Privacy & Data Security 27.10.21

5) Data Anonymization

Recall: - Anonymization eliminates personal data

- Anonymized data no longer contains personal data
- Not subject to privacy laws

- Publish anonymized data
- Anonymized data cannot be associated to individuals
- Presence or absence of one individual's data cannot be detected

5.1) Anonymization and utility

Tradeoff:

- Full dataset has complete utility
- Anonymous dataset has no —"—

- Measures for privacy or anonymity

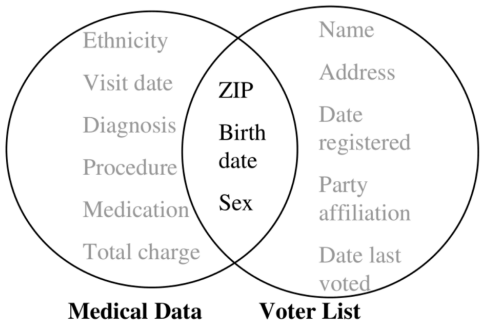
(- Measures for utility)

Anonymization is difficult

- Additional / background data
- Correlation with such external knowledge
- Predicting how much extra information will exist seems impossible

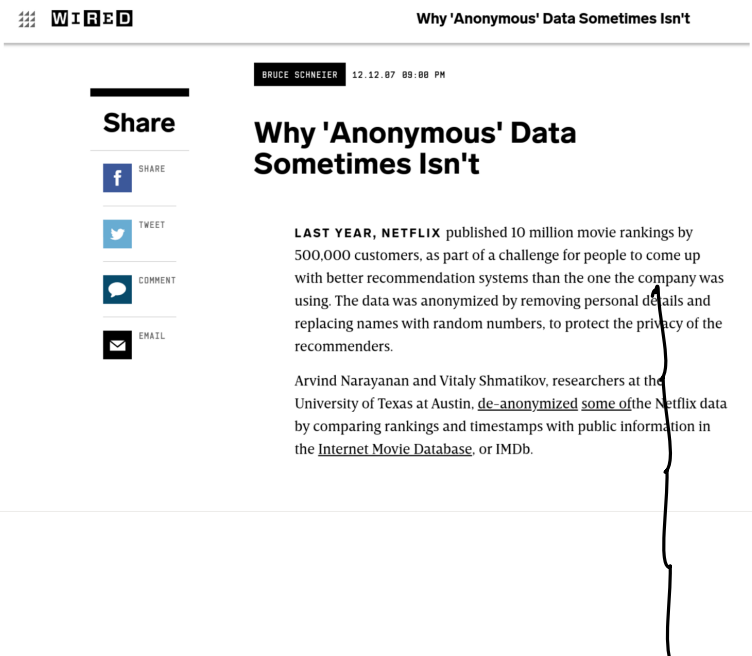
Reidentification of medical records

- In Massachusetts, 1998, insurance collected patient-specific data, removed obvious names, and released it
- Study identified individuals, including state's Governor, by correlating with public voter list



Netflix prize

- Netflix offered a prize to engineer a prediction algorithm for user ratings
  - Training data set of 100M ratings from 500K users, pseudonymized
- Researchers identified single users based on external data (e.g., IMDb)



AOL search logs

- AOL released 20M search queries in 2006, with pseudonymized users
- NY Times journalists identified a surprised searcher from GA

TECHNOLOGY

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

5.2) k-Anonymity

- Dataset is a table
  - attribute set A
- Each row (entry) is personal data
- Partition attribute set A into
  - Sensitive attr. S: attributes of interest
  - Identifiers I: individual data

- Quasi - Identifiers QI : features of data, may identify individual

$A = I \cup QI \cup S$

Ex.

Name	Firstname	PLZ	Points	System OS
Cachin	Christian	8600	9	Linux
I	I	QI	QI	S ↑ sensitive

Def: Partition dataset by quasi-identifiers s.t. each equivalence class (by QI) contains at least k entries: then the dataset is **k-anonymous.**

- Identifiers (I) are removed
- Only QI considered

Techniques to anonymize a dataset:

- Suppression: remove entries with infrequent values of QI
- Generalization: replace QI values by more general values
  - numerical data: intervals
  - qualitative data: using a semantic hierarchy

Last name I	First name I	PLZ QI	Points QI	System S

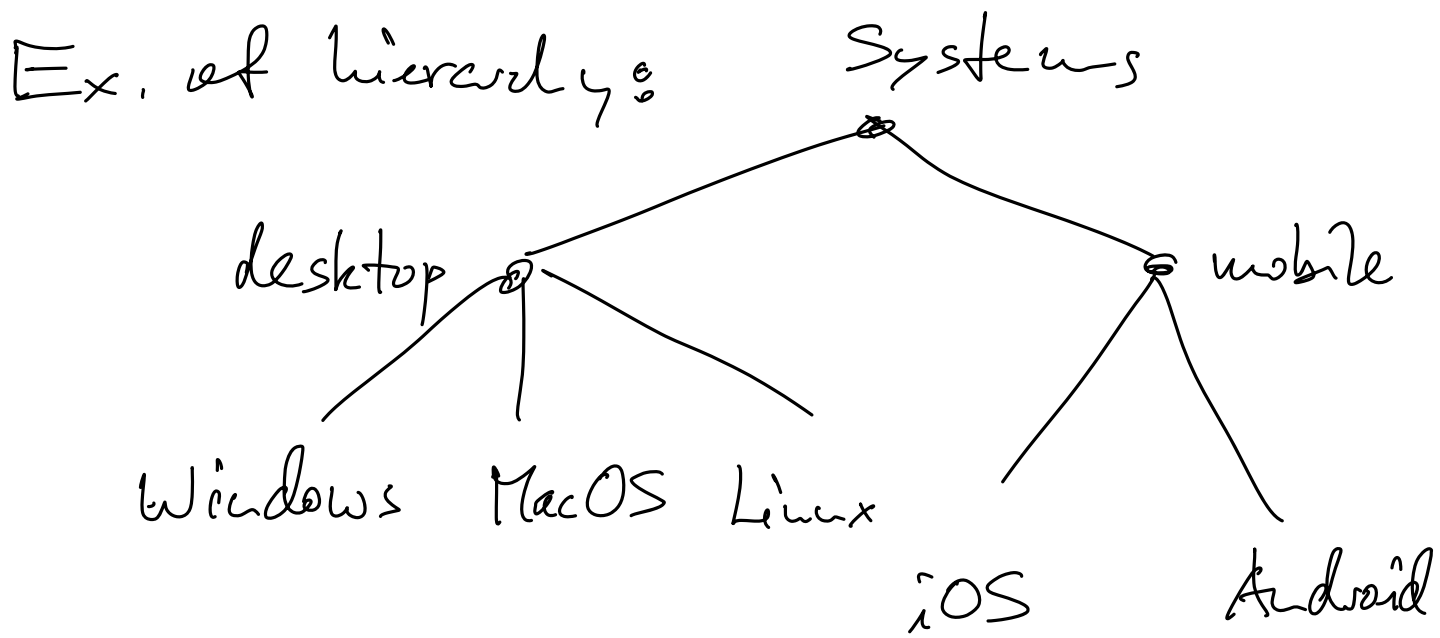
Sample data set

Andreasyan	Narek	3270	89	iOS
Asadauskas	Marius Paulius	3294	77	Android
Ayinkamiye	Leïla	3400	90	MacOS
Berger	Reto	2608	42	Windows
Bucheli	Philippe	3177	38	Linux
Bühlmann	Noah Florian	2740	35	Windows
Brunner	Julien Pierre	3763	25	MacOS
Egger	Dominic Mathias	3860	33	Windows
Gerig	Pascal Dominik	3770	30	Android

3-Anonymous data set

eg. class

3200-3299	75-90	iOS
3200-3299	75-90	Android
3200-3299	75-90	MacOS
2600-3199	35-45	Windows
2600-3199	35-45	~~Linux~~ Windows
2600-3199	35-45	Windows
3700-3899	25-34	MacOS
3700-3899	25-34	Windows
3700-3899	25-34	Android



=> Hides sensitive attribute values

=> Protect against identity disclosure

Problems with k-anonymity

- Background knowledge

  Observer can narrow down possible options and learn a sensitive attribute.

- Homogeneity attack

  Not enough diversity among sensitive attribute values, then class reveals more about sensitive value.

  ↳ not enough diversity

5.3) l-Diversity

Def: Data is partioned according to quasi-identifiers.

An equivalence class is l-diverse if it contains at least l "well-represented" values for the sensitive attr.

A dataset is l-diverse whenever each equivalence class is l-diverse.

How "well-represented"?

- Distinct l-diversity:
  At least l different values.

- Probabilistic ℓ-diversity:

  The relative frequency of a value is at most $1/\ell$.

- Many more notions exist.

Ex. 2

Last name I	First name I	PLZ QI	Points QI	System S
Sample data set				
Andreasyan	Narek	3270	89	iOS
Asadauskas	Marius Paulius	3294	77	Android
Ayinkamiye	Leïla	3400	90	MacOS
Berger	Reto	2608	42	Windows
Bucheli	Philippe	3177	38	Linux
Bühlmann	Noah Florian	2740	35	Windows
Brunner	Julien Pierre	3763	25	MacOS
Egger	Dominic Mathias	3860	33	Windows
Gerig	Pascal Dominik	3770	30	Android
2-Diverse data set (but not 3-diverse)				
		3200-3299	75-90	iOS
		3200-3299	75-90	Android
		3200-3299	75-90	MacOS
		2600-3199	35-45	Windows
		2600-3199	35-45	Linux
		2600-3199	35-45	Windows
		3700-3899	25-34	MacOS
		3700-3899	25-34	Windows
		3700-3899	25-34	Android

Ex. 3

Last name I	First name I	PLZ QI	Points QI	System S
Sample data set				
Andreasyan	Narek	3270	89	iOS
Asadauskas	Marius Paulius	3294	77	Android
Ayinkamiye	Leïla	3400	90	MacOS
Berger	Reto	2608	42	Windows
Bucheli	Philippe	3177	38	Linux
Bühlmann	Noah Florian	2740	35	Windows
Brunner	Julien Pierre	3763	25	MacOS
Egger	Dominic Mathias	3860	33	Windows
Gerig	Pascal Dominik	3770	30	Android
3-Diverse data set				
		2600-3299	35-90	iOS
		2600-3299	35-90	Android
		2600-3299	35-90	MacOS
		2600-3299	35-90	Windows
		2600-3299	35-90	Linux
		2700-3899	25-36	Windows
		2700-3899	25-36	MacOS
		2700-3899	25-36	Windows
		2700-3899	25-36	Android

=> Hides sensitive data

=> Protects against attribute disclosure

Problems with ℓ-diversity

- Similarity of sensitive values

  Ex. 2, class 2600... used Windows and Linux, this leaks they are all on desktops

- Skewed data set

  Suppose dataset of 10000 entries, only two values of S (pos./neg.):

  $$\begin{cases} \text{neg.}: & \text{for } 99\% \\ \text{pos.}: & \text{for } 1\% \quad \text{"sensitive"} \end{cases}$$

  The data is not balanced.

  - With 2-diverse classification, there can be at most 100 classes, and information loss is large.

  - If the classes are not representative

of full data set
(Ex: 1 class has 50% pos.),
then membership in this class
is "sensitive" and discloses a lot of
information about that individual.

Problem here is that distribution
of sensitive attribute values is
not taken into account.

Outlook:

$$S = \{ \text{Android}, \text{iOS}, \ldots$$

Initial assumption P_A:

$$A \in S$$

Generalized data P_Q

Leaked data P_L