UMBC
Dr. Roberto Yus
https://robertoyus.com/

Data Privacy

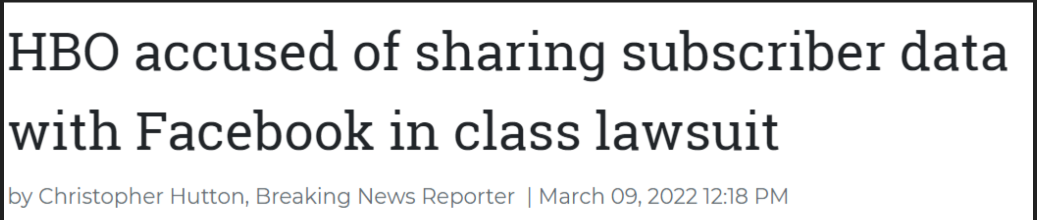
CMSC 491/691

L06 – k-anonymity, ℓ-diversity, t-closeness

Icons from https://thenounproject.com/

Previously on…

- Access Control to represent user preferences
- Policies and mechanisms
- AC models:
  - DAC, MAC, RBAC, ABAC
- Challenges: scalability, inference problem, semantics…



HBO accused of sharing subscriber data with Facebook in class lawsuit

by Christopher Hutton, Breaking News Reporter | March 09, 2022 12:18 PM

In the news!

The Need to Share Data

- For research purposes
  - E.g., social, medical, technological, etc.
- Mandated by laws and regulations
  - E.g., census
- For security/business decision making
  - E.g., network flow data for Internet-scale alert correlation
- For system testing before deployment
- …

- **Publishing data may result in privacy violations**

When Things go Wrong

The Netflix Prize

dannypeled.com

AOL Search Data

- Anonymizing datasets (e.g., removing user identifiers) **does not preserve privacy!**
- De-anonymization attacks
    - E.g., use background knowledge (IMDB for Netflix prize)

How to publish data to satisfy privacy while providing utility?

Classification of Attributes

- **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.
  - Can be used for linking anonymized dataset with other datasets
- **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	**Age**	**Sex**	**Zipcode**	**Disease**
Alice	29	Female	47677	Ovarian Cancer
Beth	22	Female	47602	Ovarian Cancer
Andre	27	Male	47678	Prostate Cancer
Dan	43	Male	47905	Heart Disease
Ellen	52	Female	47909	Heart Disease
Eric	47	Male	47906	Heart Disease

k-Anonymity: Intuition

- Each **record is indistinguishable from at least k-1 other records** when only quasi-identifiers are considered
    - *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.*
- The k records form an **equivalence class**

Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"

Achieving *k*-Anonymity

- Main methods:
  - **Generalization**: Replace with less-specific values
  - **Suppression**: Remove outliers
- Many other methods in the literature…

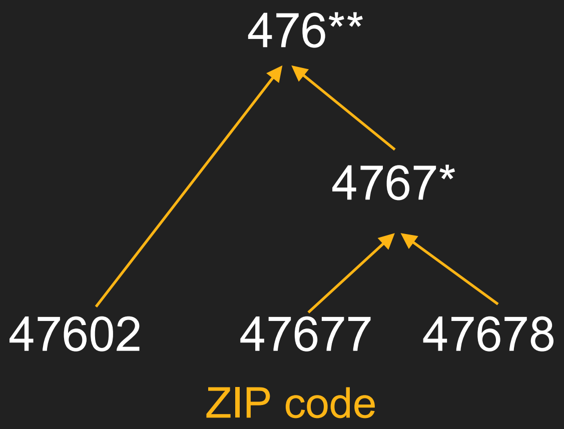
Age	Sex	Zipcode	Disease
2*	*	476**	Ovarian Cancer
2*	*	476**	Ovarian Cancer
2*	*	476**	Prostate Cancer
[43,52]	*	4790*	Heart Disease
[43,52]	*	4790*	Heart Disease
[43,52]	*	4790*	Heart Disease

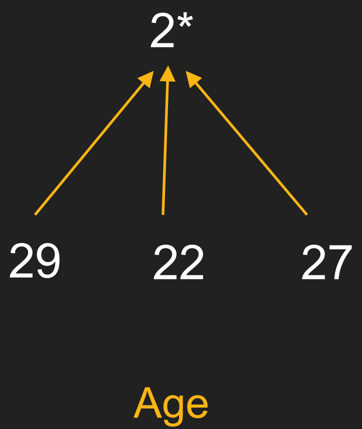
Generalization

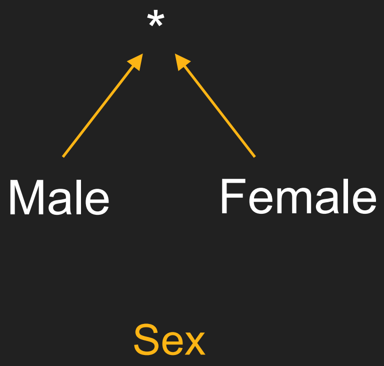
Suppression (cell-level)

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

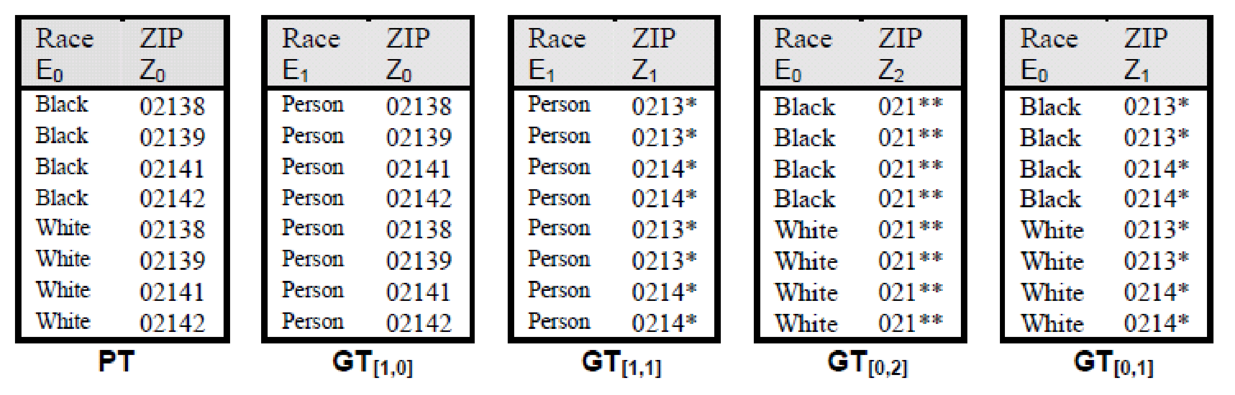






k-Minimal Generalizations

- There are many *k*-anonymizations – which one to pick?
    - Intuition: The one that does not generalize the data more than needed (decrease in utility of the published dataset!)
- **K-minimal generalization**: A *k*-anonymized table that is not a generalization of another *k*-anonymized table



Race	ZIP
E_0	Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
E_1	Z_0
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT$_{[1,0]}$

Race	ZIP
E_1	Z_1
Person	0213*
Person	0213*
Person	0214*
Person	0214*
Person	0213*
Person	0213*
Person	0214*
Person	0214*

GT$_{[1,1]}$

Race	ZIP
E_0	Z_2
Black	021**
Black	021**
Black	021**
Black	021**
White	021**
White	021**
White	021**
White	021**

GT$_{[0,2]}$

Race	ZIP
E_0	Z_1
Black	0213*
Black	0213*
Black	0214*
Black	0214*
White	0213*
White	0213*
White	0214*
White	0214*

GT$_{[0,1]}$

Figure 4 Examples of generalized tables for PT

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

2-minimal Generalizations

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

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

NOT a 2-minimal Generalization

Example *k*-anonymization

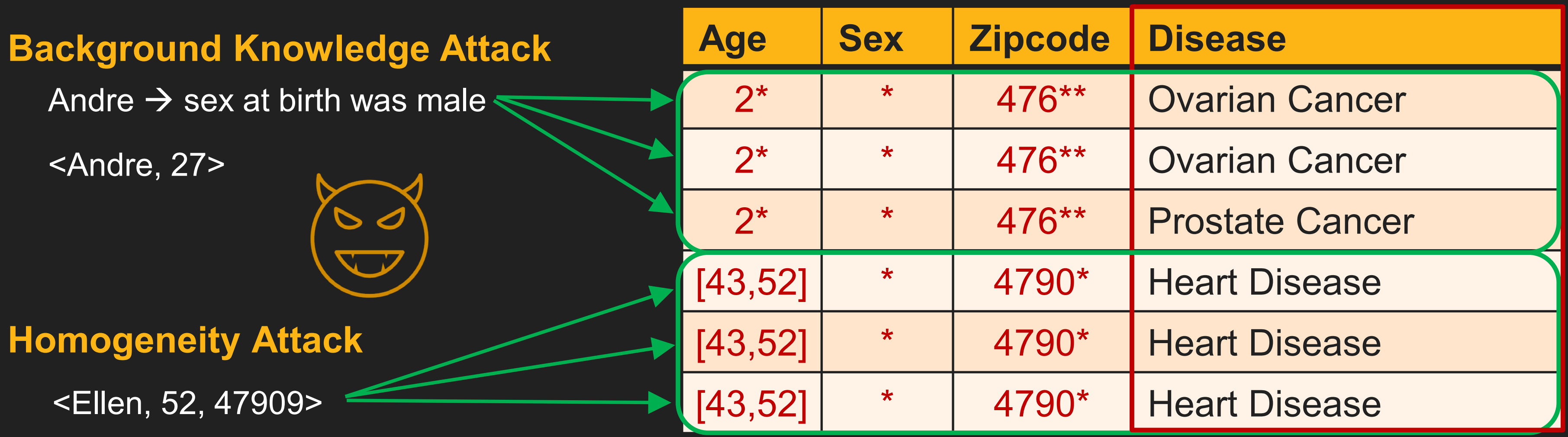
Age	Sex	Zipcode	Disease
2*	*	476**	Ovarian Cancer
2*	*	476**	Ovarian Cancer
2*	*	476**	Prostate Cancer
[43,52]	*	4790*	Heart Disease
[43,52]	*	4790*	Heart Disease
[43,52]	*	4790*	Heart Disease

- **3-Anonymous** table
  - The adversary knows Alice's QI values (47677, 29, F)
  - The adversary does not know which one of the first 3 records corresponds to Alice

Problems?

Attacks on *k*-Anonymity

- *k*-anonymity does not provide privacy if:
  - ○ Sensitive values **lack diversity**
  - ○ The attacker has **background knowledge**



Background Knowledge Attack

Andre → sex at birth was male

<Andre, 27>

Homogeneity Attack

<Ellen, 52, 47909>

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

Other Attacks

- **Complementary Release Attack**
    - Different releases of the same private table can be linked together to compromise k-anonymity
- **Unsorted Matching Attack**
    - Records appear in the same order in the released table as in the original table
- …

Group Activity

- Releasing *k*-anonymous reviews for professors by students

Name	Age	Nationality	Class	Level	Grade	Prof.
Alice	21	U.S. citizen	CMSC331	Junior	B	Smith
Beth	20	U.S. citizen	CMSC334	Junior	F	Miller
Andre	22	U.S. citizen	CMSC331	Senior	A	Smith
Dan	21	U.S. citizen	CMSC491	Senior	C	Anderson
Ellen	20	U.S. citizen	CMSC203	Sophomore	F	Miller
Eric	19	U.S. citizen	CMSC101	Sophomore	A	Williams

Privacy?

Utility?

l -Diversity

- *Recall* → *k*-anonymity, k records form an **equivalence class**
- l-diversity is a stronger definition of privacy
- Principle
  - Each equivalence class contains at least l **well-represented** sensitive values
- Instantiations
  - Distinct l-diversity
    - Each equivalence class contains distinct l sensitive values
  - …

A. Machanavajjhala, et al. "l-diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007): 3-es.

	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

4-anonymous table

	Zip	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

4-anonymous and 3-diverse table

What's Bob's (31yo/American/13053) disease?

What's Umeko's (21yo/Japanese/13068) disease?
****BK: Japanese are less prone to heart disease***

Limitations of l-Diversity

Similarity attack

Bob	
Zip	***Age***
47678	27

Zip	Age	Salary	Condition
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
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	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!

Limitations of l-Diversity

- Skewness Attack
- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)

	Zip	Age	Condition
1	476**	< 30	HIV+
2	476**	< 30	HIV+
3	476**	< 30	HIV-
4	476**	< 30	HIV-

2-diverse table

Bob	
Zip	***Age***
47678	27

- *Before l-diversity:*

probability of Bob being HIV+ = 1%

- *After 2-diverse table*

probability of Bob being HIV+ = 50%!

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

t-Closeness

- Principle:
  - Distribution of sensitive attribute value in each equi-class should be “close” to that of the overall dataset (distance ≤ t)

Can we always do this?

How would it affect utility?

Race	Zip	Condition
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

L. Ninghui, et al. "t-closeness: Privacy beyond k-anonymity and l-diversity.“ IEEE 23rd international conference on data engineering, 2007.

Combining Everything

Race	Zip	HIV status	Condition
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

Bob is Caucasian and I've heard he was admitted to a hospital with flu…

Imagine a table which is:

- k-anonymous,
- l-diverse,
- and t-close table

Perfect privacy?

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

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

k-Anonymity ≠ Privacy

- **Syntactic**
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - **"*k*-anonymous" dataset can leak sensitive information**

    Background knowledge exists!

- **"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