Building privacy-preserving web applications with the Sharemind framework

Dan Bogdanov\textsuperscript{1,2} and Riivo Talviste\textsuperscript{1}

\textsuperscript{1} University of Tartu, Institute of Computer Science
\textsuperscript{2} AS Cybernetica, Institute of Information Security

1 Introduction

Estonian Association of Information Technology and Telecommunications (ITL) [http://www.itl.ee] is a non-governmental non-profit organization whose primary focus is to unite Estonian information technology and telecommunication companies and promote their co-operation. ITL has over 60 members – Estonian enterprises that engage in the field of information and communication technology (ICT).

ICT is a fast-growing industry and thus ITL members would like to be able to compare themselves with other companies in that sector more often than once a year when the Ministry of Economic Affairs and Communications releases the annual economic reports. Making business decisions based on this annual report is not sufficient for such a dynamic industry as ICT.

2 A possible solution

Now, there is an initiative that ITL should collect some basic economic indicators from its members (only IT companies) twice a year. Based on these collected indicators, ITL can release an anonymized benchmarking info for each collected column to its members. This enables businesses to compare themselves with the best in the ICT sector.

ITL collects these economic indicators from its members with the following frequency:

\begin{center}
\begin{tabular}{ll}
Indicator & Collected \\
total return & semi-annually \\
number of employees & semi-annually \\
proportion of export & semi-annually \\
labour costs & annually \\
training costs & annually \\
profit & annually \\
added value & annually \\
\end{tabular}
\end{center}

All these seven indicators will be released to the public half a year later anyway, as they are part of the annual economic report that the ministry releases. However, ITL members would like to compare themselves with other members based on more up-to-date information.
2.1 User story

All ITL members have a username and a password that they can use to log in to the member area of the ITL web page. There they can choose an open form to start filling it. A form is “open” if it can be filled at this point in time. There are two kinds of forms:

1. Annual forms containing all 7 indicators are opened for 45 days after the economic year has ended (Jan 1 – Feb 14)
2. Semi-annual forms contain only 3 indicators and are opened for 30 days after the corresponding half-year has ended (July 1 – July 30 and Jan 1 – Jan 30)

After the user has filled in the form, he or she clicks on a button to submit it. The inserted values are sent back to the ITL web server where they are stored in a database.

When the collection period has ended a report of this period is automatically generated. Each economic indicator (database column) without the identifying information (company ID) is taken separately and sorted in decreasing order. This sorted vector is then published for each indicator.

For example, let’s assume that three companies with ID numbers 1, 2 and 3 insert such records into the database:

(company_id=1, field1=111, field2=1234, field3=42314)
(company_id=2, field1=222, field2=2431, field3=12345)
(company_id=3, field1=212, field2=3132, field3=54321)

This results in these three tables being published:

<table>
<thead>
<tr>
<th>field1</th>
<th>field2</th>
<th>field3</th>
</tr>
</thead>
<tbody>
<tr>
<td>222</td>
<td>3132</td>
<td>54321</td>
</tr>
<tr>
<td>212</td>
<td>2431</td>
<td>42314</td>
</tr>
<tr>
<td>111</td>
<td>1234</td>
<td>12345</td>
</tr>
</tbody>
</table>

As we can see here, a company being first in one table might be third in another and so on.

However, we know that when collecting sensitive information (like economic results) then anonymization might not be enough. There are numerous examples [1] where individual persons or organizations have been identified based solely on the anonymized data released to the public.

Moreover, since the collected data is stored in a database, then whoever has access to it, also has access to the info provided by the companies. This holds even if the anonymization is done right at the moment when data is inserted. The system still needs to store the information about the companies who have already inserted their data and those who have not. Even if this information is kept separate from the inserted values, the system administrators still have access to the database logs and can use timestamps in those transaction logs to identify individual companies’ data.
3 A solution with stronger privacy guarantees

To overcome the risks of using data anonymization, we instead use the Sharemind framework [2] to collect and analyse the necessary information. Sharemind is a distributed virtual machine that is able to perform privacy-preserving computations on secret shared data. Sharemind virtual machine consists of three independent data miners, each of which holds a piece of the secret shared data. Using secure multiparty computation protocols these three miners can perform various operations on their data, revealing only the computation result and no intermediate individual values. The protocols of the Sharemind virtual machine are provably secure in an honest but curious model with no more than one malicious party.

Sharemind uses additive secret sharing scheme in the ring $\mathbb{Z}_{2^{32}}$. Suppose that we have a secret value $s$ that we want to share. Let us also assume that we have a function $\text{random}()$ that gives us uniformly distributed random values from the same ring $\mathbb{Z}_{2^{32}}$. In Sharemind, secret sharing works like this:

\[
\begin{align*}
    s_1 & \leftarrow \text{random}() \\
    s_2 & \leftarrow \text{random}() \\
    s_3 & \leftarrow s - s_1 - s_2
\end{align*}
\]

All calculations are done modulo $2^{32}$, so $s_3$ is also a non-negative number. Values $s_1$, $s_2$ and $s_3$ are distributed among three Sharemind miners. Independently, all those shares look like white noise.

Three Sharemind data miners are hosted by three ITL members, because:

1. They are motivated to host the miners, as this project would also be beneficial for themselves.
2. They are independent and will not collude with each other as they are also inserting their own data into the system and want to keep it private.
3. Also, ITL members act on the field of information technology, thus they have the necessary infrastructure and competence to host a server that runs a Sharemind miner.

3.1 Web-based data entry for Sharemind

The privacy guarantees of the Sharemind virtual machine hold only when the original data is distributed into shares at the source, which in the case on web-based data entry means that it should be done in the web browser in the end user’s computer. Using web frontends to insert data into the Sharemind virtual machine is described in [3,4]. The latter introduces an architecture which consists of the following components:

- An initial web server, which serves the form that should be filled in when entering the data. In this case, the form is accessible from the member area of the ITL web page www.itl.ee.
- The PrivateSurvey API written in JavaScript that is used by the web form and is responsible for secret sharing the original values, collecting randomness to perform it and sending shares directly to different data miners.
– Miners’ web frontends that receive shares from the PrivateSurvey API and store them in a buffer database.
– Three Sharemind data miners.
– A proxy application with each of the data miners that relays shares from buffer database to miner’s inner database. The proxy applications are used for increased system robustness and security.

Figure 1 shows all of these components, where they are hosted and how they interact.

As seen from the figure, the miners are hosted by Cybernetica, Microlink and Zone who are all ITL members. Both Microlink and Zone will deploy a virtual machine that acts as a Sharemind miner in their domain. Cybernetica uses its research server as the miner. All of those servers have an installation of the Sharemind miner, a web server (Apache) and an SQL database engine (MySQL).

3.2 Gathering data

Like before, the user first logs in to the ITL web page member area and chooses one of the opened forms.
The form itself consists of a simple form for inserting the values and logos of the companies that host the Sharemind data miners (see Figure 2). When the web page is loaded, the PrivateSurvey API immediately starts collecting randomness from the miners. This process is explained in detail in [4]. When the randomness is received from a miner, a text saying “Ready” appears below the corresponding miner host’s logo. When randomness is collected from all miners, the form submit button, which was initially disabled, becomes enabled and the user can submit the form. If the API fails to receive randomness from some of the miners, then the submit button will stay disabled and the user will be notified with a modal popup window.

![Image of the web form integrated into the ITL web site](image)

**Fig. 2.** A screenshot showing the web form integrated into the ITL web site together with the logos of miner hosts.

When user has finished filling in the form he or she can click the “Submit” button. Upon submitting the form, all the inserted economic indicators are secret shared and each share is sent to a different miner, along with the necessary public information like session ID, company ID, year and half-year number. To guarantee integrity and confidentiality of the data, all communications with the miners are done using HTTPS (HTTP over SSL/TLS) protocol.

The progress of sending the data and receiving an acknowledgement is shown below each miner’s logo, just like in the case of collecting randomness. If the API fails to receive some acknowledgements, the user is again notified with a popup dialog and has to try again later. When all the data is successfully sent, the user sees a notification popup with a button that takes the user back to the previous
page (list of opened forms). Throughout this process, the user-filled form stays on the screen and at the end, the user has to explicitly click on a button to return to a previous page. This gives the user more confidence, that the form he or she filled was never sent to the ITL web server (i.e. the inserted data did not go to one server).

[Technical remark] The Sharemind virtual machine supports only a single data type – unsigned 32-bit integer (uint32). Mostly, this is not a problem as the collectable data is numeric, but there are a couple of things we have to keep in mind:

1. The percentile values that are usually given with one or two decimal places have to be rounded to the nearest integer value. As we perform statistical analyses on the given data, the rounding does not influence the outcome very much.
2. Negative values must also be stored as uint32. To accomplish this, we take all negative values modulo $2^{32}$, that is, for each $s < 0$, we use $s' = s + 2^{32}$ instead. For all non-negative values $s' = s$. Of course, we have to take this conversion into account later when performing computations on the values. For example, computing an average of the values can be done like this:
   (a) Calculate the sum of elements: $sum = \sum_{i=1}^{n} s'_i$, where $n$ is the number of elements.
   (b) If $sum > 2^{31} - 1$ then $sum = sum - 2^{32}$.
   (c) Average is $\frac{sum}{n}$.

3.3 Generating reports

After the forms are closed (30 days after the half-year has ended or 45 days after the year has ended) a report of the previous period is automatically generated. After the collecting period has ended, the proxy application is executed synchronously in all three miners. This application orders all the shares of the just ended period in its local buffer database by session ID and saves the intersection of the records into the local Sharemind miner’s database. This initial buffering is needed as the Sharemind miners assume that the data records in all three miners’ databases are in the same sequence. Due to the asynchronous nature of the internet it is hard to achieve without a buffer database and a proxy.

Next, one miner initiates the report computation process. All three miners have an identical copy of each compiled SecreC [5] script that it has to execute. At the moment we calculate only average and decreasingly sorted values for each economical indicator (database column), but any other aggregated values are possible.

The results of all the computations are stored in a public database that can be used to make various kinds of reports. For example, using Google Visualization API, we can make a simple line chart showing averages for each of the economical indicators in the course of many years. These reports are made available to the ITL members via ITL web page member area.
Using the aforementioned proxy application also increases security. A malicious miner could initiate the report calculation in very short intervals and save the outcome on each step. If only one company inserts its economic indicators between two subsequent steps then that malicious miner could use the difference in the computation outcomes and derive the inserted values from it. However, since the proxy applications have to be run synchronously on all three miners, then all of them have to have the same report computation interval.

Moreover, all the communication between Sharemind miners, computation controller applications and proxy applications uses public key cryptography to provide integrity and confidentiality of the messages.

4 Conclusion

The project described in this paper is an ongoing project led by AS Cybernetica and Software Technology and Applications Competence Centre (STACC). The prototype application is finished and at the moment we are integrating the solution with the existing content management system of the ITL web page. This allows users to use single sign-on and thus make it more usable and also more trusted.

The solution will be deployed in the beginning of January 2011 when Estonian ICT companies can start to fill two forms – their economical indicators for the second half of 2010 and for the whole year of 2010. First reports will be generated in the middle of February when the first data collection period ends.

This is a pilot project using secure multiparty computation in real life situation in Estonia. Hopefully, we will get some better understanding about the suitability of the Sharemind framework to solve this kind of problems.

References