EN4720: Security in Cyber-Physical Systems
Exercise — Privacy

Name: Thalagala B. P.
Index No: 180631J

June 24, 2023

This is an individual exercise!
Due Date: 24 June 2023 by 11.59 PM

Section 1

Provide answers to questions 1 through 7 in the given space referring to the Tables 1 and 2 given below.

Table 1: Example dataset on the weekly usage of taxis by certain individuals

Name	**Age**	**Gender**	**Occupation**	**Average No. of taxi trips per week**
943145	21	Female	Legal Counsel	15
416765	38	Male	Data Privacy Officer	2
356891	44	Female	Database Administrator	3
723145	25	Female	Administrative Assistant	1
239976	31	Male	Data Privacy Officer	5
562396	42	Female	Programmer	3
964825	22	Female	Administrative Assistant	4
873892	30	Female	Legal Counsel	2

1. What type of data anonymization technique is used for the dataset given in Table 1?

   ***De-identification**: actual names of the individuals have been replaced by a numerical identifier. Because, the attribute 'Name' of a data record is a direct identifier which can be used to identify a given individual uniquely.*

2. Is this technique sufficient to protect the privacy of the associated individuals? If not, why?

   *No. Removing only the name can not guarantee the preservation of privacy, because it can be **susceptible to linkage attacks** which leads to re-identification of the individuals by linking the information with other external information (zip, sex, birth date).*

3. Briefly describe three other data anonymization techniques that can be used to anonymize data.

   - ***Generalization**: Rather than including the specific information about an attribute (eg: age = 21) in a record, it can be replaced with more generalized value (eg: range of ages = (20 to 30)).*

- ***Suppression***: *Some fields of the data base can be entirely removed, if that field contains highly unique identifiers such as national ID card numbers and telephone numbers of the individuals.*
- ***Data swapping***: *some of the values of a given record can be swapped with some other record's value while keeping the statistical properties of the database intact.*

Table 2: Modified example dataset on the weekly usage of taxis by certain individuals

Age	**Gender**	**Occupation**	**Average No. of taxi trips per week**
21 to 30	Female	Legal Counsel	15
31 to 40	Male	Data Privacy Officer	2
41 to 50	Female	IT	3
21 to 30	Female	Administrative Assistant	1
31 to 40	Male	Data Privacy Officer	5
41 to 50	Female	IT	3
21 to 30	Female	Administrative Assistant	4
21 to 30	Female	Legal Counsel	2

4. The example dataset given in Table 1 was modified to improve anonymity. The new dataset is provided in Table 2. Mention all the data anonymization techniques that were used to achieve this.

   - ***Generalization***: *ages have been replaced with ranges of ages, some value of the occupation filed hs been replaced with its parent set's value* `IT = {Database Administrator, Programmer, ...}` *the latter is also known as 'data aggregation'*
   - ***Suppression***: *Name field has been removed entirely*

5. Can k-anonymity be observed in the data given in Table 2? If so, what is the value of k?

   Yes. $K = 2$

6. Is privacy guaranteed for the data given in Table 2? Justify your answer.

   Although the above methods reduce the risk of re-identification, it is not possible to grantee the preservation of privacy of data. The database can be still ***susceptible to attribute disclosure attacks****, if a given equivalence class (the group of k records with the same quasi-identifiers) has the same value for the given sensitive attribute. Let's assume the sensitive attribute in the Table 2 is the Average No. of taxi trips per week. Then if we consider the 3^{rd} and 6^{th} records in that table they have the same value for the sensitive attribute.*

7. Calculate the risk of re-identification for data given in Table 2 (mention as a percentage). Justify your answer.

Consider below tables with rearranged records for easy comparison about the linkages between them. Let's define the risk of re-identification as the ratio between the number of actual records (always one) and the number of potential record matches. The matching procedure depends only on the quasi-identifiers (age, gender, occupation)

As an example take the record of the name 943145. The risk R of re-identifying the name corresponding to the average taxi trips of 15, using the records given in the Table 4,

$$R = \frac{1}{2} \times 100\% = 50\%$$

Similarly, risk is calculated for all the other names and are given in the Table 3. For the names in 5^{th} and 6^{th} this risk is 100% as there is no diversity in the values in the Average No. of taxi trips per week column.

Table 3: Risk of re-identification

Name	**Age**	**Gender**	**Occupation**	Risk (%)
943145	21	Female	Legal Counsel	50
873892	30	Female	Legal Counsel	50
416765	38	Male	Data Privacy Officer	50
239976	31	Male	Data Privacy Officer	50
356891	44	Female	Database Administrator	100
562396	42	Female	Programmer	100
723145	25	Female	Administrative Assistant	50
964825	22	Female	Administrative Assistant	50

Table 4: Rearranged example dataset

Age	**Gender**	**Occupation**	**Average No. of taxi trips per week**
21 to 30	Female	Legal Counsel	15
21 to 30	Female	Legal Counsel	2
31 to 40	Male	Data Privacy Officer	2
31 to 40	Male	Data Privacy Officer	5
41 to 50	Female	IT	3
41 to 50	Female	IT	3
21 to 30	Female	Administrative Assistant	1
21 to 30	Female	Administrative Assistant	4

8. Suggest ways to enhance the privacy of this dataset considering l-diversity.

L-diversity ensures that withing each equivalence group (quasi-identifier group) there are at least l distinct values for each attribute. This make it harder to learn the sensitive attribute through attribute disclosure attacks. In order to achieve l-diversity, generalization and suppression techniques has to be used on the original data set recursively until we get l distinct values under a given attribute, within every quasi-identifier group.

Section 2

1. What is differential privacy?

Differential privacy ensures that the membership of a particular individual in a given database is not disclosed. That is, the data that can be inferred about an individual is roughly the same regardless of the fact that the data about that individual is actually available in the database or not. This functionality is achieved by adding a carefully calibrated noise to the output of the queries to the database.

2. Briefly describe what is **protected** and **not protected** by applying differential privacy.

- PROTECTED:
  - ***Membership privacy*** *of an individual (described above)*
  - ***Ability to withstand linkage attacks****, by making no assumptions about the availability of auxiliary information. That is even an attacker has some auxiliary info about a given individual, he will not be able to re-identify the individual by using the outputs of the queries.*
- NOT PROTECTED:
  - ***Utility of the data*** *is destroyed if unnecessary amount of noise is added to the real data. This can make accurate decision making impossible. (eg: a particular application may provide false or less useful information to the user depending on the data that the app accessed via a differential private protected database - location based suggestions)*
  - ***Privacy of groups****: Even though an attacker is unable to infer accurate information about a given individual, the retrieved information can still be used to gain insights about the various groups/ teams inside a database.*

3. The following mechanism is used to satisfy differential privacy where Z denotes the noise added.

$$F(x) = f(x) + Z$$

(a) Provide an equation to show the relationship between ϵ (privacy budget) and Z? Are they directly or inversely related?

$$Z \sim Laplace\,(b)\,,\ where\ b = \frac{\Delta}{\epsilon}$$

Here Δ denotes the global sensitivity whereas ϵ denotes the privacy budget. Z and privacy budget are inversely related.

(b) What does it mean to have $\epsilon = 0$?

$\epsilon = 0$ indicates the perfect membership privacy. That is output of a query must be exactly the same regardless of the fact that record about an individual is included or excluded in the database. This is practically not achievable and ϵ is always a trade off between privacy and the utility.

(c) What would be the result of adding a larger amount of noise?

Large noise will destroy the utilization of the data, and make it impossible to make accurate or useful decisions.