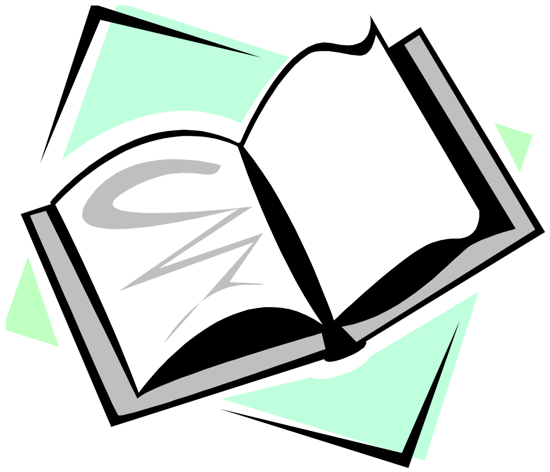
DATA SECURITY AND PRIVACY

Introduction to Differential Privacy

PURDUE UNIVERSITY

Optional Readings for This Lecture

- Differential Privacy: From Theory to Practice
  - Chapter 2: A Primer on Differential Privacy



Differential Privacy [Dwork et al. 2006]

- Definition: A mechanism A satisfies ε-Differential Privacy if and only if
  - for any **neighboring** datasets D and D’
  - and any possible transcript t ∈ Range(A),

$$\Pr[A(D)=t] \le e^{\epsilon} \Pr[A(D')=t]$$

  - For relational datasets, typically, datasets are said to be **neighboring** if they differ by a single record.
- Intuition:
  - Privacy is not violated if one’s information is not included in the input dataset
  - Output does not overly depend on any single record

Laplace Mechanism Calibrating noise to sensitivity [DMNS'06]

Given a function f: $D \to R^d$ over an arbitrary domain D, the *sensitivity* of f is

$$S(f) = \max_{A,B \text{ where } A\Delta B=1} \| f(A) - f(B) \|_1$$

Examples:

1. Count: for $f(D)=|D|$, $S(f)=1$.
2. Sum: for $f(D)=\Sigma d_i$, where $d_i \in [0,\Lambda]$, $S(f)=\Lambda$.

Given a function f: $D \to R^d$ over an arbitrary domain D, the computation

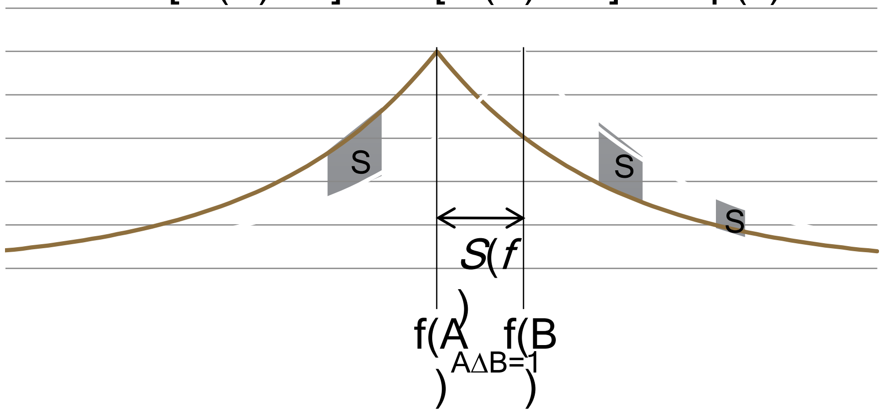
$$M(X) = f(X) + (\mathrm{Lap}\,(S(f)/\varepsilon))^d$$

provides ε-differential privacy.

Examples:

1. NoisyCount(D) = $|D|$+Laplace($1/\varepsilon$).
2. NoisySum(D) = Σd_i +Laplace(Λ/ε).

Pr[M(A)∈S] ≤ Pr[M(B)∈ S] x exp(ε).



Example of Laplace Mechanism

- Consider an example table of N=23,450 records with schema to the right?
- How many tuples are from IN?
  - True count: 546
- Answer while satisfying ε_1-DP: 546 + Lap(Δ/ε_1)
  - $\Delta = 1$
- How many people have score above 23?
- How many …..?

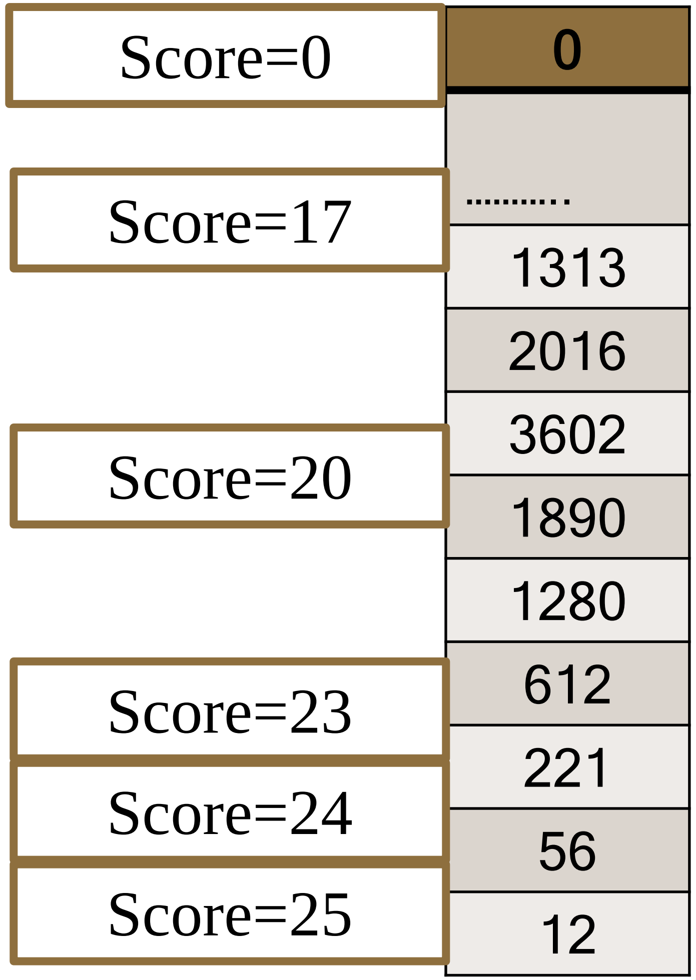
Name	Score	State
Alice	20	CA
Bob	23	CA
Carl	25	IN
David	18	NY
………	……..	……
Frank	20	TX
Jane	14	IN

Counting Queries

- In general, counting queries can be answered relatively accurately
  - Since one tuple affects the result by at most 1
  - A small amount of noise (following the Laplace distribution) can be added to achieve DP

Publishing a histogram

- Suppose we are interested only in the score distribution, then we want to publish the histogram to the right.
- Add Lap(Δ/ε) to each of the cell
- What is the sensitivity Δ?



PURDUE UNIVERSITY

Difference Between Bounded and Unbounded DP

- In unbounded DP, D has one more record than D’
  - Δ(histogram) = 1
- In bounded DP, D and D’ have the same number of records, and only one of them differ
  - Δ(histogram) = 2

PURDUE UNIVERSITY

Exponential Mechanism [MT'07]

Let $q{:}D^n \mathrm{x} R \rightarrow \mathbb{R}$ be a query function that, given a database $d \in D^n$, assigns a score to each outcome $r \in R$.
Then the exponential mechanism M, defined by

$M(d,q)$ = {return r with probability $\propto \exp(\varepsilon q(d,r)/2S(q))$},

maintains ε-differential privacy.

Reminder: $$S(q) = \max_{A,B \text{ where } A\Delta B=1} \left\| q(A) - q(B) \right\|_1$$

Motivation: $$\Pr(r) \propto \exp\left(\varepsilon \boxed{\frac{q(d,r)}{2S(q)}}\right)$$

Impact of changing a single record is within ±1

Example – private vote what to order for lunch:

Option	Score (votes) Sensitivity=1	Sampling Probability ε=0	ε=0.1	ε=1
Pizza	27	0.25	0.4	0.88
Salad	23	0.25	0.33	0.12
Hamburger	9	0.25	0.16	10^{-4}
Pie	0	0.25	0.11	10^{-6}

PURDUE UNIVERSITY

Example of Exponential Mechanism

- What is the median score?
  - Define q(D,x) = – |# of students with score higher than x – # of students with score lower than x|
  - What is the sensitivity?
  - I.e., what is max(|q(D,x) – q(D’,x) |)?

Properties of DP

- Sequential Composability
  - If A_1 satisfies ε_1-DP, and A_2 satisfies ε_2-DP, then outputting both A_1 and A_2 satisfies $(\varepsilon_1+\varepsilon_2)$-DP
- Parallel Composability
  - If D is divided into two parts, applying A_1 and A_2 on the two parts satisfy $(\max(\varepsilon_1,\varepsilon_2))$-DP
- Post-processing Invariance
  - If A_1 satisfies ε_1-DP, then $A_2(A_1(\cdot))$ satisfies ε_1-DP for any A_2

Privacy Budget

- When designing a multiple-step algorithm for ε-DP, one needs to divide ε into portions so that each step consumes some

Some queries are hard to answer

- Some queries are hard to answer
  - E.g., max, since it can be greatly affected by a single tuple

Four Settings of Satisfying DP

- Local setting
  - Do not trust server, perturb data before sending to server
- Interactive setting
  - Answer queries as they come, not knowing what the rest of the queries are
- Single workload
  - Learn a few parameters
- Non-interactive publishing
  - Able to answer a broad range of queries

Limitation of Interactive Setting

- Answering each query consumes some privacy budget
- After answering a pre-determined number of queries, one exhausts the privacy budget, and cannot answer any question anymore
- Problem especially intractable when dealing with multiple users of data

Privacy Preserving Data Publishing

- Design a mechanism A, such that given D, one publishes T=A(D).
- Requirements
  - Privacy friendly
    - Preventing adversaries from learning (individual) information from O=A(D) and A
  - Useful (fidelity-preserving)
    - Allow data users (researchers) to learn (aggregated) information from O=A(D) and A