Privacy

Privacy Techniques

Guillaume Raschia — Nantes Université
Last update: January 24, 2025

original slides from J. Near (Univ. of Vermont, CS211: Data Privacy)

An Overview of Privacy Techniques

Technique	Setting
Anonymization	Synthetic data
SDC	Synthetic data
k-Anonymity	Synthetic data
ℓ-Diversity	Synthetic data
Differential Privacy	Query answering

Synthetic Data vs. Query Answering

Synthetic data *looks like* the original **microdata**

Name	DOB	Gender	Zip
Rashad Arnold	02/26/2010	M	73909
Alyssa Cherry	05/08/2010	M	14890
Myra Ford	05/11/2010	NB	73821
Meredith Perry	03/31/2011	F	73909
Aimee Thornton	04/26/2010	F	14825

$\Longrightarrow$

DOB	Gender	Zip
2011	F	73***
2010	NB	73***
2010	M	73***
2010	F	14***
2010	M	14***

Query Answering

Query answering is an **interactive setting**

Name	DOB	Gender	Zip
Rashad Arnold	02/26/2010	M	73909
Alyssa Cherry	05/08/2010	M	14890
Myra Ford	05/11/2010	NB	73821
Meredith Perry	03/31/2011	F	73909
Aimee Thornton	04/26/2010	F	14825

- Q: *How many people were born in 2010?* A: **4**
- Q: *Are all males in the same neighborhood?* A: **No**
- Q: ...

Synthetic Data vs. Query Answering

Synthetic data

- Allows re-using existing data analyses (e.g. DBMS)
- One approach works for all query workloads (no advance knowledge of workload required)
- Makes things easier for the analyst
- **Impossible** to achieve perfect utility and strong privacy

Query answering

- Exact opposite of "Synthetic data pros & cons"
- Specialization to *one query* enables better **utility/privacy trade-off**

Outline

What does Utility Mean?

Informally	Formally
"how useful is the answer?"	depends on what the answer will be used for

"how many people have the last name Ford?"

- Anonymized data $\longrightarrow$ impossible to answer
- Differential privacy $\longrightarrow$ can answer $\pm$ 1 person

More examples

- For numerical queries, how different is the "private" answer from the "true" answer?
- For machine learning, what is the difference in testing error between "private" and "non-private" models?

Goals of De-identification

- De-identification removes the association between a person and a dataset, altering **identifying information**
- Goals:
  - Reduce the risk of privacy violation
  - Maximize data utility
- Techniques include:
  - Suppression (remove the id's)
  - Variation (scramble the id's)
  - Data swapping
  - Masking

De-identification: Examples

suppression

DOB	Gender	Zip
02/26/2010	M	73909
05/08/2010	M	14890
05/11/2010	NB	73821
03/31/2011	F	73909
04/26/2010	F	14825

swapping

Name	DOB	Gender	Zip
Alyssa Cherry	02/26/2010	M	73909
Meredith Perry	05/08/2010	M	14890
Aimee Thornton	05/11/2010	NB	73821
Rashad Arnold	03/31/2011	F	73909
Myra Ford	04/26/2010	F	14825

scrambling (hashing)

Name	DOB	Gender	Zip
A23C	02/26/2010	M	73909
85E1	05/08/2010	M	14890
B066	05/11/2010	NB	73821
45FF	03/31/2011	F	73909
3D28	04/26/2010	F	14825

masking

Name	DOB	Gender	Zip
R*****	02/26/2010	M	73909
A*****	05/08/2010	M	14890
M*****	05/11/2010	NB	73821
M*****	03/31/2011	F	73909
A*****	04/26/2010	F	14825

Re-identification

Process of associating a person with de-identified data: it is the outcome of a linkage attack to perform identity disclosure

Name	DOB	Gender	Zip
M*****	05/11/2010	NB	73821
M*****	03/31/2011	F	73909
A*****	04/26/2010	F	14825

joined with (Aimee Thornton, F), reveals the full record

Name	DOB	Gender	Zip
Aimee Thornton	04/26/2010	F	14825

Re-identification (cont'd)

- Requires **auxiliary data** to join with
- Linking de-identified data to auxiliary data can reveal sensitive information
- Could be seen as *record linkage*

Anonymization

Several definitions

- a synonym for de-identification...
- Replace identifiers with pseudo-identifiers (*pseudonymization*)
- A process which is **irreversible** and prevents re-association—linkage attack—of a person with a data sample

Limitation

True anonymization is **mainly not possible**

See the many de-identification use cases of the introductory lecture

Anonymization: A Stupid Example

Name	DOB	Gender	Zip
Rashad Arnold	02/26/2010	M	73909
Alyssa Cherry	05/08/2010	M	14890
Myra Ford	05/11/2010	NB	73821

becomes

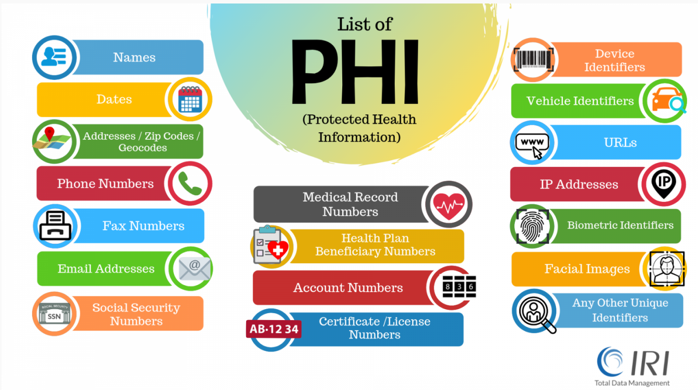
Name	DOB	Gender	Zip
*****	**/**/****	**	*****
*****	**/**/****	**	*****
*****	**/**/****	**	*****

Anonymization is actually a pretty vague term

Why Should We Care About Anonymization?

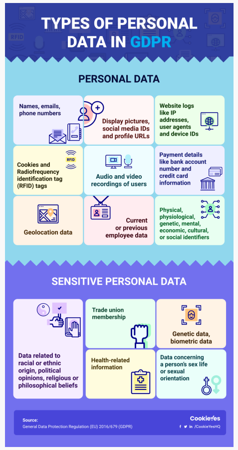
It get used **a lot**, commonly required by legal frameworks

HIPAA (Health Insurance Portability and Accountability Act) in the US

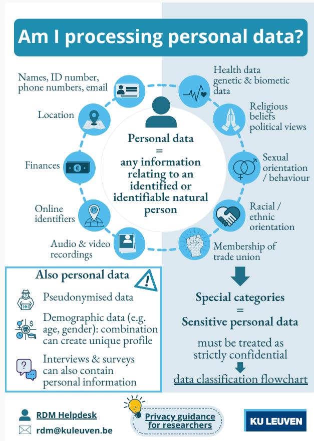


Why Should We Care About Anonymization?

GDPR (General Data Protection Regulation) in Europe



Why Should We Care About Anonymization?



Why Should We Care About Anonymization?

Those attributes are called **Personally Identifiable Information (PII)**

- Removing PII makes re-identification harder but **not impossible**
- Definitions of PII vary and then, they are also vague

What is the Goal of SDC?

Statistical Disclosure Control takes a **systematic approach** to de-identification in order to minimize the risk of re-identification

NIC-612092-Q0Y6F+admissions_for_assault_suppressed_2024_02

RP_START	RP_END	RP_TYPE	ORG_TYPE	ORG_CODE	ORG_DESCRIPTION	MEASURE_ID	MEASURE_NAME	DEMOGRAPHIC_GROUP	MEASURE_VALUE	SUPPRESSION	PROVISIONAL
01/02/2022	28/02/2022	MONTH	PFA	E23000001	Metropolitan Police	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	60		
01/02/2022	28/02/2022	MONTH	PFA	E23000002	Cumbria	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	0	Y	
01/02/2022	28/02/2022	MONTH	PFA	E23000003	Lancashire	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	10		
01/02/2022	28/02/2022	MONTH	PFA	E23000004	Merseyside	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	10		
01/02/2022	28/02/2022	MONTH	PFA	E23000005	Greater Manchester	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	25		
01/02/2022	28/02/2022	MONTH	PFA	E23000006	Cheshire	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	0	Y	
01/02/2022	28/02/2022	MONTH	PFA	E23000007	Northumbria	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	15		
01/02/2022	28/02/2022	MONTH	PFA	E23000008	Durham	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	0	Y	
01/02/2022	28/02/2022	MONTH	PFA	E23000009	North Yorkshire	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	0	Y	
01/02/2022	28/02/2022	MONTH	PFA	E23000010	West Yorkshire	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	15		
01/02/2022	28/02/2022	MONTH	PFA	E23000011	South Yorkshire	AFAS01	ASSAULT_BY_SHARP_OBJECTS_FAE	ALL	10		

Hospital admissions for assault by sharp objects February 2024 (3 995 records, Feb. 2022 - Feb. 2024)
Source: NHS England

Demographic group (all, under 25, etc.) and *measure value* have been altered

What Else Can We Do?

- Data use agreements
- Access control restrictions
- Audits
- **More systematic approach to making data private**

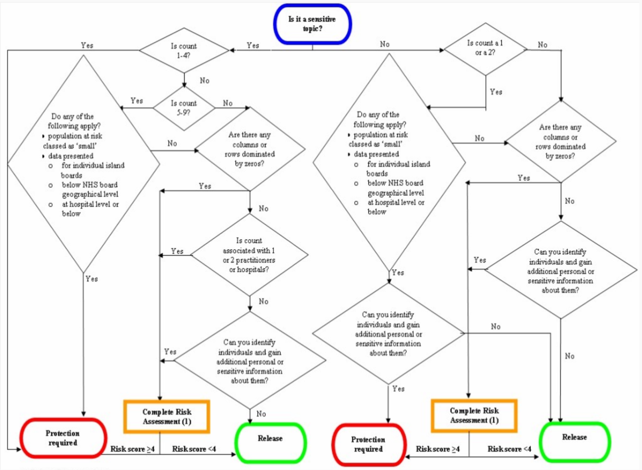
SDC Approach

Consider

- Likelihood of an attempt at disclosure
- Impact of disclosure
- Auxiliary data available to attackers
- Cell values and table design, e.g. counts of 1 or 0 represent high risk

Represents a **subjective judgment** about risk—no formal guarantee

Rule-based SDC for Scottish NHS



Generalization (Coarsening)

ORIGINAL MICRODATA

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer
13068	33	Chinese	Cancer

▷

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥ 40	Any	Cancer
1485*	≥ 40	Any	Heart
1485*	≥ 40	Any	Viral
1485*	≥ 40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Equivalence Class: block of k-anonymous records that share the same quasi-identifier value

k-Anonymity

Main Idea [Samarati and Sweeney, 1998]

Any individual is member of a block of size **at least** k over its quasi-identifier

- **Formal guarantee**, following the principle "hiding in the crowd"
- Parameter k gives the "degree" of anonymity
- Still requires to define quasi-identifier
- In SQL, table T is k-anonymous if each value from

```
SELECT COUNT(*)
FROM T
GROUP BY Quasi-Identifier
```

  is $\geq k$

Quasi-Identifier

PII attributes of a given dataset are either:

- Direct Identifier: removed
- Quasi-Identifier (QID): transformed
- Sensitive: preserved

How to set up QID?

- QID is a combination of attributes (that an adversary may know) that uniquely identify a large fraction of the population
- There can be many sets of QID: if $Q = \{A, B, C\}$ is a quasi-identifier, then $Q \cup \{D\}$ is also a quasi-identifier
- Need to guarantee k-anonymity against the largest QID

Attack 1: Homogeneity

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Viral
130**	<30	Any	Viral
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Name	Zip	Age	Nat.
Bob	13053	35	French

- Bob has cancer

ℓ-Diversity

In addition to k-Anonymity, require:

ℓ-Diversity Principle [Machanavajjhala et al., 2006]

A q^*-block is ℓ-diverse if it contains at least ℓ well-represented values for the sensitive attribute S. A table is ℓ-diverse if every q^*-block is ℓ-diverse.

Prevents attack #1 (homogeneity)

If all values are equally represented, all rows are equally likely to be the target's record

Increases resistance against attack #2 (background knowledge)

- Protects the target, even if the attacker knows $\ell - 2$ negation statements about the block ("*Umeko does not have cancer*")
- If the attacker knows $\ell - 1$ negation statements, then the attacker eliminates all rows but one

Attack 2: Background Knowledge

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Heart
130**	<30	Any	Flu
130**	<30	Any	Flu
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

- Japanese have a very low incidence of Heart disease
- Umeko has flu

Attack 2: Background Knowledge

4-ANONYMOUS RELEASE

Zip	Age	Nationality	Disease
130**	<30	Any	Heart
130**	<30	Any	Diabetes
130**	<30	Any	Cancer
130**	<30	Any	Flu
1485*	≥40	Any	Cancer
1485*	≥40	Any	Heart
1485*	≥40	Any	Viral
1485*	≥40	Any	Viral
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer
130**	[30,40)	Any	Cancer

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

- Umeko does not have cancer
- Umeko does not have heart disease
- Umeko does not have diabetes
- Umeko has flu

k-Anonymity & ℓ-Diversity

- **Formal privacy models** to prevent *identity disclosure* through linkage attack
- **Big improvement** over ad-hoc approaches
- Not yet covered: **high computation cost**
  - Given table T, find a k-anonymous table T' that maximizes utility
  - NP-hard problem [Meyerson and Williams, 2004]

Exposition to Attribute Disclosure

- Homogeneity Attack
- Background Knowledge Attack

Lots of Extended Models

- t-Closeness [Li et al., 2007]
- m-Invariance [Xiao and Tao, 2007]
- τ-Safety [Anjum et al., 2017]
- etc

Privacy protection depends on adversary's auxiliary information

What is Differential Privacy?

Definition (Differential Privacy [Dwork et al., 2006])
An algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ preserves ε-differential privacy if for any pair of neighboring databases $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$, and for any output o among the possible outputs:

$$\Pr[\mathcal{A}(\mathbf{x}) = o] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(\mathbf{y}) = o]$$

In other words...

$$\frac{\Pr[\mathcal{A}(\mathbf{x}) = o]}{\Pr[\mathcal{A}(\mathbf{y}) = o]} \leq e^{\varepsilon}$$

First proposed in [Dwork et al., 2006] by **Dwork, McSherry, Nissim and Smith who won the Gödel prize in 2017**

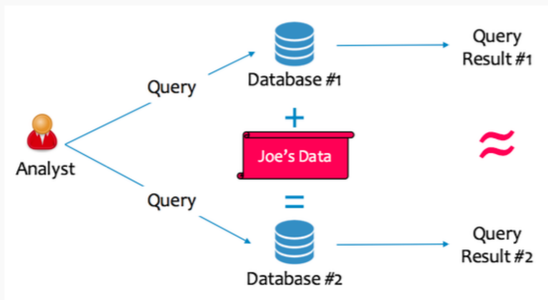
Back to the Attempt at Privacy Definition

Definition (Privacy)
"An analysis of a dataset is private if what can be learned about an individual in the dataset **is not much more** than what would be learned **if the same analysis was conducted without him/her in the dataset**."

Intuition
Cannot infer the presence/absence of an individual in the dataset, or anything "specific" about an individual

Here, "specific" refers to information that cannot be inferred unless the individual's data is used in the analysis

What Does the Guarantee Mean?



- Two neighboring DBs are identical except for data of one individual
- $\mathcal{A}$ algorithm's output does not enable adversary to distinguish between the two neighboring databases
- **Outcome is the same whether or not an individual participates**

Why is it a Good Guarantee?

- Matches a "pretty good" intuitive definition of privacy:
  nothing bad happens to me as a result of my participation in an analysis
  - i.e. if a bad thing happens, it would have happened even if I did not participate
- Formal definition enables **proving** that an algorithm satisfies differential privacy
- Holds **regardless of adversary's auxiliary knowledge**
  - Including case where the adversary knows the entire database except the target's row
  - Prevents from the attacks on k-Anonymity and its extensions
- **Only way we know** to come close to "true anonymization"

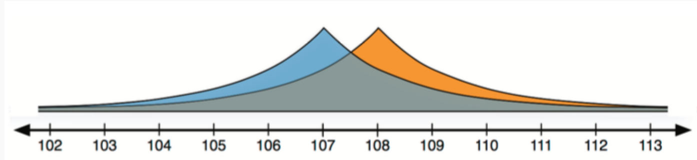
What are the Downsides?

- **No synthetic data, only query answering**
  - DP is a property of an algorithm (i.e. the analysis itself), not a property of data
    But in many cases, those algorithms can generate "good enough" synthetic data
- **Hard to interpret the guarantee**
  - Strength of guarantee parameterized by ε: "how hard is it to distinguish two neighboring databases?"
  - What ε is sufficient? too low $\rightarrow$ poor utility • too high $\rightarrow$ re-identification becomes possible
  - We don't really know the answer yet

Interpreting the Formal Definition

$$\frac{\Pr[\mathcal{A}(\mathbf{x}) = o]}{\Pr[\mathcal{A}(\mathbf{y}) = o]} \leq e^{\varepsilon} \quad \Leftrightarrow \quad \ln \frac{\Pr[\mathcal{A}(\mathbf{x}) = o]}{\Pr[\mathcal{A}(\mathbf{y}) = o]} \leq \varepsilon$$

This is called the **privacy loss** (or "privacy budget")



A differentially private mechanism **should produce probability distributions like these** over its outputs

Takeaways (1/3)

De-identification / Anonymization

- Suppresses PII to reduce risk of re-identification
- Ad-hoc approach means high risk of mistakes
- Most commonly used technique

SDC

- Makes de-identification systematic
- Considers size of groups in output data
- Still no formal guarantee

Takeaways (2/3)

k-Anonymity

- Formalizes systematic de-identification
- Requires groups to be at least size k
- Subject to homogeneity and auxiliary knowledge attacks

ℓ-Diversity

- Requires groups to be diverse
- Prevents homogeneity attack
- Prevents auxiliary knowledge attacks when the adversary knows fewer than $\ell - 2$ negative facts about the group

Takeaways (3/3)

Differential Privacy

- Formal property of a mechanism (e.g. algorithm or analysis or query)
  - Not a process to generate private data
- Corresponds to notion of **indistinguishability**: same outcome, whether I participate or not
- Guarantee holds regardless of adversary's auxiliary knowledge
  - Only family of approaches we know with this property

References i

Anjum, A., Raschia, G., Gelgon, M., Khan, A., Malik, S. U. R., Ahmad, N., Ahmed, M., Suhail, S., and Alam, M. (2017).
τ-safety: A privacy model for sequential publication with arbitrary updates.
Comput. Secur., 66:20–39.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006).
Calibrating noise to sensitivity in private data analysis.
In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, page 265–284, Berlin, Heidelberg. Springer-Verlag.

References ii

Li, N., Li, T., and Venkatasubramanian, S. (2007).
t-closeness: Privacy beyond k-anonymity and l-diversity.
In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006).
L-diversity: privacy beyond k-anonymity.
In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24.

References iii

Meyerson, A. and Williams, R. (2004).
On the complexity of optimal k-anonymity.
In *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, page 223–228, New York, NY, USA. Association for Computing Machinery.

Samarati, P. and Sweeney, L. (1998).
Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression.
Technical Report SRI-CSL-98-04, SRI International.

References iv

Xiao, X. and Tao, Y. (2007).
M-invariance: towards privacy preserving re-publication of dynamic datasets.
In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, page 689–700, New York, NY, USA. Association for Computing Machinery.