Hands-on Projects of Genetic Privacy Breaching Techniques

**Project 1.**

The goal of this project is to help students examine how the inference attack works based on the Bayesian network constructed from GWAS [2].  It focuses on disclosure risk evaluation under different scenarios, e.g., the attacker knows either some traits or SNP data about the target.  For example, in the trait inference attack,  the attacker such as an insider from organizations has access to the whole or part of a target individual's SNP profile and aims to predict some sensitive trait  of the target individual by using the constructed Bayesian network.

1.       Assume the  attacker knows the partial SNPs of the target (Alice) as shown in the sample file (sub\_genotyping\_file.txt), using “Top K Trait Prediction” from the STIP platform [1], show the top three traits with the largest changes between the posterior probability and the prior probability (determined by the population).

2.       Assume the attacker knows the target (Alice) has traits, Type 1 diabetes and Skin Pigmentation, using “SNP-Categorical Trait Inference” from the STIP platform [1], show her probability of having Lung cancer.

3.       Continue from 2, further assume the attacker knows the first 10 SNP values of Alice shown the sample file (sub\_genotyping\_file.txt), find out her probability of having lung cancer.

[1] Q. Pan, L. Zhang, and X. Wu. "STIP: An SNP-Trait Inference Platform".  Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, Nov 13-16, 2017.

[2] L. Zhang, Q. Pan, Y. Wang, X. Wu and X. Shi, "Bayesian Network Construction and Genotype-Phenotype Inference Using GWAS Statistics," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi: 10.1109/TCBB.2017.2779498 URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8126853&isnumber=4359833>

**Project 2.**

The second project focuses on whether and to what extent the attacker may exploit public resources to learn individual traits and sensitive attributes like disease.

1.       It has been shown that private traits and attributes of individuals are predictable from easily accessible digital records of behavior such as Facebook Likes [1]. Other patient social networks and online communities like `patientlikeme.com' provide a platform for users (mostly patients) to connect with others who have the same disease or condition and share their own experiences. Many sensitive information such as abortions, sexual abuse, illegal drug use, alcoholism, clinical depression are revealed there. Please choose one platform and examine what kind of background knowledge an attacker may learn about the target individual in practice.

2.       The openSNP users share both of their trait and genotype profiles online [2]. Please identify the overlapped traits and SNPs contained in both the GWAS catalog and openSNP. Among all the identified traits and SNPs, we further require that the odds ratio of the trait-SNP pair to be larger than 2 so that the effect of the SNP on the trait is significant. Show the identified trait-SNP pairs.  Then select the openSNP users who have reported at least one of the three traits and all the SNPs associated with the reported traits.

3.       For each identified user from 2, blend his/her SNP profile  into the CEU dataset.  The CEU dataset  is extracted from the genotype profiles in the 1000 Genomes Project [3].  It consists of 99 HapMap individuals from Utah residents with Northern and Western European ancestry (CEU) in the 1000 Genomes Project. Assume the attacker only knows the individual’s traits and the individual’s SNP file is contained in the blended dataset, use identify inference attack to show the likelihood of correctly finding the target’s SNP record from the blended dataset.

 [1] Michal Kosinski, David Stillwell, and Thore Graepel. “Private traits and attributes are predictable from digital records of human behavior”. *Proceedings of the National Academy of*

*Sciences*, 110(15):5802–5805, 2013.

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

[3] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature, 491:1, 2012.

**Project 3.**

The third project aims to help students explore advanced attacks. When only a partial of the target's DNA data is available, students can examine whether the attacker may apply genotype imputation technique to restore missing genotypic values in the data. It has been shown that the imputation technique, which takes advantage of the LD between markers and uses reference panels with complete genetic information, is possible when genealogical information is available in addition to genetic data. For instance, the adversary may learn the target's genetic data by exploiting the known genetic data of his relatives. Students expect to study imputation techniques and evaluate how feasible the adversary can exploit social media sites to find relatives of the target who self-identified their genetic data.