K-ANONYMITY

LATANYA SWEENEY

Xiaoyong Zhou

DATA releasing: Privacy vs. Utility

- Society is experiencing exponential growth in the number and variety of data collections containing person-specific information.
  - Search engines
  - Credit card companies
  - Hospitals
- These collected information is valuable both in research and business. Data sharing is common.
- Publishing the data may put the respondent's privacy in risk.

- Objective:
  - Maximize data utility while limiting disclosure risk to an acceptable level

Related Works

- Statistical Databases
  - The most common way is adding noise and still maintaining some statistical invariant.

Disadvantages:

- Destroy the integrity of the data
- Adding noise itself is not a easy problem.

Related Works(Cont'd)

- Multi-level Databases
  - Data is stored at different security classifications and users having different security clearances. (Denning and Lunt)
  - Eliminating precise inference. Sensitive information is suppressed, i.e. simply not released. (Su and Ozsoyoglu)

  Disadvantages:
  - It is impossible to consider every possible attack
  - Many data holders share same data. But their concerns are different.
  - Suppression can drastically reduce the quality of the data.

Related Works (Cont'd)

- Computer Security
  - Access control and authentication ensure that right people has right authority to the right object at right time and right place.
  - That's not what we want here. A general doctrine of data privacy is to release all the information as much as the identities of the subjects (people) are protected.

Adversary Model

- Attacker can use information from other source and link to the released data and identify sensitive information of an individual. It’s unclear what information the attacker has.

K-Anonymity

Sweeny came up with a formal protection model named k-anonymity

- What is K-Anonymity?
  - If the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release.
  - Ex.

    If you try to identify a man from a release, but the only information you have is his birth date and gender. There are k people meet the requirement. This is k-Anonymity.

Classification of Attributes

- Key Attribute:
  - Name, Address, Cell Phone
  - which can uniquely identify an individual directly
  - Always removed before release.
- Quasi-Identifier:
  - 5-digit ZIP code,Birth date, gender
  - A set of attributes that can be potentially linked with external information to re-identify entities
  - 87% of the population in U.S. can be uniquely identified based on these attributes, according to the Census summary data in 1991.
  - Suppressed or generalized

Classification of Attributes(Cont'd)

Hospital Patient Data

DOB	Sex	Zipcode	Disease
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Vote Registration Data

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

- Andre has heart disease!

Classification of Attributes(Cont'd)

- Sensitive Attribute:
  - Medical record, wage,etc.
  - Always released directly. These attributes is what the researchers need. It depends on the requirement.

K-Anonymity Protection Model

- PT: Private Table
- RT,GT1,GT2: Released Table
- QI: Quasi Identifier (Ai,…,Aj)
- (A1,A2,…,An): Attributes

.

Let RT(A_1,...,A_n) be a table, QI_{RT} =(A_i,…, A_j) be the quasi-identifier associated with RT, A_i,…,A_j $\subseteq$ A_1,…,A_n, and RT satisfy k-anonymity. Then, each sequence of values in RT[A_x] appears with at least k occurrences in RT[QI_{RT}] for x=i,…,j.

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Attacks Against K-Anonymity

- Unsorted Matching Attack
  - This attack is based on the order in which tuples appear in the released table.
  - Solution:
    - ing.

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Figure 3 Examples of *k*-anonymity tables based on PT

Attacks Against K-Anonymity(Cont'd)

- Complementary Release Attack
  - Different releases can be linked together to compromise k-anonymity.
  - Solution:
    - Consider all of the released tables before release the new one, and try to avoid linking.
    - Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

Attacks Against K-Anonymity(Cont'd)

- Complementary Release Attack (Cont'd)

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Attacks Against K-Anonymity(Cont'd)

- Complementary Release Attack (Cont'd)

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Attacks Against K-Anonymity(Cont'd)

- Temporal Attack (Cont'd)
  - Adding or removing tuples may compromise k-anonymity protection.

Attacks Against K-Anonymity(Cont'd)

- k-Anonymity does not provide privacy if:
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

A. Machanavajjhala et al. l-Diversity: Privacy Beyond k-Anonymity. ICDE 2006