CS295B: Data Privacy, Lecture 2

Joe Near (jnear@uvm.edu)

8/29/2018

An Overview of Privacy Techniques

Technique	Functionality
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 data

Name	**DOB**	**Gender**	**Zip**
Rashad Arnold	02/26/2018	M	73909
Alyssa Cherry	05/08/2018	M	14890
Myra Ford	05/11/2018	F	58821
Meredith Perry	03/31/2019	F	465113
Aimee Thornton	04/26/2018	F	90825

$\Downarrow$

Name	**DOB**	**Gender**	**Zip**
*****	02/26/2018	M	73909
*****	05/08/2018	M	14890
*****	05/11/2018	F	58821
*****	03/31/2019	F	465113
*****	04/26/2018	F	90825

Synthetic Data vs Query Answering

Query answering *requires* a specific query

Name	**DOB**	**Gender**	**Zip**
Rashad Arnold	02/26/2018	M	73909
Alyssa Cherry	05/08/2018	M	14890
Myra Ford	05/11/2018	F	58821
Meredith Perry	03/31/2019	F	465113
Aimee Thornton	04/26/2018	F	90825

$$+$$

How many people were born in 2018?

$$\Downarrow$$

4

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

- Often requires modifying data analyses
- Approach depends on query workload
- Makes things harder for the analyst
- Specialization to *one query* enables better utility/privacy tradeoff

What does Utility Mean?

Informally: "how useful is the answer?"

Formally: depends on what the answer will be used for

Example: "how many people have the last name *Ford*?"

- Anonymized data $\rightarrow$ impossible to answer
- Differential privacy $\rightarrow$ can answer ± 1 person

Other 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?

Outline

1. Anonymization / De-identification
2. Statistical Disclosure Control
3. k-Anonymity & ℓ-Diversity
4. Differential Privacy

Goals of De-identification

De-identification is a process which removes the association (via personal information) between a person and a data set.

Goals:

- Reduce risk of privacy violation
- Maximize data utility

Techniques:

- Suppression (remove the data)
- Variation (scramble the data)
- Swap data items
- Masking

De-identification: Example

We saw an example of de-identified data earlier:

Name	DOB	Gender	Zip
*****	02/26/2018	M	73909
*****	05/08/2018	M	14890
*****	05/11/2018	F	58821
*****	03/31/2019	F	465113
*****	04/26/2018	F	90825

In this data, **names have been masked**.

Re-identification

Re-identification is a process that re-associates a person with a data sample.

Name	DOB	Gender	Zip
*****	02/26/2018	M	73909
*****	05/08/2018	M	14890
*****	05/11/2018	F	58821
*****	03/31/2019	F	465113
*****	04/26/2018	F	90825

Re-identification

Re-identification is a process that re-associates a person with a data sample.

Name	DOB	Gender	Zip
*****	02/26/2018	M	73909
*****	05/08/2018	M	14890
*****	05/11/2018	F	58821
*****	03/31/2019	F	465113
*****	04/26/2018	F	90825

+

Name	DOB	Gender	Zip
Rashad Arnold	05/08/2018	*	*****

=

Name	DOB	Gender	Zip
Rashad Arnold	05/08/2018	M	14890

Re-identification

Re-identification is a process that re-associates a person with a data sample.

Name	DOB	Gender	Zip
*****	02/26/2018	M	73909
*****	05/08/2018	M	14890
*****	05/11/2018	F	58821
*****	03/31/2019	F	465113
*****	04/26/2018	F	90825

+

Name	DOB	Gender	Zip
Rashad Arnold	05/08/2018	*	*****

=

Name	DOB	Gender	Zip
Rashad Arnold	05/08/2018	M	14890

Relies on **auxiliary data**

Also called **record linkage**

Anonymization

Some definitions:

- Same as de-identification
- Replace identifiers with pseudoidentifiers (pseudonymization)
- A process which is **irreversible** and prevents the re-association of a person with a data sample

The last one is **not really possible**

Anonymization: Example

Name	DOB	Gender	Zip
Rashad Arnold	02/26/2018	M	73909
Alyssa Cherry	05/08/2018	M	14890
Myra Ford	05/11/2018	F	58821
Meredith Perry	03/31/2019	F	465113
Aimee Thornton	04/26/2018	F	90825

$$\Downarrow$$

Anonymization: Example

Name	DOB	Gender	Zip
Rashad Arnold	02/26/2018	M	73909
Alyssa Cherry	05/08/2018	M	14890
Myra Ford	05/11/2018	F	58821
Meredith Perry	03/31/2019	F	465113
Aimee Thornton	04/26/2018	F	90825

$\Downarrow$

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

Anonymization is a pretty vague term

Why Should We Care About Anonymization & De-identification?

It gets used **a lot**.

HIPAA (Health Insurance Portability and Accountability Act) requires removing:

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
   a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people.
   b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Facsimile numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.

Why Should We Care About Anonymization & De-identification?

GDPR (General Data Protection Regulation) requires removing:

Table 1. Examples of personal identifiers and personal characteristics

Personal identifiers	**Personal characteristics**
Name	Ethnic background
ID (social security or driver's license number)	Political views
Physical address	Religion
E-mail address	Physiological data (e.g., DNA)
Photo	Medical conditions
IP address	
Geographical location (GPS) of mobile phone	
*Browser cookie	

Why Should We Care About Anonymization & De-identification?

GDPR (General Data Protection Regulation) requires removing:

Table 1. Examples of personal identifiers and personal characteristics

Personal identifiers	**Personal characteristics**
Name ID (social security or driver's license number) Physical address E-mail address Photo IP address Geographical location (GPS) of mobile phone *Browser cookie	Ethnic background Political views Religion Physiological data (e.g., DNA) Medical conditions

These identifiers are called **personally identifiable information (PII)**.

- Removing PII makes re-identification harder
- Removing PII does **not** make re-identification impossible
- PII is another vague term

What Else Can We Do?

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

Outline

1. Anonymization / De-identification
2. Statistical Disclosure Control
3. k-Anonymity & ℓ-Diversity
4. Differential Privacy

What is the Goal of SDC?

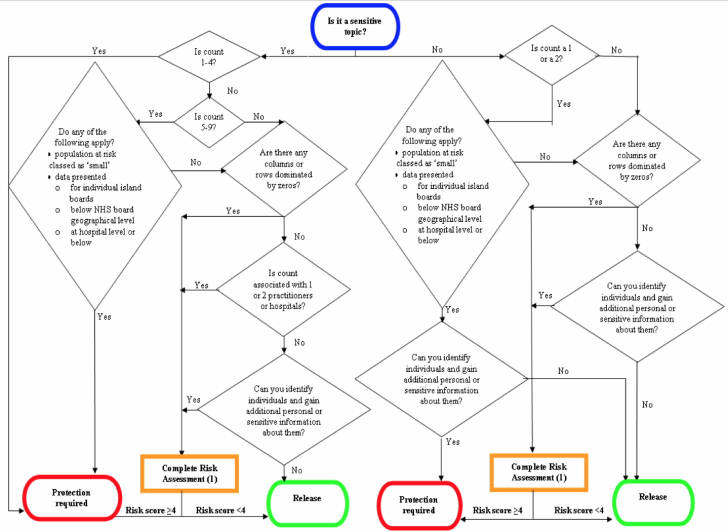
Statistical disclosure control takes a **systematic approach** to de-identification in order to minimize the risk of re-identification.

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

What Does SDC Look Like?



SDC: Example (ISD Scotland example for health data)

Table 1: Number of emergency hospital admissions due to assault by sharp object[1] in 0-17 and 18+ year olds, by council area of residence; discharged during financial years 2002/2003 to 2006/2007

Age Group	Council Area of residence	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
0-17	Council 1	1	1	1	1	1
	Council 2	-	1	2	1	-
	Council 3	3	-	-	-	-
	Council 4	1	3	-	2	1
	Council 5	10	5	5	10	7
	Council 6	1	-	-	-	-

⇓

Table 1: Number of emergency hospital admissions due to assault by sharp object[1] in 0-17 and 18+ year olds, by council area of residence; discharged during financial years 2002/2003 to 2006/2007

Age Group	Council Area of residence	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
0-17	Council 1	*	*	*	*	*
	Council 2	*	*	*	*	*
	Council 3	*	*	*	*	*
	Council 4	*	*	*	*	*
	Council 5	10	5	5	10	7
	Council 6	*	*	*	*	*

Outline

What is k-Anonymity?

Definition 2.3 (k-anonymity) *Let $T(A_1, \ldots, A_n)$ be a table and QI_T be the quasi-identifiers associated with it. T is said to satisfy k-anonymity iff for each quasi-identifier $QI \in \mathsf{QI}_T$ each sequence of values in $T[QI]$ appears at least with k occurrences in $T[QI]$.*

[Pierangela and Sweeney, 1998].

- Ensures no individual is uniquely identifiable from a group of size k
- **Formal guarantee**
- Still requires identifying **quasi-identifiers**
  - But we can include *lots of them*
- In SQL, a table **T** is k-anonymous if:

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

$\geq k$

k-Anonymity: Example (Generalization)

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

$\Rightarrow$

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

k-Anonymity Attack #1: Homogeneity

Name	Zip	Age	Nat.
Bob	13053	35	??

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

k-Anonymity Attack #1: Homogeneity

Name	Zip	Age	Nat.
Bob	13053	35	??

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

We learn: **Bob has cancer**

k-Anonymity Attack #2: Auxiliary Data

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

k-Anonymity Attack #2: Auxiliary Data

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

We learn: **Umeko has flu**

ℓ-Diversity

In addition to k-Anonymity, require:

Principle 2. (*ℓ-Diversity Principle*). A $q^\star$-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 $q^\star$-block is ℓ-diverse.

[Machanavajjhala et al., 2006].

Prevents attack #1 (homogeneity)

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

ℓ-Diversity

In addition to k-Anonymity, require:

Principle 2. (*ℓ-Diversity Principle*). A $q^\star$-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 $q^\star$-block is ℓ-diverse.

[Machanavajjhala et al., 2006].

Prevents attack #1 (homogeneity)

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

Increases resistance against attack #2 (auxiliary data)

- Protects the target, even if the attacker knows $\ell - 2$ *negation statements* about the block
  - Negation statements are of the form: "Umeko does not have cancer"

ℓ-Diversity

In addition to k-Anonymity, require:

Principle 2. (*ℓ-Diversity Principle*). A $q^{\star}$-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 $q^{\star}$-block is ℓ-diverse.

[Machanavajjhala et al., 2006].

Prevents attack #1 (homogeneity)

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

Increases resistance against attack #2 (auxiliary data)

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

ℓ-Diversity Attack: Auxiliary Data

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Umeko does not have cancer

Umeko does not have heart disease

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cancer
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

ℓ-Diversity Attack: Auxiliary Data

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Umeko does not have cancer

Umeko does not have heart disease

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cancer
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Umeko could have diabetes or flu

ℓ-Diversity Attack: Auxiliary Data

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Umeko does not have cancer

Umeko does not have heart disease

Umeko does not have diabetes

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cancer
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

ℓ-Diversity Attack: Auxiliary Data

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Umeko does not have cancer

Umeko does not have heart disease

Umeko does not have diabetes

+

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Diabetes
130**	<30	*	Cancer
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

We learn: **Umeko has flu**

Lessons: k-Anonymity & ℓ-Diversity

- **Formal**, **systematic** approaches to de-identification
- Big improvement over ad-hoc approaches
- **Still** subject to attacks
  - Privacy protection depends on **adversary's auxiliary information**

Lessons: k-Anonymity & ℓ-Diversity

- **Formal**, **systematic** approaches to de-identification
- Big improvement over ad-hoc approaches
- **Still** subject to attacks
  - Privacy protection depends on **adversary's auxiliary information**
- Not yet covered: high **computational cost**
  - Given a table T, find a table T' that satisfies k-Anonymity and maximizes utility
  - NP-hard (Meyerson & Williams, 2004)

Outline

What is Differential Privacy?

Definition (Differential privacy)

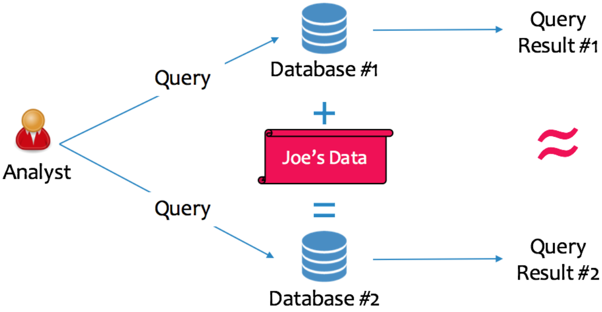
A randomized mechanism $\mathcal{K} : D^n \rightarrow \mathbb{R}^d$ preserves ϵ-differential privacy if for any pair of databases $x, y \in D^n$ such that $d(x, y) = 1$, and for all sets S of possible outputs:

$$\Pr[\mathcal{K}(x) \in S] \leq e^{\epsilon} \Pr[\mathcal{K}(y) \in S]$$

In other words...

$$\frac{\Pr[\mathcal{K}(x) \in S]}{\Pr[\mathcal{K}(y) \in S]} \leq e^{\epsilon}$$

What Does the Guarantee Mean?



- Two **hypothetical** DBs are **identical** except for data of one individual
- Mechanism's **output** does not enable adversary to **distinguish** between the two 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 a mechanism 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 the linking attacks on k-Anonymity and ℓ-Diversity
  - *Only way we know* to come close to "true anonymization"

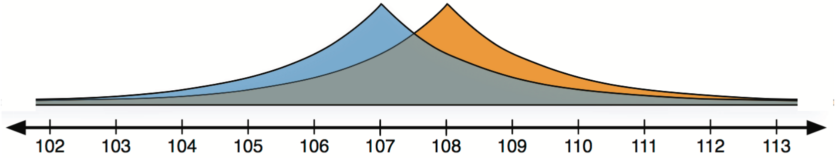
What are the Downsides?

- **No synthetic data, only query answering**
  - Differential privacy is a property of a *mechanism* (i.e. the analysis itself), not a property of *data*
  - In many cases, mechanisms *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?
    - ϵ too low $\rightarrow$ poor utility
    - ϵ too high $\rightarrow$ re-identification becomes possible
    - We don't really know the answer yet

Interpreting the Formal Definition

$$\frac{\Pr[\mathcal{K}(x) \in S]}{\Pr[\mathcal{K}(y) \in S]} \leq e^{\epsilon} = \ln \frac{\Pr[\mathcal{K}(x) \in S]}{\Pr[\mathcal{K}(y) \in S]} \leq \epsilon$$

This is called the **privacy loss**



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)
  - 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