School of Computing

COMP3300 Assignment Project

Assignment marks: 40% of overall unit marks. (Marked out of 100.)

Objective: To gain experience in evaluating a dataset for privacy and utility so that the relevant privacy law is respected; select a privacy preserving technique to protect identified vulnerabilities.

Please note: This assignment specification aims to provide as complete a description of this assessment task as possible. However, as with any specification, there will always be things we should have said that we have left out and areas in which we could have done a better job of explanation. As a result, you are strongly encouraged to ask any questions of clarification you might have, either by raising them during a lecture or by posting them on the iLearn discussion forum devoted to this assignment.

Financial Data

Many financial organisations keep records of every individual's transactions using their bank account, both for customer account-keeping and for the bank's records.

Such a collected dataset typically contains information such as that shown in the table below.

Transaction id	Amount	Date and Time	Vendor	Category	Place
24152674	$356	5 June, 22:35	Physio	Health	Diagon Alley
79216552	$2089	7 April, 11:49	Surgery	Health	Godrics Hollow
86229173	$126	4 August, 7:51	Divorce	Legal	Godrics Hollow
95399264	$598	19th May, 15:24	Divorce	Legal	Hogsmeade

This table records the amount spent, the date it was spent, the vendor, the category of spending and the location at which the amount was spent.

As we have learned in the unit so far, although this has had personal identifying information removed, it is highly revealing because knowing the time, place and category of a purchase is likely enough to identify many individuals' rows — here the *sensitive information is the vendor*, because that identifies whether a person had surgery, or legal problems of some kind.

However there is useful information that can be shared from this dataset to provide information for organisations wishing to improve public and retail services. For example, it is useful to be able to compute a breakdown of total consumer spending in terms of the various vendors within a time interval; or to compute summary statistics such as the number of times each vendor is visited per location. Bear in mind though that these are only two examples of the kind of useful information that could be extracted. What is always a challenge is to try and balance the diversity of information that can be shared, versus the protection afforded for individuals.

For a *privacy professional*, the challenge is to find a way to share as much of *this type* of useful information without infringing the prevailing privacy laws, which for the purposes of this assignment are:

"*No individual can be re-identified in any publicly-released dataset. Re-identified means that a record can be linked with very high likelihood to an individual using information which could be plausibly obtained*."

In this assignment you will be asked to apply a number of techniques you have learned about in this unit, and to evaluate their suitability for the task. You will also be asked to explore the privacy-utility tradeoff, namely whether the proposed techniques provide sufficient privacy in order for the law to be respected whilst providing a reasonable level of utility.

Splitting the Raw Data

As mentioned above the raw table above is very revealing since many, if not most of the transactions can be quasi-identifiable if for example, the amount, date, time and place of a purchase becomes known. These details could very well become public knowledge and therefore there risk of re-identification appears to be high.

One technique to decrease this vulnerability is to split the raw data into several smaller datasets which collectively contain the same data as the original, but which separately are protected through other privacy defences. Although k-anonymity is known to have inherent privacy vulnerabilities, it is still used today. In this assignment you'll explore why anonymisation must be strengthened for this kind of data.

You are first asked to demonstrate the vulnerability of splitting the data using anonymisation techniques, and then you'll be asked to provide alternative solutions, and to produce some evidence to demonstrate how effective they are.

Task 1: Reconstructing split data

A major vulnerability in methods to disguise the raw data by, for example, splitting it into several components, is the possibility for *reconstructing the original dataset*. If this can be done either fully or partially, all the original re-identification vulnerabilities are actually *still present*.

You are provided with 2 datasets generated from some raw transaction data, which have had some generalisation and suppression applied to some of the attributes. One dataset (ADV.csv) contains only the amount spent (rounded to the nearest $10), the date (with the time field suppressed) and the vendor. The second dataset (VCP.csv) contains the category, place and vendor.

The combined effect of the generalisation, suppression and splitting is that both datasets are now been k-anonymised (for different values of k>1). They will look something like the following (although not exactly these values):

Dataset ADV.csv

Amount	Date	Vendor
$360	5 June	Physio
$2080	7 April	Surgery

Amount	Date	Vendor
$120	4 August	Divorce
$590	19th May	Divorce

Dataset VCP.csv

Vendor	Category	Place
Divorce	Legal	Godrics Hollow
Divorce	Legal	Hogsmeade
Surgery	Health	Godrics Hollow
Physio	Health	Diagon Alley

Notice that the two datasets together contain the same information as the original raw transaction dataset (subject to the generalisation used). Notice though that although both tables still share the original Vendor information the rows no longer correspond. For example row 1 in dataset ADV.csv does not match row 1 in dataset VCP.csv, since the Vendors do not correspond. However we can see that in spite of these precautions, some of the rows in the original raw transaction dataset can be partially reconstructed by linking then together via Vendor. For example, Row 1 in dataset ADV.csv corresponds to row 4 in dataset VCP.csv.

In this task you are given information about **3 individuals** and you need to see if you can identify each of them by using a join on the datasets ADV.csv and VCP.csv. Do be aware that the join on corresponding rows will not work because, once separated, the rows in each of ADV.csv and VCP.csv have been independently permuted.

In this assignment we refer to ADV.csv as table1.csv, and VCP.csv as table2.csv.

Task 1 - Basic Attacks

For the following tasks you are given two datasets called table1.csv and table2.csv which you can download from iLearn. These two datasets represent the original dataset split into two pieces as described above. Notice that each is k-anonymous for some value of k. table1.csv is 3-anonymous wrt attributes Amount and Date, and table2.csv is at least 3-anonymous wrt Category and Place.

1. In the iLearn quiz for this assignment you are given descriptions of 3 individuals' transactions. (Please see special instructions below for submitting answers in the iLearn Quiz to growth this assignment.) You can find descriptions of your three characters by opening the quiz for this assignment. For each description you must use an appropriate attack to discover their sensitive information, i.e. Vendor. Put your answers in the iLearn quiz for this assignment. **[15 marks]**

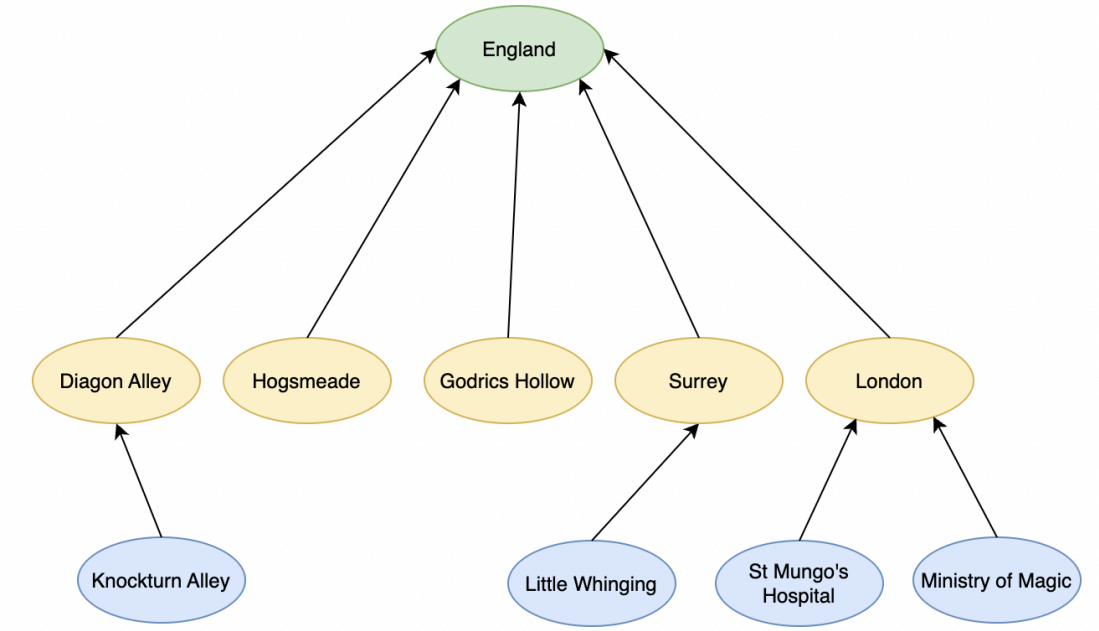
2. Describe the reconstruction attack. Put your answer in iLearn. **[5 marks]**

3. Using your answers above, what is the maximum value that k could be in the reconstructed (original) dataset wrt. the attributes Category, Place, Date and Amount? You must explain your answer. **[5 marks]**

4. Identify the vulnerability in the two datasets that enable the partial reconstruction of the data. **[5 marks].**

Task 2 — How does generalisation work?

You are provided with the following generalisation hierarchy for Place.



The top of the hierarchy (England) is the most general and the bottom (in blue) is the least general.

5. Using this generalisation hierarchy, generalise all of the places in the table2.csv to their corresponding Yellow place names (eg. Little Whinging will generalise to Surrey, but Hogsmeade will remain as is). Now repeat the attack from Q1 and specify which individuals are still identifiable. **[15 marks]**

6. You have been instructed to apply some generalisation to increase privacy for your individuals. In iLearn for this question, you are given a list of generalisations to pick from consistent with the above hierarchy. Choose the most most informative generalisation (="lowest in security classification") on Place from the list you are given so that it increases privacy for all your three individuals. If it is not possible to increase privacy, choose the most informative generalisation. **[5 marks]**

7. Using the utility: "Count of the number of times each place is involved in a transaction", describe how your answer above (Q6) affects this utility measure. **[5 marks]**

Now upload the two original datasets (table1.csv and table2.csv, not the generalised version) to a Jupyter notebook to build 2 channels, one for each dataset. Make sure that the columns are labelled by the non-sensitive attributes and the rows are labelled by the row number and the sensitive attribute (in this case Vendor).

8. Compute Bayes' vulnerability for each of the two channels and put your answer in iLearn. **[5 marks]**

Task 3 - Bounded Noise and Differential Privacy

Next you will use some alternative privacy-preserving method and compare their effectiveness on the datasets.

9. Apply the bounded noise with B = 20 to the *Amount* attribute in the channel you constructed using table1.csv. Compute the Bayes'vulnerability for the resulting channel. Put your answer in iLearn. **[5 marks]**

10. What is the *smallest* value of the bound B that you need to choose in order to reduce the Bayes' vulnerability to less than 0.35? **[5 marks]**

11. Assume that the goal of the data release is to answer: "Compute the average amount spent at each Vendor". How does increasing the bound in Question 10 affect the utility of the data release compared with the bound specified in Question 9? How could you reduce the Bayes' vulnerability to 0.35 without affecting the utility? **[5 marks]**

12. A synthetic dataset is produced from a dataset containing the place Vendor Stadium using bounded noise with B = 20. The average amount spent at Vendor Stadium was found to be $27. What is the maximum and minimum value that the true answer could be? Put your answer into iLearn. **[5 marks]**

13. Repeat Question 9 using Geometric noise, with differential privacy setting eps=log 2, with 100 rows and step size 10, and add your answers to iLearn. **[10 marks]**

Recall that Geometric noise is defined by three parameters: privacy (eps), number of rows (r) and step size (s). The privacy parameter is related to the indistinguishability properties, the step size means that the possible outputs are 0, s, 2s, 3s, etc. Finally the rows determine the number of outputs; in fact the upper bound on outputs is r x k.

14. What would you recommend: Bounded noise or Geometric noise for balancing privacy and utility? What else might you need to consider? Give reasons for your answer. (Write 3 or 4 sentences). **[10 marks]**

Instructions for obtaining materials and submitting your solutions

1. Details of your individuals (used as clues) are found in the iLearn specification for this assignment.

2. Your two csv files that you should use are in a zipped file to download.

3. The Jupyter notebook containing functions to help you compute some of the answers is available to download.

4. Please add your answers to the iLearn quiz which is found in the submission section.

5. Please note that **you have exactly one chance to complete the quiz**, so please only submit your answers when you have completed all of the sections you wish to attempt.

Mark allocation for this assignment

This assignment is structured to allow you to decide how much effort you want to expend for the return in marks that you might hope for. You can choose whether to implement only the P level functionality, or to put in more effort and complete the full problem. So you can decide upfront whether you are shooting for a pass or a high distinction and know exactly how much work will be required to obtain that mark.

Here is what is roughly what is required to obtain marks in one of the performance bands for this assignment:

- **Pass:** A successful implementation of Tasks 1&2 (Questions 1—8). (Marks up to 60/100.)
- **Credit:** The P level plus a successful implementation of Task 3 Questions 9&10. (Marks up to 70/100.)
- **Distinction** The CR level plus Task 3 Questions 11&12. (Marks up to 80/100.)
- **High Distinction:** The D level, plus Task 3 Questions 13&14. (Marks up to 100/100.)