[[1]](#footnote-1)

Aggregate Query and Analysis While Maintaining Personally Identifying Information Privacy

Andrew N. Abbott, [anabbott@smu.edu](mailto:anabbott@smu.edu)*, Joseph Stoffa,* [jstoffa@smu.edu](mailto:jstoffa@smu.edu)*, Nathan Mowat,* [nmowat@smu.edu](mailto:nmowat@smu.edu)

*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. Researchers have been able to de-anonymize private information with very little information about data sets by using other publicly available data sets. The goal of this project is to identify methods to further anonymize data sets to securely perform certain types of analysis and aggregation without exposing user’ PII.

# INTRODUCTION

D

atabase administrators are the custodians of vast amount of data, much of which may be customer’s Personally Identifiable Information (PII). The task of protecting that PII from theft, abuse or negligence is of vital importance. Analysts and business users of reports and data warehouses are often able to query user’ PII directly, putting customers at risk. Despite these risks, analysts must be able to learn properties of the population while preserving that privacy. To address this problem, this paper seeks to first identify PII and describe the various methods used to protect PII, then to detail a method to securely protect or anonymize the data by introducing noise to the data without changing the aggregate statistics or the data set, and finally to answer the question of how much noise or modification is necessary to protect the privacy of the data without significantly changing the statistics used for analysis.

# 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.

The second step is to review the different strategies for protecting PII. These include but are not limited to: de-identifying the data, anonymizing the data, and differential privacy [1].

Our goal is aggregate query and analysis while maintaining PII protection. In pursuit of this goal, Cynthia Dwork introduced the concept of differential privacy [2]. Rather than attempting to guarantee absolute protection of individual data, differential privacy reduces the ability of disclosures to identify individuals. The presence or absence of an individual record will not affect the calculation output and thus be unidentifiable. This method involves the modification of the data by introducing noise.

The final step is to test different levels of data modification to maximize the usefulness of the resulting statistics while also minimizing the ability to differentiate any individual record.

# Previous and Related Research

There has been much research done on the topic of protecting PII. 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. Conclusion

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. C. Dwork, “Differential privacy,” in ICALP, 2006, pp. 1-12.
3. https://en.wikipedia.org/wiki/Differential\_privacy
4. Handbook for Safeguarding Sensitive Personaly Identifiable Information, DHS 2012.
5. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), NIST Special Publication 800-122, 2010.

1. Proposal submitted January 31, 2017

   A. N. Abbott, N. Mowat, and J. Stoffa are students at Southern Methodist University, Dallas, TX 75275 USA. [↑](#footnote-ref-1)