Infringement of Individual Privacy via Mining GWAS Statistics

Yue Wang, Xintao Wu and Xinghua Shi
College of Computing and Informatics, University of North Carolina at Charlotte

Background

Data privacy in genome-wide association studies (GWAS) is a critical yet under-exploited research area. To illustrate the importance of data privacy in GWAS, in this paper, we introduce several attacks that demonstrate the potential risk to disclose privacy of not only GWAS participants but also general population by mining aggregate GWAS statistics.

Results

- Trait Inference

Table. 1. Attack Background Information.

Index	$Trait$	$S_j - s_j$	O_{kj}	f^t_{kj}	$\mathbf{f}^c_{kj}$	$P(t_k)$
1	COPD[1]	$rs9394152 - C$	22.22	0.41	**0.9392**	0.05
2		$rs73717741 - G$	11.9	0.07	**0.4725**	
3		$rs10928927 - C$	17.54	0.16	**0.7696**	
4	Flucloxacillin [2]	$rs2395029 - G$	45	0.05	**0.703**	0.00008
5	Jaw Osteonecrosis	$rs1934951 - T$	12.75	0.12	**0.63**	0.056
6	Osteoarthritis	$rs12982744 - C$	11.11	0.61	**0.9456**	0.036
7	Height [3]	$rs12982744 - G$	33.33	0.4	**0.9569**	0.10
8		$rs7853377 - G$	50.0	0.23	**0.9372**	
9		$rs7567288 - C$	33.33	0.2	**0.8929**	
10	Eye color [4]	$rs12913832 - A$	8.43	0.23	**0.7158**	0.16

Table. 2. Posterior Probability of Certain Trait Conditional on a Single SNP.

Index	n_1	$P(T\|r_{ij} = s_j)$	n_0	$P(T\|r_{ij} = \bar{s}_j)$	$\overline{P}(t_k\|r_{ij})$
1	58	0.1076	27	0.0054	**0.0751**
2	15	0.2621	70	0.0290	**0.0701**
3	20	0.2020	65	0.0142	**0.0584**
4	10	0.0011	75	$2.5E-5$	**1.5E-4**
5	27	0.2389	58	0.0240	**0.0923**
6	28	0.0546	57	0.0203	0.0316
7	57	0.1744	28	0.0078	**0.1195**
8	6	0.3117	79	0.0100	0.0313
9	3	0.3316	82	0.0175	0.0286
10	37	0.3721	48	0.0657	**0.1991**

- Identity Inference
  - CEU dataset: the 85 HapMap individuals from Utah residents with Northern and Western European ancestry (CEU) in the 1000 Genomes Project (The 1000 Genomes Project Consortium, Nature, 2012).
  - Random dataset: 85 randomly selected individuals from the 1,092 samples in the 1000 Genomes Project.

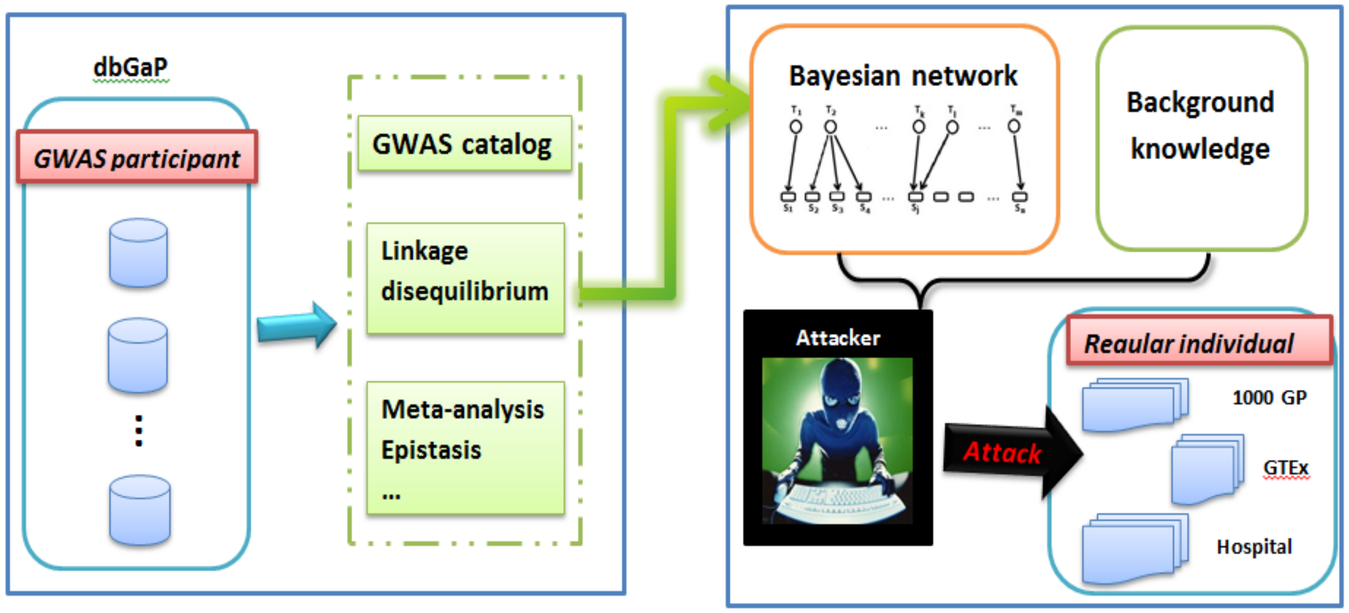

Fig. 1. Preserving privacy of GWAS participants and regular individuals.

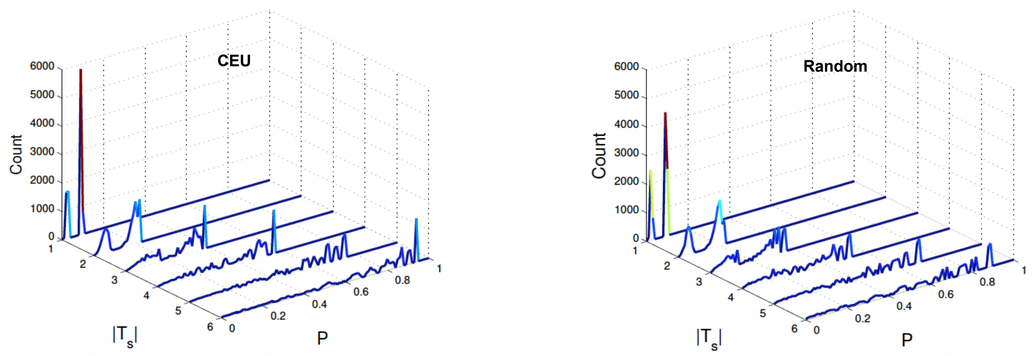

Fig. 3. Probability Distribution of Identity Inference Attack with Different Amount of Background Knowledge on (a) CEU and (b) random individuals.

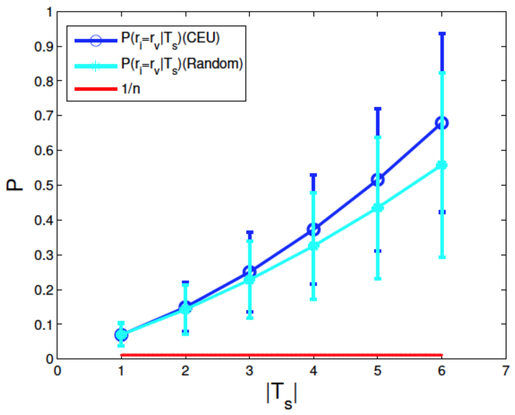

Fig. 2. Average Probability of Identity Inference Attack with Different Amount of Background Knowledge.

Methods

- We first provide a method to construct a two-layered Bayesian network explicitly revealing the conditional dependency between single-nucleotide polymorphisms (SNPs) and traits, from the public GWAS catalog.
- We then develop efficient algorithms for two attacks (***identity inference attack***, and ***trait inference attack***) based on reasoning with the dependency relationship captured in the constructed Bayesian network.

Algorithm 1 *Trait Inference*

Input: The GWAS bayesian network G, the trait set $\boldsymbol{T}$, and the genotype profile $\boldsymbol{r}_v$ of an individual v

Output: The probability $P(T_k|\boldsymbol{r}_v)$ that the individual v has any trait in the trait set $\boldsymbol{T}$

1: **for** each trait T_k in $\boldsymbol{T}$ **do**
2: Search G for T_k and obtain the associated SNPs $\{S_j\}$ (j=1..m) and corresponding risk allele;
3: Extract the subgraph of T_k, SNP set $\{S_j\}$ (j=1..m) and all the other parent traits of these SNPs from the constructed bayesian network.
4: Obtain the binary values of r_{vj} for each j from 1 to m according to whether the victim has the risk allele type of each S_j in $\boldsymbol{r}_v$;
5: Calculate $P(T_k|\boldsymbol{r}_v)$ following $P(\boldsymbol{S}_x, \boldsymbol{T}_x|\boldsymbol{S}_y, \boldsymbol{T}_y) = \frac{P(\boldsymbol{S}_x, \boldsymbol{T}_x, \boldsymbol{S}_y, \boldsymbol{T}_y)}{P(\boldsymbol{S}_y, \boldsymbol{T}_y)}$
6: **end for**

Algorithm 2 *Identity Inference*

Input: The genotype profile dataset $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_n\}$ containing the target individual's genotype record $(\boldsymbol{r}_v)$, the trait set $\{T_1, T_2, ..., T_l\}$ that the target individual has, the GWAS catalog bayesian network G.

Output: The probability of each record in R belonging to the target individual $P(\boldsymbol{r}_i = \boldsymbol{r}_v)$.

1: **for** each trait T_k in set $\{T_1, T_2, ..., T_l\}$ **do**
2: Search G for T_k and obtain the associated SNPs $S_i (j \in [1, m])$ and the corresponding risk allele type;
3: **end for**
4: **for** each record $\boldsymbol{r}_i$ in $\boldsymbol{R}$ **do**
5: Calculate the probability that $\boldsymbol{r}_i$ belongs to the target individual following $P(\boldsymbol{r}_i = \boldsymbol{r}_v|\boldsymbol{T}_S) = \frac{\prod_{j=1}^{|\boldsymbol{r}_i|} P(r_{ij}|\boldsymbol{T}_{S_j})}{\sum_{i=1}^{|\boldsymbol{R}|} \prod_{j=1}^{|\boldsymbol{r}_i|} P(r_{ij}|\boldsymbol{T}_{S_j})}$
6: **end for**

Summary

Infringement of genetic privacy is a concern in human genetic studies.

+ Future work:
Extend the models to include correlation of SNPs and traits; formalize background information.

Reference:
"Using Aggregate Human Genome Data for Individual Identification", Wang Y, Wu X, and Shi X, In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013), Shanghai, China, December 2013 (**Best Paper Award**).