Advanced Topics on Privacy Enhancing Technologies
CS-523
Privacy-preserving Data Publishing I

Exercise 1

Are the following statements TRUE or FALSE:

(a) The columns of a synthetic dataset can be considered quasi-identifiers.

(b) Enforcing l-diversity on a k-anonymous dataset increases the number of k-anonymous equivalent classes.

(c) t-closeness is not a sufficient condition for anonymity.

(d) Anonymization always impacts utility.

Exercise 2

You are working at Quitter, a social network for people that want to change their job. As those people will soon go to the job market, you think you could monetize the information you have by selling it to advertisers and job brokers. Of course, you are aware that there is regulation that constrains what you can share, and also that if you share too much you may lose the trust of your users and they may leave your social network. The information you want to share is the social graph of Quitter users, where nodes represent users and edges relationships between users.
Are the following good anonymization strategies for this scenario? Justify your answer (*Hint: Think about the auxiliary information a potential adversary could use, and also about the impact on utility of the shared information for others*):

1. Publish all nodes and edges, remove user names, but leave detailed profile information such as previous employment history, date of birth, skills, attached to each node.

2. Publish all nodes and edges, remove any user and profile information from the nodes.

3. Publish all nodes but only a subset of edges, remove any user or profile information from the nodes.

Exercise 3

A famous restaurant has released an anonymized version of the guests seated at their most popular table for dinner over the course of one week:

Monday	customer1, customer5, customer7, customer14
Tuesday	customer10, customer5
Wednesday	customer2, customer6, customer9
Thursday	customer1, customer5, customer4
Friday	customer10, customer11, customer12, customer13, customer8
Saturday	customer10, customer3

(a) Ally has gone to this restaurant once during this week for a couples dinner with her partner and two friends where they were seated at the most popular table. Looking at this anonymized data, what can she learn about her partner and her friends beyond what she might have already known before?

(b) Can k-anonymity or l-diversity help the restaurant prevent Ally's inferences?

(c) Can publishing a synthetic version of these data, where the pattern learned by the generative model would be the distribution of eaters over the week, help the restaurant prevent Ally's inferences?

Exercise 4

The two tables shown below have been released by two hospitals. You know that Bobby was a patient in both hospitals. What can you learn about Bobby, who just shared a social media post about celebrating his 25th birthday?

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Hospital A (4-anonymous)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

Hospital B (6-anonymous)

Solutions to the Exercises

Solution 1

(a) False, the attribute values of records in a synthetic dataset do not have a direct link to real records. The columns of a synthetic dataset thus can *not* be used as quasi-identifiers in a re-identification attack that tries to match a real record to a specific synthetic record.

(b) False, l-diversity is a property of an equivalent class and does not affect the number of k-anonymous groups.

(c) True. We have seen in the lecture that anonymity is a fragile property because it is not possible for the data holder to know which attributes a privacy adversary will exploit as quasi-identifier.

(d) True. Anonymization methods always modify the data and hence, in general, affect data utility. It can happen that for a concrete utility function, e.g. "count the number of records in a database", we can find an anonymization mechanism that does not reduce the utility for this specific analysis.

Solution 2

- Employment history and skill are a quasi-identifier. An adversary with access to these user attributes from another data source, e.g., LinkedIn, can easily de-anonymize users.
- Users have the same friends in different networks. The shape of the clusters (nodes and edges linking them) are very similar across social networks. Knowing another social network with the identity of the nodes, can be used to de-anonymize a social network with the same users.
- Removing edges does decrease the chance of the attack that uses auxiliary information from other social networks. However, to prevent linkage

attacks one has to remove a large fraction of edges. At that point, the utility of the graph is severely reduced.

More about graph de-anonymization in: `https://www.cs.cornell.edu/~shmat/shmat_oak09.pdf`

Solution 3

(a) Among others, Ally can learn:

- She knows what day she went to the restaurant, so she can identify which entry belongs to their group (in any case there is only one day with four customers, Monday).
- That either her partner or one of her two friends went for dinner on Tuesday with someone else that was not part of the group on Monday.
- That either her two friends or her partner and one of her friends went for dinner with a third person on Thursday.
- That the person that had dinner with her partner or one of her friends went for dinner with another four people on Friday.
- That the person that had dinner with her partner or one of her friends went for another potentially romantic dinner on Saturday.

(b) No, k-anonymity (and therefore l-diversity) does not apply in this case. What would the restaurant owner make anonymous? There are no records associated to people.

(c) Yes, a synthetic version of this data would help. The data would represent the distribution of customers per week, but not necessarily per day. Ally would not be able to identify her table, and therefore could not continue performing the chain of inferences.

Note that this is a very different case to the ones seen in the course, both in the nature of the data being published, and in the nature of the inferences we are worried about. If the worry would be: can Ally (or someone with auxiliary information about Ally) learn whether her data is in this dataset, synthetic data would not help.

Solution 4

You can learn that Bobby has AIDS. This is the only condition that appears in both tables for patients under 30 year old.