Novel Techniques in Private Telemetry Analysis

Trey Scheid
tscheid@ucsd.edu

Tyler Kurpanek
tkurpane@ucsd.edu

Bradley Nathanson
bnathanson@ucsd.edu

Christopher Lum
cslum@ucsd.edu

Yu-Xiang Wang
yuxiangw@ucsd.edu

Abstract

Is the abstract supposed to be in light gray as well? Seems difficult to read. This research investigates the practical implementation of differential privacy mechanisms for telemetry data analysis, with a focus on real-world applications. We propose a comprehensive framework that employs various privacy-preserving techniques, including randomized response and the Laplace mechanism, to protect sensitive information while maintaining analytical utility. Our methodology encompasses multiple statistical tasks, from user-level rate analysis to logistic regression classification. The study utilizes AutoDP for precise privacy loss measurement and documents the inherent tradeoffs between privacy guarantees and analytical accuracy in production environments. By demonstrating the feasibility of differential privacy in telemetry analysis, we provide a roadmap for organizations seeking to enhance their privacy practices.

Website: https://endurable-gatsby-6d6.notion.site/DP-Telemetry-14556404e74780818747cbe76de2e04a?pvs=4[Notion until actual site is created]
Code: https://github.com/Trey-Scheid/Novel-Techniques-in-Private-Data-Analysis

1 Introduction

The implementation of differential privacy in production environments presents significant challenges in balancing privacy guarantees with analytical utility. This research addresses these challenges by developing practical privacy-preserving mechanisms for existing telemetry analysis tasks while maintaining the usefulness of their systems. We identify comprehensive frameworks that integrate various differential privacy mechanisms, including Guassian Composition and the Laplace mechanism, to protect sensitive information in telemetry data. Our methodology encompasses multiple statistical tasks, from user-level rate analysis to logistic regression classification, and utilizes AutoDP for precise privacy loss measurement. By evaluating the tradeoffs between privacy guarantees and analytical accuracy in production settings, we provide a roadmap for organizations looking to enhance their privacy practices.

1.1 Motivation

Despite the growing importance of privacy-preserving data analysis, many practitioners perceive differential privacy implementation as complex and challenging **?**. This perception stems from several factors: the mathematical complexity of privacy definitions, the need to carefully calibrate privacy parameters, and concerns about reduced utility **?**. A survey by Smith et al. found that only 23% of data scientists felt confident implementing differential privacy mechanisms in their workflows **?**. However, recent developments have significantly lowered these barriers to entry. Tools like Google's Privacy on Beam [1], Microsoft's SmartNoise [2], and various open-source libraries like AutoDP[3] provide accessible frameworks for implementing differential privacy. These tools abstract away much of the underlying complexity while maintaining rigorous privacy guarantees. Additionally, educational resources and practical tutorials have emerged to guide practitioners through implementation challenges **?**. This research builds upon these recent developments by providing a practical demonstration of differential privacy mechanisms in telemetry data analysis. By implementing privacy-preserving techniques for existing tasks, we aim to show that differential privacy can be seamlessly integrated into production systems without significant utility loss. Our work focuses on two key objectives: privatizing existing telemetry analysis tasks and evaluating the privacy-utility tradeoffs in production settings.

[1] linkneeded

[2] linkneeded

[3] linkneeded

1.2 Background and Literature Review

1.3 Differential Privacy

Differential privacy is a framework for data privacy that gives a mathematical guarantee that the sensitivity of each individual in the dataset is preserved. The core idea is to introduce random noise to the output of algorithms so that any single individual's data does not significantly affect the overall result. Mathematically, a mechanism is considered (,) differentially private if for all datasets D and D' which differ by at most 1 element when $\mathbb{P}[M(D) \in S] \leq e^{\epsilon}\mathbb{P}[M(D') \in S)] + \delta$ where ϵ and δ are privacy loss parameters. Smaller and imply stronger privacy guarantees.

Differential privacy is applied to algorithms, not datasets. One common and foundational algorithm is logistic regression. Many privatized implementations of logistic regression exist, leaving data scientists with a host of convoluted choices and complex language about parameters they may not fully understand. We hope to show some examples that will help practitioners implement this model on their own datasets.

1.4 Intel Telemetry Data

Differential privacy methods and guarantees are attractive for many domains. Telemetry is the remote data transfer of automated system measurements. As people use technology everyday their machines track usage diagnostics which are used by hardware and software manufacturers to reduce bugs and increase efficiency. System usage information is recorded at regular intervals and usually results in massive quantities of measurements. The identifiability of the specific machine or user of an event is a concern regardless of PIID tags. Dinur Nissim **?** and linkage attacks can be used to recover or reconstruct the original information: the source. This is a breach of privacy for a user which depending on the sensitivity of the information can be concerning. For example, personal laptops may send diagnostics to intel given that the user opts in to the program [Intel telemetry].

We use a secure research database shared be Intel Corporation with consent of its users to generate real results....

2 Methods

2.1 Data Preprocessing

Add some language here about steps that were universal between all tasks if any

For each of the follow sections we will describe the task, the algorithm used, and the implementation details.

2.1.1 Logistic Regression (DP-SGD)

This paper[4] investigates how privacy affects different mini-batch stochastic gradient descent algorithms for logistic regression classification. It is shown that privacy affects the batch size for optimal performance.

2.1.2 LASSO Regression (DP-FW)

will add lots of detail about lasso, then talk about adapting franke-wolfe to be differentially private.

2.2 Tyler task

2.3 Bradley task

3 Results

Should we do a results section for each task separately again?

3.1 Combined Results

We have discussed with Yu-Xiang a plot we can create which combines all the tasks into 1.

[4] needs citation not a footnote

4 Discussion

4.1 Interpretation

4.2 Limitations

5 Conclusion

5.1 Summary

5.2 Impact

5.3 Future Direction

6 Contributions

6.1 Author Contributions

: T.S. focused on task22 LASSO Regression to highlight the exploratory capabilities of private data while implementing a previously theoretical framework (Franke-Wolfe). C.L. implemented the algorithms in ... B.N. analyzed the experimental results ... T.K. analyzed the experimental results ... Y.W. supervised the research and provided guidance on the mathematical foundations. All authors contributed to writing and reviewing the manuscript.

6.2 Task Details

Trey Scheid

- Replication of
- Implementation of non-private franke-wolfe lasso regression
- Ethics considerations webpage
  Todo: Implementation of private franke-wolfe lasso regression

Tyler Kurpanek

Bradley Nathanson

Christopher Lum

Yu-Xiang Wang

- Concept ideation
- Data Access
- Provided guidance on the mathematical foundations

- Proofing and editing all content

6.3 Acknowledgements

We would like to recognize the support of our instructor, Yu-Xiang Wang, for his guidance and feedback throughout the project. We would also like to thank the teaching staff Umesh Bellur and Shriniwas Kulkarni for their support and feedback. The tasks database was a foundational part of our work and was created by another student researcher: Qiyu Li.

We also would like to thank the authors of the papers we referenced in our literature review. Their work was instrumental in our understanding of the topic and the development of our project. Our understandings of differential privacy has been built on the work of many researchers in the field such as: __, __, __, and __. Especially those which engaged in discussion with us about the field (Smith, Ulman, Guatam et al.). We are grateful for their contributions.

Appendices

A.1 Project Proposal

Need to add project_proposal.tex to folder then uncomment here.

A.2 Appendix A: Additional Results

example

A.3 Appendix B: Training Details

example

A.4 Appendix C: Additional Figures

example