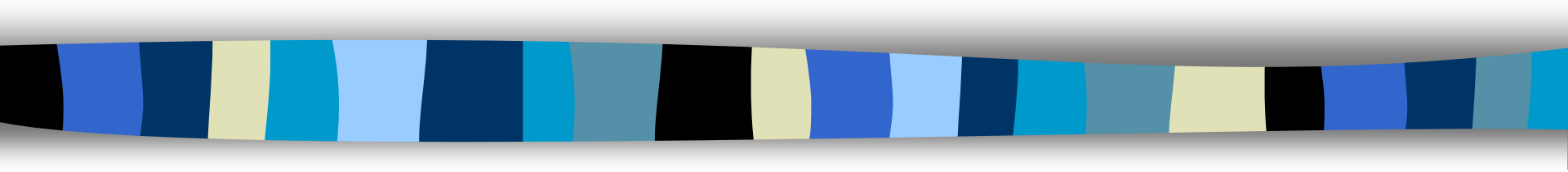
Data Security and Privacy



Topic 19: Differential Privacy

Reading

- Dwork. “Differential Privacy” (invited talk at ICALP 2006).

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

What is Privacy?

It is complicated!

Some concepts from the book “Understanding Privacy” by Daniel J. Solove:

1. the right to be let alone
2. limited access to the self
3. secrecy—the concealment of certain matters from others;
4. control over others' use of information about oneself
5. personhood—the protection of one’s personality, individuality, and dignity;
6. intimacy—control over, or limited access to, one’s intimate relationships or aspects of life.

Impossibility of "Privacy as Secrecy"

- Dalenius [in 1977] proposes this as privacy notion: "*Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access*."
  - Similar to the notion of semantic security for encryption
  - Not possible in the context if one wants utility.
- The Terry's height example:
  - *Adversary knows "Terry is two inches shorter than the average Lithuanian woman"*
  - *Published data reveal average height of Lithuanian woman*
  - *Seeing published info enable learning Terry's height*

Another Example

- Assume that smoking causes lung cancer is not yet public knowledge, and an organization conducted a study that demonstrates this connection. A smoker Carl was not involved in the study, but complains that publishing the result of this study affects his privacy, because others would know that he has a higher chance of getting lung cancer, and as a result he may suffer damages, e.g., his health insurance premium may increase.

- Can Carl legitimately complain about his privacy being violated?

Different Manifestation of the Impossibility Result

- Dwork & Naor: "*absolute disclosure prevention (while preserving utility at the same time) is impossible because of the arbitrary auxiliary information the adversary may have*".
- Kifer and Machanavajjhala: "*achieving both utility and privacy is impossible without making assumptions about the data*."
- Li et al. (Membership privacy framework): "*without restricting the adversary's prior belief about the dataset distribution, achieving privacy requires publishing essentially the same information for two arbitrary datasets*"

- Dwork & Naor: On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy, Journal of Privacy and Confidentiality, 2008.
- Kifer and Machanavajjhala: No Free Lunch in Data Privacy, SIGMOD 2011.
- Li et al.: Membership privacy: a unifying framework for privacy definitions, CCS 2013.

Analogies with Crypto

- Why privacy similar to semantic security is not possible, while semantic security can?
  - There are two kinds of recipients in encryption, but only one in the setting for privacy.
- What about order/property-preserving encryption?
  - Security defined as simulating an ideal world
- "Real-world-ideal-world" approach also used in Secure Multiparty Computation

Differential Privacy

The risk to my privacy should not substantially increase as a result of ***participating in*** a statistical database.

With or without including me in the database, my privacy risk should not change much

(In contrast, the Dalenius definition requires that using the database will not increase my privacy risk, including the case that the database does not even include my record).

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 $t \in$ Range(A),
    $$\Pr[A(D) = t] \leq e^{\epsilon} \Pr[A(D') = t]$$
  - 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

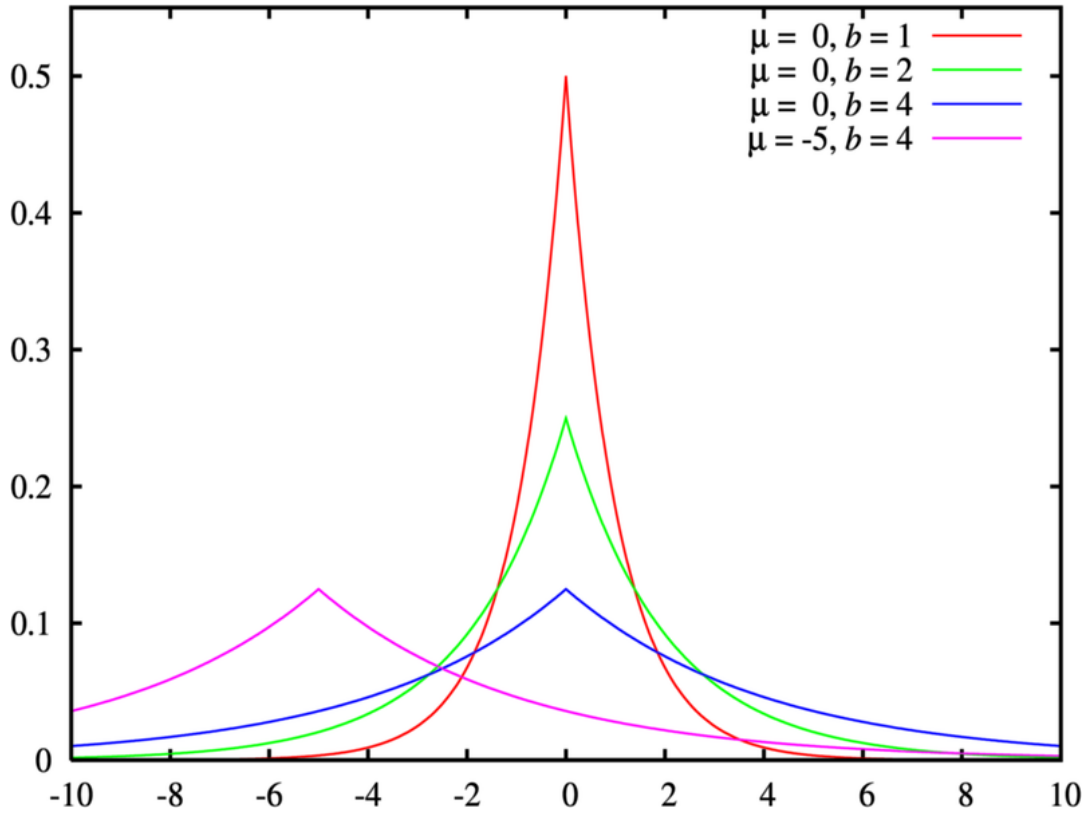
Example of Laplace Mechanism

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

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

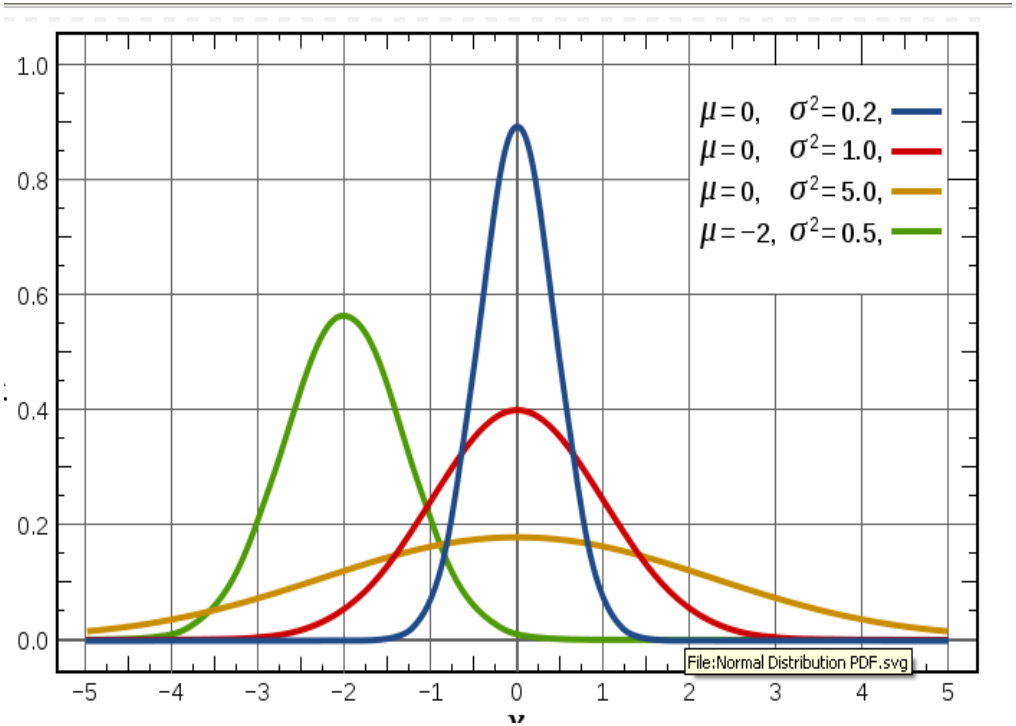
Laplace distribution noise

Using laplacian distribution to generate noise.



parameters:	μ location (real) $b > 0$ scale (real)
support:	$x \in (-\infty; +\infty)$
pdf:	$\frac{1}{2\,b} \exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$
cdf:	*see text*
mean:	μ
median:	μ
mode:	μ
variance:	$2\,b^2$
skewness:	0
ex.kurtosis:	3
entropy:	$\log(2\,e\,b)$

Similar to Guassian noise



notation:	$\mathcal{N}(\mu, \sigma^2)$
parameters:	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
support:	$x \in \mathbf{R}$
pdf:	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
cdf:	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$
mean:	μ
median:	μ
mode:	μ
variance:	σ^2
skewness:	0
ex.kurtosis:	0
entropy:	$\frac{1}{2}\ln(2\pi e\,\sigma^2)$

Laplace Mechanism

Calibrating noise to sensitivity [DMNS'06]

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

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

Examples:

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

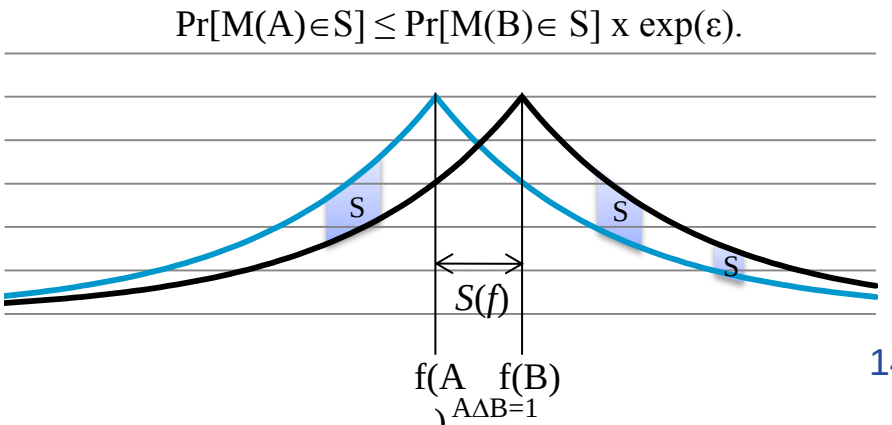
Given a function $f:D \to P^d$ over an arbitrary domain D, the computation

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

provides ε-differential privacy.

Examples:

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



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?

Score=0	0
	…….
Score=17	1313
	2016
	3602
Score=20	1890
	1280
	612
Score=23	221
Score=24	56
Score=25	12

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

Exponential Mechanism [MT'07]

Let q:D^nx$R \to \mathbb{R}$ 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: $S(q) = \max_{A,B \text{ where } A \Delta B = 1} \|q(A) - q(B)\|_1$

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

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}

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

Example of Exponential Mechanism

- Median:
  - Divide the domain into a number of discrete ranges, each range's quality based on difference of tuples above & below the region
  - Can be repeated in a few steps

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

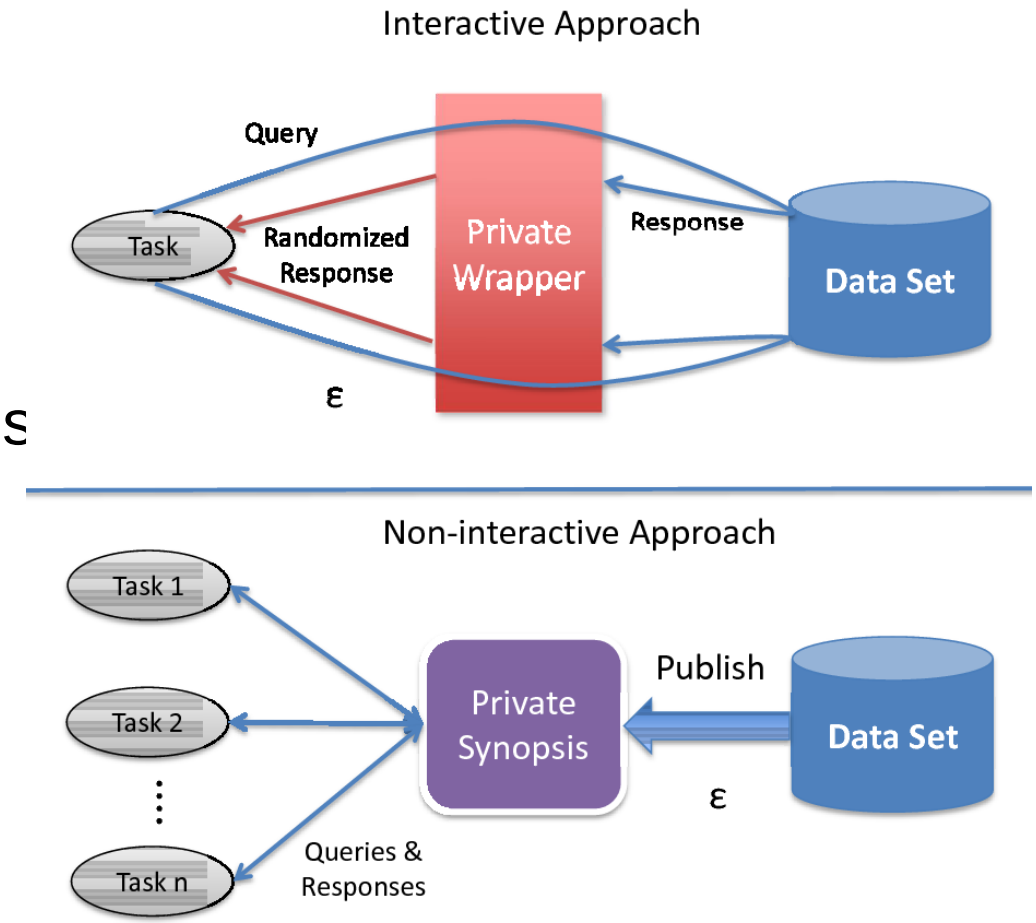
- Local setting
  - Do not trust server, perturb data before sending to server
- Interactive setting
  - Answer queries as they come, not knowing what the rest of the queries are
- Single workload
  - Learn a few parameters
- Non-interactive publishing
  - Able to answer a broad range of queries

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

Our Focus

- Non-interactive data publishing rather than interactive query answering
- Methodology: Combining analysis of how algorithm performs with experimental validation
- Diverse problem domains require different methods
  - e.g., number of dimensions



Next Lecture

- Meanings and caveats of DP