k-Anonymity and Other Cluster-Based Methods

Slides mostly from

Vitaly Shmatikov
and Li Xiong

Background

- Large amount of person-specific data has been collected in recent years
  - Both by governments and by private entities
- Data and knowledge extracted by data mining techniques represent a key asset to the society
  - Analyzing trends and patterns.
  - Formulating public policies
- Laws and regulations require that some collected data must be made public
  - For example, Census data

Public Data Conundrum

- Health-care datasets
  - Clinical studies, hospital discharge databases …
- Genetic datasets
  - $1000 genome, HapMap, deCode …
- Demographic datasets
  - U.S. Census Bureau, sociology studies …
- Search logs, recommender systems, social networks, blogs …
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon …

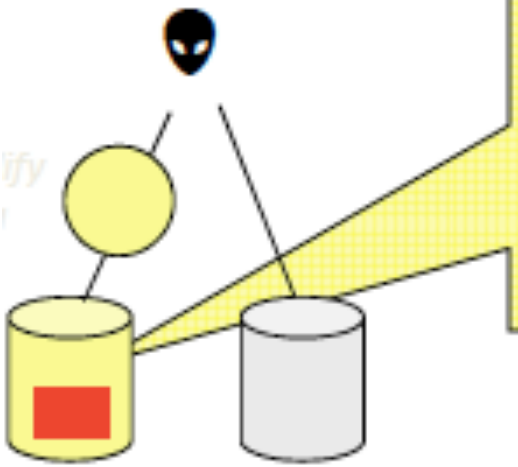
Inference Control

- **Inference control: protecting private data while publishing useful information**

Related aspects

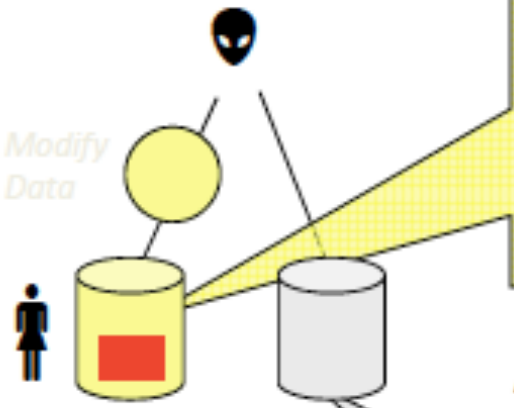
- Access control: protecting information from unauthorized access and use.
- Disclosure Control: *modification of data*, containing confidential information about individual entities *in order to prevent third parties working with these data to recognize individuals in the data*

Motivating Example



	Non-Sensitive Data			Sensitive Data	
#	Zip	Age	Nationality	Name	Condition
1	13053	28	Brazilian	Ronaldo	Heart Disease
2	13067	29	US	Bob	Heart Disease
3	13053	37	Indian	Kumar	Cancer
4	13067	36	Japanese	Umeko	Cancer

Motivating Example (continued)



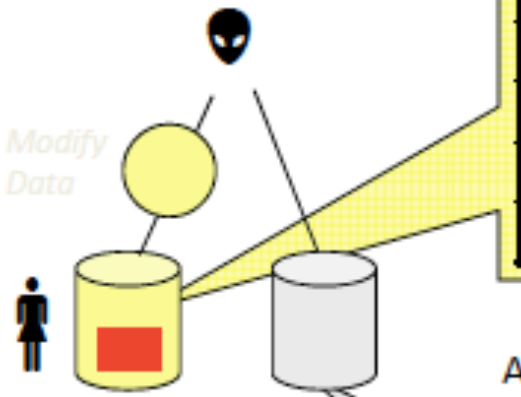
Published Data: Alice publishes data without the Name

	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
1	13053	28	Brazilian	Heart Disease
2	13067	29	US	Heart Disease
3	13053	37	Indian	Cancer
4	13067	36	Japanese	Cancer

Attacker's Knowledge: Voter registration list

#	Name	Zip	Age	Nationality
1	John	13067	45	US
2	Paul	13067	22	US
3	Bob	13067	29	US
4	Chris	13067	23	US

Motivating Example (continued)



Published Data: Alice publishes data without the Name

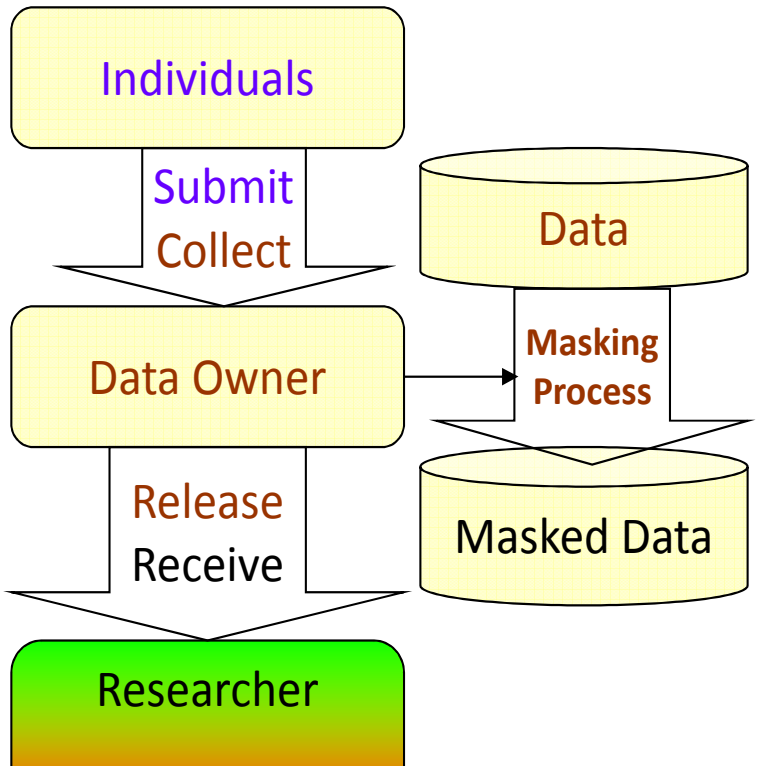
	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
1	13053	28	Brazilian	Heart Disease
2	13067	29	US	**Heart Disease**
3	13053	37	Indian	Cancer
4	13067	36	Japanese	Cancer

Attacker's Knowledge: Voter registration list

#	Name	Zip	Age	Nationality
1	John	13067	45	US
2	Paul	13067	22	US
3	**Bob**	13067	29	US
4	Chris	13067	23	US

Data Leak !

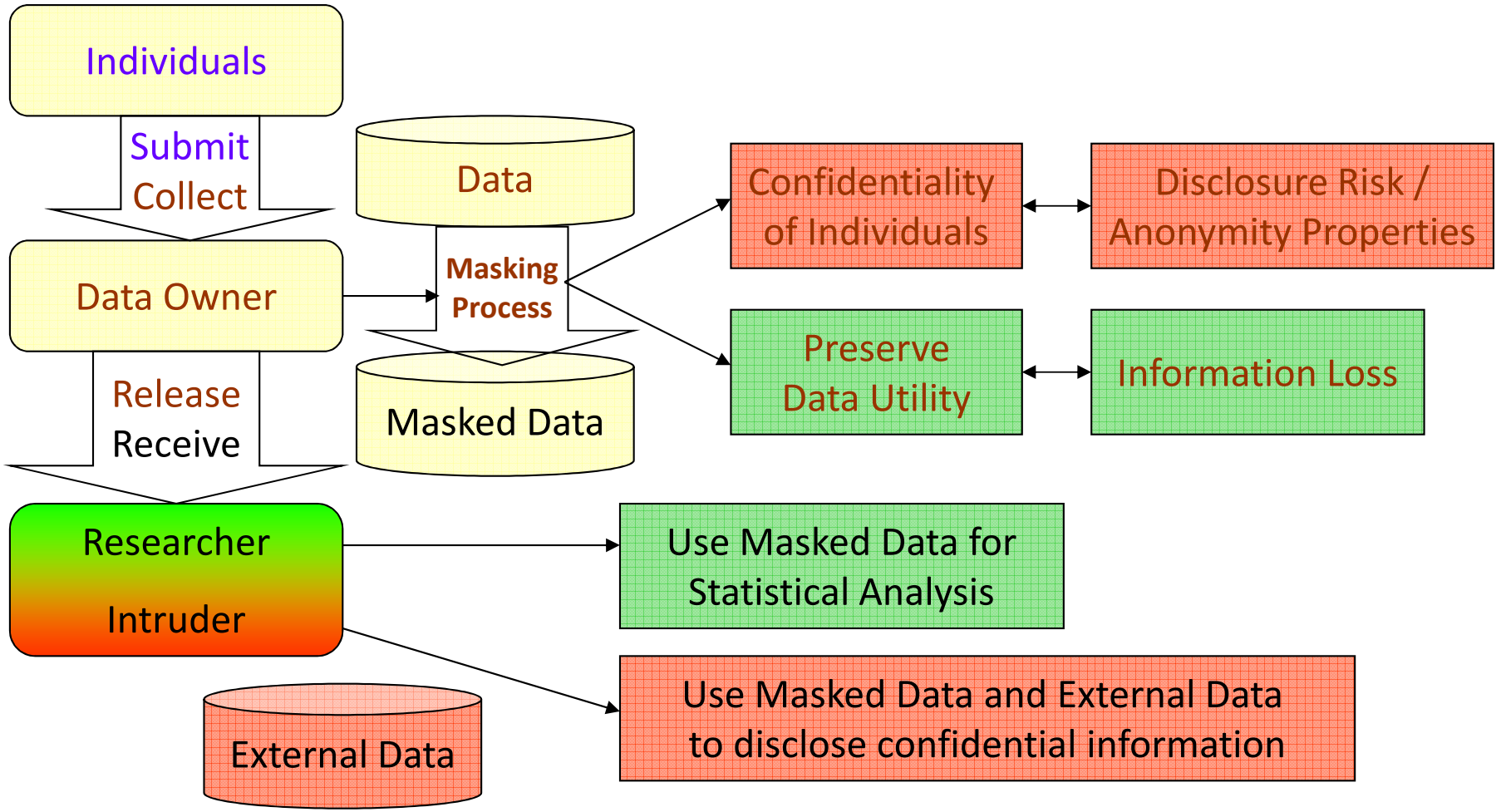
Inference Control?



Disclosure Control

is concerned with the modification of data, containing confidential information about individual entities such as persons, households, businesses, etc. in order to prevent third parties working with these data to recognize individuals in the data

Disclosure Control Problem



What About Privacy?

Types of disclosure

- Identity disclosure - identification of an entity (person)
- Attribute disclosure - the intruder finds something new about the Target person

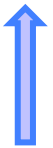
- Anonymize the data. How?
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address… what else?
  - Anything that identifies the person directly

What About Privacy?

- Anonymize the data. How?
- Remove "personally identifying information" (PII)

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease



Not public

Latanya Sweeney's Attack -1997

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

Microdata Macrodata and external info

- Microdata represents a series of records, each record containing information on an individual unit such as a person, a firm, an institution, etc
- Macrodata: contains computed data (e.g statistics)
- Masked Microdata names and other identifying information are removed or modified from microdata
- External Information any known information by a presumptive intruder related to some individuals from initial microdata

Source of the problem

- Even if we do not publish the individuals data
- There are some fields that may uniquely identify some individual
- The attacker can use them to join with additional sources and identify the individuals

Microdata protection

Re-identification by Linking

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Not public

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Other public data

The private info on Alice's diseases can be inferred by voter registration data and ZIP codes

Macrodata protection

- Macrodata: contains computed data (e.g statistics, count numbers)

	Hypertension	Obesity	Chest Pain	Short Breath	Tot
M	1	2	2	1	6
F	1	2	0	2	5
Tot	2	4	2	3	11

(a) number of respondents with a disease

Macrodata protection

- Several macrodata protection
- First step: discover sensitive cells, that is, cells that can be easily associated with a specific respondent.
  - Example: if there is only one person with HIV then this might be easily defined
- For count and frequency tables, the most important strategy used to detect sensitive cells is the **threshold rule** , according to which

a cell is sensitive if the number of respondents is less than a given threshold.

Macrodata protection

- Cell suppression consists in protecting sensitive cells by removing their values.
- These are called primary suppressions.
- However, a problem can arise when also the marginal totals of the table are published. In this case, even if it can be possible to calculate an interval that contains the suppressed cell. If the size of such an interval is small, then the suppressed cell can be estimated rather precisely.
- To block such inferences, additional cells may need to be suppressed (secondary suppression) to guarantee that the intervals are sufficiently large.

Macrodata protection

- For magnitude macrodata, there are many rules that can be used to detect sensitive cells.
- For instance, the (n,k)-rule states that a cell is sensitive if less than n respondents contribute to more than k % of the total cell value.
- Example: the (1,50)-rule. A cell is therefore sensitive if one respondent contributes to more than 50% of its value.

Macrodata protection

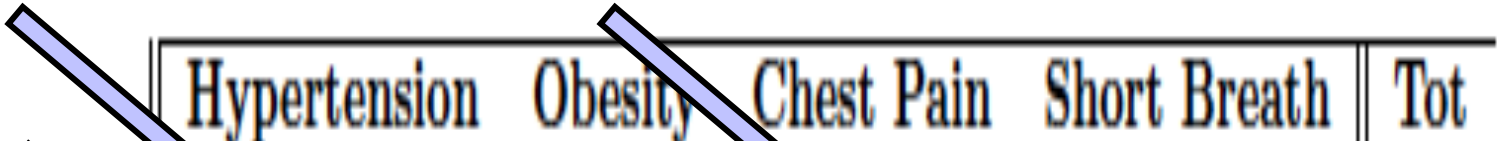
- Example, consider the macrodata table and suppose to apply the (1,50)-rule. A cell is therefore sensitive if one respondent contributes to more than 50% of its value
- Self identification: even if there is no identification one person can identfy himslf/herself. This is a problem wrt the acceptance of the data

	Hypertension	Obesity	Chest Pain	Short Breath	Tot
M	1	2	2	1	6
F	1	2	0	2	5
Tot	2	4	2	3	11

(a) number of respondents with a disease

Macrodata protection

- Example, consider the macrodata table and suppose to apply the (1,50)-rule. A cell is therefore sensitive if one respondent contributes to more than 50% of its value.



	Hypertension	Obesity	Chest Pain	Short Breath	Tot
M	1	2	2	1	6
F	1	2	0	2	5
Tot	2	4	2	3	11

(a) number of respondents with a disease

Macrodata protection

- Other rules are the p-percentage rule
- The p-percentage states that a cell is sensitive if the total value t of the cell minus the largest reported value v1 minus the second largest reported value v2 is less than (p /100)$*$v1 .
- Intuitively, this rule means that a user can estimate the reported value of some respondent too accurately.
- Example: if we have statistics about salary of the area where Bill Gates lives and other people have a mall income we can guess the salary of Bill Gates

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S. (ZIP = CAP in Italy)
  - Can be used for linking anonymized dataset with other datasets

Classification of Attributes

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes is what the researchers need, so they are always released directly

Key Attribute | **Quasi-identifier** | **Sensitive attribute**

Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

K-Anonymity Protection Model

- Private table
- Released table: RT
- Attributes: A_1, A_2, …, A_n
- Quasi-identifier subset: A_i, …, A_j

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 ⊆ 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.

- The record with Bob's information has k-1 similar records

Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Main techniques
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values
- Suppression: do not release a value at all

Generalization

Main techniques

- Generalization: replace quasi-identifiers with less specific, but semantically consistent values
- Example: change age =29 with age < 40

Generalization group values (age), eventually ignore (nationality)

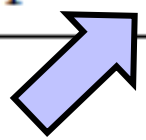
#	Zip	Age	Nationality	Condition
1	41076	< 40	*	Heart Disease
2	48202	< 40	*	Heart Disease
3	41076	< 40	*	Cancer
4	48202	< 40	*	Cancer

Generalization

Suppression

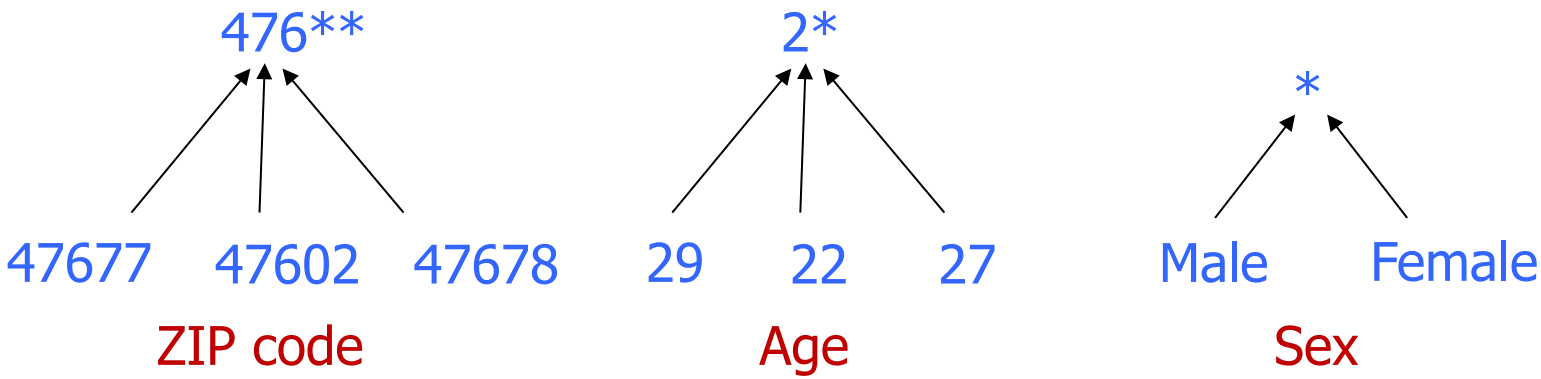
- Suppression: do not release a value at all in some tuple (sensitive)

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Disease	DH	Chol	Temp
		Asian	64/09/27	F	9413*	Divorced	Hypertension	3	260	nf
		Asian	64/09/30	F	9413*	Divorced	Obesity	1	<195	f
		Asian	64/04/18	M	9413*	Married	Chest pain	>30	200	f
		Asian	64/04/15	M	9413*	Married	Obesity	7	280	f
		Black	63/03/13	M	9413*	Married	Hypertension	2	<195	nf
		Black	63/03/18	M	9413*	Married	Short breath	3	<195	f
		Black	64/09/13	F	9414*	Married	Short breath	5	200	nf
		Black	64/09/07	F	9414*	Married	Obesity	>30	290	hf
		White	61/05/14	M	9413*	Single	Chest pain	7	<195	f
		White	61/05/08	M	9413*	Single	Obesity	10	300	hf
		White	61/09/15	F			Short breath	5	200	nf



Generalization

- Generalization: replace quasi-identifiers with less specific, but semantically consistent values



Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- Suppression
  - When generalization causes too much information loss
    - This is common with “outliers”
- Lots of algorithms in the literature
  - Aim to produce “useful” anonymizations

    ... usually without any clear notion of utility

Generalization in Action

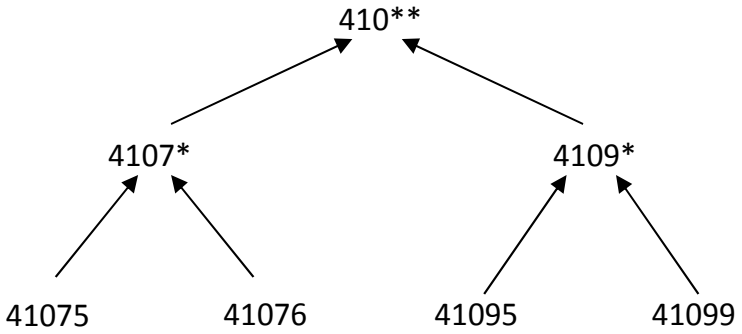
Z2 = {410**}

↑

Z1 = {4107*. 4109*}

↑

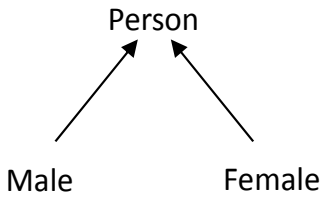
Z0 = {41075, 41076, 41095, 41099}



S1 = {Person}

↑

S0 = {Male, Female}



Generalization in Action: lattice

S1 = {Person}

↑

S0 = {Male, Female}

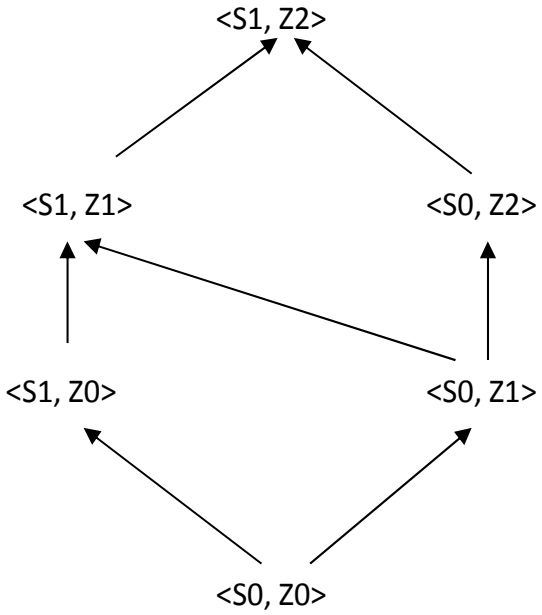
Z2 = {410**}

↑

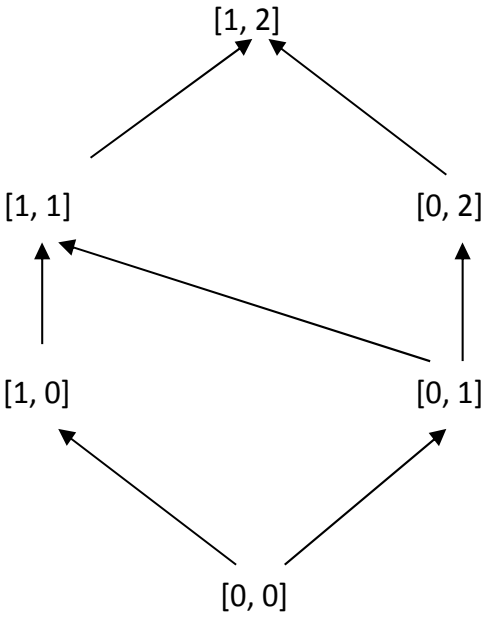
Z1 = {4107*, 4109*}

↑

Z0 = {41075, 41076, 41095, 41099}



Generalization Lattice



Example of a 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

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

Example of Generalization (1)

Released 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

External data Source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

!!

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

K-anonimity: objective

The objective in this context is to generalize or to suppress data in a given database until it becomes k-anonymized, **while incurring a minimal loss of information**.

Let **Π(D, g(D)) be a measure of the amount of information that is lost** by replacing a database D with a corresponding generalization/ suppression g(D).

Then the problem of k- anonymization is as follows.

Let $D = \{R1, \ldots, Rn\}$ be a database with public attributes Aj, $1 \leq j \leq r$. Given collections of attribute values, $Aj \subseteq P(Aj)$, and a measure of information loss Π, find a corresponding k-anonymization, $g(D) = \{R1,\ldots,Rn\}$, that minimizes $\Pi(D,g(D))$.

Not surprisingly the problem is NP-hard foor natural objective functions

Health Data: HIPAA Privacy Rule (US)

Importance of publicizing statistical health data for legal and economic issues (insurance)

Examples

- Smoking and prob. of cancer
- Statistics concerning health data close to a pollutted area (e.g. Italsider in Taranto)

HIPAA privacy rules (adopted in the US)

"*Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information*."

HIPAA Privacy Rule: Safe Harbor

"The identifiers that must be removed include direct identifiers, such as **name, street address, social security number**, as well as other identifiers, such as **birth date**, admission and discharge dates, and **five-digit zip code**. The safe harbor … [allows] the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."

Safe Harbor's 18 Identifiers

Names

All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes

- **Except for the initial three digits of a zip code if according to the currently available data from the Bureau of the Census:**
  - **The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and**
  - **The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people are changed to 000;**

All elements of dates (except year) or dates directly relating to an individual, including:

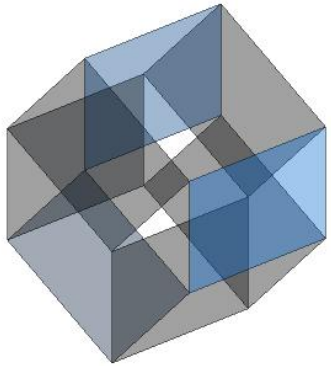
- **birth date, admission date, discharge date, date of death;**
- **and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;**

- **Telephone numbers;**
- **Fax numbers;**
- **Electronic mail addresses;**
- **Social security numbers;**
- **Medical record numbers;**
- **Health plan beneficiary numbers;**
- **Account numbers;**
- **Certificate/license numbers;**
- **Vehicle identifiers and serial numbers, including license plate numbers;**
- **Device identifiers and serial numbers;**
- **Web Universal Resource Locators (URLs);**
- **Internet Protocol (IP) address numbers;**
- **Biometric identifiers, including finger and voice prints;**
- **Full face photographic images and any comparable images; and**
- **Any other unique identifying number, characteristic, or code.**

Problem: Curse of Dimensionality

[Aggarwal VLDB '05]

- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info ⇒ k-anonymized datasets are useless



Two (and a Half) Interpretations

- Membership disclosure: Attacker cannot tell that a given person is in the dataset
- Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute
- Identity disclosure: Attacker cannot tell which record corresponds to a given person

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers

But this does not imply any privacy!

Example: k clinical records, all HIV+

Unsorted Matching Attack

- Problem: records appear in the same order in the released table as in the original table
- *Solution: randomize order before releasing*

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

Complementary Release Attack

- Different releases of the same private table can be linked together to compromise k-anonymity

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

Linking Independent Releases

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 on k-Anonymity

- 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

l-Diversity (ℓ-Diversity)

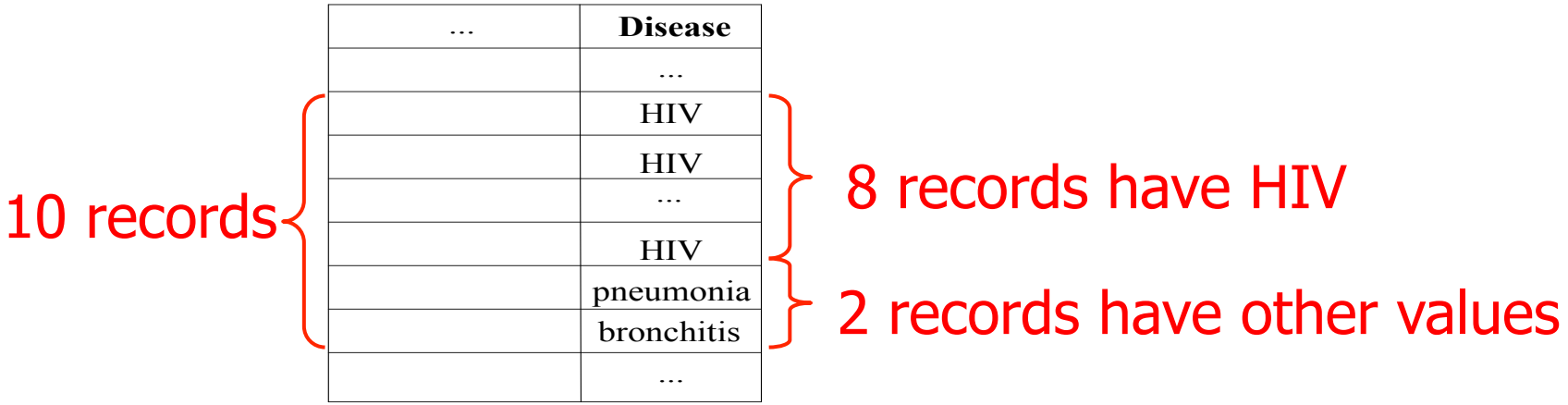
[Machanavajjhala et al. ICDE '06]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Distinct ℓ-Diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks:
  in the example attacker knows HIV with 80% prob.



ℓ-Diversity: example

Original dataset

…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Flu

99% have cancer

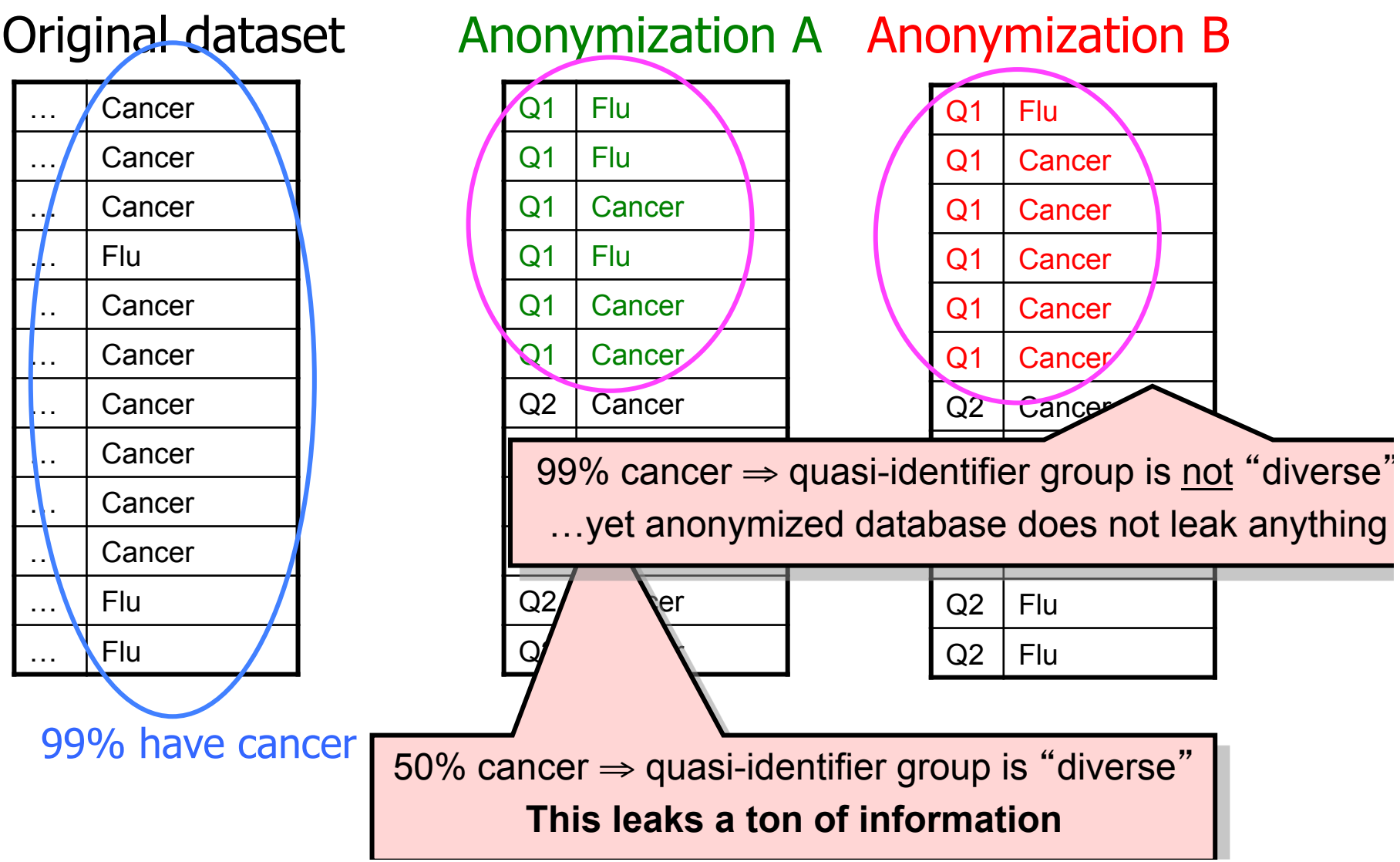
Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer
Q2	Cancer
Q2	Cancer
Q2	Cancer
Q2	Cancer
Q2	Cancer

Anonymization B

Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	Cancer
Q2	Cancer
Q2	Cancer
Q2	Cancer
Q2	Flu
Q2	Flu

ℓ-Diversity: example



Limitations of l-Diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
  - Very different degrees of sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

l-diversity does not consider overall distribution of sensitive values!

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	⩾40	50K	Gastritis
4790*	⩾40	100K	Flu
4790*	⩾40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

Other Versions of ℓ-Diversity

- Probabilistic ℓ-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by 1/l
- Entropy ℓ-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least log(ℓ)
- Recursive (c, ℓ)-diversity
  - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$ where r_i is the frequency of the i^{th} most frequent value
  - Intuition: the most frequent value does not appear too frequently

t-Closeness

[Li et al. ICDE '07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

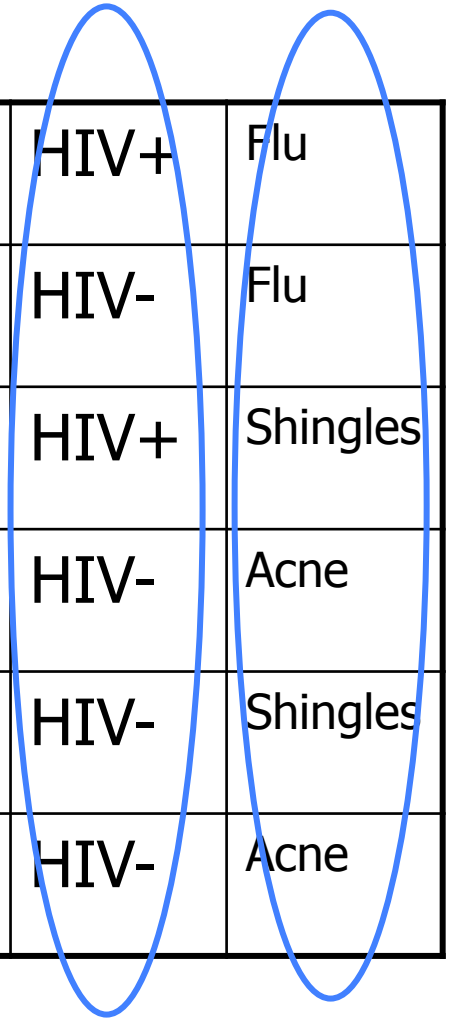
Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Trick question: Why publish quasi-identifiers at all??

Recall: we want to publish useful information that increases our knowledge

Anonymous, "t-Close" Dataset

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne



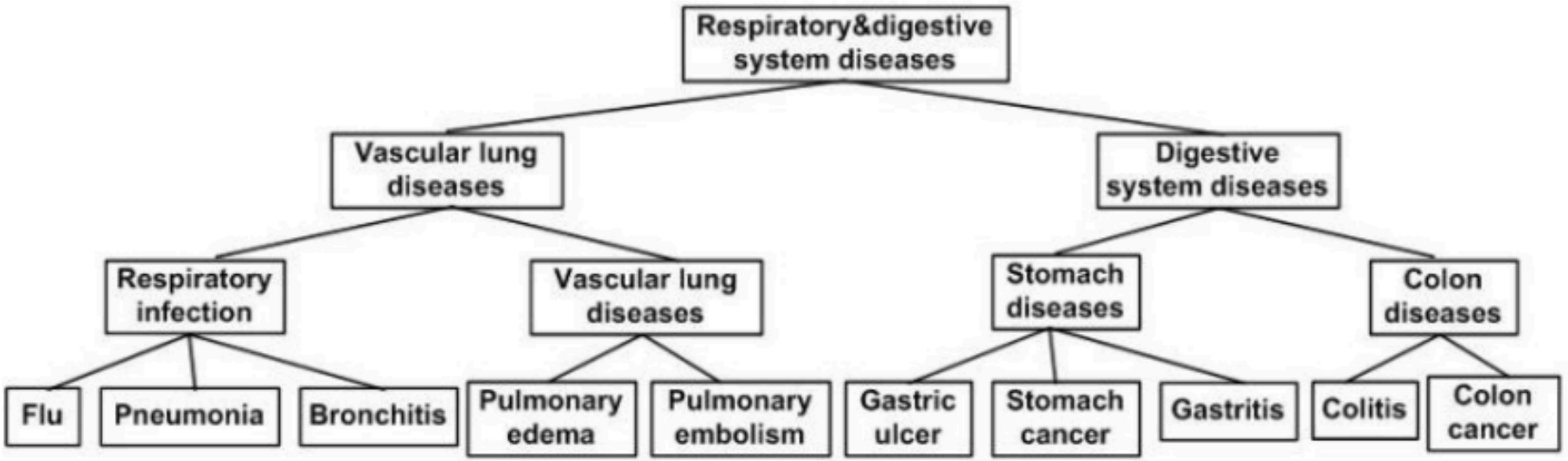
This is k-anonymous, l-diverse and t-close…

…so secure, right?

T-closeness:

Measure semantic distance between concepts

It requires generalization of sensitive information



K-anonymity: conclusions

Pros of k-Anonymity

- The cost of incurred in establishing this method is considerably lesser compared to the cost of another anonymity method such as cryptographic solution
- Algorithms of *k*-anonymity such as Datafly, Incognito, and Mondrian are used extensively, especially in PPDP.

Cons of k-Anonymity

- many limitations that have been identified:
  - unsorted matching, complementary release, Homogeneity attack
  - this technique can cause high utility loss if it is employed in high-dimensional data and / or if the released data has already undergone anonymization more than once

l-Diversity : conclusions

Pros of l-Diversity

- Provides a greater distribution of sensitive attributes within the group, thus increasing data protection.
- Protects against attribute disclosure, an enhancement of *k*-anonymity technique.

Cons of l-Diversity

- *l*-diversity can be redundant and laborious to achieve.
- Prone to attacks such as skewness attack and similarity attack as it is inadequate to avoid attribute exposure due to the semantic relationship between the sensitive attributes

T-closeness: conclusions

Pros of t-Closeness

- It interrupts attribute disclosure to protects data privacy
- Protects against homogeneity and background knowledge attacks mentioned in *k*-anonymity.
- It identifies the semantic closeness of attributes, a limitation of *l*-diversity.

Cons of t-Closeness

- Using Earth Mover's Distance (EMD) measure in *t*-closeness, it is hard to identify the closeness between *t*-value and the knowledge gained.
- Necessitates that sensitive attribute spread in the equivalence class to be close to that in the overall table

Why k-anonymity (and related concepts) are not sufficient

K-anonymity and related concepts are based on the concept of PII – personally identifiable information; PII given and fixed set of attributes

1. that identifies a person or
2. With respect to which there is a reasonable basis to believe the information can be used to identify the individual.

In 2 above which kind of information can we use?

- K-anonymity and related concepts are **purely synctatical**; they focus only on the data assuming PII ("*every combination of quasi-identifier values occurring in the data- set must occur at least k times*")
- HIPAA rules: 18 attributes are sufficient only if there is "*no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information*" (see also rules in UK)
- AOL fiasco (and other experience) shows that remaining attributes can be used to re-identification

Conclusion: PII has no guarantee and has no precise definition

Why k-anonymity (and related concepts) are not sufficient

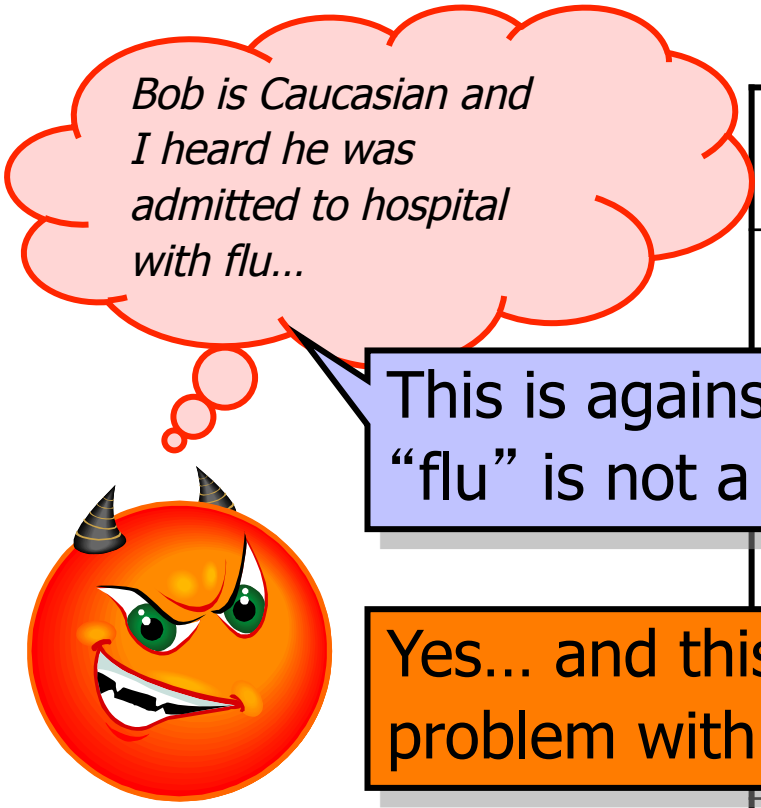
Re-identification without PII: AOL fiasco shows that remaining attributes can be used to re-identification

- Other experiments show that many other attributes can be used: previous transactional behaviour, location information, writing style, web browsing, search history, etc.
- Key properties of these set of attributes
  - they are reasonably stable across time and contexts,
  - the corresponding data attributes are sufficiently numerous and fine-grained that no two people are similar, except with a small probability
  - Example: consider books; a single book that a person read and evelauted does not mean too much**; the collection of books read by a person identify that person (if she/he has read a sufficient number of books).**
  - The above properties are verified in many cases (curse of dimensionality)

Why k-anonymity (and related concepts) are not sufficient

- HIPAA rules: 18 attributes are sufficient only if there is "*no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information*"
- Actual experience has shown that *any remaining attributes can be used for re-identification, as long as they differ from individual to individual*.
- Conclusion: PII has no meaning even in the context of the HIPAA Privacy Rule. It allows to increase the complexity of the attacker
- We need more sophisiticated approach

If attacker knows something else?



Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
		IV+	Shingles
Caucas	787XX	HIV-	Acne
		IV-	Shingles
Caucas	787XX	HIV-	Acne

This is against the rules!
"flu" is not a quasi-identifier

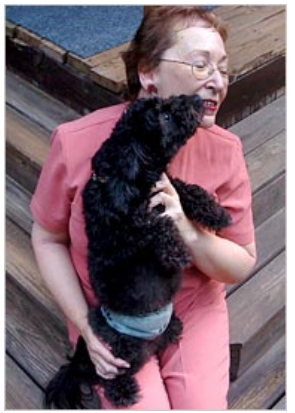
Yes… and this is yet another problem with k-anonymity

we use additional information that is not quasi identifier to identify….

AOL Privacy Debacle

- In August 2006, AOL released anonymized search query logs
  - 657K users, 20M queries over 3 months (March-May)
- Opposing goals
  - Analyze data for research purposes, provide better services for users and advertisers
  - Protect privacy of AOL users
    - Government laws and regulations
    - Search queries may reveal income, evaluations, intentions to acquire goods and services, etc.

AOL User 4417749



- AOL query logs have the form

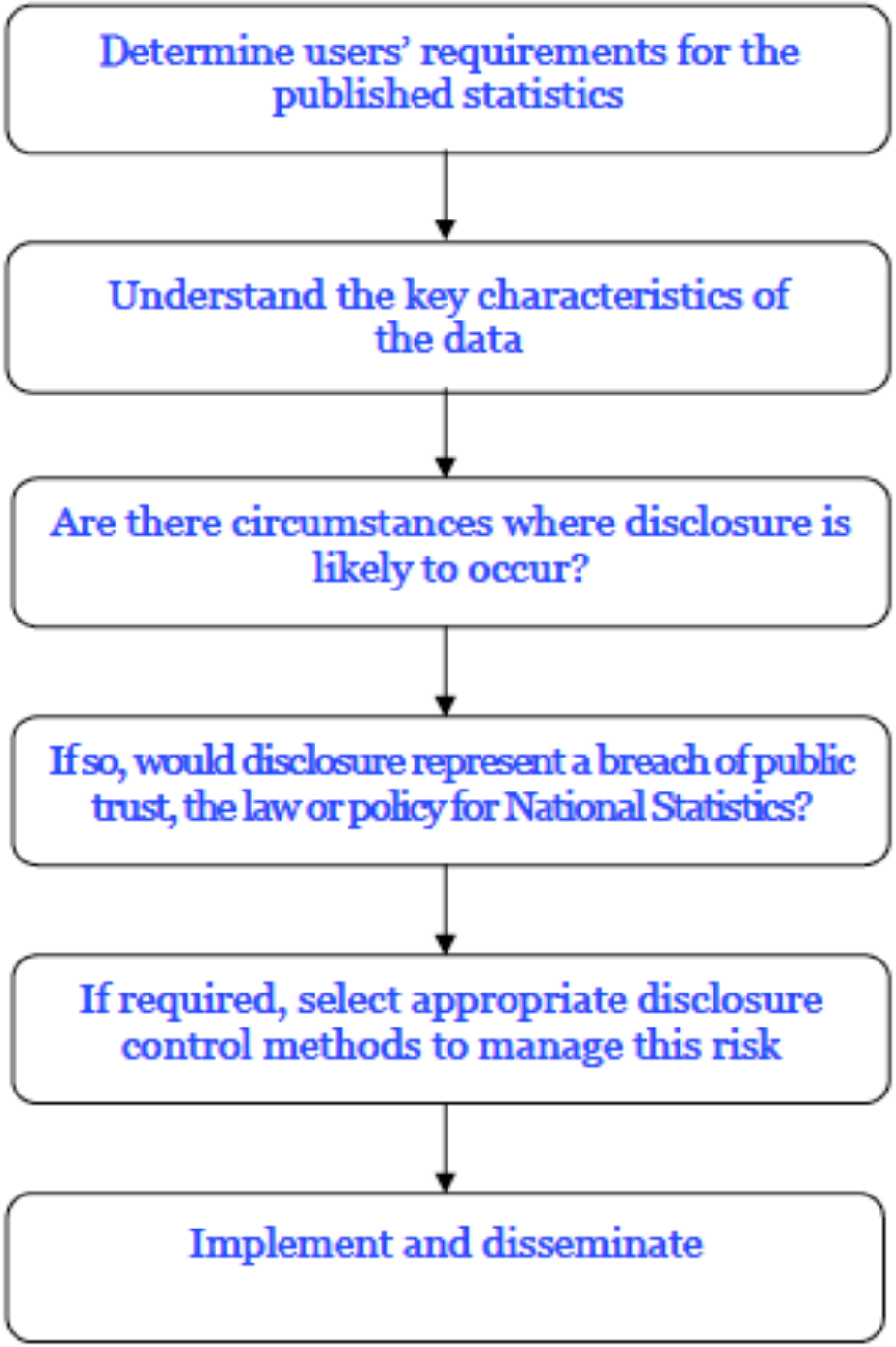
  <AnonID, Query, QueryTime, ItemRank, ClickURL>
  - ClickURL is the truncated URL
- NY Times re-identified AnonID 4417749
  - Sample queries: “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, GA”, several people with the last name Arnold
    - Lilburn area has only 14 citizens with the last name Arnold
  - NYT contacts the 14 citizens, finds out AOL User 4417749 is 62-year-old Thelma Arnold

k-Anonymity Considered Harmful

- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - "k-anonymous" dataset can leak sensitive information
- "Quasi-identifier" fallacy
  - Assumes a priori that attacker will not know certain information about his target
- Relies on locality
  - Destroys utility of many real-world datasets

UK confidentiality guidance

- not limited to synctactic rules
- try to understand semantic of data
- flowchart



UK confidentiality guidance

Example

- A table of statistics for psychiatric services at a hospital shows admissions by single years of age, and diagnosis. Attribute disclosure has occurred if someone, who knows their neighbour was admitted for such service, discovers from the statistic that they are schizophrenic
- Questions: Disclosure in the following table?

Table 2: Treatment, by type and age

Treatment	Age				Total
	< 12	12–15	16–19	> 19	
Type 1	1	0	7	1	9
Type 2	0	0	18	19	37
Type 3	0	12	5	0	17
Total	1	12	30	20	63

UK: confidentiality guidance

Example

- A table of statistics for psychiatric services at a hospital shows admissions by single years of age, and diagnosis. Attribute disclosure has occurred if someone, who knows their neighbour was admitted for such service, discovers from the statistic that they are schizophrenic

Answers

- Small values. Explain
- Rows with many zero. Explain

UK: confidentiality guidance

The motivated intruder

- An intruder with a special interest in conception statistics discovers from a table that only a small number of very young women have conceived in a particular local area. The small number in the cell doesn't tell the intruder who the women are but it may prompt them to follow up other sources of information to locate the individuals and discover – and disclose - more details

Question: under which conditions this can occur?

UK: confidentiality guidance

The motivated intruder

- An intruder with a special interest in conception statistics discovers from a table that only a small number of very young women have conceived in a particular local area. The small number in the cell doesn't tell the intruder who the women are but it may prompt them to follow up other sources of information to locate the individuals and discover – and disclose - more details.

Answer: as in HIPAA and K- anonimity large aggregation imply large numbers

- This situation may occur when small values are reported for particular cells. In a large population (for example, a country or region), the effort and expertise required to discover more details about the statistical unit may be deemed to be disproportionate. As the base population is decreased by moving to smaller geographies or sub-populations, it becomes easier to find units and discover information.

UK: confidentiality guidance

Self identification

- A statistic showing attendance at a drug misuse clinic by age and sex has a count of 1 for a particular ward. The individual may in fact be the only person who knows who this 1 is but they may feel exposed by the statistic.
- Answer: this might be a confidentiality problem

  The individual might think data can be discolsed and there is no privacy protection; if this fear is communicated to their peers, it may spread, and the result may be a lack of trust in the confidentiality of their use of the clinic.

UK: Risk category

Risk categories are defined in terms of the likelihood of an attempt to identify individuals, and the impact of any identification

- Medium Risk (majority of health statistics) it will be sufficient to consider all cells of size 1 or 2 unsafe. Care should also be taken where a rowor column is dominated by zeros
- High Risk: likelihood of an identification attempt will be higher, and the impact of any successful identification would be great (eg abortions, AIDS/HIV, ..). All cells of size 1 to 4 are considered unsafe and care should be taken where a row or column is dominated by zeros. Higher levels of protection may be required for small geographical levels or for particular variables with an extremely high level of interest and impact.

UK confidentiality guidance: methods

Table redesign is recommended as a simple method that will minimise the number of unsafe cells and preserve original counts.

- grouping categories within a table aggregating to a higher level geography or for a larger population subgroup
- aggregating tables across higher gepgraphic level (e.g Zip code) or a number of years/ months (e.g. age)

UK confidentiality guidance: methods

Modify cell values

- Cell suppression Unsafe cells are not published. They are suppressed and replaced by a special character
- Rounding involves adjusting the values in all cells in a table to a specified base. This creates uncertainty about the real value for any cell (similar to aggrgating: exampe: express rounded income)
- Barnardisation: cells of every table are adjusted by +1, 0 or -1, according to probabilities

UK confidentiality guidance: methods

Adjust the data

- Swap pairs of records within a micro-dataset that are partially matched to alter the geographic locations attached to the records but leave all other aspects unchanged

Exercise: transform 2 anonymity

In the figure you can see some records of a health-care database. In the course ***generalization*** was mentioned as one possibility how to anonymize data. Another approach is to ***suppress*** some records.

i) Let the quasi identifiers be *Day of birth*, *sex*, *ZIP*, *Civil status*. Generalize the data in such a manner that $k = 2$ holds. The generalization should be as minimal as possible.

ii) Sometimes its better to suppress some records before doing the generalization step in order to keep more knowledge of the original data. Find a way to suppress as less records as possible in order to get more information after the subsequently generalization.

Exercise: transform 2 anonymity

Name	Day of birth	Sex	ZIP	Civil stand	Duration	Diagnosis
Hans Glück	11.03.59	male	1072	married	1	HIV
Robert Liebling	17.03.59	male	1276	married	7	Hepatitis
Emma Peel	01.07.60	female	1073	unmarried	2	Hepatitis
Isolde Isenthal	07.09.64	female	1077	unmarried	0	Chest pain
John Steed	02.07.69	male	1016	divorced	2	Tuberculosis
Lola Kornhaus	21.09.71	female	1267	divorced	4	Anemia
Molly Moon	24.12.78	female	1268	divorced	4	HIV

Zip cod erase 2 digits not sufficient considering sex

So either generalize sex or erase three digit of zip code

Exercise: transform 2 anonymity

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Make this table 4- anonymous

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

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

Exercise: 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

Complementary Release Attack

- PT is the original table
- GT1 and GT3 are 2-anonymous released tables

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
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

Complementary Release Attack

- PT is the original table
- LT is the linked table
- LT is not 2 anonymous

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

Exercise: This table is 4 anonymous; is it l–diverse?

	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

Solution The table is 2–diverse

2-diverse Block

2-diverse Block

2-diverse Block

	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

Exercise: l-diversity

This table is not 2 –diverse; make it 2- diverse

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

Exercise: l-diversity

This table is 2 –diverse; but it does not contain any useful information

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

exercise

The following table is also available

- Assume you know that Alice's data is within the released dataset. Where is she most likely born and what is most likely her relationship status? Describe how you deanonymized her. (Hint: You can find enough evidence for a unique solution)
- Can you deanonymize Charlie as well? If so: describe how. If not: describe why.
- Can you deanonymize Bob as well? If so: describe how. If not: describe why.

exercise

Dataset_1

Age	Gender	Fav.Show
19-25	female	Friends!
19-25	male	Friends!
19-25	male	Friends!
12-15	female	Friends!
19-25	male	G.o.T.
19-25	female	G.o.T.
19-25	male	G.o.T.

Dataset_2

Age	Gender	Fav.Show
19-25	female	Grey's A.
19-25	female	Simpsons
19-25	female	Futurama
19-25	female	Friends!
19-25	female	G.o.T.
19-25	female	C.Minds
19-25	female	Br.Ba.

Dataset_3

Age	Gender	Fav.Show
19	female	Friends!
19	male	Friends!
19	male	Friends!
19	female	Friends!
20	male	G.o.T.
20	male	G.o.T.
20	male	G.o.T.

- Consider the following quasi-identifiers
- Does Dataset1, Dataset2, Dataset3 satisfy k-anonymity? If so: what is the maximal k for which it satisfies k-anonymity? Explain your answer!

exercise

- i) Let the quasi identifiers be *Day o f birth*, *sex*, *ZIP*, *Civil status*. Generalize the data in such a manner that $k = 2$ holds. The generalization should be as minimal as possible.
- ii) Sometimes it is better to suppress some records before doing the generalization step in order to keep more knowledge of the original data. Find a way to suppress as less records as possible in order to get more information after the subsequently generalization.

Project

Discuss and criticize different approaches based on k-anonymity and its extensions

Readings

- k-Anonymity
  V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, Advances in Information security 2007
- Microdata Protection
  V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, Advances in Informatin security 2007