Workshop Week 10 - COMP20008 2018

Questions

1. Consider the 3 party protocol for privacy preserving linkage with exact matching, discussed in lectures.

    - What is a salt?
    - Explain why a salt is used.
    - Who chooses and knows the salt?
    - What assumptions should be made about the level of trust required for the 3rd party (person C)?

2. Consider the 3 party protocol for privacy preserving linkage with approximate matching, discussed in lectures.

    - What is a bloom filter?
    - How is it used to assist the matching process?

3. A bloom filter is used to store the set of 2-grams from a string. Two strings are then compared for similarity by computing the Dice coefficient for their respective bloom filters (formula in Lecture 19) $sim(b1, b2) = \frac{2h}{b1+b2}$. Consider the following two alternative similarity measures that might be used. Explain their advantages/disadvantages compared to the Dice coefficient for evaluating bloom filter similarity.

    - Hamming similarity: $sim(b1, b2) = s/l$, where s is the number of bits which are the same in $b1$ and $b2$ and l is the bit vector length.
    - Jaccard similarity: $sim(b1, b2) = \frac{h}{l}$ where l is the bit vector length and h is the number of bits set to 1 in both bloom filters..

4. For bloom filters of length l and using k hash functions. Consider the ratio l/k.

    - As l/k increases, would you expect the matching accuracy of the system to become better or worse? Would you expect the robustness of the system to frequency attack (by the trusted 3rd party) to become better or worse as l/k increases? Why?

5. Suppose a bank wishes to perform data linkage to match the customers in its loan application database, against public twitter feeds (to help the bank more accurately assess customer risk).

    - Based on your knowledge of Twitter, how feasible do you believe this would be?
    - What legal and ethical issues could be relevant here?

Revision: 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.

Exam 2018 - Question 12

Business X and Business Y have decided to conduct a joint marketing campaign. For this marketing campaign, they need to determine how many customers they have in common (how many people are in the customer list of both businesses). They implement the following 2 party privacy preserving protocol, making use of the SHA-256 one way hashing function.

```
#In the following, the '+' symbol indicates string concatentation (joining two strings)

#Business X does the following
SetX=empty
For each customer at Business X
    SetX=SetX ∪ SHA-256("First Name"+"Last Name")
Send SetX to Business Y

#Business Y does the following
SetY=empty
For each customer at Business Y
    SetY=SetY ∪ SHA-256("First Name"+"Last Name")
result=count(SetX ∩ SetY)
Share result with Business X
```

a) Explain a privacy drawback of this protocol.

b) Explain how the protocol could be modified to eliminate this privacy drawback.

Discussion

For Phase 3, you will be giving a 5 minute oral presentation and have been asked to cover the following points:

- What is the research question?
- Why is it worth tackling (i.e. motivation)?
- What are the datasets you used and why?
- What data wrangling methodologies have you used to investigate your research question?
- What did you find? Why is it interesting? What have you learnt?
- What have been the challenges and what (if anything) would you have done differently?

Sketch and discuss a plan for how you will cover these points (number of slides, time for each slide, where to include figures, number of points per slide). How can you achieve high clarity for your talk?