Initial Plan

Implementation of a Data Privacy Protection Tool for Relational Data

Author: Benjamin G Lourence
Student No: 1111753
Supervisor: Dr. J Shao
Moderator: Dr. D Tsaneva
Module: CM3203
Date: 01/02/15

Project Description

Modern society continues to collect, store and analyse ever increasing amounts of data from a wide variety of sources. The issues surrounding protection of individuals privacy within information has become more prevalent. Hospital patient data, voting registers and user preferences for online video streaming services are all examples of data that has been released to the public. The reasons for organisations releasing this data can range from regulatory compliance with no further intention to inviting third parties to perform data mining and specialist analysis. However the privacy of the individuals featured within this data is of paramount importance and as such no person should be susceptible to re-identification. This presents the problem of Privacy Preserving Data Publishing (PPDP) in which an organisation wishes to publish data protecting the privacy of individuals contained while maintaining the usefulness of original data.

This project aims to produce a solution that will enable users to anonymise data in a manner that will retain the value of the information contained while ensuring the privacy of individuals. More specifically the idea behind this project is to create a tool that facilitates anonymisation of relational data. The project will implement one (or more depending on time constraints) algorithm(s) for achieving k-anonymity within relational data. Implementation details and listing of the chosen algorithm(s) will be specified in the design stages of the project.

k-Anonymity describes a property of data that forms protection against re-identification via record linkage attacks. A set of records is said to be k-anonymous if information pertaining to an individual cannot be distinguished from at least k-1 individuals also in the dataset, where k is some nominal value. Therefore all explicit identifiers should be removed and care must be taken to obfuscate quasi-identifiers that could be used to infer the identity of an individual. When a dataset is said to be k-anonymous no single individual can be distinguished from other records in the table. There are two methods for achieving k-anonymity in a data set.
Suppression: Entails removing distinguishing information from the dataset, for example replacing the attribute ‘Name’ with an asterisk ‘*’.
Generalisation: Requires specific values to be transformed to a boarder category that includes the original information, for example transforming age specific value 21 to a category 21-30.

Initially the system should operate as a stand-alone piece of software. Although in the future the tool may form one component of a larger project that could be used to model attacks on anonymised data in order to form a greater understanding of the protection mechanisms required. Throughout the design and implementation phases of this solution care will be taken to ensure the tool can be easily integrated into other projects. At the initial stages of discussion Java has been suggested as a suitable language to implement the system.

Project Aims & Objectives:

- Develop a suite of tools that will allow users to perform data-anonymisation algorithms on a set of 'well prepared' data (Other developers will handle the creation of tools to format and load data).

- Create an extensible solution in which future developers can simply 'plug in' their new data anonymisation algorithms and use the existing system infrastructure.

- The tool will aim to support post-processing activities and analysis.
    - Inclusion and comparison of defined metrics.
    - Time profiling relating to algorithm efficiency.
    - Graphing suite to facilitate result visualisations.

- Creation of a basic graphical user interface to support post analysis activities such as graphing of performance metrics.

- Design and implementation will be approached in a modular fashion so that the completed solution can be integrated as a component of future data anonymisation projects.

Initial Background Research

Anonymisation Terminology**:**

Explicit Identifiers: Information that explicitly identifies an individual e.g. National insurance number.

Quasi-Identifiers: An attribute that could potentially identify an individual e.g. Combination of postcode and DOB.

Sensitive Attributes: Data that relates to a specific individual within the record set e.g. Salary.

Non-Sensitive Attributes*:* Encompasses any attribute that does not fit the definition of the three terms above.

Latanya Sweeny reports "87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}." [1]

The current trend for organisations to release information containing these attributes suggests a potential that numerous individuals be could be re-identified from the supposedly 'anonymised data'. These cases demonstrate a strong need for increased privacy protection and the development of tools to implement the protection mechanisms.

There are a number of various techniques used to anonymise data such as l-diversity, t-closeness which are both themselves an extension of k-anonymity. However for this aim of this project focuses on creating and implementing a tool that will utilise k-anonymisation algorithm(s) that protects from record linkage attacks.

"In the attack of record linkage, some value qid on QID [Quasi-identifier] identifies a small number of Records in the released table T, called a group. If the victim's QID matches the value qid, the victim is vulnerable to being linked to the small number of records in the group." [2]

As explained by Benjamin Fung, a record linkage attack attempts to re-identify individuals by grouping quasi-identifiers to extract a finite number of records that could lead to identifying specific individuals. One such example was William Weld former governor of Massachusetts, who was re-identified through a combination of medical data and voter registration using leveraging the quasi-identifier triple previously referenced in the article by Latanya Sweeny. The tool to be designed and created by this project will help protect individual privacy by implementing a k-anonymous algorithm to protect data against such linkage attacks.

Ethics:

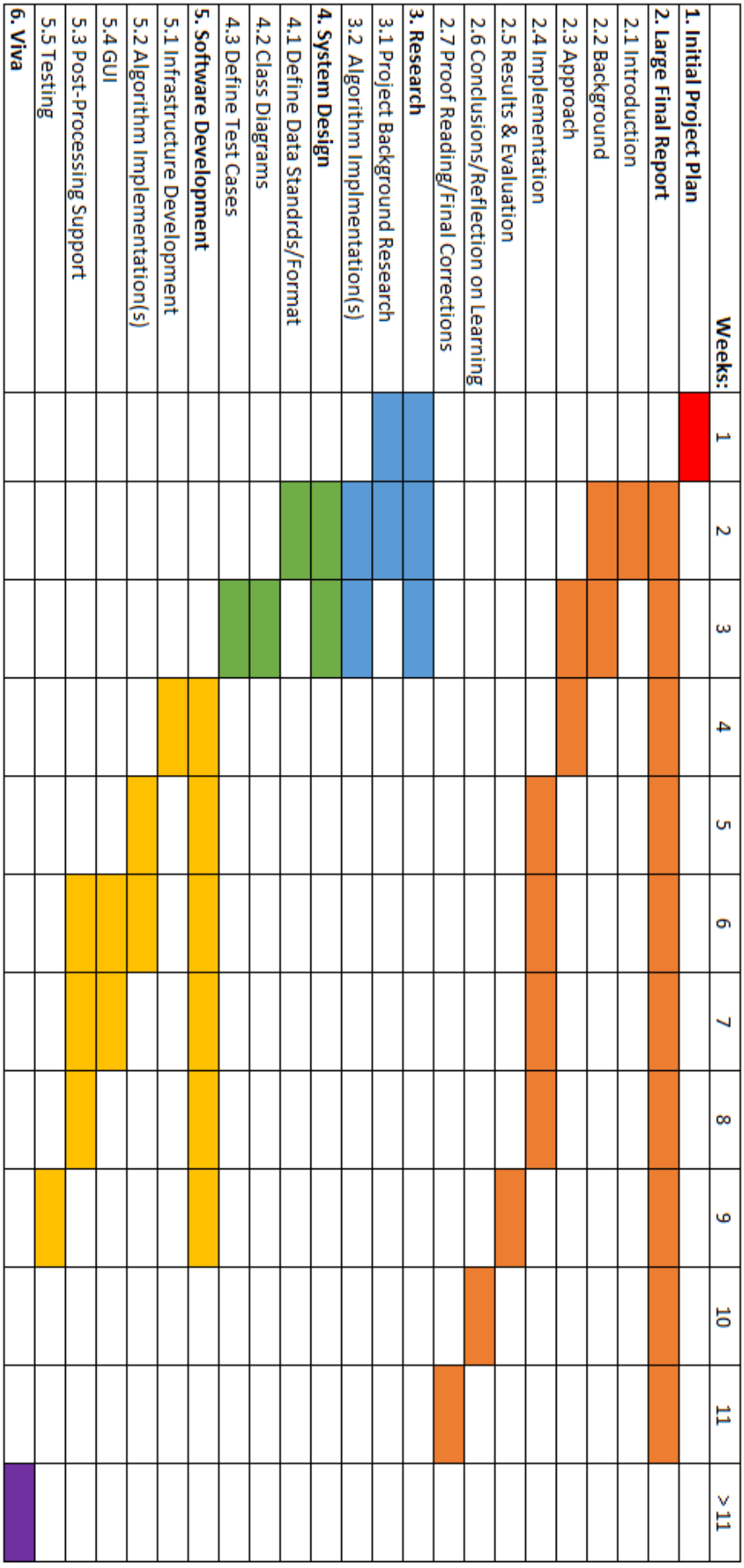
The project aims to produce a system that can anonymise relational data for adequate protection against re-identification. A test set of 'well prepared' data will be assumed while developing the system, work from other developers will focus on the process of transformation, formatting and loading data in an appropriate manner which would then pass it's result to the aspects of the project I will be working on. Data that may be passed through this process could include 'real world' data for example openly available hospital patient data (which not may have adequate anonymisation protection against record linkage attacks). Working with this type of data requires ethical caution. However the project is centred around increasing privacy for those individuals and on the sets of data being used within the system. Ensuring that any such input data is handled with the care then the system should ensure no re-identification can occur. This prevents any ethical malpractice due to the restriction on re-identification. However as previously mentioned ideally a few controlled 'well formed' sets of test data would ideally be generated to effectively to test functionality of the system. Therefore I will not delve further into the ethics of handling this type of data unless it is seemed necessary and beneficial for the project. Should I decide to leverage a set of 'real world' data to test my system I will then ensure I am compliant with any regulations and the ethics of handling such data, at which point I would include a section of how the information has been ethically handled in the final report.

Similar Pre-Existing System

ARX – Data Anonymisation Tool [3]

ARX is an open source anonymisation software tool. The application allows users to select an anonymisation technique to obfuscate personal information. However the system I am proposing will be a framework where future developers will be able test new ideas while leveraging the pre-existing infrastructure and post-processing facilities. One benefit of the tool I am proposing would be to leverage a 'pluggable' architecture that would allow developers to create, test and compare their own algorithms. Instead of one fixed pre-defined algorithm implementation for k-anonymisation as in the ARX system.

Work Plan


Weeks:
1
2
3
4
5
6
7
8
9
10
11
>11
1. Initial Project Plan
2. Large Final Report
2.1 Introduction
2.2 Background
2.3 Approach
2.4 Implementation
2.5 Results & Evaluation
2.6 Conclusions/Reflection on Learning
2.7 Proof Reading/Final Corrections
3. Research
3.1 Project Background Research
3.2 Algorithm Implmentation(s)
4. System Design
4.1 Define Data Standrds/Format
4.2 Class Diagrams
4.3 Define Test Cases
5. Software Development
5.1 Infrastructure Development
5.2 Algorithm Implementation(s)
5.4 GUI
5.3 Post-Processing Support
5.5 Testing
6. Viva

References

[1] Sweeny L. (2002). K-ANONYMITY: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 10 (5), p2.

[2] B. C. M. Fung. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys. 42 (4), 7.

[3] ARX – Data Anonymisation Tool . Available: http://arx.deidentifier.org/. Last accessed 30th Jan 2015.