A Research Tool to analyze and compare various Privacy Preserving Techniques used for Big Data

Vineet Santosh Karkera
Msc Computer Science, Memorial University of Newfoundland
MUN# 201281896, e-mail - vsk053@mun.ca

BACKGROUND

There is currently great enthusiasm, concern around how Big Data is going to change the way business decisions are made in the future. Many organizations and companies have accumulated large volumes of data over the years, and can now make predictions based on this history of data. Typically, big data involves data mining, data analysis and data storage of enormous volumes of data, usually generated on a daily basis. This may be looked at as a perk, especially by the marketing sector, but along with this comes an array of security and privacy issues related to big data.

Currently, there is a lot of research being done to improve or create new privacy preserving algorithms. In this paper, a proposal is made for a tool that will help analyze the efficiency of various privacy preserving algorithms.

With big data growing at lightning speed, Information Technology has a lot of catching up to do in terms of providing a cheap, efficient, fast and secure storage service. This paper introduces a tool that will help researchers, students, professionals alike, to analyze the performance of their algorithm used. It can also be used by organizations to verify the level of security provided to a certain data set before it is released to the public.

PROBLEM, MOTIVATION AND PROPOSED SOLUTION

"*Big data is usually accompanied by privacy concerns, on the other side of the same coin as symbolized by Big Brother*". [1]

The harvesting of large sets of personal data and the use of state of the art analytics implicate growing privacy concerns. Protecting privacy will become harder as information is multiplied and shared ever more widely among multiple parties around the world. As more information regarding individuals' health, financial, location, electricity use, and online activity percolates, concerns arise regarding profiling, tracking, discrimination, exclusion, government surveillance, and loss of control. [3]

Thus, several privacy preserving algorithms that have been proposed.

But, there is no tool yet written to figure out if a particular algorithm or technique is better than the other, although there are several metrices that are used to measure.

Following is a list of few metrics that are widely used by researchers to help measure, quantify the quality, efficiency of the results after the implementation of a privacy preserving technique.

1. **Hiding Failure (HF)**

The percentage of sensitive information that can still be effectively discovered after sanitizing data is called the Hiding failure(HF).

2. **Misses Cost (MC)**

Misses cost (MC) measures the percentage of non-sensitive information that is hidden after the sanitization process.

3. **Accuracy and Information Loss**

Accuracy and Information Loss are closely linked to each other. Lesser the information loss, more is the data quality and vice versa i.e more the information loss, less is the quality of the data.

4. **Classification**

It is defined as the sum of the individual penalties for each row in the dataset table normalized by the total number of rows.

5. **Discernibility**

Discernibility metric assigns a penalty to each tuple based on how many tuples in the transformed dataset are indistinguishable from it.

6. **Misclassification Error**

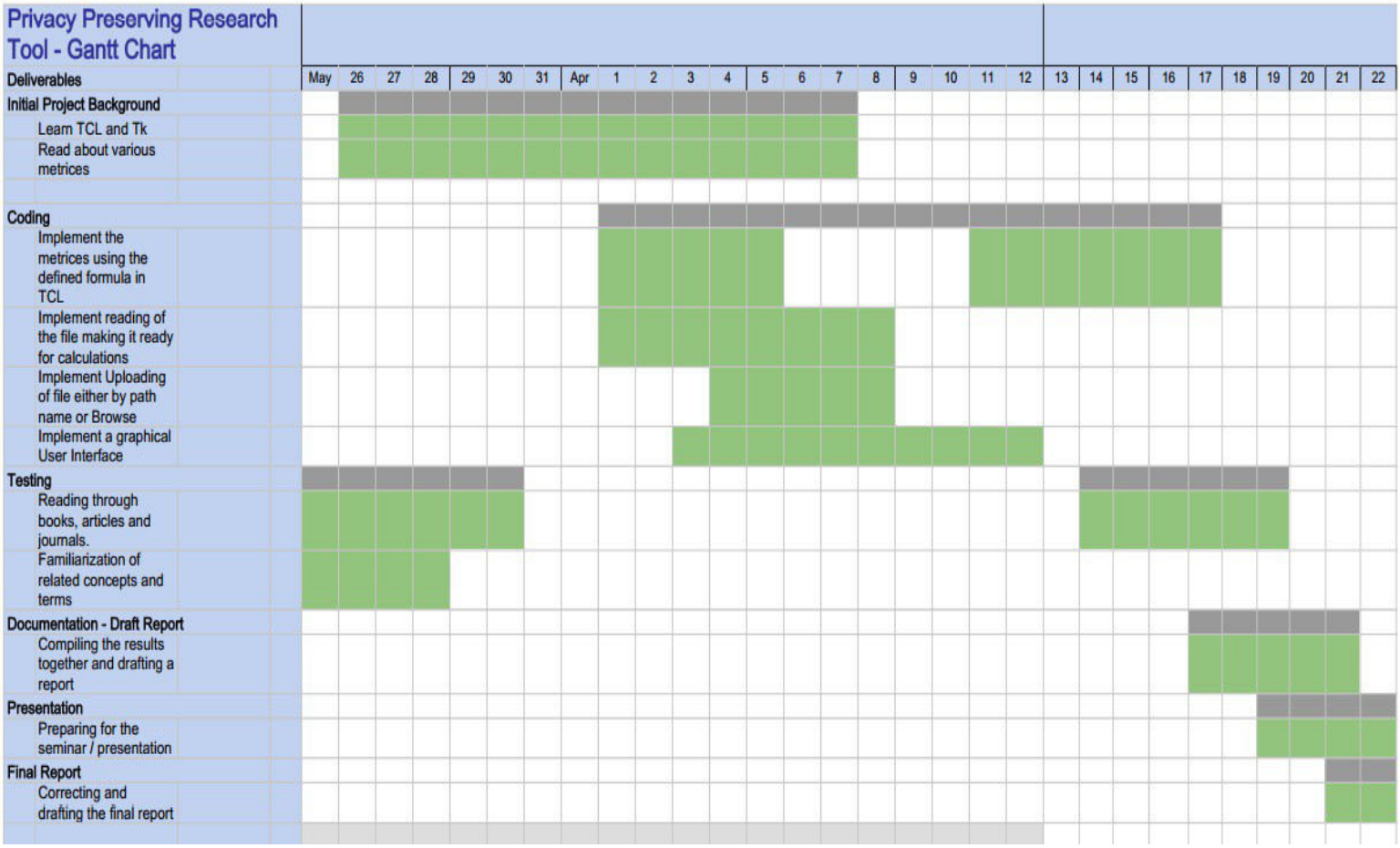
The quality of privacy preserving techniques for data mining is judged by the percentage of legitimate data points that are not well classified after the sanitization process, called the Misclassification metric. [2]

This usually serves as the first step to error analysis.

With the help of the above metrices, given a data table, it will calculate the Hiding Failure, Misses Cost, Accuracy and Information Loss, Classification for the data set. The tool will create a graph representing the trade off between Accuracy and Privacy.

The GUI, file handling and metrices will be implemented in TCL/Tk language.

8. PROJECT SCHEDULE – GANTT CHART



If the above picture is unreadable, please use the link below for the Gantt Chart -

https://docs.google.com/spreadsheet/ccc?key=0AjddRmKKR-QRdFpQZ3NnQl9tbm9yeDNXSkhCaUN2Q3c&usp=sharing

REFERENCES FOR THIS DOCUMENT

[1] “Social Issues of Big Data and Cloud: Privacy, Confidentiality, and Public Utility” by Koichiro Hayashi, Ph.D., LL.D. Institute of Information Security, IEEE, 2013.

[2] Oliveira, S.R.M., Zaiane, O.R.: Privacy preserving clustering by data transformation. In: 18th Brazilian Symposium on Databases (SBBD 2003), pp. 304–318 (2003)

[3] “A Taxonomy of Privacy” by Daniel Solove, University of Pennsylvania, 2006.