DEPARTMENT OF
NETWORKED SYSTEMS
AND SERVICES

Differential Privacy in Practice

WITSEC, Budapest, 2019

Szilvia Lestyán

CrySyS Lab, BME
lestyan@crysys.hu

www.crysys.hu

MŰEGYETEM 1782

What is privacy?

- **Privacy** is the right to private and family life, home and communications, to be autonomous, to be let alone
  - universal human right
- **Information privacy** is the right to *have some control* over how your personal information is used

Why is it important?

- Identity fraud
- Your data is valuable
- Profiling and surveillance
- Stigmatization, discrimination
- Freedom of thought and speech
- Everybody has something to hide

Why is it important?

- [illegible]
- [illegible]
- [illegible]
- St[illegible]
- Freed[illegible]
- Everybody ha[illegible]

GDPR

- Personal data:
  any information related to an **identified** or **identifiable** natural person

- (re-)identification is achieved through “identifiers”, which holds a particularly privileged and close relationship with the individual

Identifiers

- **Direct identifiers** unambiguously identify a person
  - “Prime Minister of Hungary in 2017”

Identifiers

- **Direct identifiers** unambiguously identify a person
  - "Prime Minister of Hungary in 2017"

- **Indirect (quasi-)-identifier** may ambiguously identify a person
  - "A prime minister in Europe"

Identifiers

- **Direct identifiers** unambiguously identify a person
  - "Prime Minister of Hungary in 2017"

- **Indirect (quasi-)-identifier** may ambiguously identify a person
  - "A prime minister in Europe"

"A prime minister in Europe"
+
"born on May 31, 1963"

Some identifiers

Direct identifiers	Indirect (quasi) identfiers
Full name	First name only
Date of birth	Last name only
Residential Address	A portion of address
Telephone number	Age
Email address	Place of work
Social Security number	IP address
Banking card number	Device Id
ID number	Gender
Passport number	Visited locations

- *GDPR refers to all personal data as identifiers which together unambiguously identify a person in the givencontext*

Identifiable

- A person is identifiable:

"To determine whether a natural person is identifiable, account should be taken of all the means ***reasonably likely*** *to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. "*

What does it mean?

- A person is identifiable, if:
  - **Plausible attack:** The attacker has enough motivation to launch the attack…
  - **Reasonable chance of succeeding:** the success probability of the attack is high enough
- There are ***NO explicit pre-defined thresholds*** of plausibility and reasonable chance in GDPR, as it is context-dependent

Example

- A hospital in Michigan publishes a medical dataset with 3 attributes: (1) ZIP, (2) Age, (3) Sex, (4) Diagnosis
- **It is personal data** according to the GDPR
  - Microdata (individuals are identified)

Name	Zipcode	Age	Sex	Disease
Alice S.	47677	29	F	Ovarian Cancer
Betty Q.	47602	22	F	Ovarian Cancer
Charles D.	47678	27	M	Prostate Cancer
David E.	47905	43	M	Flu
Emily J.	47909	52	F	Heart Disease
	47906	47	M	Heart Disease

Removing names?

Example

- Re-identification attack:
  1. Purchase the voter registration data for $10

Microdata

Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Example

- Re-identification attack:
  1. Purchase the voter registration data for $10
  2. Associate a voter record with the corresponding medical record (along with matching ZIP, Age, Sex attributes)

Microdata

Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Example

- Re-identification attack:
  1. Purchase the voter registration data for $10
  2. Associate a voter record with the corresponding medical record (along with matching ZIP, Age, Sex attributes)
- Success probability of linking: 63%*

Microdata

Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Example II. – Query Auditing

- Given a database with some disclosure policy
  - HIV is private, but aggregated values of HIV records may be available
    - e.g. SUM, COUNT, MEDIAN, MAX

Name	Sex	ZIP	Blood sugar	HIV
John S.	Male	1123	4.3	True
John D.	Male	1123	5.2	False
Jerry K.	Male	1114	6.1	True
Jack. D.	Male	8423	3.2	False
Eve A.	Female	1234	7.1	True

Example II. – Query Auditing

solve a system of linear equations

- SUM(HIV) WHERE ZIP < 8000
  - $x_1 + x_2 + x_3 + x_4 = 2$
- SUM(HIV) WHERE ZIP = 1123
  - $x_1 + x_2 = 1$
- SUM(HIV) WHERE ZIP > 1200
  - $x_4 + x_5 = 1$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

where $x_i \in \{0, 1\}$ for all i

Example II. – Query Auditing

solve a system of linear equations

- SUM(HIV) WHERE Z
  - $x_1 + x_2 + x_3 +$
- SUM(HIV) WH
  - $x_1 + x_2$
- SUM(HIV)
  - x_4

MATHS MAGIC

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

where $x_i \in \{0, 1\}$ for all i

Example II. – Query Auditing

solve a system of linear equations

- SUM(HIV) WHERE Z
  - $x_1 + x_2 + x_3 +$
- SUM(HIV) WH
  - $x_1 + x_2$
- SUM(HIV)
  - $x_4 +$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & \\ 0 & 0 & 1 & 1 & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

$\in \{0, 1\}$ for all i

MATHS MAGI

PRIVACY BREACH

Philosophy of Differential privacy

- **Absolute (Perfect) privacy:** access to the published data should not enable the adversary to learn ***anything*** extra about any individual compared to no access to the data

Philosophy of Differential privacy

- **Absolute (Perfect) privacy:** access to the published data should not enable the adversary to learn ***anything*** extra about any individual compared to no access to the data

There is **always** some background knowledge which allows absolute privacy breach

Philosophy of Differential privacy

- AVG = 175 cm
- John is +10

Privacy breach?

PRIVACY BREACH!

(in absolute sense)

- John has cancer with 50% and smokes

- Study shows smoking causes cancer in 92%

BUT!

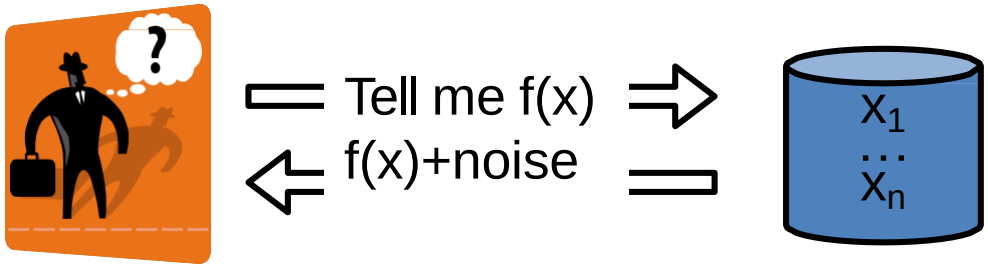
Does removing John from the database cause any change in the outcome?

Philosophy
Differential privacy

- The outcome of the anonymization scheme should be more or less independent of the value of a single record
  - *It does not "leak" information about any* ***single record***
  - *It can formally be proven!*

- Differential privacy aims to hide only those information that is ***specific to John*** (or any single individual in the dataset)

Differential privacy



A privacy mechanism $\mathcal{S}$ gives ε-differential privacy if for any database D and D' differing in at most one record, and for any possible output $O \in Range(\mathcal{S})$,

$$e^{-\varepsilon} \leq \frac{Pr[\mathcal{S}(D) = O]}{Pr[\mathcal{S}(D') = O]} \leq e^{\varepsilon}$$

where the probability is taken over the randomness of $\mathcal{S}$.

GDPR vs. Differential Privacy

- **DP implies GDPR** (not vica versa)
- Identifiers vs noise
- In GDPR DP is a method
- The confidence of ALL inferences are bounded
- Formally provable!
- Apple, Uber, Google

- **Aggregates vs Microdata**?

- ”Syntactic” guarantees are
  - *Not sufficient in practice*
  - Does not defend from future attacks
  - Not formally provable

Conclusions

So what to use?

Conclusions

So what to use?

Experts!

DEPARTMENT OF
NETWORKED SYSTEMS
AND SERVICES

Thank you!

Szilvia Lestyán

CrySyS Lab, BME
lestyan@crysys.hu

www.crysys.hu

MŰEGYETEM 1782