

# Applications and Implementation of Differential Privacy on Statistical Databases

Jonathan Sumner Evans\*, Victoria Girkins<sup>†</sup>, and Sam Sartor<sup>‡</sup>

Department of Computer Science, Colorado School of Mines

Golden, Colorado

Email: \*jonathanevans@mines.edu, <sup>†</sup>vgirkins@mines.edu, <sup>‡</sup>ssartor@mines.edu,

**Abstract**—Statistical databases have long been known to be vulnerable to inference attacks. We describe the nature of inference attacks and analyze one method of mitigating these attacks called Differential Privacy. We discuss two mechanisms for differential privacy. We then explain our implementation of a simple Differential Privacy statistical database and the experiments we performed on the database. We then analyze how well our implementation achieves the objectives of Differential Privacy databases. We conclude by calling on industry to create differential privacy implementations that can be used in general applications.

## I. INTRODUCTION

Statistical databases are designed for statistical analysis of the datasets contained within the database. One of the desired properties of statistical databases is that nothing about an individual should be learnable from the database that cannot be learned without access to the database [4]. In other words, gathering information about the records underlying the statistics using allowed queries on the database should be impossible.

Without any protections, well-crafted queries to a statistical database can implicitly reveal private information. This attack is known as an *inference attack*. To prevent these attacks, the key case to consider is that of two databases which differ by only a single row. Consider, for example, a database which associates people with their incomes. The query

---

```
SELECT SUM(income) FROM table
```

---

seems harmless at first. But if a new record were added and we ran the query again, the difference between the two results represents the exact income of the new entry. If the attacker knows the identity of the person who was added, the privacy leak becomes even more severe.

Of course, an attacker may not have the luxury of always knowing when a new record will be added to the database, and of timing his queries accordingly. He may, however, achieve the same results by narrowing down his query with conditionals.

The query

---

```
SELECT SUM(income)
FROM table
WHERE zip != 80127
```

---

may be sufficient in a small enough database, or more conditionals may need to be added to ensure the queries jointly isolate a single row. However, since the danger occurs when information about a single entry is returned, regardless of how the attacker managed it, we will assume for simplicity's sake that the database is queried before and after the addition of a single row.

### A. Previous Attempts at Anonymizing Statistical Data

Many attempts have been made to privatize data in the past and many are still in use today. However everything from simply removing columns containing personally identifiable information to advanced techniques such as *k*-anonymity and *l*-diversity have been shown to be vulnerable to attack [2]. *K*-anonymity, for example, does not include any randomization, attackers can still make inferences about data sets [1].

### B. Differential Privacy

In 2006, Cynthia Dwork proposed a new technique to reduce the risk of a successful inference attack called *Differential Privacy* [6]. The goal of Differential Privacy is to maximize the statistical accuracy of queries on a statistical database while also maximizing the privacy of the individual records. It does this by introducing noise into the dataset. This randomness reduces the likelihood of obtaining meaningful information about the individual records in the database.

The real power of Differential Privacy is when it is used in conjunction with a privacy bound,  $\epsilon$  as described in Section I-B1. The methods for generating random noise are discussed in Section I-B2. Details about how the one of the mechanisms is used to ensure differential privacy are explained in Section I-B3.

1) *Epsilon-Differentially Private*: One useful way of discussing Differential Privacy is to talk about what it means for a database to be *epsilon-differentially private*. A database falls under this category if it is private as bounded by a chosen value epsilon. As opposed to the mathematical definition, here we give a more intuitive definition as stated by Atokar [2]:

This [epsilon-differentially private] can be translated as meaning that the risk to one's privacy should not substantially (as bounded by epsilon) increase as a result of participating in a statistical database.

For the sake of rigorousness, the formal mathematical definition of epsilon-differential privacy described by Dwork [4] follows.

**Definition 1.** A randomized function  $\mathcal{K}$  gives  **$\epsilon$ -differential privacy** if for all data sets  $D$  and  $D'$  differing on at most one row, and all  $S \subseteq \text{Range}(\mathcal{K})$ ,

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D') \in S] \quad (1)$$

For now, all that concerns us about this equation is that  $D$  and  $D'$  are the above-mentioned databases differing by a single row, and  $\mathcal{K}$  is our mechanism for adding noise drawn from our distribution as described in Section I-B2.

2) *Differential Privacy Mechanisms*: There are two primary Differential Privacy Mechanisms: the Exponential Mechanism and the Laplace Mechanism. The Exponential Mechanism should be used on categorical, non-numeric data. For example queries of the form “what is the most common eye color in this room” should be protected using the Exponential Mechanism [2]. The Laplace Mechanism should be used with non-categorical, real-value data [5]. Because of the nature of our data, our primary focus for this project is the Laplace Mechanism.

Before we rigorously examine the mathematics, let us review our inputs and our goals:

**Inputs** A statistical database with sensitive information, an attacker poised with a malicious query, and a desired epsilon, or privacy parameter.

**Goals** To add enough Laplace-distributed noise to the results of database queries that our database

is mathematically considered differentially private.

Mathematically, the Laplace Distribution is defined as follows:

**Definition 2.** The **Laplace Distribution**,  $\text{Lap}(\mu, b)$ , also known as a symmetric exponential distribution, is the probability distribution with p.d.f.:

$$P(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \quad (2)$$

where  $\mu$  is the location of the center of the distribution and  $b$  is the scale parameter.

Figure 1 is a visualization of the p.d.f. for a few values of  $\mu$  and  $b$ .

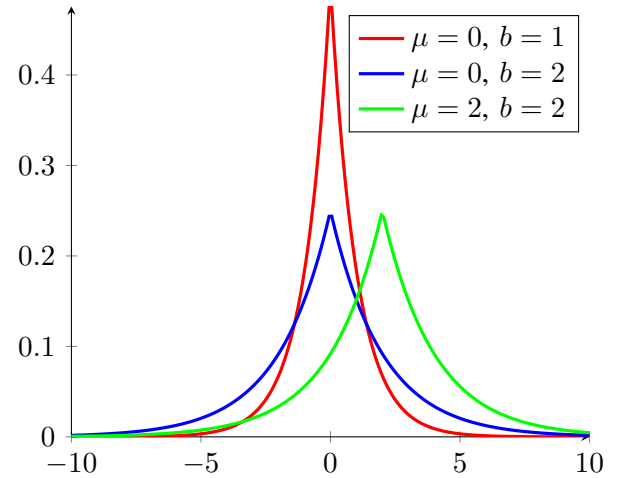


Figure 1. Probability Density Function for the Laplace Distribution

For us,  $\mu$  is 0 so we can use a variant of the Laplace Distribution called the Zero-Centered Laplace Distribution which has only one parameter,  $b$ , the scale.

To draw a value from the Zero-Centered Laplace Distribution efficiently, we use the random variate for the Laplace Distribution which is defined as follows:

**Definition 3.** The **Laplace Random Variate** is

$$X(u) = \begin{cases} -b \log |u| & \text{if } u > 0 \\ b \log |u| & \text{if } u \leq 0 \end{cases} \quad (3)$$

where  $u$  is uniformly distributed in the range  $(-1, 1)$ .

3) *Using the Laplace Differential Privacy Mechanism*: In choosing the value for  $b$ , it makes sense that we should consider  $\epsilon$ , which quantifies the level of privacy

Table I  
COLUMNS IN OUR SAMPLE DATASET

Field	Column Name	Generation Method
Name	name	the names Python library
Age	age	uniformly distributed between 18 and 65
Income	income	distributed log-normal
Zip Code	zip	from a random set of ~40 five-digit numbers
Net-worth	net_worth	from age, income, and a random offset

we desire. Another factor comes into play, however: the “sensitivity” of the database. This is defined as the maximum difference between the results of any query run on two databases which differ by only one row, as described above [2]. We shall assume

---

```
SELECT SUM(income) FROM table
```

---

is the only query that this database will accept, so the “sensitivity” of our database is clearly the highest income in the table, since this value will be the difference between two queries run on the database before and after the record associated with this income is added.

So, we must add noise while considering both  $\epsilon$  and the database’s above-defined sensitivity. In fact, in a paper by Dwork [3], it was proven that choosing the “scale” parameter ( $b$ ) of our Laplace distribution according to this equation

$$b = \Delta f / \epsilon \quad (4)$$

where  $\Delta f$  is sensitivity, our database will mathematically satisfy epsilon-differential privacy. That is, we should add noise to the data drawn from a 0-centered Laplace distribution with scale parameter  $\Delta f / \epsilon$ .

The one further consideration we must make is the possibility of multiple queries to the database. The Laplace distribution is symmetrical, so if unlimited queries were allowed, the attacker could simply run his query many times and take the average to extract sensitive data. So we must limit the number of queries allowed, and fortunately the calculation is linear. Assume we draw all noise from the same distribution (it would be possible to vary it but this unnecessarily complicates things), and that this distribution is Laplace, 0-centered, with a privacy parameter  $\epsilon_i$  (that is, with scale parameter  $\Delta f / \epsilon_i$ ). If we limit learners to  $k$  queries, our overall privacy budget is  $k\epsilon_i$ . This privacy budget is our actual epsilon; as stated by Atcock, “it reflects the maximum privacy release allowable for the total query session” [2].

So to sum it up practically: if we wish to make our database epsilon-differentially private as bounded by

some value epsilon, our database has sensitivity  $\Delta f$ , and we wish for learners to be allowed  $k$  queries before being locked out, we should add a random variable to each query, drawn from the 0-centered Laplace distribution with scale parameter  $k \times \Delta f / \epsilon$ .

Of course, the lower of a privacy budget we allow, the noisier the data returned by the queries will become [2]. If our queries become useless due to the noise in the data, we may choose to raise the privacy budget—that is, to increase the value of epsilon—or to further limit the number of queries each user is allowed—that is, to lower the value of  $k$ . As with many issues in information security, this is a balancing act between security and usability, and we should consider the nature of the database, its contents, and its purpose when choosing where to draw this line.

\*\*\*

For our final project, we implemented and analyzed a differential privacy database which can perform a limited set of aggregate functions on a dataset. We describe our approach and experiment in Section II. The results of our experiment are summarized in Section III and we discuss some of the limitations of our implementation in Section IV. In Section V we describe some of the potential future work in the field of differential privacy. We conclude in Section VI and call on industry to create effective, easy-to-use differential privacy implementations for inclusion in general industry applications.

## II. APPROACH AND EXPERIMENT

To perform our experiment, we generated a large sample dataset suitable for statistical queries (Section II-A). Then we built a simple Differential Privacy database as detailed in Section II-B and crafted a set of queries with which to attack our database as described in Section II-D.

### A. Dataset

To test our database, we created a dataset with 10,000 data points (rows). The columns for each row are listed in Table I. The dataset was generated using a `python3`

program and is output to a comma separated value (CSV) file. We chose to use a CSV file for simplicity, however future iterations of our project could use a more sophisticated data store.

We chose the generation methods to roughly model the distributions of names, age, income, zip code, and net worth of the US population. For example, the income is distributed log-normal which approximates the distribution of incomes in the United States.

### B. Database Implementation

Our actual database implementation was written in Python to speed up development time. If we were implementing this in an application that needed high performance, we would use a compiled language such as C or Rust. To help us test our implementation, our database allows the user to enable and disable Differential Privacy. When Differential Privacy is enabled, the user is also able to configure  $\epsilon$  and query limit.

The database reads the data from a CSV file which is specified by the user and then allows the following aggregate, or summary, functions on the data:

- **SUM**: the numerical sum of the values in a column
- **COUNT**: the number of rows
- **MIN**: the minimum numerical value in the column
- **MAX**: the maximum numerical value in the column
- **MEAN**: the average value in the column
- **VARIANCE**: the statistical variance of the column
- **SD**: the standard deviation of the column

Our database implementation allows the user to specify a `WHERE` clause. Currently, we implement this as a directly evaluated Python expression which is executed via the `eval` function in Python. This is a code injection vulnerability, but our goal is to study Differential Privacy, not code injection prevention. Future iterations of our project could wrap this functionality and prevent code injection attacks.

The statistic and  $\Delta f$  are calculated using the functions described in section II-C. Noise is added according to the Laplace Distribution as described in Section I-B3. This privatized result is then displayed. For testing purposes, we also output the actual statistic, but this can easily be removed in future implementations.

### C. Summary Functions

The statistic  $f$  over values  $D$  is  $f(D)$ . The sensitivity of the statistic is  $\Delta f(D) = |f(D) - f(D')|$  where  $D' \subset D$  and  $|D'| = |D| - 1$ .  $D'$  is always chosen to maximize  $\Delta f(D)$ .

Generally,  $\Delta f(D)$  can be found by iterating over all possible  $D'$  sets to find the maximum  $|f(D) - f(D')|$ . This often has  $O(n^2)$  complexity and is very slow on large datasets. To allow for more performant queries, our implementation individually defines  $f$  and  $\Delta f$  for each of the following summary functions.

1) *Count*: Removing one row from the dataset decreases the row count by one.

$$\begin{aligned} \text{count}(D) &= |D| \\ \Delta \text{count}(D) &= 1 \end{aligned} \quad (5)$$

2) *Sum*: The largest element in a set has the most influence on the set's sum.

$$\begin{aligned} \text{sum}(D) &= \sum D \\ \Delta \text{sum}(D) &= \max(D) \end{aligned} \quad (6)$$

3) *Mean*: Mean is most changed by the addition or removal of elements with a high variance from the mean. So, let  $d$  be some  $d \in D$  where  $|d - \text{mean}(D)|$  is maximal.

$$\begin{aligned} \text{mean}(D) &= \frac{\sum D}{|D|} \\ \Delta \text{mean}(D) &= \left| \frac{\sum D}{|D|} - \frac{\sum(D) - d}{|D| - 1} \right| \end{aligned} \quad (7)$$

4) *Maximum*: Let  $l_n$  be the  $n^{\text{th}}$  largest element in  $D$ . Since  $\max(D)$  is dependent only on  $l_0$ ,  $D'$  must not contain  $l_0$ .

$$\begin{aligned} \max(D) &= l_0 \\ \Delta \max(D) &= l_0 - l_1 \end{aligned} \quad (8)$$

5) *Minimum*: Let  $s_n$  be the  $n^{\text{th}}$  smallest element in  $D$ . Since  $\min(D)$  is dependent only on  $s_0$ ,  $D'$  must not contain  $s_0$ .

$$\begin{aligned} \min(D) &= s_0 \\ \Delta \min(D) &= s_1 - s_0 \end{aligned} \quad (9)$$

6) *Variance & Standard Deviation*: We use the brute force method for these functions. A more efficient method certainly exists, but we have not yet implemented it.

## D. Crafted Queries

### III. RESULTS

With Differential Privacy disabled.

---

```
use dp (y/n): n
database csv: bigdata.csv
=====
summarize field: age
summary function: sum
where: zip == 31643
10412
=====
summarize field: age
summary function: sum
where: zip == 31643 and name != "Abel Woods"
10379
```

---

From these queries, we can determine the total age of the people in ZIP code 31643 is 10412. However, we can also infer that Abel Woods is  $10412 - 10379 = 33$  years old. This is a violation of privacy.

With Differential Privacy enabled.

---

```
use dp (y/n): y
epsilon: 5
query limit: 2
database csv: bigdata.csv
=====
summarize field: age
summary function: sum
where: zip == 31643
10409
=====
summarize field: age
summary function: sum
where: zip == 31643 and name != "Abel Woods"
10435
```

---

From these queries, we can determine that the total age of the people in ZIP code 31643 is 10409, a 0.2% error. If we attempt to use these queries to find the age of, Abel Woods, we will get  $10409 - 10445 = 49$ , a 48% error. In this way, Abel Woods' privacy is protected.

### IV. LIMITATIONS

### V. FUTURE WORK

Differential Privacy is a relatively new method of protecting data in a statistical database. As such, there are not many implementations of it outside of academia and large companies like Microsoft and Apple. If differential privacy was adopted by a larger percentage of the industry, more statistical databases with sensitive information would be resistant to inference attacks.

Software engineers have a great opportunity to create implementations of differential privacy suitable for use

in general industry applications. Such an implementation could be a library which other programmers could include in their project using dependency managers such as `pip`, `npm`, or `cargo`. The creation of these general-purpose libraries could help make differential privacy more accessible and cost effective for industry professionals. This would help increase adoption of differential privacy and result in an overall improvement of user's privacy in the industry.

### VI. CONCLUSION

### REFERENCES

- [1] Charu Aggarwal. "On k-Anonymity and the Curse of Dimensionality." In: *VLDB 2005 - Proceedings of 31st International Conference on Very Large Data Bases*. Vol. 2. Jan. 2005, pp. 901–909.
- [2] Atokkar. "Differential Privacy: The Basics". In: (2014). URL: <https://research.neustar.biz/2014/09/08/differential-privacy-the-basics/>.
- [3] Cynthia Dwork. "A Firm Foundation for Private Data Analysis". In: *Communications of the ACM* (Jan. 2011). URL: <https://www.microsoft.com/en-us/research/publication/a-firm-foundation-for-private-data-analysis/>.
- [4] Cynthia Dwork. "Differential Privacy". In: *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*. Vol. 4052. Springer Verlag, July 2006. ISBN: 3-540-35907-9. URL: <https://www.microsoft.com/en-us/research/publication/differential-privacy/>.
- [5] Quan Geng and Pramod Viswanath. "The optimal mechanism in differential privacy". In: *2014 IEEE International Symposium on Information Theory* (2014). DOI: 10.1109/isit.2014.6875258. URL: <https://arxiv.org/pdf/1212.1186.pdf>.
- [6] Michael Hilton. *Differential Privacy: A Historical Survey*. URL: <http://www.cs.uky.edu/~jzhang/CS689/PPDM-differential.pdf>.