Diff: An Open Platform for Differentially Private Datasets

Juan Ortiz, Arlene Siswanto, William Rodriguez
{dkogut, juanmoo, siswanto, wrod}@mit.edu

https://diffdatasets.herokuapp.com/

May 16, 2018

1 Abstract

As the world of big data continues to expand, the digital footprints identifying individual users will continue to grow. In order to work with and analyze this data, however, valid concerns should be made about the privacy of these users. As a tangible solution, differential privacy presents itself as a feasible yet nontrivial challenge to implement concept for the average data analyst. Through *Diff*, we aim to create an easy-to-use platform for the average user to upload their datasets and yield differentially private query results. We expand upon knowledge of differential privacy and take advantage of the Multiplicative Weights Exponential Mechanism (MWEM) within a web application framework.

2 Introduction

As consumers and the organizations that serve them adopt more technological services and systems, they consequently create and collect more information. We find ourselves thus in this ever-evolving world of "big data." This contemporary information-rich ecosystem opens many avenues for novel technologies to take advantage of large datasets that document behaviors and interactions. For example, machine learning algorithms and applications of artificial intelligence can become more feasible and provide more accurate results when there is a wider set of training data available to work with.

However, with such wide availability of data, there are many concerns over the privacy and security of the information being collected and used for analysis by

researchers. Although people may want to publish large data tables to share interesting and/or important results with larger communities, they may be restricted in doing so since the numeric data left behind after removing personally identifying data (e.g. names, addresses, social security numbers) can still be used to trace back to individual users.

As a novel approach to this challenge, *differential privacy* has come to prominence. Mathematically, the result of an algorithm A is ϵ-deferentially private if for two datasets D_1 and D_2 that differ by exactly one element:

$$Pr[A(D_1) \in S] \leq e^{\epsilon} * Pr[A(D_2) \in S]$$

More intuitively, differential privacy describes that given the result of an algorithm on a dataset, it is hard to identify or obtain information pertaining to an individual entry of this dataset from simply looking at the result.

Having differentially private algorithms that can securely release results from large datasets with sensitive information can thus provide a tangible solution to the problem of individual privacy within the frame of big data. However, differential privacy can be challenging for individuals to attain. There are few, if any, publicly available platforms on the internet that deliver differential privacy for manipulations of uploaded datasets.

To that motive, we created *Diff* – an open platform that allows anyone upload their pre-processed datasets and execute queries with differentially private results.

3 Background

3.1 Other Tools and Platforms

Several researchers have conducted previous work in creating tools that aim to provide the ability to query databases and manipulate their contents in a differentially private manner. Although these initiatives have lowered the barrier of entry in implementing data distribution techniques with strong privacy guarantees, at the current state, the average user still encounters hurdles in incorporating differential privacy.

We will now proceed to a presentation of a few tools that aim to provide such services, while briefly explaining their contributions. Lastly, we will demonstrate how *Diff* can fill in some of the gaps they leave.

3.1.1 Microsoft PINQ

At its core, PINQ is a platform that provides a programmatic interface to unmodified data through a SQL-like language [3]. PINQ imposes limitations to what can be learned about the data by performing queries that provide formal guarantees in the form of differential privacy to the users of the platform. It does this by doing the following: First, it allocates a privacy budget that is fixed and correlated to how many clients of the platform are allowed to learn about a particular dataset. This budget is then consumed through queries. After the budget has been exhausted, any further queries are left unanswered.

PINQ's greatest contribution is the creation of a framework that permits differentially private interactions on datasets managed by developers that do not require expert knowledge on the topic. It achieves this by abstracting away the complexity that comes by having to enforce privacy away from developers, allowing them to solely focus on the logic of the application.

Although PINQ undoubtedly lowers the barrier of entry of using differentially private methods from solely experts to any developer, it is still unrealistic for an average user to take advantage of differential privacy. On the other hand, Diff, although not as powerful or expressive, allows users to interact with it in the form of a simple web application that does not demand extensive technical knowledge.

3.1.2 Airavat MapReduce

Airavat is a MapReduce based platform that provides strong privacy and security guarantees for distributed computations [4]. Its objective enables the usage of untrusted or not thoroughly inspected code to analyze sensitive information and generate aggregate computations that stem from input datasets. It does this without exposing information about particular entries on the set. By using mandatory access control in conjunction with added differential privacy in the form of added Laplacian noise, Airavat provides an end-to-end privacy guarantee.

Although the end goals of Airavat and *Diff* are similar in the sense that they attempt to simplify the process of applying differential privacy, the scope in which they try to achieve this objective is fundamentally different. While Airavat establishes a foundation on using arbitrary code to simply provide privacy guarantees in a distributed architecture, *Diff* aims to simplify the user experience such that the average user would be able to use it and take advantage of its benefits. To this extent, future iterations of *Diff* could try to incorporate systems like Airavat. Moreover, it could expand upon its scalability as well as the types of computations

it can handle.

4 Differential Privacy

4.1 Overview

The goal of Diff is to give providers of data a platform to easily share their sensitive data while ensuring that the privacy of the entries that compose their data-sets remains protected. To this end, we will use the framework of differential privacy to provide a concrete definition of what guarantees *Diff* offers to its users and their data.

Formally, a process $\mathcal{F}$ is said to be (ϵ, δ)-differentially private if the following holds true: For any two datasets that differ by exactly one element that is present in D_1 but not in D_2, the probability that the output of of the operation in the two sets is contained within a subset of the possible outputs of $\mathcal{F}$ differs by at most e^ϵ multiplicatively and δ additively. Alternatively, for all $S \subseteq Range(\mathcal{F})$

$$Pr[\mathcal{F}(D_1) \in S] \leq e^\epsilon \cdot Pr[\mathcal{F}(D_2) \in S] + \delta$$

Intuitively, this means that a process is differentially private if for any set of inputs, any element of the set does not affect the computation significantly. Here, the ϵ parameter serves as a measure of the levels of the privacy guarantee. The lower the value of ϵ, the greater the level of privacy offered. The parameter δ serves to relax the definition for computations in certain situations. For example, if one of the probabilities were zero, the multiplicative bound would not be achievable regardless of the value of ϵ even if the other probability is small and is not a problem in the scope of the problem.

We choose the definition of privacy given by the framework of differential privacy because of its great ability to generalize due to its strong nature. One of its main advantages is that it does not rely on assumptions made about the capabilities or knowledge of the adversary [4]. This means that if it is satisfied, it holds for any threats arbitrary in nature. Additionally, we can measure the effects of performing multiple operations because the composition of differentially private operations is also differentially private and at worst additive (i.e. ϵ and δ will at most be the sum of the respective values of the operations being composed).

4.2 General Privacy Concerns with Differential Privacy

Our service serves as a platform to host datasets for a public or restricted set of users. Inevitably, this requires trust in our platform to host the data. Thus, it should be assumed that data uploaded to our web app will be encrypted and that the burden of protecting the privacy in the true dataset lies with the differentially private results that we would publish from queries sent to our platform. An inherent flaw in differential privacy is the susceptibility of the privacy of data to subsequent queries. In addition, there is a direct trade-off in the privacy of data when considering the accuracy of the data. This trade-off is concisely represented by the epsilon parameter. We offer a dynamic value of the epsilon parameter per dataset to allow the dataset owner to define the trade-offs themselves according to their own needs. Differential privacy also only applies to queries that aggregate and do not release the data themselves. For example, queries that simply select columns are not included in our service since that revelation would effectively break differential privacy invariants that exclude microdata release.

4.3 MWEM

Our choice of a differentially private algorithm should to take into account the trade-off between privacy and accuracy. We know that releasing datasets to a public or semi-public audience will inevitably reduce the privacy when subsequent queries are ran on a dataset and the results are conglomerated. Thus, our choice in a differentially private algorithm will need to be resilient to subsequent queries while still giving reasonably accurate results. The Multiplicative Weights and Exponential Mechanism algorithm fits these needs perfectly. The general strategy of this algorithm is to randomize a true dataset to some initial distribution that reasonably mocks the statistical behavior of the true dataset. This initial distribution is very inaccurate and thus very private. When queries are executed, they run on the new "synthetic" dataset and return the appropriate result. Afterwards, the synthetic dataset is updated to better reflect the true behavior of the original dataset. Essentially, this algorithm provides a dataset that mocks the original dataset in a manner that gives a more accurate view of the data after each subsequent query. There are two components of this algorithm that distinguish it from other differentially private algorithms.

The first is the randomization strategy. In order to provide a probability guarantee to the privacy of a dataset, we need to randomize the data in a way that preserves its general properties so as to maintain some accuracy but add randomized noise to protect the specific elements. The exponential mechanism is a tool used in MWEM that assigns a score to each possible result of a query and returns a

particular result according to its score (results with a higher score are returned with exponentially higher probability than those with lower scores). This score would favor results that are closer to the result of a query on the original dataset, and so this mechanism, if scaled correctly, returns a result that is approximately the most accurate. This mechanism itself maintains epsilon differential privacy. The Laplacian mechanism is a tool that adds noise to each element in the dataset so that the approximate mean of the dataset is maintained but each element is different enough so that the true value cannot be learned by a set of queries. This mechanism is also epsilon-differentially private.

The second is the multiplicative weights method. This is the method that changes the synthetic dataset to better reflect the true dataset without compromising privacy. It does so by weighing each element by its sensitivity (an element is considered sensitive to a particular query in a given dataset if its value with drastically change the result of the query) and updating the synthetic dataset with emphasis on these elements.

4.4 Optimizations for Privacy

As any other approach that utilizes differential privacy, using MWEM to protect entries in a dataset is vulnerable to attacks of repeated queries. Even if a process provides a set guarantee on the upper limit of how much can be learned about the dataset through its execution, using enough executions, more and more information can be obtained. In [5], Dinur and Nissim conclude that a dataset can be decoded with a linear number of queries in the size of the dataset.

One common way to deal with this in complex systems, with the ability to express a variety of different procedures such as Airavat and PINQ, is to allocate a privacy budget whose size is inversely related to the level of privacy desired. As described in Section 3.2.1, each execution consumes a portion of the allocated budget until it is depleted in its entirety. In the case of *Diff*, the restricted repertoire of operations that can be applied to data allow us to take a rather simple approach. If there is a global cache of queries per dataset, frequently-run queriescan use the result of the cached query instead of producing another differentially private result. This has two-fold benefits: query time is faster since the pipeline of the query does not have to be executed and no additional information is re-leased, thereby improving privacy concerns. This optimization can be exploited by changing the query slightly to avoid receiving a cached response. This exploit can be solved on a case-by-case basis. For example, if this were a query that tookthe average of a particular set of values in a column, we could return a query that took the average on a similar range of values in the same column. If the query summed a particular range of

values, we could do a similar approach by returning the result of a previous query and adding or subtracting the sampled differentially private mean according to the change in span of the range.

In addition to creating a "smart query cache", we can improve privacy by restricting the number of queries per user. This restriction can be set by the owner of the dataset as a parameter at upload time. By restricting the number of queries and bounding the number of users on a dataset, we effectively can construct a value of the epsilon parameter since these two values would be the only parameters informing the privacy of a particular dataset.

4.5 Limitations

Since the privacy of a dataset can be compromised from a series of queries, the amount of times a particular user can query a particular dataset will have to be limited. Malicious users who aim to curb this restriction by making multiple accounts present a clear privacy concern. This can be fixed in two ways: either make it so the uploader of the dataset grants permissions to various accounts or the accounts are linked to a specific and verifiable user.

5 Diff

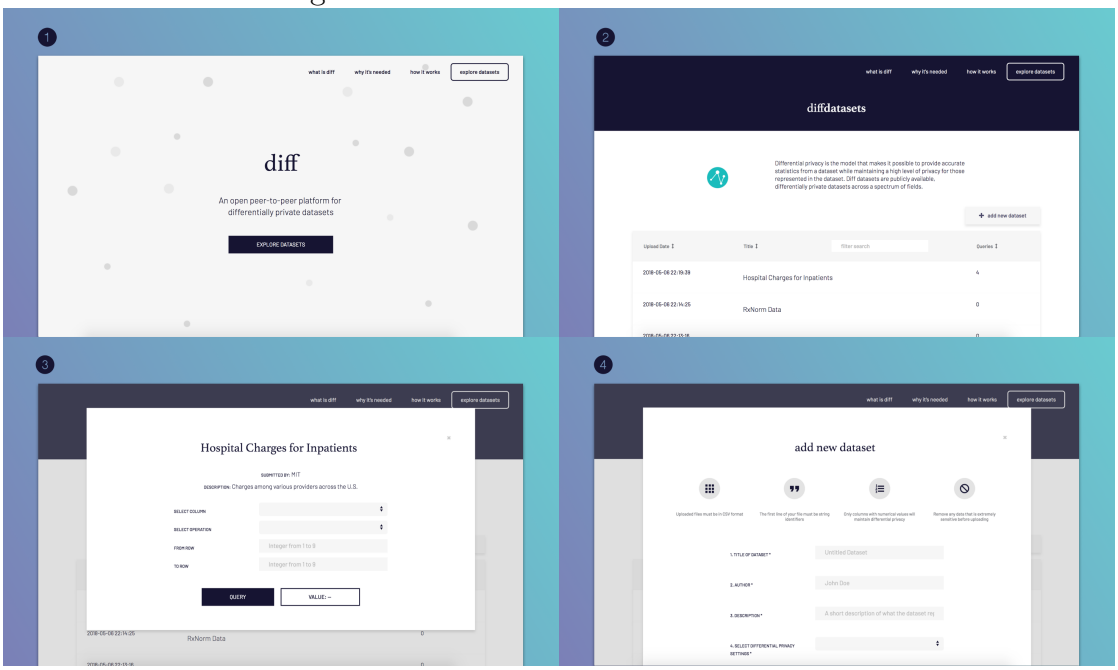
Our product, *Diff*, is an open web application that allows users to (1) obtain differentially private responses to queries on available datasets and (2) upload their own datasets to the platform to contribute to openly accessible data. Target users of *Diff* are the researchers, data scientists, and holders of data who need or own datasets useful for research and analysis. These users may not necessarily know what differential privacy is. For this reason, we created *Diff* to be explanatory and user friendly to help visitors to the website understand differential privacy and its potential use cases so they are informed when sharing datasets.

5.1 Our Application

As seen on Figure 1, there are two main components of our website. The homepage gives a quick introduction to *Diff*. The page with datasets allows users to search through available datasets, make specific queries to chosen datasets, then add their own datasets to the platform.

1. The homepage gives a brief overview of what *Diff* is, defines the motivations for creating the platform, and specifies several use cases for which

Figure 1: Screenshots of the Diff website



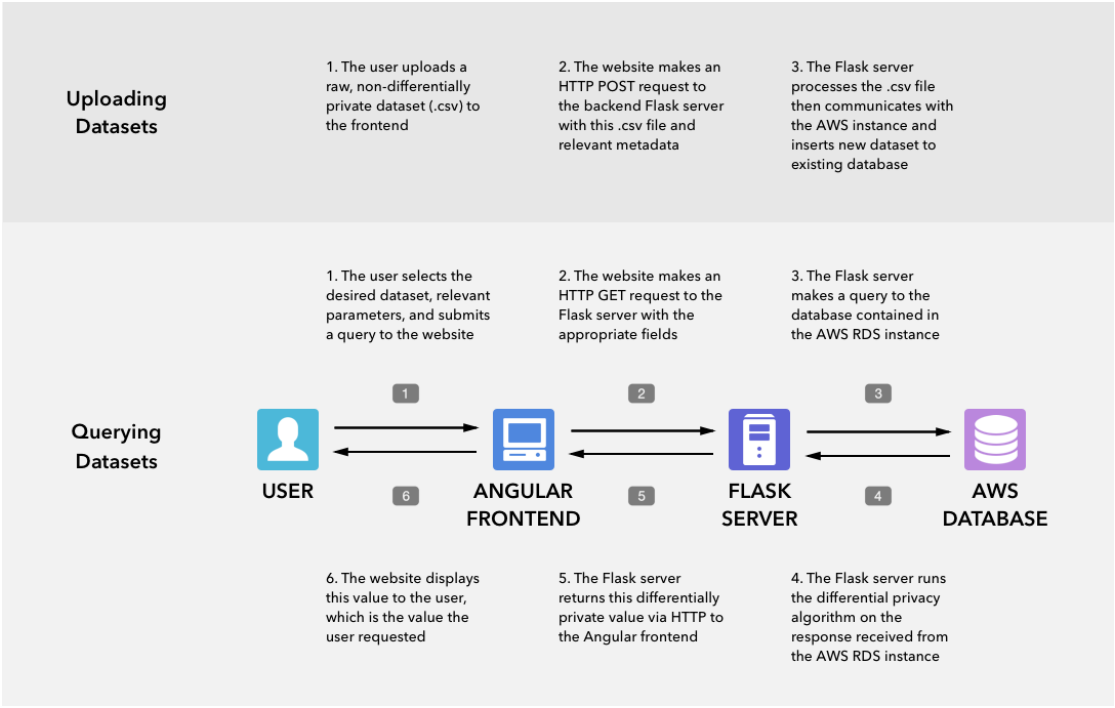
differentially private datasets are useful. It also describes our algorithm for differential privacy as described in Section 4.

2. The datasets page displays all available datasets. The user can filter datasets by name to find relevant material, and can sort through datasets by upload date, name, and popularity.

3. After selecting a desired dataset, users can query that dataset by selecting a relevant column, an aggregate operation (max, min, average, sum), and relevant rows.

4. Holders of data must follow certain constraints to add their dataset to our platform. First, they must upload their data in the format of a CSV file. All values except the first row (columns) must be numerical because we cannot make strings differentially private. In addition, any sensitive identifying data that should not be queried should be removed beforehand. Along with this CSV file, then, those uploading should include a title, write a short description, and specify a differential privacy setting. Because there exists a tradeoff between privacy and accuracy of data, uploaders can determine whether privacy or accuracy is more important to them based on the sensitivity of the content in their data.

Figure 2: The user flow for uploading and querying datasets



5.2 How It Works

There are four main components involved in our application. There is a **user** who interacts with our website. This website serves as an Angular **frontend** that is deployed to Heroku. When a dataset is queried or added, the frontend communicates with a **backend** Flask server. This Flask server communicates with the **AWS database** that stores datasets and can answer relevant queries.

5.2.1 Uploading a Dataset

As shown in Figure 2, the process for uploading a dataset is simple.

1. A user uploads a raw dataset that follows specified constraints, listed in Section 5.1.4.

2. The frontend makes an HTTP POST request to the backend Flask server with the uploaded CSV file and other information, such as the title, description, and differential privacy setting.

3. The Flask server processes the CSV file and communicates with an AWS instance. The new dataset is added to the database containing all datasets.

Note that *Diff* stores the raw, user-inputted version of the dataset, as differential privacy is attained from adding noise after a query is made, not from modifying the dataset itself.

5.2.2 Querying a Selected Dataset

After a user chooses a dataset he or she would like to query, the process for querying a selected database is also simple.

1. The user selects a desired dataset from the list of available datasets. After specifying the parameters described in Section 5.1.3, the user submits the desired query.
2. The website makes an HTTP GET request to the backend server with the provided parameters.
3. The Flask server interacts with the AWS RDS instance, and performs the differential privacy procedure as outlined in Section 4.
4. Following this same differential privacy procedure, the AWS instance returns to the server the results from the queried dataset. This value is differentially private and can be safely returned to the user.
5. The Flask server returns this differentially private value via HTTP to the frontend.
6. The website reveals the differentially private value to the user.

6 Examples and Use Cases

We created *Diff* to provide a wide range of people the ability to obtain differentially private query results from their own datasets. As such, there are many different use cases and examples for using *Diff*. We describe these below.

6.1 Health

A researcher at a hospital may own data quantifying immunological weaknesses that current patients of an experimental drug are using.

She wants to publish this information to a leading journal due to promising results, but does not want to violate the HIPAA law of privacy and security of health

information. The HIPPA law[1] describes that: *the Privacy Rule protects all "individually identifiable health information."* With this strict definition, it almost seems impossible to share health data without making it differentially private.

6.2 Finance

Financial institutions manage enormous sums of data every hour. Everything from client balance sum requests and valuations to bond and stock acquisitions yield prodigious spreadsheets, rife of sensitive numeric data. To maintain their reputations as trustful organizations, financial institutions have little leeway for any breaches of security. It is thus of paramount importance for these institutions to prevent any leakage of financial information of individuals or groups. Otherwise, they risk leaking financial information that can be used to steal money from bank accounts or even, in worse cases, identity theft.

For example, a bank may have ran a new pilot loan program with relatively few users. In the end, the changes in the mean credit scores for participants were promising, so they want to publish these results. Since relatively few users participated in the pilot program, sharing a mean credit score anonymously could still leave an adversary with the opportunity to trail back to the original user from which the credit came from. Differential privacy can be used here to protect the users in the pilot program.

6.3 Policy and Economics

The social sciences are a field in which data analysis and computation methods are becoming more and more common for inquiry.

A given use case can be data-based investigations within economics studies that later justify policy measures that become instituted into law. For example, some economists may be looking into exploring the change in average height of students in a school district after a new meal program was adopted. Releasing the data obtained from minors should be done in a way that anonymizes them and prevents individual pieces of data from ever personally identifying a single student. Here, differential privacy can help further protect student's privacy, adding another layer of security after anonymization.

[1] https://www.hhs.gov/sites/default/files/privacysummary.pdf

6.4 Whistleblower

Interestingly enough, differential privacy can bolster the efforts of whistle-blowers in tight contexts. For instance, there may be a journalist in a region with high censorship and dangerous retaliation practices.

A journalist may discover that workers in a factory have ethical, advertised working hours, but in practice, are threateningly pressured to exceed these hours. This journalist may then record the data timesheets of such workers. The journalist, however, does not want to simply release these data, because it strictly identifies individual workers that may then face disciplinary action and harsh punishment. When releasing statistics of the data, using differential privacy, the journalist can limit the chances for potential punishers to trace back the data to individual workers, helping bring justice to exploited people.

7 Discussion

Our design of *Diff* trades off the flexibility of the types of queries it can service for simplicity of both implementation and interaction. We take advantage of the restrictions imposed on valid queries to use caching to protect against attacks based on repeated queries. This, however, does not provide guarantees against cleverly combining the results from an arbitrary number of distinct queries like a protection based on a privacy budget would. This design choice inherently creates a prioritization of availability over security which could be later amended through the integration of a privacy budget protocol similar to that described for PINQ [3].

8 Conclusion

Along with the growth of big data and machine learning comes the increasing need to enforce mechanisms that maintain client privacy in datasets. We created a publicly available platform that allows the 'average' user to add differential privacy guarantees to their datasets. Using *Diff*, holders of data can upload their datasets for access to researchers and data scientists, who can then query the data in differentially private manner.

The creation of *Diff* is a step in the right direction in the creation of a tool that makes sharing data with strong privacy guarantees simple. Notwithstanding, there are a degree of ways in which it could be improved in both the security of its implementation and the scope in which it can be used. For the first concern,

future versions of *Diff* could benefit from access control and user verification. This would allow for results to datasets to be separated by verified users and therefore modify their responses based on only their own previous queries. For the second, restructuring *Diff* such that it uses Airavat to generate its aggregate computations would expand the types of operations it can perform as well as allow it to easily scale due to the distributed nature of MapReduce. In all, we have forged a path towards democratizing differential privacy.

References

[1] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy.* Now Publishers. 2014.

[2] M. Hardt, K. Ligett, F. McSherry, *A Simple and Practical Algorithm for Differentially Private Data Release.* 2012.

[3] F. McSherry, *PINQ: Privacy Integrated Queries.* Association for Computing Machinery, Inc. 2009.

[4] Indrajut Roy, Srinath Setty, Ann Kilzer, Vitaly Shmatikov, Emmet Witchel. *Airavat: secirity and privacy for MapReduce.* USENIX Association 2010

[5] I. Dinur and K. Nissim. *Revealing information while preserving privacy.* PODS, 2003.