Privacy-Preserving Techniques for Scanner and Mobile Phone Data Analysis

Liina Kamm, PhD
@liinakamm
Information Security Research Institute
About Cybernetica

- **Estonian ICT company**, founded in 1997
  - Successor of the Institute of Cybernetics of Estonian Academy of Sciences
- We develop and sell **mission-critical** e-government, information security, radio communications and surveillance **software products and systems**
- Our goal is to inspire **new areas of advancement through interactions between research and development**
- As of 2017 **we are 140 people** and **10% with PhD degrees**
Privacy Techniques Useful in Statistics

- **Computations on Encrypted Data** uses cryptography to make re-identification of data subjects infeasible and reduces the risks of insider attacks without reducing the accuracy of results.
  - **Example technologies**: Secure Multi-party Computation, Homomorphic Encryption, Trusted Execution Environments

- **Anonymisation** adds noise makes re-identification of data subjects harder, but can also reduce the accuracy of results.
  - **Example technologies**: Differential Privacy, k-anonymisation
Where to Use Privacy Technologies in Statistics

Input parties

Organisations

People

Direct data collection (surveys, scraping etc) using Privacy Enhancing Technologies

Data collection via intermediaries (telco data, payment data, observation data) using Privacy Enhancing Technologies

Computing parties

National Statistics Office

Other Organisations

Secure, aggregation statistical analysis and result disclosure using Privacy Enhancing Technologies

Result parties

Community
Secure Multi-party Computation

- Data owners encrypt data on-site and upload it to the server(s).
- Data analysts build and run queries without accessing the data.
- The secure MPC platform processes the queries without removing the protection.
- Authorised users receive query results in an encrypted form which they can then decrypt.

Confidential data

Analyse without ever seeing the data

Privacy controls for third parties

New knowledge
Additive Secret Sharing

Input parties

\( IP_1 \)
\( \quad \cdots \quad \)
\( IP_k \)

Computing parties

\( CP_1 \)
\( CP_2 \)
\( CP_3 \)

Result parties

\( RP_1 \)
\( \quad \cdots \quad \)

Step 1: secret sharing of inputs

Step 2: secure multiparty computation

Step 3: reconstruction of results

reconstruction of results
Concept of the Sharemind System

BUILD DATA-DRIVEN SERVICES WITH END-TO-END ENCRYPTION

DATA OWNERS
Encrypt data on site and upload to Sharemind

SHAREMIND
Runs queries without removing the protection

AUTHORISED USERS
Receive results in an encrypted form

SHAREMIND
Can give proofs of its activities to third parties
# The Sharemind Model Has Two Implementations

<table>
<thead>
<tr>
<th></th>
<th>Sharemind MPC</th>
<th>Sharemind HI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>Secure Multi-party Computation</td>
<td>Hardware Isolation using Trusted Execution Environments</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>Low to medium performance overhead</td>
<td>Minimal performance overhead</td>
</tr>
<tr>
<td><strong>Deployment</strong></td>
<td>Multi-party application server (three servers needed)</td>
<td>Single-node application server (one server with modern CPU)</td>
</tr>
<tr>
<td><strong>Usage Model</strong></td>
<td>Analytical tools and SDK available</td>
<td>Tailor-made applications only</td>
</tr>
<tr>
<td><strong>Requirements</strong></td>
<td>Deployable in any data centers or private/public clouds</td>
<td>Requires modern servers to run (available on some clouds)</td>
</tr>
</tbody>
</table>
The Rmind Tool

```r
'citation()' on how to cite R or R packages in public
Type 'demo()' for some demos, 'help()' for on-line help,
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> subject <- read.csv("subject1000.csv", header = TRUE)
> salary <- read.csv("avg-salaries.csv", header = TRUE)
> edu <- merge(subject, salary)
> age <- edu$age
> sal <- edu$avgSalary100
> plot(age, sal)
```
The Rmind Tool
Data Analysis Workflow

1. **Tax data**
   - Extract data
   - Employment tax payments
   - Aggregate

2. **Employment record of a person**
   - Merge by person's ID
   - Compute additional attributes
   - Complete record of a person
   - Analysis table
   - Analysis results

3. **Higher study events**
   - Extract data
   - Aggregate
   - University career of a person

4. **Education data**

Statistical analyst
Recipe for Converting a Workflow

- Adapt algorithm to support secure computation
  - Optimise for the chosen secure computation technique
  - To avoid timing attacks
  - Reusable libraries exist

- Create data import tools
- Use existing or create custom query tools
  - Data analysts cannot view individual values
Data Analysis Workflow Using Secure MPC

- Tax data
- Employment tax payments
- Higher study events
- Education data
- Aggregate
- Employment record of a person
- Compute additional attributes
- Complete record of a person
- Analysis table
- Analysis results
- Recover results from shares

13 21.10.2018
Example: Linking Tax and Education Registries

![Bar chart showing the number of enrolled students and students who quit studies before 2012 from 2006 to 2012.](chart.png)
Regulation Prevented a Data-Driven Answer

Tax records

Has the student worked?
In which period?
In an IT company?

How is working related to not graduating on time?

Barriers
Data Protection
Tax Secrecy

Education records

When did student enrol?
When did he/she graduate?
In an IT curriculum?
Source data:
- 10 million tax records,
- 600,000 education records.

Sharemind hosted by government agencies and Cybernetica.

Data owners used the Sharemind encryption tools to upload data.

Data never existed outside the source in a decrypted state.
Privacy-Preserving Study of Students’ Working Habits

- Data scientists used the Rmind tool to run the analysis.
- Sharemind prevented queries outside the study plan.
- Reports were given to industry, universities and the government.
- **Result**: no clear relation between working during studies and not graduating.
Regulatory precedents in Europe

- The Estonian Data Protection Agency stated that the combination of technology and processes ensured that **private data was not processed** and the requirements of the **Data Protection Act need not apply**.
  - Assumption: no identifiable records are published.
- The Internal Supervision of the Tax and Customs Board agreed to provide unmodified tax records after a code and process review.
- A German legal research team extended the precedent to work under the GDPR.
- We are now preparing for validation with other DPAs within new projects.
Example: Privacy-Preserving Mobile Data Analysis

LIVE DATA SOURCES
- Roaming phones in Estonia
- New data added every day
- Encrypted and uploaded with Sharemind HI tools

SECURE AGGREGATION
- Sharemind HI provides secure storage and processing
- 700 MB of data aggregated in a Trusted Execution Environment in under ten minutes

INTERACTIVE VISUALIZATION
- Explore inbound tourism data
- Statistical methodology provided by Positium
Privacy-Preserving Mobile Data Analysis Demo
Example: Privacy-Preserving Scanner Data Analysis

**SYNTHETIC DATA**
- Artificial price data provided by Stats NZ
- Encrypted and uploaded with Sharemind MPC tools

**SECURE STATISTICS**
- Sharemind MPC provides secure storage and processing
- Complete workflow performed in the privacy-preserving environment

**CUSTOM INTERFACE**
- Part of the UNGlobalPlatform
- Returns the price indices
- Methodology provided by Stats NZ
Privacy-Preserving Scanner Data Cloud Deployment
# Privacy-Preserving Scanner Data Analysis

<table>
<thead>
<tr>
<th>Data Profile</th>
<th>No. of Rows</th>
<th>No. of Sharemind Instances</th>
<th>Time Spent</th>
<th>Cloud Cost (Bandwidth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>230 products for 26 months</td>
<td>3843</td>
<td>2 (11 parallel regressions)</td>
<td>8m 57s</td>
<td>$31.50</td>
</tr>
<tr>
<td>230 products for 26 months</td>
<td>3843</td>
<td>4 (~5 parallel regressions)</td>
<td>7m32s</td>
<td>$31.50</td>
</tr>
<tr>
<td>460 products for 26 months</td>
<td>7654</td>
<td>4</td>
<td>22m 24s</td>
<td>$165</td>
</tr>
</tbody>
</table>
Find more information at sharemind.cyber.ee