Differential Privacy

# Introduction

Data Science’s core resource is data, and the need to gather more of it to improve the models is constant. But data is a double edged sword, it’s like the new oil: it’s very valuable but leaking it causes a lot of damage. Even if the damage is usually not environmental, it can cause a lot of damage to the individual at the source of the data. For example, the Equifax breach dating back to the summer 2017 leaked various key information (names, dates of birth, Social Security numbers and addresses) of about 143 millions Americans; these information are also used as unique identifier across many industries like banking or utilities. The immediate consequence of this breach has been the skyrocketing sale of Identity Protection services, but the long term consequences are still unknown. Another form of leakage, which is usually more insidious, is the possibility to infer the identity of individuals or entities through key markers in data aggregates. An example can be found with the Netflix Prize challenge: even though the data was anonymized, it was possible to deanonymize it with a high level of confidence by matching the data provided in this dataset (date of liking a certain movie, rating of the movie, etc.) with publicly available data.

This raise the question of how to protect individuals from potential data leakages in aggregate reports, what techniques are available to us today, what are their advantages and drawbacks, what can be done about these drawbacks and what other non-technical approaches are in use around the world nowadays.

We are going to focus in this paper on Differential Privacy, a mathematical method developed in the 2005 by Cynthia Dwork and Frank McSherry[[1]](#endnote-1), and largely in use by major technology companies like Apple and Google.

# Differential Privacy

Data aggregation can leak information from the original data, exposing the entities constituting the original dataset. This has been observed numerous time, and even though the complexity of these data models hasn’t ceased to increase, the method to avoid them hasn’t evolved much: peoples are in charge of picking and choosing which information is released to the public. As expected, this has become impossible with the scale and complexity of the existing datasets. More systematic methods are then needed to protect the users’ data, possibly even before it reaches the service provider to guarantee that even with the worst intention, it’s not able to deanonymize the user. This is particularly important in the case of sensitive data like rare conditions which can affect only a few users, drastically augmenting the confidence with which you can identify them based only on a few properties.

Differential Privacy approach to this problem is to introduce randomness even before it reaches the services (ex: Google or Facebook). This will create noise which will be canceled out during the aggregation of the large number of records, allowing the service providers to have accurate models while still protecting the users’ data and privacy. The quantity of noise introduced depends on the parameter ε. The lower this parameter is, the more noise will be introduced. A value lower than 1 is advised by the creators of the method, even if a higher value simply means that the quantity of data needed to deanonymize a user is reduced. This is then a balancing act between the quantity of noise in the data and the quantity of data the user is sending.

## Benefits

First of all, the method is fairly simple to put in place: neither does it require infrastructure changes since the data will be in the same format and only the values might be randomly different, nor does it require any different algorithm to create the models. The only required changes are on the users’ devices where you have to randomly alter data before sending it to the service provider. The method also relies on simple mathematical concepts, making it both easy to understand and easy to implement.

Second, the loss in precision is intrinsically minimal because of the nature of the random changes: since they are normally distributed, the error they introduce will be eliminated with the aggregation of the large number of records. This guarantee that there will be minimal losses in the precision of the created models, while still maintaining the privacy of the users.

Finally the method is mathematically proven, giving the service provider and the users a great confidence into their effectiveness, as long as the users trust the service providers to implement it correctly.

## Drawbacks

The first drawback comes directly from one of the benefits: since it’s only a mathematical method, the service provider is free to implement it however it wants. A few service providers have taken the approach of releasing their implementation in the open (i.e. Open Source); this greatly augment the trust with the users since they are now able to verify by themselves the quality and effectiveness of the implementation. This is not a silver bullet, as has been observed with OpenSSL – the most used SSL and TLS library - which was victim to major security issues despite the fact it has been open source for years and has been used by millions of service providers.

This approach might also go against the business goals of the company. In the case of Google for example, its main revenue generator is based on serving advertisement which is notoriously hungry on users data, and which effectiveness is highly dependent on the accuracy of the users data. The value of ε can then be set to any value over 1 which reduces the anonymity of the users. For example, Apple is notorious to values as high as 14, and this can have major implications for users of their HealthKit product:

*Say someone has told their phone’s health app they have a one-in-a-million medical condition, and their phone uploads that data to the phone’s creator on a daily basis, using differential privacy with an epsilon of 14. After one upload obfuscated with an injection of random data, the company’s data analysts would be able to figure out with 50 percent certainty whether the person had the condition. After two days of uploads, the analysts would know about that medical condition with virtually 100 percent certainty*

*(Frank McSherry, co-creator of Differential Privacy)*

Overall, this approach highly depends on the trust a user gives to the service provider, and it is up to the service provider to do everything in its power to deserve and conserve this trust.

# Trust

Since this method is fundamentally based on trust, this raises the question as to how service providers are held accountable for the trust we put in them,

1. [US 7698250](https://worldwide.espacenet.com/textdoc?DB=EPODOC&IDX=US7698250), Cynthia Dwork & Frank McSherry, "Differential data privacy", published 2010-04-13 [↑](#endnote-ref-1)