Privacy-Preserving Natural Language Processing

Lecture 4 – Pure Differential Privacy Mechanisms

Prof. Dr. Ivan Habernal

May 8, 2025

www.trusthlt.org
Chair of Trustworthy Human Language Technologies (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

RUHR UNIVERSITÄT BOCHUM RUB

TrustHLT 1001100110101

CENTER FOR TRUSTWORTHY DATA SCIENCE AND SECURITY

Recap Lecture 3

Example: How many persons in the database take illegal drugs?

Name	Hospitalized in year	Age	Illegal drug use
Alice	2024	32	yes
Bob	2020	21	no
Charlie	2023	45	no
...			
Xander	2020	31	yes

Public information: the size of the dataset (say $n = 100$)

The true answer to the query: 2

If Xander reveals his drug use, and we know Alice is in the database, her privacy is revealed!

Irreversibly alter the true query result by randomness

The randomized output should be likely close to the true value, and unlikely far away

We picked the Laplace distribution with scale $b = 1$

We explored two possible scenarios

- D: Alice was in the dataset (and then querying)
- D': Alice was removed from the dataset (and then querying)

For any observed y, was it sampled from Y or from Y'?

The underlying database was D for Y and D' for Y', respectively

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp(1) \qquad \frac{\Pr[Y' = y]}{\Pr[Y = y]} \leq \exp(1)$$

No matter what value we get after privatizing the counting query – we can only get some limited "information" about whether it came from Y or Y'.

Summary

Four counting queries, the maximum difference of the query result is 1 (when a particular person is not in the dataset)

Scale $b = 1$ for the Laplace distribution gives us the following upper bound on privacy loss

$$\frac{\Pr[Y=y]}{\Pr[Y'=y]} \leq \exp(1) \qquad \frac{\Pr[Y'=y]}{\Pr[Y=y]} \leq \exp(1)$$

Neighboring datasets

Defining neighboring datasets

We saw that the bound was 'symmetric'

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp(1) \qquad \frac{\Pr[Y' = y]}{\Pr[Y = y]} \leq \exp(1)$$

On the left: One entry less in the denominator (= one entry more in the nominator)

On the right: One entry less in the nominator (= one entry more in the denominator)

The bound holds for any two datasets that **differ in the presence of one entry** (e.g., one row removed, or on row added) → **neighboring datasets**

Neighboring dataset examples (choice of Alice is arbitrary!)

Name	Hospitalized in year	Age	Illegal drug use
Alice	2024	32	yes
Bob	2020	21	no
Charlie	2023	45	no

Table 1: Database 1, including Alice

Name	Hospitalized in year	Age	Illegal drug use
Bob	2020	21	no
Charlie	2023	45	no

Table 2: Database 2, excluding Alice

Neighboring dataset examples

Databases 1 and 2 from the previous slide are **neighboring** as they differ in presence or absence of a single row

We denote them 1: D and 2: D' **(an arbitrary choice)**

The neighboring dataset definition allows us to simplify

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp(1) \qquad \frac{\Pr[Y' = y]}{\Pr[Y = y]} \leq \exp(1)$$

into one inequality (where again Y is a random variable based on D and Y' is a random variable based on D')

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp(1)$$

Restating our bound with neighboring datasets

Neighboring datasets = presence or absence of a single individual

Our bound $\frac{\Pr[Y=y]}{\Pr[Y'=y]} \leq \exp(1)$ holds for any neighboring datasets (we proved it, but convince yourself again)

So the absence or presence of **any single individual**

- will influence the result of the counting query (but we want to keep this influence small → **utility**)
- for any 'all-mighty' adversary, the likelihood to find out the extra person's 'private bit' (e.g., drug use) is upper bounded (the person wants to keep this bound small → **privacy**)

Controlling privacy strength

What if 2.718 is not strong enough?

For $y \sim y_{\text{true}} + \text{Lap}(b = 1)$

The likelihood for preferring one hypothesis over the other

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp(1) \approx 2.718$$

What would be the minimum of $\frac{\Pr[Y=y]}{\Pr[Y'=y]}$?

What would be the maximum of $\frac{\Pr[Y=y]}{\Pr[Y'=y]}$?

What if 2.718 is not strong enough?

For $y \sim y_{\text{true}} + \text{Lap}(b = 1)$

The likelihood for preferring one hypothesis over the other

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp(1) \approx 2.718$$

What would be the minimum of $\frac{\Pr[Y=y]}{\Pr[Y'=y]}$?

What would be the maximum of $\frac{\Pr[Y=y]}{\Pr[Y'=y]}$?

$$1 \leq \frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \infty$$

Why?

Playing with the scale parameter b

For $y \sim y_{\text{true}} + \text{Lap}(\mu = 0; b = 1)$

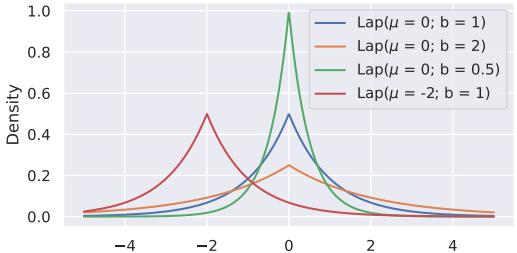


Figure 1: Laplace PDF $\text{Lap}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$

What to do with the scale for stronger privacy?

Intuition: Larger b gives stronger privacy

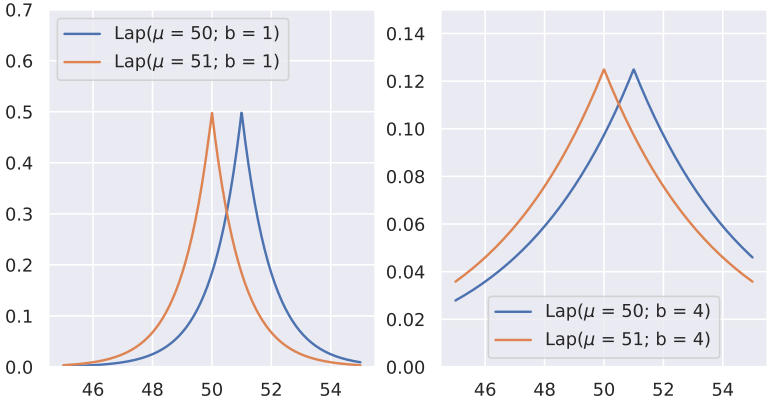


Figure 2: Laplace PDF (probability density function)

How does that relate to the maximum privacy loss?

Recall: likelihood of any output (x-axis) coming from D' as opposed do D (and vice versa)

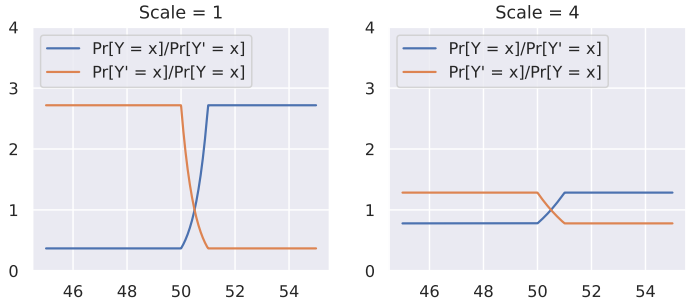


Figure 3: Privacy loss for two Laplace distributions for a counting query, varying scale b

Let's try to formalize our intuition

Counting query true value $y = 51$, one person removed

$y = 50$

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} = \frac{\frac{1}{2b}\exp\left(-\frac{1}{b}|y - 51|\right)}{\frac{1}{2b}\exp\left(-\frac{1}{b}|y - 50|\right)} = \exp\left(\frac{1}{b}|y - 50| - |y - 51|\right)$$

$$\leq \exp\left(\frac{1}{b}|(y - 50) - (y - 51)|\right)$$

$$= \exp\left(\frac{1}{b}|y - 50 - y + 51|\right)$$

$$= \exp\left(\frac{1}{b}|1|\right) = \exp\left(\frac{1}{b}\right)$$

Will be the same for $\frac{\Pr[Y'=y]}{\Pr[Y=y]}$ (prove it! check the appendix)

What have we discovered so far

Our choice of 50 and 51 was arbitrary, the same proof will hold for any true answers differing by 1

For a **counting query**, the database curator (**trusted curator**, trusted holder) protects privacy of each individual by reporting

$$y \sim y_{\text{true}} + \text{Lap}(b)$$

This ensures that for **any two neighboring datasets** the privacy loss is **bounded**

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} \leq \exp\left(\frac{1}{b}\right)$$

Formalizing differential privacy

1. Recap Lecture 3
2. Neighboring datasets
3. Controlling privacy strength
4. Formalizing differential privacy
5. Laplace mechanism
6. Exponential mechanism

Let's generalize our findings

We had $\frac{\Pr[Y=y]}{\Pr[Y'=y]} \leq \exp(\frac{1}{b})$

We saw, that $\frac{1}{b}$ controls the strength of privacy. Let's generalize this notion and call it

Privacy budget

Denoted as $\varepsilon \in [0, \infty)$

$\varepsilon = 0$ is complete privacy but completely random

$\varepsilon = \infty$ is no privacy whatsoever

So for our counting query example, we would have $\frac{\Pr[Y=y]}{\Pr[Y'=y]} \leq \exp(\varepsilon)$ where $\varepsilon = \frac{1}{b}$

Let's formalize pure differential privacy

Let's just rewrite $\frac{\Pr[Y=y]}{\Pr[Y'=y]} \leq \exp(\varepsilon)$:

$$\Pr[Y = y] \leq \exp(\varepsilon) \Pr[Y' = y]$$

Let's generalize the random variables Y and Y' by saying thery are **random variables parametrized by the dataset** (randomized algorithms, **randomized mechanisms**), e.g., $\mathcal{M}(D)$ or $\mathcal{M}(D')$

$$\Pr[\mathcal{M}(D) = y] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') = y]$$

Let's formalize pure differential privacy

We had $\Pr[\mathcal{M}(D) = y] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') = y]$

To make this defintion work for **any** random variable (categorical, discrete finite, countably infinite, uncounble), we must generalize the co-domain of $\mathcal{M}$

The co-domain of $\mathcal{M}(D)$ can be some arbitrary set $\mathcal{Z}$

We want that our privacy guarantees hold for **any possible output** $\mathcal{Y}$ which is any subset of the co-domain, i.e. $\mathcal{Y} \subseteq \mathcal{Z}$:

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}]$$

C. Dwork and A. Roth (2013). **"The Algorithmic Foundations of Differential Privacy".** In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407, Definition 2.4

$(\varepsilon, 0)$ differential privacy (aka. pure DP)

C. Dwork and A. Roth (2013). **"The Algorithmic Foundations of Differential Privacy".** In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407, Definition 2.4

A randomized algorithm (mechanism) $\mathcal{M}$ is $(\varepsilon, 0)$-**differentially private** if for any two neighboring datasets D, D' and any output $\mathcal{Y} \subseteq \mathcal{Z}$ this guarantee holds:

$$\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}]$$

Laplace mechanism

Laplace mechanism for counting query is $(\varepsilon, 0)$-DP

C. Dwork and A. Roth (2013). **"The Algorithmic Foundations of Differential Privacy".** In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407, Definition 2.4

Counting query: Report a number of persons with a certain attribute

Draw a random number $y_{\text{lap}} \sim \text{Lap}(\mu = 0; b = \frac{1}{\varepsilon})$

Report $y = y_{\text{true}} + y_{\text{lap}}$

Proof: We did it already! (but practice it now backwards from the definition)

Laplace mechanism

General numeric queries

From counting query to numeric query

Counting query: Report a number of persons with certain attribute(s)

$$f : \mathcal{X} \mapsto \mathcal{Y} \subseteq \mathbb{N}$$

Real-valued query: Report any value computed over certain attribute(s)

$$f : \mathcal{X} \mapsto \mathcal{Y} \subseteq \mathbb{R}$$

Working example of numeric query

Name	Debt (in €)
Alice	2,800,798.00
Bob	7,000.00
Charlie	1.56
Xander	0.00

Table 3: Debts of account holders in a bank

Important constraint: Maximum debt the bank offers per person is 10,000,000 €

Query: What is the total debt in €?

Naive attempt with the Laplace mechanism for counting queries

Laplace mechanism for counting query

Draw $y_{\text{lap}} \sim \text{Lap}(\mu = 0; b = \frac{1}{\varepsilon})$ and report $y = y_{\text{true}} + y_{\text{lap}}$

Will it ensure $\Pr[\mathcal{M}(D) \in \mathcal{Y}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') \in \mathcal{Y}]$ for all neighboring datasets and any outputs?

Name	Debt (in €)
Alice	2,800,798.00
Bob	7,000.00
Charlie	1.56
Xander	0.00

Naive attempt with the Laplace mechanism for counting queries

Name	Debt (in €)
Alice	2,800,798.00
Bob	7,000.00
Charlie	1.56
Xander	0.00

Let's try for $D = \{A, B, C, X\}$ and $D' = \{B, C, X\}$

- $y_{\text{true}}(D) = 2,807,799.56$
- $y_{\text{true}}(D') = 7,001.56$

Naive solution with the Laplace mechanism for counting queries

We had $y_{\text{true}}(D) = 2,807,799.56$, $y_{\text{true}}(D') = 7,001.56$

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} = \frac{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}}|y - 2,807,799.56|\right)}{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}}|y - 7,001.56|\right)} = \frac{\exp\left(-\varepsilon|y - 2,807,799.56|\right)}{\exp\left(-\varepsilon|y - 7,001.56|\right)}$$

$$= \exp\left\{\varepsilon \underbrace{\left(|y - 7,001.56| - |y - 2,807,799.56|\right)}_{\text{Exceeds 1 for some large } y}\right\}$$

$$\not\leq \exp(\varepsilon)$$

This naive solution violates DP! How to fix this?

Laplace mechanism

Global ℓ_1 sensitivity

How different neighboring datasets influence the output

Name	Debt (in €)
Alice	2,800,798.00
Bob	7,000.00
Charlie	1.56
Xander	0.00

$D = \{A, B, C, X\}$

- $D' = \{B, C, X\} = D \setminus \{A\}$
  - $y_{\text{true}}(D) = 2,807,799.56$
  - $y_{\text{true}}(D') = 7,001.56$
  - $y_{\text{true}}(D) - y_{\text{true}}(D') = 2,800,798$

How different neighboring datasets influence the output

Name	Debt (in €)
Alice	2,800,798.00
Bob	7,000.00
Charlie	1.56
Xander	0.00

- $D' = D \setminus \{B\}$
  - $y_{\text{true}}(D) = 2,807,799.56$
  - $y_{\text{true}}(D') = 2,800,799.56$
  - $y_{\text{true}}(D) - y_{\text{true}}(D') = 7,000$

You should now see the pattern for this sum query

How different neighboring datasets influence the output

Name	Debt (in €)
Alice	2,800,798.00
Bob	7,000.00
Charlie	1.56
Xander	0.00

- $D' = D \setminus \{A\}$: $y_{\text{true}}(D) - y_{\text{true}}(D') = 2,800,798$
- $D' = D \setminus \{B\}$: $y_{\text{true}}(D) - y_{\text{true}}(D') = 7,000$
- $D' = D \setminus \{C\}$: $y_{\text{true}}(D) - y_{\text{true}}(D') = 1.56$
- $D' = D \setminus \{X\}$: $y_{\text{true}}(D) - y_{\text{true}}(D') = 0$

Each person can change the result of the query by its value

Can we use this max value to fix our problem?

Rescale the Laplace distribution

Previous attempt

We had $y_{\text{true}}(D) = 2,807,799.56$, $y_{\text{true}}(D') = 7,001.56$

$$\begin{aligned}\frac{\Pr[Y = y]}{\Pr[Y' = y]} &= \frac{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}}|y - 2,807,799.56|\right)}{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}}|y - 7,001.56|\right)} \\ &= \exp\left\{\varepsilon\left(|y - 7,001.56| - |y - 2,807,799.56|\right)\right\} \\ &\not\leq \exp(\varepsilon)\end{aligned}$$

How to make the term ≤ 1?

Rescale the Laplace distribution

How to make the term ≤ 1?

$$|y - 7,001.56| - |y - 2,807,799.56|$$

Recall: Triangle inequality

$a, b \in \mathbb{R}$: $|a| - |b| \leq |a - b|$

$$|y - 7,001.56| - |y - 2,807,799.56| \leq |(y - 7,001.56) - (y - 2,807,799.56)|$$

$$|y - 7,001.56| - |y - 2,807,799.56| \leq 2,800,798 \text{ (Alice's value!)}$$

$$\frac{|y - 7,001.56| - |y - 2,807,799.56|}{2,800,798} \leq 1$$

Now plug it back

Rescale the Laplace distribution

$$\begin{aligned}\frac{\Pr[Y=y]}{\Pr[Y'=y]} &= \frac{\frac{1}{2\frac{1}{\varepsilon}}\exp\left(-\frac{1}{\frac{1}{\varepsilon}}\frac{|y-2{,}807{,}799.56|}{2{,}800{,}798}\right)}{\frac{1}{2\frac{1}{\varepsilon}}\exp\left(-\frac{1}{\frac{1}{\varepsilon}}\frac{|y-7{,}001.56|}{2{,}800{,}798}\right)} \\ &= \exp\left\{\varepsilon\frac{(|y-7,001.56|-|y-2,807,799.56|)}{2,800,798}\right\} \\ &\leq \exp(\varepsilon)\end{aligned}$$

By scaling the Laplace distribution proportionally to the difference we achieved privacy (for this particular D and D' only!)

Will it work for the worst case scenario?

Remember: we said the maximum debt the bank offers per person is 10,000,000 €

We need to protect privacy even in the worst case, even when one person is a complete outlier

Name	Debt (in €)
Alice	0
Bob	0
Charlie	0
Xander	10,000,000

Our previous scale $2,800,798$ would not be enough

Let $D = \{A, B, C, X\}$, $D' = \{A, B, C\}$

$y_{\text{true}}(D) = 10,000,000$, $y_{\text{true}}(D') = 0$

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} = \frac{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}} \frac{|y-10,000,000|}{2,800,798}\right)}{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}} \frac{|y-0|}{2,800,798}\right)}$$

$$= \exp\left\{ \varepsilon \underbrace{\frac{(|y - 10,000,000| - |y - 0|)}{2,800,798}}_{\text{Exceeds 1 for some } y} \right\}$$

$$\not\leq \exp(\varepsilon)$$

Introducing global ℓ_1 sensitivity

Remember: we said the max debt/person = 10,000,000 €

We also saw that

- One person can change the sum query by the max value
- The max value was essential for scaling the Laplace distribution

Global ℓ_1 sensitivity

The ℓ_1-sensitivity of a function $f : \mathcal{X} \to \mathbb{R}^k$

$$\Delta = \max_{D,D'} \|f(D) - f(D')\|_1$$

for any neighboring datasets D, D'

Global ℓ_1 sensitivity

Global ℓ_1 sensitivity

The ℓ_1-sensitivity of a function $f : \mathcal{X} \to \mathbb{R}^k$

$$\Delta = \max_{D,D'} \|f(D) - f(D')\|_1$$

for any **possible** neighboring datasets D, D'

See local sensitivity by K. Nissim, S. Raskhodnikova, and A. Smith (2007). **"Smooth Sensitivity and Sampling in Private Data Analysis".** In: *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing - STOC '07*. San Diego, CA, USA: ACM Press, p. 75

ℓ_1-sensitivity of a function f captures the magnitude by which a single individual's data can change the function f in the worst case

Important: Global sensitivity is defined apriori by the domain (and is public information), not depending on actual data

Let's fix our previous example

Let $D = \{A, B, C, X\}$, $D' = \{A, B, C\}$

$y_{\text{true}}(D) = 10,000,000$, $y_{\text{true}}(D') = 0$

$$\frac{\Pr[Y = y]}{\Pr[Y' = y]} = \frac{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}} \frac{|y-10,000,000|}{\Delta}\right)}{\frac{1}{2\frac{1}{\varepsilon}} \exp\left(-\frac{1}{\frac{1}{\varepsilon}} \frac{|y-0|}{\Delta}\right)}$$

$$= \exp\left\{ \varepsilon \underbrace{\frac{(|y - 10,000,000| - |y - 0|)}{\Delta}}_{\text{Always} \leq 1} \right\} \leq \exp(\varepsilon)$$

Will work for all D, D'!

The correct scale of Laplace mechanism

Our previous observations (scale of the Laplace is proportional to the maximum difference of neighboring datasets, aka. global sensitivity) lead us to the following

Laplace mechanism for numeric queries

Draw $y_{\text{lap}} \sim \text{Lap}(\mu = 0; b = \frac{\Delta}{\varepsilon})$ and report $y = y_{\text{true}} + y_{\text{lap}}$

This correct Laplace mechanism is $(\varepsilon, 0)$-differentially private

Formal proof? We have all building blocks $\rightarrow$ homework! (or during exercise)

Laplace mechanism

Multi-dimensional numeric queries

Numerical query function might be $\mathcal{X} \mapsto \mathbb{R}^k$

Recall: Global ℓ_1 sensitivity

The ℓ_1-sensitivity of a function $f : \mathcal{X} \to \mathbb{R}^k$

$\Delta = \max_{D,D'} \|f(D) - f(D')\|_1$

for any neighboring datasets D, D'

For $\mathcal{X} \mapsto \mathbb{R}$ it is just an absolute value of max difference

For $\mathcal{X} \mapsto \mathbb{R}^k$ it **must** be ℓ_1 norm

ℓ_1 norm (city block distance, Manhattan distance)

For two vectors x and y in $\mathbb{R}^k$:

$\|x - y\|_1 = \sum_{i=1}^{k} |x_i - y_i|$

Laplace mechanism for $\mathcal{X} \mapsto \mathbb{R}^k$

Laplace mechanism for numerical queries

For each position $i \in \{1, \ldots, k\}$ in the output vector $\mathbf{y}$:

- Draw $\mathbf{y}_{\text{lap}}[i] \sim \text{Lap}(\mu = 0; b = \frac{\Delta}{\varepsilon})$
- Output $\mathbf{y}[i] = \mathbf{y}_{\text{true}}[i] + \mathbf{y}_{\text{lap}}[i]$

Important: The random variables (draws) for each i are **independent**

Formal proof? Very similar as before but now use the above fact of independence for the Laplace PDFs (homework/exercise)

Exponential mechanism

Motivation

Person	City
Alice	Los Angeles
Bob	New York
Charlie	Washington
Xander	Toronto

Table 4: The Secret Society

Given a known list of (all) cities, pick a city that is in the "middle" for the next meeting of the secret society

All cities C = {Ottawa, Toronto, New York, Washington, Memphis, Los Angeles, La Habana}

Motivation

City closest to the mean GPS coordinates – does it protect privacy of the members?

Example more concretely

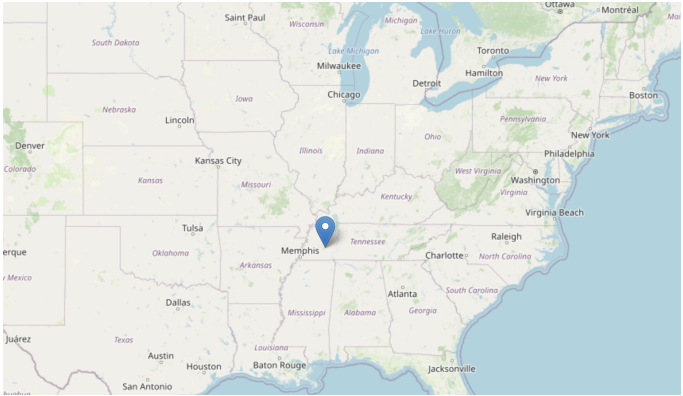
All cities C = {Ottawa (45, -75), Toronto, New York, Washington, Memphis (35, -90), Los Angeles, La Habana (23, -82)}

Person	City	(Latitude, Longitude)
Alice	Los Angeles	(34, -118)
Bob	New York	(40, -74)
Charlie	Washington	(38, -77)
Xander	Toronto	(43, -79)

Table 5: The Secret Society, database D

Mean = 36, -88, closest city = Memphis (35, -90)

Example more concretely



Mean = 36, -88, closest city = Memphis (35, -90)

Let's define a utility function

Based on: Distance (Euclid) of the database's D GPS mean to a known city c GPS

$$d(D, c)$$

(the smaller, the closer = better answer)

This is OK, but we want our utility u to be higher for closer cities, so

$$u(D, c) = \frac{1}{d(D, c)}$$

(ignoring zero in the denominator for simplicity)

Our query should return an element with highest utility

Utility u: $u(D, c) = \frac{1}{d(D,c)}$

Take the database D, compute the mean coordinates, and choose city $c \in C$ with the highest utility.

What happens for a neighboring dataset D'?

Does our utility function have sensitivity?

Utility formally: for a database and a range

$$u : \mathcal{X} \times \mathcal{R} \to \mathbb{R}$$

Sensitivity

$$\Delta u = \max_{r \in \mathcal{R}} |u(D, r) - u(D', r)|$$

for any neighboring datasets $D, D' \in \mathcal{X}$

How to choose the city privately?

What we wanted

Take the database D, compute the mean coordinates, and choose city $c \in C$ with the highest utility.

We want this process to be DP!

$\Pr[\mathcal{M}(D) = c] \leq \exp(\varepsilon) \Pr[\mathcal{M}(D') = c]$ for all neighboring datasets and any c

We need to sample c somehow randomly

Sample from a probability distribution over $\mathcal{R}$

The Exponential mechanism

$$\Pr[\mathcal{M}(D) = r] = \frac{\exp\left(\frac{\varepsilon \cdot u(D,r)}{2\Delta u}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{\varepsilon \cdot u(D,r')}{2\Delta u}\right)}$$

Homework: Implement the example

Exercise: Proof that this is $(\varepsilon, 0)$-DP

License and credits

Credits

Ivan Habernal

Content from ACL Anthology papers licensed under CC-BY
`https://www.aclweb.org/anthology`

Partly inspired by lectures from Antti Honkela, Aurélien Bellet, Gautam Kamath

Appendix: Proof for slide 16

$$\frac{\Pr[Y' = y]}{\Pr[Y = y]} = \frac{\frac{1}{2b}\exp\left(-\frac{1}{b}|y - 50|\right)}{\frac{1}{2b}\exp\left(-\frac{1}{b}|y - 51|\right)} = \exp\left(\frac{1}{b}|y - 51| - |y - 50|\right)$$

$$\leq \exp\left(\frac{1}{b}|(y - 51) - (y - 50)|\right)$$

$$= \exp\left(\frac{1}{b}|y - 51 - y + 50|\right)$$

$$= \exp\left(\frac{1}{b}|-1|\right) = \exp\left(\frac{1}{b}\right)$$