"You have zero Privacy. Get over it."

Scott McNealy, 1999

Maintaining Privacy in a World of services - revisited

Prof. Johann-Christoph Freytag, Ph.D.

Datenbanken und Infomationssysteme (DBIS)
Institut für Informatik (CS Department)
Humboldt-Universität zu Berlin
freytag@dbis.informatik.hu-berlin.de

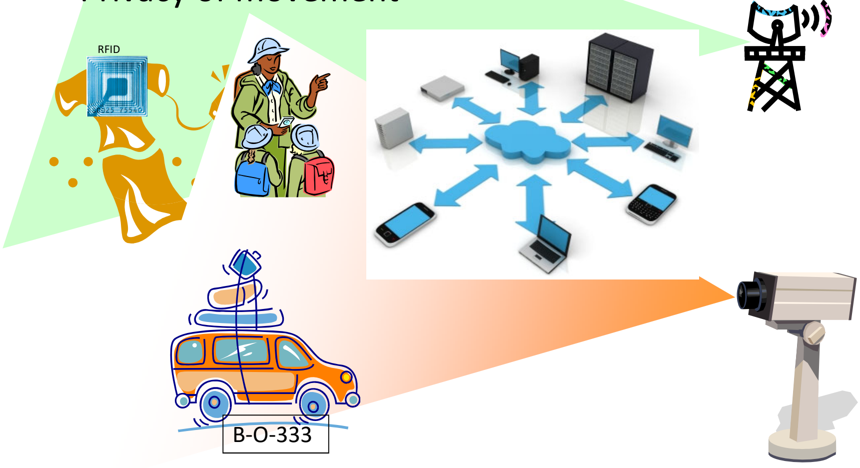
Overview

- What's the problem with privacy?
- Privacy & services
- Brief intro to k-anonymity
  - other concepts building on k-anonymity
  - Queries and what you learn…..
- Using differential privacy – DP & LDP
  - What is it
  - What's different
  - Where used

What‘s the Problem with Privacy??

Privacy violation ...

- Privacy of movement



Sensitive and Personal Data/Information

- Sensitive Information (slightly changed)

  information which through loss, or misuse, or unauthorized access to, or modification of which could adversely affect the interests of groups, organizations (such as the government or businesses), or the privacy to which individuals are entitled to by national or international law.

- Personal (private) data/information

  ***Personal data** is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data.*

  *Personal data that has been de-identified, encrypted or pseudonymised **but can be used to re-identify a person remains personal data** and falls within the scope of the GDPR.*

FEDSTD-1037C

2018 European General Data Protection Regulation (GDPR)

SummerSoc 2019 - Privacy Talk 6

Main changes

Scope of the General Data Protection Regulation (GDPR)

What will change against the former 1995 EU Data Protection Directive ?

Broader territorial scope	Applies to players **not established in the EU** but whose activities consist of targeting data subjects in the EU
Enforcement	Data Protection Authorities will be entitled to impose **fines ranging between 2 to 4% of annual turnover** and **increased powers**
Accountability	Explicit obligation on controller AND processor to **demonstrate their GDPR compliance (eg data protection officer, privacy impact assessments (PIA))**
Expanded definitions	Personal data now **explicitly includes location data, IP addresses, online and technology identifiers**;
Data subjects rights	**Reinforced rights**: Access, rectification, restriction, erasure, objection to processing; no automated processing and profiling, data portability
Consent	**Spelled out more clearly** and focus on ability of individuals to distinguish a consent. Need for affirmative action
Data breach notification	Report a personal data breach to the Data Protection Authority within 72h...
One-stop shop	**Data Protection Authorities (DPA) of main establishment can act as lead DPA**, supervising processing activities throughout the EU
International data transfers	Binding Corporate Rules as tools for data transfers outside the EU and EEA are now embedded in law

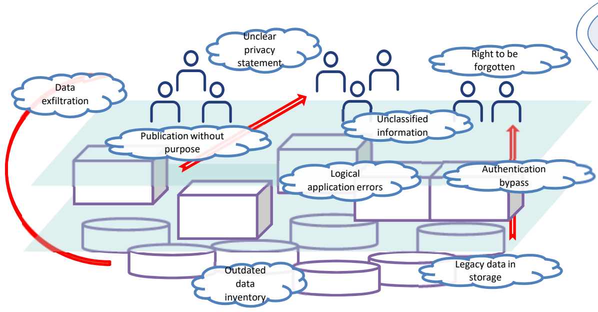
Source: https://www2.deloitte.com/

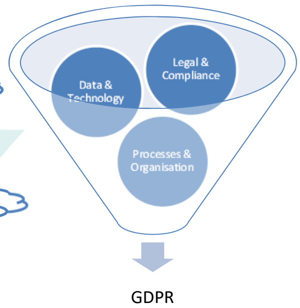
SummerSoc 2019 - Privacy Talk

GDPR - Holistic approach

GDPR is not only about legal aspects of data protection
GDPR is not only about technical aspects of data protection

GDPR calls for a combined approach





Source: https://www2.deloitte.com/

What is Privacy in the context of DBMS?

- **Definition 1:** [Sweeney, 2002]
  "**Privacy** reflects the ability of a person, organization, government, or entity to control its own space, where the concept of space (or "privacy space") takes on different contexts."
  – Physical space, against invasion
  – Bodily space, medical consent
  – Computer space, spam
  – Web browsing space, Internet privacy

- **Definition 2:** [Agrawal et al., 2002]
  "**Privacy** is the right of individuals to determine for themselves when, how, and to what extent information about them is communicated to others."
  (We shall call this data/information privacy)

(data) security vs. (data) privacy

Data security

≠

Data Privacy

(data) security vs. (data) privacy

- **Data security comprises of all means, techniques, and approaches to protect data from destructive forces and unwanted actions of non-authorized users.**
- **Data privacy comprises of all means, techniques, and approaches to secure the rights of individual to determine for themselves when, how, and to what extend to share data about themselves with others."**
  - **Definition holds for both analog and digital data**
  - **Data privacy implies data security**
  - **Protecting (data) privacy is necessary**
    - **Personal data is shared with third parties**
    - **At the same time guaranteeing/protecting the privacy of the person described (for example by protecting his/her identity).**

Is it always obvious when privacy is violated?

- Is it always obvious that privacy is violated or breached?
- Sweeney's Finding [Sweeney, 2002]
  - In Massachusetts, USA, the *Group Insurance Commission* (GIC) is responsible for purchasing health insurance for state employees
  - GIC has to publish the data:

GIC					
ZIP	Date of birth	Sex	Diagnostic	Medication	…

http://lab.privacy.cs.cmu.edu/people/sweeney/

Sweeney's Finding (1)

- Sweeney paid $20 to buy the voter registration list for Cambridge, MA:

Voter					
Name	Address	…	**ZIP**	Date of birth	**Sex**

GIC					
ZIP	Date of birth	**Sex**	Diagnostic	Medication	…

- William Weld (former governor) lives in Cambridge, hence is in VOTER
- 6 people in VOTER share his **date of birth**
- only 3 of them were man (same **sex**)
- Weld was the only one in that **zip**
- Sweeney learned Weld's medical records!
- 87 % of population in U. S. can be identified by ZIP, dob, sex

Sweeney's Finding (2)

- **Observation:** *All systems worked as specified, yet an important data has leaked*
  - "Information leakage" occurred
  - Despite the observation that all "participating sites" worked as specified
  - Beyond correctness!
  - What's missing/causing the problem?
- How do we protect against this kind of "lack (leakage) of privacy"?

Privacy-Preserving Data Publishing Challenge

- Objective
  - Publish privacy-relevant data
    - e.g., personal data
  - Preserve privacy of data subjects
    - e.g., individuals
- Purpose
  - e.g., statistic analyzes, legal regulations
- Challenge
  - **Given**
    - privacy-relevant data in microdata table *T*
      - attribute types: **identifying**, **sensitive**, **other**
  - **Goal**
    - generate privacy-preserving public release table *T**
      - information should remain practically useful

Name	Zipcode	Age	Sex	Disease
Alison	10000	18	F	Asthma
Ben	11000	19	M	Bronchitis
Clark	12000	20	M	Cold
Debra	12000	21	F	Diabetes
Elaine	12000	22	F	Earache
Fiona	12000	23	F	Flu
Gary	14000	24	M	Earache

Microdata table *T*

Privacy-Preserving Data Publishing
Insufficient Approach

- Insufficient approach
  - remove only identifying attributes
- Problem
  - set of other attributes could be used to identify individuals
    - call these attributes quasi-identifier
- Example
  - combination of Zipcode, Age, Sex is unique
  - with help of external data (e.g., voter list) identify individuals

Name	Zipcode	Age	Sex	Disease
Alison	10000	18	F	Asthma
Ben	11000	19	M	Bronchitis
Clark	12000	20	M	Cold
Debra	12000	21	F	Diabetes
Elaine	12000	22	F	Earache
Fiona	12000	23	F	Flu
Gary	14000	24	M	Earache

Microdata table *T*

Zipcode	Age	Sex	Disease
10000	18	F	Asthma
11000	19	M	Bronchitis
12000	20	M	Cold
12000	21	F	Diabetes
12000	22	F	Earache
12000	23	F	Flu
14000	24	M	Earache

Insufficient release table *T**

Name	Zipcode	Age	Sex
Alison	10000	18	F
Ben	11000	19	M

external data

Name	Disease
Alison	Asthma
Ben	Bronchitis

ϟ

Privacy-Preserving Data Publishing
Improved Approach

- Improved Approach
  - remove identifying attributes
  + generalize quasi-identifier
    - replace value with a less specific but semantically consistent value
- *k*-anonymity
  - for each tuple there exist *k*–1 other tuples which share the same values for all quasi-identifiers

Name	Zipcode	Age	Sex	Disease
Alison	10000	18	F	Asthma
Ben	11000	19	M	Bronchitis
Clark	12000	20	M	Cold
Debra	12000	21	F	Diabetes
Elaine	12000	22	F	Earache
Fiona	12000	23	F	Flu
Gary	14000	24	M	Earache

Microdata table *T*

Zipcode	Age	Sex	Disease
10–12000	18–20	*	Asthma
10–12000	18–20	*	Bronchitis
10–12000	18–20	*	Cold
12–14000	21–24	*	Diabetes
12–14000	21–24	*	Earache
12–14000	21–24	*	Flu
12–14000	21–24	*	Earache

3-anonymous release table *T**

Name	Zipcode	Age	Sex
Alison	10000	18	F

external data

Name	Disease
Alison	Asthma
Alison	Bronchitis
Alison	Cold

Privacy-Preserving Data Publishing
Better Approach

- Problem
  - tuples in QI-group with same sensitive value
    - QI-group: set of tuples with *same values for all quasi-identifiers*
- Better Approach
  - Restrict sensitive values in each QI-group
    - e.g., *distinct l-diversity*: ≥ *l* distinct sensitive values
    - many other approaches

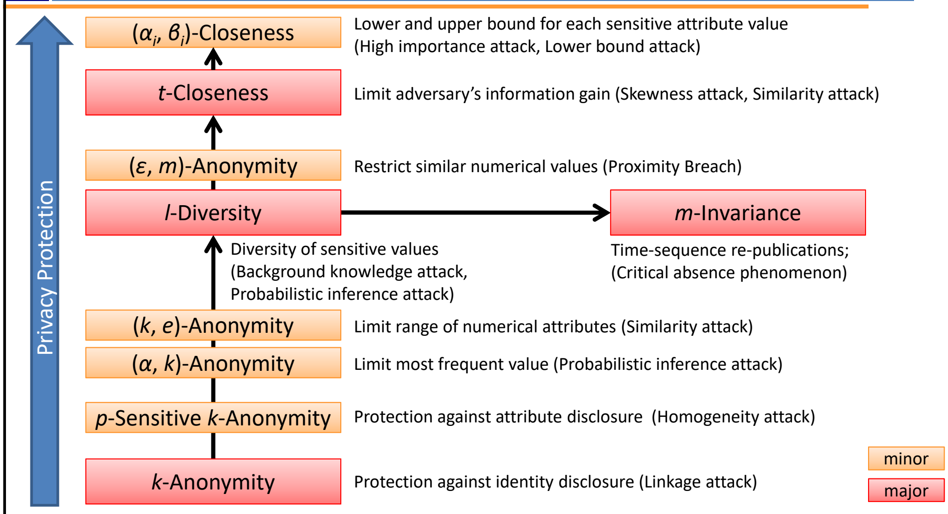
Name	Zipcode	Age	Sex	Disease
Alison	10000	18	F	Asthma
Ben	11000	19	M	Bronchitis
Elaine	12000	22	F	Earache
Gary	14000	24	M	Earache

Microdata table *T*

Zipcode	Age	Sex	Disease	QI-groups
10–12000	18–20	*	Asthma	2-anonymous ✓
10–12000	18–20	*	Bronchitis	distinct 2-divers ✓
12–14000	21–24	*	Earache	2-anonymous ✓
12–14000	21–24	*	Earache	distinct 2-divers ↯

Release table *T**

Anonymization Methods
Overview



What you learn from queries on anonymized data

Work done with my form student Lukas Dölle

Privacy-Preserving Request (Query) Processing
Scenario

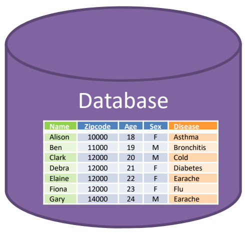
User
(potential adversary)

`SELECT … FROM T …`

Requests Q_1, Q_2, …

(anonymous) Results
R_1, R_2, …

Zipcode	Age	Sex	Disease
10–12000	18–20	*	Asthma
10–12000	18–20	*	Bronchitis
10–12000	18–20	*	Cold



His knowledge: R_1, R_2, …

Goal: Combination of user knowledge (R_1, R_2, …) **comply with privacy criteria** (e.g., distinct *l*-diversity)

Example

Name	Age	Disease
Alison	18	Asthma
Ben	19	Bronchitis
Clark	20	Cold
Debra	21	Diabetes
Elaine	22	Earache
Fiona	23	Flu
Gary	24	Earache

Microdata table *T*

```
SELECT Age, Disease
FROM T …
```

Q_1: … `WHERE Age BETWEEN 18 AND 20`

Age	Disease
18–20	Asthma
18–20	Bronchitis
18–20	Cold

R_1: distinct 3-divers

Q_2: … `WHERE Age BETWEEN 20 AND 23`

Age	Disease
20–23	Cold
20–23	Diabetes
20–23	Earache
20–23	Flu

R_2: distinct 4-divers

Q_3: … `WHERE Age BETWEEN 22 AND 24`

Age	Disease
22–24	Earache
22–24	Flu
22–24	Earache

R_3: distinct 2-divers

Example
Reasoning

Name	Age	Disease
Alison	18	Asthma
Ben	19	Bronchitis
Clark	20	Cold
Debra	21	Diabetes
Elaine	22	Earache
Fiona	23	Flu
Gary	24	Earache

Microdata table *T*

Q_1

Age	Disease
18–20	Asthma
18–20	Bronchitis
18–20	Cold

R_1: distinct 3-divers

Q_2

Age	Disease
20–23	Cold
20–23	Diabetes
20–23	Earache
20–23	Flu

R_2: distinct 4-divers

Q_3

Age	Disease
22–24	Earache
22–24	Flu
22–24	Earache

R_3: distinct 2-divers

Conclusion 1: Clark – Cold

Conclusion 2: Gary – Earache

Knowledge of adversary
- Anonymous results of queries (R_i)
- Quasi-identifier values of all tuples in *T*

Adversary wants to link individuals to sensitive attribute values

Clark

If an adversary knows that Clark is 20 years old, then he concludes:
- tuple for Clark in R_1
- tuple for Clark in R_2
- only one sensitive value in R_1 and R_2: Cold

Gary

If an adversary knows that Gary is 24 years old, then he concludes:
- tuple for Gary in R_3
- sensitive values in R_3: Earache, Flu
- assume Gary-Flu
- → in R_3: Elaine-Earache + Fiona-Earache
- → in R_2: 2 × Earache ↯

Query Graph
1st Query

ID	SA
1	A
2	B
3	C
4	D
5	E
6	F
7	E

Data

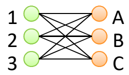
Query

Q_1

ID	SA
1, 2, 3	A, B, C

Graph

G_1

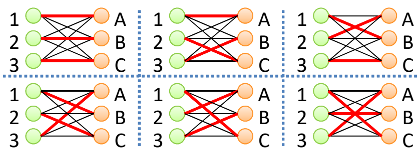


- Query graph $G_1 = (V_1, E_1)$
  - model query/result as graph
    - V_1: vertex for each tupel (ID) and each SA value
    - E_1: edges between tuple and SA vertices
- G_1 is bipartite

Assignments

ID	1	2	3	4	5	6
1	A	A	B	B	C	C
2	B	C	A	C	A	B
3	C	B	C	A	B	A

Perfect Matchings



- Each value assignment
  - = one perfect matching in G_1
- matching := set of edges without common vertices
- **perfect** := each vertex in one edge

Result (PhD by Lukas Dölle)

- We can detect when k-anonymity (or other privacy criteria) are violated
  - In polynomial time only for a limited case
    - Can be nicely characterized by ring structure
- Algorithm simplifies
  - When no duplicates are present

Differential Privacy

https://www.seas.harvard.edu/directory/dwork

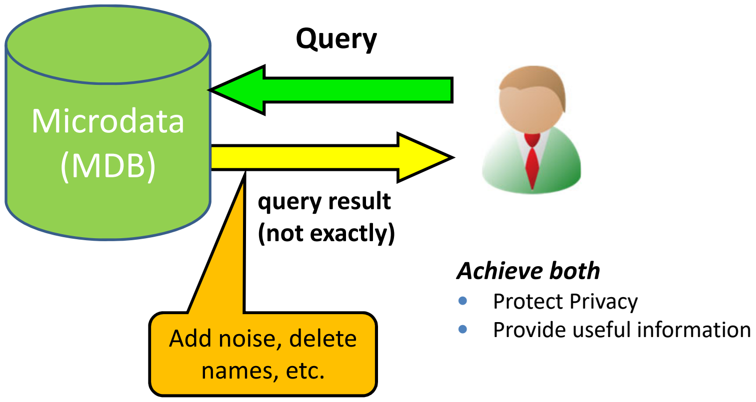
Motivated by Netflix problem in 2009

- Netflix Recommends Movies to its Subscribers
  - Offers $1,000,000 for 10% improvement in its recommendation system
    - Not concerned here with how this is measured
- Solve, see [here]

The Netflix Prize

- Netflix Recommends Movies to its Subscribers (cont.)
  - Publishes training data
    - Nearly 500,000 records, 18,000 movie titles
    - "The ratings are on a scale from 1 to 5 (integral) stars. To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids have been replaced by randomly-assigned ids. The date of each rating and the title and year of release for each movie are provided."
    - Some ratings not sensitive, some may be sensitive
      - OK for Netflix to know, not OK for public to know
- Despite all efforts scientists developed a probabilistic algorithm for re-identification
  - With small amount of background knowledge on the individual
  - See https://arxiv.org/PS_cache/cs/pdf/0610/0610105v2.pdf

Sanitization of Databases



Differential Privacy (informal)

- Output of a query is similar whether any single individual's record is included in the database or not

Query: # of persons with a cold?

Database D

Name	Disease
Chris	Arthritis
David	Cold
Ethan	Heart problem

Query → R1 ≈ R2 ← Query

Database D'

Name	Disease
Chris	Arthritis
Ethan	Heart problem

- David is no worse off because his record is/is not included in the output of a query

Basic Definitions

Definition 1:

Two databases D, D' are **neighbors** if they differ by at most one tuple

Definition 2:

A randomized algorithm G provides **ε-differential privacy** if:

- for all neighboring databases D and D', and
- for any outputs O:

$$\Pr[G(D) = O] \leq e^{\varepsilon} * \Pr[G(D') = O]$$

Differential Privacy – additional remarks

- $\Pr[G(D) = O] \leq e^{\varepsilon} * \Pr[G(D') = O]$

$$= \frac{\Pr[G(D) = O]}{\Pr[G(D') = O]} \leq e^{\varepsilon} \approx 1 \pm \varepsilon$$

ε is a privacy parameter

- Epsilon is usually small: e.g. if $\varepsilon = 0.1$ then $e^{\varepsilon} \approx 1.10$

⇩ epsilon = ⇧ stronger privacy

Query sensitivity

Definition 3: The **sensitivity** of a query Q is

$\Delta q = \max |Q(D) - Q(D')|$

where D, D' are any two neighboring databases

Query Q	Sensitivity Δq
Q1: Count tuples	1
Q2: Count (patients with "Cold")	1
Q3: Count (patients with property X)	1
Q4: Max (age of patients)	max age

Differential privacy

[Dwork, ICALP06]

- How to add noise: ***Laplace distribution***

$$\Pr[\eta = x] = \frac{1}{2\lambda}\, e^{-|x-\mu|/\lambda}$$

- with
  - μ is the mean of the distribution (usually $\mu = 0$)
  - λ (referred to as the noise scale) is a parameter that controls the degree of privacy protection
  - $\lambda = \Delta q / \varepsilon$,
    i.e. sensitivity (of query) / strength of protection

Calibrate Noise & Sensitivity (1)

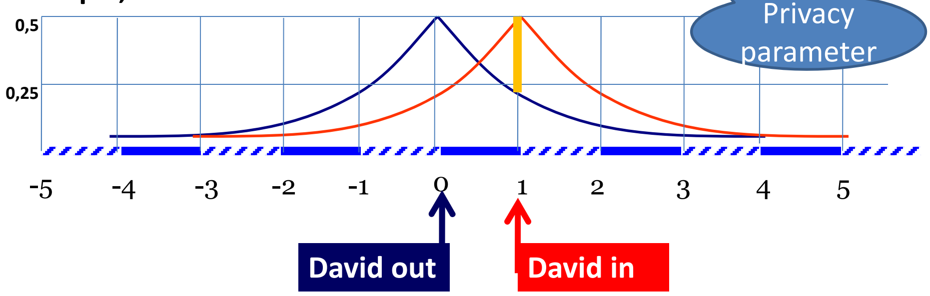
- Example 1:

Sensitivity

Q(D) + Laplace(Δq / ε)

Privacy parameter

Δq=1, ε=1.0



Calibrate Noise & Sensitivity (2)

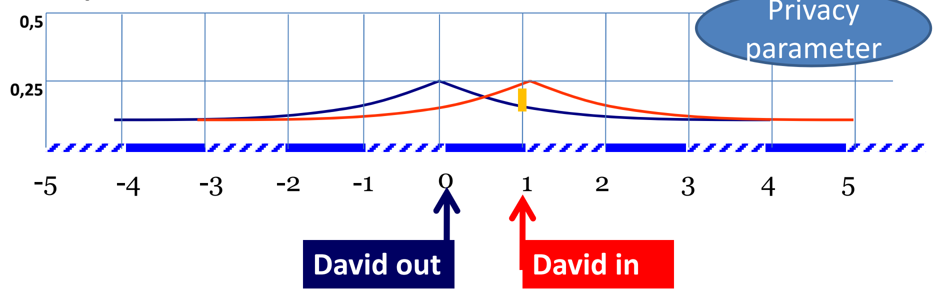
- Example 2:

Q(D) + Laplace(Δq / ε)

Sensitivity

Privacy parameter

Δq=1, ε=0,5



Differentially private algorithms

- Any (*statistical*) query can be answered (but perhaps with lots of noise)
- Noise determined by privacy parameter epsilon and the sensitivity (both public)
  - Increasing Δq/ε flattens curve; more privacy
  - *Noise depends on Δq and ε, not on the database*
- Privacy guarantee does not depend on assumptions about the adversary (caveats omitted, see **[Kifer, SIGMOD 11]**)
- Survey paper on differential privacy: **[Dwork, CACM 11]**

Multiple Queries

- For query sequence $Q_1, \ldots, Q_d$ ε-privacy achieved with increasing noise for each response
- Naively, more queries mean noisier answers
- Noise must increase with the sensitivity of the query sequence
- Problem of Non-Interactive Setting
  - Any non-interactive solution permitting "too accurate" answers to "too many" questions is vulnerable to privacy attack.
- Dinur Nissim Result:
  - A *vast majority of records* in a database of size n can be reconstructed when $n \log(n)^2$ queries are answered by the database …

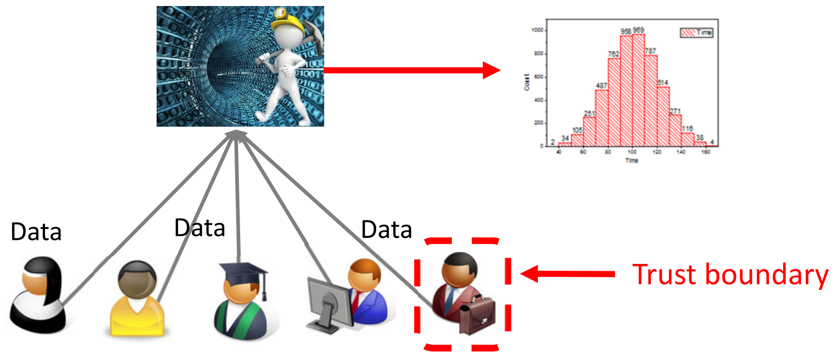
SummerSoc 2019 - Privacy Talk 39

Local Differential Privacy (LDP)

Based on tutorial: Privacy at Scale: Local Differential Privacy in Practice,
(Cormode, Jha, Kulkarni, Li, Srivastava, and Wang) Sigmod 2018, Houston, TX

SummerSoc 2019 - Privacy Talk 40

Local Differential Privacy - Model



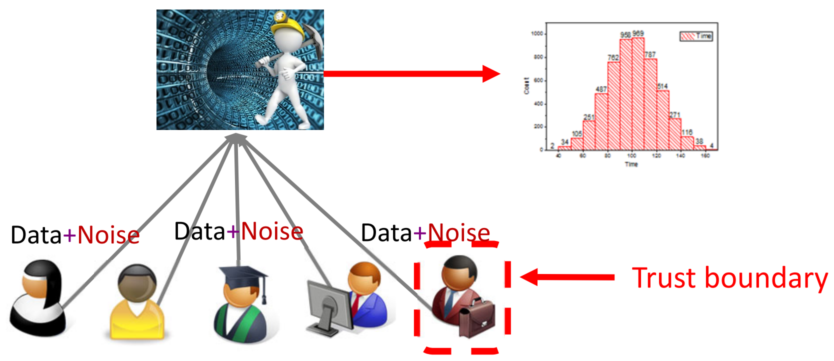
Trying to Reduce Trust

- Most work on differential privacy assumes a trusted party
  - Data aggregator (e.g., organizations) that sees the true, raw data
  - Can compute exact query answers, then perturb for privacy
- A reasonable question: can we reduce the amount of trust?
  - Can we remove the trusted party from the equation?
  - Users produce ***locally private output***, to answer aggregate queries
- One approach is to use homomorphic encryption
  - Merge encrypted data, and add noise for privacy inside encryption
  - Complex to get right, and very *high computational overhead*

Local Differential Privacy

- What about having each user run a DP algorithm on their data?
  - Then combine all the results to get a final answer
- On first glance, this idea seems crazy
  - Each user adds noise to mask their own input
  - So surely the noise will always overwhelm the signal?
- But … noise can cancel out or be subtracted out
  - We end up with the true answer, plus noise which can be smaller
  - However, noise is still larger than in the centralized case

Local Differential Privacy - Model



From DP to LDP: Formal Definition

Idea of DP: Any output should be about as likely regardless of whether or not I am in the dataset

A randomized algorithm $\boldsymbol{A}$ satisfies $\boldsymbol{\varepsilon}$-differential privacy, iff for any two neighboring datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$ and for any output $\boldsymbol{O}$ of $\boldsymbol{A}$,

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

A randomized algorithm $\boldsymbol{A}$ satisfies $\boldsymbol{\varepsilon}$-local differential privacy, iff for any two inputs $\boldsymbol{x}$ and $\boldsymbol{x'}$ and for any output $\boldsymbol{y}$ of $\boldsymbol{A}$,

$$\Pr[\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{y}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{x'}) = \boldsymbol{y}]$$

Run by ... person

ε is also called privacy budget
Smaller ε → stronger privacy

Idea of LDP: Any output should be about as likely regardless of my secret

Local Differential Privacy: Example

- Each of N users has 0/1 value, estimate total population sum
  - Each user adds independent Laplace noise (DP): mean 0, variance $2/\varepsilon^2$
- Adding user results: true answer + sum of N Laplace distribution values
  - Error is random variable, with mean 0, variance $2N/\varepsilon^2$
  - Confidence bounds: ~95% chance of being within 2σ of the mean
  - So error looks like $\sqrt{N}/\varepsilon$
- Numeric example: suppose true answer is $N/2$, $\varepsilon = 1$, $N = 10^6$
  - We see 500K ± 2800 : about 1% uncertainty
  - Error in centralized case would be close to 500K ± 1 (0.001%)

Local Differential Privacy

- We can achieve LDP, and obtain reasonable accuracy (for large N)
  - The error typically scales with √N
- Generic approach: apply centralized DP algorithm to local data
  - But error might still be quite large
  - Unclear how to merge private outputs (e.g. private clustering)
- So we seek to design new LDP algorithms
  - Maximize the accuracy of the results
  - Minimize the costs to the users (space, time, communication)
  - Ensure that there is an accurate algorithm for aggregation

Properties of (Centralized) DP

A randomized algorithm $\boldsymbol{A}$ satisfies $\boldsymbol{\varepsilon}$-differential privacy, iff for any two neighboring datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$ and for any output $\boldsymbol{O}$ of $\boldsymbol{A}$,

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

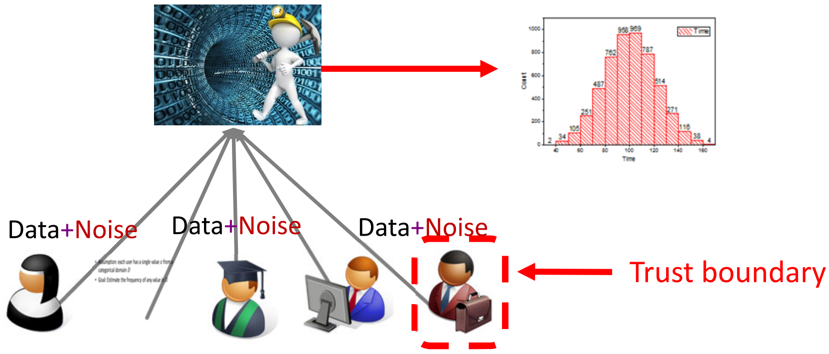
- Post-processing (of the output) is free
  - does not consume privacy budget

What about LDP?

- Parallel composition
  - partition the dataset into subsets, each applying an ε_i-DP algorithm, the overall result satisfies $\max(\varepsilon_i)$-DP
- Sequential composition
  - apply k DP algorithms, each using ε_i, result satisfies $\sum \varepsilon_i$-DP

Frequency Estimation

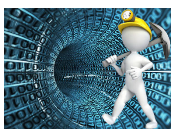
- Assumption: each user has a single value x from a categorical domain D
- Goal: Estimate the frequency of any value in D



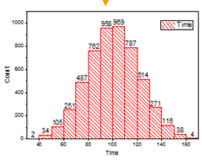
Frequency Oracle Framework

- $x := E(v)$ takes input value v from domain D and outputs an encoded value x
- $y := P(x)$ takes an encoded value x and outputs y.

y



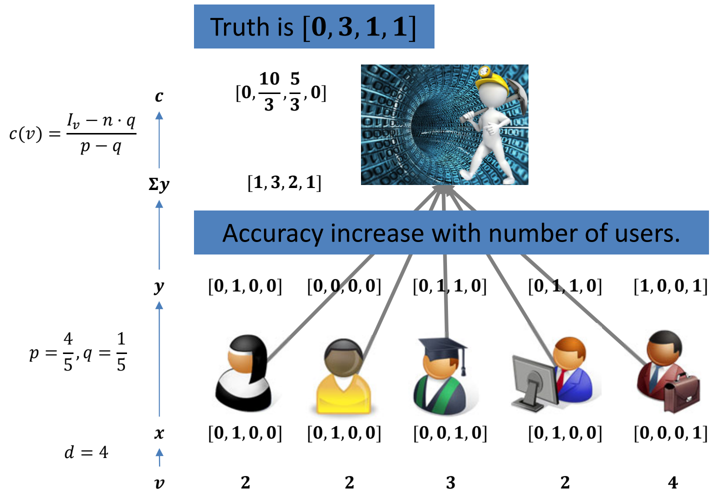
- $c := Est(\{y\})$ takes reports $\{y\}$ from all users and outputs estimations $c(v)$ for any value v in domain D



P is ε -LDP iff for any v and v' from D, and any valid output y,

$$\frac{\Pr[P(E(v))=y]}{\Pr[P(E(v'))=y]} \leq e^{\varepsilon}$$

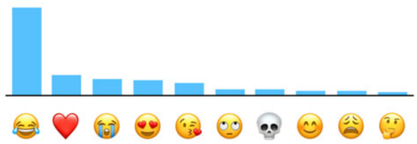
Frequency Oracle Framework (Example)



Privacy in practice

- Differential privacy based on coin tossing is widely deployed!
  - In Google Chrome browser, to collect browsing statistics
  - In Apple iOS and MacOS, to collect typing statistics
  - In Microsoft Windows to collect telemetry data over time
  - From Snap to perform modeling of user preference
  - This yields deployments of over 100 million users each
- All deployments are based on **R**andom **R**esponse (RR), but extend it substantially
  - To handle the large space of possible values a user might have
  - Randomized response invented in 1965: five decades ago!

Apple’s Differential Privacy in Practice



The Count Mean Sketch technique allows Apple to determine the most popular emoji to help design better ways to find and use our favorite emoji. The top emoji for US English speakers contained some surprising favorites.

- Apple uses their system to collect data from iOS and OS X users
  - Popular emojis: (heart) (laugh) (smile) (crying) (sadface)
  - “New” words: bruh, hun, bae, tryna, despacito, mayweather

Privacy for/in DBMS

- Several „add-ons“ exist:
  - Diffix by Aircloak (Germany)
  - PINQ (Microsoft prototype)
    - Extends the programming language interface
  - SAP HANA DA
    - k-anonymity & LDP (local differential privacy) (April 2018)
    - l-diversity (April 2019)
    - Industrial paper with more details will appear at VLDB2019

Questions???

Thank you!!