Differential privacy – basics, tools, application scenarios

Sabidur Rahman

Friday Group Meeting

Network Research Lab, CS, UC Davis

Nov. 15, 2019

Differential privacy – introduction

- The problem of privacy-preserving data analysis has a long history spanning multiple disciplines
- Electronic data about individuals are becoming increasingly detailed
- Technology enables ever more powerful collection and curation of these data
- Need increases for a robust, meaningful, and mathematically rigorous definition of privacy, together with a computationally rich class of algorithms that satisfy this definition
- *Differential Privacy* is such a definition

C. Dwork and A. Roth, "*The Algorithmic Foundations of Differential Privacy,"* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy - introduction

- Differential privacy is a meaningful and mathematically rigorous definition of 'privacy' useful for quantifying and bounding privacy loss
- Differential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population
- Differential privacy ensures that the same conclusions, for example, smoking causes cancer, will be reached, independent of whether any individual opts into or opts out of the data set
- It ensures that any sequence of outputs (responses to queries) is "essentially" equally likely to occur, independent of the presence or absence of any individual

C. Dwork and A. Roth, "*The Algorithmic Foundations of Differential Privacy,"* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy – example scenario

- Data cannot be fully anonymized and remain useful. The richer the data, the more interesting and useful it is.
- This has led to notions of “anonymization” and “removal of personally identifiable information”; but not enough
- ‘linkage attack’ to match “anonymized” records with non-anonymized records in a different dataset has led to violation of privacy
- Medical records of the governor of Massachussetts were identified by matching anonymized medical encounter data with (publicly available) voter registration records

C. Dwork and A. Roth, “*The Algorithmic Foundations of Differential Privacy,”* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy – example scenario

- Netflix subscribers whose viewing histories were contained in a collection of anonymized movie records published by Netflix as training data for a competition on recommendation were identified by linkage with the Internet Movie Database (IMDb)
- Differential privacy neutralizes 'linkage attacks': since being differentially private is a property of the data access mechanism, and is unrelated to the presence or absence of auxiliary information available to the adversary

C. Dwork and A. Roth, "*The Algorithmic Foundations of Differential Privacy,"* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy – example scenario

- Queries over large sets are not protective
- Forcing queries to be over large sets is not a panacea, as shown by the following **differencing attack**
- Suppose it is known that Mr. X is in a certain medical database
- Taken together, the answers to the two large queries "How many people in the database have the sickle cell trait?" and
- "How many people, not named X, in the database have the sickle cell trait?" yield the sickle cell status of Mr. X

C. Dwork and A. Roth, "*The Algorithmic Foundations of Differential Privacy,"* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy – example scenario

- “Ordinary” facts are not “OK.”
- Revealing “ordinary” facts, such as purchasing bread, may be problematic if a data subject is followed over time
- For example, consider Mr. T, who regularly buys bread, year after year, until suddenly switching to rarely buying bread
- An analyst might conclude Mr. T most likely has been diagnosed with Type 2 diabetes
- The analyst might be correct, or might be incorrect; either way Mr. T is harmed

C. Dwork and A. Roth, “*The Algorithmic Foundations of Differential Privacy,”* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy - definition

- A *database* is a set of rows, each row containing the data of a single individual
- Opting in or out of the database is formalized by considering pairs of databases *x*,*y* differing in at most one row, meaning, one database is a subset of the other and the larger database contains exactly one additional row
- Parameter $\varepsilon > 0$ is public, and its' selection is a social question. Smaller ε yields a stronger privacy guarantee

Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

- Probability of *M* returning the same set S
- Outcome from two datasets should be indistinguishable
- Depends on ε and δ

- Close ε and δ are to zero, more private *M is*
- common way to achieve such M is to add some special noise (e.g., binomial noise) to the original queries

C. Dwork and A. Roth, "*The Algorithmic Foundations of Differential Privacy,*" Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

Differential privacy - definition

- For a given computational task **T** and a given value of **ε** there will be many differentially private algorithms for achieving **T** in an ε-differentially private manner
- Some will have better accuracy than others. When ε is small, finding a highly accurate ε-differentially private algorithm for **T** can be difficult, much as finding a numerically stable algorithm for a specific computational task can require effort

C. Dwork and A. Roth, "*The Algorithmic Foundations of Differential Privacy,"* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2013.

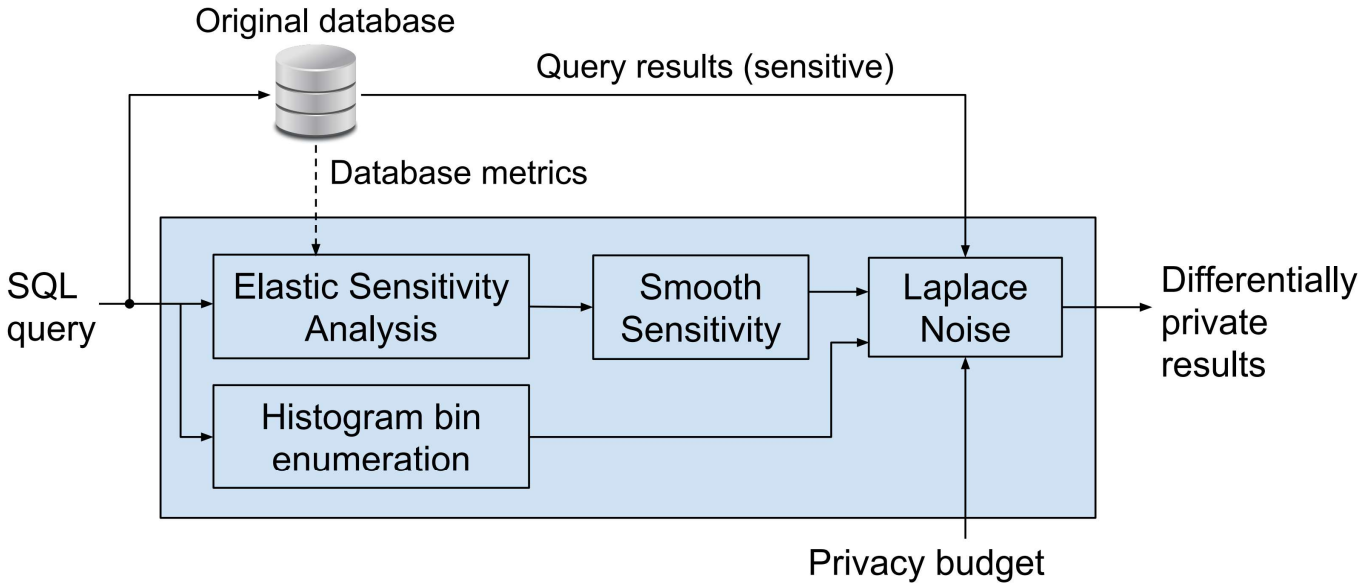
Differential privacy - tools

Uber Elastic Sensitivity

Differential privacy is enforced by adding noise to a query's result, but some queries are more sensitive to the data of a single individual than others. To account for this, the amount of noise added must be tuned to the sensitivity of the query, which is defined as the maximum change in the query's output when an individual's data is added to or removed from the database. A major challenge for practical differential privacy is how to efficiently compute the sensitivity of a query.

Elastic sensitivity calculates the sensitivity of a query in a few milliseconds even for large databases

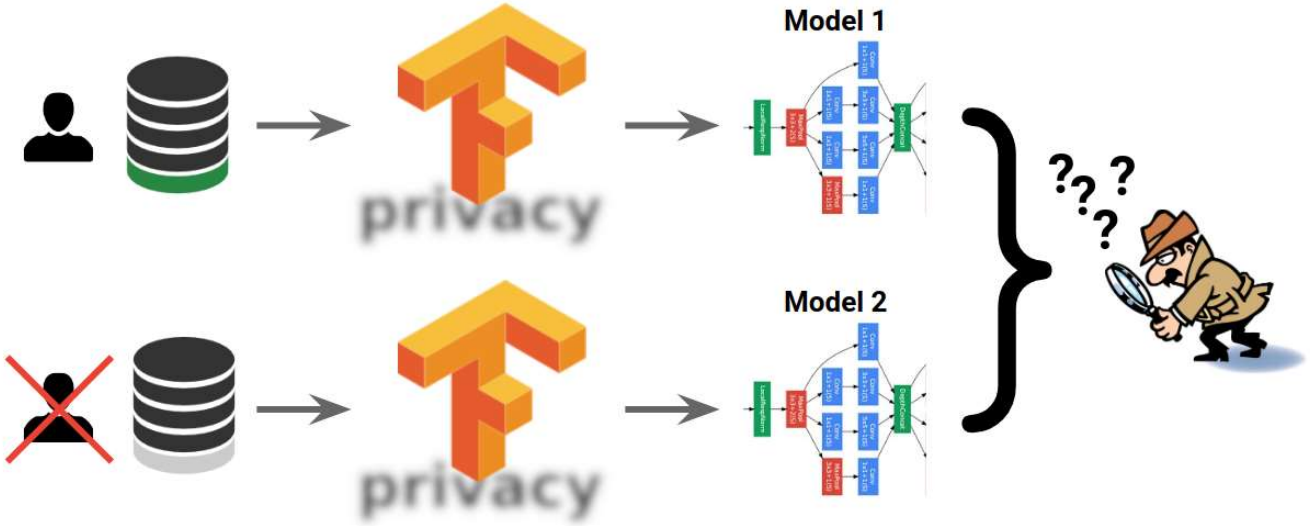
Microservice: differential privacy as a service



https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6

TensorFlow Privacy

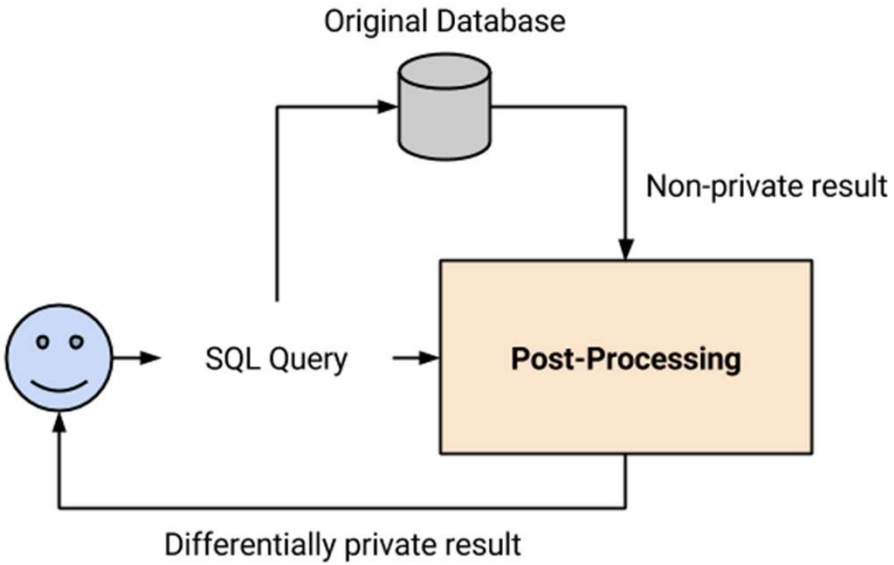
TensorFlow Privacy can prevent memorization of user details and can guarantee that two machine-learning models will indistinguishable whether some examples (e.g., some user's data) was used in their training.



TensorFlow Privacy is to set three new hyperparameters that control the way gradients are created, clipped, and noised. Setting these three hyperparameters can be an art, but the TensorFlow Privacy repository includes guidelines for how they can be selected for the concrete examples.

https://medium.com/tensorflow/introducing-tensorflow-privacy-learning-with-differential-privacy-for-training-data-b143c5e801b6

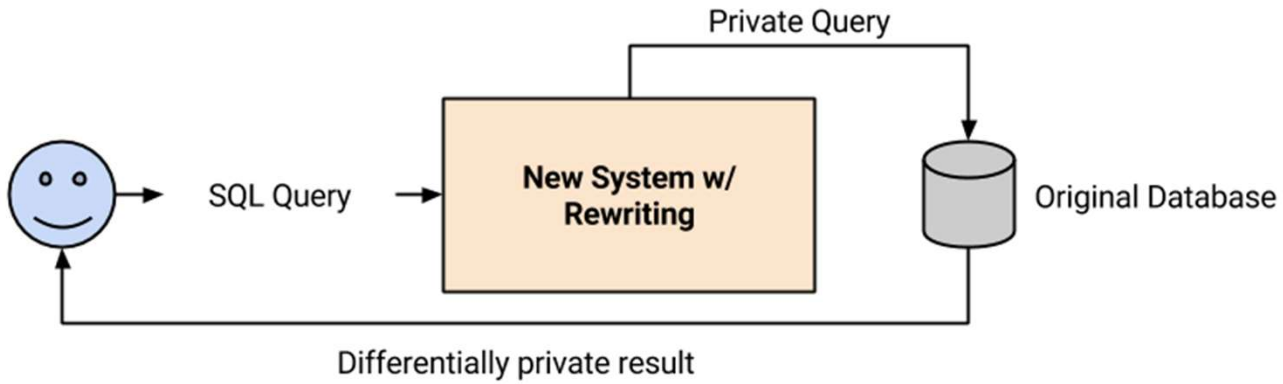
Uber Elastic Sensitivity-update



Post-processing, including Elastic Sensitivity, is DBMS-agnostic, but only supports a limited number of post-processing mechanisms.

Deeply integrated systems support many differential privacy mechanisms, but implementation either necessitates complex changes to the underlying DBMS or a purpose-built DBMS *for each mechanism*.

The key advancement in this release is to embed the differential privacy mechanism into the SQL query itself, before execution, so the query enforces differential privacy on its own output.



https://medium.com/uber-security-privacy/uber-open-source-differential-privacy-57f31e85c57a

Harvard Privacy Tools Projects

- PSI-A Private Data Sharing Interface: a prototype system allows researchers to:
  - upload private data to a secured Dataverse archive,
  - decide what statistics they would like to release about that data, and
  - release privacy preserving versions of those statistics to the repository,
  - that can be explored through a curator interface without releasing the raw data
  - allows interactive queries.
- OpenDP: a community effort to build a system of tools for enabling privacy-protective analysis of sensitive personal data, focused on an open-source library of algorithms for generating differentially private statistical releases.

http://psiprivacy.org/static/about/index.html

Differential privacy – application scenarios

DP – application: Smart-Grid data

- The smart grid introduces new privacy implications to individuals and their family due to the fine-grained usage data collection.
- Smart metering data could reveal highly accurate real-time home appliance energy load, which may be used to infer the human activities inside the houses.
- Even though some techniques has been demonstrated useful and can prevent certain types of attacks, none of existing works can provide probably privacy-preserving mechanisms.
- The following study investigates the privacy of smart meters via differential privacy.

J. Zhao et al. "Achieving differential privacy of data disclosure in the smart grid." *IEEE INFOCOM, 2014*.

DP – application: Edge computing for IoT

- Edge computing makes full use of the computing power of edge nodes, which greatly reduces the computing pressure of data centers, and brings great convenience to the storage and processing of big data
- However, it is easy to become the object of hacker attacks due to the lack of centralized management of distributed nodes
- Once these nodes are compromised, a series of privacy issues can happen
- Authors prodive overview the architecture of MEC for IoT
- Authors discuss privacy issues in the MEC, especially in data aggregation and data mining
- They consider machine learning privacy preserving as a case study in the application of MEC

Miao Du et al. "Big data privacy preserving in multi-access edge computing for heterogeneous internet of things." *IEEE Communications Magazine, vol.* 56, no. 8 pp. 62-67, 2018

DP – application: Indoor positioning data

- Most city dwellers spend over 80% of daily life indoor
- Data containing users' indoor positioning information is a critical asset for understanding the indoor behavior pattern of users: shopping behavior pattern of customers in a large department store
- There is also a risk of leakage of personal information, because it is feasible to infer the users' sensitive information by tracking and analyzing the users' indoor positions
- Local differential privacy (LDP) is the state-of-the-art approach that is used to protect individual privacy in the process of data collection
- LDP ensures that the privacy of the data contributor is protected by perturbing her/his original data at the data contributor's side
- Thus, the data collector cannot access the original data, but is still able to obtain population statistics
- This study focuses on the application of LDP to the collection of indoor positioning data
- We experimentally evaluated the utilization of indoor positioning big data collected by leveraging LDP for estimating the density of the specified indoor area
- Experimental results with both synthetic and actual data sets demonstrate that LDP is well applicable to the collection of indoor positioning data for the purpose of inferring population statistics

J. Kim et al., "Application of local differential privacy to collection of indoor positioning data." IEEE Access, vol. 6, pp. 4276-4286, 2018.

Summary

- Privacy-aware data analysis and AI is a must in usage of user data
- As network research is getting increasingly data dependent, researchers should pay attention to privacy mechanisms and tools available
- Usecases for differential privacy in network research is still evolving
- Identifying important problem that can benefit from DP, proposing solutions that preserves privacy without impacting performance