Analysis of Individual Characterizing Information Leakage in Gene Expression and Genotype Datasets

Arif Harmanci, Jieming Chen, Dov Greenbaum, Mark Gerstein

ABSTRACT

WD

C:HAR

~~The~~Genomic privacy is gaining much attention with the unprecedented increase in the breadth and depth of personal "-omic" datasets ~~enforces the data sharing mechanisms to adapt to~~ . While most of the ~~risks associated with leakage~~ studies are focused on protection of genomic variants in personal genomes, the analysis of sensitive ~~personal medical~~ information~~. The genome-wide studies on association between the genetic variants and the~~ leakage in molecular phenotypic ~~profiling data have identified correlations~~ datasets, like functional genomics datasets, is in its inception. A significant amount of leakage can be caused by the phenotype –to-genotype correlations identified by the genome-wide studies, where associations between large number of genetic loci and different molecular phenotypes~~.~~ are discovered. Although these correlations are valuable for ~~biological~~ understanding ~~of~~ how phenotype and genotype ~~interacts, they can serve to~~ interact, an adversary ~~as a backdoor~~ can utilize the phenotype-to-genotype correlations for predicting the variant genotypes ~~from phenotypes or vice versa.~~for the individuals, for which only phenotype information is available. When ~~the prediction is done~~performed over ~~very~~a large number of ~~genetic loci, this allows~~predicted genotypes the adversary ~~to~~can accurately link the entries in ~~genotype and~~ phenotype datasets ~~so as~~to the entries in a genotype dataset to reveal sensitive ~~phenotypic information about individuals. Even though majority the genomic privacy studies has focused solely on protection of genetic variants, it is necessary to analyze how these correlations can lead to a linking attack with other datasets and lead to privacy breach~~information.

In this paper, we study the characterizability of individuals in the context of linking attacks, where an adversary aims at revealing an individual's sensitive information by ~~matching, or~~ linking~~,~~ the entries in phenotype and genotype datasets. While doing this the attacker utilizes a third dataset that contains the ~~genotype-to-~~phenotype-to-genotype correlations. We focus on the correlations between genotypes and gene expression levels reported in eQTL datasets. We first ~~perform a quantitative analysis~~quantitatively assess the relation between the amount of information leakage the adversary can cause by genotype prediction and ~~the correct predictability of the genotypes~~how accurately the leakage can be performed. We propose two quantification metrics that can be used for evaluating the amount of leakage at different levels of ~~prediction.~~correct predictability. We then present a ~~generalized~~general framework for analysis of the individual characterization and evaluate the fraction of characterizable individuals in a general setting on the representative dataset. For ~~a~~ illustrating the practicality of these analyses, we present a simple practical genotype prediction method, which, when employed on a representative dataset, yields a significant fraction of individuals characterizable. Overall, the quantification metrics and

GL?

SINCE

the analysis framework can be utilized ~~to~~in analysis of individual characterizability in other ~~genotype-to-~~studies where phenotype~~-correlation studies~~-to-genotype correlations are investigated.