Meenakshi Nagarajan

Differential Privacy guarantee using RAPPOR

Introduction

Big Data, Machine learning and cloud computing in state-of-the-art world is being considered as the new opportunity for analytics [1]. Statistical data obtained from users are major source of information for machine learning & analytics operations. However, application of different statistical queries on the obtained data should not reveal the identity of the individual to preserve privacy. Publishing such statistical data on public platform (for eg. Kaggle) may lead to inference attack on database [2]. Any updates to the published data, may result in difference in the statistical parameters between old and new data. This may cause background knowledge attack. Often with manipulative queries and SQL injection, attacker can identify a single user from statistical database which may lead to inferential attacks. Hence publishing the user information by maintaining user's privacy is essential.

In this project, differential privacy, a mechanism to preserve user's identity by adding random noise factors to the statistical database, RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response) will be discussed.

Problem Discussion:

The database is assumed to be present on a trusted server with authorization and access controls on sensitive data. Statistical data which is to be published for common use will be taken from this trusted database by writing a query which will aggregate user statistics. To protect privacy of the user, a noise can be added at the end of the obtained results. But it can be easily identified by attackers if the noise is a fixed value. Adding Laplacian noise or gaussian distributed noise over the dataset are some well-known techniques which are used only in theory may improve privacy at the cost of accuracy.

Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) is a technique which improves accuracy by preserving crowdsourced information obtained from surveys, questionnaires, feedback etc. Also, on randomizing the response from participants and approximating the probability of correctness in response, privacy of the user is preserved. Applying this technique on any data would significantly contribute to accuracy and privacy of published information [3].

Mental disease data surveys were conducted across different communities working in several industrial sectors [4]. This data is open to raising awareness, educating and providing resources to support mental well - being in technology and open source communities. Using these data, researchers intend to find the effect of the frequency of mental health diseases and attitudes to mental health varying by geographical location [4]. Data from one geographical region which involves a small range of audience knowing each other, there are many changes for the attacker to find details about an individual. Publishing this mental health data on a public platform can

result in privacy concerns. Hence, we will be using differential privacy technique on the above dataset to preserve privacy and accuracy.

Using epsilon differential privacy technique by adding Laplacian noise on the input data will result in increased privacy [2]. On a typical survey, the data may be categorical. Adding Laplacian noise on a categorical data will result in loss of semantic information. Adding exponential noise on client's side might not help the analysts to do data sampling [3]. Using dimensionality reduction techniques like l-anonymity and k-diversity may result in increased privacy but utility of the information published is compromised [3]. The goal of this project is to use RAPPOR differential privacy technique on mental illness data to preserve privacy and improve accuracy of the published information.

Goals

1. Data Analysis

Data Source: https://www.kaggle.com/osmi/mental-health-in-tech-survey

The above link will be used as data source for this project. Queries will be the primary mode to access data from the database. Different types of inference attacks will be employed on the data. Background knowledge attack, SQL Injection and inferential queries will be used to identify individuals on the data. If no inferential attacks are possible on the above data, additional data will be introduced for specific scenarios.

2. Implementation

Python will be used for adding RAPPOR noise to the original data. RAPPOR source code will be imported from GitHub repository for specific implementation. There might be a possibility where the data is too large. In that situation, data sampling will be used to reduce the dataset.

3. Validation

The goal of this task is to measure the privacy and accuracy of statistical data after applying RAPPOR noise to the original data. The queries which was used in the original data for inferential attacks will again be used on the perturbed data to measure accuracy and privacy. Results will be drafted in the report which will be submitted.

4. Report Submission

Steps and Timeline

S.NO	Goal	Timeline	Status

1.	Data Analysis	Oct 14	Backlog
2.	Implementation	Oct 30	Backlog
3.	Validation	Nov 7	Backlog
4.	Report Submission	Dec 1	Backlog

Deliverables

1. Executable program with RAPPOR mechanism applied on obtained data
2. Project Report

Comparatives

1. Preserving Anonymity
2. Maximizing Privacy
3. Closeness

References:

1. https://ieeexplore.ieee.org/document/8399125/
2. http://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf
3. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/42852.pdf
4. https://www.kaggle.com/osmi/mental-health-in-tech-survey
5. https://github.com/google/rappor