Aggregate Query and Analysis While Maintaining Personally Identifying Information Security

Andrew N. Abbott, *anabbott@smu.edu, MSDS 7330*

***Abstract*—Many electronic record keeping systems necessarily collect personally identifiable information (PII), such as social security numbers, dates of birth, or addresses. Protecting this information from undesired users is important. It is also important to be able to analyze the data in ways that combine user' information together to do things like intelligent disease cluster identification or anomaly detection. The goal of this project is to find a way to securely perform certain types of analysis and aggregation without exposing user' PII.**

I. Introduction

DATABASE administrators are the custodians of vast amount of data, much of which may be customer's PII (Personally Identifiable Information). The task of protecting that PII from theft, abuse or negligence is of vital importance. Analysts and business users of reports and datawarehouses are often able to query user' PII directly, putting customers at risk. To address this problem, this papers seeks to first identify PII and then securely protect or anonomize it to remove the possibility of personal identification.

II. Research Methodology

The first step will be to clearly define PII. There have been differing definitions used in literature and regulations. A review of the regulations and risks to consumers due to the current state of the technology used to gain unauthorized access to PII leads to a broader rather than narrower definition. "Any information that distinguishes one person from another can be used for re-identifying anonymous data [1]." I would go further to describe PII as any information related to a person that distinguished one person from another and is not generated solely by the enterprise. Once PII is defined, the goal is to systematize the identification of PII fields for protection.

The second step is to describe and review different strategies for protecting PII. These strategies include but are not limited to: anonymizing the data (scrambling data, hash functions, encryption, tokenization), an interactive query based approach.

A third step is to build and test a query based approach. This may require an interface or some layer between the user and the data. Testing should be for protection, usability, and performance.

III. Previous and Related Reasearch

There has been much research done on the topic of protecting personally identifiable information. The topics range from discussion of the definition of PII, different methods of protecting PII, to the different types of protection required for each stage in the life cycle of the data.

IV. Research Schedule

September 21: Proposal submission.

October 5: Complete prior research review and establish working PII definition.

October 19: Have strategy background complete. Plan query based approach.

November 2: First query prototype complete.

November 9: Refine and test.

November 16: Finalized testing methods.

November 23: Write up results and presentation.

November 30: Presentation.

December 7: Final project submission.

V. Resources Required

To conduct this research, sample data sets containing fictional personal information will be used. MySQL will be used for database administration and query execution. For statistical analysis R will be utilized.

References

[1] A. Narayanan, V. Shmatikov. (2010, June). Privacy and Security, Myths and Fallacies of "Personally Identifiable Information". *Communications of the ACM*. [Online]. *53(6),* pp. 24-26. Available: https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf

[2] Handbook for Safeguarding Sensitive Personaly Identifiable Information, DHS 2012.

[3] Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), NIST Special Publication 800-122, 2010.

Proposal submitted September 21, 2016

A. N. Abbott is a student at Southern Methodist University, Dallas, TX 75275 USA (e-mail: anabbott@smu.edu).