Enabling privacy-preserving biomedical data analytics in the cloud and across institutions

Haixu Tang, Xiaofeng Wang and Yan Huang
School of Informatics and Computing
Indiana University, Bloomington
Xiaoqian Jiang, Shuang Wang and Lucila Ohno-Machado
Department of Biomedical Informatics
University of California, San Diego
Project Objectives

(U01 EB023685: 2016-2019 Encryption Methods and Software for Privacy-Preserving Analysis of Biomedical Data)

- Develop encryption methods and software for outsourcing biomedical computations to public/commercial clouds.

- Develop encryption methods and software for collaborative biomedical computations across multiple institutions.

- Develop customized, interactive software for targeted applications of encryption methods in biomedical research.
Secure Computation on Biomedical Data: Two Scenarios

• **Secure outsourcing**
  - Motivation: to leverage the cheap computing and storage resources of commercial clouds (e.g., Amazon EC2) for biomedical computation
  - Privacy concern: the inadvertent exposure of biomedical (e.g. human genomic) data to unauthorized users.
  - Approach: data owner conduct secure computation on encrypted data stored on clouds, and decrypt the outcome to get the results
  - Examples: Secure signature search; Secure learning on encrypted data

• **Secure cross-institutional collaboration**
  - Motivation: patient centered research networks (PPRN) and clinical data research networks (CDRN) like to collaborate
  - Privacy concern: about data being used beyond agreed research scope and being processed in untrusted computational environments.
  - Approach: two (or more) institutions jointly compute a task without exposing compute a task without exposing to each other individuals’ input
  - Examples: Secure similar patients search; Privacy-preserving record linkage.

• **Adversary model:** semi-trusted (*honest-but-curious*) parties
Secure Computation: General approaches

• Secure outsourcing
  – Approach: data owner conduct secure computation on encrypted data stored on clouds, and decrypt the outcome to get the results
  – Homomorphic Encryption (HME): fully homomorphic encryption (Gentry, 2009)

• Secure cross-institutional collaboration
  – Approach: two (or more) institutions jointly compute a task without exposing individuals’ input
  – Secure Multiparty Computation (SMC): Yao’s Garbled circuits, secret sharing, etc
Fully Homomorphic Encryption

Dec_{sk}(f(Enc_{pk}(x, y))) = f(x, y)

Partially homomorphic encryption:
with respect to +: Paillier
with respect to ×: RSA, ElGamal

Fully homomorphic encryption (FHE, Gentry 2009):
A bootstrapping transformation begins with a somewhat homomorphic encryption (SWHE) encryption scheme that supports a limited number of operations, and can homomorphically evaluate its own decryption circuit plus at least one more computation Dec_{sk}(f(c)). The process can be bootstrapped: in each step, at least one more computation can be achieved. All constructions of FHE follow Gentry’s lattice-based methods, but most uses a layered (instead of bootstrapping) construction.

Implementations: HElib (IBM), SEAL (Microsoft).
Free for academic users

Gentry, 2009;
Credit: Wu, 2015
Garbled Circuits-based SMC

\[ f(x_0, x_1, \ldots, x_m; y_0, y_1, \ldots, y_m) \]

\[ X_{0,0} = K_i^0 \oplus K_j^0 \oplus K_k^0 \]
\[ X_{0,1} = K_i^0 \oplus K_j^1 \oplus K_k^0 \]
\[ X_{1,0} = K_i^1 \oplus K_j^0 \oplus K_k^0 \]
\[ X_{1,1} = K_i^1 \oplus K_j^1 \oplus K_k^1 \]

Authentication function

\[ k_i^0 k_i^1 \]
\[ k_j^0 k_j^1 \]

Oblivious transfer (OT)

Implementations: fastGC (Huang), EMP (U Maryland)

Yao, 1982.
Multiparty Computation

The Lead research institution will conduct the largest fraction of computation.
Challenges when applied to real-world biomedical tasks

- Heavy computation overhead
  - FHE: $10^8$-$10^9$, SMC: $10^5$-$10^6$
  - For DP-based edit distance computation
    - FHE: 27.5 s for 8 DNA bases (Lauter et. al., 2014)
    - SMC: ~4 s for 100 DNA bases (Huang, et. al., 2011), 4.7 hrs for 5K segments (fastGC)

- Additional communication overhead for SMC
  - The garbled circuit needs to be sent from one party to another

- Already work for some biomedical tasks (e.g., signature search)

- Usability and customization
  - Existing software tools are designed for general purposes of secure computation, and need to be customized and optimized for biomedical computation tasks
Hardware-based secure computation

- **Intel Software Guard Extensions (SGX):** A secure computation architecture using a protected area inside the CPU (i.e., the secure enclave), for dedicated computation with sensitive codes and/or on private data.
  - Software tools have been developed for secure analysis of biomedical data (e.g., human genomic data) using SGX.
  - Much lower (negligible in some cases) computation and communication overhead.
- Using SGX platform will retain the efficiency of analyzing massive biomedical data while protecting the privacy of human subjects.
  - Can potentially be used for sharing sensitive biomedical data.
- Careful security analyses of the SGX platform are needed for understanding privacy risks of using SGX.
  - Attacks are known on the page tables and cache.
- Combining hardware-based and software-based approaches may provide an optimal solution to secure computation of biomedical data.
Project Objectives

(U01 EB023685: 2016-2019 Encryption Methods and Software for Privacy-Preserving Analysis of Biomedical Data)

• Develop encryption methods and software for outsourcing biomedical computations to public/commercial clouds.

• Develop encryption methods and software for collaborative biomedical computations across multiple institutions.

• Develop customized, interactive software for targeted applications of encryption methods in biomedical research
Secure Genome Challenge
(http://humangenomeprivacy.org)

• Annual Competition on Secure Methods for Genome Data Analysis and Sharing
  – A community effort to facilitate development of new privacy-preserving techniques for sharing and analyzing biomedical data, and evaluate those techniques on real-world data.
  – Organized by the iDASH center at UCSD and Indiana University

• 2014: Differential privacy for data sharing
• 2015: Secure computation
2016 Challenge

• Results announced at a iDASH privacy workshop, 11/11/2016, Chicago

• Over 50 teams from 13 countries registered, and 20 teams submitted solutions.

• Track 1: Practical Protection of Genomic Data Sharing through Beacon Services (Privacy-preserving data dissemination);

• Track 2: Privacy-Preserving Search of Similar Cancer Patients across Organizations (Secure collaboration)

• Track 3: Testing for Genetic Diseases on Encrypted Genomes (Secure outsourcing)
Plan for the Future

• Evaluation of Hardware-based Secure Algorithms
• Privacy-preserving Learning Algorithms
• Extending to Biomedical Data Beyond Human Genomic data
• Repository for Software Tools
  • Engaging user community
  • Engaging industrial partners (Cloud providers, Microsoft, IBM, Intel, etc.)
• Suggestions are very welcome! Please contact us.
  – Haixu Tang: hatang@indiana.edu
Acknowledgements

• U01EB023685 (2016-2019) from the BD2K project
• R01HG007078 from NIH/NHGRI.