K-Anonymity: A Model for Protecting Privacy

HUM 4441: Engineering Ethics

Dr. Mohammad Rezwanul Huq
Adjunct Faculty, IUT

Introduction

- The paper, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY by L. Sweeney, addresses disclosures based on inferences that can be drawn from released data (vs. access control and authentication protections).
- Problem Statement:

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 Proposed Work

1) A formal protection model named *k*-anonymity *(**key contribution**):*

   - ❑ The released information is enforced to map to many *(**k**)* possible “people”*.*
   - ❑ The greater k is made, the more anonymous the released information become.

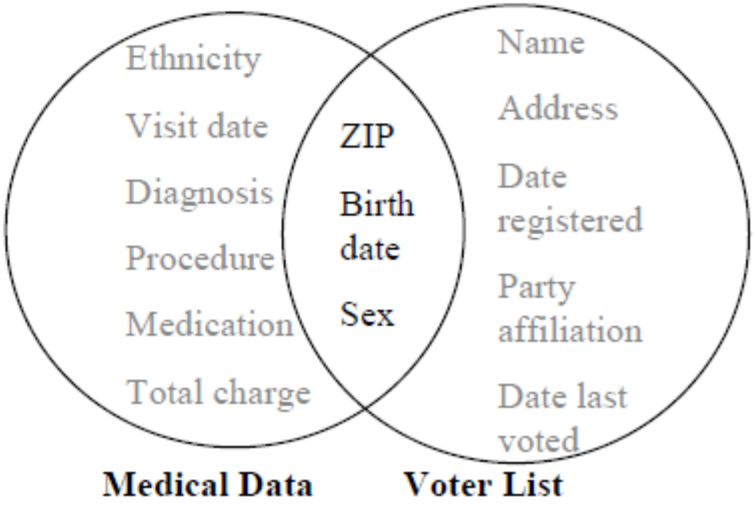
2) Some re-identification attacks that can be realized on releases that adhere to *k*-anonymity.

3) A set of accompanying policies that if deployed can thwart the presented attacks.

Talk roadmap

- Example: Re-identification by linking.
- The k-anonymity protection model:
  - ❑ Quasi-identifier.
  - ❑ K-anonymity: Definition and Example.
- Attacks against k-anonymity:
  - ❑ Unsorted Matching.
  - ❑ Complementary Release.
  - ❑ Temporal.
- Concluding remarks (strengths and weaknesses).

Re-Identification By Linking

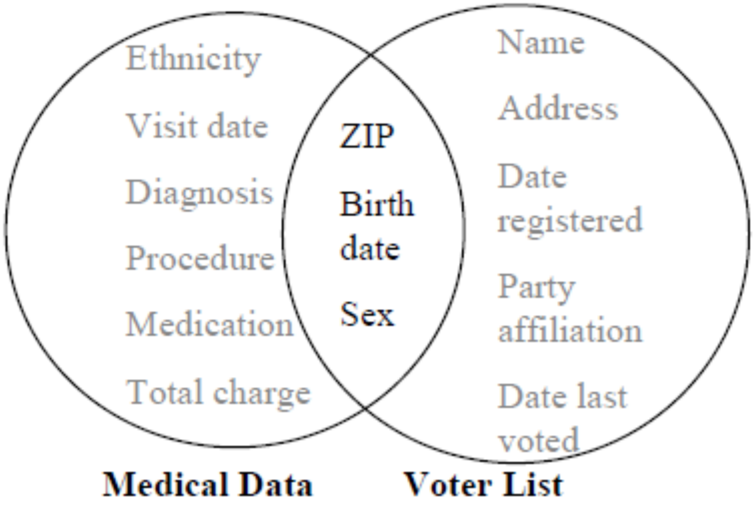


This information can by linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, **procedures, and medications to particularly named individuals.**

K-anonymity Model

- **Objective:** Released information limits what can be revealed about properties of the entities that are to be protected.

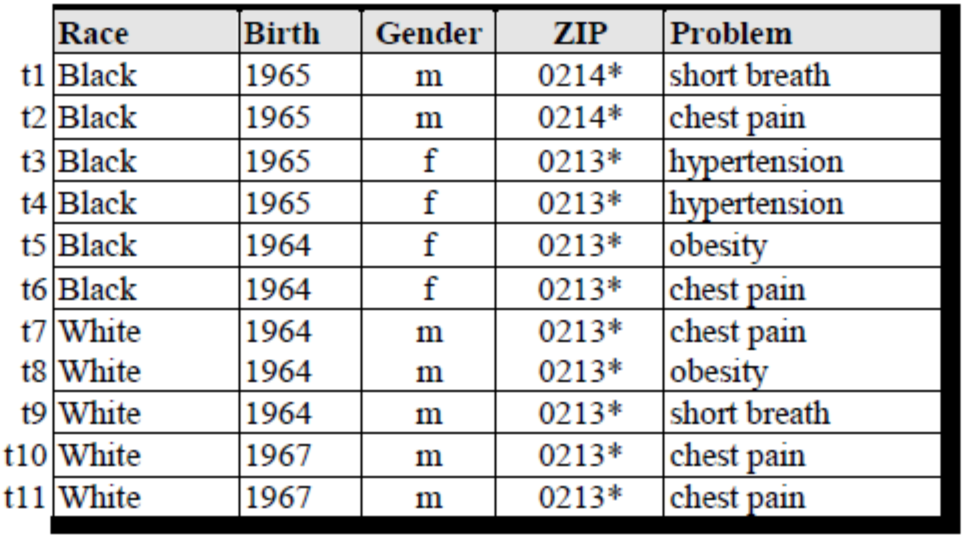
- **Quasi-identifier:** set of attributes that can be lined with external data to uniquely identify individuals in the population (e.g., ZIP code, gender, and date of birth).



The data holder can accurately identify quasi-identifiers.

K-anonymity Model

- **Definition:** Let RT(A_1, …, A_n) be a table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT[QI_{RT}] appears with at least k occurrences in RT[QI_{RT}].



	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} and K = 2.

Attacks Against k-Anonymity

- **Unsorted Matching:** based on the order in which tuples appear in the released tables.

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

QI_{PT} = {Race, ZIP} & k = 2.

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

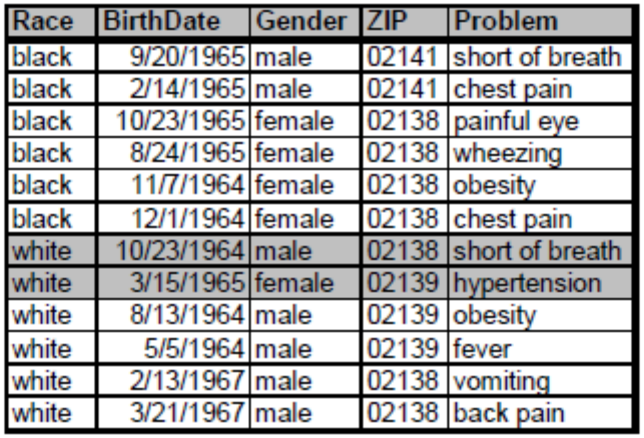
Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Solution: Randomly sort the tuples of the released tables.

Attacks Against k-Anonymity

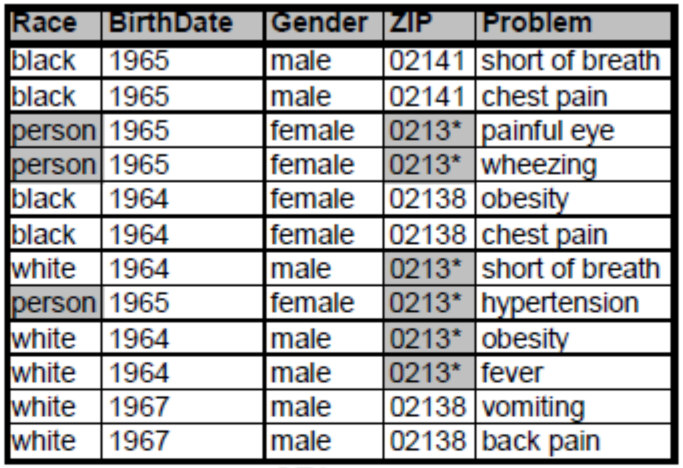
- **Complementary Release:** based on the common fact that the attributes that constitute the quasi-identifier are themselves a subset of the attributes released.



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

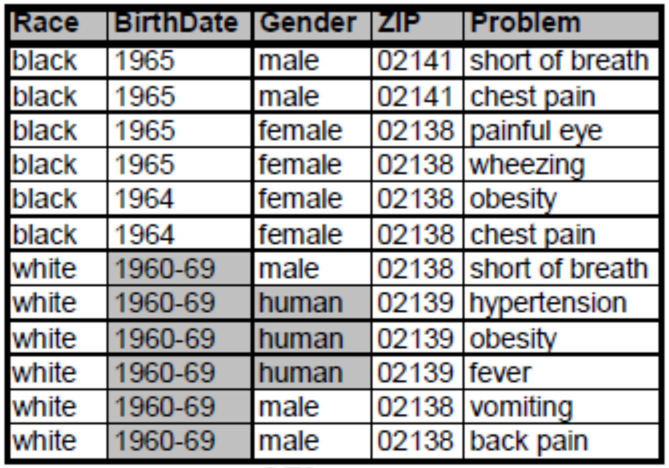
PT

QI_{PT} = {Race, BirthDate, Gender, ZIP} & k = 2.



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

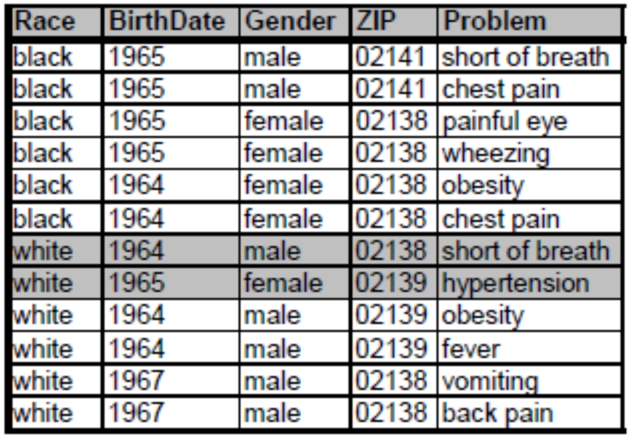


Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human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

- **Linking GT1 and GT3 on {Problem} reveals the table LT.**



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

Solution: Subsequent releases of the same privately held information must consider "all" of the released attributes of T a quasi-identifier, or subsequent releases themselves would be based on T.

Attacks Against k-Anonymity

- **Temporal attack:** based on the fact the data collections are dynamic:

  a) At time t_0, let table T_0 be the original privately held table.
  b) Release RT_0.
  c) At time t_1, additional tuples are added to T_0 → T_t.
  d) Release $RT_{t.}$
  e) Because no requirement that RT_t respect RT_0, linking the tables RT_0 and RT_t may reveal sensitive information and thereby compromise k-anonymity protection.

- **Solution:** Subsequent releases of the same privately held information must consider "all" of the released attributes of RT_0 a quasi-identifier, or subsequent releases themselves would be based on RT_0.