Differential Privacy in practice - Expose your epsilons!

A paper by Cynthia Dwork, Nitin Kohli and Deirdre Mulligan

Table of Contents

Introduction

- Differential privacy implementations successfully leveraged in various industries
  - This allows scientists, engineers, and researchers to learn about populations of interest without specifically learning about individuals.

- Examples of industrial use include:
  - in the Google Chrome browser to identify vectors for malware
  - in Microsoft Windows’ collection of usage and error statistics
  - in all releases of Apple’s macOS and iOS since 2016 to “identify things like the most popular emoji, the best QuickType suggestions, and energy consumption rates in Safari.”
  - Open source implementations are available as well

- Differential privacy allows us to **quantify cumulative privacy loss**
  - Could be a watershed moment for privacy in systems relying on personal information
  - Differential privacy allows us to assess the relative quality of a firm’s privacy practices prior to purchase or participation

- Devil is in the details → implementation details matter
  - When implemented well -→ insights into data with minimal privacy loss
  - To maximize learning with a meaningful degree of privacy → privacy parameter ϵ must be chosen well
  - But what is the optimal value of ϵ for a given system/data?
  - How to figure it out?

Purpose of this paper

To understand how to choose the privacy parameter in practice, the authors:

- ➔ Explore the current differential privacy implementations
- ➔ Conducted interviews with differential privacy practitioners to learn from their experiences.

Results:

- ➔ No clear consensus on how to choose ϵ
- ➔ No agreement on how to approach this and other key implementation decisions
- ➔ Need for shared learning amongst the differential privacy community

Proposal:

Create Epsilon Registry → a publicly available communal body of knowledge about differential privacy implementations

Differential Privacy Implementations

In this section:

- Define differential Privacy
- Explore important properties of differential privacy
- Clarify its scope

1.1. Defining Differential Privacy

➔ Differential privacy hides the presence or absence of any individual in a dataset

➔ For each individual, any conclusion reached from the analysis would be essentially as likely to have been reached, whether the given individual joined, or refrained from joining, the dataset

➔ The decision to opt in or opt out of the data set will not significantly change the risk.

➔ Here “significantly” is controlled by the parameter ϵ, where a smaller ϵ means less change and hence better privacy.

Defining Differential Privacy

- ➔ A smaller ϵ means less change in result if individual is in or not and hence better privacy.
- ➔ The change (increase or decrease) has a maximum bound e^{ϵ}
- ➔ When this bound e^{ϵ} is close to one (i.e., ϵ is close to 0), anything that can be learned about an individual who has participated is almost equally likely to be learned about an individual who has not participated.

Differential privacy separates learning about a population as whole from learning idiosyncrasies of individual people.

- ➔ This is accomplished by introducing a controlled amount of randomness into the computation.
- ➔ This means that the output of a differentially private analysis depends not only on the data but also on the randomness

1.2. The Opportunity for Meaningful Evaluation of Institutional Differential Privacy Practices

Four features of differential privacy that make it particularly valuable:

- The privacy loss in differential privacy implementations can be objectively measured via ϵ.
    - This permits comparative privacy risk assessment between systems.

- The need to select ϵ creates an opportunity for reflecting on values of organizations:
    - The ability to tune ϵ forces institutions to select a privacy-utility mix and document it, allowing an adjustment of preference for privacy and utility.

- Being differentially private is independent of:
    - what a privacy adversary might or might not know
    - other sources of information such an adversary may or may not have access, at any given time, even in the future

→differential privacy is future-proof and adversary-agnostic

1.3. What Differential Privacy Doesn't Do

- Differential privacy is the wrong tool to use to study outliers, as it hides their presence or absence.

- It is not the right tool for analyzing small datasets.
  - Adding or removing an individual from a small dataset is more likely to significantly change the value of the statistical estimator as compared to when the dataset is large.

- Depending on the choice of epsilon, differential privacy may hide important differences in small populations or subpopulations of interest. While this may be construed as a limitation, it is actually a feature.
  - Despite lower accuracy, differential privacy is indeed working as intended – hiding the presence or absence of an individual over the outcomes of the analysis.

1.4. Reasoning about the Privacy Provided by Differential Privacy Implementations

- ➔ The ϵ is not sufficient to measure the privacy of a differentially private system
- ➔ Also important are the implementation choices below:

The potential impact of implementation choices on the privacy provided by a differentially private implementation:

1.4.1. The Importance of ϵ for Assessing Privacy Quality:

- ➔ Not all data are of equal sensitivity
- ➔ Not all data usage is of equal value

The parameterization of differential privacy allows different decisions about how much privacy to ensure based on data type

1.4.2. Limitations on Total Privacy Loss

- ➔ Overly accurate answers to too many questions can destroy privacy
- ➔ To protect against this, differential privacy in practice requires a privacy budget to limit harm
- ➔ This budget is the (declared) maximum acceptable privacy loss allowed before no more queries are permitted.
- ➔ Once this budget has been exhausted, the dataset is retired, never to be queried again

1.4.3. At What Point is Differential Privacy Applied?

- ➔ Two differentially private systems can provide radically different privacy protection depending upon when differential privacy is introduced

Example: Applying differential privacy to the raw data and computing on the result and computing the result and then applying differential privacy

1.4.4. Privacy Over What.

- This question is the granularity at which differential privacy is applied.
- Consider movie recommendation system that protects the privacy of the viewer at the single-movie level
  - Meaning that, for any given movie the system will provide differential privacy for individual movie-watching events
- This provides less protection than a recommendation system that provides differential privacy for the entire movie watching history of the user

1.5. The Meaning of Epsilon - uncertainty in choosing ϵ

1. Lack of clarity of what is the right degree of privacy loss in a given context.
2. No formula for determining, for a given privacy-utility trade off, what is the judicious choice of ϵ

Thus, while we can cap the privacy loss of a given algorithm, if we do not know how small an ϵ is possible

Case Studies - Differential Privacy in Practice

There are many implementation decisions that alter privacy provided by a differentially private system even with the same ϵ:

2.1. Impetus for Differential Privacy

➔ Learn organizations’ needs for privacy protection, and what spurred them to choose differential privacy at all

➔ Results varied in organizations but many chose differential privacy because it has both privacy and strong utility guarantees

2.2. Data Characteristics & Avenues for Privacy Loss

➔ The avenues for privacy loss involve domain-specific attributes.

Example: In operating system work, “as you bring in more and more types of data, even as you try to make it impossible to link, now you can use correlation through metadata.”

2.3. Granularity of Protection

- ➔ Differential privacy over what?
- ➔ Does the implementation seek to protect event level data, individual data, or something else?
- ➔ Generally, organizations were aware of the granularity of data they were trying to protect.
- ➔ The granularity used in one implementation may be completely inappropriate in another.

2.4. Limitations on Privacy Loss

- ➔ Determining an appropriate privacy budget is context specific.

One interviewee noted, “[We are] thinking about on the order of 1,2, and 4 [over a year]” while another interviewee, from a different organization, stated the “overall budget is 0.1.”

2.5. Algorithm Specifics.

- With respect to techniques for determining ϵ, a wide range of methods being utilized.
- The approaches ranged from fairly sophisticated to, more-or-less, a random choice.

Some sophisticated methods used:

- Experimental methods:
  - Simulation was used to find a value of ϵ that provided sufficient utility to meet an institution's specific product and business goals
- Others used a value of ϵ that was rooted in a previous differential privacy implementation
  - This involved computing a new value of ϵ to meet the same utility produced from a prior differential privacy implementation

4. The Epsilon Registry

- ➔ Authors propose the creation of the Epsilon Registry – a publicly available communal body of knowledge about differential privacy implementations that can be used by various stakeholders
- ➔ Firms using differential privacy would disclose the choice of ϵ, as well as several critical related policies and practices. This information would be publicly available.
- ➔ The knowledge shared through the Epsilon Registry will advance privacy in two ways:
  - ◆ It will support the identification of judicious parameter ϵ and other privacy preserving design choices and best practices among practitioners
  - ◆ By enabling stakeholders to compare the quality of privacy offered by various firms, create pressure on firms to reduce privacy losses while assuring utility gains.

Thank you!