k-anonymity

Privacy Enhancing Technologies

KARLSTAD
UNIVERSITY
SWEDEN

Leonardo A. Martucci

The PETS Module Structure

1. Introduction
   Day 1

2. Secure Communications
   Day 2

3. Anonymous Communications
   Day 3

4. Databases
   Day 4

5. Additional Topics
   Day 5

Part 4: Database Privacy

- Why do we need database privacy?
- k-anonymity 

- Differential Privacy

Not Easy to Release Data

- Data anonymization is a difficult

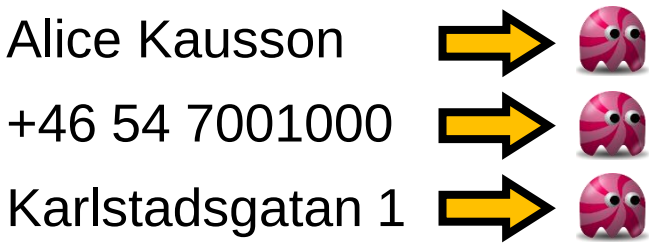
released a dataset of search queries from ca. 650K users ⇨ replaced usernames with numbers

The New York Times

Types of Identifiers

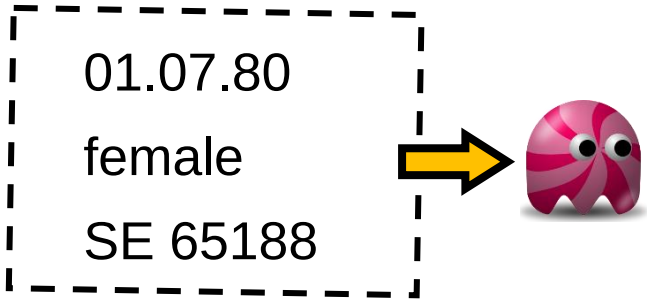
Explicit Identifiers

- Uniquely attributable
  - name
  - phone number
  - address



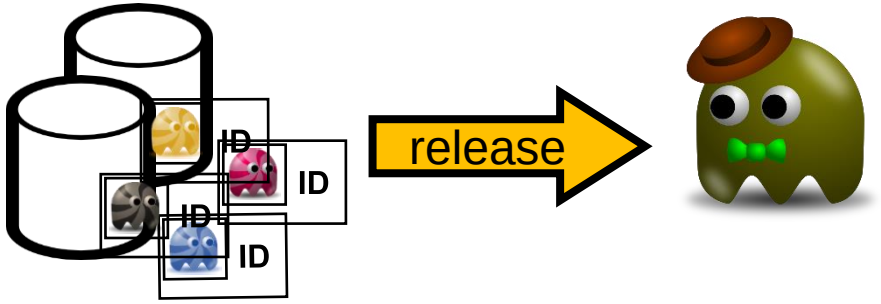
Quasi-Identifiers

- In combination, can uniquely identify
  - birth date
  - gender
  - ZIP code



k-anonymity

- Goal: to prevent re-identification of individuals when releasing data



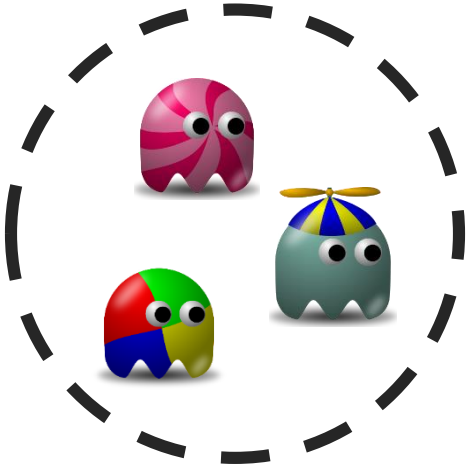
- k-anonymity property:

on data release, information about a subject cannot be distinguished from at least k-1 individuals

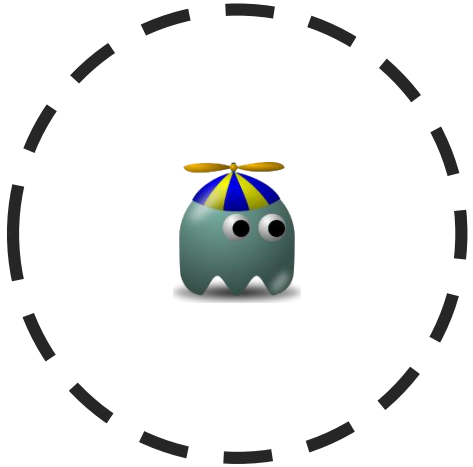
k

- Measure for the anonymity set

where min(k) = 2 (k = 1 means NO anonymity)



k = 3



k = 1

no anonymity!

Example: building a k=2 release

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	11.03.79	male	1072	married	1	A
	17.03.79	male	1276	married	7	B
	01.07.80	female	1073	single	2	B
	07.09.84	female	1077	single	0	C
	02.07.89	male	1016	single	2	D
	21.09.91	female	1267	it‘s complicated	4	E
	24.12.98	female	1268	it‘s complicated	4	A

Remove Name Field

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	11.03.79	male	1072	married	1	A
	17.03.79	male	1276	married	7	B
	01.07.80	female	1073	single	2	B
	07.09.84	female	1077	single	0	C
	02.07.89	male	1016	single	2	D
	21.09.91	female	1267	it‘s complicated	4	E
	24.12.98	female	1268	it‘s complicated	4	A

Generalize Birth date to Range

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	female	1073	single	2	B
	1980‘s	female	1077	single	0	C
	1980‘s	male	1016	single	2	D
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

The Gender Field

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	female	1073	single	2	B
	1980‘s	female	1077	single	0	C
	1980‘s	male	1016	single	2	D
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

NOT k=2 here

Generalize Gender Field

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	ghost	1073	single	2	B
	1980‘s	ghost	1077	single	0	C
	1980‘s	ghost	1016	single	2	D
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

OR Suppress Information

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1072	married	1	A
	1970‘s	male	1276	married	7	B
	1980‘s	female	1073	single	2	B
	1980‘s	female	1077	single	0	C
*	*	*	*	*	*	*
	1990‘s	female	1267	it‘s complicated	4	E
	1990‘s	female	1268	it‘s complicated	4	A

Generalize ZIP data

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	A
	1970‘s	male	1***	married	7	B
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
	1980‘s	ghost	10**	single	2	D
	1990‘s	female	12**	it‘s complicated	4	E
	1990‘s	female	12**	it‘s complicated	4	A

Civil Status Field is k=2!

Name	Birth date	Gender	ZIP	Civil Status	Duration	Diagnosis
	1970‘s	male	1***	married	1	A
	1970‘s	male	1***	married	7	B
	1980‘s	ghost	10**	single	2	B
	1980‘s	ghost	10**	single	0	C
	1980‘s	ghost	10**	single	2	D
	1990‘s	female	12**	it‘s complicated	4	E
	1990‘s	female	12**	it‘s complicated	4	A

l-diversity and t-closeness

l-diversity

- Addresses two attacks on k-anonymity
  - Homogeneity attack
  - Background knowledge attack

BUT

- Difficult, sometimes unnecessary
- Insufficient to prevent attribute disclosure

{ it does not consider overall data distribution
{ it does not consider semantics

t-closeness

- Addresses l-diversity limitations
- Metric is the attacker's information gain

BUT

- No computational procedure
- Limitations on the utility of data releases

If you want to know more

- Sweeney, L.: k-Anonymity: a Model for Protecting Privacy. Int. J. Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
- Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: Int Conf Data Engineering, ICDE 2006.
- Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Int Conf Data Engineering, ICDE 2007.

Part 4: Database Privacy

- Why do we need database privacy?
- k-Anonymity
- Differential Privacy