\log^3 n/\varepsilon^2)$

Summary of Wavelet Mechanism

- Query Strategy: use wavelet coefficients
- Can be computed in linear time
- Noise in each range query: $O(\log^3 n/\varepsilon^2)$

Outline

- **Constrained inference**
  - Ensure that the returned answers are consistent with each other.

- **Query Strategy**
  - Answer a different set of ***strategy*** queries A
  - Answer original queries using A
  - **Universal Histograms**
  - **Wavelet Mechanism**
  - **Matrix Mechanism**

Linear Queries

- A set of linear queries can be represented by a matrix
- **X** = [x1, x2, x3, x4] is a vector representing the counts of 4 values
- $\mathbf{H_4}$ **X** represents the following 7 queries
  - x1+x2+x3+x4
  - x1+x2
  - x3+x4
  - x1
  - x2
  - x3
  - x4

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{H}_4$$

Query Matrices

Identity

$$\begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{bmatrix}$$

$$\mathbf{I}_4$$

Binary Index

$$\begin{bmatrix}1 & 1 & 1 & 1\\1 & 1 & 0 & 0\\0 & 0 & 1 & 1\\1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{bmatrix}$$

$$\mathbf{H}_4$$

Haar Wavelet

$$\begin{bmatrix}1 & 1 & 1 & 1\\1 & 1 & -1 & -1\\1 & -1 & 0 & 0\\0 & 0 & 1 & -1\end{bmatrix}$$

$$\mathbf{Y}_4$$

Sensitivity of a Query Matrix

- How many queries are affected by a change in a single count?

Sensitivity = 1

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{I}_4$$

Sensitivity = 3

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{H}_4$$

Sensitivity = 3

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

$$\mathbf{Y}_4$$

Laplace Mechanism

Sensitivity

$$\mathcal{L}(\mathbf{W}, \mathbf{x}) = \mathbf{W}\mathbf{x} + (\frac{\Delta \mathbf{w}}{\epsilon})\tilde{\mathbf{b}}.$$

Noise Vector of Laplace(1)

Matrix Mechanism

Original Data

Noisy Representation

$$\mathbf{x} \longrightarrow \mathbf{y} = \mathcal{L}(\mathbf{A}, \mathbf{x})$$

$$\downarrow$$

Reconstructed Data

$$\hat{\mathbf{x}}_{\mathbf{A}} = \mathbf{A}^{+}\mathbf{y},$$

$$\textit{where } \mathbf{A}^{+} = (\mathbf{A}^{t}\mathbf{A})^{-1}\mathbf{A}^{t} \textit{ is the } \text{pseudo-inverse} \textit{ of } \mathbf{A}.$$

$$\downarrow$$

Final query answer

$$\mathbf{w}\hat{\mathbf{x}}_{\mathbf{A}}$$

Reconstruction

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

$\mathbf{I}_4$

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

(a) $\mathbf{I}_4{}^{-1}$

$$\begin{bmatrix}1&1&1&1\\1&1&-1&-1\\1&-1&0&0\\0&0&1&-1\end{bmatrix}$$

$\mathbf{Y}_4$

$$\begin{bmatrix}0.25&0.25&0.5&0.0\\0.25&0.25&-0.5&0.0\\0.25&-0.25&0.0&0.5\\0.25&-0.25&0.0&-0.5\end{bmatrix}$$

(c) $\mathbf{Y}_4{}^{-1}$

$$\begin{bmatrix}1&1&1&1\\1&1&0&0\\0&0&1&1\\1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

$\mathbf{H}_4$

$$\frac{1}{21}\times\begin{bmatrix}3&5&-2&13&-8&-1&-1\\3&5&-2&-8&13&-1&-1\\3&-2&5&-1&-1&13&-8\\3&-2&5&-1&-1&-8&13\end{bmatrix}$$

(b) $\mathbf{H}_4^{+}$

Matrix Mechanism

$$\begin{aligned}\mathcal{M}_{\mathbf{A}}(\mathbf{W},\mathbf{x}) &= \mathbf{W}\mathbf{A}^{+}\mathcal{L}(\mathbf{A},\mathbf{x}).\\ &= \mathbf{W}\mathbf{A}^{+}(\mathbf{A}\mathbf{x} + (\frac{\Delta_{\mathbf{A}}}{\epsilon})\tilde{\mathbf{b}})\\ &= \mathbf{W}(\mathbf{x} + (\frac{\Delta_{\mathbf{A}}}{\epsilon})\mathbf{A}^{+}\tilde{\mathbf{b}})\end{aligned}$$

Error analysis

$$\textsc{Error}_{\mathbf{A}}(\mathbf{w}) = Var(\mathbf{w}\hat{\mathbf{x}}_{\mathbf{A}}) = Var(\mathbf{w}\mathbf{x} + (\frac{\Delta_{\mathbf{A}}}{\epsilon})\mathbf{w}\mathbf{A}^{+}\tilde{\mathbf{b}})$$

$$= (\frac{\Delta_{\mathbf{A}}}{\epsilon})^2 Var(\mathbf{w}\mathbf{A}^{+}\tilde{\mathbf{b}}).$$

$$\begin{aligned}
Var(\mathbf{w}\mathbf{A}^{+}\tilde{\mathbf{b}}) &= \mathbf{w}\mathbf{A}^{+} Var(\tilde{\mathbf{b}})(\mathbf{w}\mathbf{A}^{+})^t \\
&= \mathbf{w}\mathbf{A}^{+} 2\mathbf{I}_m (\mathbf{w}\mathbf{A}^{+})^t \\
&= 2\mathbf{w}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\mathbf{A}((\mathbf{A}^t\mathbf{A})^{-1})^t\mathbf{w}^t \\
&= 2\mathbf{w}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{w}^t,
\end{aligned}$$

$$\textsc{TotalError}_{\mathbf{A}}(\mathbf{W}) = (\frac{2}{\epsilon^2})\,\Delta_{\mathbf{A}}^2\ trace((\mathbf{A}^t\mathbf{A})^{-1}\mathbf{W}^t\mathbf{W})$$

Extreme strategies

- Strategy A = In
  - Noisily answer each xi
  - Answer queries using noisy counts

$$\textsc{TotalError}_{\mathbf{I}_n}(\mathbf{W}) = (\frac{2}{\epsilon^2})\ trace(\mathbf{W}^t\mathbf{W})$$

Good when each query hits a few values.

- Strategy A = W
  - Add noise to all the query answers

$$\textsc{TotalError}_{\mathbf{W}}(\mathbf{W}) = (\frac{2}{\epsilon^2})\ \Delta^2_{\mathbf{W}}\ n.$$

Good when sensitivity is small

Duke UNIVERSITY

Finding the Optimal Strategy

- Find A that minimizes $\text{TotalError}_A(W)$
  - Reduces to solving a semi-definite program with rank constraints
  - $O(n^6)$ running time.

- See paper for approximations and an interesting discussion on geometry.

Summary

- A linear query workload and strategy can be modeled using matrices
- Previous techniques to find a better strategy to answer a batch of queries is subsumed by the matrix mechanism
- General mechanism to answer queries.
- Noise depends on the sensitivity of the strategy and A^tA^{-1}

Next Class

- Sparse Vector Technique
  - Answering a workload of “sparse” queries

References

X. Xiao, G. Wang, J. Gehrke, “Differential Privacy via Wavelet Transform”, ICDE 2009

C. Li, M. Hay, V. Rastogi, G. Miklau, A. McGregor, “Optimizing Linear Queries under Differential Privacy”, PODS 2010

Duke UNIVERSITY