ROYAL INSTITUTE
OF TECHNOLOGY

K-Anonymity
Enhancing information privacy

Module TA: **Thanassis Giannetsos**, athgia@kth.se

Networked Systems Security Group
Panos Papadimitratos
www.ee.kth.se/nss

ROYAL INSTITUTE OF TECHNOLOGY

Share data

- How do you publicly release a database without compromising individual privacy?

The Wrong Approach:

- Just leave out any *unique* identifiers like name, personnummer, mobile number, etc. and hope that this works.

- The triple (date of birth, gender, zip code) suffices to uniquely identify at least 87% of US citizens in publicly available databases (Sweeney).

ROYAL INSTITUTE OF TECHNOLOGY

Definitions

- *Database* – a table with n rows (records) and m columns (attributes)
- *Alphabet of a Database* (Σ) – the range of values that individual cells in the database can take.
- Note that the alphabet of the k-anonymized database is Σ U {*}

ROYAL INSTITUTE OF TECHNOLOGY

Share data – the right way

- Privacy is required when it comes to data storage
  - Encryption is not always a solution
    - We want some forms of data to be accessible
    - We want to prevent linking various informations
      - Imagine a medical database!
- Models: **K-Anonymity** (Sweeney), Output Perturbation
  - Output perturbation algorithms severely distort query answers

ROYAL INSTITUTE OF TECHNOLOGY

Share data with k-Anonymity

How do you publicly release a database without compromising individual privacy?

- K-Anonymity: attributes are **suppressed** or generalized until each *row is identical with at least k-1* other rows. At this point the database is said to be k-anonymous.
- K-Anonymity thus prevents definite database linkages. At worst, the data released narrows down an individual entry to a group of k individuals.
- Unlike Output Perturbation models, K-Anonymity guarantees that the data released is accurate.

ROYAL INSTITUTE OF TECHNOLOGY

Methods

- **Suppression** – can replace individual attributes with a *
- **Generalization** – replace individual attributes with a broader category
  - Example: (Age: 26 => Age: [20-30])

ROYAL INSTITUTE OF TECHNOLOGY

K-Anonymity (1)

Original Database to Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
14	Fred Thompson	Male	1967	-ve

Protecting Privacy Using k-Anonymity

K-Anonymity (2)

	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

Protecting Privacy Using k-Anonymity

K-Anonymity (3)

Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

- In order to go back we need the identification database
- In this scenario we used 2-anonymity (gender, decade)
- We can expose only the table in the previous slide
  - For example for research purposes

Protecting Privacy Using k-Anonymity