COMP20008
Week 10 Tutorial 9

Data Linkage Privacy

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

- What is a salt?
  - An extra string that is appended to the information being encoded, so that hashed value is not susceptible to a dictionary attack. The two parties doing the linkage would agree on a salt, the 3rd party would not know it.
- Explain why a salt is used.
  - It is used to prevent a dictionary attack by a 3rd party.
- Who chooses and knows the salt?
  - The two parties A and B who are integrating their data choose and know the salt.
- What assumptions should be made about the level of trust required for the 3rd party (person C)?
  - C does not know the salt.

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

- What is a bloom filter?
  - A bloom filter is an array of l bits, with all bits initially set to zero.
  - We use it to store a set of strings $\{X_1, X_2, X_3, \dots, X_n\}$
    - Each element X_j is hash coded using the k hash functions and all bits having indices $H_i(X_j)$ are set to 1 ($for\ 1 \leq i \leq k$ and $for\ 1 \leq j \leq n$).
    - If a bit was set to 1 before, no change is made.

- How is it used to assist the matching process?
  - A bloom filter can be used to comparing two strings for approximate similarity in a private manner
    - The 2-grams of the first string are stored in bloom filter B1, the 2-grams of the second string are stored in bloom filter B2. Both bloom filters are the same length and use the same hash functions.
    - The two strings are then compared for similarity by computing the Dice coefficient: $sim(B1, B2) = 2h / b1 + b2$
      - Where h is the number of bits set to 1 in both bloom filters
      - b1 is the number of bits set to 1 in bloom filter B1
      - b2 is the number of bits set to 1 in bloom filter B2
    - If the two strings have a lot of 2-grams in common, then their bloom filters will have a large number of identical bit positions set to 1.

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): $sim(b1, b2) = 2h/b1 + b2$.

- 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: $sim(b1; b2) = h/l$ where l is the bit vector length and h is the number of bits set to 1 in both bloom filters.

Solution

- Example: having B1= 11000010001111 and B2=11100001101111 , Calculate the three similarity measures
- Hamming similarity: Overestimates the similarity artificially due to the matching of the 0's
- Jaccard similarity: Underestimates the similarity artificially due to the division by the number of bits l

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?

3 FACTORS: SPEED – ACCURACY - PRIVACY

- The ratio results in sparse or less sparse representations.
- When the representation is sparse (number of bits is LARGE and number of hash functions is SMALL then
  - The privacy is lower;
  - The accuracy of membership queries is increased, since there are less collisions.
  - The speed is also slower (more bits to scan).
- The opposite extreme case (the number of bits is SMALL and the number of hash functions is LARGE then
  - The privacy is high since we can't determine the original records.
  - The accuracy of membership queries is very low. (the bloom filter will consist of mostly ones)
  - The filter is relatively faster because there are fewer bits.

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?

6. (Discussion) For Phase 4, 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?