Analysis and Protection of Sensitive Information in Gene Expression Datasets

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

ABSTRACT [[OUTLINE-ish]]

With the unprecedented increase in the size of genomic datasets, the quantification and protection of privacy-sensitive information is a vital issue to be addressed for protection of anonymity of the participants of the scientific studies. [[Differential privacy, different types of attacks, inversion attack, linking attack]]

In this paper, we present a comprehensive framework for analysis of sensitive information in the gene expression datasets. We present a general scenario where the gene expression datasets can be exploited to predict eQTL genotypes to link independently distributed anonymized datasets by an adversary to re-identify individuals.

[[1st act::We first analyze the amount of genotype information in each eQTL SNP that can be extracted using the gene expression datasets.]] Motivation: How much of the SNPs can be predicted?

[[2nd act::We next present a 3-step general framework for individual identification.]] Motivation: This analysis can bring important insight into the extent of vulnerability of individuals and what can be predicted, which is important for designing differentially private release algorithms for analysis of gene expression datasets. In addition, the framework that we are presenting can be utilized for the analysis of vulnerability in the future eQTL studies.

It has been shown that the differential privacy approach for releasing genomic information may lead to very poor utility~\cite{XX,XX}. For this, this study the understanding of the predictable sensitive genetic information from gene expression datasets.

This is also the first time a systematic analysis of genetic leakage is analyzed with respect to prediction from gene expression. We also do a cross-dataset analysis of the reproducibility of genetic leakage attack using different datasets.

[[3rd act::We finally propose a practical linking attack method. (Extremity attack)]] Motivation: An example for the practicality of all the analysis.

1 BACKGROUND

[[Define sensitive information: Anything that the individuals do not want leaked]]

[[Previous work: Homer, Schadt, Erlich, …]]

[[Genetic leakage protection: Several of these: De-identification based (removal of names), Encryption based, more complicated de-identification techniques, differential privacy based (makes a very high compromise of utility for privacy sake). Last two are active field of research.]]

[[In this paper, we analyze identifiability of SNP genotypes and identifiability of individuals in the context of linking attacks. These are the most prevalent attacks that can affect the currently generated genomics datasets.]]

[[First, we present an analysis framework that formalizes the analysis of genetic leakage. Our framework decomposes the linking attack into 3 steps that we study in detail.

-- We make the assumption that the attacker recovers the conditional probabilities perfectly, which enables us to be as stringent about what the attacker can predict as possible.

-- We evaluate the incorporation of auxiliary information.

-- Simulate suboptimal conditional probabilities?

This framework can be used for leakage analysis in the future studies.

We finally present a practical attack for prediction of genotypes from gene expression levels.]]

2 RESULTS

2.1 Overview of the Privacy Breaching Scenario

Figure 1 illustrates the privacy breaching scenario that is considered. The breach occurs by linking two datasets such that one of the datasets contains the individual identities and corresponding genotypes and the second dataset contains the gene expression levels and sensitive information (e.g. disease status) about each individual. The second dataset is assumed to be anonymized by removal of the individual identities to protect the individuals. The adversary gains access to both datasets and links the datasets to associate the sensitive information to individuals. While performing the linking “attack” the adversary utilizes publicly available databases. In the considered scenario, the eQTL databases are utilized which enable linking the expression levels to the genotypes.

[[Notations:

The gene expression and genotype datasets are stored in $n_g \times N$ and $n_e \times N$ matrices g and e, respectively, where N represents the number of individuals and n_g and n_e denote the number of variants and genes, respectively. k^{th} row of e, e_k, contains the expression values for k^{th} gene and e_k^j represents the expression of the k^{th} gene for j^{th} individual. Similarly, l^{th} row of g, g_l, contains the genotypes for l^{th} variant and g_l^j represents the genotype ($g_l^j \in \{0,1,2\}$) of l^{th} variant for j^{th} individual. We will denote the random variables (RVs) whose values represent that the gene expression of k^{th} gene and the variant genotypes for l^{th} variant with $\{E_k\}$ and $\{G_l\}$, respectively. These random

The eQTL dataset contains n eQTLs as a set of gene and variant RV pairs $\{(E_{a_i}, G_{b_i})\}$, $i < n, a_i < n_e, b_i < n_g$ such that there is significant correlation between E_{a_i} and G_{b_i}. We will denote the correlation with $\rho(E_{a_i}, G_{b_i})$. The sign of $\rho(E_{a_i}, G_{b_i})$ represents the direction of association, i.e., which genotype corresponds to higher expression and the magnitude represents the strength of the association. Figure 2a shows the fraction of eQTLs with different correlation cutoffs.

]]

2.2 General Individual Identification Model

[[Introduce the 3-step individual identification model]]

2.2.1 eQTL Selection and Genotype Prediction Accuracy

[[eQTL selection stats and genotype prediction accuracy]]

2.2.2 Individual Identification Accuracy

[[Individual identification accuracy: Nearest neighbor matching using the predicted genotypes]]

[[Auxiliary Information: Gender and/or Population]]

2.2.3 Population Identification Accuracy

[[Population accuracy statistics]]

2.2.4 [[OPTIONAL]] Cross Study Individual Identification Accuracy

[[Stranger et al eQTLs on GEUVADIS expression data results]]

Identifiability of individuals requires generating features that are unique to certain individuals in the sample. The eQTLs are not suitable for identifying individuals since they are common variants. Although each eQTL genotype is common, the frequencies for genotypes of SNP combinations can be small and these combinations, which can be well predicted, can be utilized for identifying individuals. By using a similar approach in Section 2.1, we first quantify the amount of individual identifying information that is leaked in the gene expression datasets. This quantification enables us to evaluate the bounds on the amount of information that can be extracted from the expression levels about the genotype data, which

can then be utilized for anonymization of expression dataset so as to guarantee a quantifiable privacy level in the released dataset.

Individual identification is basically generating a feature that can distinguish, or discriminate, an individual from all others in the dataset. Since we are aiming to do this via prediction of genotypes from expression levels, we will utilize surprisal estimated in terms of self-information. This measure captures discriminative power of the genotypes and also enable us to measure predictability in terms of mutual information.

[[Figure 4a shows the predictability versus discrimination power of the top eQTLs]]

[[Figure 4bc individual identifying information and leakage quantification]]

2.3 Genotype Prediction by Extremity Attack

This analysis is useful for getting quantification of leaked genetic information from gene expression datasets. To predict the eQTL genotypes from gene expression levels, we propose using a method that we name "extremity attack". In this attack, given one gene whose expression level correlates with a variant. The prediction utilizes a statistic we termed $extremity$ of gene expression level which quantifies how extreme an individual's gene expression level is away from the mean of the distribution. Given the gene expression level, e, for a $extremity$ is defined as following:

$$extremity(e) = \frac{rank\ of\ e}{N} - 0.5.$$

Extremity is bounded between -0.5 and 0.5. Figure 3a illustrates the extremity attack. The adversary utilizes the extremity and the gradient of association between the gene expression level to assign a genotype to the associated variant.

[[Figure 3bc shows the accuracy of extremity attack with different extremity and correlation thresholds.]]

2.4 Individual Identification with Extremity Attack

[[Fig. 5a; Distribution of the maximum of absolute extremity over all the samples. How well does expression extremity identify individuals? It is mostly uniform except for some samples.]]

We analyze the

To formalize the analysis using the low frequency multi-SNP genotypes, we utilize the k-anonymization framework. K-anonymization formalizes a way to identify the number of vulnerable individuals and also to ensure the anonymization, which is presented in Section 2.5. Briefly, in order to identify the individuals that are vulnerable to the linking attack, we identify the individuals that have the low

frequency multiple SNP genotypes such that all the SNP genotypes are highly predictable using the expression dataset.

[[External information: 1 bits of gender information can be easily predicted from ; how does this change vulnerability; this justifies the fact that we need “buffering” in anonymization to protect against unaccounted external information that may cause increased vulnerability.]]

2.5 Anonymization

[[Do anonymization for all possible parametrizations to decrease the privacy loss to minimum]]

[[k-anonymization formality for guaranteeing anonymity]]

3 METHODS

3.1 Quantification of Genotype Information Content and Loss of Privacy

[[MI and entropy based definition of IC and Loss of Privacy]]

[[Must justify with MI computation method]]

3.2 Extremity Attack

[[Define the extremity attack: Correlation and extremity parameters]]

3.3 K-Anonymization

[[Define k-anonymization]]

[[Present in detail the anonymization procedure that we propose]]

4 CONCLUSION AND DISCUSSION

In this paper we present a simple framework for quantification of the sensitive information leakage in the linking attack scenarios. The premise of sharing genomic information is that there is always an amount of leakage in the sensitive information. We believe that this quantification methodology can be utilized for more extensive analysis of the leakage in sensitive information for high level correlations in the genomic datasets. The quantification can be further developed for guaranteeing bounds on anonymized datasets.

[[How does this framework compare to other formalities? For example differential privacy? It is similar but differential privacy does not enable quantification of the leakage.]]

We also presented a simple attack that is based on using extremity statistic to predict genotypes that can implicate the sensitive information. Compared to previous approaches, this statistic is very easy to compute.

5 Datasets

[[GEUVADIS dataset, and eQTLs, 1000 genomes dataset]]

[[Other eQTL datasets?]]

6 Quantification of Individual Identifying Information [[OBSOLETE]]

To quantify the individual identifying information, we use surprisal, measured in terms of self-information of the genotypes:

$$III(g) = I(G = g) = -\log\left(p(G = g)\right)$$

where G is an eQTL variant and g ($g\epsilon\{0,1,2\}$) is a specific genotype for G, $p(G = g)$ is the probability frequency of the genotype in the sample set and III denotes the individual identifying information. Assessing this relation, the genotypes that have low frequencies have high identifying information, as expected. Given multiple eQTL genotypes, assuming that they are independent, the total individual identifying information is simply summation of those:

$$III(\{G_1 = g_1, G_2 = g_2, \dots, G_N = g_N\}) = -\sum_{i=1}^{N} \log\big(p(G_i = g_i)\big).$$

The individual identifying information after the gene expression levels are revealed is basically the conditional III given the gene expression levels:

$$III_{remaining}(\{G_1 = g_1, G_2 = g_2, \dots\} \mid \{E_1 = e_1, E_2 = e_2, \dots\}) = -\sum_{i=1}^{N} \log\left(p(G_i = g_i \mid E_i = e_i)\right)$$

where E_i represents the gene expression level for the i^th gene, which is associated with the genotype of G_i. The leakage in III is the remaining III after expression levels are revealed:

$$III_{leaked} = III - III_{remaining}.$$