Infringement of Individual Privacy via Mining Differentially Private GWAS Statistics

Yue Wang[1], Jia Wen[1], Xintao Wu[2], and Xinghua Shi[1](✉)

[1] University of North Carolina at Charlotte, Charlotte, NC, USA
{ywang91,jwen6,x.shi}@uncc.edu
[2] University of Arkansas, Fayetteville, AR, USA
xintaowu@uark.edu

Abstract. Individual privacy in genomic era is becoming a growing concern as more individuals get their genomes sequenced or genotyped. Infringement of genetic privacy can be conducted even without raw genotypes or sequencing data. Studies have reported that summary statistics from Genome Wide Association Studies (GWAS) can be exploited to threat individual privacy. In this study, we show that even with differentially private GWAS statistics, there is still a risk for leaking individual privacy. Specifically, we constructed a Bayesian network through mining public GWAS statistics, and evaluated two attacks, namely trait inference attack and identity inference attack, for infringement of individual privacy not only for GWAS participants but also regular individuals. We used both simulation and real human genetic data from 1000 Genome Project to evaluate our methods. Our results demonstrated that unexpected privacy breaches could occur and attackers can derive identity information and private information by utilizing these algorithms. Hence, more methodological studies should be invested to understand the infringement and protection of genetic privacy.

1 Introduction

In the era of genomic medicine, it is critical to share genomic information with minimal worries about genetic privacy. To achieve this goal, we need to investigate human genetic data such as individual genotypes, and explore if and to what extent genetic privacy [1–4] can be breached. Human genotype data is sensitive by nature and belongs to the data type that should be dealt with scrutiny and specific restrictions. For example, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) is deployed protects the privacy of individually identifiable health information in the USA [5]. In response to the HIPAA privacy rule, data collectors and supervisory organization must meet the requirements that ensure the data analysts agree with privacy restrictions according to USA Genetic Information Nondiscrimination Act of 2008 (GINA),

Y. Wang et al. (Eds.): BigCom 2016, LNCS 9784, pp. 355–366, 2016.
DOI: 10.1007/978-3-319-42553-5_30

and organizations should protect against all forms of genetic discrimination from using individuals' genetic information.

However, studies have shown that publicly available data not covered by HIPAA protection (e.g. allele frequency of genetic variants) can be used to infer identifiable personal information [1,6]. Taking Homer et al.(2008)'s research as an example, they developed a method that can identify straightforward whether a target with some known SNPs comes from an population with known allele frequency [7]. It attracted more and more attention on the privacy disclosure of the public dissemination of the genotype-related data and aggregate statistics from the genome-wide association studies (GWAS) [8–10]. Hence, the database of Genotypes and Phenotypes (dbGaP) was deployed to manage controlled access to genotype data. However, even without raw genotype or sequence data, summary statistics can be exploited by attackers where such public information is combined with health records and other online information [1,6]. One recent study [11] showed that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. They considered a scenario in which the genomic data are available with the target's year of birth and state of residency, two identifiers not protected by HIPAA.

Our previous work [12] studied whether and to what extent the unperturbed GWAS statistics can be exploited to breach the privacy of regular individuals who are not GWAS participants. We introduced a framework based on Bayesian networks that captures the associations between SNPs and traits mined from public GWAS statistics in the GWAS catalog [13]. Two attacks, namely trait inference attack and identify inference attack, which can be exploited to breach genetic privacy of non-participant individuals, were formalized based on the Bayesian network inference and empirically evaluated.

Several research works [14,15] have been conducted for the safe release of aggregate GWAS statistics without compromising a participant's privacy. Their ideas were based on differential privacy [16]. According to [6,16,17], differential privacy is defined as a paradigm of post-processing the output and is agnostic to auxiliary information an adversary may possess, and provides guarantees against arbitrary attacks. A differentially private algorithm provides an assurance that the output cannot be exploited by the attacker to derive whether or not any individual's record is included.

In this paper, we focus on examining whether and to what extent the differentially private GWAS statistics can still be exploited by attackers to breach the privacy. As the differentially private GWAS statistics are perturbed with noise, one conjecture is that the perturbed GWAS statistics will do no harm to regular individuals. To examine this conjecture, we construct the Bayesian network using the differentially private GWAS statistics, develop efficient formulas to infer the probability of conducting these two attacks, and conduct empirical evaluations of our formulas and algorithms on simulation and real human genetic data. Our results reveal that these privacy protected statistics under differential privacy can still be employed by attackers to identity individuals or derive private information.

2 Differentially Private GWAS Statistics

2.1 Differential Privacy

We first revisit the formal definition and mechanism of differential privacy [17]. In prior work on differential privacy, a database is treated as a collection of *rows*, with each row corresponding to the data of an individual. Here we focus on how to compute GWAS statistics under differential privacy. The goal is to ensure that the inclusion or exclusion of an individual in the GWAS dataset makes no statistical difference to the results.

Definition 1 *(Differential Privacy). A GWAS algorithm Ψ that takes as input a GWAS dataset D, and outputs $\Psi(D)$, preserves (ϵ)-differential privacy if for all closed subsets S of the output space, and all pairs of neighboring datasets D and D' from $\Gamma(D)$,*

$$Pr[\Psi(D) \in S] \leq e^{\epsilon} \cdot Pr[\Psi(D') \in S], \tag{1}$$

where D and D' are two neighboring datasets that differ in only one record.

A general method for computing an approximation to any function f while preserving ϵ-differential privacy is given in [16]. The mechanism for achieving differential privacy computes the sum of the true answer and random noise generated from a Laplace distribution. The magnitude of the noise distribution is determined by the sensitivity of the computation and the privacy parameter specified by the data owner. The sensitivity of a computation bounds the possible change in the computation output over any two neighboring datasets (differing at most one individual's record).

Theorem 1 *(The Mechanism of Adding Laplace noise [16]). An algorithm A takes as input a dataset D, and some $\epsilon > 0$, a query Q with computing function $f : D \rightarrow \boldsymbol{R}^d$, and outputs*

$$\boldsymbol{A}(D) = f(D) + (Y_1, ..., Y_d) \tag{2}$$

where the Y_i are drawn i.i.d from $Lap(GS_f(D)/\epsilon)$ and $GS_f(D) := \max_{D,D' s.t. D' \in \Gamma(D)} ||f(D) - f(D')||_1$ is the global sensitivity of a function f. The mechanism satisfies ϵ-differential privacy.

Differential privacy applies equally well to an interactive process, in which an adversary adaptively questions the system about the data. Differential privacy maintains composability, i.e., differential privacy guarantees can be provided even when multiple differentially private releases are available to an adversary.

2.2 GWAS Catalog and Statistics

Case-control studies under the GWAS framework are usually conducted by comparing the genotypes of two groups of participants: individual with the trait (case group) and matched individuals without the trait (control group). Dependent on genotyping platform, the number of SNPs genotyped in a GWAS setting typically ranges from tens of thousands to tens of millions. From genotype data, we can view that an SNP locus has two possible alleles, a risk allele and a non-risk allele. The risk allele is the allele that is more frequent in the case group comparing with the control group. The odds ratio, which is defined as the ratio of the proportion of individuals in the case group having a specific allele, and the proportion of individuals in the control group having the same allele, is often used to report the difference. When the allele frequency in the case group is higher than in the control group, the odds ratio will be higher than 1. Additionally, a p-value for the significance of the odds ratio is typically calculated using a simple chi-squared test. Those SNPs whose odds ratios are significantly different from 1, along with the statistics (e.g. p-value and odds ratio) are curated as the GWAS catalog [13].

Specifically, we can extract the following data from the GWAS catalog: a trait set $\mathcal{T}$, which contains m traits, and an SNP set $\mathcal{S}$, which contains n SNPs. For each specific trait $T_k \in \mathcal{T}$, we have a subset of associated SNPs. For each associated SNP S_j, we can extract its corresponding risk allele type ($rSNP_{kj}$) associated trait T_k, the odds ratio O_{kj} of the association test, and the risk allele frequency in the control group f^t_{kj}.

Though not directly given in the GWAS catalog, the risk allele frequency in the case group can be derived from the corresponding odds ratio and the risk allele frequency in the control group. For an SNP S_j associated with a trait T_k, with the released odds ratio (O_{kj}) and the risk allele frequency in the control group f^t_{kj}, the risk allele frequency in the case group f^c_{kj} can be derived as

$$f^c_{kj} = \frac{O_{kj} \cdot f^t_{kj}}{O_{kj} \cdot f^t_{kj} + 1 - f^t_{kj}}. \tag{3}$$

2.3 Differentially Private GWAS Statistics

Differential privacy has been significantly studied from a theoretical perspective [18–21]. Enforcing differential privacy in genomic data has been recently proposed [14,15], where classical GWAS statistics and models (e.g., the allele frequencies of cases and controls, chi-square statistic and p-values) were explored.

We use $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n_c+n_t}\}$ to denote a GWAS data set that contains n_c cases and n_t controls. Each SNP profile $\mathbf{x}_i$ contains N SNPs. The purpose of a typical GWAS study is to identify K SNPs that are significantly associated with the trait under study. For each SNP, we can easily derive that the risk allele frequency in the case (control) group f^c (f^t) has a global sensitivity of $\frac{1}{n_c}$ ($\frac{1}{n_t}$) where $n_c(n_t)$ is the number of individuals in the case (control) group. The sensitivity of various statistics used for statistical tests between a given SNP and the trait can also be derived straightforwardly. For example, the sensitivity values of chi-square statistic and p-values were derived in [14] and those sensitivity values are small. For those statistics with large sensitivity values (e.g., the sensitivity of odds ratio is infinity), we can use perturbed risk allele frequencies to indirectly calculate them.

One naive approach for differentially private releasing K significant SNPs based on a given statistics $\varPhi$ (e.g., chi-square statistic) is to add the Laplace noise $Lap(\frac{N}{\epsilon}GS_{\varPhi})$ to the true statistic value of each of N SNPs and then output K SNPs with most significant perturbed statistics values. However, this naive approach is infeasible in GWAS because the noise magnitude of $Lap(\frac{N}{\epsilon}GS_{\varPhi})$ is very large due to the large number of SNPs (N). In [18], the authors developed an effective differential privacy preserving method on how to release the most significant patterns together with their frequencies in the context of frequent pattern mining. The authors in [14] adapted this method to GWAS and aimed to release K most significant SNPs. This algorithm achieves ϵ differential privacy, with the magnitude of added noise proportional to K rather that to N. This is more efficient since that the number of significant SNPs (K) is much smaller than the number of total SNPs (N).

Here, we assume we are not able to access the raw SNP genotype data, while we have access to significant SNPs $\mathbf{\Gamma}$ associated with a trait via the released GWAS catalog. Thus we add the Laplace noise directly to the statistics of those SNPs $\mathbf{\Gamma}$. In particular, for each significant SNP, we add the Laplace noise of mean zero and magnitude of $Lap(\frac{2K}{\epsilon n_c})$ ($Lap(\frac{2K}{\epsilon n_t})$) to the risk allele frequency in the case group f^c (in the control group f^t), and then use the perturbed frequencies to calculate the odds ratio. Recall that the risk allele frequency in the case (control) group f^c (f^t) has a global sensitivity of $\frac{1}{n_c}$ ($\frac{1}{n_t}$). Algorithm 1 shows our detailed algorithm. The perturbed odds ratio values are used to construct the two-layered Bayesian network.

Algorithm 1. *Differentially Private Genome-wide Association Study.*

```
Input: The genotype profile dataset $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n_c+n_t}\}$ containing $n_c$ cases and $n_t$ controls in terms of a total number of $N$ SNPs; the number of most relevant SNPs to be released $K$; the sufficient statistic function $F$; the privacy parameter $\epsilon_0, \epsilon$.
Output: The $K$ most relevant SNPs with corresponding noisy statistics.
1: Compute the sufficient statistics $F(\mathbf{x})$ for each of the $N$ SNPs and perturb each
      real value with the Laplace noise of mean zero and magnitude of $Lap(\frac{4K}{\epsilon_0}GS_F)$.
2: Pick $K$ most relevant SNPs in terms of the noisy $F(\mathbf{x})$. Let this set be denoted as $\mathbf{S}$.
3: Perturb the true value $F(\mathbf{x})$ with the new Laplace noise with mean zero and mag-
      nitude of $Lap(\frac{2K}{\epsilon_0})GS_F$ and output $\mathbf{S}$.
4: Calculate and output other related statistics to be released for the SNPs in $\mathbf{S}$, for
      example risk allele frequency in control $f^t$ and that in case $f^c$, under differ-
      ential privacy with additional amount of privacy parameter $\epsilon$ based on their
      corresponding global sensitivity.
```

3 Attack Inference Based on a Bayesian Network

3.1 Constructing a Bayesian Network from Perturbed GWAS Statistics

A Bayesian network $G = (V, E)$ is a Directed Acyclic Graph (DAG), where the nodes in V represent the variables and the edges in E represent the dependence relationships among the variables. The dependence/independence relationships are graphically encoded by the presence or absence of direct connections between pairs of variables. Hence a Bayesian network shows the (in)dependencies between the variables qualitatively, by means of the edges, and quantitatively, by means of conditional probability distributions which specify the relationships.

In GWAS, we distinguish between two different sets of variables: the set $\mathcal{T}$ of the m traits, T_k, and the set $\mathcal{S}$ of the n SNPs, S_j. Each trait T_k is a binary random variable taking values in the set $\{1, 0\}$, where 1 stands for the presence of the trait of a participant and 0 stands for the absence. Similarly, each SNP S_j has its domain in the set $\{1, 0\}$, where 1 stands for the SNP has the risk allele and 0 otherwise. Throughout this paper, we use upper-case alphabets, e.g., X, to represent a variable; bold upper-case alphabets, e.g., $\mathbf{X}$, to represent a subset of variables. We use lower-case alphabets, e.g., x, to represent a value assignment of X; bold lower-case alphabets, e.g., $\mathbf{x}$ to represent a value assignment of $\mathbf{X}$.

We adopt the approach [12] to build a two-layered Bayesian network from the aforementioned perturbed GWAS statistics. The constructed network is composed of two layers, the trait layer and the SNP layer, with edges only going from trait nodes to SNP nodes. Each node at the top level denotes a specific trait; while each node at the second level denotes an SNP. If an SNP(S_j) is associated with a trait(T_k), a directed edge is added from T_k to S_j. The conditional probability table associated with each node is populated with the derived information from the perturbed GWAS statistics.

With the Bayesian network constructed from the perturbed GWAS statistics, we can calculate the joint probability for any desired assignment of values to variables sets $\mathbf{S}$ (SNPs), $\mathbf{T}$ (traits) by

$$P(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{T}'} \Big(\prod_{S \in \mathbf{S}} P(s|Par(S)) \cdot \prod_{T \in \mathbf{T}} P(t) \cdot \prod_{T' \in \mathbf{T}'} P(t') \Big) \tag{4}$$

where lowercase $\mathbf{s}$ and $\mathbf{t}$ denote value assignment to variable sets $\mathbf{S}$ and $\mathbf{T}$, $\mathbf{T}'$ denotes the set of all the parent traits of the SNPs in $\mathbf{S}$ except for those already contained in $\mathbf{T}$, i.e., $\mathbf{T}' = Par(\mathbf{S}) \backslash \mathbf{T}$, and $\sum_{\mathbf{X}} f(\mathbf{x})$ means to sum up all $f(\mathbf{x})$ going through all instances of attributes $\mathbf{X}$ (i.e., all value combinations of attributes in $\mathbf{X}$).

Additionally, we can calculate the conditional joint probability for any *desired* assignment of values to variables sets $\mathbf{S}_x, \mathbf{T}_x$ given the *observed* assignment of variables sets $\mathbf{S}_y, \mathbf{T}_y$ by

$$P(\mathbf{s}_x, \mathbf{t}_x | \mathbf{s}_y, \mathbf{t}_y) = \frac{P(\mathbf{s}_x, \mathbf{t}_x, \mathbf{s}_y, \mathbf{t}_y)}{P(\mathbf{s}_y, \mathbf{t}_y)} \tag{5}$$

where $\mathbf{S}_x$ and $\mathbf{S}_y$ denote the set of SNPs, $\mathbf{T}_x$, $\mathbf{T}_y$ denote the set of traits, and the joint probability $P(\mathbf{S}_x, \mathbf{T}_x, \mathbf{S}_y, \mathbf{T}_y)$ and $P(\mathbf{S}_y, \mathbf{T}_y)$ can be calculated Eq. 4.

Equations 4 and 5 are straightforwardly derived by following the marginalization strategy in the reasoning process of the Bayesian network. Note that we do not need to involve all variables in our summation to calculate $P(\mathbf{S}, \mathbf{T})$ and we can apply marginalization by summing out 'irrelevant' variables. In our two-layer Bayesian network, irrelevant variables include all nodes that are not in the ancestor subgraph for the set of variables of interest $(\mathbf{S}, \mathbf{T})$.

3.2 Inference Attacks Based on a Two-Layered Bayesian Network

The constructed Bayesian network, which captures the conditional dependency between SNPs and their associated traits, is used as background knowledge for two attacks.

Trait Inference Attack. We assume that an attacker has stolen genotype profile of the target and aims to derive the probability that the victim has a specific trait using the constructed Bayesian network. The probability of the prevalence of a specific trait, which is retrievable from the literature or the internet, is used as the prior probability that the target has the specific trait. The attacker can improve his/her guess by calculating the posterior probability of the target having the trait by inferring from with the target's genotypes. Formally, we represent the genotype of a target v as a vector, $\mathbf{r}_v = (r_{v1}, r_{v2}, \cdots, r_{vn})$, with each entry r_{vj} denoting the allele type of SNP j. The attacker aims to learn the posteriori probability $P(t_k|\mathbf{r}_v)$ that the target has a specific trait T_k given the target's genotype profile $\mathbf{r}_v$ using the constructed Bayesian network. The posteriori probability $P(t_k|\mathbf{r}_v)$ can be calculated by

$$P(t_k|\mathbf{r}_v) = P(t_k|Chd(T_k)) = \frac{P(t_k) \cdot \prod_{S \in Chd(T_k)} P(s|t_k)}{\sum_{T_k} P(t_k) \cdot \prod_{S \in Chd(T_k)} P(s|t_k)}, \tag{6}$$

where $Chd(T_k)$ denotes children SNP nodes of trait T_k.

Instead of conducting inference based on the whole Bayesian network G, the attacker can simply identify the subgraph G_k that contains all children SNPs of the target trait T_k, and then calculate the posterior probability following Eq. 6.

Identity Inference Attack. We assume that the attacker has access to an anonymized genotype dataset that contains the target's genotype record and the attacker knows a subset of traits the target has. Formally, we denote the anonymized genotype profile dataset as $\mathbf{R}$, where each record $\mathbf{r}_i = (r_{i1}, r_{i2}, \cdots, r_{in})$ represents the genotype profile of an anonymized individual i. We assume that the genotype profile of the target $\mathbf{r}_v$ is contained in $\mathbf{R}$, and the attacker knows $\mathbf{T}^\star$, a subset of traits the target has. The attacker aims to learn the posteriori probability $P(\mathbf{r}_i = \mathbf{r}_v|\mathbf{t}^\star)$ that the genotype record $\mathbf{r}_i$ corresponds to the target using the constructed Bayesian network.

For each genotype record $\mathbf{r}_i \in \mathbf{R}$, the posterior probability $P(\mathbf{r}_i|\mathbf{t}^\star)$ is

$$P(\mathbf{r}_i|\mathbf{t}^\star) = \sum_{\mathbf{T}'} \Big(\prod_{j=1}^{|\mathbf{r}_i|} P(r_{ij}|Par(S_j)) \cdot \prod_{T' \in \mathbf{T}'} P(t') \Big), \tag{7}$$

and the probability that $\mathbf{r}_i$ belongs to the target v is

$$P(\mathbf{r}_i = \mathbf{r}_v|\mathbf{t}^\star) = \frac{P(\mathbf{r}_v|\mathbf{t}^\star)}{\sum_{i=1}^{|\mathbf{R}|} P(\mathbf{r}_i|\mathbf{t}^\star)} = \frac{\sum_{\mathbf{T}'} \Big(\prod_{j=1}^{|\mathbf{r}_v|} P(r_{vj}|Par(S_j)) \cdot \prod_{T' \in \mathbf{T}'} P(t') \Big)}{\sum_{i=1}^{|\mathbf{R}|} \sum_{\mathbf{T}'} \Big(\prod_{j=1}^{|\mathbf{r}_i|} P(r_{ij}|Par(S_j)) \cdot \prod_{T' \in \mathbf{T}'} P(t') \Big)} \tag{8}$$

where $\mathbf{T}' = \mathcal{T} \backslash \mathbf{T}^\star$.

Since the calculation of $P(\mathbf{r}_i = \mathbf{r}_v|\mathbf{t}^\star)$ shown in Eq. 8 involves summation over $\mathbf{T}'$. We present a simplified formula. For each genotype record, the probability that $\boldsymbol{r}_i$ belongs to the target v is

$$P(\boldsymbol{r}_i = \boldsymbol{r}_v|\boldsymbol{t}^\star) = \frac{\prod_{j=1}^{|\boldsymbol{r}_i|} P(r_{ij}|Par(S_j))}{\sum_{i=1}^{|\boldsymbol{R}|} \prod_{j=1}^{|\boldsymbol{r}_i|} P(r_{ij}|Par(S_j))}. \tag{9}$$

Identity inference attack describes a possible approach an attacker could take to identify the target individual's record in the dataset. Based on this attack, the attacker can also infer other private information of the target individual. For example, after deriving the probability that each record in the genotype dataset belongs to the target individual, the attacker can further derive any other trait that the target may have, based on the genotype information contained in the dataset. Assume the attacker also knows the target individual has a subset of traits, $\mathbf{T}_S$. The probability that the target has a new trait T_{new} can be derived as

$$P(t_{new}|\mathbf{r}_v \in \mathbf{R}, \mathbf{t}^\star) = \sum_{i=1}^{|\mathbf{R}|} P(\mathbf{r}_i = \mathbf{r}_v|\mathbf{r}_v \in \mathbf{R}, \mathbf{t}^\star) \times P(t_{new}|\mathbf{r}_i). \tag{10}$$

4 Evaluation

We conduct experiments to evaluate how the trait inference attack and the identity inference attack work based on the Bayesian network constructed from the differentially private statistics. Our evaluation is based on the 85 Utah residents with ancestry from northern and western Europe (CEU) from the 1000 Genomes Project. In our experiments, we choose two privacy threshold values, $\epsilon = 2$ and $\epsilon = 0.2$, which represent two settings for reasonable privacy preservation in GWAS. For each ϵ, we follow the procedure in Sect. 2.3 to derive the differential privacy preserving statistics and then construct the Bayesian network.

Table 1 shows the comparison of the trait inference attack. Column $P(t_k = 1)$ shows the prevalence of the trait in population. Columns $\overline{P}(t_k = 1|r_{ij} = 1)$, $\overline{P}(t_k = 1|r_{ij} = 1)(\epsilon = 2)$, and $\overline{P}(t_k = 1|r_{ij} = 1)(\epsilon = 0.2)$ show the average probability that the 85 CEU participants from the 1000 Genomes Project has each

Table 1. Differential private posterior probability of certain trait considering one SNP.

Index	$P(t_k = 1)$	$\overline{P}(t_k = 1 \mid r_{ij} = 1)$	$\overline{P}(t_k = 1 \mid r_{ij} = 1)(\epsilon = 2)$	$\overline{P}(t_k = 1 \mid r_{ij} = 1)(\epsilon = 0.2)$
1	0.05	0.0751	0.0749	0.0749
2		0.0701	0.0670	0.0679
3		0.0584	0.0581	0.0571
4	8E-5	$1.54E-4$	$1.59E-4$	$1.49E-4$
5	0.056	0.0923	0.0934	0.2637
6	0.036	0.023	0.023	0.023
7	0.10	0.2031	0.2054	0.2055
8		0.0303	0.0360	0.0360
9		0.0258	0.0300	0.0301
10	0.16	0.1991	0.1992	0.1986

trait under three compared scenarios, using directly released GWAS statistics, 2-differentially private statistics, and 0.2-differentially private statistics, respectively. The results from Table 1 shows that most of the average probabilities are significantly different than the corresponding prior probability of having a trait. We are interested in how the derived posterior probabilities using perturbed statistics are different from those using the original statistics. We define the average absolute relative error as $\gamma(\epsilon) = \frac{1}{K}\sum_{j=1}^{K} \frac{|\overline{P}(t_k=1|r_{ij}=1) - \overline{P}_\epsilon(t_k=1|r_{ij}=1)|}{\overline{P}(t_k=1|r_{ij}=1)}$. Our results show $\gamma(2) = 0.0408$ and $\gamma(0.2) = 0.2282$, which indicate the more rigorous privacy protection incurs more loss of attack performance in terms of accuracy.

We also use the differentially private statistics to run the identity inference attack again on 'CEU' dataset. In Table 2, each row corresponds to some certain number of traits the target individual has. The columns under label 'Original', '$\epsilon = 2$' and '$\epsilon = 0.2$' denote the average probability of correctly identifying

Table 2. Average probability of identity inference attack with different amount of background knowledge.

$\lvert\mathbf{T}^\star\rvert$	$\overline{P}(\mathbf{r}_i = \mathbf{r}_v \mid \mathbf{T}^\star)$					
	Original		$\epsilon = 2$		$\epsilon = 0.2$	
	ave	std	ave	std	ave	std
1	0.0697	0.0321	0.0645	0.0275	0.0325	0.0075
2	0.1493	0.0312	0.1320	0.0576	0.0646	0.0229
3	0.2503	0.1138	0.2118	0.0916	0.0978	0.0497
4	0.3725	0.1578	0.3032	0.1348	0.1230	0.0923
5	0.5158	0.2047	0.4079	0.1911	0.1360	0.1484
6	0.6792	0.2565	0.5323	0.2657	0.1340	0.2200

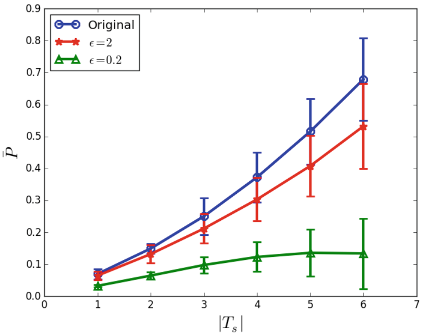

Fig. 1. Average Probability of Identity Inference Attack with Different Amount of Background Knowledge. (Color figure online)

the target individual $\overline{P}(\mathbf{r}_i = \mathbf{r}_v|\mathbf{t}^\star)$ calculated with original value of GWAS statistics, the 2-differentially private GWAS statistics, and the 0.2-differentially private GWAS statistics respectively. For each scenario, we use ‘ave’ and ’std’ to denote the mean and the standard deviation. We can easily observe from Table 2 and Fig. 1 that the average probability of correctly identifying the target individual $\overline{P}(\mathbf{r}_i = \mathbf{r}_v|\mathbf{t}^\star)$ increases as the number of known traits increases under three scenarios. This observation shows that the more background knowledge the attacker has, the more likely the target individual can be identified. We are interested in how the performance of the identity inference attack is affected by the perturbed GWAS statistics. We can see that the attack performance is significantly decreased when GWAS statistics are distorted under rigorous privacy protection. For example, as the last row shows, when $|\mathbf{T}^\star| = 6$, the accuracy of the attack decreases from 0.6792 to 0.5323 ($\epsilon = 2$), and further to 0.1340 ($\epsilon = 0.2$). However, the probability (0.1340) that the target individual being correctly identified under $\epsilon = 0.2$ is still an order high than the probability of random guess (0.0116).

5 Conclusions and Future Work

In summary, we constructed a Bayesian network from perturbed GWAS catalog and explored whether an attacker can get the private information from public population and to what extent if so. We evaluated two types of attacks, trait inference attack and identity inference attack respectively. Both of these two attacks derive private information by using the GWAS public catalog data that capture the relationship between SNPs and their associated traits. Using both simulated and real human genetic data, we found that both of these two attacks can be real threat to the privacy of general population, even when the GWAS statistics are already perturbed under differential privacy. In our future work, we will further incorporate trait-trait relationships and/or SNP-SNP correlations into our perturbed Bayesian network and develop new inference algorithms on the network. We aim to develop methods that could protect data privacy or could release GWAS statistics with less threat for general population.

Acknowledgements. The work is supported in part by US National Science Foundation (DGE-1523115 and IIS-1502273 to XW, and DGE-1523154 and IIS-1502172 to XS).

References

1. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**(6), 409–421 (2014)
2. Greenbaum, D., Gerstein, M.: Genomic anonymity: have we already lost it? Am. J. Bioeth. **8**(10), 71–74 (2008)
3. Greenbaum, D., Gerstein, M.: Social networking and personal genomics: suggestions for optimizing the interaction. Am. J. Bioeth. **9**(6–7), 15–19 (2009)
4. Greenbaum, D., Sboner, A., Mu, X.J., Gerstein, M.: Genomics and privacy: implications of the new reality of closed data for the field. PLoS Comput. Biol. **7**(12), e1002278 (2011)
5. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). http://www.hhs.gov/hipaa/
6. Shi, X., Wu, X.: Genetic privacy: risks, ethics, and protection techniques. In: The Workshop on Data Science Learning and Applications to Biomedical and Health Sciences, pp. 57–62, New York, NY (2016)
7. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. **4**(8), e1000167 (2008)
8. Masca, N., Burton, P.R., Sheehan, N.A.: Participant identification in genetic association studies: improved methods and practical implications. Int. J. Epidemiol. **40**(6), 1629–1642 (2011)
9. Wang, R., Li, Y.F., Wang, X., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: 16th ACM Conference on Computer and Communications Security, pp. 534–544. ACM (2009)
10. Zhou, X., Peng, B., Li, Y.F., Chen, Y., Tang, H., Wang, X.F.: To release or not to release: evaluating information leaks in aggregate human-genome data. In: Atluri, V., Diaz, C. (eds.) ESORICS 2011. LNCS, vol. 6879, pp. 607–627. Springer, Heidelberg (2011)
11. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)
12. Wang, Y., Wu, X., Shi, X.: Using aggregate human genome data for individual identification. In,: IEEE International Conference on Bioinformatics and Biomedicine, pp. 410–415. IEEE, Shenzhen, China (2013)
13. Hindorff, L.A., MacArthur, J., Morales, J., Junkins, H.A., Hall, P.N., Klemm, A.K., Manolio, T.A.: A Catalog of Published Genome-wide Association Studies. http://www.genome.gov/gwastudies
14. Fienberg, S.E., Slavkovic, A., Uhler, C.: Privacy preserving GWAS data sharing. In: 11th International Conference on Data Mining Workshops, pp. 628–635. IEEE (2011)

15. Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: 19th ACM International Conference on Knowledge Discovery and Data Mining, pp. 1079–1087. ACM, Chicago, IL (2013)
16. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
17. Dwork, C.: A firm foundation for private data analysis. Commun. ACM **54**(1), 86–95 (2011)
18. Bhaskar, R., Laxman, S., Smith, A., Thakurta, A.: Discovering frequent patterns in sensitive data. In: 16th ACM International Conference on Knowledge Discovery and Data Mining, pp. 503–512. ACM, Washington, DC (2010)
19. Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: 23rd Annual Conference on Neural Information Processing Systems, pp. 289–296. Citeseer, Vancouver, B.C., Canada (2008)
20. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: 17th ACM International Conference on Knowledge Discovery and Data Mining, pp. 193–204. ACM, San Diego, CA (2011)
21. Lee, J., Clifton, C.: Differential identifiability. In: 18th ACM International Conference on Knowledge Discovery and Data Mining, pp. 1041–1049. ACM, Beijing, China (2012)