# Project Title: Initial Plan- Implementation of a data privacy protection method for transaction data

*Author*: Dominyque Mohammed (c1839296)

*Supervisor and reviewer*: Jianhua Shao

*Module*: CM2303

*Module Title*: One Semester Individual Project

*Credits*: 40

# Project Description

Everyday data is collected by companies, governments and individuals for the main purpose of analysing the information so that informed decisions can be made from it in order to benefit their customers and the public. This information can be termed as transaction data. However, the data collected will contain sensitive and private information such as race, health issues, religious beliefs, sexual orientation and mental health status to name a few. Consequently, publishing this raw data violates the privacy of those individuals as attackers can link attributes from different tables to reveal the identities of those people. With this, comes the difficult task of balancing maintaining the privacy of individuals and ensuring that the published data is not distorted in such a manner that it is unusable for the intended data recipients.

Methods and tools have been developed to assist in this aim of publishing data that is useful while individual privacy is preserved and it is known as 'privacy-preserving data publishing'. One major part of PPDP is the anonymization approach which seeks to hide the sensitive data of record owners, i.e., guaranteeing privacy protection meaning that attackers should not be able to learn anything extra about any target victim given the published data. To achieve this privacy models and anonymization algorithms have been created to aid in anonymising the data in a way that suits the specific situation. For example, there are algorithms that work best for large data sets etc.

For this project, one of the existing algorithms for anonymising transaction data will be implemented in Python. This project will then be a collaborative effort with a fellow student's project, who would be implementing a different anonymising algorithm, to make an educational tool. This tool would have an interface where users can input the raw data they want to anonymize, and they can then pick which anonymising algorithm they want to run the data through, therefore allowing them to see how different approaches work and presents lecturers/ teachers a tool which gives them a visual aid in explaining how anonymising works and the importance of doing so, i.e., the risks that can arise from having raw data made public.

With data collections rapidly growing creating massive amounts of opportunities for knowledge, it is important to ensure that whilst we become smarter about using and mining the data, creating important statistics etc, that we do not forget the significance of maintaining

the privacy of individuals for safety reasons and to prevent cases of fraud, identity theft and losing trust.

## Project Aims and Objectives

- To implement an existing anonymization algorithm to protect transaction data
  - To produce the correct anonymized data from the raw transaction data being inputted.
  - To detect whether or not an attacker can link any of the anonymized data together to uncover the identity of the individuals, violating their privacy.
  - To ensure that the anonymised data is still as useful for data mining and obtaining important statistics as the raw data was.
- Highlighting the downfalls of the chosen algorithm
  - Discuss the attacks that are still possible with the anonymization algorithm chosen.
  - Discuss what more needs to be done in this field to achieve the aim of having privacy alongside useful data.
- Creating an educational tool
  - Creating an interface to allow for the easy usage of the different algorithms
  - Ensure the results presented are clear, readable and easy to understand
- Conduct research
  - Conduct background research on the other anonymization algorithms that exist so that a comparison can be done as well as for the purpose of highlighting the other approaches in this field in the report, giving the user a more well rounded view of this field.
  - Discuss and identify the risks that can arise from not anonymizing data

# Work Plan

| Week | Work to be done |
|---|---|
| **February:** | |
| **1**: Mon 31st- Fri 4th | - Initial plan to be done<br>- First meeting with supervisor to hash out all the details (frequency of meetings, implementation specifics etc) |
| **2**: Mon 7th – Fri 11th | - Submission of Initial plan on Monday<br>- Background research to be done on anonymization algorithms<br>• Introduction of the final report should be done<br>- Meeting with supervisor |
| **3**: Mon 14th – Fri 18th | - Continued background research on approaches and anonymization algorithms<br>- Planning out the software development design (the raw data that will be needed, flowcharts for grasping how the algorithm should flow)<br>• The above two points will then be written up in the Background and Approach section for the final report.<br>- Meeting with supervisor |
| **4**: Mon 21st – Fri 25th | - Start on programming of the chosen algorithm: getting the basic structure done, i.e. reading in the input data and storing in a table (python)<br>- Meeting with supervisor |
| **March:** | |
| **5**: Mon 28th -Fri 4th | - Continued programming- adding the anonymisation code and calculations<br>- Testing as programming is done<br>- Adding notes to the Implementation section of the report<br>- Meeting with supervisor |
| **6**: Mon 7th – Fri 11th | - Continued programming<br>- Review of code by supervisor<br>• Should have majority of the anonymization algorithm done by the end of this week |
| **7**: Mon 14th – Fri 18th | - With the main part of the coding done, the coding of the presentation of the results of the algorithm is the focus.<br>- Meeting with the supervisor |
| **8**: Mon 21st – Fri 25th | - Collaborating with the other student for the creation of the interface<br>- Meeting with supervisor |

| | |
|---|---|
| | ❖ Note: This is a busy week, there is the Hackathon for my Emerging Tech module this coming weekend. |
| **9**: Mon 28th – Fri 1st | - Finish the coding of the presentation of the results<br>- Go through the code looking for any anomalies.<br>- Meeting with supervisor<br>• Code should be completed here |
| **April:** | |
| *Mon 4th – Fri 22nd | EASTER<br>*This period will be used as a time buffer. No work will be scheduled but allows for catching up on work if fallen behind on schedule. |
| **10**: Mon 25th – Fri 29th | - Writing up the results and evaluation of the report using program screenshots and referring to the output data to explain what was done and what could be improved.<br>- Meeting with supervisor- asking for review for everything done so far.<br>❖ Note: Busy week, coursework deadline coming up for Emerging Tech module. |
| **May:** | |
| **11**: Mon 2nd – Fri 6th | - Write up the rest of the report: Future Work, Conclusions, Reflection and the support sections.<br>- Meeting with supervisor<br>• The write up of the report should be completed here |
| **12**: Mon 9th – Fri 13th | - Proof reading of the report from top to bottom<br>- Ensuring references have been added in<br>- Submit final report on Thursday (due Friday) |