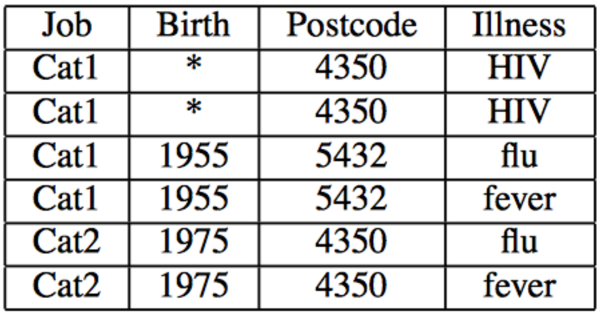
Workshop Week 12 - COMP20008 2021

Questions

1. Consider the quasi-identifier {job,birth,postcode}.

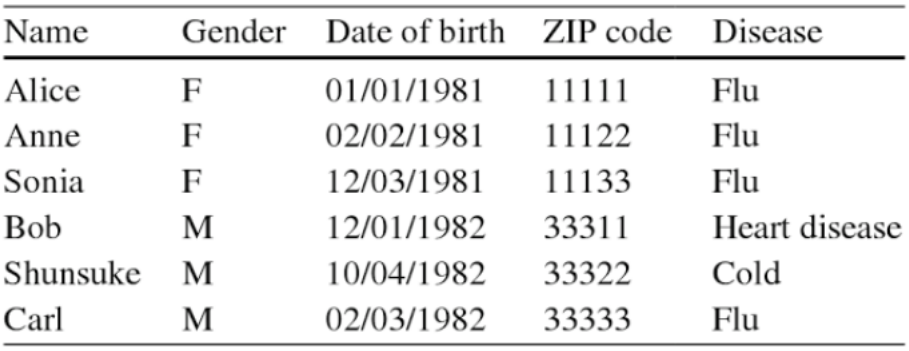


Job	Birth	Postcode	Illness
Cat1	*	4350	HIV
Cat1	*	4350	HIV
Cat1	1955	5432	flu
Cat1	1955	5432	fever
Cat2	1975	4350	flu
Cat2	1975	4350	fever

   (a) What is the highest k for which this data is k-anonymous. **2-anonymous**

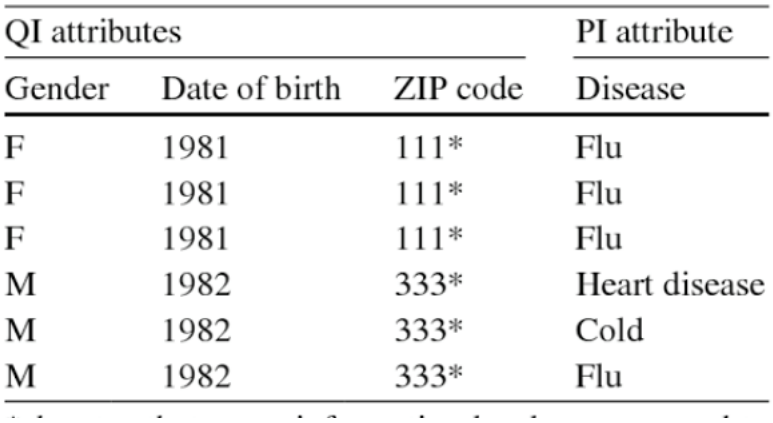
   (b) Describe one possible privacy attack on this data **Homogeneity attack - If I know from other background information that someone lives in postcode 4350 and was not born in 1975, I can conclude they have HIV.**

2. Consider the quasi-identifier {gender,date of birth,zipcode}. Apply generalisation to the following table to make it 3 anonymous.



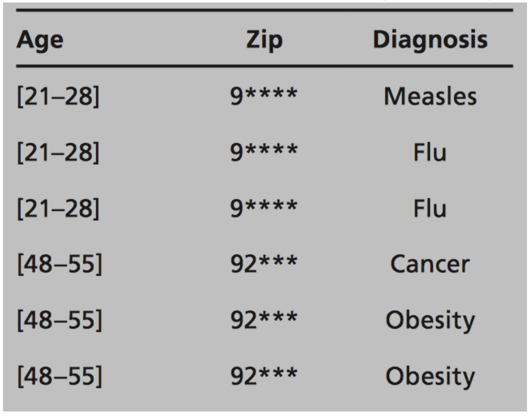
Name	Gender	Date of birth	ZIP code	Disease
Alice	F	01/01/1981	11111	Flu
Anne	F	02/02/1981	11122	Flu
Sonia	F	12/03/1981	11133	Flu
Bob	M	12/01/1982	33311	Heart disease
Shunsuke	M	10/04/1982	33322	Cold
Carl	M	02/03/1982	33333	Flu

One possible answer. We would first get rid of name attribute before doing any further processing, we then just work with the quasi identifiers and sensitive (attribute)



QI attributes			PI attribute
Gender	Date of birth	ZIP code	Disease
F	1981	111*	Flu
F	1981	111*	Flu
F	1981	111*	Flu
M	1982	333*	Heart disease
M	1982	333*	Cold
M	1982	333*	Flu

3. Consider the quasi identifier {Age,Zip} for the table below.



Age	Zip	Diagnosis
[21–28]	9****	Measles
[21–28]	9****	Flu
[21–28]	9****	Flu
[48–55]	92***	Cancer
[48–55]	92***	Obesity
[48–55]	92***	Obesity

   (a) What is the highest k for which this data is k-anonymous? **3-anonymous**

   (b) What is the highest l for which this data is l-diverse? **2-diverse**

   (c) Describe one possible privacy attack on this data **One possibility: Background attack. For example, if I know that someone aged 48-55 in postcode 92*** doesn't have Cancer, I can conclude they have Obesity**

4. In the context of providing differential privacy:

   - What is global sensitivity G? What is the privacy budget k?
   - How does the G/k ratio affect the noise level?
     Global sensitivity is evaluating the maximum possible change in query output due to a presence of a single record. The privacy budget determines how close the query result for a database with the record is expected to be compared to query result for a database without the record. For smaller k, or larger global sensitivity, more noise will be added to the query result.

5. Consider a survey that collects two values from the respondents, e.g., marital status and sex.

- Consider a query that takes the survey database as input and outputs a pair of counts (CountNumberFemale,CountNumberMarried). How much can adding or removing an individual affect the output? What is the global sensitivity?
- Consider a query that takes the survey database as input and outputs the quadruplet of counts (CountMaleMarried,CountMaleSingle,CountFemaleMarried,CountFemaleSingle). How much can adding or removing an individual affect the output? What is the global sensitivity?

  In the first case

  Adding or removing any individual can affect the count of each column by maximum 1. The maximum difference a single record can make query is 1+1=2

  In the second case

  Adding or removing any individual can affect the count of each column by maximum 1. If it affects the count of some column by one, then it will affect the counts of other columns by 0 (since the columns are mutually exclusive, you can't have a 1 in both columns)
  The maximum difference a single record can make To F is thus 1+0+0+0=1

Other Previous Exam Questions

Exam 2018 - Question 4

University X is planning to build a recommender system for its students. Based on subjects they have enrolled in, the system will recommend new subjects they might consider studying in the future.

A table showing a fragment of the data input to the system is below. The columns correspond to the codes of all the subjects in the University handbook. Rows correspond to students and whether they have enrolled in a subject ("Yes" if they have previously enrolled and "-" otherwise). The dataset covers the period 2010-2017, with 100,000 students and 3000 subjects. It is proposed that subject recommendations should be made using user based collaborative filtering.

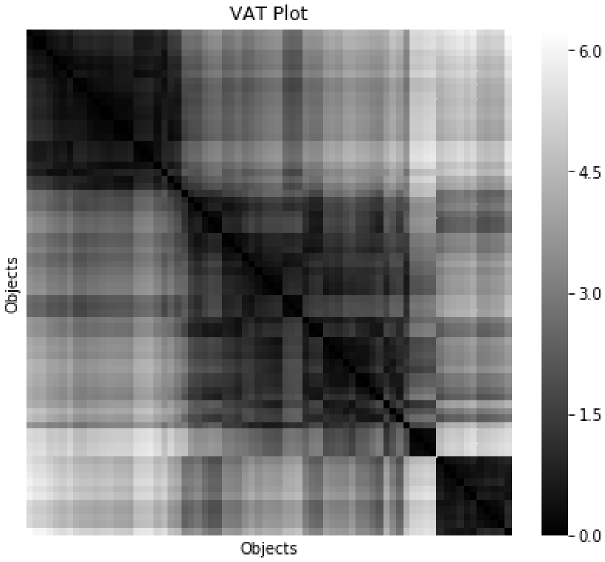
PersonName	Subject1	Subject2	Subject3	Subject4	Subject5	...
Alice	Yes	-	-	-	-	...
Bob	Yes	Yes	-	Yes	-	...
Margaret	Yes	Yes	-	-	-	...
...	...	...	...	...	...	...

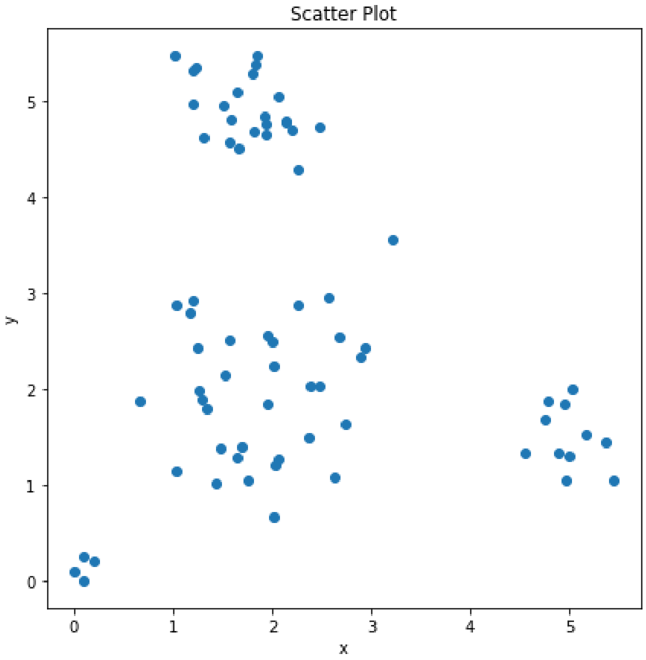
Explain three challenges for making this recommendation approach effective.
-cold start - how to handle new students with no history of subjects, eg. starting in 1st year
-how to avoid over-recommending popular subjects and under recommending rare ones

-how to incorporate domain knowledge about degree requirements and compulsory subjects

Need to make three plausible points. Each point needs to be clear. Points should be distinct from each other

Exam 2019 - Question 3





(a) (1.5 mark) Are these likely to have come from the same dataset? Give two reasons why or why not.

(b) (1.5 marks) Consider a dataset with 10000 rows and 500 features. Give three reasons why we might want to apply PCA while analysing the dataset.

Part a: Yes.

- **Upper cluster in VAT corresponds to upper cluster in scatter plot**
- **Centre cluster corresponds on VAT to large cluster in the middle of the dataset.**
- **Centre cluster less well defined = higher spread in centre cluster on scatter plot**
- **Well defined section at bottom of centre cluster on VAT corresponds to points around (0,0)**
- **Bottom cluster on VAT corresponds to right cluster on scatter plot**

Part b

- **Computational efficiency**

- **Reduce noise**
- **Allow visualisations**

Exam 2019 - Question 2b

(3 marks) Consider the following regular expression:

```
int[1-9]+=[A-Z][a-z]*
```

Which of the following terms would matched by this regular expression (write down all that apply):

```
int=Abc
int2=A
int3=Abc
int44=Azz
int5Abc
int6=abc
```

int2=A
int3=Abc
int44=Azz