[[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 database systems collect personally identifiable information (PII), such as social security numbers, dates of birth, or addresses. Protecting this information from undesired users is vital. It is also important to be able to analyze the data in ways that aggregate user’ information together to do things like disease 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 paper is to explore differential privacy methods and to both accurately and securely perform certain types of analysis and aggregation without exposing user’ PII.

# INTRODUCTION

D

atabase administrators are the protectors of an ever growing amount of data, much of which is customer’s Personally Identifiable Information (PII). All this data is continuously being analyzed, bought and sold. The task of protecting that PII from theft, abuse or negligence is critical. Despite these risks, analysts must maintain the ability to calculate statistics based on these data sets while still preserving that privacy. To address this problem, this paper seeks to first identify PII and describe some of the methods used to protect PII, then to detail the method to securely protect or anonymize the data by introducing noise to the data without changing the aggregate statistics of the data set known as “differential privacy”. In pursuit of this goal, Cynthia Dwork introduced the concept of differential privacy [1]. Rather than attempting to guarantee absolute protection of individual data, differential privacy reduces the ability of disclosures to identify individuals. Finally, we will seek 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

We begin by defining PII for the purposes of this paper. Definitions vary across and within domains. Generally, “any information that distinguishes one person from another can be used for re-identifying anonymous data [2]”. This definition provides us with too broad a metric. For our purposes, we will define PII as “any information related to a person that distinguishes one person from another and is not intended for the public.”

Second, we will develop a test dataset which we will use to examine the effects of noise on accuracy. This dataset will be sufficient in number of observations and scope without being unnecessarily complex. Unlike a real-life data set it will also provide “factual” responses in order to gauge true accuracy.

The Third step is to research the most current and promising methods of providing differential security and to evaluate their effectiveness. This will include a comparison and contrast of the different methodologies.

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. Ideally, 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. In effect, we will be trying to determine the ideal amount of noise needed to provide the required anonymity without significantly altering the statistics produced by the aggregation.

# Previous and Related Research

Dwork published a procedure for differential privacy which has served as the basis for most other research in the field. Two of Dwork’s key contributions are the concepts of ε known as the *privacy budget* and *sensitivity*. The privacy budget limits the number of queries a user can execute before they exceed the budget and therefore can begin to compromise the data’s confidentiality [1]. Sensitivity “measures how much a query amplifies the distance between two inputs” or more plainly it is the max number of queries that can be made on two databases that differ by only one row (referred to as *neighbors*). Both terms are critical to the discussion of differential privacy data aggregation optimization.

Two competing mechanisms for differential privacy of linear aggregate data are the Matrix Mechanism [8] and the Low- Rank Mechanism [9]. Both of which are the main mechanisms used in this paper. G. Yuan et al, state that while theoretically sound the Matrix Mechanism is unpractical and, according to their tests, produces no more accuracy then more naïve methods of differential privacy [12].

Apart from applying differential privacy to linear and batch aggregation there has also been increased interest its application to other areas as well. L. Fan et al and E. Shi et al have both made noteworthy contributions to differential privacy of aggregations of time-series data [7, 10]. Recently in the arena of machine learning, M. Ababi et al have attempted to develop techniques for conducting analysis while maintaining privacy [11].

# Contrasting Differentially Private Data aggregation

# Testing

Here we will test different levels of noise and various

# Conclusion

References

1. PII C. Dwork, “Differential privacy,” in ICALP, 2006, pp. 1-12.
2. 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
3. C. Dwork, “Differential privacy,” in ICALP, 2006, pp. 1-12.
4. https://en.wikipedia.org/wiki/Differential\_privacy
5. Handbook for Safeguarding Sensitive Personaly Identifiable Information, DHS 2012.
6. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), NIST Special Publication 800-122, 2010.
7. L. Fan and L. Xiong. Adaptively sharing time-series with differential privacy. CoRR, abs/1202.3461, 2012.
8. Li, Chao, "Optimizing Linear Queries Under Differential Privacy" (2013). Dissertations. 805. http://scholarworks.umass.edu/open\_access\_dissertations/805
9. G. Yuan, Z. Zhang, M. Winslett, X. Xiao, Y. Yang and Z. Hao. Low-Rank Mechanism: Optimizing Batch Queries under Differential Privacy. VLDB, 2012.
10. Shi, E., Chan, H. T. H., Rieffel, E., Chow, R., & Song, D. (2011). Privacy-preserving aggregation of time-series data. In Annual Network & Distributed System Security Symposium (NDSS). Internet Society..
11. Abadi, M. Chu, A, Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, Li. (2016). Deep Learning with Differential Privacy. In ACM Conference on Computer Communicatios Security.
12. G. Yuan, Z. Zhang, M. Winslett, X. Xiao, Y. Yang and Z. Hao. Low-Rank Mechanism: Optimizing Batch Queries under Differential Privacy. VLDB, 2012.
13. W. Yuxiang, "Differential Privacy: a short tutorial,". [Online]. Available: https://www.cs.cmu.edu/~yuxiangw/docs/Differential%20Privacy.pdf. Accessed: Mar. 1, 2017.
14. M. Abadi et al., "Deep Learning with Differential Privacy," in arxiv.org, 2016. [Online]. Available: https://arxiv.org/pdf/1607.00133.pdf. Accessed: Mar. 1, 2017.
15. N. Zumel, "A simpler explanation of differential privacy," in Win-Vector Blog, Win-Vector Blog, 2015. [Online]. Available: http://www.win-vector.com/blog/2015/10/a-simpler-explanation-of-differential-privacy/. Accessed: Mar. 1, 2017.
16. M. J. Kusner, J. R. Gardner, R. Garnett, and K. Q. Weinberger, "Differential Private Bayesian Optimization,". [Online]. Available: http://www.cs.cornell.edu/~kilian/papers/private\_GP.pdf. Accessed: Mar. 1, 2017.
17. L. Fan and L. Xiong, "Adaptively Sharing Real-Time Aggregate with Differential Privacy," 2013. [Online]. Available: https://arxiv.org/pdf/1202.3461.pdf. Accessed: Mar. 1, 2017.
18. videosfromIAS, "The definition of differential privacy - Cynthia Dwork," in YouTube, YouTube, 2016. [Online]. Available: https://www.youtube.com/watch?v=lg-VhHlztqo. Accessed: Mar. 1, 2017.

1. [↑](#footnote-ref-1)