Building Secure Multi-Party Data Systems

A private-information network (e.g. Health Information Exchange, HIE) is an emerging federated big-data platform where autonomous data owners store, exchange and process their data. The data sharing helps establish holistic views about distributed data and allows to further extract new insights. However, there exist a major data-sharing barrier, caused by privacy concerns and data-protection laws (e.g. HiPAA).

Intellectual merits:

This proposal aims at enabling secure data sharing and practical global computation, by marrying the multi-party computation technique (MPC) and big-data systems ("MPC meets SparkSQL").

In order to achieve the goal, there are two key challenges that have to be tackled: 1) MPC performance optimization and 2) practical MPC-database architecture design and implementation. For Challenge 1), big-data systems feature a large dataset and support of complex queries, while MPC is known to be expensive even for small-scale use. The key challenge lies in effectively minimizing the *use* of MPC protocols without sacrificing security. Fortunately in our overall system, we can leverage the *query*-level semantics to minimize the use of MPCs (query-aware MPC).

For Challenge 2), the system building of existing databases and MPC is fundamentally different; for instance, the database's iterator model executes query in data-dependent fashion, while MPC has to decouple data-dependence from the execution flow for security (i.e. obliviousness). To reconcile the difference, we treat security/MPC as first-class citizen and on this basis, study the systems research problem; how to reuse existing query processing framework in databases for the MPC-based system.

We propose a MPC query processing framework including components of MPC-query optimizer, compiler and executor. In this framework, we systematically apply our motivating idea (i.e. the query-aware MPC as mentioned) to different query-processing components and propose a series of techniques: leaking cardinality for small MPC circuits, MPC localization, database-style MPC optimizer, oblivious similarity-join compilation. We formally specify and analyze the security implication of these optimization techniques. We will implement the system on existing MPC software and integrate it with SparkSQL alike data systems.

Broader impacts: The broader impacts of this research are two-fold: 1) The proposed research is expected to have an immediate impact on the research community of security and systems, and to promote wider adoption of secure data sharing and federation in financial data exchange, health-care marketplace, smart-city, etc. The research results will be disseminated through peer-reviewed publications. The developed software will be released in open source and packed in Docker alike images for easy deployment. The PI will actively seek for opportunities with security companies and government agencies for technology transfer and system integration in broader sectors. 2) Based on the results of the research, new course materials, such as hands-on labs, will be developed for undergraduate and graduate students. Professional training tutorials will be developed to help future practitioners deploy and use the system to related domains. The proposed research will also create opportunities for undergraduate and under-represented groups to participate in research activities in cyber-security.

Keywords: secure system, applied cryptography, multi-party computation, database, query processing