Securing Data with Differential Privacy: A Developer's Guide

Josh Queja, Leah Sarlez, Alexander Specht

December 2, 2024

Background

In today's world, technology is embedded in nearly every aspect of our lives, leading to the production of large, rapidly growing volumes of data [1]. According to IBM, "Data is a collection of facts, numbers, words, observations or other useful information" [2]. As data has become an essential part of daily life, the protection of privacy in the digital age has gained significant attention. Privacy, in this context, can be understood as an individual or group's capacity to control the sharing and use of personal information selectively.

As organizations collect vast amounts of data to drive innovation and improve services, companies face increasing responsibility to safeguard user data and comply with privacy regulations and standards [3]. Differential privacy has emerged as a leading technique for protecting user data in a way where user's data is protected and organizations can still conduct analysis on the data. To support users of any experience level interested in implementing differential privacy, we will create a guide at applying differential privacy to various datasets. This guide is for anyone, ranging from beginner to more advanced users who want to explore privacy concerns with their data.

Current Work

As privacy in data practices and technologies becomes increasingly more important, open source differential privacy libraries must provide impactful utility to users of all skill levels and for various different mathematical applications to remain effective [4]. Google DP – Google's library under the Apache-2.0 license – offers a suite of DP analytics queries through a Python wrapper around C++, providing an interface layer for non-experts based on Apache Beam while also allowing experts to directly access and utilize the DP "building blocks"[5]. In particular, Google DP implements a Laplace mechanism – an algorithm drawing from a Laplace distribution with adjustable query sensitivity and privacy budget inputs – to safely release numeric data while protecting individual data points, and a Gaussian Mechanism utilizing a Gaussian distribution with similar sensitivity and privacy budget parameters, but instead emphasizing increased flexibility in application and utility at the cost of a small privacy failure probability for larger datasets that might not require the exact privacy of a Laplace mechanism. However, Google DP's general utility and approachability for all skill levels limits its applications to primarily to production applications of data science with no

machine learning utilities [5].

Meanwhile, IBM's diffprivlib, despite being a general purpose library, also includes a plethora of DP mechanisms beyond the basic Gaussian and Laplacian, like both the truncated and bounded forms of the Geometric mechanism for a discrete probability distribution particularly beneficial for discrete, non-negative data due to its efficiency and simplicity in achieving differential privacy guarantees for those types of datasets [6]. Diffprivlib offers a vast collection of mechanisms, analytics queries, and differential privacy machine learning algorithms that make up for the vulnerability from no floating-point safety implementation, making it a slightly riskier but much more comprehensive library for both experimental data science and machine learning [7]. Each library serves to fulfill a specific set of goals and datasets, along with varying levels of security and limitations implemented throughout a series of trade-offs between technological complexity, privacy risks, and data utility, resulting in the need for a wide variety of differential privacy technologies.

Motivation

Our team decided to go the route of implementing differential privacy because differential privacy is one of the most popular techniques used today that reduces the identifiability of individuals within datasets. Most big companies including Amazon, Google, Facebook, Apple, and many more currently use Differential Privacy today [8]. Because of this, our team wants to know the level of utility that differential privacy brings to data sets while still protecting individuals data so we can better understand how our data is protected within these popular applications today. Similarly, there have been risks and errors when companies adopt differential privacy and by studying the algorithm we can better understand what causes said risks and where errors might occur [9].

Contribution

- Introduction to different types of noise-adding algorithms
- Explanation of each algorithm's working mechanism
- Scenarios for choosing the most suitable noise-adding method
- User-friendly guide for practitioners of all experience levels

Timeline

1. Week 8
   - Finish Project Write up
   - Find three open source libraries regarding to Differential Privacy
   - Research Algorithms
2. Week 9
   - Start Paper
   - Find Datasets and implement algorithms - Write the walk through

3. Week 10
   - Get feedback
   - Implement feedback

4. Finals
   - Final touches
   - Submit Project

References

[1] Uthayasankar Sivarajah, Mohammad M. Kamal, Zahir Irani, and Vishanth Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of Big Data*, 4(1):1–24, 2017.

[2] Mary Theofanos. Differential privacy: A Q&A with NIST's Mary Theofanos, 2019. Accessed: 2024-11-11.

[3] TrustCloud Community. Data protection in technological advancements: Balancing between innovation and privacy. `https://community.trustcloud.ai/docs/grc-launchpad/grc-101/governance/data-protection-in-technological-advancements-balancing-between-innovation-and-priva` 2023. Accessed: 2024-11-14.

[4] Patrick Song. From theory to implementation: How open-source dp libraries shape mental models of privacy concepts, 2024.

[5] Gonzalo Munilla Garrido, Joseph Near, Aitsam Muhammad, Warren He, Roman Matzutt, and Florian Matthes. Do I get the privacy I Need? Benchmarking Utility in Differential Privacy Libraries, 2021. Accessed: 2024-11-11.

[6] Naoise Holohan, Stefano Braghin, Pol Mac Aonghusa, and Kilian Levacher. Diffprivlib: The ibm differential privacy library. *arXiv*, 2019. Accessed: 2024-11-14.

[7] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: On the trade-off between utility and information leakage. In *Lecture Notes in Computer Science*, pages 39–54. 2012.

[8] Damien Desfontaines. A list of real-world uses of differential privacy - ted is writing things, October 1 2021. A List of Real-World Uses of Differential Privacy.

[9] Kai Chen and Qiang Yang. Differential privacy. In *Privacy-Preserving Computing: For Big Data Analytics and AI*, pages 80–104. Cambridge University Press, 2023.