11-2018

Protecting Patient Privacy in Genomic Analysis

David Wu
Stanford University

Follow this and additional works at: https://repository.upenn.edu/admindata_conferences_presentations_2018

https://repository.upenn.edu/admindata_conferences_presentations_2018/28

DOI https://doi.org/10.23889/ijpdxs3i5.1046

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_conferences_presentations_2018/28
For more information, please contact repository@pobox.upenn.edu.
Abstract
Patient genomes are interpretable only in the context of other genomes. However, privacy concerns over genetic data oftentimes deter individuals from contributing their genomes to scientific studies and prevent researchers from sharing their data with the scientific community. In this talk, I will describe how we can leverage secure multiparty computation techniques from modern cryptography to perform useful scientific computations on genomic data while protecting the privacy of the participants’ genomes. In multiple real scenarios, our methods successfully identified the disease-causing genes and even discovered previously unrecognized disease genes, all while keeping nearly all of the participants’ most sensitive genomic information private. We believe that our techniques will help make currently restricted data more readily available to the scientific community and enable individuals to contribute their genomes to a study without compromising their personal privacy.

The material from this talk is based on joint works with Gill Bejerano, Bonnie Berger, Johannes A. Birgmeier, Dan Boneh, Hyunghoon Cho, and Karthik A. Jagadeesh.

Comments
DOI https://doi.org/10.23889/ijpds.v3i5.1046
Protecting Patient Privacy in Genomic Analysis

David Wu

based on joint works with:
Gill Bejerano, Bonnie Berger, Johannes A. Birgmeier, Dan Boneh, Hyunghoon Cho, and Karthik A. Jagadeesh
The Era of “Big Data”
Data Collection and Data Breaches

<table>
<thead>
<tr>
<th>Entries</th>
<th>Database</th>
<th>Category</th>
<th>Dump Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>358,676,097</td>
<td>Myspace.com</td>
<td>Social Media</td>
<td>2013-06</td>
</tr>
<tr>
<td>153,004,874</td>
<td>Adobe.com</td>
<td>Software</td>
<td>2013-10</td>
</tr>
<tr>
<td>117,046,470</td>
<td>LinkedIn.com</td>
<td>Social Media</td>
<td>2012</td>
</tr>
<tr>
<td>77,039,888</td>
<td>Edmodo.com</td>
<td>Education</td>
<td>2017-05</td>
</tr>
<tr>
<td>68,743,269</td>
<td>Neopets.com</td>
<td>Gaming</td>
<td>2013-10</td>
</tr>
<tr>
<td>36,397,296</td>
<td>AshleyMadison.com</td>
<td>Dating</td>
<td>2015-07</td>
</tr>
<tr>
<td>16,500,334</td>
<td>Zomato.com</td>
<td>Food &amp; Drink</td>
<td>2017-05</td>
</tr>
<tr>
<td>6,054,459</td>
<td>Xat.com</td>
<td>Chatroom</td>
<td>2015-11</td>
</tr>
<tr>
<td>5,960,654</td>
<td>Adobe.com Common Passwords</td>
<td>Software</td>
<td>2013-10</td>
</tr>
</tbody>
</table>

Database breaches have become the norm rather than the exception...

[Data taken from Vigilante.pw]
Genomics in the Era of Big Data

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore’s law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.

Source: Nature, 2014
Genealogy Databases Enable Naming Of Anonymous DNA Donors

CAMBRIDGE, MASSACHUSETTS—One afternoon in March last year, Yaniv Erlich sat down at his computer to do an experiment. Before Privacy concerns have been raised about publicly accessible genome data before. A study 5 years ago showed that individuals

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,1,2,3,4 Amy L. McGuire,5 David Golan,6 Eran Halperin,7,8,9 Yaniv Erlich1*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.
Privacy-Preserving Genomics

Finding a tradeoff between functionality and privacy
Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

What gene causes a specific (rare) disease?

Patients with Kabuki Syndrome

Each patient has a list of 200-400 rare variants over \(\approx 20,000\) genes
Rare Disease Diagnosis

Patients with Kabuki Syndrome

Each patient has a list of 200-400 rare variants over ≈20,000 genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>A1BG</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>⋮</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZZZ3</td>
<td>0</td>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Each patient has a vector \( \mathbf{v} \) where \( v_i = 1 \) if patient has a rare variant in gene \( i \)

**Goal:** Identify gene with most variants across all patients

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]
Rare Disease Diagnosis

Each patient has a list of 200-400 rare variants over \( \approx 20,000 \) genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>A1BG</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>ZZZ3</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Each patient has a vector \( \nu \) where \( \nu_i = 1 \) if patient has a rare variant in gene \( i \).

Goal: Identify gene with most variants across all patients.

Patients with Kabuki Syndrome

Works well for Mendelian (monogenic) diseases (estimated to affect \( \approx 10\% \) of individuals).
Patients often in geographically-diverse locations

Question: Can we perform this computation without seeing complete patient genomes?
Rare Disease Diagnosis

Each patient has a list of 200-400 rare variants over ≈20,000 genes

Patients “secret share” their data with two non-colluding hospitals

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]
Rare Disease Diagnosis

Each patient has a list of 200-400 rare variants over ≈20,000 genes

Hospitals run a multiparty computation (MPC) protocol on pooled inputs

Patients “secret share” their data with two non-colluding hospitals

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]
Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

Patients with Kabuki Syndrome

Each patient has a list of 200-400 rare variants over \( \approx 20,000 \) genes

MPC Protocol

Top variants (sorted): \textbf{KMT2D}, COL6A1, FLNB

Known cause of disease
Rare Disease Diagnosis

Jagadeesh-W-Birgmeier-Boneh-Bejerano [Science 2017]

- Each patient has a list of 200-400 rare variants over ≈20,000 genes

- Top variants (sorted): KMT2D, COL6A1, FLNB

- Other variants that the patients possess are kept hidden
Experimental benchmarks for identifying causal gene in small disease cohort
  • Simulated two non-colluding entities with 1 server on East Coast and 1 on West Coast
Modern cryptographic tools enable useful computations while protecting the privacy of individual genomes.
Modern cryptographic tools enable useful computations while protecting the privacy of individual genomes.
Conclusions

- Privacy and functionality are not inherently incompatible
- Modern cryptographic tools enable computation on private data
- **Question:** What privacy challenges are there in your area, and what kind of cryptographic tools can we use to address them?

**Project Website:**

https://crypto.stanford.edu/~dwu4/genomepriv-project.html

Thank you!