Internship project description
Differentially private queries for analysis of medical data

Jan Ramon

11/2023

1 Motivation and context

Over the last decades, there has been an increasing interest in exploiting data. On the other hand, recently there has also been an increasing awareness of the risks of collecting sensitive data centrally, given the frequency of data leaks, hacking or abuse. INRIA's Magnet team is interested in decentralized privacy-preserving machine learning where the sensitive data remains with the data owners, and machine learning is performed collaboratively by these data owners by participating in collaborative algorithms which (through the use of differential privacy and/or encryption) generate the desired statistical models but prevents sensitive data from being revealed. Important ongoing research projects in the team include the TIP, TRUMPET and FLUTE projects. This internship fits into the larger research program including these projects.

In the Horizon Europe projects TRUMPET and FLUTE we will develop a platform for privacy-preserving federated learning on medical data. With such platform, multiple hospitals owning patient data will collaborate on their joint data to learn together a statistical model (which is more accurate than what they can learn with their own, smaller dataset), however they will do so without revealing any of their data. The medical researcher hence will not see the data directly but can interact with it by asking queries. As the answer to queries can reveal sensitive information, we will protect the answer using differential privacy (or later similar notions). A group of medical researchers have now listed what kind of queries they want to ask.

2 Objectives

The objective of this project is to propose in a uniform way algorithms to answer the queries needed by medical researchers. Differently from classic queries to a database, we here will need additionally to keep track of the impact of a query on the privacy budget. Moreover, if we can write for the same query multiple

query plans, and if we can bring queries in some canonical form, this may have added value as it may allow for optimization of the query budget.

3 Plan

Here is a tentative work plan:

- Getting familiar with differential privacy and query answering (2 weeks)
- Getting familiar with the SmartNoise SQL (or other, similar) library (2 weeks)
- Propose a generalization for honest fraction statistical privacy (which generalizes over differential privacy) (6 weeks)
- Propose new algorithms in this framewrk for medical queries needed in the TRUMPET project (4 weeks)
- Optionally, develop a strategy to optimize groups of queries by sharing partial (correlating) query plans (6 weeks)
- Completion of the internship report (2 weeks)

The timing (here shown on a scale of 22 weeks) can be adapted according to the personal preferences of the student or the requirements of his school or type of project. The optional item can be attempted by strong students or in longer projects. If the student is insufficiently familiar with SQL, the project can also be performed using more tensor-based representations (as in libraries such as numpy, pytorch, tensorflow, etc.)

4 Environment

The project will be conducted in the INRIA MAGNET team. The student will collaborate and interact with various other collaborators in the team. It is possible the student will be asked to present progress to the team or to partners in ongoing projects. An application for ZRR access will need to be made to the FSD.

5 Requirements

We expect the student has a strong knowledge of basic computer science concepts, e.g., data structures and algorithms, and of statistics / machine learning. Knowledge of Python is required.