Using Aggregate Human Genome Data for Individual Identification

Yue Wang, Xintao Wu, Xinghua Shi
College of Computing and Informatics
University of North Carolina at Charlotte, USA
Email: {ywang91,xwu,x.shi}@uncc.edu

***Abstract*—Data privacy in genome-wide association studies (GWAS) is a critical yet under-exploited research area. In this paper, we first provide a method to construct a two-layered bayesian network explicitly revealing the conditional dependency between SNPs and traits, from the public GWAS catalog. Then we 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. Different from previously proposed attacks, the possible target of our attacks may be any common people, not limited to GWAS participants. The empirical evaluations show that unprotected statistics released from GWAS can be exploited by attackers to identify individual or derive private information. Thus we show that mining GWAS statistics threatens the privacy of a much wider population and privacy protection mechanisms should be employed.**

Keywords*—Genome-wide association study; privacy; Bayesian network;

I. Introduction

Genome-wide association studies (GWAS) have received intensive attention due to the rapid decrease of genotyping costs and promising potential in genetic diagnostics. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and human traits like common diseases. It has been shown that many chronic diseases and various cancer types have genetic disposition factors.

In the general biomedical community, it is usually assumed that statistics (e.g., allele frequency) from SNP data can not be used to identify individuals and hence be safe to release. However, the findings in [1] showed that GWAS statistics do not completely conceal identity since it is straightforward to assess the probability a person participated in a GWA study. The proposed method [1] measures the difference between the distance of the individual from a reference population and that from the mixture.

In this paper, we are focused on a related but different privacy protection problem, i.e., whether and to what extent GWAS statistics can be exploited by an attacker to learn private information of regular people (rather than those GWAS participants). Specifically, we study two potential attacks: 1) *trait inference attack* that aims to infer the probability of a target developing some private trait when the target's SNP profile is available to the attacker; and 2) *identity inference attack* that aims to infer the probability of a record in an anonymized genebank database belonging to the target when some traits of the target are available.

Both attacks pose a serious threat to individuals when their SNP profiles are exploited by attackers. Many organizations such as biobanks, hospitals, research consortia and pharmaceutical companies collect and publish DNA sequence and SNP data. For example, 1000 genome project [2] provides the public with free services like browsing and downloading DNA sequence, SNP genotypes and other types of data from over a thousand anonymous participants in different populations. In *trait inference attack*, the attacker such as an insider from the organizations is assumed to know the whole or part of a target individual's SNP profile and aims to predict some sensitive trait (e.g., disease) of the target individual. In *identity inference attack*, we assume the attacker such as an outsider has access to the anonymized DNA sequence dataset which contains the target individual's record and aims to identify the target individual's record from the anonymized dataset. We also assume the attacker knows some traits of the target individual. For example, it was shown in [3] that private traits and attributes of individuals are predictable from easily accessible digital records of behavior such as Facebook Likes. 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. The data generated in such process may also have potential risks in that the data can be used by the attacker to learn the private traits and attributes of individuals.

Our contributions are as follows.

- We develop a method to build a two-layer bayesian network from the released GWAS statistics. The constructed bayesian network explicitly reveals the conditional dependency between SNPs and traits, and can be used to compute the probability distribution for any subset of network variables given the values or distributions for any subset of the remaining variables.
- We then formulate two attacks, namely trait inference attack and identity inference attack, as two inference problems based on the dependency relationship captured in the bayesian network, and develop efficient formulas and algorithms to infer the probability of attacks.
- We conduct empirical evaluations of the proposed methods. Our results show that unexpected privacy breaches can occur because aggregation statistics provide no explicit security guarantees and these statistics could be exploited by attackers to identify individuals

978-1-4799-1310-7/13/$31.00 ©2013 IEEE

or derive private information.

II. Background on 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 disease (case group) and matched individuals without the disease (control group). Each individual is genotyped by microarray or sequencing platforms. 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.

In a GWAS process, SNP profile data is firstly generated by genotyping the individuals in cases and controls. Secondly, Allele frequency for each of those SNPs over the case group and the control group is calculated respectively and a statistical test is performed on a contingency table to investigate if the allele frequencies are significantly different in cases versus controls. 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 much 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. Finding SNPs whose odds ratios are significantly different from 1 is the objective of the GWAS because those SNPs are associated with the trait. Finally, those SNPs that are associated with the trait, along with the statistics (e.g. *p*-value and odds ratio) are reported. These reported SNPs, along with information about the study, the trait, specific SNP information (e.g. identifier, position, and the risk allele type), and statistics, are later collected and curated at the National Human Genome Research Institute (NHGRI) GWAS catalog [4].

III. Constructing a Bayesian Network from GWAS

In this section, we present how to build a two-layered bayesian network from the aforementioned GWAS catalog. The constructed bayesian network, which explicitly captures the conditional dependency between SNPs and traits, will be used as background knowledge for inference attacks.

A. Knowledge from GWAS Catalog

Some of the information publicly available from the NHGRI GWAS catalog, can be directly used to construct the bayesian network. Such information includes trait/disease name, the related SNPs and corresponding risk allele type, the risk allele frequency in control group and statistics like odds ratio in the association test of each SNP. Formally, we can extract the following data from the GWAS catalog: a trait set $\boldsymbol{T}$, which contains m traits, and an SNP set $\boldsymbol{S}$, which contains n SNPs. For each specific trait $T_k \in \boldsymbol{T}$, 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, its odds ratio is

$$O_{kj} = \frac{f^c_{kj}(1 - f^t_{kj})}{f^t_{kj}(1 - f^c_{kj})} \tag{1}$$

With the released values of the 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

$$f^c_{kj} = \frac{O_{kj} \cdot f^t_{kj}}{O_{kj} \cdot f^t_{kj} + 1 - f^t_{kj}} \tag{2}$$

Lemma 1. *The background knowledge that an attacker can obtain from the GWAS catalog [4] includes: a trait set $\boldsymbol{T}$, an SNP set $\boldsymbol{S}$, the risk allele type ($rSNP_{kj}$), the odds ratio O_{kj}, and the risk allele frequency in the control group f^t_{kj} for each pair of trait and its associated SNPs.*

B. Two-layered Bayesian Network Construction

In GWAS, we can distinguish between two different sets of variables: the set $\boldsymbol{T}$ of the m traits, T_k, and the set $\boldsymbol{S}$ of the n SNPs, S_j. Each trait T_k is a binary random variable taking values in the set $\{t_k, \bar{t}_k\}$, where $t_k(\bar{t}_k)$ stands for the presence (absence) of the trait of a participant. Similarly, each SNP S_j has its domain in the set $\{s_j, \bar{s}_j\}$, where s_j stands for the SNP has the risk allele and $\bar{s}_j$ otherwise.

We construct a bayesian network to represent the conditional dependencies between traits and SNPs, with the background knowledge shown in Lemma 1. 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. As shown in Figure 1, in such a network, 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.

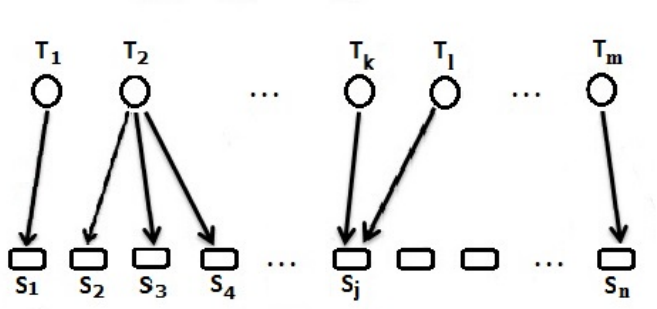

Fig. 1. A Two-layered Bayesian Network of Traits and Associated SNPs

The next step to completely specify a bayesian network is to determine the conditional probability table stored at each node. Firstly, we need to acquire the prior probability of each

TABLE I. Probability Function for SNPs with single Parent Trait

$P(S_1\|T_1)$		T_1	
		t_1	$\bar{t}_1$
S_1	s_1	f_{11}^c	f_{11}^t
	$\bar{s}_1$	$1-f_{11}^c$	$1-f_{11}^t$

trait node at the top level of the network. The prevalence of a trait T_k in the population can be obtained from the literature or internet. We treat it as the prior probability that one individual has this trait, denoted as $P(t_k)$. Secondly, we need to determine the conditional probability table of each SNP node at the second level. There are two types of SNP nodes in terms of dependency relationship with traits: a) SNP nodes which have a single parent trait node, e.g., S_1, S_2 in Figure 1; and b) SNP nodes which have more than one parent trait node, e.g., S_j in Figure 1.

For those SNP nodes with a single parent trait node, we can specify the values of the conditional probability table associated with each SNP node by its risk allele frequency in the control group and the risk allele frequency in the case group. An example for node $S1$ is shown in Table I. The probability of the risk allele of SNP $S1$ given the presence of the trait $T1$, $P(s_1|t_1)$ equals the risk allele frequency in the case group f_{kj}^c. Note that the conditional probability table of S_2 can be specified in the same way as S_1, although S_2 shares the parent trait node T_2 with other SNP nodes (i.e., S_3 and S_4).

For the SNP nodes with multiple parent trait nodes, the conditional probability table cannot be built from the GWAS catalog directly. Instead, the value of each cell in the conditional probability table, which represents one of the possible combinations of its parent nodes being true or false, can be calculated based from Lemma 2.

Lemma 2. *(Conditional probability for SNPs with multiple parent trait nodes) For a specific risk SNP S_j associated with a subset of traits $Parent(S_j)$, we have*

$$P(S_j|Parent(S_j)) = \frac{\prod_{T_k \in Parent(S_j)} P(S_j|T_k)}{P^{q-1}(S_j)}, \quad (3)$$

where q is the number of elements in $Parent(S_j)$. Specifically, when S_j is associated with two traits T_k, T_l, we have

$$P(S_j|T_k, T_l) = \frac{P(S_j|T_k) \cdot P(S_j|T_l)}{P(S_j)} \quad (4)$$

In Lemma 2, we assume that the traits are conditionally independent with each other, given the SNPs they are both associated with. The probability that one allele of an SNP appears in the population can be found from an NCBI website[1].

With the bayesian network constructed from the GWAS catalog, we can calculate(predict) the joint probability for any desired assignment of values to variables sets $\boldsymbol{S}$ (SNPs), $\boldsymbol{T}$ (traits), for example $< s_1, s_2, ..., s_{|\boldsymbol{S}|}, t_1, t_2, ...t_{|\boldsymbol{T}|} >$, following Lemma 3. In the original reasoning process in the bayesian network, we need to involve all of the other variables to calculate $P(\boldsymbol{S}, \boldsymbol{T})$. By marginalization of summing out 'irrelevant' variables, we achieve the form in Lemma 3.

Lemma 3. *The joint probability for any desired assignment of values to variables sets $\boldsymbol{S}$ (SNPs), $\boldsymbol{T}$ (traits), for example $< s_1, s_2, ..., s_{|\boldsymbol{S}|}, t_1, t_2, ...t_{|\boldsymbol{T}|} >$, can be calculated using Equation 5. Note that we can apply this equation to any other assignment of values, simply changing s_i/t_j to $\bar{s}_i$/$\bar{t}_j$ accordingly.*

$$\begin{aligned} P(\boldsymbol{S}, \boldsymbol{T}) &= P(s_1, s_2, ..., s_{|\boldsymbol{S}|}, t_1, t_2, ...t_{|\boldsymbol{T}|}) \\ &= \sum_{T_k \in \{t_k, \bar{t}_k\}} (\prod_{i=1}^{|\boldsymbol{S}|} P(s_i|Parent(S_i)) \prod_{j=1}^{|\boldsymbol{T}|} P(t_j) \prod_{k=1}^{|\boldsymbol{T'}|} P(T_k)) \end{aligned} \quad (5)$$

where $\boldsymbol{T'}$ denotes the set of all the other parent traits of the SNPs in $\boldsymbol{S}$ except for those already contained in $\boldsymbol{T}$.

Additionally, we can calculate(predict) the conditional joint probability for any desired assignment of values to variables sets $\boldsymbol{S}_x, \boldsymbol{T}_x$ given the observed assignment of variables sets $\boldsymbol{S}_y, \boldsymbol{T}_y$ following Theorem 1. Note that $\boldsymbol{S}_x$ and $\boldsymbol{S}_y$ denote the set of SNPs; while $\boldsymbol{T}_x$, $\boldsymbol{T}_y$ denote the set of traits.

Theorem 1. *(Inference via a GWAS Bayesian Network) The joint probability for any desired assignment of values to variables in $\boldsymbol{S}_x, \boldsymbol{T}_x$ given the (observed) assignment of values to variables in $\boldsymbol{S}_y, \boldsymbol{T}_y$ can be derived from Equation 6.*

$$P(\boldsymbol{S}_x, \boldsymbol{T}_x | \boldsymbol{S}_y, \boldsymbol{T}_y) = \frac{P(\boldsymbol{S}_x, \boldsymbol{T}_x, \boldsymbol{S}_y, \boldsymbol{T}_y)}{P(\boldsymbol{S}_y, \boldsymbol{T}_y)} \quad (6)$$

where the joint probability $P(\boldsymbol{S}_x, \boldsymbol{T}_x, \boldsymbol{S}_y, \boldsymbol{T}_y)$ and $P(\boldsymbol{S}_y, \boldsymbol{T}_y)$ can be calculated following Lemma 3.

IV. Inference Attacks based on a Two-layered Bayesian Network

A. Trait Inference Attack

In this attack scenario, we assume that an attacker has stolen genotype profile of the target. Formally, we represent the genotype of a target v as a vector, $\boldsymbol{r}_v = (r_{v1}, r_{v2}, \cdots, r_{vn})$, with each entry r_{vj} denoting the allele type of SNP j. The attacker aims to derive the probabilities that the victim has specific traits using the constructed bayesian network.

Definition 1. *(Trait Inference Attack) Assume that the attacker has the genotype profile $\boldsymbol{r}_v$ of the target v. The attacker aims to learn the posteriori probability $P(t_k|\boldsymbol{r}_v)$ that the target has a specific trait T_k given the target's genotype profile $\boldsymbol{r}_v$ 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.

The attacker can calculate the posterior probability of the target having a particular trait ($P(t_k|\boldsymbol{r}_v)$), using the victim's genotype information ($\boldsymbol{r}_v$) and Lemma 4.

Lemma 4. *(Trait Development Risk Estimation With Several Related SNPs) The posteriori probability $P(t_k|\boldsymbol{r}_v)$ can be*

[1] $http://www.ncbi.nlm.nih.gov/snp/$

calculated following Equation 6, specifically with $\boldsymbol{S}_x, \boldsymbol{T}_y = \emptyset$, $\boldsymbol{T}_x = T_k = t_k$. Based on conditional independence, we have $\boldsymbol{S}_y$ that contains only the SNPs associated with T_k, where the value assignment of SNP genotypes is equal to the corresponding genotype record of the target individual.

In the attack scenario described in Algorithm 1, the attacker intends to find out the possibility that the victim has certain trait, according to his/her genotype and the GWAS catalog information.

Algorithm 1 *Trait Inference*

Input: The genotype profile $\boldsymbol{r}_v$ of an individual v, the GWAS Bayesian Network G, the trait set $\boldsymbol{T}$
Output: The probability $P(T_k|\boldsymbol{r}_v)$ that the individual has any trait in $\boldsymbol{T}$

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 type;
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 Lemma 4.
6: **end for**

B. Identity Inference Attack

In identity inference attack, we assume that the attacker has access to an anonymized genotype dataset that contains the target's genotype record. Formally, we denote the anonymized genotype profile dataset as $\boldsymbol{R}$, where each record $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \cdots, r_{in})$ represents the genotype profile of an anonymized individual i. We assume that the genotype profile of the target $\boldsymbol{r}_v$ is contained in $\boldsymbol{R}$, and the attacker knows $\boldsymbol{T}_S$, a subset of traits the target has.

Definition 2. *(Identity Inference Attack) Given the anonymized genotype profile dataset $\boldsymbol{R}$ which contains the target's genotype record $\boldsymbol{r}_v$, and a subset of the target's traits, $\boldsymbol{T}_S$, the attacker aims to learn the posteriori probability $P(\boldsymbol{r}_i = \boldsymbol{r}_v|\boldsymbol{T}_S)$ that the genotype record $\boldsymbol{r}_i$ corresponds to the target using the constructed bayesian network.*

With the bayesian network constructed in the previous section, we can naturally acquire the probability that an individual has a specific allele type for an SNP given his/her associated trait information. Lemma 5 shows how to calculate the possibility that a record in the dataset belongs to the target given his specific traits. The proof is straightforward based on Theorem 1 and we skip it due to space limit.

Lemma 5. *For each genotype record, the probability that $\boldsymbol{r}_i$ belongs to the target v is*

$$P(\boldsymbol{r}_i = \boldsymbol{r}_v|\boldsymbol{T}_S) = \frac{\prod_{j=1}^{|\boldsymbol{r}_i|} P(r_{ij}|\boldsymbol{T}_{S_j})}{\sum_{i=1}^{|\boldsymbol{R}|} \prod_{j=1}^{|\boldsymbol{r}_i|} P(r_{ij}|\boldsymbol{T}_{S_j})} \tag{7}$$

where $\boldsymbol{T}_{S_j}$ denotes the parent trait nodes of S_j in the bayesian network. $P(r_{ij}|\boldsymbol{T}_{S_j})$ can be acquired from the bayesian network.

Algorithm 2 describes a possible approach an attacker could take to identify the target individual's record in the dataset. Based on this approach, the attacker can also infer other private information of the target individual.

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 $S_j (j \in [1, m])$ 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 Lemma 5.
6: **end for**

C. New Trait Inference

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. We formalize such new trait inference in Lemma 6.

Lemma 6. *Assume that the genotype profile of the target, $\boldsymbol{r}_v$, is contained in a genotype profile dataset $\boldsymbol{R}$. The attacker has access to $\boldsymbol{R}$ where each record $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \cdots, r_{in})$ denotes the genotype profile of an individual i. The attacker also knows the target individual has a subset of traits, $\boldsymbol{T}_S$. The probability that the target has a new trait T_{new} can be derived as*

$$\begin{aligned} P(T_{new}|\boldsymbol{r}_v \in R, \boldsymbol{T}_S) &= \sum_{i=1}^{|\boldsymbol{R}|} P(\boldsymbol{r}_i = \boldsymbol{r}_v) \times P(T_{new}|\boldsymbol{r}_i) \\ &= \sum_{i=1}^{|\boldsymbol{R}|} P(\boldsymbol{r}_i = \boldsymbol{r}_v|\boldsymbol{T}_S) \times P(T_{new}|\boldsymbol{r}_i) \end{aligned} \tag{8}$$

where $P(T_{new}|\boldsymbol{r}_i)$ can be derived following Lemma 4 and $P(\boldsymbol{r}_i = \boldsymbol{r}_v|\boldsymbol{T}_S)$ can be derived following Lemma 5.

V. Evaluation

We evaluate our methods using data extracted from the online NHGRI GWAS catalog [4] as of May 21, 2013. This version of the GWAS catalog includes 1,607 publications and 12,520 records about 10,133 SNPs associated with 834 traits. Publications included in such a catalog are limited to those attempted to assay at least 100,000 SNPs in the initial stage. SNP-trait associations listed are limited to those with p-values less than 10^{-5}.

Table II shows the information and statistics of a snapshot of our constructed two-layer bayesian network from the GWAS catalog. There are 6 traits and 9 associated SNPs, which were

TABLE II. Attack Background Information

Index	$Trait$	$S_j - s_j$	f^t_{kj}	O_{kj}	$\mathbf{f}^c_{kj}$	$P(t_k)$
1	Chronic obstructive pulmonary disease	$rs9394152 - C$	0.41	22.22	**0.9392**	0.05
2		$rs73717741 - G$	0.07	11.9	**0.4725**	
3		$rs10928927 - C$	0.16	17.54	**0.7696**	
4	Drug-induced liver injury (flucloxacillin)	$rs2395029 - G$	0.05	45	**0.703**	0.00008
5	Jaw Osteonecrosis	$rs1934951 - T$	0.12	12.75	**0.63**	0.056
6	Osteoarthritis	$rs12982744 - C$	0.61	11.11	**0.9456**	0.036
7	Height(taller than 90% of population)	$rs12982744 - G$	0.4	33.33	**0.9569**	0.10
8		$rs7853377 - G$	0.23	50.0	**0.9372**	
9		$rs7567288 - C$	0.2	33.33	**0.8929**	
10	Eye color (Green)	$rs12913832 - A$	0.23	8.43	**0.7158**	0.16

TABLE III. Posterior Probability of Certain Trait Considering one 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**

reported from from six previous GWAS publications. For each SNP-trait pair, the risk allele type, risk allele in the control group and the odds ratio are shown in Columns 3-5. We calculate the risk allele frequency in the case group for each SNP-trait pair and show the result in Column 6. Note that there is a big gap (around 0.5) between the risk allele frequency in the case group and that in the control group. We also acquire the prior probability (prevalence) of each trait, $P(t_k)$, from the original studies or Wikipedia.

A. Trait Inference Attack

With the constructed bayesian network, the attacker can then run the trait inference attack(Algorithm 1) to calculate the posterior probability that the target individual has a trait given his/her genotype profile. In our evaluation, we use the genotype profiles of the 85 HapMap individuals from Utah residents with Northern and Western European ancestry (CEU) in the 1000 Genomes Project [2]. In addition to their genotype profiles, the race and gender information is also publicly available.

Table III shows the estimation results calculated from Lemma 4. Each row in Table III corresponds to the row with the same index in Table II. Column $\overline{P}(t_k|r_{ij})$ gives the average probability that the 85 CEU participants from the 1000 Genomes Project has each trait. We can see that most of the average probabilities (with bold font) are higher than the corresponding prior probability of having a trait. Columns n_1 and n_0 respectively represent the number of individuals who have and doesn't have the risk allele type listed in the corresponding row of Table II. Columns $P(t_k|r_{ij} = s_j)$ and $P(t_k|r_{ij} = \bar{s}_j)$ respectively represent the the posterior probability of one individual has a trait if he/she has the risk allele type, or doesn't have the risk allele type of one specific SNP corresponding to the trait.

B. Identity Inference Attack

In the example of the identity inference attack illustrated in Algorithm 2, we also use the genotype data of 85 CEU individuals from the 1000 Genomes Project [2]. In our experiment, we assume that the target has traits of green eyes and top 10% height and all other traits list in Table II. That is to say, the trait set for the target has six elements.

We randomly generate the genotype record for the target individual. The generating strategy is that for each SNP S_j associated with one trait T_k, we generate $r_{ij} = s_j$ with the probability $P(s_j|t_k)$. Next we blend the generated record into the dataset containing the genotype records of the 85 CEU individuals. Finally, we calculate the possibility that the generated record is identified as belonging to the target individual, given the background trait information. We also compare the inference capability with different amount of background knowledge, i.e., with the size of trait set ranging from one to six.

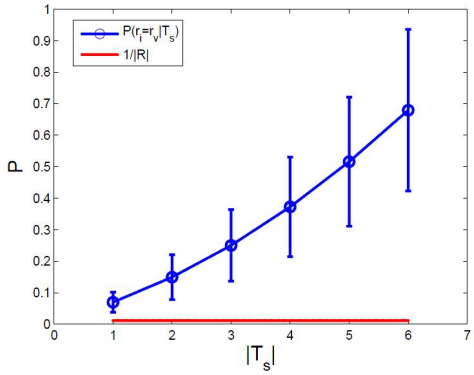

Fig. 2. Average Probability of Identity Inference Attack with Different Amount of Background Knowledge

We run this whole process for 10,000 times and Figure 2 shows the average value of the resulted possibilities. As shown in Figure 2, the red line (1/86) is the baseline representing the probability that the generated record is inferred as belonging to the target individual without background knowledge. The first blue point represents the average probability that the

generated record is correctly referred given any one of the six traits. Similarly, the second blue point represents the average probability that the generated record is correctly referred given any two different traits from all the six traits, and so on. We can see that in general, the probability of correctly identifying the target individual increases as the background knowledge increases, while the inference probability given only one trait is much larger than that of the situation without background knowledge. The bar at each point shows the standard deviation of the probability among 10,000 times of test.

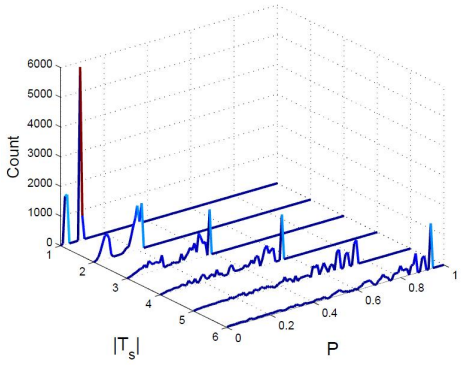

Fig. 3. Probability Distribution of Identity Inference Attack with Different Amount of Background Knowledge

Figure 3 shows the distribution of the inference probability among the 10,000 times of identity inference test. As the amount of background traits increases, the peaks of the process count would be located at positions with larger identifying probability. This indicates that in general, the more background knowledge we have, the more probably that the target individual's record is correctly identified. Specifically, the highest peak in the wave of $|T_s| = 6$ locates near line of $P = 1$, which represents that if knowing all six traits, the attacker could successfully identify the target individual with a confidence of more than 90% in most times of test. On the other hand, the multiple peaks in each line represent the different identifying probabilities due to the different combinations of background traits as well as the different possible genotype records being randomly generated.

VI. Related Work

Sharing de-identified genotype or genomic information has become a common practice in human genetics. [5] demonstrated end-to-end identification of individuals with only public information and 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 that are not protected by HIPAA. From a different angle but along the same line, our study here further shows that the re-identification of anonymized genotype data still hold a real threat to normal individuals who are not GWAS participants using published data.

Homer et al. in [1] developed a method to determine whether a person with known genotypes at a number of markers was part of a sample from which only allele frequencies are known. This study prompted concerns about the public dissemination of genotype data and aggregate statistics from GWAS. Consequently NIH regulated that the database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/gap) has to be accessed by controlled access. An study in [6] proposed two attacks based on statistics release in GWAS. The first attack extended Homer's attack by utilizing a more powerful statistics (r^2) which describes the correlation among different SNPs, rather than the allele frequencies in Homer's attack. The other attack gave the way to recover the un-released SNPs of participants by analyzing the r^2 between pairwise SNPs. All the above papers are focused on the privacy protection of GWAS participants.

VII. Conclusions and Future Work

In summary, we studied whether and to what extent GWAS statistics can be exploited by an attacker to learn private information of general population, not limited to GWAS participants. We developed two potential attacks, *trait inference attack* and *identity inference attack*. Both attacks exploit the released GWAS statistics about the associations between SNP genotypes and human traits. Our evaluations showed that the proposed attacks have made re-identification of anonymized genotype data a real threat. In our future work, we will study how to extend our two-layered bayesian network to capture trait-trait associations and/or SNP-SNP correlations. We will study how to formalize various types of background knowledge that an attacker may have in practice and evaluate how well data perturbation and agglomeration techniques with background knowledge can protect privacy when releasing GWAS statistics. Our goal is to develop methods to enable researchers to safely release aggregate GWAS data without compromising the anonymity of both study participants and non-participants.

Acknowledgments

The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported in part by U.S. National Institute of Health (1R01GM103309).

References

1. N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLoS genetics*, vol. 4, no. 8, p. e1000167, 2008.
2. The 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, p. 1, 2012.
3. M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.
4. L. Hindorff, J. MacArthur, J. Morales, H. Junkins, P. Hall, A. Klemm, and T. Manolio. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies.
5. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.
6. R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers: information leaks in genome wide association study," in *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 534–544.