Robust Traceability from Trace Amounts

Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, Salil Vadhan
A motivating story

Is TCS related to a genetic mutation?

Case Group
What the data looks like (in theory)

A Berkeley genome-wide association study has discovered a link between TCS and part of the human genome.

Lead scientist Dr. Ekaf Nosrep said "identifying the genetic markers of TCS may prove useful in developing a cure."
Privacy concerns

Can I share my data with other TCS researchers?

That’s not OK with us!

Perhaps I can just remove identifiers to protect subjects.
# Aggregate data

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>0</td>
<td>-.5</td>
<td>.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-.5</td>
<td>0</td>
<td>-1</td>
<td>.5</td>
</tr>
</tbody>
</table>
Aggregate data

I can safely share the aggregates, right?

Wrong!
[HSR+08, SOJH09, DN03, BUV14, ..., This Work]

Led to changes in how NIH deals with releasing genetic data.
Fundamental law of information recovery

Releasing “overly accurate” estimates of “too many” aggregate statistics is not private.

[DN03, DMT07, HSR+08, DY08, SOJH09, MN12, BUV14, SU15, ...]
Genetics work [HSR+08, SOJH09,...]

Given the exact aggregate statistics for the case group and the data of one individual, and a large reference sample, I can determine whether that individual is in the case group.

Requires $d=\Theta(n)$ attributes.

$m=\Theta(n)$ samples from the same population as the case group.

“Tracing”
Differential privacy

By releasing approximate instead of exact aggregate statistics, we can prevent tracing (and other privacy attacks) for up to \( d = \tilde{O}(n^2) \) attributes by using differential privacy [DMNS06, DKM+06,...].
Limits of differential privacy [CFN94,BS95,Tar03,BUV14,SU15,...]

Given approximate aggregate statistics for the case group and full knowledge of the population, I can identify at least one person in the case group with high probability assuming $d \geq \tilde{\Omega}(n^2)$ and an artificial population.

“Fingerprinting codes”

Motivating Question: Is tracing possible when the database comes from a realistic distribution and the tracer has realistic side-information?

assuming $d \geq \tilde{\Omega}(n^2)$ and an artificial population.

strong assumptions
Our results

Given approximate aggregate statistics for the case group and a single reference sample, I can identify at least one person in the case group with high probability assuming $d \geq \tilde{O}(n^2)$ assuming and a population drawn from a rich family of distributions.
The model

Distribution of $p$: “well-spread” product distribution on $[-1,1]^d$

Population: Product distribution $p$ on $[-1,1]^d$

Reference sample: sample from $p$ $z$ in $[-1,1]^d$

Case group: $n$ samples from $p$ $x_1...x_n$ in $[-1,1]^d$

Target $y$

$q=M(x)$

Tracer: $A(y,z,q)$

Accuracy assumption:
$$\|q - \frac{1}{n} \sum_{i=1}^{n} x_i \| \leq \alpha$$

Our result: If $y$ is OUT, tracer says OUT whp. Whp, for some of $y=x_i$, tracer says IN.
### Comparison

<table>
<thead>
<tr>
<th>Approach</th>
<th>“Statistics”</th>
<th>“Fingerprinting”</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[HSR+08,SOJH09]</td>
<td>[CFN94,BS95,T03,BUV14]</td>
<td>“Natural fingerprinting codes”</td>
</tr>
<tr>
<td>Approach</td>
<td>“Statistics” [HSR+08,SOJH09]</td>
<td>“Fingerprinting” [CFN94,BS95,T03,BUV14]</td>
<td>This work</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------</td>
<td>----------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Exact</td>
<td>Approximate</td>
<td>Approximate</td>
</tr>
</tbody>
</table>
### Comparison

<table>
<thead>
<tr>
<th>Approach</th>
<th>“Statistics” [HSR+08,SOJH09]</th>
<th>“Fingerprinting” [CFN94,BS95,T03,BUV14]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Exact</td>
<td>Approximate</td>
<td>Approximate</td>
</tr>
<tr>
<td>Reference</td>
<td>(m = \Theta(n))</td>
<td>(m \cong \infty)</td>
<td>(m = 1)</td>
</tr>
<tr>
<td>Approach</td>
<td>“Statistics” [HSR+08,SOJH09]</td>
<td>“Fingerprinting” [CFN94,BS95,T03,BUV14]</td>
<td>This work</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------</td>
<td>---------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Exact</td>
<td>Approximate</td>
<td>Approximate</td>
</tr>
<tr>
<td>Reference</td>
<td>$m=\Theta(n)$</td>
<td>$m\approx\infty$</td>
<td>$m=1$</td>
</tr>
<tr>
<td>Dimension</td>
<td>$d\geq\Theta(n)$</td>
<td>$d\geq\tilde{\Theta}(n^2)$</td>
<td>$d\geq\tilde{\Theta}(n^2)$</td>
</tr>
</tbody>
</table>
### Comparison

<table>
<thead>
<tr>
<th>Approach</th>
<th>&quot;Statistics&quot; [HSR+08,SOJH09]</th>
<th>&quot;Fingerprinting&quot; [CFN94,BS95,T03,PUV14]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Exact</td>
<td>Approximate</td>
<td>Approximate</td>
</tr>
<tr>
<td>Reference</td>
<td>(m = \Theta(n))</td>
<td>(m \leq \infty)</td>
<td>(m = 1)</td>
</tr>
<tr>
<td>Dimension</td>
<td>(d \geq O(n))</td>
<td>(d \geq \tilde{O}(n^2))</td>
<td>(d \geq \tilde{O}(n^2))</td>
</tr>
<tr>
<td>Identifies</td>
<td>Everyone</td>
<td>1 person</td>
<td>1 person</td>
</tr>
</tbody>
</table>

“Natural fingerprinting codes”
<table>
<thead>
<tr>
<th>Approach</th>
<th>“Statistics”</th>
<th>“Fingerprinting”</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[HSR+08,SOJH09]</td>
<td>[CFN94,BS95,T03, BUV14]</td>
<td></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Exact</td>
<td>Approximate</td>
<td>Approximate</td>
</tr>
<tr>
<td>Reference</td>
<td>$m=\Theta(n)$</td>
<td>$m\approx\infty$</td>
<td>$m=1$</td>
</tr>
<tr>
<td>Dimension</td>
<td>$d\geq\Theta(n)$</td>
<td>$d\geq\tilde{\Theta}(n^2)$</td>
<td>$d\geq\tilde{\Theta}(n^2)$</td>
</tr>
<tr>
<td>Identifies</td>
<td>Everyone</td>
<td>1 person</td>
<td>1 person</td>
</tr>
<tr>
<td>Population</td>
<td>No assumption</td>
<td>“Artificial”</td>
<td>“Rich Family”</td>
</tr>
</tbody>
</table>
### Extension

<table>
<thead>
<tr>
<th>Approach</th>
<th>“Statistics” [HSR+08,SOJH09]</th>
<th>“Fingerprinting” [CFN94,BS95,T03,BUV14,SU15]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>Exact</td>
<td>Approximate</td>
<td>Approximate</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
<td>$m = \Theta(n)$</td>
<td>$m \approx \infty$</td>
<td>$m = 1$</td>
</tr>
<tr>
<td><strong>Dimension</strong></td>
<td>$d \geq O(n)$</td>
<td>$d \geq \tilde{O}(n^2)$</td>
<td>$d \geq \tilde{O}(n^2)$</td>
</tr>
<tr>
<td><strong>Identifies</strong></td>
<td>Everyone</td>
<td>1 person</td>
<td>1 person</td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td>No assumption</td>
<td>“Artificial”</td>
<td>“Rich Family”</td>
</tr>
</tbody>
</table>

Accuracy assumption:

$$\| q - \frac{1}{n} \sum_{i=1}^{n} x_i \| \leq \alpha$$

Smoothly interpolates between extremes
Our tracer

Reference: $z \in \{-1,1\}^d$

Target $y$

Case group: $x_1...x_n \in \{-1,1\}^d$

$q = M(x)$

Tracer: $A(y,z,q)$

IN/OUT
Very simple tracer

\[ A(y, z, q) : \]

Input: \( y, z \in \{-1, 1\}^d, \ q \in [-1, 1]^d \).

Compute \( s = \langle y - z, q \rangle = \langle y, q \rangle - \langle z, q \rangle \).

If \( s \geq \sqrt{8d \log(1/\delta)} \), output \text{IN}; else output \text{OUT}.

"Is \( y \) more correlated with \( q \) than \( z \)?" [HSR+08]
Why does our tracer work?

**Soundness:** If OUT, say OUT whp.

\[ \mathbb{E}[y] = \mathbb{E}[z] \text{ and } q \text{ is independent from } y \text{ and } z. \]

Thus \( \mathbb{E}[\langle y - z, q \rangle] = 0. \)

**Chernoff:** \( \mathbb{P}\left[\langle y - z, q \rangle < \sqrt{8d \log(1/\delta)}\right] \geq 1 - \delta. \)
Why does our tracer work?

Completeness: Say IN for some $y = x_i$.

Lemma: $q$ accurate $\implies \sum_{i=1}^{n} \mathbb{E}[\langle x_i - z, q \rangle] \geq \Omega(d)$.

Azuma: $\mathbb{P}[\sum_{i=1}^{n} \langle x_i - z, q \rangle \geq \Omega(d)] \geq 1 - \delta$.

Thus $\exists i \langle x_i - z, q \rangle \geq \Omega(d/n) \geq \sqrt{8d \log(1/\delta)}$.

So, if $d = O(n^2 \log(1/\delta))$, say IN for some $y = x_i$ whp.
Why does our tracer work?

Lemma: $q$ accurate $\implies \sum_{i=1}^{n} \mathbb{E} [\langle x_i - z, q \rangle] \geq \Omega(d)$.

Intuition ($d = 1$): $q = M(x)$. If $x_1 = \cdots = x_n = b \in \{\pm 1\}$, then $q \approx b$. So “on average” changing $x_i$ changes $q$ by $2/n$. If bias is well-spread, we get correlation on average.
Comparing exact and approximate statistics

Simulation:
- \( n = 100 \)
- \( m = 200 \)
- \( d = 5000 \)

Exact vs. Rounded

Compare to:
Likelihood ratio test
[SOJH09]
Conclusion

We provide a simple and robust tracer that needs less auxiliary information than previous work.

Build on work in genetics and cryptography. Simplified proofs.

Clearer picture of what can(not) be released privately.

Tells us differential privacy correctly quantifies privacy here.

Releasing “overly accurate” estimates of “too many” aggregate statistics is not private.
Experimental results
Experimental results

Simulation:  
- $n=100$
- $m=1$
Rounded to 0.1

(Here we are varying the IN/OUT threshold.)
More experimental results

Simulation:
- $n=100$
- $m=200$
- Rounded to 0.1

The graph shows ROC curves with different values of $d$:
- $d = 200$, $auc = 0.6182$
- $d = 1000$, $auc = 0.8185$
- $d = 5000$, $auc = 0.9596$
- $d = 10000$, $auc = 0.9993$
Why does our tracer work?

Soundness:

In **OUT** case, $y-z$ and $q$ are independent. $E[y]=E[z]$.

Thus $s=<y-z,q> \equiv 0$ whp by Chernoff-Hoeffding bound.

$P[A(y,z,q) \text{ says } \textbf{OUT}] \geq 1-\delta$. 
Why does our tracer work?

**Completeness:**

In **IN** case, y = x_i and q = M(x) are correlated.

Lemma: \( \Sigma_i E[<x_i-z,q>] \geq \Omega(d) \).

Whp \( \Sigma_i <x_i-z,q> \geq \Omega(d) \).

Whp \( <x_i-z,q> \geq \Omega(d/n) \) for at least one \( i \in \{1,2,\ldots,n\} \).

\( P[\exists i \ A(x_i,z,q)=\text{IN}] \geq 1-\delta \).
Very simple tracer (m=1 case)

A(y,z,q): 

Input: y, z ∈ {-1,1}^d, q ∈ [-1,1]^d

Compute s=<y-z,q>=<y,q>-<z,q>

If s ≥ \sqrt{8 \ d \ log(1/\delta)}, output IN; otherwise output OUT.

"Is y more correlated with q than z?"
A(y, z, q):

Input: \(y, z_0, z_1, \ldots, z_m \in \{-1,1\}^d\), \(q \in [-1,1]^d\)

Compute \(s = <y - z_0, [q - \hat{z}] >\)

If \(s \geq 4\alpha \sqrt{d \log(1/\delta)}\), output IN; otherwise output OUT.
More precisely...

Given the aggregate statistics for the case group and the data of one individual,

I can determine whether that individual is in the case group.*

Requires auxiliary information and accuracy assumption.
Tightness

There exists a method to release $\alpha$-approximate aggregate statistics even when $d=\tilde{O}(\alpha^2 n^2)$ that prevents tracing and other attacks.

Corollary: We cannot do better than differential privacy in this setting.

Differential Privacy

[DMNS06, DKM+06, ...]

i.e. Differential privacy is tight.
Simple tracer \((m=O(\log(n)/\alpha^2))\) case

\[ A_\alpha(y, z, q) : \]

**Input:** \( y, z_0, z_1, \ldots, z_m \in \{-1, 1\}^d, \ q \in [-1, 1]^d \).

Let \( \bar{z} = \frac{1}{m} \sum_{i=1}^{m} z_i \).

**Compute** \( s = \langle y - z_0, [q - \bar{z}] \rangle \).

If \( s \geq 4\alpha \sqrt{d \log(1/\delta)} \), output IN; else output OUT.