Sharemind - practical privacy-preserving analytics

Sander Siim
Cybernetica AS
sander.siim@cyber.ee
About Sharemind

Sharemind uses MPC to analyse data that was not accessible before.

Sharemind resolves trust issues by removing centralised control and unwanted data access points.
Application Server paradigm

- sharemind interfaces
- application servers
- database backends

Host 1
- Mobile apps
- SQL queries
- Java/JavaScript/C/C++/Haskell

Host 2
- Web apps
- Rmind statistics package

Host n
- Desktop apps
- Mobile apps
- Web apps
- SQL queries
- Java/JavaScript/C/C++/Haskell

Mobile apps
Web apps
Desktop apps
Encrypted computing

Data owners
- People
- Industry

Acquisition channels
- Mobile applications
- Online services
- Existing databases

<table>
<thead>
<tr>
<th>ID</th>
<th>sex</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>M</td>
<td>23</td>
</tr>
<tr>
<td>106</td>
<td>F</td>
<td>38</td>
</tr>
<tr>
<td>118</td>
<td>M</td>
<td>19</td>
</tr>
<tr>
<td>143</td>
<td>M</td>
<td>32</td>
</tr>
</tbody>
</table>

Access channels

Data users
- Decisionmakers
- Researchers
- General population

Data are collected and stored in an encrypted form

Data are not decrypted for processing

Only the results of allowed queries can be published
Model of secure computing

Input parties

\[ x_1 \rightarrow x_{k_1} \rightarrow x_{k_i} \rightarrow x_{k_l} \rightarrow \ldots \]

\[ \ldots \]

\[ IP_1 \rightarrow IP_k \]

Computing parties

\[ CP_1 \rightarrow y_{1j} \rightarrow y_{i} \rightarrow y \]

\[ \ldots \]

Step 1: upload and storage of inputs

Step 2: secure computation

Step 3: publishing of results

Result parties

\[ RP_1 \rightarrow RP_m \]
<table>
<thead>
<tr>
<th>Name</th>
<th>num of input parties</th>
<th>num of computing parties</th>
<th>num of result parties</th>
<th>Technology</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>shared3p</td>
<td>any</td>
<td>3</td>
<td>any</td>
<td>LSS MPC, (Yao)</td>
<td>In commercial use</td>
</tr>
<tr>
<td>shared2p</td>
<td>any</td>
<td>2</td>
<td>any</td>
<td>LSS MPC, (Yao)</td>
<td>Under development</td>
</tr>
<tr>
<td>sharednp</td>
<td>any</td>
<td>3 or more</td>
<td>any</td>
<td>LSS MPC</td>
<td>Under development</td>
</tr>
</tbody>
</table>
The shared3p core

- **Storage**: additive and bitwise secret sharing
- **Computing**: three-party MPC based on LSS
- **Data types**: 13 types (boolean, signed and unsigned integers, fixed point, floating point)
- **Operations**: 650 machine-optimized protocols

- Protocols developed by Cybernetica over the last 10 years, heavily tuned and optimized
- Powers all our commercial applications and most R&D prototypes
Protocol DSL and compiler

- Our newest and fastest protocols are implemented with a special-purpose compiler
- DSL(high-level description of $\pi$) = machine-code that runs $\pi$
- Easy to test and implement new protocols
- Optimizes protocol structure and communication — up to 40x speed-up
- Helps maintain our growing library of protocols
- Can use also in 2-party/n-party case

Cores in development

shared2p
- **Storage**: additive and bitwise secret sharing
- **Computing**: two-party secure MPC
- Combination of shared2p techniques with Beaver triples

sharednp
- **Storage**: Shamir’s secret sharing
- **Computing**: $n$-party secure MPC
- Classic Shamir protocols + custom designs
Controlling computations

Sharemind only runs computations deployed by all computing parties.

Allowed outputs are defined by the queries.

If a computing party does not agree to run an application, it cannot be run.

Data owners:
- Database
- Policy

Data users:
- Published results
The SecreC language

// Import module for the secure protocol suite
import shared3p;

// Data in private domain is processed via MPC
domain private shared3p;

void main () {
    // Perform secure computations
    private int a = 2, b = 3;
    private int c = a * b;
    // Must explicitly declare publishing c
    print (declassify (c));
}
Polymorphic functions

template <domain D>
D int scalarProd(D int[][1] x, D int[][1] y) {
    return sum(x*y);
}

domain private3 shared3p;
domain private2 shared2p;

void main () {
    private3 int[][1] x3(100) = 2, y3(100) = 3;
    private2 int[][1] x2(100) = 2, y2(100) = 3;
    print (declassify (scalarProd(x3, y3)));
    print (declassify (scalarProd(x2, y2)));
}
SecreC standard library

- A library of privacy-preserving algorithms.
- Array and matrix operations, oblivious access, statistical testing, sorting, linking, regression modelling, aggregation, etc.
- 15 000 lines of reusable SecreC code
Demo!
Prototype an MPC application in minutes
Sharemind SDK

• Free open-source prototyping tools available: http://sharemind-sdk.github.io/

• Includes SecreC and the standard library
• An emulated Sharemind run-time that estimates online performance
• Excellent for quick prototyping
Case study: Government data analytics
IT training has a failure rate

By 2012, a total of 43% of students enrolled in the four largest IT higher learning institutions in Estonia during 2006-2012 had quit their studies. Source: Estonian Ministry of Education and Research, CentAR.
Barriers for assessing the situation

**Tax records**

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

**Barriers**

- Data Protection
- Tax Secrecy

**Education records**

- When did student enrol?
- When did he/she graduate?
- In an IT curriculum?

January 2014: Estonian Data Protection Agency declared that Sharemind technology and processes protect data so well that the Personal Data Protection Act doesn’t apply.

January 2015: after a code audit, the internal oversight at the Tax Board agreed to upload actual income tax records into the Sharemind-based analysis system.

February 2015: the Tax Board, Ministry of Education, Information Systems Authority, Ministry of Finance IT Center and Cybernetica signed the world’s first secure multi-party data analysis agreement.
Step 1: Import data

- Data owners uploaded data with the Sharemind importer to a shared3p core.
- Each value was encrypted at the source, private data never left the data owner.
- Over 600,000 study records (100 MB) used.
- Over 10 million tax records (1 GB) used.
- Largest MPC application on real-world data.
Step 2: Run the analysis

- Statisticians used Rmind to post queries.
- Sharemind ensured that only queries in the study plan were actually executed.
- Additional microdata protection controls were enforced.
Operations performed

- Tax and Customs Board
- Employment tax payments
- Secret share and upload
- Higher study events
- Ministry of Education and Science

Data stored with secret sharing and processed with secure multi-party computation

- Aggregate by year
  - Monthly income
  - Aggregate by month
  - Employment tax payments
  - Aggregate by person
  - Higher study events

- Average yearly income
  - Expand by years and aggregate by person
  - Employment record of a person
  - Merge by person's ID
  - University career of a person

- Complete record of a person
  - Analysis results
    - Analysis table
    - Compute additional attributes and align tax payments

- Recover results from shares

- Statistical analyst
Sharemind Analytics Engine

LXTerminal

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

> subject <- read.csv("subject1000.csv")
> salary <- read.csv("avg-salaries.csv")
> edu <- merge(subject, salary)
> age <- edu$age
> sal <- edu$avgSalary100
> plot(age, sal)

LXTerminal

[sharemind@sm-build-vm rmind]$ ./rmind
Rmind
Copyright (C) Cybernetica AS
Type ‘q()’ to quit
Connecting to Sharemind...
Connected
> salary <- load("DS1", "salaries")
> subject <- load("DS1", "subjects")
> edu <- merge(subject, salary)
> age <- edu$age
> sal <- edu$avgSalary100
> heatmap (age, sal)
Sharemind Analytics Engine

R Graphics: Device 2 (ACTIVE)

Gnuplot window 0

avgSalary100

age

-1.76700, -2.41657e+07
IT is harder to graduate

Joonis 1. Nominaalajaga lõpetajate osakaal immatrikuleerimisaastate lõikes, IKT- ja mitte-IKT õppekavade, bakalaureuseõpe.
All students are working

Joonis 4. Nominaalaja jooksul töötanud tudengite osakaal kõigist tudengitest aastati, IKT- ja mitte-IKT õppekavade, bakalaureuseõpe
Practice makes perfect

• After successfully ending the project, we went back to the lab to see if we can do better

• The new protocol DSL gave a “conservative” 20% performance improvement

• It turned out we could significantly optimize the aggregation algorithms through better parallelization
Major speed-ups

Protocol DSL: 345h
Parallelized aggregation: 5h

6 ms latency for one server, 1Gbps bandwidth

More gains from high-level algorithm optimizations than low-level protocols
Case study: A privacy-preserving survey system
Privacy-preserving surveys

• Traditional survey systems do not hide individual answers from organizer/server
• Use MPC to remove centralised trusted service provider
• We built a secure survey system in the PRACTICE project together with Alexandra Institute and Partisia
• Has both Sharemind and Fresco/SPDZ back-ends
Demo! A happy employee answering a survey anonymously
Case study: Tax fraud detection
Estimate of unpaid VAT

<table>
<thead>
<tr>
<th>Tax Type</th>
<th>MEUR 2011</th>
<th>MEUR 2012</th>
<th>MEUR 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAT</td>
<td>208.5 MEUR</td>
<td>216.2 MEUR</td>
<td>300.6 MEUR</td>
</tr>
<tr>
<td>Social tax</td>
<td>67.5 MEUR</td>
<td>95.8 MEUR</td>
<td>101.3 MEUR</td>
</tr>
<tr>
<td>Income tax</td>
<td>101.3 MEUR</td>
<td>69.4 MEUR</td>
<td>71.7 MEUR</td>
</tr>
<tr>
<td>Alcohol excise</td>
<td>10.5 MEUR</td>
<td>9.4 MEUR</td>
<td>9.3 MEUR</td>
</tr>
<tr>
<td>Tobacco excise</td>
<td>31.3 MEUR</td>
<td>35.3 MEUR</td>
<td>25.0 MEUR</td>
</tr>
<tr>
<td>Fuel excise</td>
<td>12.3 MEUR</td>
<td>11.2 MEUR</td>
<td>25.7 MEUR</td>
</tr>
<tr>
<td>Packaging excise</td>
<td>13.0 MEUR</td>
<td>10.7 MEUR</td>
<td>14.4 MEUR</td>
</tr>
</tbody>
</table>
Attempted fix to the gap

- In 2013, the Estonian parliament ratified the Value-Added Tax Act and the Accounting Act Amendment Act that would force enterprises to report transactions to the Tax and Customs Board (MTA).
- MTA would then match outgoing invoices to the incoming invoices reported by others and find companies trying to get refunds for fraudulently declared input VAT.
The story of the 1000 € law

Ilves Blocks Amendment for Sweeping Disclosures in Tax Filing

12/19/2013 9:12 AM
Category: Politics

President Toomas Hendrik Ilves has blocked an amendment to the VAT law - which would require all transactions greater than 1,000 euros to be declared - on the grounds that it is unconstitutional.
Implementation using MPC

- The Tax Board was worried enough after the veto that they were willing to hear us out.

- It also helped that Cybernetica was the company who won the tender to build the actual system.

- We agreed with the Tax Board that Cybernetica will build a research prototype that implements four risk analyses and will test its performance and that they will look at our results.

- We borrowed a systems analyst and an architect from our tax team to build the prototype.
Secure implementation

Benefits

- Encryption is applied on the data directly at the source.
- The data is cryptographically protected during processing.
- No need to unconditionally trust a single organization.

Transactions

Taxpayers

sharemind
Secure implementation

Encryption is applied on the data directly at the source.
The data is cryptographically protected during processing.
No need to unconditionally trust a single organization.

Benefits

secure multi-party computation system with database

Watchdog NGO server

Taxpayer's association's server

Transactions

Taxpayers

sharemind

sharemind
Secure implementation

Benefits
- Analyze, combine and build reports without decrypting data.
- Confidentiality is guaranteed against all servers and against malicious hackers.
- Values are only decrypted when all hosts agree to do so.

Encryption is applied on the data directly at the source.
The data is cryptographically protected during processing.
No need to unconditionally trust a single organization.

secure multi-party computation system with database

Tax Office server

Risk scores

Taxpayer's association's server

Transactions

Watchdog NGO server

Analyze, combine and build reports without decrypting data.
Confidentiality is guaranteed against all servers and against malicious hackers.
Values are only decrypted when all hosts agree to do so.

Benefits
- Encryption is applied on the data directly at the source.
- The data is cryptographically protected during processing.
- No need to unconditionally trust a single organization.
Using fork-join parallelism

- **Secret sharing of transactions**
- **Distribute inputs between tasks**
- **Tasks aggregate transaction tables**
- **Finalize aggregation and calculate scores**

---

Company A

<table>
<thead>
<tr>
<th>A</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1</td>
<td>2</td>
<td>...</td>
<td>n</td>
</tr>
</tbody>
</table>

Host 1

MPC task

Host 2

MPC task 2

Host 3

MPC task n

---

Host 1

<table>
<thead>
<tr>
<th>MPC task 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
</tbody>
</table>

Host 2

<table>
<thead>
<tr>
<th>MPC task</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
</tbody>
</table>

Host 3

<table>
<thead>
<tr>
<th>MPC task</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
</tbody>
</table>

---

Risk analyst at MTA

Secure multi-party computation

- **Send risk scores to analyst**
- **End-user communication with Sharemind**

---

organizational boundary

---

Sharemind

---
Experiments on AWS cloud

Note: actual deployment should run on three different clouds. However, we had a humble research grant from AWS.
Much improved parallelism

12 computing nodes running a total of 80 Sharemind processes

eu-central  eu-west

Client instance

Sharemind
## Computing environment

<table>
<thead>
<tr>
<th>Setup</th>
<th>Client</th>
<th>Computing parties</th>
<th>Latency (round-trip)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>us-east – c3.8xlarge</td>
<td>us-east – 12x c3.8xlarge</td>
<td>&lt; 0.1ms between all nodes</td>
</tr>
<tr>
<td>2</td>
<td>eu-west – c3.8xlarge</td>
<td>eu-west – 8x c3.8xlarge, eu-central – 4x c3.8xlarge</td>
<td>&lt; 0.1ms inside eu-west (19ms (eu-west/eu-central))</td>
</tr>
<tr>
<td>3</td>
<td>us-east – c3.8xlarge</td>
<td>us-east – 4x c3.8xlarge, us-west – 4x c3.8xlarge, eu-west – 4x c3.8xlarge</td>
<td>77ms (us-east/us-west), 133ms (us-west/eu-west), 76ms (us-east/eu-west)</td>
</tr>
</tbody>
</table>
Realistic data sizes

<table>
<thead>
<tr>
<th>No. of companies</th>
<th>No. of transaction partner pairs</th>
<th>Total no. of transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 000</td>
<td>200 000</td>
<td>25 000 000</td>
</tr>
<tr>
<td>40 000</td>
<td>400 000</td>
<td>50 000 000</td>
</tr>
<tr>
<td>80 000</td>
<td>800 000</td>
<td>100 000 000</td>
</tr>
</tbody>
</table>

The source data for 100 000 000 transactions had a total size of 35 GB in XML format (about 1 GB in the secret-shared database).
Better running times

Technical issues prevented the completion of this test and budgetary constraints did not allow for a repeat.
Significantly lower price

The diagram illustrates the number of companies and their associated prices in different deployment regions. The prices are as follows:

- **2-us, 1-eu**
  - 20k: $126
  - 40k: $223
  - 80k: $150

- **2-eu**
  - 20k: $61
  - 40k: $91
  - 80k: $71

- **us**
  - 20k: $27
  - 40k: $49
  - 80k: $71

The deployment regions are color-coded as follows:
- **us** (red)
- **2-eu** (green)
- **2-us, 1-eu** (blue)
Conclusion

Our dream is to see MPC becoming an ubiquitous tool in applications where privacy is important.

We can already demonstrate solving privacy issues for real-world users and organizations on a large scale.
We build applications

Learn about Sharemind and request an academic license
http://sharemind.cyber.ee/

Open source prototyping tools (under development)
http://sharemind-sdk.github.io/

Contact us for more information and collaborations
E-mail: sharemind@cyber.ee
Twitter: @sharemind