Initial Plan

Implementation of a data privacy protection method for transaction data

Author: Lim Chun Kuan (C1855672)

Supervisor: JianHua Shao

Moderator: Neetesh Saxena

Module: CM3203 – One Semester Individual Project

Credits: 40

Project Description

The collection of digital information by governments, corporations and individuals has created tremendous opportunities for knowledge-based decision making. Either driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties [1]. The published data in its original form normally contains sensitive information about individuals, and publishing such data means violating individual privacy. A task of the utmost importance is to develop methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved. This undertaking is called privacy-preserving data publishing (PPDP)[1]. In this project, my main task is to implement an anonymization algorithm proposed by Manolis et al. in a research paper [2].

Manolis et al. proposed an anonymization technique termed *disassociation* that focus on protection against identity disclosure in the publication of sparse multidimensional data. It preserves the original terms but hides the fact that two or more different terms appear in the same record. We protect the users' privacy by disassociating record terms that participate in identifying combinations [2].

Manolis et al. suggest that applying k-anonymity [3] on sparse multidimensional data result in huge information loss, hence the anonymity guarantee is opted to k^m-anonymity. k^m-anonymity guarantees an adversary who knows up to m items from any records, will not be able to distinguish any record from other k-1 records [2]. The project aims to implement the algorithm mentioned above into a program using Python. The program should take in a dataset as input and produces a disassociated k^m-anonymous dataset as output. The program will be tested on an extended dataset that appeared in research paper [2].

Ethics

After discussing with my supervisor, we decided that the project does not require ethical approval. A dataset will be provided for testing purpose. The dataset that would be used to test the program does not correspond to any real individuals' information, hence ethical approval is not required.

Project Aims and Objectives

The overall aims of the project is to implement the dissociation algorithm into a python program and it is able to produce the correct k^m-anonymous output.

The main objectives would be the following:

- Understand different part of the algorithm (vertical partitioning, horizontal partitioning and refining)
- implementation of each part of the algorithm using Python
- Program is able to take a dataset as input and produce an appropriate output
- Evaluation of the correctness and performance of the program
- Produce a detailed report based on the program above

Work Plan

There will be a weekly scheduled meeting with my supervisor to discuss the problem I face when working on the project and to discuss my progression. Further meetings will be scheduled if needed. A rough plan for each week is shown below:

Week 1 – 27th January – 2nd February

- Work on initial plan
- Study the research paper [2] the project is based on
- Study on the background of privacy-preserving data publishing method

Deliverables: Initial Plan

Week 2 – 3rd February – 9th February

- Get familiar with each stage of the algorithm and their respective pseudocode
- Improve my Python coding skill necessary for the project

Week 3 – 10th February – 16th February

- Start working on implementation of “horizontal partitioning” stage of the algorithm

Week 4-5 – 17th February – 1st March

- Complete the implementation of ““horizontal partitioning” stage of the algorithm
- Testing using the provided dataset
- Code review meeting with supervisor

Week 6 – 2nd March – 8th March

- Start working on implementation of “vertical partitioning” stage of the algorithm

Week 7-8 – 9th March – 22nd March

- Complete the implementation of “vertical partitioning” stage of the algorithm
- Testing using the provided dataset
- Code review meeting with supervisor

Week 9 – 23rd March – 29th March

- Start working on implementation of “refining” stage of the algorithm

Week 10 – 30th March – 26th April (including Easter break)

- Complete the implementation of “refining” stage of the algorithm
- Testing using the provided dataset
- Perform evaluation on performance and correctness of the program
- Initial draft of the final report
- Information research for final report

Deliverables: Completed Python program code

Week 11-12 – 27th April – 7th May

- Code/Program review and feedback from supervisor
- Code refining if needed
- Final testing with the completed program
- Completion of final report
- Submission of final report

Deliverables: Final Report

Reference

[1] Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, *42*(4), 1-17.

[2] Terrovitis, M., Liagouris, J., Mamoulis, N., & Skiadopoulos, S. (2012). Privacy preservation by disassociation. *arXiv preprint arXiv:1207.0135*.

[3] L. Sweeney. k-anonymity: a model for protecting privacy. IJUFKS, 10(5):557-570, 2002.