Homomorphic Encryption for Protecting Genome Privacy

Miran Kim
School of Biomedical Informatics
University of Texas, Health Science Center at Houston
Miran.Kim@uth.tmc.edu

***Abstract*—Homomorphic encryption has emerged as one of the promising solutions to address privacy and security issues in outsourcing computation on sensitive data. We introduce recent progresses on homomorphic encryption. Further we summarize the state-of-art benchmarks of the encryption systems in real-world applications such as machine learning and genome-wide association study.**

***Index Terms*—homomorphic encryption, genome privacy**

I. Introduction

LARGE amount of human genomic data are being generated and used for driving novel scientific discoveries and promoting biomedical research. However, the effective and responsible utilization of data remain to be a big challenge. The analysis of large genomic data requires intensive computing resources, but this issue might be alleviated by outsourcing to public cloud service providers. The remaining problem is the privacy and security of outsourcing genomic data. In this paper, we introduce how to protect sensitive information by deploying a special cryptosystem, called *homomorphic encryption* (HE).

II. Homomorphic Encryption

Over the past 10 years, many HE schemes have been suggested following Gentry's blueprint [2]. This cryptosystem enables to perform arithmetic operations on encrypted data without decryption. That is, computational analysis performed on encrypted data would return the same result from computing on plaintext, thereby enabling secure outsourcing computation of data in an untrusted cloud environment. There are three types of HE schemes in terms of message spaces: (1) word encryption scheme supporting exact arithmetic over encrypted finite numbers; (2) bitwise encryption scheme that allows bitwise operations; and (3) approximate homomorphic encryption scheme with support for real numbers arithmetic. Several HE schemes support the *ciphertext packing* technique, which allows multiple values to be encrypted a single ciphertext, thereby performing parallel computation on encrypted vectors. Recently, Jiang et al. [3] developed a practical solution to encrypt a matrix into a single ciphertext and perform arithmetic operations on encrypted matrices.

III. Applications of Homomorphic Encryption

HE is starting to demonstrate its feasibility in offering rigorous yet practical solutions to biomedical applications. In particular, the approximate HE scheme [1] has shown remarkable performance in real-world applications that require arithmetic over real numbers (where small numeric errors in the least significant bits will not affect results, e.g. machine learning, statistical testing, and control systems). For example, Kim et al. [6] presented the first secure outsourcing method to train a logistic regression model on encrypted data and the follow-up showed their feasibility with real data [4]. The authors proposed efficient methods to encrypt genomic data and evaluate gradient descent algorithm on encrypted data by applying ciphertext packing and parallelization techniques. It took about six minutes to obtain a logistic regression model given the dataset consisting of 1579 samples, each of which has 18 features with a binary outcome variable. Recently, Kim et al. [5] developed a method for outsourcing computation of genome-wide association studies (GWAS) on encrypted data. The authors proposed a practical protocol to assess logistic regression model to compute p-values of different Single Nucleotide Polymorphisms (SNPs), where the associations of genotypes/phenotypes were adjusted on the basis of given covariate features (e.g. age, weight and height). The protocol proceeds in two steps: (1) build a secure logistic regression model for encrypted covariates by applying the method of [4], and (2) perform one additional update of Newton's method on all encrypted individual SNPs in parallel. This approach enables to efficiently obtain logistic regression based models for thousands of SNPs all in one, thereby achieving the best performance of HE system for GWAS. It took less than one minute given a dataset consisting of 245 samples, each of which has 10643 SNPs and 3 covariates.

References

[1] J. H. Cheon, A. Kim, M. Kim, and Y. Song. Homomorphic encryption for arithmetic of approximate numbers. In *Advances in Cryptology – ASIACRYPT*, pages 409–437. Springer, 2017.

[2] C. Gentry. Fully homomorphic encryption using ideal lattices. In *Stoc*, volume 9, pages 169–178, 2009.

[3] X. Jiang, M. Kim, K. Lauter, and Y. Song. Secure outsourced matrix computation and application to neural networks. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1209–1222. ACM, 2018.

[4] A. Kim, Y. Song, M. Kim, K. Lee, and J. H. Cheon. Logistic regression model training based on the approximate homomorphic encryption. *BMC medical genomics*, 11(4):83, 2018.

[5] M. Kim, Y. Song, B. Li, and D. Micciancio. Semi-parallel logistic regression for GWAS on encrypted data. Cryptology ePrint Archive, Report 2019/294, 2019. https://eprint.iacr.org/2019/294.

[6] M. Kim, Y. Song, S. Wang, Y. Xia, and X. Jiang. Secure logistic regression based on homomorphic encryption: Design and evaluation. *JMIR medical informatics*, 6(2), 2018.