Privacy in Genome Studies & Differentially Private Regression

CompSci 590.03
Instructor: Ashwin Machanavajjhala
Outline

• Genome Wide Association Studies
  – Privacy Attacks

• Private Regression
GENOME WIDE ASSOCIATION STUDIES
Gene6c

• Although any two unrelated people are the same at about 99.9% of their DNA sequences, the remaining 0.1% is important because it contains the genetic variants that influence how people differ in their risk of disease or their response to drugs.

• Discovering the DNA sequence variants that contribute to common disease risk offers one of the best opportunities for understanding the complex causes of disease in humans.
2007 SCIENTIFIC BREAKTHROUGH OF THE YEAR
Science Magazine, December 21, 2007

“It’s all about me!”

Single Nucleotide Polymorphisms (SNPs)

Individual 1

Individual 2

Individual 3

Individual 4

Lecture 19: 590.03 Fall 13
SNPs

- Sites in the genome where the DNA sequences of many individuals differ by a single base are called single nucleotide polymorphisms (SNPs).

  .... A C C A C C T ....
  .... A C C G C C T ....

- Each form (A or G) is called an allele

- About 10 million SNPs exist in human populations, where the rarer SNP allele has a frequency of at least 1%.
Haplotypes

• Alleles of SNPs that are close together tend to be inherited together.

• A set of associated SNP alleles in a region of a chromosome is called a **haplotype**.

  ... A ... C ... A ... T ... G ... T ...
  ... A ... C ... C ... G ... C ... T ...
  ... G ... T ... C ... G ... G ... A ...

  **The two SNPs in bold are sufficient to identify (tag) each of the three haplotypes**

• Only a few common haplotypes (each with a frequency of at least 5%) account for most of the variation in a population.
HapMap Project

Genome Wide Association Study

DNA from different individuals sequenced

Variation at a single nucleotide

Some individuals will have one version of the SNP, some the other

Sample with disease

Normal population

A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)

In a population, a certain percentage will have one version, the rest the other

© Gibson & Muse, A Primer of Genome Science
Genome Wide Association Studies

• Pick a disease
• Pick a subsets of the population with (case) and without (control) the disease
• Extract SNPs at selected sites
  – Usually 2 alleles are found (major – 1 and minor – 0)

• Compute allele frequencies in both the case and control groups

• Perform association test using these frequencies
  – Logistic Regression, Chi Square Test, ...

• Putative SNPs: highly associated with the disease (small p-value)
Genome Wide Association Studies

[ from Wang et al CCS ‘09]

- Linkage Disequilibrium analysis:
  Map associations of putative SNPs with other “nearby” SNPs
  - Compute correlations (e.g. $r^2$ coefficient) between SNPs
  - Helps identify other SNPs also related to the disease

\[
r^2 = \frac{(C_{00}N - C_{0*}C_{0*})^2}{C_{0*}C_{1*}C_{0*}C_{1*}}
\]

<table>
<thead>
<tr>
<th>Alleles</th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$C_{00}$</td>
<td>$C_{01}$</td>
<td>$C_{0*}$</td>
</tr>
<tr>
<td>1</td>
<td>$C_{10}$</td>
<td>$C_{11}$</td>
<td>$C_{1*}$</td>
</tr>
<tr>
<td>Sum</td>
<td>$C_{*0}$</td>
<td>$C_{*1}$</td>
<td></td>
</tr>
</tbody>
</table>
# Genetic determinants of diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene or Loci</th>
<th>Date Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate Cancer</td>
<td>8q24</td>
<td>April 1, 2007</td>
</tr>
<tr>
<td>Acute Lymphoblastic Leukemia</td>
<td>PAX 5 and others</td>
<td>April 12, 2007</td>
</tr>
<tr>
<td>Obesity</td>
<td>FTO</td>
<td>April 12, 2007</td>
</tr>
<tr>
<td>Multiple Solid Tumors</td>
<td>CASP8</td>
<td>April 22, 2007</td>
</tr>
<tr>
<td>Diabetes, Type II</td>
<td>CDKAL1 and 6 others</td>
<td>April 26, 2007</td>
</tr>
<tr>
<td>Myocardial Infarction, Coronary Artery Disease</td>
<td>9p21</td>
<td>May 3, 2007</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>FGFR2, TNCR9, MAP3K1, LSP and others</td>
<td>May 27, 2007</td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>IRGM</td>
<td>June 7, 2007</td>
</tr>
<tr>
<td>Diabetes, Type I</td>
<td>12q24 and others</td>
<td>June 7, 2007</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>16p12</td>
<td>June 7, 2007</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>6p21, 1p13</td>
<td>June 7, 2007</td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>4q25</td>
<td>July 1, 2007</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
PRIVACY & GENOME DATA
Privacy issues

• If identified, genome data can be used to predict individual’s susceptibility to diseases

• Can be used by employers to deny health insurance coverage

• May disclose information about immediate family
De-identified genome data

A number of re-identification attacks even if only the genome is published (in full) using

• Genotype/Phenotype inference

• Familial structures
Genotype/Phenotype Re-identification

• Health(Name, Address, Birthdate, Gender, Zip Code, Hospital Visit Date, Diagnosis, Treatment)
• Genomic(DNA)

• There exist many diseases to which DNA mutations are directly related.
  – We can construct mappings between DNA and diseases
• Relationships between genetic variations and the ability to process drugs and treatments
  – We can construct mappings between DNA and treatment
• Can predict the age of onset of progressive disorders from genome.
Re-identification through Familial Structures

- Many genome datasets include familial structures
- One can identify disease/gender patterns from the de-identified genome data
- This can be matched with (identified) familial structures extracted from public and official records (e.g., ancestry.com and social security death index)

[Malin AMIA ‘06]
NIH Policy (before 2008)

• Controlled access to de-identified genetic data (genotypic and phenotypic)

• Public access to aggregate genotypic data.

Statistical re-identification attacks on aggregate genetic data

• Statistical methods can accurately predict whether an individual’s DNA is in the mixture just using aggregate statistics about the mixture.

  – Using allele frequencies [Homer et al PLOS Genetics 2008]
  – Using p-values from research publications [Wang et al CCS 2009]
"We have removed aggregate genotype data for GWAS studies from public access, but may make them available through the controlled access DAR/DAC process."
Homer et al attack

Suppose a study published the allele frequencies for a sample $M$ of the population having disease $D$.

Want to find out whether a particular individual $X$ is in $M$ (and therefore $X.\text{disease} = D$).

Solution:

- Attacker obtains a high density SNP profile (knows hundreds of thousands of SNPs) of the individual $X$.
- $Y_j = 0$ or $1$ (major / minor) for SNP $j$ of the individual
- $M_j = \text{fraction of the sample having SNP } j = 1$
- $\text{Pop}_j = \text{fraction of a reference population having SNP } j = 1$
Homer et al attack

\[ D(Y_{ij}) = |Y_{ij} - Pop_j| - |Y_{ij} - M_j| \]

- \( Y_{ij} = \) allele freq of individual I at SNP j
- \( M_j = \) allele frequency of the mixture at SNP j
- \( Pop_j = \) allele frequency of ref popula

- Use statistical hypothesis testing to check if mean of \( D(Y_j) > 0 \).

\[ T(Y_i) = \frac{\mu - \mu_0}{\sigma / \sqrt{s}} \]

- \( \mu_0 = \) mean of \( D(Y_{ij}) \) over individuals not in the mixture
- \( \sigma = \) standard deviation of \( D(Y_{ij}) \)
- \( s = \) number of SNPs
Homer et al attack

- Compute: \( D(Y_j) = |Y_j - \text{Pop}_j| - |Y_j - M_j| \)
- Use statistical hypothesis testing to check if mean of \( T(Y_j) > 0 \).
Using p-values from research papers

• Typically, all allele frequencies are not published (especially after NIH’s revised guidelines)

• Research papers publish:
  – Allele frequencies of putative (not all) SNPs for a disease
  – p-values and $r^2$ statistics only are published for other SNPs.

• Can’t use Homer et al attack.

[Wang et al CCS ‘09]
Statistical Attack

\[ r_{ij} = \frac{C_{11}C_{00} - C_{01}C_{10}}{\sqrt{C_{1*}C_{0*}C_{*1}C_{*0}}} \]

• We can compute \( r^R \) (for a reference population)
  – where the \( C_{ij} \) are the allele counts for SNPs \( i \) and \( j \) in the reference population
• We know the \( (r^C)^2 \) values (which are published) in the case population
  – To estimate \( r^C \), we need the sign … can be computed using integer programming using the allele frequencies of the putative SNPs.
Statistical Attack

- Compute $T_r$
- Use statistical hypothesis testing to check if $T_r > 0.$

$$T_r = \sum_{1 \leq i < j \leq N} (r_{ij}^C - r_{ij}^R) \cdot (Y_{ij}^{00} + Y_{ij}^{11} - Y_{ij}^{01} - Y_{ij}^{10})$$

Lecture 19: 590.03 Fall 13
Summary of GWAS

• Genome wide association studies help find genetic determinants of diseases.
  – Statistical methods like regression are used widely

• However, publishing genetic data (even aggregate counts and statistical models) can allow re-identification of individuals.
PRIVATE REGRESSION
Private Techniques for Regression

- In the genome setting, need to account for adversaries who may know significant background knowledge about target individual
- Use differential privacy
Linear Regression

- Set of records \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}
- X : predictor variables (e.g., race, gender, genetic markers)
  - X = (x_1, x_2, \ldots, x_d)
- y : response variable (e.g., medical expenses)

- Want to predict y given new values for X

- Estimation: \[ w^* = \arg\min_w \sum_{i=1}^{n} (y_i - X_i^T \cdot w)^2 \]

  \[ w = (w_1, w_2, \ldots, w_d) \]
Logistic Regression

• Set of records \{(X_1, y_1), (X_2, y_2), ..., (X_n, y_n)\}
• \(X\) : predictor variables (e.g., race, gender, genetic markers)
• \(y\) : binary response variable (e.g., disease outcome)

• Want to predict \(y\) given new values for \(X\)

• Estimation: 
\[
\hat{w} = \arg\min_w \sum_{i=1}^{n} \left( \log \left( 1 + e^{X_i^T \cdot w} \right) - y_i X_i^T \cdot w \right)
\]
Computing w* privately

- Can use Laplace mechanism
  - Add noise to the output w*
  - But, analyzing the sensitivity of the optimization problem is complex.

- Idea: Rather than perturbing the output optimal w*, return the optimal solution to a perturbed optimization problem.
1. Represent objective function as a polynomial in $w_1, w_2, ..., w_d$.

$$f_D(w) = \sum_{x \in D} (y_i - x^T w)^2$$

$$= \sum_{x \in D} y_i^2 - 2 \sum_{j=1}^d \left( \sum_{x \in D} y_i x_j \right) w_j$$

$$+ \sum_{1 \leq j, l \leq d} \left( \sum_{x \in D} x_j x_l \right) w_j w_l$$

$$= a + \sum_{j=1}^d b_j w_j + \sum_{1 \leq j, l \leq d} c_{jl} w_j w_l$$
Functional Mechanism

1. Represent objective function as a polynomial in $w_1, w_2, ..., w_d$.
2. Compute sensitivity:

$$\Delta = 2 \max_{(x,y) \in D} \left( \sum |\text{coefficients of } f_D(w)| \right)$$

$$\leq 2 \max_{(x,y) \in D} \left( y^2 + 2 \sum_{j=1}^{d} yx_j + \sum_{1 \leq j,l \leq d} x_j x_l \right)$$

$$\leq 2(1 + 2d + d^2)$$

$$\sqrt{\sum_{i=1}^{d} x_{id}^2} \leq 1 \text{ and } y_i \in [-1, 1]$$
Functional Mechanism

1. Represent objective function as a polynomial in \( w_1, w_2, \ldots, w_d \).
2. Compute sensitivity.
3. Perturb coefficients

\[
\tilde{f}_D(w) = \tilde{a} + \sum_{j=1}^{d} \tilde{b}_j w_j + \sum_{1 \leq j, l \leq d} \tilde{c}_{jl} w_j w_l
\]

\[
\tilde{a} = a + \text{Lap}\left(\frac{\Delta}{\varepsilon}\right)
\]

\[
\tilde{b}_j = b_j + \text{Lap}\left(\frac{\Delta}{\varepsilon}\right)
\]

\[
\tilde{c}_{jl} = c_{jl} + \text{Lap}\left(\frac{\Delta}{\varepsilon}\right)
\]
Functional Mechanism

1. Represent objective function as a polynomial in $w_1, w_2, \ldots, w_d$.
2. Compute sensitivity.
3. Perturb coefficients
4. Solve the perturbed optimization problem.

As $n$ becomes large, perturbed objective function value tends to real objective value.

All continuous and differentiable functions can be written as (potentially infinite) polynomial of $w_1, \ldots, w_d$. 

Lecture 19: 590.03 Fall 13
Unbounded Objective Functions

• The perturbed objective function may not have a bounded minimum

\[ f_d(w) = w^2, \quad \text{has a minimum} \]

• Example:

\[ \tilde{f}_d(w) = -w^2, \quad \text{doesn't have a minimum} \]

• Solution: Repeatedly perturb the function till you get one with a bounded minimum (with budget \( \varepsilon \)).
  – Guaranteed \( 2\varepsilon \) differential privacy.
Summary

• Genome wide association studies help find genetic determinants of diseases.
  – Statistical methods like regression are used widely

• However, publishing genetic data (even aggregate counts and statistical models) can allow re-identification of individuals.

• Technique to perform regression privately.

• Perturb the objective function rather than the output model.
References

