CS 528 (Fall 2021)
Data Privacy & Security

Yuan Hong
Department of Computer Science
Illinois Institute of Technology

Chapter 2
Data Anonymization (Structured Data)

OUTLINE

Anonymization for Centralized Data

1. **k-Anonymity**
2. **l-Diversity**
3. **t-Closeness**
4. **Other Anonymity Models**

PRIVACY MODELS

Start our study of privacy models

- Ways to quantify privacy
- Methods to achieve this
- Measure loss of utility (if data is obfuscated)

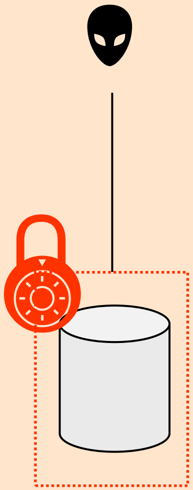
Primary data model

- Centralized data release (i.e., one party holds the entire dataset, anonymizes, and then releases/shares it)

DEFINING PRIVACY IN DATA PUBLISHING

Privacy in this lecture, IS NOT traditional security of data

e.g. hacking,
access control,
theft of disk etc.



NO FOUL PLAY

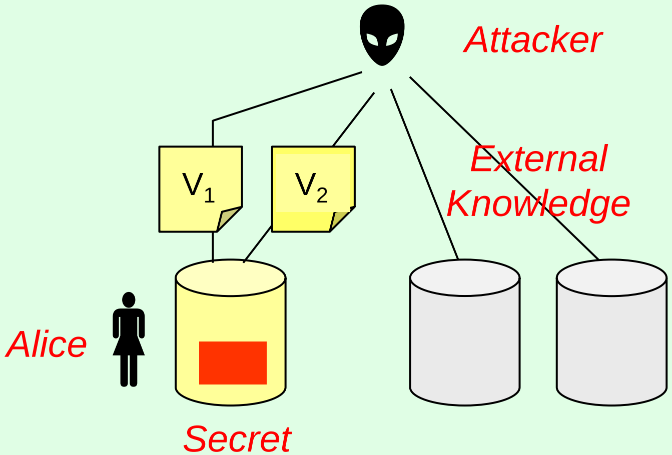
DEFINING PRIVACY IN DB PUBLISHING

Privacy in this lecture IS logical security of data

If the attacker uses *legitimate* methods,

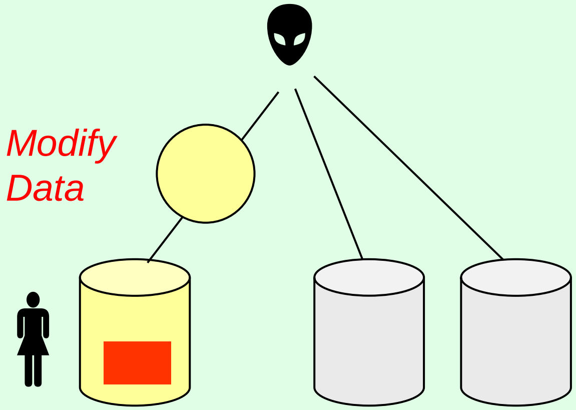
- can he/she infer the data I want to keep private?

Decision Problem



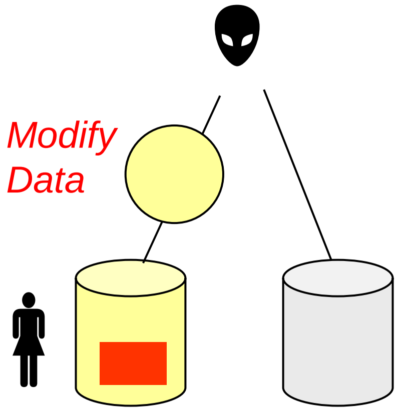
- how can I keep some data private while publishing useful info?

Optimization Problem



NEED FOR PRIVACY IN DB PUBLISHING

ILLINOIS INSTITUTE OF TECHNOLOGY



- Alice is a owner of person-specific data
  - Public health agency, Telecom provider, Financial Organization
- The person-specific data contains
  - Attribute values which can **uniquely identify** an individual

    { zip-code, gender, date-of-birth } or/and {name} or/and {SSN}
  - **Sensitive information** corresponding to individuals

    medical condition, salary, location
- Great demand for sharing of person-specific data
  - Medical research, new telecom applications
- Alice wants to publish this person-specific data s.t.
  - Information remains practically useful
  - Identity of the individual cannot be determined

PRIVACY-PRESERVING DATA PUBLISHING

ILLINOIS INSTITUTE OF TECHNOLOGY

Two opposing goals

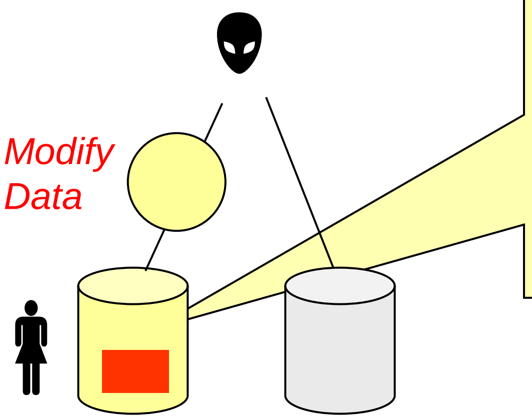
- To allow researchers to extract knowledge about the data
- To protect the privacy of every individual

Microdata/Tabular Data

- Identifier (ID), Quasi-Identifier (QID), Sensitive Attribute (SA)

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

MOTIVATING EXAMPLE



Secret: Alice wants to publish hospital data, while the correspondence between name & disease stays private

	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

The Optimization Problem

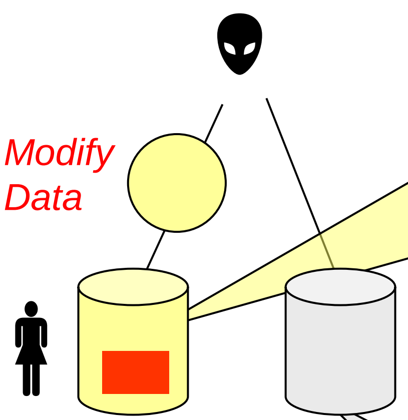
- Anonymization
  - Remove identifiers!

MOTIVATING EXAMPLE (CONTINUED)

ILLINOIS INSTITUTE OF TECHNOLOGY

The Optimization Problem

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

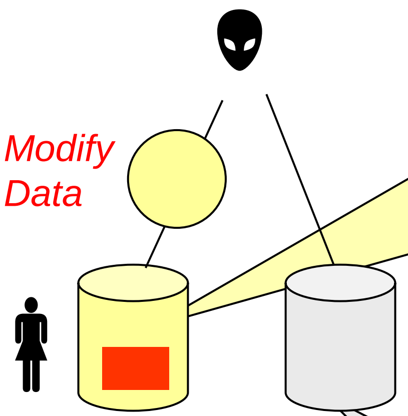
#	***Name***	***Zip***	***Age***	***Nationality***
1	John	13067	45	US
2	Paul	13067	22	US
3	Bob	13067	29	US
4	Chris	13067	23	US

MOTIVATING EXAMPLE (CONTINUED)

The Optimization Problem

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 !

SOURCE OF THE PROBLEM: DATA LINKAGE

Even if we do not publish the identities:

- There are some fields that may *uniquely* identify some individual

	Non-Sensitive Data			*Sensitive Data*
#	***Zip***	***Age***	***Nationality***	***Condition***
…	…	…	…	…

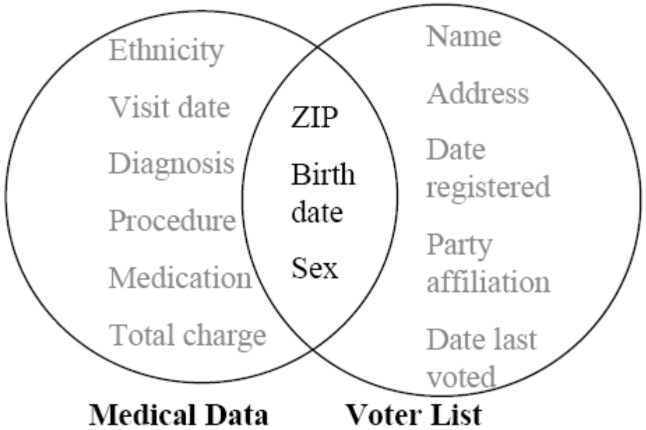
$\underbrace{\text{Zip, Age, Nationality}}_{\text{Quasi Identifier}}$

Quasi Identifier

- The attacker can use them to *join* with other sources and identify the individuals

REAL THREATS OF LINKING ATTACKS

- Fact: 87% of the US citizens can be uniquely linked using only three attributes <Zipcode, DOB, Sex>
- Sweeney [Sweeney, 2002] managed to re-identify the medical record of the government of Massachusetts.



- Census data (income), medical data, transaction data, tax data, etc.

QUASI-IDENTIFIERS

ILLINOIS INSTITUTE OF TECHNOLOGY

Wikipedia

- Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier.

- Quasi-identifiers can thus, when combined, become personally identifying information (PII). This process is called re-identification. As an example, Latanya Sweeney has shown that even though neither gender, birth dates nor postal codes uniquely identify an individual, the combination of all three is sufficient to identify 87% of individuals in the United States.

QUASI-IDENTIFIERS

ILLINOIS INSTITUTE OF TECHNOLOGY

- The term was introduced by **Tore Dalenius** in 1986. Since then, QIs have been the basis of several attacks on released data.
- For instance, Sweeney linked health records to publicly available information to locate the then-governor of Massachusetts' hospital records using uniquely identifying quasi-identifiers.
- Sweeney, Abu and Winn used public voter records to re-identify participants in the Personal Genome Project.

FIRST-CUT SOLUTION: K-ANONYMITY

P. Samarati, L. Sweeney: Generalizing data to provide anonymity when disclosing information
P. Samarati: Protecting Respondents' Identities in Microdata Release
L. Sweeney: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression

Instead of returning the original data:

- *Change the data* such that for each tuple in the results there are at least k-1 other tuples with the same value for the quasi-identifier, e.g.,

#	Zip	Age	Nationality	Condition
1	13053	28	Brazilian	He
2	13067	29	US	He
3	13053	37	Indian	Ca
4	13067	36	Japanese	Ca

Original Table

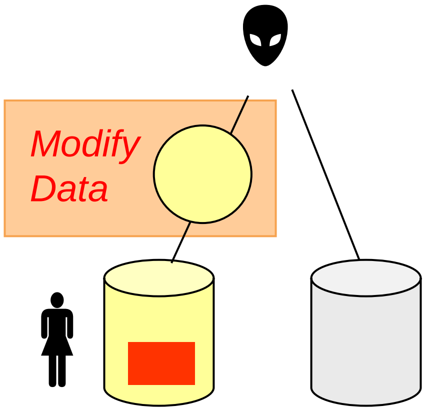
#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Cancer
4	130**	< 40	*	Cancer

4-anonymous Table

K-ANONYMITY

- Each record is indistinguishable from at least k-1 other records
- These k records form an equivalence class
- k-Anonymity ensures that linking cannot be performed with confidence > 1/k.

GENERALIZATION AND SUPPRESSION



Different ways of modifying data:

- Randomization
- Data-Swapping

…

- Generalization
  Replace the value with a less specific but semantically consistent value

- Suppression
  Do not release a value at all

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13067	< 40	*	Heart Disease
3	13053	< 40	*	Cancer
4	13067	< 40	*	Cancer

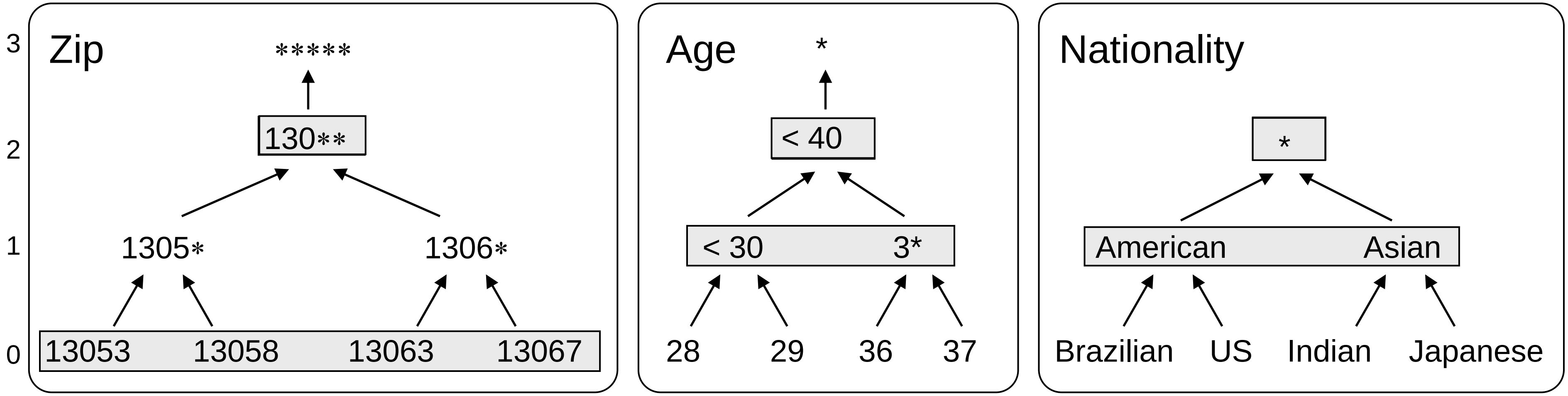
GENERALIZATION AND SUPPRESSION

- Advantages
  - Reveals what was done to the data
  - Truthful (no incorrect implications)
  - Trade-off between anonymity and distortion
  - Adjustable to the recipient's needs (only one's)

- Disadvantages
  - May be possible to distort less data by modifying information in incorrect ways
  - May be difficult to maintain basic statistics

GENERALIZATION HIERARCHIES

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

e.g.,

k= 2 or 4

#	*Zip*	*Age*	*Nationality*	*Condition*
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Heart Disease
3	130**	3*	Asian	Cancer
4	130**	3*	Asian	Cancer

K-MINIMAL GENERALIZATIONS

ILLINOIS INSTITUTE OF TECHNOLOGY

- There are *many* k-anonymizations. Which to pick?

  The ones that do not generalize the data more than needed

k-minimal Generalization: A k-anonymization that is not a generalization of another k-anonymization

e.g.,

✓ *2-minimal Generalization*

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13067	< 40	*	Heart Disease
3	13053	< 40	*	Cancer
4	13067	< 40	*	Cancer

✓ *2-minimal Generalization*

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Heart Disease
3	130**	3*	Asian	Cancer
4	130**	3*	Asian	Cancer

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Cancer
4	130**	< 40	*	Cancer

✗ *Non-minimal 2-anonymization*

K-MINIMAL DISTORTIONS

The Optimization Problem for k-Anonymity

ILLINOIS INSTITUTE OF TECHNOLOGY

- There are *many* k-minimal generalizations. Which to pick?
  The ones that create the minimum distortion to the data

k-minimal Distortion: A k-minimal generalization that has the least distortion

$$\text{Distortion } D = \frac{\displaystyle\sum_{\text{attrib } i} \frac{\text{Current level of generalization for attribute i}}{\text{Max level of generalization for attribute i}}}{\text{Number of attributes}}$$

e.g.,

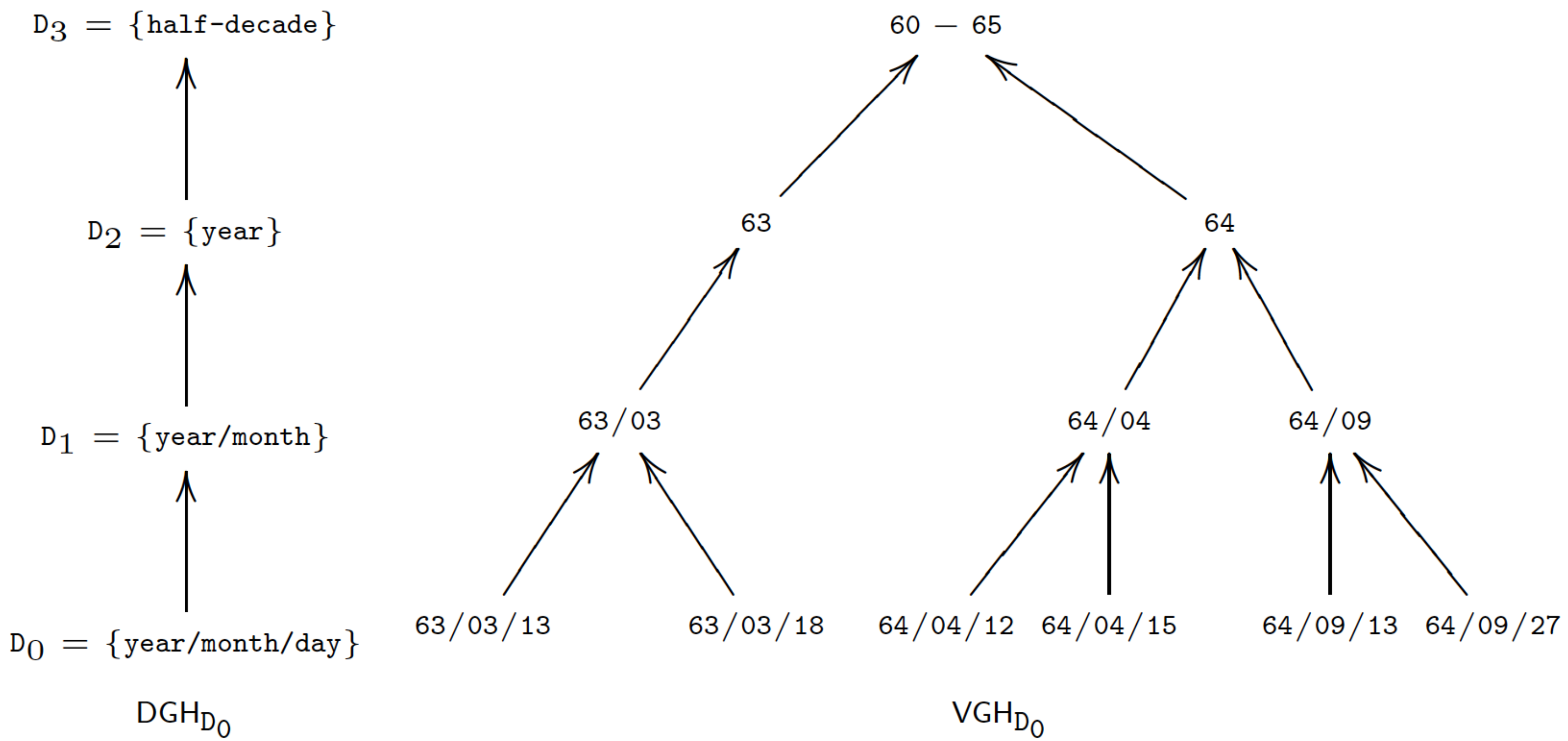
#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13067	< 40	*	Heart Disease
3	13053	< 40	*	Cancer
4	13067	< 40	*	Cancer

$$D = \left(\frac{0}{3} + \frac{2}{3} + \frac{2}{2}\right) / 3 = 0.56$$

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Heart Disease
3	130**	3*	Asian	Cancer
4	130**	3*	Asian	Cancer

$$D = \left(\frac{2}{3} + \frac{1}{3} + \frac{1}{2}\right) / 3 = 0.5$$

DOMAIN GENERALIZATION HIERARCHY (DGH) AND VALUE GENERALIZATION HIERARCHY (VGH)



(e) Date of birth

PRECISION

Precision: average height of generalized values, normalized by Value Generalization Hierarchy (VGH) depth per attribute per record

- N_A : number of attributes
- |PT| : data set size
- h: height of generalized value
- $|DGH_{Ai}|$: depth of the VGH for attribute A_i

$$Prec(RT) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N} \frac{h}{|DGH_{Ai}|}}{|PT| \bullet |N_A|}$$

PRECISION (CONTD)

- Notice that precision depends on DGH/VGH
- Different DGHs result in different precision measurements for the same table
- Structure of DGHs might determine the generalization of choice
- DGHs should be semantically meaningful
  - i.e., created by domain experts

MINGEN ALGORITHM

Steps:

- Generate all generalizations of the private table
- Discard those that violate k-anonymity
- Find all generalizations with the highest precision
- Return one based on some preference criteria

Unrealistic

- Even with attribute level generalization/suppression, there are too many candidates

COMPLEXITY & ALGORITHMS

ILLINOIS INSTITUTE OF TECHNOLOGY

Search Space:

- Number of generalizations = $\prod_{\text{attrib } i}$ (Max level of generalization for attribute i + 1)

If we allow generalization to a different level for each value of an attribute:

- Number of generalizations = $\prod_{\text{attrib } i}$ [(Max level of generalization for attribute i + 1)^ #tuples]

Problem is NP-hard!

1. Naïve Brute Force algorithm
2. Heuristics: Datafly, μ - Argus

DATAFLY ALGORITHM

Steps:

- Heuristically select an attribute to generalize (select the greatest number of distinct values)
- Continue until < k records remain (suppression)

Too much distortion due to attribute level generalization and greedy choices

k-anonymity is guaranteed

μ-ARGUS ALGORITHM

Steps:

- Generalize until each QI attribute appears k times
- Check k-anonymity over 2/3-combinations
- Keeps generalizing according to data holder’s choices
- Suppress any remaining *outliers*

k-anonymity is not guaranteed

Faster than DataFly

K-ANONYMITY SUMMARY

K-Anonymity: attributes are suppressed or generalized until each row is identical with at least k-1 other rows.

K-Anonymity thus can prevent definite external table linkages. At worst, the data released narrows down an individual entry to a group of k individuals.

K-Anonymity guarantees that the released data is accurate.

OPEN ISSUES

How to identify a proper quasi-identifier is a hard problem.

- It depends on what the external table looks like.
- It is hard to predict what external tables will be used to infer the sensitive information.

How to find a k-anonymity solution with suppressing fewest cells?

- We can suppress every cell, but this makes the data useless.
- A minimum cost k-anonymity solution suppresses the fewest number of cells necessary to guarantee k-anonymity.

K-ANONYMITY VULNERABILITIES

Even when sufficient care is taken to identify the QI, K-Anonymity is still be vulnerable to attacks.

Attacks

- Unsorted Matching Attack
- Complementary Release Attack
- Temporal Attack

Fortunately, these attacks can be prevented by following some best practices.

UNSORTED MATCHING ATTACK

ILLINOIS INSTITUTE OF TECHNOLOGY

This attack is based on the order in which tuples appear in the released table.

Solution:

- Randomly sort the tuples 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 can be linked together to compromise k-anonymity.

Solution:

- Consider all of the released tables before releasing 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.

COMPLEMENTARY RELEASE ATTACK (CONT'D)

ILLINOIS INSTITUTE OF TECHNOLOGY

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

- Both of them are 2-anonymized and QI is {Race, Birth, Gender, ZIP}.
- But linking them on {Problem} will generate LT. See next slide.

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

In LT, {White, 1964, male, 02138} and {White, 1965, female, 02139} are unique.

So LT doesn't satisfy 2-anonymity.

TEMPORAL ATTACK

Adding or removing tuples may compromise k-anonymity protection.

Solution: subsequent releases must use the **already released table**.

MORE SERIOUS ATTACKS ON K-ANONYMITY

k-Anonymity alone does not provide privacy if:

- Attacker has background knowledge
- Sensitive attributes lack diversity

K-ANONYMITY ATTACK EXAMPLE

Original Data

	Quasi-Identifier			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

The attacker knows:

- About quasi-identifiers:

Umeko		
Zip	***Age***	***National***
13068	21	Japanese

Bob		
Zip	***Age***	***National***
13053	31	American

- Other background knowledge:

  Japanese have low incidence of heart disease

K-ANONYMITY ATTACK EXAMPLE

4-anonymization

	Quasi-Identifiers			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	**Viral Infection**
4	130**	< 30	*	**Viral Infection**
5	1485*	> = 40	*	Cancer
6	1485*	> = 40	*	Heart Disease
7	1485*	> = 40	*	Viral Infection
8	1485*	> = 40	*	Viral Infection
9	130**	3*	*	**Cancer**
10	130**	3*	*	**Cancer**
11	130**	3*	*	**Cancer**
12	130**	3*	*	**Cancer**

Umeko		
Zip	***Age***	***National***
13068	21	Japanese

Umeko has Viral Infection!

Data Leak !

Bob		
Zip	***Age***	***National***
13053	31	American

Bob has Cancer!

ADVANCED MODEL: L-DIVERSITY

Principle

- Each equivalence class has at least *l* <u>well-represented sensitive values</u>

Distinct *l*-diversity

- Each equivalence class has at least *l* distinct sensitive values
- Probabilistic inference

...	**Disease**
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

10 records

8 records have HIV

2 records have other values

HOMOGENEITY ATTACKS ON K-ANONYMITY

Observation 1. k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute.

k-Anonymity focuses on generalizing the quasi-identifiers but does not address the sensitive attributes which can reveal information to an attacker.

HOMOGENEITY ATTACKS

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Since Alice is Bob's neighbor, she knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob's record number is 9,10,11, or 12. She can also see from the data that Bob has cancer.

BACKGROUND KNOWLEDGE ATTACKS

Observation 2. k-Anonymity does not protect against attacks based on background knowledge.

Depending on other information available to an attacker, an attacker may have increased probability of being able to determine sensitive information.

BACKGROUND KNOWLEDGE ATTACKS

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Alice knows that Umeko is a 21 year-old Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko's information is contained in record number 1, 2, 3, or 4. With additional information, Umeko being Japanese and Alice knowing that Japanese have an extremely low incidence of heart disease, Alice can conclude with near certainty that Umeko has a viral infection.

WEAKNESSES IN K-ANONYMOUS TABLES

Given these two weaknesses, there needs to be a stronger method to ensure privacy.

Based on this, the l-diversity model is proposed

MODEL AND NOTATION

Basic Notation:

- Let $T = \{t_1,t_2,...,t_n\}$ be a table with attributes $A_1,...,A_m$. We assume that T is a subset of some larger population Ω where each tuple represents an individual from the population. For example, if T is a medical dataset then Ω could be the population of the United States.
- Let A denote the set of all attributes $\{A_1,A_2,...,A_m\}$ and $t[A_i]$ denote the value of attribute A_i for tuple t. If $C = \{C_1,C_2,...,C_p\} \subseteq A$, then we use the notation t[C] to denote the tuple $(t[C_1],...,t[C_p])$, which is the projection of t onto the attributes in C.
- All actual identifiers such as name, SSN, address, etc., are removed from the data leaving sensitive attributes and non-sensitive attributes (quasi-identifiers).

L-DIVERSITY PRINCIPLE

Given the previous discussions we arrive at the l-Diversity principle for k-anonymous tables:

- q*-block: equivalence class
- A q*-block is l-diverse if contains at least l *"well-represented"* values for the sensitive attribute S.
- A table is l-diverse if every q*-block is l-diverse.

REVISITING THE EXAMPLE

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	$*$	Heart Disease
4	1305*	≤ 40	$*$	Viral Infection
9	1305*	≤ 40	$*$	Cancer
10	1305*	≤ 40	$*$	Cancer
5	1485*	> 40	$*$	Cancer
6	1485*	> 40	$*$	Heart Disease
7	1485*	> 40	$*$	Viral Infection
8	1485*	> 40	$*$	Viral Infection
2	1306*	≤ 40	$*$	Heart Disease
3	1306*	≤ 40	$*$	Viral Infection
11	1306*	≤ 40	$*$	Cancer
12	1306*	≤ 40	$*$	Cancer

Using a 3-diverse table, we no longer are able to tell if Bob has heart disease or cancer. We also cannot tell if Umeko has a viral infection or cancer.

L-DIVERSITY PRINCIPLE

The l-Diversity principle advocates ensuring well represented values for sensitive attributes but does not define *what well represented values mean.*

L-DIVERSITY INSTANTIATIONS

- Entropy *l-Diversity*
- Recursive *(c, l)-Diversity*
- Positive Disclosure-Recursive *(c, l)-Diversity*
- Negative/Positive Disclosure-Recursive *(c1, c2, l)-Diversity*

ENTROPY L-DIVERSITY

ILLINOIS INSTITUTE OF TECHNOLOGY

Definition 4.1 (Entropy ℓ-Diversity) *A table is* Entropy ℓ-Diverse *if for every $q^\star$-block*

$$-\sum_{s \in S} p_{(q^\star,s)} \log(p_{(q^\star,s')}) \geq \log(\ell)$$

where $p_{(q^\star,s)} = \frac{n_{(q^\star,s)}}{\sum\limits_{s' \in S} n_{(q^\star,s')}}$ is the fraction of tuples in the $q^\star$-block with sensitive attribute value equal to s.

This implies that for a table to be entropy l-Diverse, the entropy of the entire table must be *at least log(l)*. Therefore, entropy l-Diversity may be too restrictive to be practical.

RECURSIVE (C, L)-DIVERSITY

Less restrictive than entropy l-diversity

Let $s_1, \ldots, s_m$ be the possible values of sensitive attribute S in a q*-block

Assume we sort the counts $n(q^*,s_1), \ldots, n(q^*,s_m)$ in descending order with the resulting sequence $r_1, \ldots, r_m$.

We can say a q*-block is recursive (c, 2)-diverse if $\mathbf{r_1 < c(r_2 + \ldots + r_m)}$ for a specified constant c.

RECURSIVE (C, L)-DIVERSITY (CONT.)

Definition 4.2 (Recursive (c, ℓ)-Diversity) *In a given $q^\star$-block, let r_i denote the number of times the i^{th} most frequent sensitive value appears in that $q^\star$-block. Given a constant c, the $q^\star$-block satisfies* recursive (c, ℓ)-diversity *if $r_1 < c(r_\ell + r_{\ell+1} + \cdots + r_m)$. A table $T^\star$ satisfies recursive (c, ℓ)-diversity if every $q^\star$-block satisfies recursive ℓ-diversity. We say that 1-diversity is always satisfied.*

POSITIVE DISCLOSURE-RECURSIVE (C, L)-DIVERSITY

In practice, some cases of positive disclosure may be acceptable such as when medical condition is “healthy”.

PD-Recursive (c, l)-diversity

PD-RECURSIVE (C, L)-DIVERSITY

ILLINOIS INSTITUTE OF TECHNOLOGY

Definition 4.3 (Positive Disclosure-Recursive (c, ℓ)-Diversity). *Let Y denote the set of sensitive values for which positive disclosure is allowed. In a given $q^\star$-block, let the most frequent sensitive value not in Y be the y^{th} most frequent sensitive value. Let r_i denote the frequency of the i^{th} most frequent sensitive value in the $q^\star$-block. Such a $q^\star$-block satisfies* pd-recursive (c, ℓ)-diversity *if one of the following hold:*

- $y \leq \ell - 1$ *and* $r_y < c \sum_{j=\ell}^{m} r_j$
- $y > \ell - 1$ *and* $r_y < c \sum_{j=\ell-1}^{y-1} r_j + c \sum_{j=y+1}^{m} r_j$

Allows for most sensitive values not in Y

NEGATIVE/POSITIVE DISCLOSURE-RECURSIVE (C1, C2, L)-DIVERSITY

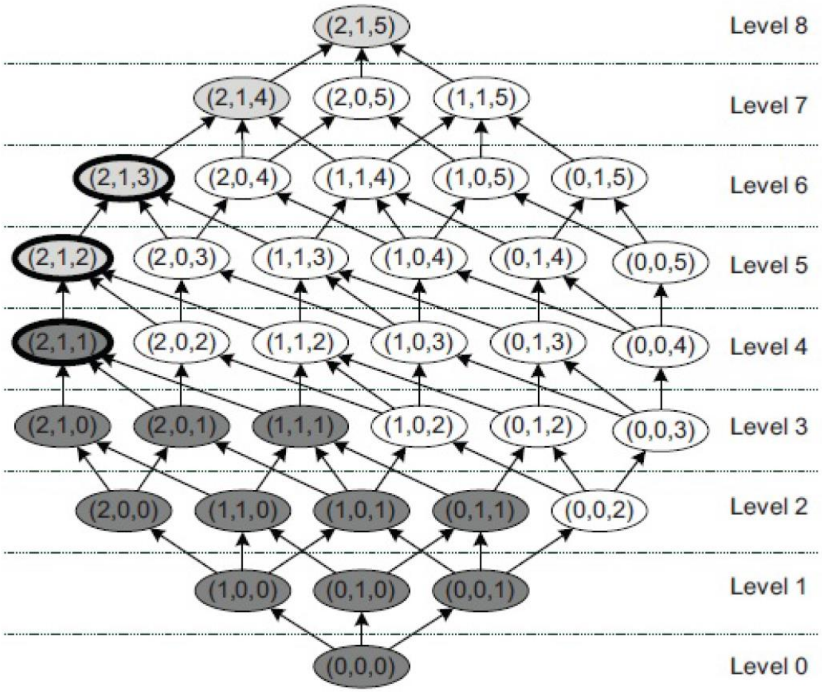
NPD-recursive (c_1, c_2, l)-diversity prevents negative disclosure

Definition 4.4 (Negative/Positive Disclosure-Recursive (c_1, c_2, ℓ)-Diversity). *Let W be the set of sensitive values for which negative disclosure is not allowed. A table satisfies* npd-recursive (c_1, c_2, ℓ)-diversity *if it satisfies pd-recursive (c_1, ℓ)-diversity and if every $s \in W$ occurs in at least c_2 percent of the tuples in every $q^\star$-block.*

IMPLEMENTING PRIVACY PRESERVING DATA PUBLISHING

Domain generalization is used to define a generalization lattice.

For discussion, all non-sensitive attributes are combined into a multi-dimensional attribute (Q) where the bottom element on the lattice is the domain of Q and the top of the lattice is the domain where each dimension of Q is generalized to a single value.



CONT

The algorithm for publishing should find the point on the lattice where the table T* **preserves privacy** and is **useful** as possible.

The usefulness (utility) of table T* is diminished as the data becomes more generalized, so the most utility is at the bottom of the lattice.

CONT

Monotonicity property is described as a stopping point in the lattice search where the privacy is protected and further generalization does not increase privacy.

An example is if zip 13065 can be generalized to 1306* and it preserves privacy, generalizing it to 130** also preserves privacy. However, the additional generalization reduces utility.

MONOTONICITY PROPERTY

k-anonymity satisfies the monotonicity property which guarantees the correctness of all efficient lattice search algorithms, so if l-diversity satisfies the monotonicity property, these algorithms can be used by l-diversity.

MONOTONICITY PROPERTY

Theorem 5.2 (Monotonicity of Entropy ℓ-diversity) *Entropy ℓ-diversity satisfies the monotonicity property: if a table $T^\star$ satisfies entropy ℓ-diversity, then any generalization $T^{\star\star}$ of $T^\star$ also satisfies entropy ℓ-diversity.*

Theorem 5.3 (Monotonicity of NPD Recursive ℓ-diversity) *npd recursive (c_1, c_2, ℓ)-diversity satisfies the monotonicity property: if a table $T^\star$ satisfies npd recursive (c_1, c_2, ℓ)-diversity, then any generalization $T^{\star\star}$ of $T^\star$ also satisfies npd recursive (c_1, c_2, ℓ)-diversity.*

All variants of l-diversity can be proven to satisfy monotonicity.

MONOTONICITY PROPERTY

Therefore, to create an algorithm for l-diversity, a k-anonymity routine can be used by substituting l-diversity processing instead of k-anonymity.

Since l-diversity is local to each q*-block and the l-diversity test as based on the counts of sensitive attributes the testing is quite efficient.

LIMITATIONS OF L-DIVERSITY

ILLINOIS INSTITUTE OF TECHNOLOGY

l-diversity may be difficult and unnecessary to achieve.

- A single sensitive attribute
  - Two values: HIV positive (1%) and HIV negative (99%)
  - Very different degrees of sensitivity
- l-diversity is **unnecessary** to achieve
  - 2-diversity is unnecessary for an equivalence class that contains only negative 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

LIMITATIONS OF L-DIVERSITY

l-diversity is insufficient to prevent attribute disclosure.

Skewness Attack

- Two sensitive values
  - HIV positive (1%) and HIV negative (99%)
- Serious privacy risk
  - Consider an equivalence class that contains an equal number of positive records and negative records
- l-diversity does not differentiate: (both satisfy)
  - Equivalence class 1: 49 positive + 1 negative
  - Equivalence class 2: 1 positive + 49 negative

l-Diversity does not consider the overall distribution of sensitive values (in the original dataset)

LIMITATIONS OF L-DIVERSITY

ILLINOIS INSTITUTE OF TECHNOLOGY

l-diversity is insufficient to prevent 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 semantic meanings of sensitive values

T-CLOSENESS: A NEW PRIVACY MEASURE

ILLINOIS INSTITUTE OF TECHNOLOGY

Adversarial belief

Belief	Knowledge
B_0	External Knowledge
B_1	Overall distribution Q of sensitive values

A completely generalized table

Age	Zipcode	……	Gender	Disease
*	*	……	*	Flu
*	*	……	*	Heart Disease
*	*	……	*	Cancer
.	.	……	.	.
.	.	……	.	.
.	.	……	.	.
*	*	……	*	Gastritis

T-CLOSENESS: A NEW PRIVACY MEASURE

Adversarial belief

A released table

Age	Zipcode	......	Gender	Disease
2*	479**	......	Male	Flu
2*	479**	......	Male	Heart Disease
2*	479**	......	Male	Cancer
.	.	......	.	.
.	.	......	.	.
.	.	......	.	.
≥50	4766*	......	*	Gastritis

Belief	Knowledge
B_0	External Knowledge
B_1	Overall distribution Q of sensitive values
B_2	Distribution P_i of sensitive values in each equi-class

T-CLOSENESS: A NEW PRIVACY MEASURE

Adversarial belief

Belief	Knowledge
B_0	External Knowledge
B_1	Overall distribution Q of sensitive values
B_2	Distribution P_i of sensitive values in each equi-class

- Rationale
  - Q should be public information
  - Knowledge gain is separated:
    - About whole population (from B_0 to B_1)
    - About individuals (from B_1 to B_2)
  - We bound knowledge gain between B_1 and B_2
- Principle
  - The distance between Q and P_i is bounded by a threshold t
  - l-diversity considers only P_i

DISTANCE MEASURES

Measure distance between

- **P**=(p_1,p_2,…,p_m), **Q**=(q_1,q_2,…,q_m)

Distance measures

- Trace-distance (differences between two matrices): $\mathrm{D}[\mathbf{P},\mathbf{Q}] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|p_i - q_i|.$
- KL-divergence (differences between two distributions): $\mathrm{D}[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} p_i \log\frac{p_i}{q_i} = H(\mathbf{P}) - H(\mathbf{P},\mathbf{Q})$

Semantic meanings

- ***Q***: {20K,30K,40K,50K,60K,70K,80K,90K,100K}

  P_1:*{20K,30K,40K}*
  P_2:*{20K,60K,100K}*

- Intuitively, $D[\mathbf{P_1},\mathbf{Q}]>D[\mathbf{P_2},\mathbf{Q}]$

SUMMARY

t-closeness protects against attribute disclosure but not identity disclosure

t-closeness requires that the distribution of a sensitive attribute in **<u>any equivalence class</u>** is close to the distribution of a sensitive attribute in the **<u>overall table</u>**.

TYPES OF INFORMATION DISCLOSURE

ILLINOIS INSTITUTE OF TECHNOLOGY

Identity Disclosure

- An individual is linked to a particular record in the published data.
- k-Anonymity [Sweeney, 2002].

Attribute Disclosure

- Sensitive attribute information of an individual is disclosed.
- l-Diversity [Machanavajjhala et al., 2006].
- t-Closeness [Li et al., 2007].

Membership Disclosure

- Information about whether an individual’s record is in the published data or not.
- δ-presence [Nergiz et al., 2007].

ILLINOIS INSTITUTE OF TECHNOLOGY

SUMMARY

Looked at several different models for privacy

- k-anonymity
- l-diversity
- t-closeness

Factoid: Optimal K-anonymization is NP-hard

- Must find efficient techniques to do this well

Other extensions

- Personalization (i.e., each user sets its own k, .etc)
- Multi-relational k-anonymity

ILLINOIS INSTITUTE OF TECHNOLOGY

CS 528 (Fall 2021) Data Privacy & Security

Yuan Hong

Department of Computer Science

Illinois Institute of Technology

Chapter 2 - Extension

Data Anonymization (Unstructured Data)

OUTLINE

Anonymization

1. **Set-valued Data (e.g., Movie Rating, and Market Basket)**
2. **GPS Locations**
3. **Social Network (Graph)**
4. **Search Queries (Text)**

SET-VALUED DATA

- "Relational data"
  - One sensitive attribute for each tuple

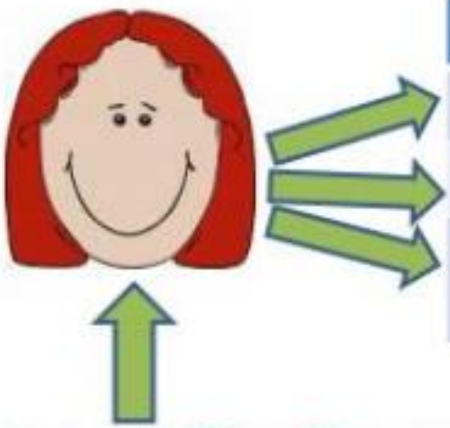
Zipcode	Gender	Age	...	Medical diagnosis
53705	male	30	...	***flu***
98072	female	40	...	***diabetes***

- "Set-valued data"
  - Logically: (*personid*, {$item_1$, $item_2$, …, $item_n$})
  - Multiple sensitive values in one record possible

Person ID	Item set
001	{milk, sunglasses, ***viagra***}
002	{beer, diapers, shampoo}
003	{beer, milk, diapers, ***pregnancy test***, ***diabetes medicine***}

AN ATTACK SCENARIO

- Retailer publishes market basket data
- The adversary knows Alice has bought milk, beer, and diapers
- The adversary infers Alice has also bought **pregnancy test** and **diabetes medicine**



PID	Item set
001	{milk, sunglasses, **viagra**}
002	{beer, diapers, shampoo}
003	{beer, milk, diapers, **pregnancy test, diabetes medicine**}

beer, milk, diapers, pregnancy test, diabetes medicine

ANONYMIZATION (1)

	Wine	Strawberries	Meat	Cream	Pregnancy Test	Viagra
Bob	X		X			X
David	X		X			
Claire		X		X	X	
Andrea		X	X			
Ellen	X		X	X		

Quasi-identifying Items (Wine, Strawberries, Meat, Cream)

Sensitive Items (Pregnancy Test, Viagra)

ANONYMIZATION (1)

	Wine	Meat	Cream	Strawberries	Pregnancy Test	Viagra
Bob	X	X				X
David	X	X				
Ellen	X	X	X			
Andrea		X		X		
Claire			X	X	X	

Band Matrix
Organization

PRESERVES

CORELATIONS!

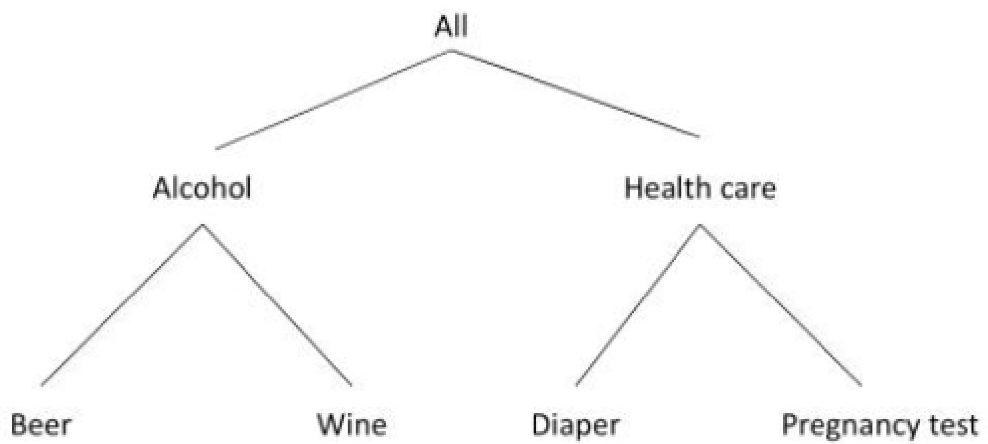
SHARED DATA

	Wine	Meat	Cream	Strawberries	Sensitive Items
Bob	X	X			Viagra: 1
David	X	X			
Ellen	X	X	X		
Andrea		X		X	Pregnancy Test: 1
Claire			X	X	

Summary of Sensitive Items

ANONYMIZATION (2)
K-ANONYMITY

Hierarchical generalization



- Transaction generalization

  T_i: {"Beer", "Wine", "Diaper"} → {"Alcohol", "Health care"}

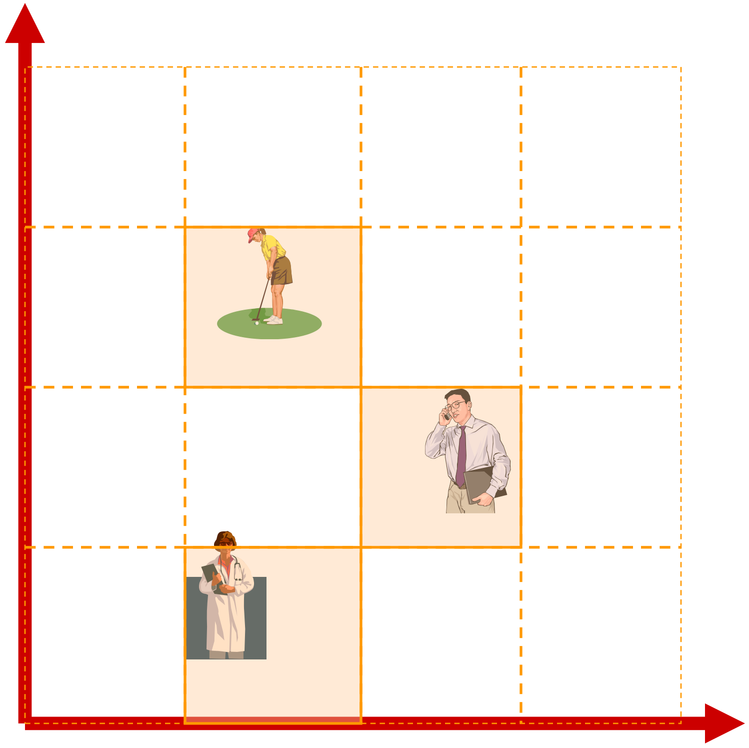
- Duplicates removed

OUTLINE

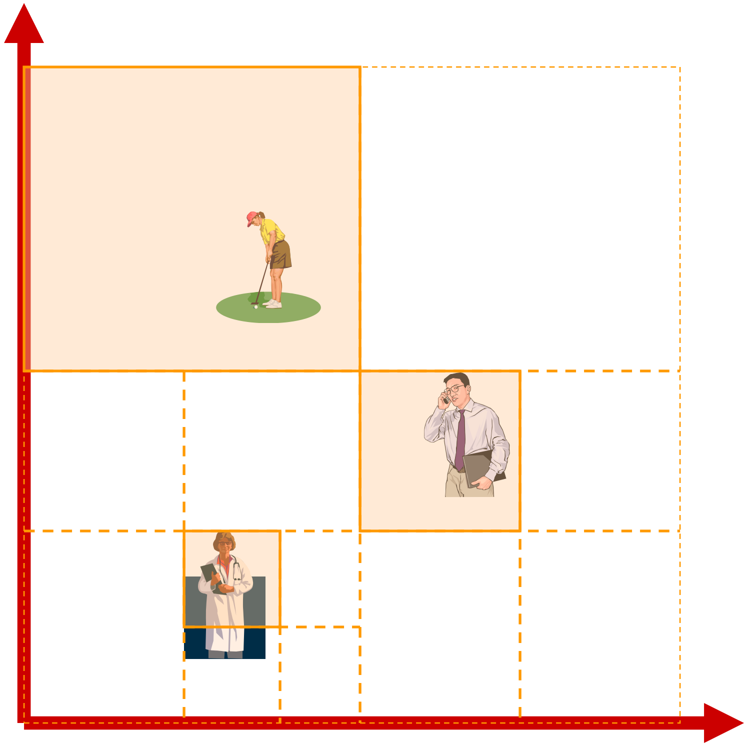
Anonymization

1. **Set-valued Data (e.g., Movie Rating, and Market Basket)**
2. **GPS Locations**
3. **Social Network (Graph)**
4. **Search Queries (Text)**

LOCATION PRIVACY



Fixed grid cloaking



Adaptive grid cloaking

ILLINOIS INSTITUTE OF TECHNOLOGY

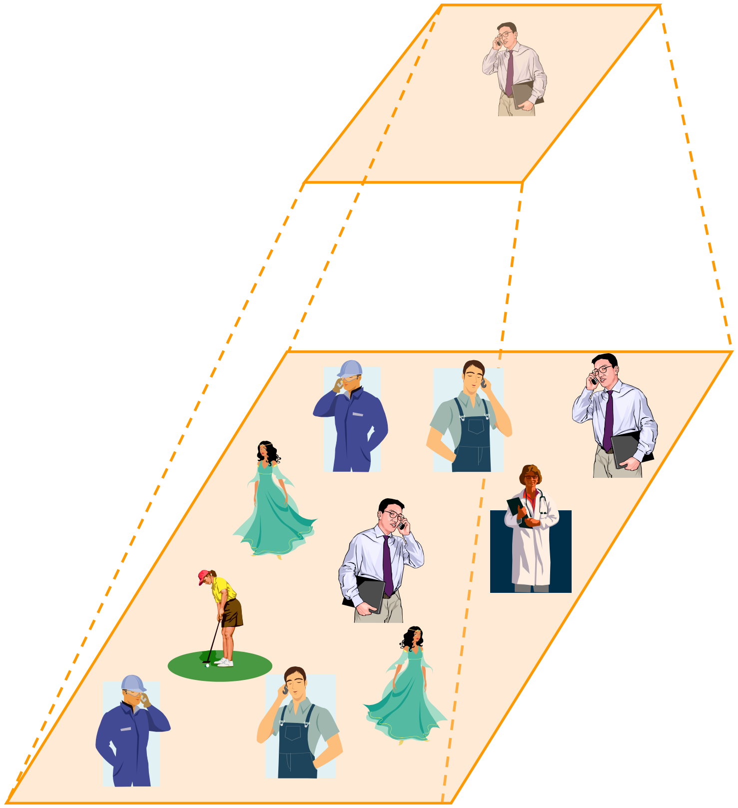
K-ANONYMITY

The *cloaked* region contains at least *k* users

The user is indistinguishable among other *k* users

The cloaked area largely depends on the surrounding environment.

A value of *k* =100 may result in a very small area if a user is located in the stadium or may result in a very large area if the user in the desert.



10-anonymity

OUTLINE

Anonymization

1. **Set-valued Data (e.g., Movie Rating, and Market Basket)**
2. **GPS Locations**
3. **Social Network (Graph)**
4. **Search Queries (Text)**

SOCIAL NETWORK (GRAPH)

http://www.touchgraph.com/

Linkedin

facebook

myspace.com
a place for friends

PRIVACY BREACHES ON GRAPH DATA

ILLINOIS INSTITUTE OF TECHNOLOGY

- Identity disclosure
  - Identity of individuals associated with nodes is disclosed
- Link disclosure
  - Relationships between individuals are disclosed
- Content disclosure
  - Attribute data associated with a node is disclosed

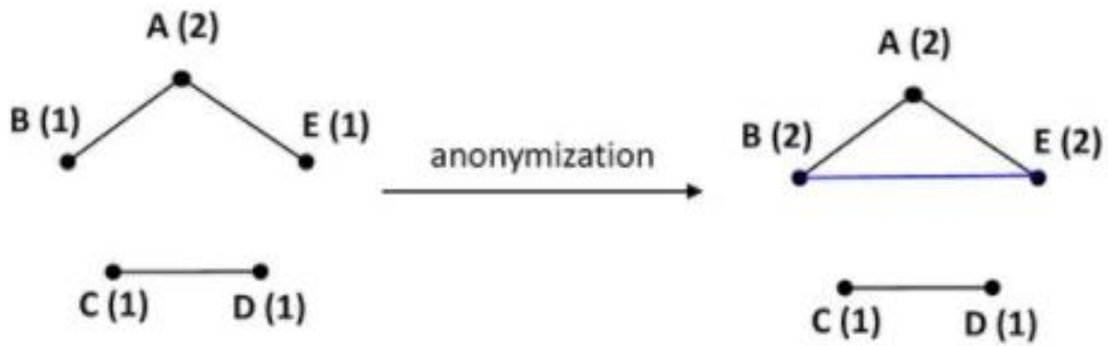
PRIVACY LEAKAGE

What if you want to prevent the following from happening

- Assume that adversary **A** knows that **B** has 327 connections in a social network!
- If the graph is released by removing the identity of the nodes
  - **A** can find all nodes that have degree 327
  - If there is only one node with degree 327, **A** can identify this node as being **B**.

PRIVACY MODEL

[k-degree anonymity] A graph *G(V, E)* is *k-degree anonymous* if every node in *V* has the same degree as *k-1* other nodes in *V*.



[Properties] It prevents the re-identification of individuals by adversaries with *a priori* knowledge of the degree of certain nodes.

K-ANONYMITY

ILLINOIS INSTITUTE OF TECHNOLOGY

Given a graph *G(V, E)* and an integer *k*, modify *G* via a minimal set of edge addition or deletion operations to construct a new graph *G'(V', E')* such that

1) *G'* is *k*-degree anonymous;

2) *V' = V*;

3) The *symmetric difference* of *G* and *G'* is as small as possible

- Symmetric difference between graphs *G(V,E)* and *G'(V,E')* :

$$\mathrm{SymDiff}(G', G) = (E' \setminus E) \cup (E \setminus E')$$

OUTLINE

Anonymization

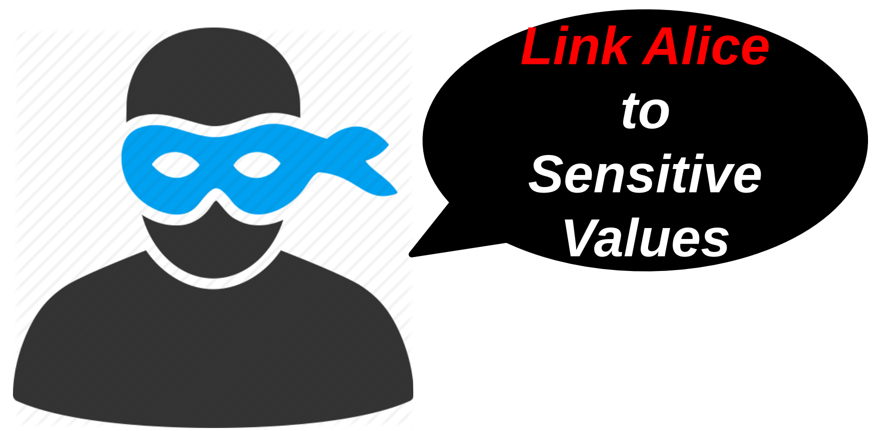
1. **Set-valued Data (e.g., Movie Rating, and Market Basket)**
2. **GPS Locations**
3. **Social Network (Graph)**
4. **Search Queries (Text)**

Dataset was published in 2006 (anonymized by only random pseudo-user-IDs)

Alice

Bob

User-ID	Query	Clicked URLs	Query Time
0001	**181 Park Ave.**	maps.google.com	02-01-2006 11:02 AM
	Pizza	pizzahut.com	02-01-2006 03:17 PM
	Honda	honda.com	02-02-2006 06:25 PM
	Pregnancy test	medicinenet.com	02-05-2006 09:39 PM
	…	…	…
000	Jobs	linkedin.com	03-01-2006 01:08 PM
	…	…	…

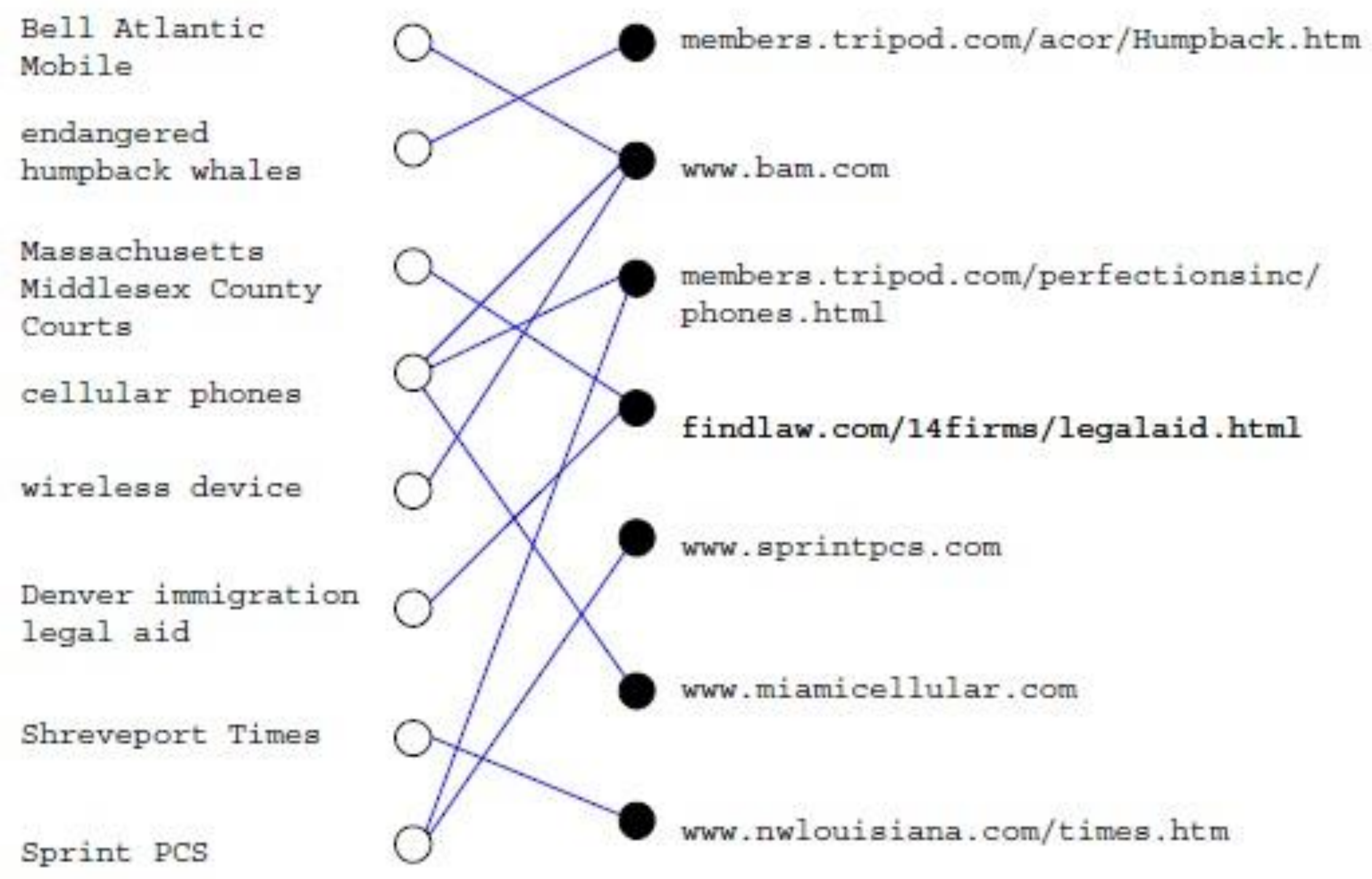


SEMANTICALLY SIMILAR QUERIES

We can use the clicked URL to represent the search (intent), and measure semantic similarity.

- E.g., a user clicked http://www.sun.com/, we know that he wants some information about Sun Microsystems rather than that star
    - Sun → click http://www.sun.com/
    - Solaris System → click http://www.sun.com/
    - Sun Inc. → click http://www.sun.com/
    - Sun Company → click http://www.sun.com/
    - Sun Unix server → click http://www.sun.com/
    - Sun → https://en.wikipedia.org/wiki/Sun
    - Solar System → https://en.wikipedia.org/wiki/Sun

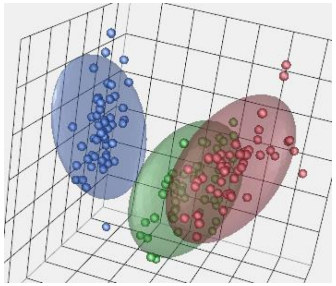
QUERY-URL BIPARTITE GRAPH



ANONYMIZATION

1. Find search engine users with Similar Search Intents

a) Clustering **Semantically Similar Queries** to identify different search intents



b) Clustering **Similar Users** with similar search intents

2. Make users' search data in the output Indistinguishable

- Adding or suppressing (**sampled**) semantically similar queries for users in the same cluster – for minimizing the utility loss

ACKNOWLEDGMENTS

Note: Some of the slides in this lecture are adapted based on materials created by

- Dr. Alessandro Acquisti at CMU
- Dr. Murat Kantarcioglu at UT Dallas
- Dr. Ninghui Li at Purdue
- Dr. Ashwin Machanavajjhala at Duke
- Dr. Latanya Sweeney at CMU
- Dr. Jaideep Vaidya at Rutgers