Master Thesis

Large-Scale Privacy-Preserving Statistical Computations for Distributed Genome-Wide Association Studies in ABY

TECHNISCHE UNIVERSITÄT DARMSTADT

The *Engineering Cryptographic Protocols Group* (`www.encrypto.de`) together with the *Computational Biology and Simulation Group* (`www.kay-hamacher.de`) is offering a master thesis on "Large-Scale Privacy-Preserving Statistical Computations for Distributed Genome-Wide Association Studies in ABY" that will be jointly supervised by Christian Weinert, Prof. Dr. Kay Hamacher, and Dr. Thomas Schneider.

Motivation

***S**ecure **M**ulti-**P**arty **C**omputation* (SMPC) allows multiple parties to compute a publicly known function while keeping their inputs private. This privacy property is desirable, e.g., in the context of distributed ***G**enome-**W**ide **A**ssociation **S**tudies* (GWAS). In GWAS, researchers try to find associations between variations in genomes and traits (e.g. diseases) by computing certain statistics on sequenced genome data of case and control groups. Since genome sequencing is still comparatively expensive and meaningful GWAS results require large participant groups, it is reasonable for research institutes to collaborate and share their data sets for this purpose. However, participants do not want their genome data to be given out of hand, since data leakage could lead, for example, to discrimination by health insurance companies.

In the past there have been several efforts to employ SMPC techniques for privacy-preserving GWAS computations (e.g. [1]), mostly as part of the yearly iDash competition. However, none of them are satisfying: they employ outdated frameworks for their implementations, evaluate performance using very small input data sets, consider only the two party case (i.e. only two research institutes collaborate), and do not take into account several practically relevant statistics.

Goal

The goal of this thesis is to implement statistical computations for GWAS (e.g. χ^2-, p-, g-, and KS-tests) in ABY (`https://github.com/encryptogroup/ABY`), a widely used state-of-the-art framework for compiling function descriptions into highly efficient privacy-preserving protocols.

Using this framework, functions can be described as circuits by plugging together different kind of high-level gates (e.g. `ADD` and `MUL`) that operate on encrypted or secret shared input data. There exist arithmetic gates for integer and IEEE 754 floating point input values. However, currently only one type of gate can be used throughout a computation, i.e. for integer inputs only integer operations and likewise for floating point inputs only floating point operations can be performed. This is not desirable in cases where floating point operations on integer inputs are only necessary in the final parts of a computation since the complexity of floating point gates leads to much higher run-times. Therefore, conversion gates should be implemented for ABY s.t. it is possible to convert between integer and IEEE 754 floating point values.

The framework additionally provides the possibility to combine different SMPC schemes within a single protocol, a feature that should be thoroughly investigated in order to maximize the efficiency of the resulting protocols.

Furthermore, an outsourcing scenario should be considered where multiple research institutes securely share their inputs with a small amount of servers that together execute the generated SMPC protocols and return the results.

Finally, the performance of the implementation should be evaluated by first comparing it to previous works and then testing scalability using large input data sets.

Requirements

- Good programming skills in C/C++
- At least basic knowledge of cryptography and statistics
- High motivation + ability to work independently
- Knowledge of the English language, Git, LaTeX, etc. goes without saying

Contact

If you are interested, please contact: M.Sc. Christian Weinert (christian.weinert@crisp-da.de)

References

[1] Scott D. Constable, Yuzhe Tang, Shuang Wang, Xiaoqian Jiang, and Steve Chapin. Privacy-preserving GWAS analysis on federated genomic datasets. *BMC Medical Informatics and Decision Making*, 15, 2015.