浙江大學

《Big Data Security and Privacy Protection》

Set 2

题　　目：	Set 2
上课时间：	August 27th
授课教师：	Rongxing Lu
姓　　名：	杜宗泽
学　　号：	3220105581
组　　别：	个人
日　　期：	8月27日

Set 2

1 Question 4

4. Consider a dataset about Hospital Covid19 Case, which mainly describes the status of covid19 patient in care at hospital with the following attributes. Now, the hospital wants to release the dataset for some potential scientific analytics, please prepare a design guideline for the hospital so that a released version of the dataset can balance the patients' privacy and the whole dataset's utility.

Data Source : csv

Name	Descriptions	Data Type	More Information
Name	Name of patient	String	
Gender	Gender	String(Enum)	Male, Female
Age	Age of patient	Int	
HospitalID	Hospital ID based on master data	Int	
HospitalName	Hospital Name	String	
IsPositive	Covid19 positive status	Boolean	
PatientCovidStatus	Covid19 suspected category	Enum (string)	*ODP, PDP*
IsTested	Patient is tested for swab test	Boolean	
TestedDate	Patient swab test date	DateTime	
PatientStatus	Patient status in hospital	String(Enum)	PassedAway, InCare, Healed
CreatedOn	Created record	DateTime	

Example:

- Fepri Putra, Male, 30, 1, RS.Sulianti Saroso, True, PDP, True, 10-05-2020 10:10:10, InCare, 10-05-2020 10:10:10

My answer:

To balance patients' privacy and the utility of the dataset when releasing a dataset about Hospital Covid19 Cases, the hospital can follow these design guidelines:

1. Anonymize Personally Identifiable Information (PII): Remove or generalize any attributes that directly identify individuals, such as names, addresses, social security numbers, or specific dates of birth. This helps protect patient privacy.(Just like the k-anonymity in question 5)

2. Aggregate Data: Instead of releasing individual-level data, aggregate the data to a higher level, such as by grouping patients based on age ranges, geographic regions, or other relevant categories. Aggregating data helps protect individual privacy while still providing useful insights.

3. Implement Access Controls: Limit access to the released dataset to authorized individuals or organizations. This helps prevent unauthorized use or disclosure of sensitive information.

4. Obtain Informed Consent: If possible, obtain informed consent from patients before releasing their data. This ensures that patients are aware of how their data will be used and allows them to make an informed decision.

5. Data Sharing Agreements: Establish data sharing agreements with the recipients of the dataset. These agreements should outline the purpose of data usage, restrictions on data sharing, and measures to protect patient privacy.

6. Regular Data Audits: Conduct regular audits to ensure compliance with privacy regulations and data protection measures. This helps identify any potential privacy risks and allows for timely corrective actions.

By following these design guidelines, the hospital can release a version of the dataset that balances patients' privacy and the utility of the dataset for scientific analytics.

2 Question 5

5. Incognito is one of approaches to implement the k-anonymity. Given a table below, the full-domain generalizations described by "domain vectors" are represented as follows.

   - $Z_0 = \{47677, 47602, 47678, 47905, 47909, 47906\} \rightarrow Z_1 = \{476**, 4790*\}$
   - $A_0 = \{29, 22, 27, 43, 52, 47\} \rightarrow A_1 = \{2*, [43, 52]\}$
   - $S_0 = \{M, F\} \rightarrow S_1 = \{*\}$

   Please apply (Z_1, A_1, S_1) to generalize the original table, and discuss what is the value of k in your generalized table? In your generalized table, if we apply the definition of distinct l-diversity (a table is l-diverse if each of its QI groups contains at least l "well-represented" values for the SA), what is the value of l? If we apply the definition of entropy l-diversity (Entropy l-diversity: for each QI group g, $entropy(g) \geq \log(l)$), what is the value of l?

QI			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

My answer:

The Generalized-table:

QI		
Zipcode	Age	Sex
476**	2*	*
476**	2*	*
476**	2*	*
4790*	[43, 52]	*
4790*	[43, 52]	*
4790*	[43, 52]	*

In the generalized table, the value of k is 2 for that each combination of quasi-identifiers (Z1, A1, S1) appears at least twice in the table.

To determine the value of l for distinct l-diversity, we need to check each QI group in the generalized table. In this case, we have one QI group, which is (Z1, A1, S1). Since each combination appears at least twice, the value of l for distinct l-diversity is 2.

To determine the value of l for entropy l-diversity, we calculate the entropy for each QI group. In this case, we have one QI group, which is (Z1, A1, S1). The entropy of this group is calculated as:

entropy(g) = log2(number of distinct values in g)

For (Z1, A1, S1), the distinct values are: Z1: 476*, 4790* A1: 2*, [43, 52] S1: *

The number of distinct values in (Z1, A1, S1) is 4. Therefore, the value of l for entropy l-diversity is 4.

Summarization:

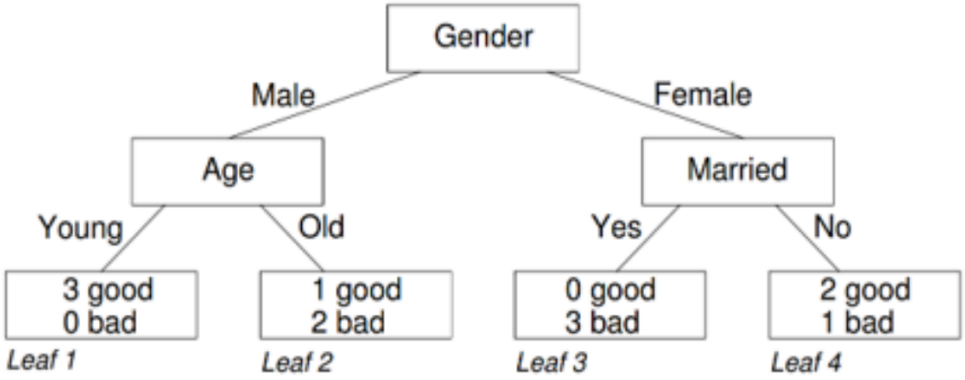
- The value of k in the generalized table is 2.
- The value of l for distinct l-diversity is 2.
- The value of l for entropy l-diversity is 4.

3 Question 6

6. Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and

machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Consider the following dataset, which describes the loan risk information of a mortgage company with the following attributes: Gender, Married, Age, Sports Car, and Loan Risk. Now, the mortgage company wants to release the table. Taking the privacy into account (balancing the privacy and utility), can you help the company to generate a k-anonymous version of the table to fit a better utility of the mortgage company decision tree, i.e., the decision tree can be correctly generated from the released k-anonymous table. Further, what is the value of k? If considering l-diversity, what is the value of l?

Mortgage company decision tree



Mortgage company data

Name	Gender	Married	Age	Sports Car	Loan Risk
Anthony	Male	Yes	Young	Yes	good
Brian	Male	Yes	Young	No	good
Charles	Male	Yes	Young	Yes	good
David	Male	Yes	Old	Yes	good
Edward	Male	Yes	Old	Yes	bad
Frank	Male	No	Old	Yes	bad
Alice	Female	No	Young	No	good
Barbara	Female	No	Old	Yes	good
Carol	Female	No	Young	No	bad
Donna	Female	Yes	Young	No	bad
Emily	Female	Yes	Young	Yes	bad
Fiona	Female	Yes	Young	Yes	bad

My answer:

To generate a k-anonymous version of the table while balancing privacy and utility, we need to ensure that each combination of quasi-identifiers (attributes that can potentially identify individuals) appears at least k times in the released table. In this case, the quasi-identifiers are Gender, Married, Age, and Sports Car.

To determine the value of k, we need to consider the sensitivity of the quasi-identifiers and the desired level of privacy. A higher value of k provides stronger privacy protection but may result in a loss of utility. Generally, a common approach is to start with a value of k=2 and then increase it if necessary to achieve the desired privacy level.

Regarding l-diversity, it refers to the requirement that each group of records with the same quasi-identifiers must have at least l distinct values for sensitive attributes. The value of l depends on the sensitivity of the sensitive attribute and the desired level of diversity.

Above all, we set the k=3 & l=2 for this model. The concrete decision tree label is:

Gender	Married	Age	Sports Car	Loan Risk
Male	*	Y	*	GOOD
Male	*	Y	*	GOOD
Male	*	Y	*	GOOD
Male	*	O	*	GOOD
Male	*	O	*	BAD
Male	*	O	*	BAD
Female	Yes	*	*	GOOD
Female	Yes	*	*	GOOD
Female	Yes	*	*	BAD
Female	No	*	*	BAD
Female	No	*	*	BAD
Female	No	*	*	BAD

4 Question 7

7. Assume a data owner $\mathcal{A}$ has a table D below, showing the one-day electricity uses of all users in one residential area. From the table, we can observe that the electricity use of each user ranges from 0 to 5. Now, in the interactive database query (IDQ) model, a client wants to launch a statistical function query

"$Sum(D)$", i.e., "What is the total electricity consumption in the residential area at that day?" Please try your best to answer the following questions.

- What is the sensitivity of the function $S(F)$ of $Sum()$ in this table D?
- Consider we set the privacy level as ϵ, how to randomly choose a random noise from a Lapacian distribution so that we can achieve ϵ-Differential Privacy in the statistical function query "$Sum(D)$". Please follow the lecture note to prove your result, i.e.,

$$\Pr[A(D+I) \in T] \leq e^{\epsilon} \Pr[A(D-I) \in T]$$

where $A(*) = Sum(*) + Lapacian\ noise.$

ID#	Name	Electricity Use (0-5/day)
1	Alice	4
2	Bob	2
3	Carlo	1
4	David	4
5	Elvas	5
6	Ford	0
7	Geoge	2
8	Hilton	4
9	Ives	2
10	Jack	5

My answer:

(a) To determine the sensitivity of the function S(F) of Sum() in this table D, we need to consider the maximum possible change in the output of the function when a single record in the table is modified. In this case, the function Sum(D) calculates the total electricity consumption in the residential area for a given day. The sensitivity of this function can be determined by finding the maximum difference in the output when a single record is changed.

Let's consider two scenarios:

1. The maximum electricity use in the table is 5, and we change the electricity use of a user from 5 to 0. In this case, the maximum change in the output would be 5 (from subtracting 5 from the sum).
2. The minimum electricity use in the table is 0, and we change the electricity use of a user from 0 to 5. In this case, the maximum change in the output would be 5 (from adding 5 to the sum).

Therefore, the sensitivity of the function S(F) of Sum() in this table D is 5.

(b) By reading some relevant book, my answer are as follows.

Consider achieving ε-differential privacy in the statistical function query "Sum(D)" by adding random noise from a Laplacian distribution.

To achieve ε-differential privacy, we need to add random noise from a Laplacian distribution with a scale parameter of Δ/ϵ, where Δ is the sensitivity of the function. In this case, the sensitivity Δ is 5, and let's assume we want to set the privacy level ϵ.

To randomly choose a random noise from a Laplacian distribution, we can use the following steps:

1. Generate a random number r from a uniform distribution between 0 and 1.
2. Calculate the noise value N as $N = -\Delta/\epsilon \times sign(r - 0.5) \times \ln^{1-2|r-0.5|}$.

By adding this noise N to the result of the function Sum(D), we can achieve ε-differential privacy in the statistical function query "Sum(D)".

Pr[A(D + I) ∈ T] ≤ e^ε * Pr[A(D - I) ∈ T], is a general result for ε-differential privacy, and it holds for the Laplacian noise mechanism as well.