Privacy-Preserving Data Analysis & Security by Design

Daniel Kraschewski
Big Techday 11
2018-05-18
Motivation

personal data is omnipresent

- internet browsing history
- cell phone movements
- smart metering, smart homes, IoT
- social media, cloud
- ...

Photo: “Base.” by Instant Vantage (CC BY-SA 2.0), clipped to fit page layout
Use Cases

- customer analytics
- city planning
- medical surveys
- social surveys

Photos by @nordwood, @hellocolor, @rawpixel, @jaseess on Unsplash
GDPR (from 2018-05-25)
- strong notion of consent
  - informed
  - freely given
  - specific
  - unambiguous
  - clear affirmative act
- high fines (up to 4% of annual turnover) for data privacy violations

Privacy-Preserving Data Analysis
- derive large-scale statistical insights
- still preserve individual’s privacy
- security by design
- provable security/privacy
Exemplary Mechanism Stack

**privacy-preserving results**
- aggregated statistics must not reveal personal information
- proper formal notion of anonymity
- Differential Privacy, Laplace Mechanism

**privacy-preserving computation**
- personal data must be unaccessible even for system owner
- Cryptography
- Secret Sharing, Secure Multi-Party Computation

**privacy-preserving environment**
- no unauthorized third party may access any data
- IT-Security
- Access Control, encrypted & authenticated channels
Example: Privacy-Preserving Averaging

\[
\begin{align*}
27 + 2 &= 29 \\
-15 + 16 &= 1
\end{align*}
\]

\[
\frac{29 + 1}{2} = 15
\]
Secure Multi-Party Computation (MPC)

**general setting**
- mutually mistrusting parties $P_1, P_2, \ldots$
- secret inputs $x_1, x_2, \ldots$
- want to compute some agreed on function value $f(x_1, x_2, \ldots)$
- nothing but $f(x_1, x_2, \ldots)$ should be revealed about $x_1, x_2, \ldots$

**a universal solution**
1. write $f$ as arithmetic circuit
2. transform each $x_i$ into unintelligible Secret Sharing
3. evaluate $f$ gate-by-gate, preserving Secret Sharing
4. recombine result

Photo by Sahmeditor on Wikimedia Commons (CC BY-SA 2.0), clipped to fit page layout
### Shared Addition

<table>
<thead>
<tr>
<th>$x_A, y_A$</th>
<th>$x_B, y_B$</th>
<th>$z_A$</th>
<th>$z_B$</th>
<th>s.t. $z_A + z_B = x + y$</th>
</tr>
</thead>
</table>

**easy:** $z_A = x_A + y_A$ and $z_B = x_B + y_B$

### Shared Multiplication

<table>
<thead>
<tr>
<th>$x_A, y_A$</th>
<th>$x_B, y_B$</th>
<th>$z_A$</th>
<th>$z_B$</th>
<th>s.t. $z_A + z_B = x \times y$</th>
</tr>
</thead>
</table>

**problematic:** $z = x_A y_A + x_A y_B + x_B y_A + x_B y_B$

### missing building block

<table>
<thead>
<tr>
<th>$v_A$</th>
<th>$v_B$</th>
<th>$w_A$</th>
<th>$w_B$</th>
<th>s.t. $w_A + w_B = v_A v_B$</th>
</tr>
</thead>
</table>
Building Block for Shared Multiplication

**Invariants**

\[ r_A r_B = s_A + s_B \]
\[ v_A r_B = s_A + w_B \]
\[ v_A v_B = w_A + w_B \]

Random \( r_A, r_B, s_A \)

\[ s_B := r_A r_B - s_A \]

\[ v_A := s_A + v_A(v_B - r_B) \]

\[ w_A := s_A + v_A(v_B - r_B) \]

\[ v_B := s_B + r_B(v_A - r_A) \]

\[ w_B := s_B + r_B(v_A - r_A) \]
**k-Anonymity**

- published data must coincide with at least $k$ individuals

**De-anonymization attack on correlated data**

- published data: number of people in mobile cell at time $t_i$

\[
\begin{array}{cccc}
963 & 733 & 990 & 830 \\
764 & 857 & 814 & 991 \\
962 & 733 & 857 & 814 \\
991 & 842 & 814 & 991 \\
62 & 831 & 843 & 815 \\
43 & 857 & 962 & 990 \\
830 & 857 & 962 & 990 \\
857 & 814 & 991 & 842 \\
990 & 842 & 814 & 991 \\
\end{array}
\]

$t_1 = 02:30 \quad t_2 = 03:00 \quad t_3 = 03:30 \quad t_4 = 04:00$

- trajectory recovery $\supseteq$ optimization problem
  - higher “costs” for sudden/far movements
  - higher “costs” for irregular movements and/or movements at night
- 50% – 91% accuracy, depending on space-time resolution

---

1 Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, Depeng Jin: Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data, 26th International Conference on World Wide Web (WWW 2017)
### Secure Anonymization

#### $\varepsilon$-Differential Privacy
- statistical similarity: $\kappa(\text{real data}) \approx \kappa(\text{real data} \setminus \text{me})$ up to factor $e^\varepsilon$

#### Laplace Mechanism
1. calculate histogram
2. add Laplace noise
3. output noisy group sizes

#### Laplace Distributions
![Laplace Distributions](image)

#### Example histogram ($\frac{1}{10}$-Differential Privacy)

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Diagnosis</th>
<th>count</th>
<th>noise</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 35</td>
<td>f</td>
<td>infection</td>
<td>48</td>
<td>9</td>
<td>57</td>
</tr>
<tr>
<td>&lt; 35</td>
<td>f</td>
<td>NCD</td>
<td>61</td>
<td>-1</td>
<td>60</td>
</tr>
<tr>
<td>&lt; 35</td>
<td>m</td>
<td>infection</td>
<td>75</td>
<td>-5</td>
<td>70</td>
</tr>
<tr>
<td>&lt; 35</td>
<td>m</td>
<td>NCD</td>
<td>44</td>
<td>-7</td>
<td>37</td>
</tr>
<tr>
<td>≥ 35</td>
<td>f</td>
<td>infection</td>
<td>165</td>
<td>6</td>
<td>171</td>
</tr>
<tr>
<td>≥ 35</td>
<td>f</td>
<td>NCD</td>
<td>127</td>
<td>-4</td>
<td>123</td>
</tr>
<tr>
<td>≥ 35</td>
<td>m</td>
<td>infection</td>
<td>228</td>
<td>2</td>
<td>230</td>
</tr>
<tr>
<td>≥ 35</td>
<td>m</td>
<td>NCD</td>
<td>168</td>
<td>-2</td>
<td>166</td>
</tr>
</tbody>
</table>
### Design Principles

<table>
<thead>
<tr>
<th>Strictness</th>
</tr>
</thead>
<tbody>
<tr>
<td>- fail-safe defaults</td>
</tr>
<tr>
<td>- need-to-know principle</td>
</tr>
<tr>
<td>- principle of least privilege</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>- separation of duties</td>
</tr>
<tr>
<td>- multi-factor/layered security</td>
</tr>
<tr>
<td>- forward secrecy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>- complete security-model</td>
</tr>
<tr>
<td>- defense in depth</td>
</tr>
<tr>
<td>- homogeneity/uniformity</td>
</tr>
</tbody>
</table>
Conclusion & Outlook

Summary

- large-scale statistics can be calculated in a privacy-preserving way
- security/privacy by design, not just by contract
- security/privacy mathematically defined and provable
- though, inefficient universal solutions

Improvements

- less generic, optimized MPC constructions
- less MPC, more IT-Security (e.g., “self-sealing” hardware)
- tailored Differential Privacy mechanisms
- ...
Thank you for your attention!