Hands-on Projects on Privacy Preservation Techniques

**Project 1.**

This project aims to evaluate how well the differential privacy preserving techniques can safely release summary statistics, and to what extent non-participant privacy can be protected when attackers exploit those perturbed statistics in their launched background knowledge attacks.

1.       The openSNP users share both of their trait and genotype profiles online [1]. Please choose one particular trait, add those profiles with the reported trait value (Yes) to case group and those profiles with the non-reported trait value to control group.

2.       Focus  on one particular SNP (say the first one).  Using the sensitivity values of chi-square statistic and p-values [2] and varying the differential privacy threshold (0.1, 0.2, 1, 2, 10), add the Laplace noise to get the perturbed statistics and compare with the unperturbed values.

3.       Continue from 2. Our aim is to release the differentially private odds ratio. However, its sensitivity is infinite. Design and implement an algorithm of using the perturbed risk allele frequencies to calculate the differentially private odds ratio.

4.       For the task of differentially private releasing K significant SNPs based on the chi-square statistic, implement the naïve algorithm of adding the Laplace noise to the true statistic value of each of N SNPs and then output K SNPs with most significant perturbed statistics values and report experimental results.

5.       Continue from 4.  Implement the algorithm in [2] that achieves differential privacy with the magnitude of added noise proportional to K rather than to N (based on the idea [3]). Compare experimental results with 4.

[1] Bastian Greshake, Philipp E Bayer, Helge Rausch, and Julia Reda. Opensnp–a crowdsourced web resource for personal genomics. PLoS One, 9(3):e89204, 2014.

[2] S. E Fienberg, A. Slavkovic, and C. Uhler.: Privacy Preserving GWAS Data Sharing. In: 11th International Conference on Data Mining Workshops, pp 628–635. IEEE (2011).

[3] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta.: Discovering Frequent Patterns in Sensitive Data. In: 16th ACM International Conference on Knowledge Discovery and Data Mining, pp 503–512. ACM, Washington, DC (2010).

**Project 2.**

The second project focuses on differential privacy preserving mechanisms tailored for advanced genetic data analysis. The classic approach of directly perturbing the output coefficients of the regression algorithms requires an explicit sensitivity analysis, which is often infeasible. Students will apply the objective perturbation approach [1] to build differentially private logistic regression model for genomic data analysis.

1.       Use the top K significant SNPs from the openSNP file (step 1 of Project 1). Randomly split the data into training data (2/3) and test data (1/3).

2.       Download the [functional mechanism code](https://sourceforge.net/projects/functionalmecha), vary differential  privacy threshold (0.1, 0.2, 1, 2, 10), train differential privacy preserving logistic regression on the training data and compare its utility-privacy tradeoff on the test data.

[1] Jun Zhang, Zhenjie Zhang**,**Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional Mechanism: Regression Analysis under Differential Privacy**.***Proceedings of the VLDB Endowment* (***PVLDB***), 5(5):1364-1375, 2012.

**Project 3.**

The third project focuses on the application of cryptographic solutions on genetic privacy protection. Students will learn 1) how to design cryptographically secure algorithms for searching genetic databases; 2) how to securely match genetic profiles between two parties based on SMC without exposing the actual genetic data; 3) how to out-source computation of GWAS contingency table through homomorphic encryption; and 4) how to generate GWAS summary statistics without a trusted party using only SMC between the data centers.

This project is directly borrowed from [Competition Tasks of IDASH Privacy & Security Workshop 2015](http://www.humangenomeprivacy.org/2015/competition-tasks.html).  Refer to [FAQ](http://www.humangenomeprivacy.org/2015/faq.html) and check the [five journal papers](http://www.humangenomeprivacy.org/2015/) related to the competition tasks. Students are encouraged to explore and use the chosen homomorphic encryption and SMC packages (e.g., [PySEAL](https://gab41.lab41.org/pyseal-homomorphic-encryption-in-a-user-friendly-python-package-e27547a0b62f)) for this competition.