Project Proposal

Indian Institute of Technology Delhi

Differential Privacy

Name: Deepanshu Entry Number: 2019CS50427
Name: Pranjal Aggarwal Entry Number: 2019CS50443

1 Introduction

The idea of differential privacy is to make available for studying and processing a data distribution/dataset in the public or to relevant stakeholders, without allowing them to extract exact data points from it. Thus it is extremely useful in places where research over personal data of people needs to be done, but at the same time privacy of users must be respected. Thus in today's data driven world, since its important to protect the personal identity of people, while at the same time allow researchers (and companies) to build better algorithms for the society, differential privacy becomes an important topic of research. For example consider that Government of India decides to use the data collected through Ayushman Bharat Digital Mission [1], to provide better facilities, provide better insights to insurance companies, or provide data to researchers, it will become important, that none of the stakeholders get hold of personally identifiable information, but yet are able to do constructive research on the data. Various mathematical models have been developed to implement differential privacy. However, implementing such a system has many challenges, such as the designer should ensure that information doesn't leaks, and at the same time not so much noise should be added in the dataset, that it becomes useless.

2 Proposal

In this project we will implement a server-client based differential privacy system with special focus on health or taxation domain. We will research various types of differential privacy and then move on to researching various algorithms present in the literature and evaluate their pros and cons for our scenario. As a starting point we plan to use open-sourced algorithms provided by big Tech Companies such as Google [2] and IBM [3]. Then we will incorporate these algorithms into our systems, and do a critical anaylsis of what risk of leaking information, the harm and the benefit they brought.

Additionally, we plan to evaluate the effect of such systems in black-box methods such as deep learning. For this we plan to use open source library by Facebook, on some health or tax domain related dataset, which we will search on various competition and dataset hosting websites such as Kaggle [4]. Henceforth, we will show the trade-off between accuracy and effect of de-anonymization using differential privacy, with the aim to show minimum drop in accuracy with sufficient privacy.

References

[1] *Ayushman Bharat Digital Mission*. URL: https://abdm.gov.in/.

[2] *Google-Differential Privacy*. 2021. URL: https://github.com/google/differential-privacy.

[3] Naoise Holohan et al. "Diffprivlib: the IBM differential privacy library". In: *ArXiv e-prints* 1907.02444 [cs.CR] (July 2019).

[4] *Kaggle*. URL: https://www.kaggle.com/.