Implementation of a Data Privacy Protection Tool for Transactional Data

Initial Plan

Author: Daniel Hendry
Supervisor: Dr Jianhua Shao
Moderator: Dr Víctor Gutiérrez Basulto

Module Code: CM3203
Module Title: One Semester Individual Project
Credits: 40

Project Description

As an increased amount of data being gathered and stored, how to protect the private information contained within such data sets becomes an important issue. One of the more recent approaches to addressing this issue is called k-anonymisation, which attempts to make any record in a data set identical to at least $k-1$ other records (hence no individual could be identified)[1].

It has been shown that k-anonymisation does not scale well to high dimensional data, (e.g transaction data) which also has variable length unlike relational data, and that in order to achieve the same level of k-anonymity, you would either have to suppress most of the data or lose the desired level of anonymity[2].

An approach to ensuring data privacy, with minimal information loss, with high dimensional data is by disassociation. This approach preserves the original data, but hides the identifying combinations. Due to k-anonymisation not being suited to high dimensional data, this method ensures k^m-anonymisation. This is defined as someone who has partial record knowledge of a record, up to m terms, will not be able to distinguish any record from other $k-1$ records[3].

This project aims to implement a software tool based on an existing k^m-anonymisation algorithm[4], to help anonymise high dimensional data, which will be transaction data. The implementation will be tested using data from [4] and open-sourced data to show the performance of the algorithm with various data sets and level of information loss.

It would be desirable to complete some association data mining on the anonymised data sets if time allows, however the most important aim of this project is to implement a privacy algorithm for high dimensional data.

Ethics

After discussion with my supervisor, I have concluded that I do not need to consider the ethics of the data I use to test the algorithm. This is because the data used in the paper [4] was created for that purpose and does not come from real world data and the further testing data is openly sourced and not sensitive information.

Project Aims and Objectives

1. **Implement algorithm for privacy protection for transaction data**
   (a) Implementation will be done in Java as object-oriented models are useful for the three stages of the algorithm and it's my most confident programming language.
   (b) Must be able to take a transaction dataset as input and output the anonymised dataset correctly.
   (c) Input can be through command line. Algorithm more important than UI.
2. **Evaluate performance and information loss**
   (a) Aim to achieve the same results as in [4].
   (b) Assess performing algorithm even for large datasets.
   (c) Assess information loss for implementation on various datasets.

Work Plan

I will be having weekly meetings with my supervisor to discuss progress and answer any of my queries. In my work plan I have dedicated weeks 12-15 as simply writing my final report. This represents the Easter recess. Although pseudo-code is provided by [4] for each part of the algorithm. This pseudo-code, especially the "Refining" part of the algorithm is quite vague and cannot be implemented straight away from the pseudo-code. After discussion with my supervisor, I have concluded more time is required to figure out the missing code from this section.

Week 1 - 28th January

- Complete Initial Plan
- Background research into Transactional data and Disassociation
- **Milestone 1: Initial Plan Complete**
- **Deliverable 1: Initial Plan**

Week 2 - 4th February

- Background research into general k-anonmysation and subsequently k^m-anonmysation.
- Background research into algorithm to implement.
  - Include high level understanding of each of the three parts to the algorithm. 1) Horizontal partitioning, 2) Vertical partitioning and 3) Refining.

Week 3 - 11th February

- Begin implementation of "horizontal partitioning" part of the algorithm.

  - Pseudo code provided by [4].

Week 4 - 18th February

- Complete implementation "horizontal partitioning" part of the algorithm.
  - Including unit tests.
- Test using data from paper.

Week 5 - 25th February

- Begin implementation "vertical partitioning" part of the algorithm.
  - Pseudo code provided by [4].

Week 6 - 4th March

- Complete "vertical partitioning" part of the algorithm.
  - Including unit tests.
- Test using data from paper.

Week 7 - 11th March

- Understand "refining" part of the algorithm.
  - Pseudo code is provided by [4], but is very vague and not enough to start the implementation.
- Refine previous code if necessary, as part of the understanding.
  - For example, if my model of the first two parts of the algorithm has to be changed in order to better suit this final part.

Week 8 - 18th March

- Begin implementation of "refining" part of the algorithm.

Week 9 - 25th March

- Finish implementation "refining" part of algorithm.
- **Milestone 2: Algorithm Implemented**
- **Deliverable 2: Completed Algorithm Code**
- **Review Meeting 1: Review implementation**

Week 10 - 1st April

- Test implementation ensures k^m-anonymity.
- Test performance of algorithm
- Refine and fix any bugs where needed.
- Begin testing on information loss.

Week 11 - 8th April

- Complete testing on information loss within the algorithm.
- Evaluate the results.
- **Milestone 3: Testing complete**
- **Review Meeting 2: Review testing results**

Week 12 - 15th April

- Write final report.

Week 13 - 22rd April

- Continue writing final report.

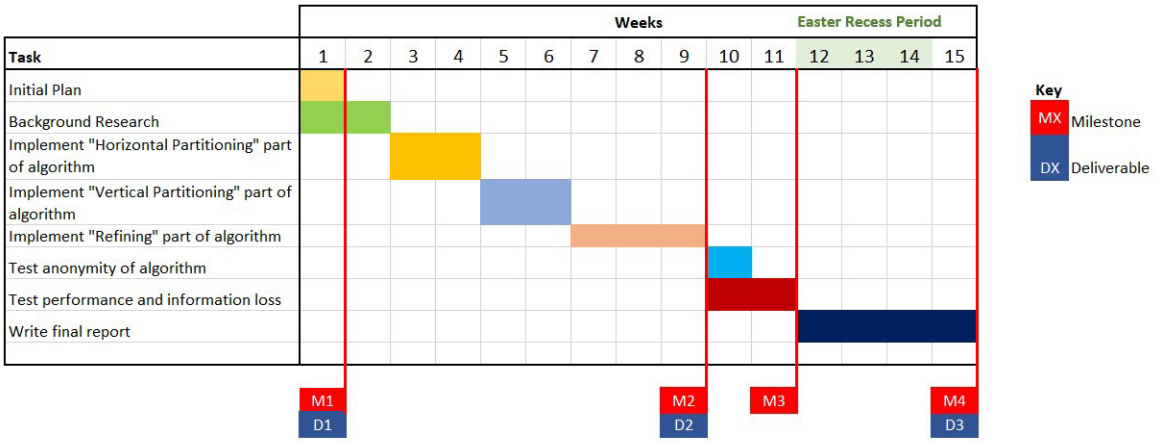
Week 14 - 29th April

- Continue writing final report.

Week 15 - 6th May

- Complete and submit final report.
- **Milestone 4: Final Report Complete**
- **Deliverable 3: Final Report**

Gantt Chart



References

[1] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 557–570, Oct. 2002.

[2] C. Aggarwal, "On k-anonymity and the curse of dimensionality.," *VLDB 2005 - Proceedings of 31st International Conference on Very Large Data Bases*, vol. 2, pp. 901–909, Jan. 2005.

[3] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," *Proc. VLDB Endow.*, vol. 1, pp. 115–125, Aug. 2008.

[4] M. Terrovitis, N. Mamoulis, J. Liagouris, and S. Skiadopoulos, "Privacy preservation by disassociation," *Proc. VLDB Endow.*, vol. 5, pp. 944–955, June 2012.