Deploying secure multi-party computation for financial data analysis (short paper)*

Dan Bogdanov[1,2], Riivo Talviste[1,2,3], and Jan Willemson[1,3]

[1] Cybernetica, Akadeemia tee 21, 12618 Tallinn, Estonia
{dan,riivo,janwil}@cyber.ee
[2] University of Tartu, Institute of Computer Science, Liivi 2, 50409 Tartu, Estonia
[3] STACC, Akadeemia tee 15A, Tallinn 12618, Estonia

Abstract. We show how to collect and analyze financial data for a consortium of ICT companies using secret sharing and secure multi-party computation (MPC). This is the first time where the actual MPC computation on real data was done over the internet with computing nodes spread geographically apart. We describe the technical solution and present user feedback revealing that MPC techniques give sufficient assurance for data donors to submit their sensitive information.

Keywords: financial data analysis, secure multi-party computation

1 Introduction

Financial metrics are collected from companies to analyze the economic situation of an industrial sector. Since this data is largely confidential, the process can not be carried out just by sending the data from one company to another. We claim that the use of secure multi-party computation (MPC) distributes the role of a trusted third party among many parties so that none of them has to be trusted unconditionally. The greatest added value for the companies is that no single data value can be seen by a single outside party after it leaves the user's computer.

In this paper we describe a secure system for collecting and analyzing financial data in an industrial consortium. The system was deployed for ITL—an Estonian non-governmental non-profit organization with the primary goal of promoting co-operation between companies engaging in the field of information and communication technology. The data collection and analysis system was built using the Sharemind secure computation framework [7].

Some of the details of this work have been omitted because of space limitations. The extended version of this paper that covers all these details can be found in the IACR ePrint Archive [8].

MPC has been studied for almost thirty years and recently, many MPC projects have started reaching practical results [9,10,1,7,13,15,4,11]. However, to

* This research was supported by the ERDF through EXCS and STACC; the ESF Doctoral Studies and Internationalisation Programme DoRa; the target funded theme SF0012708s06 and the Estonian Science Foundation, grant No. 8124.

the best of our knowledge, this is the first time where the actual secure multi-party function evaluation was done over a wide area network (the internet) using real data.

In 2004, J. Feigenbaum et al. implemented a privacy-preserving version of the Taulbee Survey[4] using MPC [11]. Their implementation used secret sharing at the data source and two parties evaluating a Yao circuit over a wide area network. However, their implementation was never used with real data [12].

MPC was first used in a large-scale practical application in Denmark in 2008 when a secure double auction system allowed Danish sugar beet farmers to trade contracts for their production on a nation-wide market [9]. The Danish system used three secure computation servers. In the farmers' computers, each share of private data was encrypted with a public key of one of the computation servers. The encrypted shares were sent to a central database for storing. In the data analysis phase, each computation node downloaded their corresponding shares from the central database and decrypted them. The actual MPC process was performed in a local area network set up between the three computation nodes.

2 Sharemind

Sharemind [7] is a distributed virtual machine that uses secure multi-party computation to securely process data. Sharemind is based on the secret sharing primitive introduced by Blakley [6] and Shamir [16]. In secret sharing, a secret value s is split into a number of shares s_1, s_2, ..., s_n that are distributed among the parties. Depending on the type of scheme used, the original value can be reconstructed only if the shares belonging to some predefined sets of parties are known. Sharemind uses the additive secret sharing scheme in the ring $\mathbb{Z}_{2^{32}}$ as this allows it to support the efficient 32-bit integer data type.

Sharemind uses three *data miners* to hold the shares of secret values. Secret sharing of private data is performed at the source and each share is sent to a different miner over a secure channel. The miners are connected by secure channels and run MPC protocols to evaluate secure operations on the data. The Sharemind protocols are secure in the *honest-but-curious* model with no more than one corrupted party. The honest-but-curious model means that security is preserved when a malicious miner attempts to use the values it sees to deduce the secret input values of all the parties without deviating from the protocol.

To set up a Sharemind application we first have to find three independent parties who will host the miner servers. In a distributed data collection and analysis scenario, it is possible to select the parties from the organizations involved in the process. Second, we have to implement privacy-preserving data analysis algorithms using a special high-level programming language called SecreC [14]. In the third step, we use the Sharemind controller library to build end-user applications that are used for collecting data, starting the analysis process and generating the reports.

[4] Computing Research Association, Taulbee Survey, http://www.cra.org/statistics

3 The application scenario

In Estonia, the Ministry of Economic Affairs and Communications publishes an economic report every year, combined from all of the annual reports of Estonian companies. However, while this report is accurate and gives a detailed overview of the country's economic situation, it is only compiled once a year and by the time it is published, the data is already more than half a year old.

Since ICT is a rapidly evolving economic sector, ITL members would like to get more up-to-date information about the sector to make better business decisions. ITL decided to collect basic financial data from its members twice a year and publish them as anonymized benchmarking results for its members. As the collected data does not have to be audited, the data collection periods can be shorter, which means that the published benchmarking results will be up-to-date.

During our interviews, ITL representatives described a solution they had imagined. They would collect the following financial indicators:

- total return, number of employees, percentage of export, added value — semi-annually;
- all of the above plus labour costs, training costs and profit — annually.

After each collection period, the values would be anonymized (i.e. the company identifiers removed) and each indicator would be sorted independently to reduce the risk of identifying some companies by just looking at a set of financial indicators. For example, combining total return, number of employees and profit, it could be easy to identify some ICT companies. However, when sorting by each indicator independently, a company that is the first when sorted by one indicator might not be the first when sorted by another indicator.

Sorting the collected data by each indicator separately gives us a slightly stronger privacy guarantee than just stripping away the identifying information. However, all of the collected data is still accessible by the ITL board, which consists of the leaders of competing ICT companies. This means that ITL members must trust the ITL board not to misuse or leak the collected information. Consequently, ITL member companies might be reluctant to participate and give away their sensitive economic information, as it can be seen by their competitors. ITL members are required to trust the board with their data and this is quite a significant assumption.

3.1 Reducing trust requirements

We proposed to use the Sharemind framework to collect and analyze the financial data to address the shortcomings of the initial solution. By using secret sharing at the source and distributing the sensitive values among the three Sharemind data miners we make sure that no single party has access to the original values. Hence, we also have a lower risk of insider attacks and unintentional disclosures (e.g. data leak via backup). Most importantly, the use of MPC reduces the trust that ITL members need to have in any single party.

After data has been collected from all of the members, the data miners engage in secure MPC protocols and sort all the collected economic indicators. The sorted indicators are then published as a spreadsheet and made accessible to the board members of ITL. The board can then calculate aggregate values and/or charts and give this edited report to the members. The data flow and visibility to different parties for this solution is shown on Figure 1.

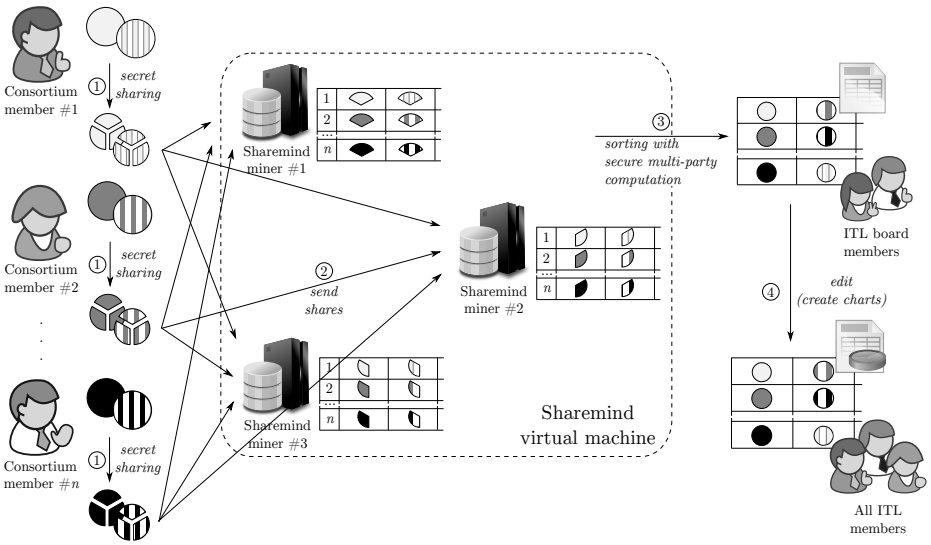


Fig. 1: Data flow and visibility in the improved solution using the SHAREMIND framework.

4 The ITL secure data aggregation system

4.1 Deployment

In the real-life deployment, the SHAREMIND miners are hosted by three Estonian companies and ITL members—Cybernetica, Microlink and Zone Media. Choosing the miner hosts among the consortium members fulfills the following requirements set for the data miners: *a*) they are motivated to host the miners, as this project would also be beneficial for themselves; *b*) 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; *c*) ITL members act in the field of information technology, thus they have the necessary infrastructure and competence to host a SHAREMIND server.

As the miner hosts provided their servers with no cost, they wished to reduce the effort needed to maintain the servers. Thus, all of the three miner hosts were

set up by a single administrator who also regularly executes the computations. Ideally, each host should be maintained by its respective owner and this should be a rule in all future deployments of the technology. We consider it an important challenge to reduce the administrative attention required for managing a Sharemind miner to a minimum as this makes miner host selection easier and makes the technology easier to deploy in practice.

4.2 Securing web-based data collection

ITL requested that the online data submission form should be integrated into their web-based member area. This way, the representatives of ITL members can access everything related to ITL from one familiar environment. It also allows us to reuse the authentication mechanisms of the ITL web page.

We have developed a JavaScript library that can be used to turn a basic HTML form into an input source for secure MPC applications with minimal effort. This library [17] performs secret sharing on the user-entered data and distributes the shares among the three miner hosts using HTTPS connections.

Security The representatives of an ITL member company can log in to the ITL member area over an HTTPS connection using either their credentials (username and password) or more securely, using the Estonian ID-card or Mobile-ID.

We use access tokens to make sure that only representatives of ITL member companies are able to send shares to the miners. A random access token is generated by the ITL web server and sent together with the form each time the financial data submission form is requested by one of the logged-in users. The JavaScript library used in the submission form sends this token together with the corresponding shares and other submission data to each miner. Before saving the received shares into the database, the miner contacts the ITL web server and confirms that this token was really generated for the current submission form, the current company and has never been used for any submission before. The latter means that access tokens also act as nonces to rule out any replay attacks.

All the communication between a miner and the ITL web server is done over the HTTPS protocol and a unique, previously agreed and pre-configured passphrase is used to identify each miner to the ITL web server. If a miner receives a positive reply from the ITL web server, it saves the received shares to its local database and notifies the submission form. If the latter receives these notifications from all three miners, it marks this submission form as "submitted" in the ITL web server. This also invalidates the used nonce.

4.3 Maintaining confidentiality during data analysis

After the data collection period has ended, the secure MPC protocols can be started. Each Sharemind miner has a copy of a SecreC script that loads shares from the miner's database and uses a secure MPC implementation of an oblivious Batcher's odd-even merge sorting network [3] to sort the underlying private data vector. All of the collected financial indicator vectors are sorted separately in that

Analysis operation	Required MPC primitives
Sorting each financial indicator vector.	Oblivious sorting algorithm using a sorting network. Requires multiplication, addition and comparison.
Privacy-preserving filtering to keep only the data values that were really submitted by the end user.	Casting boolean to integer, vector multiplication.
Calculating a new composite indicator, *added value per employee*.	Division of secret shared values.
Time series for each financial indicator over all of the three forms.	Sorting the columns in a secret shared matrix by the values in one of the rows.

Table 1: The secure analyses performed on the collected financial data.

manner and the results are published on the ITL web page member area for the ITL board members as an Excel spreadsheet. After reviewing the results, the board forwards this report to all other ITL members.

Security SHAREMIND uses the RakNet library[5] for its network layer. The RakNet library provides secure connections between the data miners using efficient 256-bit elliptic curve key agreement and the ChaCha stream cipher [5]. While the latter choice is not standard, the best known attacks against ChaCha are still infeasible in practice [2]. This technique is used to encrypt all the communication between the SHAREMIND miners as well as between the miners and the controller applications (e.g. analysis applications).

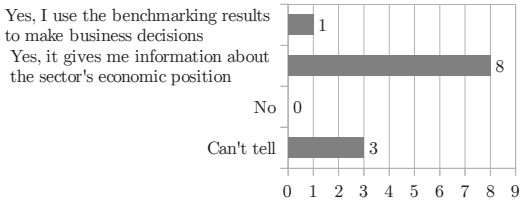
5 Secure financial statistics in practice

The described solution was deployed in the beginning of 2011 and has been already used to collect financial data for several periods. After each data collection period, the system used secure MPC protocols to sort each financial indicator vector and published the results as a spreadsheet for the ITL board.

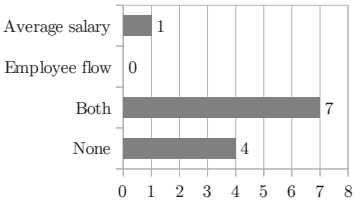
In addition to this, the ITL board requested a few extra reports. A list of the analyses performed on the collected financial data, together with the required computational routines, are listed in Table 1. The implementation was relatively effortless as we were able to create new algorithms in SECREC and deploy them at the miners. This justifies the use of a general-purpose secure MPC framework.

We conducted a survey among ITL members in the second data collection period, asking about the motivation and privacy issues of the participants. While the number of responders is not large enough to draw statistically significant conclusions, they still cover the most important players in the Estonian ICT market. As seen in Figure 2a, most of the participants feel that collecting and analyzing the sector's financial indicators is beneficial for themselves. We can also see that most of the participants are concerned about their privacy as they

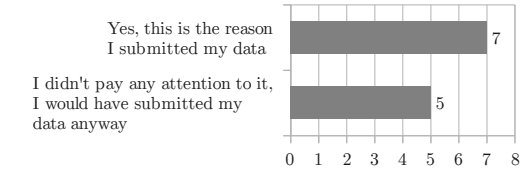
[5] RakNet – Multiplayer game network engine, http://www.jenkinssoftware.com



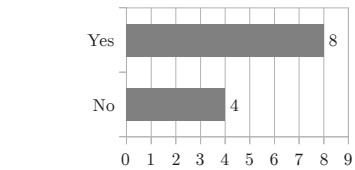
(a) Does collecting and analyzing the economic information benefit your company in any way?



(b) Which extra indicators are you willing to submit for the anonymous analysis?



(c) Did the explanation of applied security measures make it easier for you to submit your sensitive information?



(d) Did you familiarize yourself with the provided materials that explained which security measures were taken to protect the sensitive information?

Fig. 2: Results from the feedback questionnaire.

familiarized themselves with the security measures taken to protect the privacy of the collected data (Figure 2d) and about half of the participants submitted their data only because they felt that the system is secure in that matter (Figure 2c). The fact that most of the participants are willing to submit even more indicators (see Figure 2b) shows once more that ITL members are pleased with the security measures employed in this system to protect the participants' privacy.

6 Conclusions and Future Work

We have described a solution for securely collecting and analyzing financial data in a consortium of ICT companies. Companies are usually reluctant to disclose their sensitive financial indicators, as it is difficult for them to trust the parties who have access to their data for the purpose of analyzing it. The use of secure MPC means that the companies do not have to trust any one party unconditionally and their sensitive data stays private throughout the analysis process.

The system was implemented and deployed in the beginning of 2011 and is in continuous use. To the best of our knowledge, this is the first practical secure MPC application where the computation nodes are in separate geographic locations and the actual MPC protocol is run on real data over the internet.

A survey conducted together with one of the collection periods shows that ICT companies are indeed concerned about the privacy of their sensitive data and using secure MPC technology gives them enough confidence to actually participate in the collective sector analysis process. Moreover, thanks to the increased security and privacy measures, many companies are also willing to submit some extra indicators during the data collection process in the future.

Based on the experience of the ITL financial statistics application we conclude that MPC-based applications can be successfully deployed for real-life

problems. Performance of the available implementations is no more a bottleneck, but more effort needs to be put into making application deployment and administration easier. Our current setup works over open internet, but still assumes relatively well controlled environment for the miner hosts. The next logical step is to study the challenges arising from cloud-based installations, and this remains a subject for future developments.

References

1. SecureSCM. Technical report D9.1: Secure Computation Models and Frameworks. http://www.securescm.org, July 2008.
2. Jean-Philippe Aumasson, Simon Fischer, Shahram Khazaei, Willi Meier, and Christian Rechberger. New Features of Latin Dances: Analysis of Salsa, ChaCha, and Rumba. In *Proc. of FSE '08*, volume 5086 of *LNCS*, pages 470–488. Springer, 2008.
3. K. E. Batcher. Sorting networks and their applications. In *Proc. of AFIPS '68*, pages 307–314. ACM, 1968.
4. Assaf Ben-David, Noam Nisan, and Benny Pinkas. FairplayMP: a system for secure multi-party computation. In *Proc. of CCS '08*, pages 257–266. ACM, 2008.
5. D.J. Bernstein. ChaCha, a variant of Salsa20. *http://cr.yp.to/chacha.html*, 2008.
6. G.R. Blakley. Safeguarding cryptographic keys. In *Proc. of AFIPS '79*, pages 313–317. AFIPS Press, 1979.
7. Dan Bogdanov, Sven Laur, and Jan Willemson. Sharemind: A Framework for Fast Privacy-Preserving Computations. In *Proc. of ESORICS '08*, volume 5283 of *LNCS*, pages 192–206. Springer, 2008.
8. Dan Bogdanov, Riivo Talviste, and Jan Willemson. Deploying secure multiparty computation for financial data analysis. Cryptology ePrint Archive, Report 2011/662, 2011.
9. Peter Bogetoft, Dan Christensen, Ivan Damgård, Martin Geisler, Thomas Jakobsen, Mikkel Krøigaard, Janus Nielsen, Jesper Nielsen, Kurt Nielsen, Jakob Pagter, Michael Schwartzbach, and Tomas Toft. Secure multiparty computation goes live. In *Proc. of FC '09*, volume 5628 of *LNCS*, pages 325–343. Springer, 2009.
10. Martin Burkhart, Mario Strasser, Dilip Many, and Xenofontas Dimitropoulos. SEPIA: Privacy-Preserving Aggregation of Multi-Domain Network Events and Statistics. In *Proc. of USENIX Security Symposium '10*, pages 223–239, 2010.
11. J. Feigenbaum, B. Pinkas, R. Ryger, and F. Saint-Jean. Secure computation of surveys. In *EU Workshop on Secure Multiparty Protocols*, 2004.
12. J. Feigenbaum, B. Pinkas, R. Ryger, and F. Saint-Jean. Some requirements for adoption of privacy-preserving data mining. *PORTIA Project White Paper*, 2005.
13. Wilko Henecka, Stefan Kögl, Ahmad-Reza Sadeghi, Thomas Schneider, and Immo Wehrenberg. TASTY: tool for automating secure two-party computations. In *Proc. of CCS '10*, pages 451–462. ACM, 2010.
14. Roman Jagomägis. SecreC: a privacy-aware programming language with applications in data mining. Master's thesis, Inst. of Comp. Sci., Tartu University, 2010.
15. Lior Malka and Jonathan Katz. VMCrypt - modular software architecture for scalable secure computation. Cryptology ePrint Archive, Report 2010/584, 2010.
16. Adi Shamir. How to share a secret. *Commun. ACM*, 22:612–613, November 1979.
17. Riivo Talviste. Deploying secure multiparty computation for joint data analysis — a case study. Master's thesis, Inst. of Comp. Sci., Tartu University, 2011.