Workshop Week 12 - COMP20008 2018

Questions

1. Consider a simple blockchain with three blocks. Each block consists of header and some data. The ID of a block is equal to a hash of its header. The header is a string consisting of the following items in sequence, each separated by a '+' character.

- ID of the parent block
- Timestamp (DDMMYYHHMM format)
- Hash of the data stored in the block

The following table gives a partial description of the first three blocks in this blockchain. Row 1 is the first block, row 2 is the second block, row 3 is the third block. Assume there exists a dummy block zero, (not included) with ID 000.

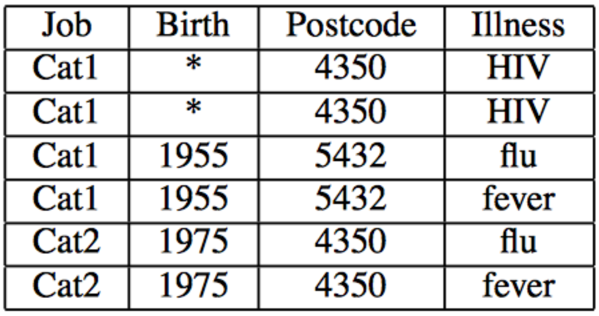
Block ID	**Header**	**Data**
a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2	000+1205170931+?	COMP20008
?	?+1305171033+?	COMP20009
?	?+1405172333+?	COMP20010

Each '?' in the table represents a 64 digit hexadecimal value. Replace each '?' by its appropriate value, to maintain the integrity of the blockchain. Use the SHA-256 function for hashing, available at this website.

2. Suppose the data for the first block was modified to read *COMP20008!*, describe (at a high level) how would this would affect the rest of the blockchain. Suppose the timestamp for the second block was modified, describe (at a high level) how this would affect the rest of the blockchain.

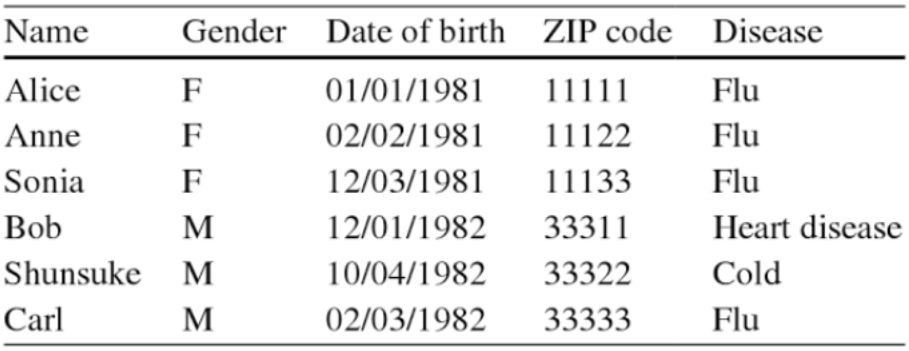
3. Suppose Bob signs a document with his digital signature. Fred receives the document and changes its contents, but leaves the digital signature unchanged. How could a third party (Alice), know that the document has been modified from its original version, by someone other than Bob?

4a. Consider the quasi-identifier {job,birth,postcode}. Is data in the following table 1-anonymous? Is it 2-anonymous? Is it 3-anonoynous? Is it 4-anonymous?



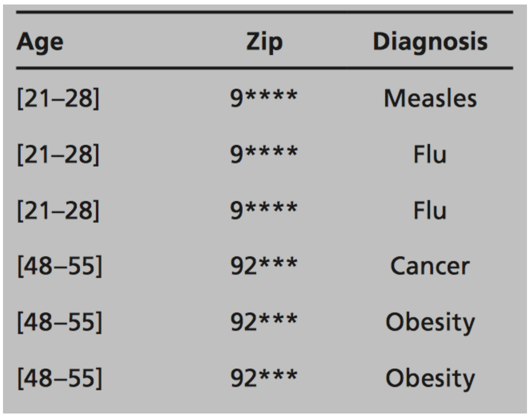
Job	Birth	Postcode	Illness
Cat1	*	4350	HIV
Cat1	*	4350	HIV
Cat1	1955	5432	flu
Cat1	1955	5432	fever
Cat2	1975	4350	flu
Cat2	1975	4350	fever

4b. 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

4c. Consider the quasi identifier {Age,Zip} for the table below. Is the data 1-anonymous? Is it 2-anonymous? Is it 3 anonymous? Is it 4 anonymous? Is it 5 anonymous? With respect to the sensitive attribute Diagnosis - is it 1-diverse? Is it 2-diverse? Is it 3-diverse? Is it 4-diverse?



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

5. In the context of providing differential privacy:

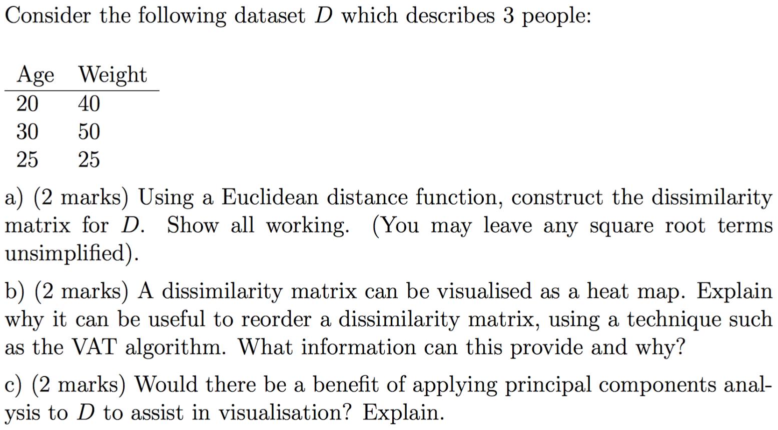
- What is global sensitivity G? What is the privacy budget k?
- How does the G/k ratio affect the noise level?

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

- 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?

From the 2016 exam

7.



Consider the following dataset D which describes 3 people:

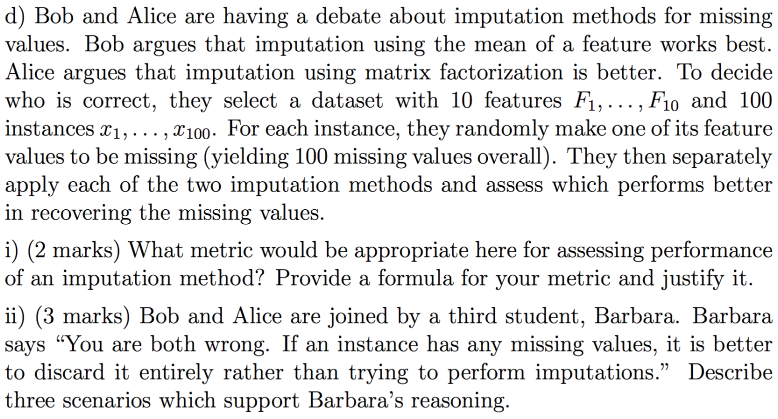
Age	Weight
20	40
30	50
25	25

a) (2 marks) Using a Euclidean distance function, construct the dissimilarity matrix for D. Show all working. (You may leave any square root terms unsimplified).

b) (2 marks) A dissimilarity matrix can be visualised as a heat map. Explain why it can be useful to reorder a dissimilarity matrix, using a technique such as the VAT algorithm. What information can this provide and why?

c) (2 marks) Would there be a benefit of applying principal components analysis to D to assist in visualisation? Explain.

8.



d) Bob and Alice are having a debate about imputation methods for missing values. Bob argues that imputation using the mean of a feature works best. Alice argues that imputation using matrix factorization is better. To decide who is correct, they select a dataset with 10 features $F_1, \ldots, F_{10}$ and 100 instances $x_1, \ldots, x_{100}$. For each instance, they randomly make one of its feature values to be missing (yielding 100 missing values overall). They then separately apply each of the two imputation methods and assess which performs better in recovering the missing values.

i) (2 marks) What metric would be appropriate here for assessing performance of an imputation method? Provide a formula for your metric and justify it.

ii) (3 marks) Bob and Alice are joined by a third student, Barbara. Barbara says "You are both wrong. If an instance has any missing values, it is better to discard it entirely rather than trying to perform imputations." Describe three scenarios which support Barbara's reasoning.