Privacy-preserving Information Sharing: Tools and Applications (Volume 2)

Emiliano De Cristofaro
University College London (UCL)
https://emilianodc.com
Prologue

Privacy-Enhancing Technologies (PETs):
Increase privacy of users, groups, and/or organizations

PETs often respond to privacy threats
Protect personally identifiable information
Support anonymous communications
Privacy-respecting data processing

Another angle: privacy as an enabler
Actively enabling scenarios otherwise impossible w/o clear privacy guarantees
Sharing Information w/ Privacy

Needed when parties with limited mutual trust willing or required to share information

Only the required minimum amount of information should be disclosed in the process
Private Set Intersection?

**DHS** (Terrorist Watch List) and **Airline** (Passenger List)
Find out whether any suspect is on a given flight

**IRS** (Tax Evaders) and **Swiss Bank** (Customers)
Discover if tax evaders have accounts at foreign banks

**Hoag Hospital** (Patients) and **SSA** (Social Security DB)
Patients with fake Social Security Number
Genomics
Cracking pace
Numbers of genomes sequenced

2014: 229,000
2015: 422,000
2016: 952,000
2017: 1,620,000

Source: Illumina

From: The Economist
Genome sequencing identifies major causes of severe intellectual disability

Christian Gilissen¹*, Jayne Y. Hehir-Kwa¹*, Djie Tjwan Thung¹, Maartje van de Vorst¹, Bregje W. M. van Bon¹, Marjolein H. Willemsen¹, Michael Kwint¹, Irene M. Janssen¹, Alexander Hoischen¹, Annette Schenck¹, Richard Leach², Robert Klein², Rick Tearle², Tan Bo¹,³, Rolph Pfundt¹, Helger G. Yntema¹, Bert B. A. de Vries¹, Tjitske Kleefstra¹, Han G. Brunner¹,⁴*, Lisenka E. L. M. Vissers¹* & Joris A. Veltman¹,⁴*
THE ANGELINA EFFECT

Angelina Jolie's double mastectomy puts genetic testing in the spotlight. What her choice reveals about calculating risk, cost and peace of mind.

BY JEFFREY KLUGER & ALICE PARK
### Genetic Risk Factors

<table>
<thead>
<tr>
<th>REPORT</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha-1 Antitrypsin Deficiency</td>
<td>Variant Absent; Typical Risk</td>
</tr>
<tr>
<td>Alzheimer's Disease (APOE Variants)</td>
<td>ε4 Variant Absent</td>
</tr>
<tr>
<td>Early-Onset Primary Dystonia (DYT1-TOR1A-Related)</td>
<td>Variant Absent; Typical Risk</td>
</tr>
<tr>
<td>Factor XI Deficiency</td>
<td>Variant Absent; Typical Risk</td>
</tr>
<tr>
<td>Familial Hypercholesterolemia Type B (APOB-Related)</td>
<td>Variant Absent; Typical Risk</td>
</tr>
</tbody>
</table>

See all 11 genetic risk factors...

### Inherited Conditions

<table>
<thead>
<tr>
<th>REPORT</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta Thalassemia</td>
<td>Variant Present</td>
</tr>
<tr>
<td>ARSACS</td>
<td>Variant Absent</td>
</tr>
<tr>
<td>Agenesis of the Corpus Callosum with Peripheral Neuropathy (ACCPN)</td>
<td>Variant Absent</td>
</tr>
<tr>
<td>Autosomal Recessive Polycystic Kidney Disease</td>
<td>Variant Absent</td>
</tr>
<tr>
<td>Bloom's Syndrome</td>
<td>Variant Absent</td>
</tr>
<tr>
<td></td>
<td>See all 43 carrier status...</td>
</tr>
</tbody>
</table>

### Traits

<table>
<thead>
<tr>
<th>REPORT</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol Flush Reaction</td>
<td>Does Not Flush</td>
</tr>
<tr>
<td>Bitter Taste Perception</td>
<td>Can Taste</td>
</tr>
<tr>
<td>Blond Hair</td>
<td>28% Chance</td>
</tr>
<tr>
<td>Earwax Type</td>
<td>Wet</td>
</tr>
<tr>
<td>Eye Color</td>
<td>Likely Brown</td>
</tr>
</tbody>
</table>

See all 41 traits...

### Drug Response

<table>
<thead>
<tr>
<th>REPORT</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proton Pump Inhibitor (PPI) Metabolism (CYP2C19-related)</td>
<td>Rapid</td>
</tr>
<tr>
<td>Warfarin (Coumadin®) Sensitivity</td>
<td>Increased</td>
</tr>
<tr>
<td>Phenytoin Sensitivity (Epilepsy Drug)</td>
<td>Increased</td>
</tr>
<tr>
<td>Sulfonylurea Metabolism</td>
<td>Greatly reduced</td>
</tr>
<tr>
<td>Abacavir Hypersensitivity</td>
<td>Typical</td>
</tr>
</tbody>
</table>

See all 12 drug response...
<table>
<thead>
<tr>
<th>List View</th>
<th>Map View</th>
<th>Surname View</th>
</tr>
</thead>
</table>

| Search matches | Show: both sides | Sort: relationship | 25 per page | 1 - 25 of 424 |

<table>
<thead>
<tr>
<th>Male</th>
<th>You</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Female</th>
<th>2nd to 3rd Cousin</th>
<th>J2a2</th>
<th>Send an Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.68% shared, 5 segments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Female</th>
<th>3rd to 4th Cousin</th>
<th>United States, Alsace-Lorraine (Strasbourg), France, Paternal</th>
<th>Public Match</th>
<th>Send a Message</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.30% shared, 3 segments</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male</th>
<th>3rd to 4th Cousin</th>
<th>H13a1a, R1b1b2</th>
<th>Send an Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.03% shared, 2 segments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Female</th>
<th>3rd to 5th Cousin</th>
<th>H7</th>
<th>Send an Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.45% shared, 2 segments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Female</th>
<th>3rd to 5th Cousin</th>
<th>H1</th>
<th>Send an Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.42% shared, 2 segments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male</th>
<th>3rd to 5th Cousin</th>
<th>United States, Reno, Nevada, San Diego, California, Tucker, Littlefield, Warga</th>
<th>Public Match</th>
<th>Send a Message</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.40% shared, 2 segments</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male</th>
<th>3rd to 5th Cousin</th>
<th>United States, fathers, prince Edward Island, United States, Tucker, Littlefield, Warga</th>
<th>Public Match</th>
<th>Send a Message</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.37% shared, 2 segments</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male, b. 1978</th>
<th>3rd to 6th Cousin</th>
<th>United States, New Jersey, Utah, California, U3b1, T</th>
<th>Send an Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.40% shared, 1 segment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
But… not all data are created equal!
Privacy Researcher’s Perspective

Treasure trove of sensitive information

Ethnic heritage, predisposition to diseases

Genome = the ultimate identifier

Hard to anonymize / de-identify

Sensitivity is perpetual

Cannot be “revoked”

Leaking one’s genome ≈ leaking relatives’ genome
Secure Genomics?

Privacy:

Individuals remain in control of their genome

Allow doctors/clinicians/labs to run genomic tests, while disclosing the required minimum amount of information, i.e.:

(1) Individuals don’t disclose their entire genome
(2) Testing facilities keep test specifics (“secret sauce”) confidential

[BBDGT11]: Secure genomics via PSI

Most personalized medicine tests in < 1 second

Works on Android too
A Strawman Approach for Paternity Test:

On average, \( \sim 99.5\% \) of any two human genomes are identical
Parents and children have even more similar genomes
Compare candidate’s genome with that of the alleged child:

Test positive if percentage of matching nucleotides is greater than \( 99.5 + \tau \)

First-Attempt Privacy-Preserving Protocol:

Use an appropriate secure two-party protocol for the comparison
Private Set Intersection Cardinality (PSI-CA)

Server

\[ S = \{ s_1, \ldots, s_w \} \]

Client

\[ C = \{ c_1, \ldots, c_v \} \]

\[ |S \cap C| \]
A Strawman Approach for Paternity Test:

On average, ~99.5% of any two human genomes are identical
Parents and children have even more similar genomes
Compare candidate’s genome with that of the alleged child:

*Test positive if percentage of matching nucleotides is > 99.5 + \( \tau \)*

First-Attempt Privacy-Preserving Protocol:

Use an appropriate secure two-party protocol for the comparison

PROs: High-accuracy and error resilience

CONs: Performance not promising (3 billion symbols in input)
      In our experiments, computation takes a few days
Wait a minute!

~99.5% of any two human genomes are identical
Why don’t we compare *only* the remaining 0.5%?
We can compare by counting how many

But… We don’t know (yet) where *exactly* this 0.5% occur!
Private RFLP-based Paternity Test

Test Result:
(#fragments with same length)
Secure Function Evaluation

- Private Set Intersection (PSI)
- Authorized PSI
- Private Pattern Matching
- Homomorphic Encryption
- Garbled Circuits

Output reveals nothing beyond test result

- Paternity/Ancestry Testing
- Testing of SNPs/Markers
- Compatibility Testing
- Disease Predisposition

Test result

Individual

Genome

Doctor or lab

Test specifics
Drugs designed for patients’ genetic features

Associating drugs with a unique genetic fingerprint
Max effectiveness for patients with matching genome
Test drug’s “genetic fingerprint” against patient’s genome

Examples:

*tmpt* gene – relevant to leukemia
(1) G->C mutation in pos. 238 of gene’s c-DNA, or (2) G->A mutation in pos. 460 and one A->G is pos. 419 cause the *tpmt* disorder (relevant for leukemia patients)

*hla-B* gene – relevant to HIV treatment
One G->T mutation (known as *hla-B*5701 allelic variant) is associated with extreme sensitivity to abacavir (HIV drug)
Privacy-preserving PM Testing ($P^3$MT)

Challenges:

Patients may refuse to unconditionally release their genomes

Or may be sued by their relatives...

DNA fingerprint corresponding to a drug may be proprietary:

- **We need privacy-protecting fingerprint matching**

But we also need to enable FDA approval on the drug/fingerprint

- **We reduce $P^3$MT to Authorized Private Set Intersection (APSI)**
Authorized Private Set Intersection (APSI)

Server

$S = \{s_1, \ldots, s_w\}$

Client

$C = \{(c_1, \text{auth}(c_1)), \ldots, (c_v, \text{auth}(c_v))\}$

Authorized Private Set Intersection

$S \cap C = \{s_j \in S | \exists c_i \in C : c_i = s_j \land \text{auth}(c_i) \text{ is valid}\}$
Reducing \( \text{P}^3\text{MT} \) to APSI

Intuition:

\[ \text{FDA} = \text{Court}, \quad \text{Pharma} = \text{Client}, \quad \text{Patient} = \text{Server} \]

Patient’s private input set: \( G = \left\{ (b_i \| i) \mid b_i \in \{A, C, G, T\} \right\}^{3 \times 10^9} \)

Pharmaceutical company’s input set: \( fp(D) = \left\{ (b_j^* \| j) \right\} \)

Each item in \( fp(D) \) needs to be authorized by FDA
Pre-Computation

Patient’s pre-processing of the genome: a few days

Optimization:

Patient applies reference-based compression techniques

Input all differences with “reference” genome (0.5%)

Online Computation

Depend (linearly) on fingerprint size – typically a few nucleotides, <1s for most tests

Communication

Depends on the size of encrypted genome (about 4GB)
Open Problems?
Micro-blogging
@Alice and @Bob – Twitter edition

There might be no mutual knowledge/trust between Alice and Bob

Follow requests are approved by default (opt-out)

Tweets are public by default
  Streamed into www.twitter.com/public_timeline, available through API
  But Bob can restrict his tweets to followers

All public tweets are searchable by hashtag
#Privacy and Twitter

Twitter.com is “trusted” to
  - Get all tweets
  - Enforce coarse-grained access control (follower-only)
  - Monitor relations between users

Privacy and Twitter
  - Targeted advertisement, PII collected and shared with third parties
  - Trending topics, real-time “news”

I don’t care about #privacy on @Twitter… but Remember @Wikileaks? Snowden?
Our proposal: Hummingbird

Follow by hashtag:

E.g., @Alice follows @Bob only on hashtag #privacy

Tweeter (@Bob)

Learns who follows him but not which hashtags have been subscribed to

Follower (@Alice)

Learns nothing beyond her own subscriptions

Server (HS)

Doesn’t learn tweets’ content or hashtags of in
(But can scale to million of tweets/users)
### Issue Request

\[(N_b, e_b)\]

\[(Alice, Bob, \mu)\]

\[r \in Z_{N_b}\]

\[\mu = H(ht) \cdot r^{e_b}\]

### Approve

\[(Alice, \mu)\]

\[(\mu')\]

\[\mu' = \mu^{d_b}\]

### Finalize Request

\[(Bob, \mu')\]

\[(Alice, Bob, t)\]

\[\delta = \mu'/ r\]

\[t = H'(\delta)\]
For all \((U,V,t)\) s.t. \(V=\text{‘Bob’}\) and \(t=t^*\):

Store and mark \((\text{Bob},t^*,ct^*)\) for delivering \((t^*,ct^*)\) to Alice.

\[
\begin{align*}
\delta &= H(ht^*)^{d_b} \\
t^* &= H'(\delta) \quad k^* = H''(\delta) \\
ct^* &= \text{Enc}_{k^*}(M)
\end{align*}
\]

\[
\begin{align*}
k &= H''(\delta) \\
M &= \text{Dec}_k(ct^*)
\end{align*}
\]
Overhead

Follow protocol: Alice wants to follow Bob on #privacy
  Bob’s computation: 1 CRT-RSA signature (<1ms) per hashtag
  Alice’s computation: 2 mod multiplications per hashtag
  Communication: 2 RSA group elements (<1KB)

Tweet: Bob tweets “I’m at #fosad!”
  Computation: 1 CRT-RSA signature (<1ms) per hashtag, 1 AES enc
  Communication: 1 hash output (160-bit)

Read
  Computation: 1 AES decryption
  Communication: 1 hash output (160-bit)

Server
  No crypto!
  Overhead: matching of PRF outputs, 160-bit
    Can do efficiently, just like for cleartexts
Collecting Statistics Privately?

Collaboratively Train Machine Learning Models, Privately?
Why are statistics important?

Examples:

1. Recommender systems for online streaming services
2. Statistics about mass transport movements
3. Traffic statistics for the Tor Network

How about privacy?
Private Recommendations

BBC keeps 500-1000 free programs on iPlayer

No account, no tracking, no ads

Still, BBC wants to collect statistics, offer recommendations to its users

  E.g., you have watched Dr Who, maybe you’ll like Sherlock Homes too!
Item-KNN Recommendation

Predict favorite items for users based on their own ratings and those of “similar” users

Consider \(N\) users, \(M\) TV programs and binary ratings (viewed/not viewed)

Build a co-views matrix \(C\), where \(C_{ab}\) is the number of views for the pair of programs \((a,b)\)

Compute the Similarity Matrix

\[
\{Sim\}_{ab} = \frac{C_{ab}}{\sqrt{C_a \cdot C_b}}
\]

Identify K-Neighbours (\(KNN\)) based on matrix
Privacy-Preserving Aggregation

Goal: aggregator collects matrix, s.t.

Can only learn aggregate counts (e.g., 237 users have watched both a and b)

Not who has watched what

Use additively homomorphic encryption?

\[ \text{Enc}_{PK}(a) \ast \text{Enc}_{PK}(b) = \text{Enc}_{PK}(a+b) \]

How can I used it to collect statistics?
Keys summing up to zero

Users $U_1$, $U_2$, ..., $U_N$, each has $k_1$, $k_2$, ..., $k_N$ s.t.

$$k_1 + k_2 + \ldots + k_N = 0$$

Now how can I use this?
User $\mathcal{U}_i$ ($i \in [1,N]$)

\[
x_i \in \mathbb{G}, y_i \equiv g^{x_i} \mod q
\]

\[
k_{i\ell} := \sum_{j \neq i} \text{H}(y_j^{x_i} || \ell || s) \cdot (-1)^{i > j} \mod 2^{32}
\]

\[
b_{i\ell} := X_{i\ell} + k_{i\ell} \mod 2^{32}
\]

\[
k'_{i\ell} := \sum_{j \neq i, j \notin \mathcal{U}_i^{on}} \text{H}(y_j^{x_i} || \ell || s) \cdot (-1)^{i > j} \mod 2^{32}
\]

\[
\text{Is this efficient?}
\]

\[
\{b_{i\ell}\}_{\ell=1}^{L} \rightarrow \mathcal{U}_i^{on}
\]

\[
\{k'_{i\ell}\}_{\ell=1}^{L} \rightarrow C'_\ell := \left( \sum_{i \in \mathcal{U}_i^{on}} b_{i\ell} - \sum_{i \in \mathcal{U}_i^{on}} k'_{i\ell} \right) \mod 2^{32}
\]
Preliminaries: Count-Min Sketch

An estimate of an item’s frequency in a stream

Mapping a stream of values (of length $T$) into a matrix of size $O(\log T)$

The sum of two sketches results in the sketch of the union of the two data streams
Security & Implementation

Security

In the honest-but-curious model under the CDH assumption

Prototype implementation:

Tally as a Node.js web server

Users run in the browser or as a mobile cross-platform application (Apache Cordova)

Transparency, ease of use, ease of deployment
Accuracy
Aggregate statistics about the number of hidden service descriptors from multiple HSDirs

Median statistics to ensure robustness

Problem: Computation of statistics from collected data can potentially de-anonymize individual Tor users or hidden services
Private Tor Statistics?

We rely on:
- A set of authorities
- A homomorphic public-key scheme (AH-ECC)
- Count-Sketch (a variant of CMS)

Setup phase
- Each authority generates their public and private key
- A group public key is computed
Each HSDir (router) builds a Count-Sketch, inserts its values, encrypts it, sends it to a set of authorities.

The authorities:

- Add the encrypted sketches element-wise to generate one sketch characterizing the overall network traffic.
- Execute a divide and conquer algorithm on this sketch to estimate the median.
How we do it (1/2)

The range of the possible values is known

On each iteration, the range is halved and the sum of all the elements on each half is computed

Depending on which half the median falls in, the range is updated and again halved

Process stops once the range is a single element
Output privacy:

Volume of reported values within each step is leaked

Provide differential privacy by adding Laplacian noise to each intermediate value
Evaluating

Experimental setup:
- 1200 samples from a mixture distribution
- Range of values in [0,1000]

Performance evaluation:
- Python implementation (*petlib*)
- 1 ms to encrypt a sketch (of size 165) for each HSDir and
- 1.5 sec to aggregate 1200 sketches
Collaborative Threat Mitigation
Collaborative Anomaly Detection

Anomaly detection is hard

Suspicious activities deliberately mimic normal behavior
But, malevolent actors often use the same resources

Wouldn’t it be better if organizations collaborated?

“It is the policy of the United States Government to increase the volume, timelines, and quality of cyber threat information shared with U.S. private sector entities so that these entities may better protect and defend themselves against cyber attacks.”

Barack Obama
2013 State of the Union Address
Problems with Collaborations

Trust
  Will others leak my data?

Legal Liability
  Will I be sued for sharing customer data?
  Will others find me negligible?

Competitive concerns
  Will my competitors outperform me?

Shared data quality
  Will data be reliable?
Solution Intuition [FDB15]

Company 1

Sharing Information w/ Privacy

Company 2

Securely assess the benefits of sharing

Securely assess the risks of sharing

Better Analytics
1. Estimate Benefits

What are good **indicators** of the fact that sharing will be beneficial?

- Many attackers in common?
- Many similar attacks in common?
- Many correlated attacks in common?
2. Select Partners

How do I **choose** who to collaborate with?

- Collaborate with the top-k?
- Collaborate if benefit above threshold?
- Hybrid?
3. Merge

Once we partnered up, what do we share?

- Everything?
- Just what we have in common?
- Just information about attacks or also metadata?
System Model

Network of $n$ entities $\{V_i\}$ (for $i=1,\ldots,n$)

Each $V_i$ holds a dataset $S_i$ of suspicious events

E.g., events in the form $\langle \text{IP}, \text{time}, \text{port} \rangle$ as observed by a firewall or an IDS
Select

Estimate Benefits
Intersection-Size($S_i, S_j$)
Jaccard($S_i, S_j$)
Correlation($S_i, S_j$)
Cosine($S_i, S_j$)

Partner
Decide
Benefit > threshold
Maximize benefits

Merge
Share
Intersection($S_i, S_i$)
Union($S_i, S_i$)

Select

Estimate Benefits
Intersection-Size($S_i, S_j$)
Jaccard($S_i, S_j$)
Correlation($S_i, S_j$)
Cosine($S_i, S_j$)

Partner
Decide
Benefit > threshold
Maximize benefits

Merge
Share
Intersection($S_i, S_i$)
Union($S_i, S_i$)
### Privacy-preserving benefit estimation

<table>
<thead>
<tr>
<th>Metric</th>
<th>Operation</th>
<th>Private Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection-Size</td>
<td>$</td>
<td>S_i \cap S_j</td>
</tr>
<tr>
<td>Jaccard</td>
<td>$\frac{</td>
<td>S_i \cap S_j</td>
</tr>
<tr>
<td>Pearson</td>
<td>$\sum_{l=1}^{N} \frac{(s_{i_l} - \mu_i)(s_{j_l} - \mu_j)}{N \sigma_i \sigma_j}$</td>
<td>Garbled Circuits (2PC)</td>
</tr>
<tr>
<td>Cosine</td>
<td>$\frac{\vec{S}_i \cdot \vec{S}_j}{| \vec{S}_i | | \vec{S}_j |}$</td>
<td>Private Cosine Similarity (PCS)</td>
</tr>
</tbody>
</table>
Privacy-preserving data sharing

<table>
<thead>
<tr>
<th>Metric</th>
<th>Operation</th>
<th>Private Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection</td>
<td>$</td>
<td>S_i \cap S_j</td>
</tr>
<tr>
<td>Intersection with Associated Data</td>
<td>${\langle \text{IP, time, port} \rangle</td>
<td>IP \in S_i \cap S_j }$</td>
</tr>
<tr>
<td>Union with Associated Data</td>
<td>${\langle \text{IP, time, port} \rangle</td>
<td>IP \in S_i \cup S_j }$</td>
</tr>
</tbody>
</table>
The Road Ahead…

This slide is intentionally left blank