CSC 495.002 – Lecture 6
Web/Social Networks Privacy: K-anonymity

Dr. Özgür Kafalı

North Carolina State University
Department of Computer Science

Fall 2017

Targeted Advertising

- Online behavioral advertising definition
- Types of targeted advertising
- Types of cookies and how they work
- Tools to mitigate privacy concerns of targeted advertising
- People's attitudes towards private browsing tools

Problem Definition

- Data owner, e.g., hospital
- Has private dataset with user specific data
- Goal: To share a version of the dataset with researchers
  - Dataset can help researchers to train better models
  - Results can help the data owner
- Provide scientific guarantees that users in the dataset cannot be re-identified
- Data should remain practically useful

Real Problem

- For, 87% (216M of 248M) of the US population
- Uniquely identifiable based only on
  - 5-digit ZIP code
  - Gender
  - Date of birth

Sweeney. Uniqueness of Simple Demographics in the US Population, 2000

Netflix Prize

- In October 2006, Netflix offered a $1M prize for a 10% improvement in its recommendation system
- Released a training dataset for competitors to train their systems
- Disclaimer: To protect customer privacy, all personal information identifying individual customers has been removed and all customer IDs have been replaced by randomly assigned IDs

- Netflix is not the only movie-rating portal on the web
- On IMDb, individuals can rate movies “not” anonymously
- Researchers from University of Texas at Austin, linked Netflix dataset with IMDb to de-anonymize the identity of some users

Differential Privacy

- Provide guarantees for your released dataset

- Formally
  - Maximize the accuracy of queries from statistical databases
  - While minimizing the chances of identifying its records

Studies

- Look at two studies
  - Originators of k-anonymity
  - De-anonymizing the Netflix dataset

K-anonymity: A model for Protecting Privacy

k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[1]

LATANYA SWEENEY

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
E-mail: latanya@cs.cmu.edu

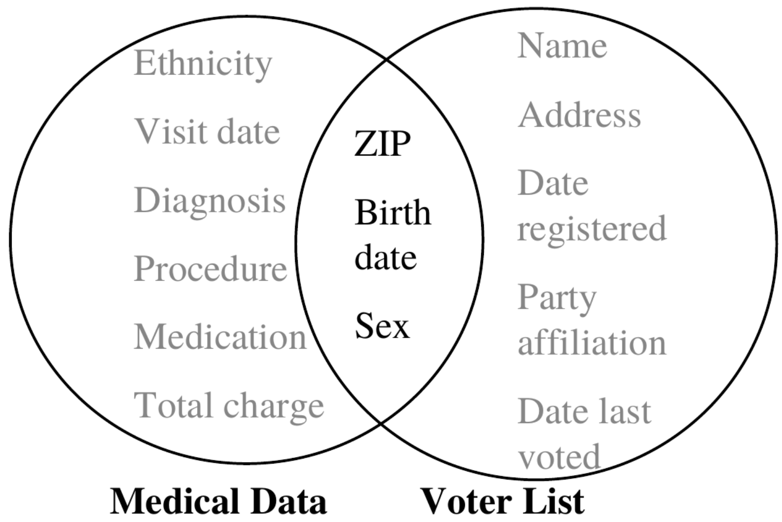
Received May 2002

Consider a data holder, such as a hospital or a bank, that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? The solution provided in this paper includes a formal protection model named k-anonymity and a set of accompanying policies for deployment. A release provides k-anonymity protection 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. This paper also examines re-identification attacks that can be realized on releases that adhere to k-anonymity unless accompanying policies are respected. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, μ-Argus and k-Similar provide guarantees of privacy protection.

Keywords: data anonymity, data privacy, re-identification, data fusion, privacy.

Sweeney. K-anonymity: A model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002

Re-identification by Linking



Re-identification of Individuals

- William Weld: Governor of MA at the time
- His medical record in the Group Insurance Commission (GIC) data
- Lived in Cambridge, MA
- From the voter list
  - Six people with his particular birth date
  - Three of them male
  - He was the only one in his ZIP code

Statistical Databases

- Data: Person-specific information organized as a table of rows and columns
- Tuple: Corresponds to a row, describes the relationship among the set of values for a person
- Attribute: Corresponds to a column, describes a field or semantic category of information

Quasi-Identifiers

- Attributes that in combination can uniquely identify individuals
- Such as ZIP, gender, and date of birth
- Data owner should identify the quasi-identifier

Sensitive vs Nonsensitive Attributes

Zip Code	Gender	Date of Birth	Medical Condition
**	**	**	**
**	**	**	**

`nonsensitive' (at least individually)

sensitive

Exercise: Column Combinations

- Table with three columns
  - Physician
  - Patient
  - Medication
- Which combinations are sensitive?
  - R(Physician, Patient): Sensitive?
  - R(Physician, Medication): Sensitive?
  - R(Patient, Medication): Sensitive?

K-Anonymity: Formal Definition

- Informally, your information contained in the released dataset cannot be distinguished from at least k-1 other individuals whose information also appear in the dataset
- Formally,
  - Let RT($A_1, \ldots, A_n$) be a table
  - Let QI_{RT} be the quasi-identifier for RT
  - RT satisfies k-anonymity if and only if each sequence of values in RT[QI_{RT}] appears with at least k occurrences

Methods to Achieve K-anonymity

- Suppression: Values replaced with '*'
  - All or some values of a column may be replaced
  - Attributes such as "Name" or "Religion"
- Generalization: Values replaced with a broader category
  - '19' of the attribute "Age" may be replaced with '≤ 20'
  - Replace '23' with '$20 < \text{Age} \leq 30$'

Example K-Anonymous Table

	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

- QI = {Race, Birth, Gender, ZIP}
- k = 2

More Examples

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

Exercise: Make This Table 4-anonymous

	Zip code	Age	Nationality	Condition
1	27609	18	Chinese	Heart Disease
2	27615	19	American	Heart Disease
3	26724	50	Indian	Cancer
4	26724	55	Chinese	Heart Disease
5	27615	21	Japanese	Viral Infection
6	26725	47	American	Viral Infection
7	27609	23	American	Viral Infection
8	27609	31	American	Cancer
9	27615	36	Japanese	Cancer
10	26725	49	American	Viral Infection
11	27609	37	Indian	Cancer
12	27615	35	American	Cancer

One Solution

	Zip code	Age	Nationality	Condition
1	276**	<30	*	Heart Disease
2	276**	<30	*	Heart Disease
3	2672*	≧40	*	Cancer
4	2672*	≧40	*	Heart Disease
5	276**	<30	*	Viral Infection
6	2672*	≧40	*	Viral Infection
7	276**	<30	*	Viral Infection
8	276**	3*	*	Cancer
9	276**	3*	*	Cancer
10	2672*	≧40	*	Viral Infection
11	276**	3*	*	Cancer
12	276**	3*	*	Cancer

L-diversity

276**	3*	*	Heart Disease
276**	3*	*	Cancer
276**	3*	*	Viral Infection
276**	3*	*	Flu

Machanavajjhala et al. l-diversity: Privacy beyond k-anonymity. International Conference on Data Engineering, 2006

L-diversity Solution

276**	3*	*
276**	3*	*
276**	3*	*
276**	3*	*

Exercise: L-diversity

	Zip code	Age	Nationality	Condition
1	276**	<30	*	Cancer
2	276**	<30	*	Cancer
3	2672*	≧40	*	Flu
4	2672*	≧40	*	Heart Disease
5	276**	<30	*	Heart Disease
6	2672*	≧40	*	Heart Disease
7	276**	<30	*	Heart Disease
8	276**	3*	*	Flu
9	276**	3*	*	Heart Disease
10	2672*	≧40	*	Flu
11	276*	3*	*	Flu
12	276**	3*	*	Heart Disease

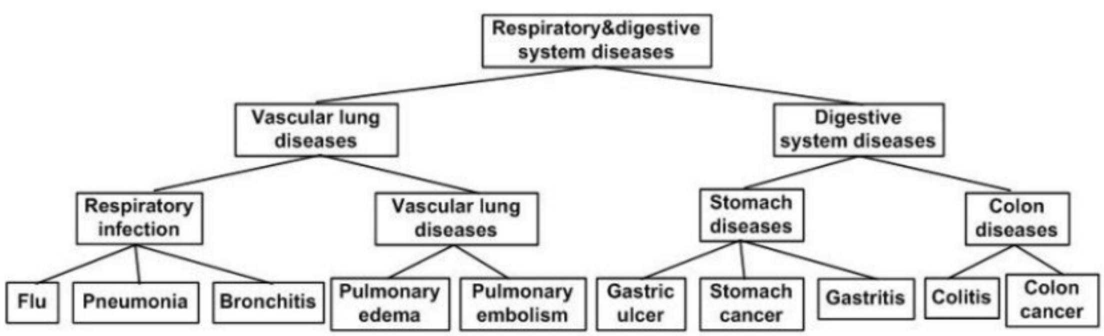
L-diversity Blocks

	Zip code	Age	Nationality	Condition
1	276**	<30	*	Cancer
2	276**	<30	*	Cancer
7	276**	<30	*	Heart Disease
5	276**	<30	*	Heart Disease
3	2672*	≧40	*	Flu
4	2672*	≧40	*	Heart Disease
6	2672*	≧40	*	Heart Disease
10	2672*	≧40	*	Flu
8	276**	3*	*	Flu
9	276**	3*	*	Heart Disease
11	276*	3*	*	Flu
12	276**	3*	*	Heart Disease

L-diversity Concerns

- Some medical conditions are more sensitive than others
- Some medical conditions may indicate same disease

T-closeness



- Measure semantic distance between concepts

Li et al. t-closeness: Privacy beyond k-anonymity and l-diversity. International Conference on Data Engineering, 2007

Example T-closeness Table

Zip code	Age	Disease
4767*	<40	Gastric ulcer
4767*	<40	Stomach cancer
4767*	<40	Pneumonia
4790*	>39	Gastritis
4790*	>39	Flu
4790*	>39	Bronchitis
2760*	<40	Gastritis
2760*	<40	Bronchitis
2760*	<40	Stomach cancer