Laboratory on Privacy
Total : 100 points

This lab is to be completed in pairs of two students. You will need to hand in one report per group.

Names: _____

_____

Date: _____

Objective: In this lab, you will gain understanding about multiple types of anonymization techniques and their tradeoffs. Additionally, you will learn to anonymize a dataset using k-anonymity and l-diversity.

Option: To perform this lab you will need to have a 32bit Windows and java run time environment (JRE) 1.5 or higher. Additionally, you will need the software provided in the privacy-lab.zip that contains the UTD Anoymization toolbox and the datasets that you are going to use.

Note: The output of anonymization toolbox is provided. You can just use the anonymized files to complete the lab.

Preliminary questions:

1. In your own words, define what a *quasi-identifier* is

   _____
   _____
   _____

2. Explain the difference between a *quasi-identifier* and an *identifier field*

   _____
   _____
   _____

3. What is an equivalence class?

   _____
   _____
   _____

4. Provide a brief overview of *k-anonymity*

   _____

5. Provide a brief overview of *L-diversity*

6. Explain the difference between *L-diversity* and *K-anonymity*

Hands-on questions:

Part A: Naïve anonymization

1. Unzip the privacy-lab.zip file in your computer. Go to folder PartA and open the file **naive.txt.** For your convenience, each column is separated using tabs. This file contains a naively anonymized dataset. This means that the name, last name and address of people in the dataset have been removed. If you know that Alice Smith is divorced and works in *Business_and_repair_services*, are you able to identify her record? If so, please write it below.

   According to the previous question, do you conclude this method of anonymization guarantees any privacy?

Part B: K-anonymity and l-diversity datasets

In this part of the lab, you will learn to anonymize datasets using different anonymization techniques. You will use the UTD anonymization toolbox provided in the zip file.

First, you will compare the effect of using different values for parameter in the anonymized data.

1. Open the file **header.txt** stored in the folder **/dataset.** This file contains the list of names of the fields of the dataset that you are going to analyze. List the first five fields of this file:

   Open the file **original-data.cvs** this is the data file, each column in the data file corresponds to the fields that were described in **header.txt.** For the first record on this file, show the mapping between the header name and the data. E.g., provide the name of the column and the value of

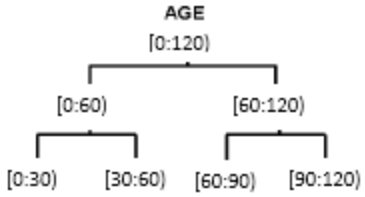
the first record.

---

2. Open the file **config.xml.** This file contains the specifications that will be used to anonymize the dataset specified in the tag <*input filename*>. Below, there is an example of such specification. In the example, **k=1,** the identifier is <*id*> **"Zipcode"** and the quasi-identifiers <*qid*> used are age and sex**.** The *tag* method specifies the type of anonymization technique use and the tag <*ouput filename*> is used to specify the output name of the anonymized data generated by the toolkit. For your convenience, these tags are underlined and shown in blue.

```
<?xml version="1.0"?>
<config method = 'Datafly' k = '6'>
    <input filename='dataset/original-data.cvs' separator=','/> <!-- If left blank, separator will
     be set as comma by default.-->
    <output filename='anonimized-k=6.txt' format ='genVals'/> <!-- Format
    options = {genVals, genValsDist, anatomy}.
        If left blank, output format will be set as genVals by default.-->
    <id> <!-- List of identifier attributes, if any, these will be excluded from the output -->
        <att index='0' name='Zipcode'/>
    </id>
    <qid>
        <att index='7' name='sex'>
            <map> <!-- Mapping of a categorical domain to a discrete-valued numeric domain.-->
                <entry cat='Female' int='0' />
                <entry cat='Male' int='1' />
            </map>
            <vgh value='[0:1]'>
            </vgh>
        </att>
        <att index='1' name ='age'>
            <vgh value='[0:120)'>
                <node value='[0:60)'>
                    <node value='[0:30)'/>
                    <node value='[30:60)'/>
                </node>
                <node value='[60:120)'>
                    <node value='[60:90)'/>
                    <node value='[90:120)'/>
                </node>
            </vgh>
        </att>
    </qid>
</config>
```

The age in the previous xml is generalized using the tree



3. From the provided xml, infer the generalization tree used for ***age*** and draw it in the flowing space.

4. Analyze the anonymization case "k=1"
   The result is provided in the output folder, see the **anonymized-k=1.txt** file.
   Option:
   If you want to generate the anonymized file by yourself, please follow the instructions:
   cd to the appropriate directory and type the following in the command line
   java -cp sqlite.jar;anon_toolbox.jar;. -Djava.library.path=. anonymizer.Anonymizer
   you will anonymize the ***original-data.cvs*** *using the configuration provided in* ***config.xml****. After running the command, the* ***anonymized-k=1.txt*** *file is created in the directory.*

   Open the file. Are you able to see the zipcode of the users in the dataset?
   _____Why? _____
5. List the first three ages of the anonymized data
   _____
6. Compare the original dataset with this anonymized dataset. Was the age modified substantially?
   _____
7. By exploring the anonymized data, given the configuration provided in **config.xml,** why was the age column modified as it was?
   _____

8. Plot the histogram of field age of the anonymized data with 4 bins ([0:25), [25:50), [50:75), and [75:100)).

Tests for K=4

9. Now, you need to repeat the experiment above, but changing the configuration file **config.xml** so that k=4, and then the input file is the output filename to **anonymized-k=4.txt.**

   The result file is also provided in the output folder.

   Open the file that was created by this action. Is the age column different from the previous anonymized data? Why?

10. List the first three ages of the anonymized output

11. Now you need to find the equivalence classes in the anonymized data. List them and clearly separate them in the following space

12. By looking uniquely the anonymized dataset (without considering the original dataset) are you able to locate the record of Anne who is 45 years old and is a widow?

13. If yes, list Anne's anonymized record below.

14. If no, why you could not identify her record?

15. By looking uniquely the anonymized dataset (without considering the original dataset) are you able to locate the record of Samantha who was born in Cuban?

16. If yes, list Samantha's anonymized record below.

17. If no, why you could not identify her record?

18. What can you conclude about the answers of questions 12 to 17?

Tests for K=6

19. Now, you need to change the configuration file **config.xml** so that k=6 and the output filename to **anonymized-k=6.txt**.

    The result file is also provided in the output folder.

    Open the file that was created by this action. Is the age column different from the previous anonymized data?

20. List the first three ages of the anonymized data

21. Now you need to find the equivalence classes in the anonymized data. List them and clearly separate them in the following space

22. Are the equivalence classes different from the ones you found in point 11? _____
    Why _____
    _____
23. Inspect uniquely the anonymized dataset (without considering the original dataset) are you able to locate the record of Anne who is 45 years old and is a widow? _____
    If yes, list the record you identified as Anne's record. If not, explain why.
    _____
24. By looking uniquely the anonymized dataset (without considering the original dataset) are you able to locate the record of Alice who is 30 years old and divorced? _____
25. If yes, list Alice's anonymized record below. If not, explain why.
    _____
    _____
26. Looking uniquely the anonymized dataset (without considering the original dataset), are you able to locate the record of Pamela who is 30 years old, single? _____
27. If yes, list her anonymized record below. If not, explain why.
    _____
    _____
28. You know that Jessica has never worked and she is 24. Looking uniquely the anonymized dataset (without considering the original dataset), are you able to locate her record ?
    _____
29. If yes, list her anonymized record below. If not, explain why.
    _____
    _____
30. In this part of the lab, you have anonymized multiple datasets using k-anonymity. According to your results, what value of k was less secure?_____and what value of *k* seems to provide more protection? _____
31. Now find the histogram of the age with two bins ([0:50) and [50:100)). Draw it below.

32. Compare the histograms for k=1, k=4 and k=6. What k value produces the closest result to the original dataset?_____Why? _____
33. Which one the worst?_____Why? _____

Comparing L-diversity with k-anonymity

In the following, you will compare k-anonymity with l-diversity.

1. Does L-diversity consider the sensitivity of the parameters to be anonymized? _____ What for? _____
2. Take the output for k=6 that you previously found. Does the output respect l-diversity if the sensitive attribute is the field place where the person was born and l=2?_____If it does not fulfill L-diversity. What records violate this property? List them below.
3. If you know George mother is from Mexico and his father is form the US. Are you able to identify the record of George in the k=6 output in point 16?_____if so, list his record
4. Would l-diversity help to avoid this type of inference? _____
5. Describe how you could avoid this problem_____
6. Anonymize the dataset manually to ensure that l-diversity is fulfilled with k=4, l=2 considering the *country_of_birth_self* as a sensitive attribute. Show the equivalence classes of your solution below

Part C: Analyzing datasets

1. Go to the folder **/dataset**. In a text editor open the file **header.txt**. This file contains the list of the names of the fields of the dataset that you are going to analyze. List the fields that you think correspond to *identifiers* in the dataset and state why you think they are *identifiers*

   _____

   _____

   _____

2. Open the file **/dataset/ anonymized-part3.txt** List the first two rows of the dataset. Were any fields used during anonymization as identifiers? If yes, which?

   _____

   _____

3. Why were you able to infer if any fields were used as identifiers during the anonymization procedure?

   _____

   _____

   _____

4. The age and sex were used as a quasi-identifier in this anonymization. Their generalization trees are shown in the xml presented at the beginning of this document. By exploring the anonymized dataset, what was the *k* specified?

   _____

5. Show the groups of equivalence classes in the following space.