THE UNIVERSITY OF MELBOURNE

Workshop 12

COMP20008

Elements of Data Processing

Zijie Xu

Agenda

- Recommender systems
  - Collaborative filtering
- Privacy
  - k-Anonymity and l-diversity
- Differential privacy
  - Privacy loss budget *k,* Global sensitivity *G, G*/*k* ratio

Recommender system

- Collaborative filtering is one of the approaches to recommend stuff to users
- Assumption: People who give similar ratings to stuff tends to like similar stuff
- Making predictions about a user according to the collective behaviour of many other users (c.f. content-based recommendations)
- Input data: Rating matrix (*m* users by *n* items)

	Item 1	Item 2	…	Item j	…	Item n
User 1	3	4	2	2		1
User 2	2					5
⋮	1	1	2	2		
User i	3	2		r_{ij} ??		
⋮				2		3
User m		5	3	4	3	

Recommender system

- Two approaches to predict rating of item *j* by user *i*, r_{ij}
  - Item-based:
    - Find *k* items similar to item *j*
    - Predict r_{ij} by combining ratings of the *k*-most similar items rated by user *i*
  - User-based:
    - Find *k* users similar to user *i*
    - Predict r_{ij} by combining ratings of the *k*-most similar users' ratings
- Need to choose: value of k, method of imputation, similarity measure, method to combine ratings
  - e.g. k=3, mean imputation, 1/(1+Euclidean), weighted average
- Imputed values should not be used in predict -> choose next best value

Privacy

- Objective: protect individuals from being identified from public data releases that contain sensitive attributes
- A person may be identified by:
  - an explicit identifier (e.g. driver's license number)
  - some combination of Quasi-Identifiers QI (e.g. {Gender, DOB, zip code})

Privacy

- Some strategies to improve privacy
  - *k*-anonymity: Each combination of QI values has at least "*k*" records
- Two methods to achieve *k*-anonymity
  - Generalisation
  - Suppression (i.e., deletion)

Age	Gender	Medical Condition
25	M	Diabetes
25	M	Diabetes
30	F	Heart Disease
30	F	Diabetes
40	M	Asthma
40	M	Asthma
45	F	Asthma
45	F	Diabetes

Privacy

- Some strategies to improve privacy
  - *l*-diversity: Each combination of QI values (i.e. each *k*-anonymous group) has at least "*l*" different sensitive attributes

Age	Gender	Medical Condition
25	M	**Diabetes**
25	M	**Asthma**
30	F	**Heart Disease**
30	F	**Diabetes**
40	M	**Asthma**
40	M	**Heart Disease**
45	F	**Asthma**
45	F	**Diabetes**

Privacy

- Consider the quasi-identifier {Favourite Genre, Year of Birth, Postcode}. The sensitive attribute is “Customer Value”.
- Q1.1: What is the highest k for which this data is k-anonymous? Explain and justify your answer.

Favourite Genre	Year of Birth	Postcode	Customer Value
Action	****	3100	High
Action	****	3100	High
Gore	1998	3104	Medium
Gore	1998	3104	Low
Action	2001	3100	Medium
Action	2001	3100	Low

Privacy

- Consider the quasi-identifier {Favourite Genre, Year of Birth, Postcode}. The sensitive attribute is “Customer Value”.
- Q1.1: What is the highest k for which this data is k-anonymous? Explain and justify your answer.
  - k=2

Favourite Genre	Year of Birth	Postcode	Customer Value
Action	****	3100	High
Action	****	3100	High
Gore	1998	3104	Medium
Gore	1998	3104	Low
Action	2001	3100	Medium
Action	2001	3100	Low

Privacy

- Consider the quasi-identifier {Favourite Genre, Year of Birth, Postcode}. The sensitive attribute is “Customer Value”.
- Q1.2: Describe one possible privacy attack on this data.

Favourite Genre	Year of Birth	Postcode	Customer Value
Action	****	3100	High
Action	****	3100	High
Gore	1998	3104	Medium
Gore	1998	3104	Low
Action	2001	3100	Medium
Action	2001	3100	Low

Privacy

- Consider the quasi-identifier {Favourite Genre, Year of Birth, Postcode}. The sensitive attribute is “Customer Value”.
- Q1.2: Describe one possible privacy attack on this data.
- Homogeneity attack

Favourite Genre	Year of Birth	Postcode	Customer Value
Action	****	3100	High
Action	****	3100	High
Gore	1998	3104	Medium
Gore	1998	3104	Low
Action	2001	3100	Medium
Action	2001	3100	Low

Privacy

- Consider the quasi-identifier {Gender, DoB, Postcode}. The sensitive attribute is “Customer Value”.
- Q2: Apply generalisation to the following table to make it 3-anonymous. Use $*$ to suppress certain identifiable fields and values.

Name	Gender	DoB	Postcode	Customer Value
Sophie	F	3/2/1998	3100	High
Jessica	F	24/12/1998	3100	High
Mia	F	4/04/1998	3104	High
Zachary	M	1/01/2001	3010	Medium
Nicholas	M	3/2/2001	3010	Medium
Joshua	M	31/12/2001	3000	Medium

Privacy

- Consider the quasi-identifier {Gender, DoB, Postcode}. The sensitive attribute is “Customer Value”.
- Q2: Apply generalisation to the following table to make it 3-anonymous. Use $*$ to suppress certain identifiable fields and values.

Name	Gender	DoB	Postcode	Customer Value
Sophie	F	**/**/1998	31**	High
Jessica	F	**/**/1998	31**	High
Mia	F	**/**/1998	31**	High
Zachary	M	**/**/2001	30**	Medium
Nicholas	M	**/**/2001	30**	Medium
Joshua	M	**/**/2001	30**	Medium

Privacy

- Consider the quasi-identifier {Age, Postcode} for the table below. The sensitive attribute is “Diagnosis”.
- Q3 discussion:
- What it means for a dataset to be ℓ-diverse.
- Why medical data should be kept private? How can an adversary use this information maliciously?

Age Range	Postcode	Diagnosis
[21-28]	3***	COVID-19
[21-28]	3***	Flu
[21-28]	3***	Flu
[48-55]	31**	Cancer
[48-55]	31**	Obesity
[48-55]	31**	Obesity

Privacy

- Consider the quasi-identifier {Age, Postcode} for the table below. The sensitive attribute is “Diagnosis”.
- Q3.1: What is the highest k for which this data is k-anonymous?

Age Range	Postcode	Diagnosis
[21-28]	3***	COVID-19
[21-28]	3***	Flu
[21-28]	3***	Flu
[48-55]	31**	Cancer
[48-55]	31**	Obesity
[48-55]	31**	Obesity

Privacy

- Consider the quasi-identifier {Age, Postcode} for the table below. The sensitive attribute is “Diagnosis”.
- Q3.2: What is the highest l for which this data is l-diverse?

Age Range	Postcode	Diagnosis
[21-28]	3***	COVID-19
[21-28]	3***	Flu
[21-28]	3***	Flu
[48-55]	31**	Cancer
[48-55]	31**	Obesity
[48-55]	31**	Obesity

Privacy

- Consider the quasi-identifier {Age, Postcode} for the table below. The sensitive attribute is “Diagnosis”.
- Q3.3: Describe one possible privacy attack on this data

Age Range	Postcode	Diagnosis
[21-28]	3***	COVID-19
[21-28]	3***	Flu
[21-28]	3***	Flu
[48-55]	31**	Cancer
[48-55]	31**	Obesity
[48-55]	31**	Obesity

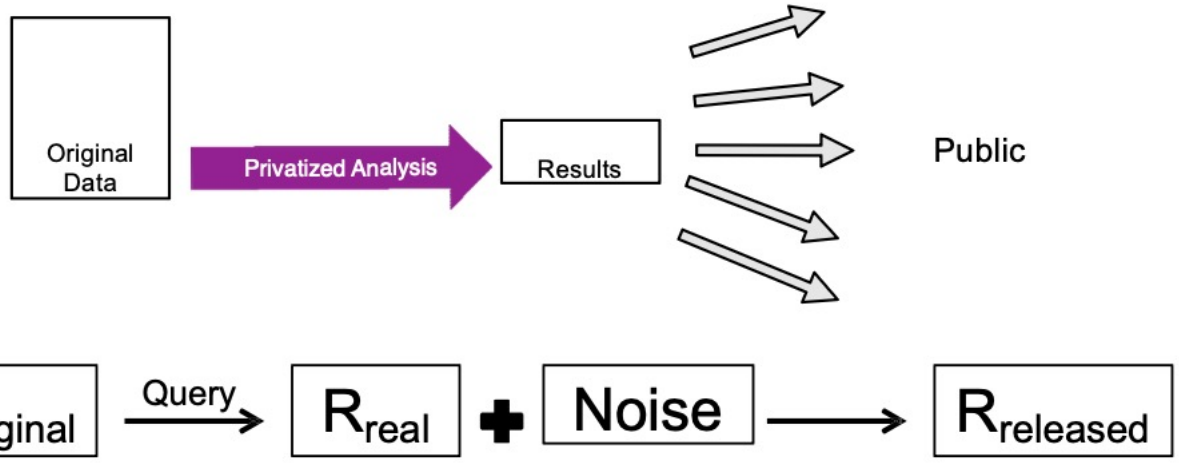
Privacy

- Consider the quasi-identifier {Age, Postcode} for the table below. The sensitive attribute is “Diagnosis”.
- Q3.3: Describe one possible privacy attack on this data
- Background attack

Age Range	Postcode	Diagnosis
[21-28]	3***	COVID-19
[21-28]	3***	Flu
[21-28]	3***	Flu
[48-55]	31**	Cancer
[48-55]	31**	Obesity
[48-55]	31**	Obesity

Differential Privacy

- Objective: ensures that the presence or absence of a particular person's data does not significantly affect the results
- Achieved by adding controlled noise to the released data



Differential Privacy

- Privacy loss budget *k*
  - A parameter on how private we want the result to be
  - Pr(R=…|I participate) ≤ Pr(R=…|I don't participate) • 2^*k*
  - Chosen by the data owner
- Global sensitivity *G*
  - The maximum change in the query/function output due to the addition or removal of a single data point in the dataset
  - Determined by the property of data + query

Differential Privacy

- *G/k* ratio
  - Controls the spread/deviation of the added random noise
  - *G/k* = scale in Laplace noise = standard deviation in Gaussian Noise
  - High *G*

    -> Output easily changed by addition/removal of one data point

    -> Need more deviation in noise
  - Low *k*

    -> small loss budget

    -> Need stronger privacy guarantee

    -> Need more deviation in noise

Differential Privacy

- Consider a query that outputs **CountFemale + CountMarried**.
- How much can adding or removing an individual affect the output? What is the global sensitivity G?

Sex	Marital Status
M	Single
M	Married
F	Single
M	Single
F	Married

Differential Privacy

- Consider a query that takes the survey database as input and outputs the statistics **CountMaleMarried + CountMaleSingle + CountFemaleMarried + CountFemaleSingle.**
- How much can adding or removing an individual affect the output? What is the global sensitivity?

Sex	Marital Status
M	Single
M	Married
F	Single
M	Single
F	Married

Slides

- Slides available on GitHub:
  ccijjj/COMP20008-23s2

Thank you and good luck to your exams!

THE UNIVERSITY OF MELBOURNE

Thank you

More Resources: Canvas