UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science
Cybersecurity Curriculum

Meril Vaht

The Analysis and Design of a Privacy-Preserving Survey System

Master's Thesis (30 ECTS)

Supervisor: Dan Bogdanov, PhD

Tartu 2015

The Analysis and Design of a Privacy-Preserving Survey System

Abstract:
There are many topics that are needed to be analyzed and, at the same time, the answers of respondents can not be public. Collecting sensitive data requires applying privacy-preserving security measures. This master's thesis describes the design and business processes of the prototype of a secure survey system using secure multi-party computation. The business processes of the system are introduced using activity diagrams, the use cases and the state machine diagram. The design of the system is also described in this paper and is illustrated with a deployment model. Based on the analysis, the prototype has been implemented and the system will be used to conduct real surveys in the near future.

Keywords: secure multi-party computation, survey system, confidentiality of data

Privaatsust säilitava küsitlussüsteemi analüüs ja disain

Lühikokkuvõte:
Vajadus konfidentsiaalseid andmeid koguda ja analüüsida nõuab privaatsust säilitavate turvameetmete kasutusele võtmist. Käesolev magistritöö kirjeldab privaatsust säilitava, turvalisel ühisarvutusel põhineva küsitlussüsteemi prototüübi analüüsi ning disaini. Süsteemi äriprotsesside kirjeldamiseks on kasutatud tegevusskeeme, kasutuslugusid ning olekumasina skeemi. Lisaks kirjeldab töö süsteemi ülesehitust ning esitleb juurutusskeem. Prototüüp on realiseeritud töös kirjeldatud analüüsi põhjal ning süsteemi on lähitulevikus plaanis kasutada ka praktiliste küsitluste läbiviimiseks.

Võtmesõnad: turvaline ühisarvutus, küsitlussüsteem, andmete konfidentsiaalsus

Contents

1 Introduction

A survey is a method for collecting information from a sample of individuals representing themselves or some organization. It is a study where data is collected by asking people questions. The same questionnaire is distributed to each respondent, data will be collected and later analyzed. Organizer is a person who represents himself or some organization and has decided to create a survey to collect data from respondents. Respondent is a person who represents himself or some organization and has decided to participate in a survey.

A sample is referring to a group of people who are selected from the population to take part of the survey. There are many different possibilities to gather the sample. The type of the sample depends on what is the target group of the survey and what kind of respondents are expected to answer the questions. Respondents of the survey might be chosen randomly (random sampling or volunteer sampling) or, for example, regarding the fact that they might be living in a particular area (cluster sampling) or be users of some service or product (purposive sampling). Respondents also might be gathered amongst people working on some particular field (purposive sampling) or the sample might be created as a national representative (stratified sampling).

There can be a variety of purposes for conducting survey research. Surveys are a simple way for getting feedback from customers about some services or products (customer satisfaction survey), discovering customers' expectations (market research survey), making better political (political satisfaction survey) or marketing decisions (market research survey), for succeeding in innovative production (market research survey), gathering statistics (statistical survey), measuring employee satisfaction (employee satisfaction survey), being able to predict the behaviour of voters (political satisfaction survey) or for gathering any kind of opinion or reaction from the respondents (event planning survey, student satisfaction survey, patient survey, etc).

There are many topics that are needed to be analyzed but at the same time, the answers of respondents can not be public. The data of the surveys that collect information about personal health or personal income or any other surveys that require revealing personal information of a respondent must be kept private and not disclosed to any third parties. It is crucial that the initial answers of a respondent will not be seen by anyone else but himself and no confidential information can be learned from the stored data. This requires applying privacy-preserving security measures.

Privacy-preserving computation and survey technologies are also a useful tool in cyber- and physical defence. For example, it could be used to improve collaborations within in coalitions and defence organizations, such as NATO (North-Atlantic Treaty Organization) and EDA (European Defence Agency). Consider the following example. If members of the EDA were to exchange information on the kind of cyber attack patterns they are witnessing in their systems, it could help bolster the response mechanisms in the coalition. However, information about which attacks target which countries can leak information about the defensive weaknesses of the said country, making the related information highly confidential. Secure data collection technology, such as the privacy-preserving survey system described in this thesis, can help improve such collaborations.

The aim of this master's thesis is to describe the design and business processes of the prototype of the secure survey system using secure multi-party computation which enables survey organizer to collect sensitive data without causing harm to participants by disclosing their answers. The main workflows that are required to function correctly and securely in this prototype are the transfer of data from the input party to the computing parties, secure multi-party computation protocols and the transfer of results from computing parties to the result party.

The business processes are modelled using Unifed Modelling Language (UML). The reason for choosing UML as a modelling language is to analyze whether designing the business processes of the applications that are based on secure multi-party computation affect the system analysis methods that are used in an IT-company that applies UML in its development process.

Authors contribution in this work was to analyze the business processes of the secure survey system and to cooperate with developers from Cybernetica AS and Partisia in the PRACTICE EU FP7 project for finding the best set of the processes that should be included in the prototype. The author has modelled all processes of the prototype as activity diagrams as well as the state machine diagram. Also, she has written the use cases that describe the functionality in more detail and supported the implementation of the prototype by testing the functionality. Additionally, the author participated in writing the supplementary report [25] of the secure survey system for the final review.

In addition to the secure survey prototype, the author of this thesis has an experience in analyzing the business processes of another prototype that is based on secure multi-party computation – the prototype of the Estonian Tax and Customs Board (MTA) value-added tax (VAT) fraud detection system. The author analyzed the business processes of the system and is the co-author of the research report of Cybernetica [15] and the article published in Financial Cryptography [16]. Based on the experience, the author introduces the key differences in the deployment of SMC-based software compared to the usual solution.

2 Privacy in online surveys

This section describes the main steps of the survey process, different types of surveys and the issues with privacy when conducting a survey. Finally, the privacy guarantees of existing online survey platforms are described.

2.1 General survey process

Surveys are conducted by an organizer – the person or organization who wants to collect the data. The organizer can also contract a research or data collection organization that conducts the survey. When a service provider is used, this organization collects the data on client's behalf and provides final results to the client. Usually, the analysis of data will be conducted by the organizer of the survey, but there is also a possibility to use the services of an organization that is specialized on data analysis (see Figure 2).

Based on [19, 29, 23] and on author's knowledge from previous work experience in the field of data collection and analysis, the process of survey generally consists of eight steps (see Figure 1).

1. **Plan.** The most important phase of conducting a survey is planning. This is also the phase where hypotheses are set. The quality of planning will define whether the research will be successful or not, whether the data will be high-quality or not. To produce good analysis at the end, it is inevitable to plan the structure of a survey and method of sampling according to the goals determined and the final use of the results. The response rate of the survey is highly related to the design of the survey. The respondent must be motivated to answer the questions. The instructions, as well as question texts, must be clearly understandable. Also, completing the survey must not take unnecessarily long time.
   Result: Initial questionnaire in written form (ready for testing or ready for programming) and sample.

2. **Design.** In the case of online surveys or computer-assisted telephone interviews it is needed to program the questionnaire. In this phase, the questionnaire will be programmed by the organizer or the conducting organization. This phase can be skipped, if there is no need to program the questionnaire. For example, if a chosen survey will be conducted by using face-to-face interviews, making telephone interviews or gathering focus groups without the use of computers.
   Result: The first version of the programmed questionnaire.

3. **Test.** In order to avoid any problems while running the survey, it is important to always have a testing phase before the launch. While testing, it is also possible to find out the estimated time for completing the questionnaire, detect possibly problematic questions, etc. Testing should be done with a small group of possible respondents to find out if everything is understandable and to find out the actual time used to complete the survey.
   Result: Final version of the survey that can be used in a live deployment.

4. **Run.** Running the survey is the phase where the actual data is collected. Depending on the survey method, it is the period when interviewers will call respondents, meet the respondents at their homes or in public places or the period while the

online survey will be active.
Result: Answers from respondents.

5. **Prepare data.** The data has to be cleaned and prepared for analyzing. For example, if data was collected on paper, but the answers will be analyzed digitally, an organizer or conducting organization has to insert the answers manually into a digital database. If the survey was made online, data might have to be exported and transformed into the requisite format.
   Result: Cleaned data that is ready for analysis.

6. **Analyze.** In this stage, collected and prepared data will be analyzed, whether digitally or manually.
   Result: Tables and charts representing the analyzed data.

7. **Report.** When data is analyzed, it is time to report the results. The report should include conclusions made based on the analysis and be clearly understandable for the target group.
   Result: Usually a report in digital form based on the analysis including tables and charts of the analyzed data and concluding notes.

8. **Act.** When the survey is completed – data is analyzed and reported – it is time to compare the results with the hypotheses set in the planning phase and use the results for the objectives stated. The organizer will use the results of the survey for making better marketing decisions, changing services according to the opinion of the respondents, using statistical information to predict the behaviour of a customer, etc.
   Result: Actions taken, changes applied, etc.

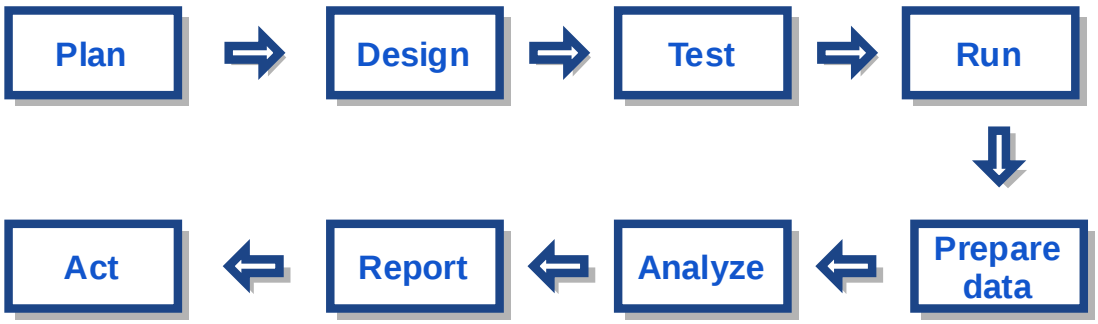

Figure 1: General survey process

Actors that can be responsible for the processes of administering different stages of the survey are organizer, conducting organization and/or analyzing organization. The associations of processes and actors can be seen on Figure 2.

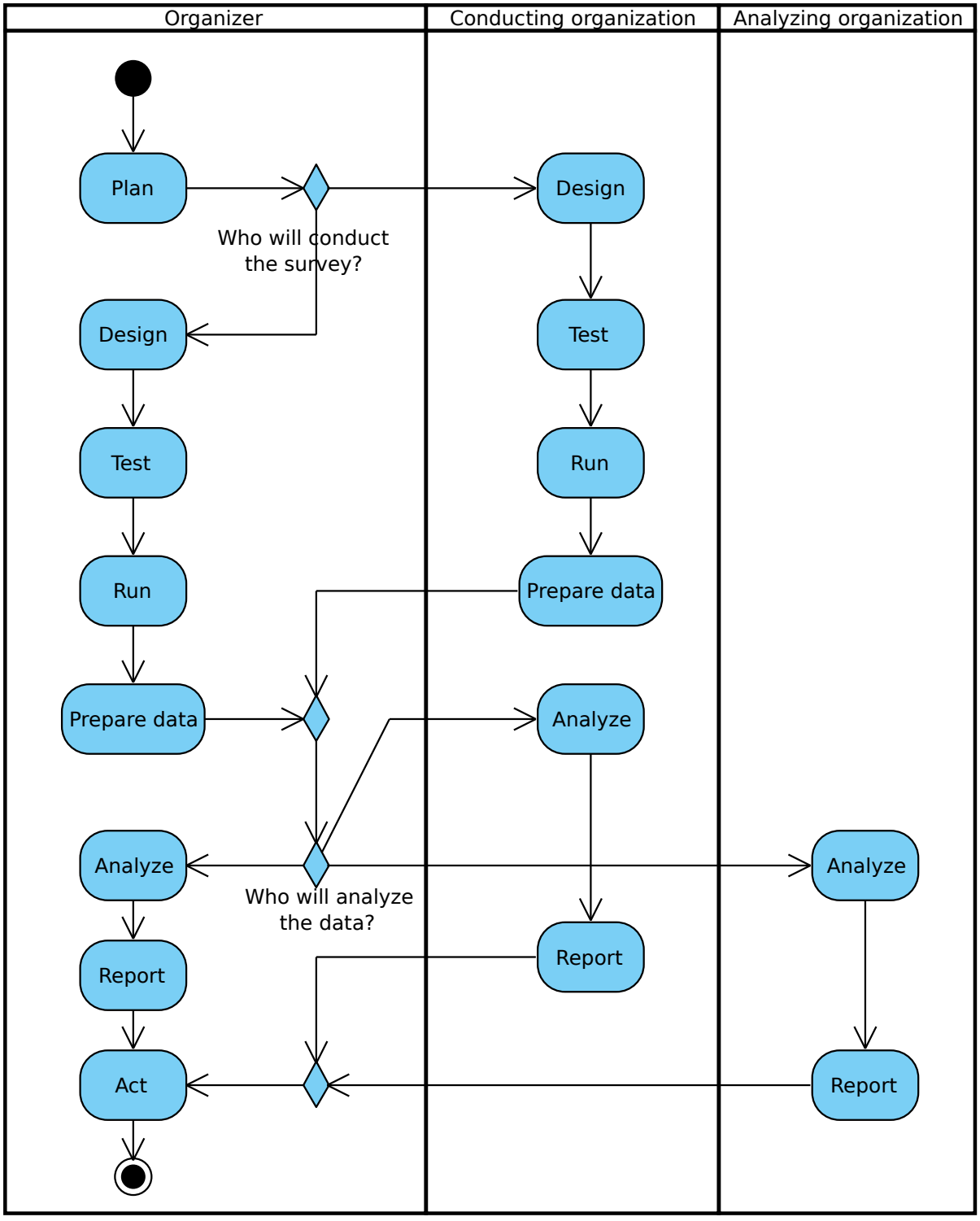

Figure 2: The activity diagram of survey process

2.2 Survey technologies

There are many different possibilities for data collection. The most popular methods are following.

1. **Face-to-face interviews** – Interviewing the respondent face-to-face either in his/her home or in a public place.
2. **Telephone interviews** – Interviewing the respondent over the phone.
3. **Online surveys** – Conducting surveys online.
4. **Focus groups** – Gathering respondents in one room and asking them questions while observers examine their reactions.
5. **Paper surveys** – Collecting the answers for a survey on a paper, sending questionnaires either by mail or, for example, handing them out in a supermarket.

The best method for a survey depends on a topic, target group, structure, question types, length of a survey, etc. See Table 1.

Type of survey	**Suitable for**	**Unsuitable for**	**Relative cost**
Face-to-face survey	Long interviews. Interviews that might need additional instructions.	Short period surveys. Surveys that require revealing sensitive data.	Expensive
Telephone survey	Medium time interviews. Interviews that might need additional instructions or encouragement for answering the survey.	Long interviews. Surveys that require revealing sensitive data.	Medium cost
Online survey	Short time to long time interviews. Surveys that require revealing sensitive data.	Research that requires answers from respondents who possibly might not have access to the internet.	Low cost
Focus group	Medium time to long time interviews. Interviews that require observation of the respondent's reaction.	Research that requires answers from respondents from a wide geographical area. Surveys that require revealing sensitive data.	Medium cost
Paper survey	Medium time to long time interviews.	Short period surveys.	Low cost

Table 1: The comparison of different survey types

2.3 Privacy in surveys

Surveys may be conducted practically in every field. The research fields of surveys include salary surveys to collect data about labor market, employee satisfaction surveys, market research surveys, health and medical care surveys, psychological surveys, scientific surveys, government surveys. Surveys about transportation, politics, legislation, media, education, economy, religious beliefs, etc are also common.

The main concerns for respondents is usually their anonymity and the confidentiality of collected data. Perceptions of privacy may be the most influential factor in a respondent's decision whether to participate in a survey or to give truthful answers. Respondents must be assured, that the survey is legitimate and their personal data will be protected.

2.3.1 Legislation

Data protection legislation is concerned with the processing of personal data and applies to anyone involved in the collection, processing and use of market research data – the organizer, conducting organization and analyzing organization. For example, according to the the European Union Personal Data Protection Act [6], personal data can only be processed (e.g. collected and further used) if the data subject has unambiguously given his consent. Researchers must take into account all state, federal and international regulations. As surveys may often be international and the regulations may differ in every state, it is important to know and apply the regulations separately in every state the survey will be conducted in, not just in the state the survey is created.

For example, the Estonian Personal Data Protection Act defines personal data as follows "Personal data is any data concerning an identified or identifiable natural person, regardless of the form or format in which such data exist" [6]. In addition, the Act also defines sensitive personal data as:

1) data revealing political opinions or religious or philosophical beliefs, except data relating to being a member of a legal person in private law registered pursuant to the procedure provided by law;
2) data revealing ethnic or racial origin;
3) data on the state of health or disability;
4) data on genetic information;
5) biometric data (above all fingerprints, palm prints, eye iris images and genetic data);
6) information on sex life;
7) information on trade union membership;
8) information concerning commission of an offense or falling victim to an offense before a public court hearing, or making of a decision in the matter of the offense or termination of the court proceeding in the matter.

Similarly, the European Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data [1], Article 8 states that Member States shall prohibit processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, and the processing of data concerning health or sex life unless the data subject has given his explicit consent to the processing of those data. In some fields, e.g. in clinical trials, it is required that the respondent

must sign a separate informed consent form.

Moreover, Article 17 of the Convention states that Member States shall provide that the controller must implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access, in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing.

To be able to transfer data from the European Union to the United States, for example, while conducting international surveys, a researcher must adhere to the U.S.-EU Safe Harbor framework for collecting data from EU [3] principles which regulate the collection, use and retention of personal information from EU member countries. The U.S.-EU Safe Harbor Framework [4] provides guidance for U.S. organizations on how to provide adequate protection for personal data from the EU as required by the European Union's Directive on Data Protection.

In case the purpose of conducting a survey is to collect personal health information in U.S., compliance with the Health Insurance Portability and Accountability Act [2] of 1996 (HIPAA) is required. HIPAA provides for the protection of individually identifiable health information that is transmitted or maintained in any form or medium. The privacy rules affect the day-to-day business operations of all organizations that provide medical care and maintain personal health information.

2.3.2 Self-regulation and standards

In addition to regulations, there are also ethical rules and other good practices that can convince the respondent about confidentiality of his responses, e.g. CASRO Code of Standards and Ethics for Survey Research [13]. CASRO (The Council of American Survey Research Organizations) is the national association established in 1975 to represent the U.S. research industry and those organizations engaged in the conduct, support, or education of market, opinion, and social research, often described as data analytics, consumer insights, or business intelligence [13].

An example of good practices according to SurveyMonkey [31]: "It is always good to disclose your privacy practices to your respondents. Doing this helps to increase response rates by putting potential respondents more at ease." According to SurveyMonkey, the privacy policy should include information about the personal information that will be collected, information about how the responses will be used, information about whether responses will be accessible for third parties and contact information.

Moreover, there is an international standard that provides a high-level framework for the protection of personally identifiable information (PII) within information and communication technology (ICT) systems – ISO/IEC 29100:2011 [7]. It is intended to help organizations define their privacy safeguarding requirements related to PII within an ICT environment by referencing following principles:

1. **Consent and choice** – provide PII principal with the opportunity to choose how their PII is handled and allow a PII principal to withdraw consent. Also PII prin-

cipal possibility to give a consent in relation to the processing of their PII at the time of collection, first use or as soon as practicable thereafter.

2. **Purpose legitimacy and specification** – ensure that the purpose(s) comply with applicable law, inform PII principal about the purpose(s) before the information is collected or used for the first time for a new purpose in a clearly understandable way.

3. **Collection limitation** – limit the collection of PII to that which is within the boudns of applicable law and strictly necessary for the specified purpose(s).

4. **Data minimization** – minimize the PII which is processed and the number of privacy stakeholders and people to whom PII is disclosed or who have access to it.Delete and dispose PII whenever the purpose for PII processing has expired or there are no legal requirements to keep the PII.

5. **Use, retention and disclosure limitation** – limit the use, retention and disclosure (including transfer) of PII to that which is necessary in order to fulfil specific, explicit and legitimate purposes. Retain PII only as long as necessary to fulfil the stated purposes, and thereafter securely destroy or anonymize it.

6. **Accuracy and quality** – ensure that the PII processed is accurate, complete, up-to-date, adequate and relevant for the purpose of use. Ensure the reliability of PII collected from a source other than from the PII principal before it is processed.

7. **Openness, transparency and notice** – provide PII principals with clear and easily accessible information about the PII controller's policies, procedures and practices with respect to the processing of PII.Notice PII principals about the fact that PII is being processed, the purpose for which this is done, the types of privacy stakeholders to whom the PII might be disclosed and the identity of the PII controller. Give notice to PII principals when major changes in the PII handling procedures occur.

8. **Individual participation and access** – give PII principals the ability to access and review their PII. Allow PII principals to challenge the accuracy and completeness of the PII and have it amended, corrected or removed if requested.

9. **Accountability** – document and communicate as appropriate all privacy-related policies, procedures and practices. Assign to a specified individual within the organization the task of implementing the privacy-related policies, procedures and practices. Provide suitable training for the personnel of the PII controller who will have access to PII. Measures to remediate a privacy breach should be proportionate to the risks associated with the breach but they should be implemented as quickly as possible.

10. **Information security** – protect PII under its authority with appropriate controls at the operational, functional and strategit level to ensure the integrity, confidentiality and availability of the PII, and protect it against risk such as unauthorized access, destruction, use, modification, disclosure and loss throughout the whole of its life cycle. Limit access to those individuals who require access to perform their duties.

11. **Privacy compliance** – Verify and demonstrate that the processing meets data protection and privacy safeguarding requirements by periodically conducting audits using internal auditors or trusted third-party auditors. Develop and maintain privacy risk assessments, have appropriate internal controls and independent supervision mechanisms in place that assure compliance with relevant privacy law and wuith their security, data protection and privacy policies and procedures.

2.4 Online surveys

For conducting an online survey it is possible to choose amongst several possible survey building tools, that can be divided into three levels.

The first level includes free online tools (like Google Forms[1] and LimeSurvey[2]) which include the basic functionality needed to create a simple questionnaire. These tools are the best option for people who rarely conduct a survey or do not have any special requirements for the questionnaire.

The second level includes online survey tools that offer basic functionality for free, but users have a possibility to get many additional features if they pay for them (tools like SurveyGizmo[3] and QuestionPro[4]). Free versions of these tools are the best option for people who also rarely conduct a survey or do not have any special requirements for the questionnaire. Free or paid versions of these tools are a great choice for scientists, smaller organizations or freelancers, who might have higher requirements for questionnaire or who might conduct surveys more regularly.

The third level includes tools not free of charge (for example Confirmit[5] and Quretec[6]). These are usually advanced systems where users are also able to include more complex programming logic to manage the conduction of a questionnaire and many other advanced features. Tools not free of charge are mostly used by organizations that are specialized to market research or any kind of data collection and which need advanced techniques for building questionnaires and analyzing the data.

2.4.1 Online survey techniques

Depending on the technique, online surveys can be additionally divided into subcategories. There are different possibilities for conducting an online survey, for example e-mail surveys, mobile surveys, on-site polls and panels.

E-mail surveys – Allows to send invitations to participants by e-mail. This is the most popular choice when conducting a survey that is directed to certain target group. This is also, a good choice if the questionnaire is quite long and time consuming. Some of the

[1]Google Forms, https://www.google.com/forms/about/
[2]LimeSurvey, https://www.limesurvey.org/
[3]SurveyGizmo, http://www.surveygizmo.com/
[4]QuestionPro, http://www.questionpro.com/
[5]Confirmit, http://www.confirmit.com/
[6]Quretec, http://www.quretec.com/

online tools that provide this functionality are QuestionPro, Google Forms and FreeOnlineSurvey.

Mobile apps – Allows to get a higher response rate due to the flexibility offered to the respondents. In addition, these apps enable to collect information like geolocation, camera, audio/video recording, etc, due to mobile devices' capacity. Data is collected and treated in real time. The respondent does not have to be by the computer, but can fill the questionnaire practically anywhere. Online tools that provide this functionality are, for example, Qualtricks, FluidSurveys and SurveyMonkey.

On-site poll widgets – Allows the organizer to embed the poll on his website, to ask site visitors about their opinions, such as whether they favour or oppose a new policy or who they think is likely to win an upcoming election. These polls typically ask site visitors about their opinions and are usually attractive features for news sites. Tools that can be used for creating on-site poll widgets are, for example, EasyPolls, PollMaker and PollSnack.

Panels – Allows respondents to take voluntarily part of the surveys. Respondents can join an online research and data collection panels, where they can collect bonus points or money as a reward for completing the survey. Respondents have possibility to fill initial questionnaire and based on the answers get regular invitations to surveys with suitable target group. Online tools that provide this functionality are, for example, iPoll, i-Say and MySurvey.

2.4.2 Most popular online survey platforms and their security and privacy features

Some of the most popular online survey tools are SurveyMonkey [31], SurveyGizmo [30], QuestionPro [28], Qualtrics [26] and Google Forms [22]. The security aspects of these tools that are described on their official websites are briefly introduced in Table 2. The table communicates the cryptographical security methods of every survey tool mentioned above. It also describes which methods are used to ensure the security of the processes, e.g. certification or compliance with the laws. The table also points out how the security in the user interface is guaranteed.

Online tool	**Cryptographical security methods**	**Security of processes**	**Security in user interface**
SurveyMonkey	Collecting data over secure SSL/TLS encryption channels; Sensitive user data is stored in encrypted format	HIPAA compliance; SSAE-16 SOC II certified data centers [8]; TRUSTe certification [10]	User authentication
SurveyGizmo	Collecting data over secure SSL/TLS encryption channels; Project data is stored in encrypted format	HIPAA compliance; Safe Harbor certification [4]	Password protected surveys; Password protected reports
QuestionPro	Collecting data over secure SSL/TLS encryption channels; Sensitive user data is stored in encrypted format; Project data is stored in encrypted format	SSAE-16 SOC II certified colocation facilities [8]; Respondent anonymity assurance [27]; TRUSTe certification [10]; Safe Harbor certification [4]	Password protected surveys; Possibility to assign unique password to respondents
Qualtrics	Collecting data over secure SSL/TLS encryption channels; Project data is stored in encrypted format	HIPAA compliance; SSAE-16 SOC II certified data centers [8]; FISMA Act of 2002 requirements guaranteed [5]; Safe Harbor certification [4]	Password protected surveys
Google Forms	Collecting data over secure SSL/TLS encryption channels	Role-based access control inside the company	Two step verification when you access Google Account

Table 2: The comparison of different online survey tools' privacy guarantees

3 Secure multi-party computation

This section describes secure multi-party computation (SMC). In more detail, it describes what is SMC and how it helps to preserve the privacy of data. This section also gives some examples of practical SMC applications.

3.1 Overview

Secure multi-party computation (SMC) is a cryptographic technology for computing a function with multiple parties where all parties know their own input value and can learn the final output value but are not able to see each other's inputs. For example, if a number of parties $\mathcal{P}_1, ..., \mathcal{P}_n$, who initially hold inputs $x_1, ..., x_n$, wish to compute the value of a function $f(x_1, x_2, x_3, ..., x_n) = (y_1, y_2, y_3, ..., y_n)$, then every party $\mathcal{P}_i$ can only learn the value of y_i. It allows to process private data without compromising anyone's privacy.

One of the most popular example of the necessity of using SMC in real life is the millionaires' problem introduced by Andrew C. Yao [12] where two millionaires wish to know who is richer. However, they do not want to reveal any additional information about each other's wealth. For example, person $\mathcal{P}_i$ knows the value of x_i and no other input values. The millionaires' problem corresponds to the case when $n = 2$ and $f(x_1, x_2) = 1$ if $x_1 < x_2$ and 0 otherwise.

There are a number of possibilities for implementing SMC. In our secure survey prototype we use SMC based on secret sharing technology. Secret sharing was introduced by Adi Shamir in [11] and George Blakley in [21]. In secret sharing, a secret value s is split into a number of shares $(s_1, s_2, s_3, ..., s_n)$ and the shares are distributed to different instances. Each share looks random to the holder and a predetermined number of shares is needed to reconstruct the value.

Conducting a survey that requires respondents revealing sensitive data is one of the many fields that would benefit from the use of SMC, so we took this opportunity to implement a practical SMC survey application prototype.

3.2 Examples of practical SMC applications

This subsection briefly introduces some examples of implemented practical SMC applications.

3.2.1 Danish sugar beet auction

One of the first practical secure multi-party application experiments, the Danish sugar beet auction, was reported in 2008 [24]. In Denmark, several thousand farmers produce sugar beets, which are sold to the company Danisco, the only sugar beets processor on the Danish market. When the EU drastically reduced the support for sugar beet production there was an urgent need to reallocate contracts to farmers whose productions are the most beneficial. This was best done via a nation-wide exchange, a double auction.

To satisfy all parties, the actual bids had to be hidden from others. It would not have been acceptable for farmers if, for example, Danisco would have acted as the only auctioneer. Therefore, the solution was to implement an electronic double auction using three-party multi-party computation, where computing parties were Danisco, DKS and the SIMAP project.

In the system, a web server was set up for receiving bids, and three servers were set up for doing the secure computation. To submit the data, every participant had to download the program to their computer together with the public keys of the computation servers. The secret shares were not actually sent directly from the computers to the computing nodes. Each share was first encrypted with a public key of one of the nodes and then stored in a database by the web server. After that, the representative for each of the involved parties triggered the computation by inserting his USB stick, where the private key material was stored, and entering his password on his own machine. For the technical description of the sugar beet auction see [20].

3.2.2 SMC application for financial data analysis

Another practical application that uses SMC is the prototype reporting system for a consortium of ICT companies (Estonian Association of Information Technology and Telecommunications – ITL) for collecting financial data twice a year to analyze the economic situation of an industrial sector [18]. The application has been already used several times for collecting financial information. To eliminate the fact that the collected data would be accessible for the board of ITL as this might reduce the number of companies that agree to submit their data, the application was built on Sharemind [14] secure computation framework as the data needed for analysis is highly confidential and must not be disclosed. This is the first application which was used to make SMC computation on real data over the internet with computing nodes spread geographically apart.

The companies hosting the three computing nodes were chosen amongst the ITL consortium (Cybernetica, Microlink and Zone Media), so they would have no intention to collude as they are also submitting their own private data in the system and want to ensure the privacy of their data. The data was submitted through an online submission form that was integrated into ITL webpage. Submitted data was secret-shared at the source and distributed among the three computing parties. After the deadline, the computing parties engaged in SMC protocols and collected all economic indicators independently. Then the indicators were published to the board of ITL as a spreadsheet that did not include any identifying information.

Together with the second collection period, there was also a simple feedback survey conducted among the members of the consortium, asking them about the motivation and possible privacy issues of participating in this data collection system. The results showed that about a half of the participants submitted their data only because they felt that the system is preserving their privacy.

3.2.3 Employee satisfaction survey

Cybernetica AS in Estonia has been using SMC application for conducting employee satisfaction survey for the past two years. This is another great example of a SMC application successfully used in real life, as the use of this application was successful both times. This application was also built on SHAREMIND [14] secure computation framework as the answers of employees must not be disclosed to third parties. Three computing parties were chosen amongst employees, as they would have no intention to collude as they are also submitting their own private data in the system and want to ensure the privacy of their data.

The data was submitted through online submission form and the secret shares of respondents' answers were stored in three different computers acting as computing nodes. After the deadline, the computing parties engaged in SMC protocols and data was aggregated. The aggregated data was published to the data collector without any identifying information.

3.3 Shortcomings of the state-of-the-art

This section describes the shortcomings of the existing online survey tools and implemented practical SMC survey tools.

The applications described in Section 2.4.2 put the users in position where they need to trust the one enterprise which controls the server(s) where the data is stored. Even if the privacy policy states that the data is securely stored and sensitive data can not be studied by third parties, it might not be enough to convince the user to trust the organization. This, however, may influence the response rate of the survey.

These applications also give the organizer the possibility to analyze the whole dataset of the initial answers by giving him/her direct access to the data. This means that it could be possible for organizer to recognize the respondent based on his/her answers if, for example, employee satisfaction survey is conducted. For example, if organizer has the information about the department, gender and position of a respondent, he/she can identify the person and with that know how that specific person answered to all the questions.

The applications described in Subsection 3.2.2 and 3.2.3 give user more confidence about privacy-preserving data storage. Even though these tools were successfully used in practice, there are some disadvantages regarding to these systems. First of all, the structure of the questionnaire is fixed. This means that the organizer of the survey can not build up complex questionnaires. The survey conducted using these systems must be very simple.

Secondly, the business processes of these systems are not analyzed thoroughly from the organizer's point of view. Therefore, the final reports' structure is fixed and very simple. Organizer can not choose, neither before activating nor after closing the survey, how he/she wants the computed data to be presented. Also, to publish the results, organizer must do it manually – report is not meant to be published directly from the system.

The secure survey system prototype described in this work, aims to reduce the shortcomings and provide more flexibility. Firstly, comparing to the existing online survey tools, it provides a new trust model by using secure multi-party computation. Moreover, the whole dataset of the initial answers will never be public, only computed results will be displayed to organizer (see Figures 3 and 4). The results can be computed only once and a minimum of five answers for each question are required.

Secondly, comparing to the implemented SMC applications, it allows the organizer to create more complex questionnaires and inspect the final results online in various formats. However, the secure survey prototype does not have as many features as a commercial survey software. The organizer can also print the report and publish it by sending the unique link to respondents (this functionality has not been completely implemented in the prototype by the time of submitting the thesis).

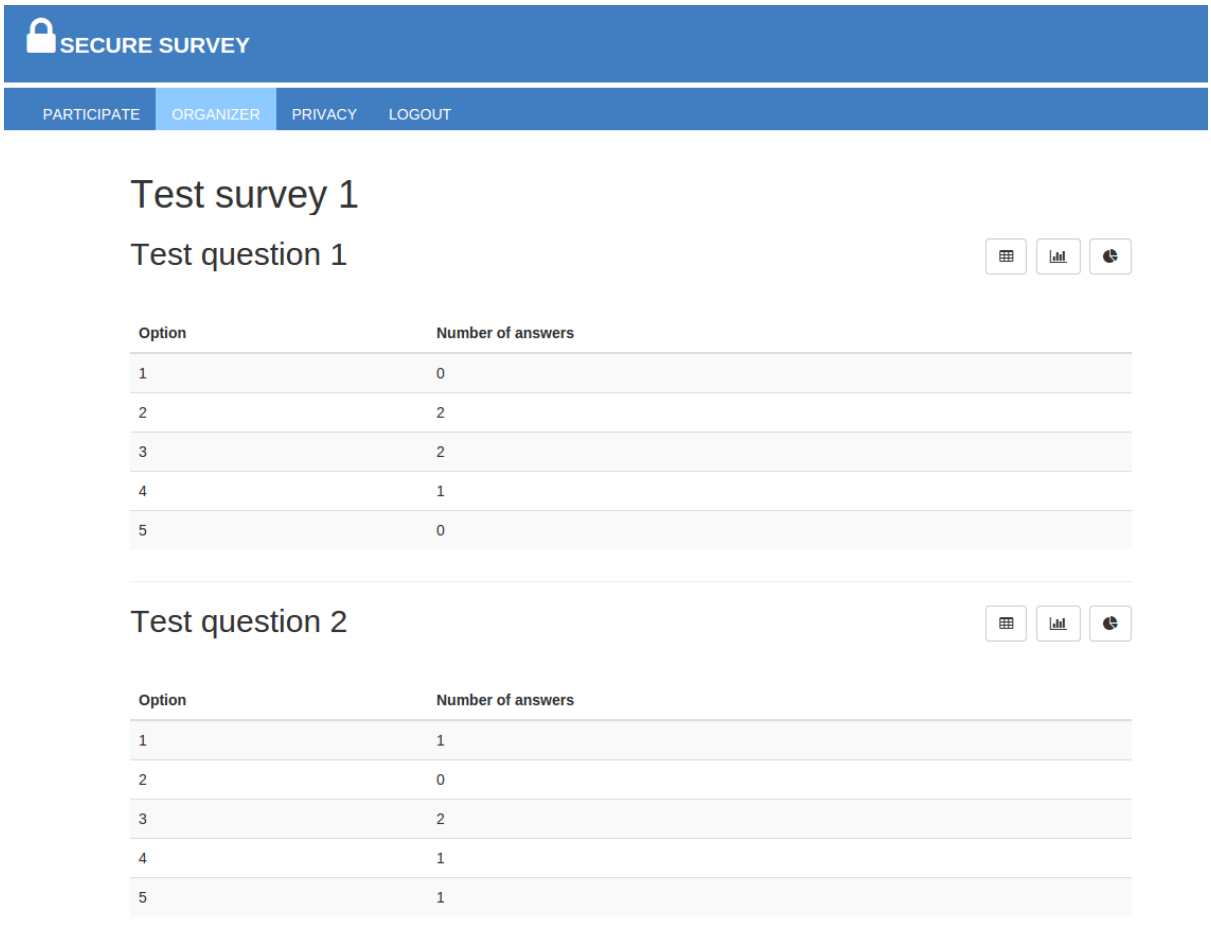

Figure 3: The table view of the report

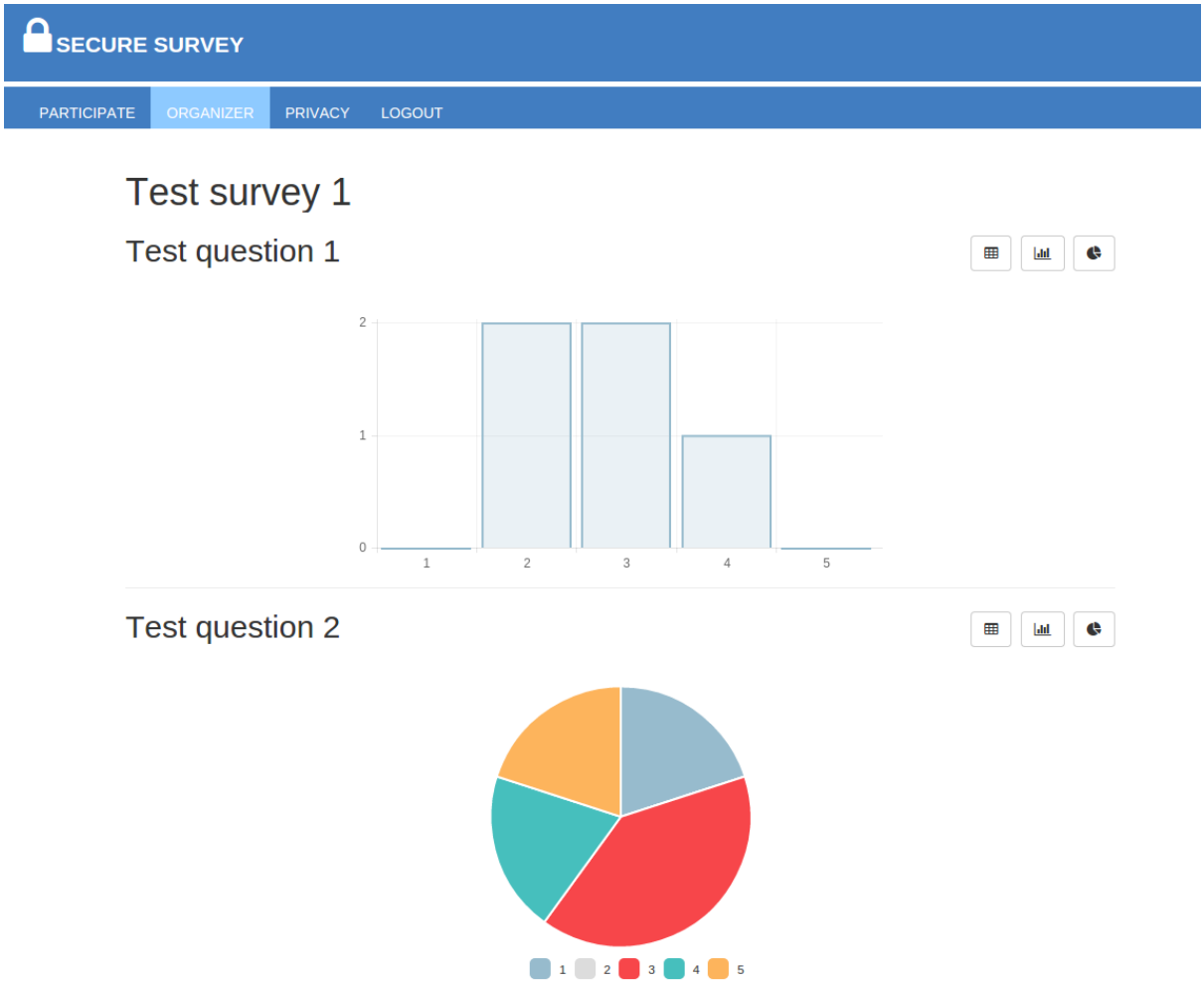

Figure 4: The chart view of the report

4 The secure survey system

The objective of the secure survey system prototype, which has been implemented during the PRACTICE EU FP7 project by Partisia, Cybernetica AS and Alexandra Institute, is to enable the survey organizer to collect sensitive data and produce statistics without causing harm to respondents by allowing third parties (intentionally or unintentionally) to study their individual answers. It is crucial that the initial answers of respondents will not be seen by anyone else but themselves and no confidential information can be learned from the stored data. Therefore, in the prototype, answer data is encrypted. Data about the survey itself (questions and possible answer options) is not encrypted in this prototype.

Given that the organizer who collects the data and participant who answers the questions are typically concerned about the sensitive data collected, it is very important that it can be clearly communicated through the secure survey system itself that no secret data will be leaked. In the system, there is a simple description about the privacy preservation techniques and information about the companies that host the servers. This information is accessible for all users of the system. When the participant is aware of the security guarantees of the system, he/she is probably more eager to participate in a survey or to give truthful answers.

The main workflows that are required to function correctly and securely in this prototype are the transfer of secret-shared answer data from the input parties (the respondents) to the computing parties, the engagement process of secure multi-party computation protocols and the transfer of secret-shared results from computing parties to the result party.

The secure survey system prototype is designed to run on two different secure multi-party computation engines: Sharemind and Fresco/SPDZ. This master's thesis concentrates on the implementation using the Sharemind 3 framework (see Subsection 4.1).

4.1 Sharemind 3

Sharemind [14] is a full framework for developing secure multi-party computation applications and it can be used for building applications that analyze confidential data in a way that privacy of the data owners will be preserved. Sharemind is using the additive secret sharing scheme with three parties connected over secure asynchronous network channels to preserve the confidentiality of data. This means that every secret value is split into three pieces called shares and every share is stored in a different server instance (see Figure 6) as a completely random bit sequence.

For example, if an input party wants to provide a secret value $x \in \mathcal{Z}$ (where $\mathcal{Z}$ is a finite integer ring) as a private input to n computing parties, it uniformly generates shares $x_1, ..., x_{n\text{-}1} \leftarrow \mathcal{Z}$ and calculates the final share $x_n = x - x_1 - ... - x_{n\text{-}1}$. Each computing party receives one share x_i (see Figure 5). As an individual share, x_i is just a uniformly distributed value and no computing party can learn anything about x without colluding with others. Computing parties can process the shares without recovering the secret. For example, if each computing party has shares x_i and y_i of secrets x and y, they can calculate $z_i = x_i + y_i$ to get the shares of $z = x + y$.

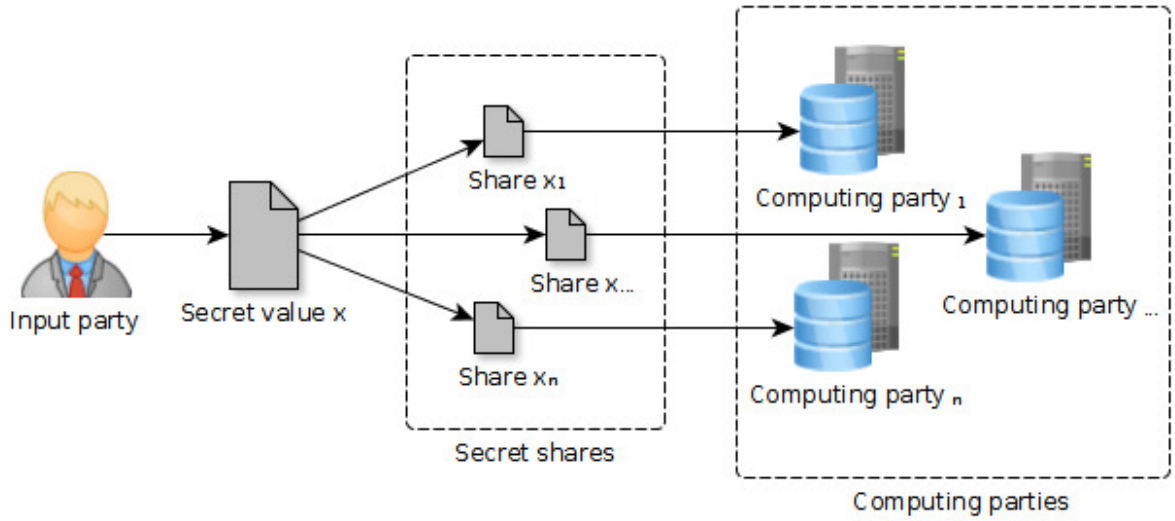

Figure 5: Example of secret-sharing

The secret sharing of secret values is performed at the source and each share is sent to a different server over a secure channel. This guarantees that no one but data owner will know the original value. Next, SHAREMIND server instances engage secure multi-party computation protocols to compute the results. When results are computed, the aggregated data will be available for the end user who can reconstruct the values of computation result.

SHAREMIND includes two different kinds of programs for different parties - server instances run the server software and other parties (data owners and end users) run the controller software. The server software consists of algorithms and protocols that perform the multi-party computation and data mining tasks. The controller software is used to send data and commands to the server instances and handles data encryption and decryption.

To ensure a high level of security, all server instances must be operated by independent parties to avoid collusion. Therefore, if shares from one server will be disclosed to adversary, he can not reconstruct the initial secret values by using only one share of them. SHAREMIND secure multi-party computation protocols are secure in the *honest-but-curious* model with no more than one passive corrupted party. Working in the passive model expects more honesty from computing parties, but makes it possible to have considerably faster computations over many other secure MPCs implementations. SHAREMIND is also capable of providing security against an active adversary, but with a lower efficiency.

SHAREMIND is using the high-level programming language SECREC("secrecy") [17] for developing privacy-preserving data mining algorithms for SHAREMIND applications. SECREC is a C-like language that allows to separate public and private data types. Variables that are typed as private are processed using secure computation whereas public values

are stored and processed as usual, this means that confidential data remains protected. The design of SecreC aims at simplifying the programming task and preventing the developer from making trivial privacy leaks.

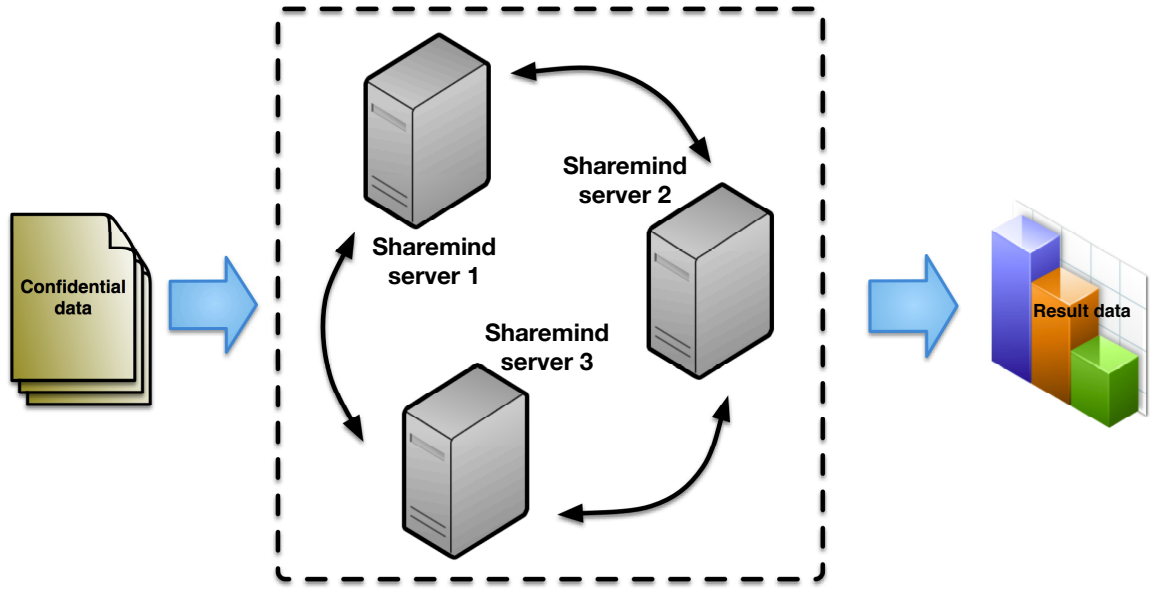

Figure 6: The deployment model of Sharemind

4.2 Actors and privacy goals

Actor	Description	Privacy goals
Organizer	A person who represents himself or some organization and has decided to create a survey to collect data from participants.	Wants to ensure (and to be able to confirm) that participants' answer data is kept confidential to collect data about sensitive topics and get higher response rate. Wants to comply with the data protection laws.
Participant	A person who represents himself or some organization and has decided to participate in a survey.	Wants that the confidentiality of his/her private data is guaranteed.
The owner of a computing instance	A person or an organization who hosts one of three computing instances and, also, is interested in keeping the data confidential and has no intention to collude with other computing instances' owners. At the same time, could also be an organizer or a participant.	Is either taking part of the survey himself or is interested in the final results of the survey and, therefore, is interested in protecting the confidentiality of private data.

The purpose of secure survey system is to guarantee that the privacy goals of the actors

described above will be fulfilled. The privacy preservation techniques that are used to accomplish the level of security that corresponds to the actors' needs are described in next subsection.

4.3 Privacy preservation techniques

This subsection highlights the privacy preservation techniques of the secure survey system that enhance the privacy of the system comparing to the existing online survey tools described in Section 2.4.2. It also describes how the increased security level is introduced to the users.

Secret-shared data
The secret sharing of respondents' answers is performed at the source and each share is sent to a different server over a secure channel. This guarantees that no one but data owner will know the original value because each share looks random to the holder and to reconstruct the initial value all shares are needed. Processing with SMC ensures that private inputs are not reconstructed even during report preparation. The whole dataset of the initial answers will never be public, only computed results will be displayed to organizer.

Distributed system
The survey system is a distributed cloud computing system that does not allow the individual cloud service provider to access any of confidential information at any time. Second, the control of the individual cloud service provider instances is distributed among independent parties, each knowing no more than the individual cloud service provider.

Minimum number of answers
A minimum of five answers for each question are required to analyze a survey and the results of the survey can only be computed once. Thus, the organizer cannot misuse the survey system to deduce otherwise confidential answers by comparison of repeatedly computed reports i.e. by comparing two results with only a single additional answer as the difference. The number of minimum answers can be modified in future versions of the system.

Public data checks
The client application is implemented in such a way that it accepts public data and results of user actions if and only if all the servers return identical results. This way it is possible to detect when public data is manipulated on any of the servers. If one of the servers wants to behave in a malicious way and edit the questions displayed to the user and thus mislead the participant, survey organizer would get inaccurate results for the survey. The implementation eliminates that risk with checking whether duplicated public data is identical.

5 Business processes of the secure survey system

This section presents the activity diagrams, use cases and state machine of the secure survey system prototype. The activity diagrams introduce the overall picture of functionalities of the survey system. In activity diagrams, the functionalities that are an important part of the privacy preservation are marked with the yellow color. Use cases describe activity diagrams in more detail. In the use case diagram, the use cases which contain the functionality that is important part of the privacy preservation are marked with the yellow color. Use cases also include the references to the state machine diagram (see Subsection 5.6). The state machine diagram gives an overview of the states and functions of the survey system.

The actor 'survey system' is connected to all use cases by default. The system is always an involved actor and is not connected to every use case in the use case diagram nor mentioned in the description of use cases' involved actors. The system is mentioned as an involved actor only if it is the only actor of this use case.

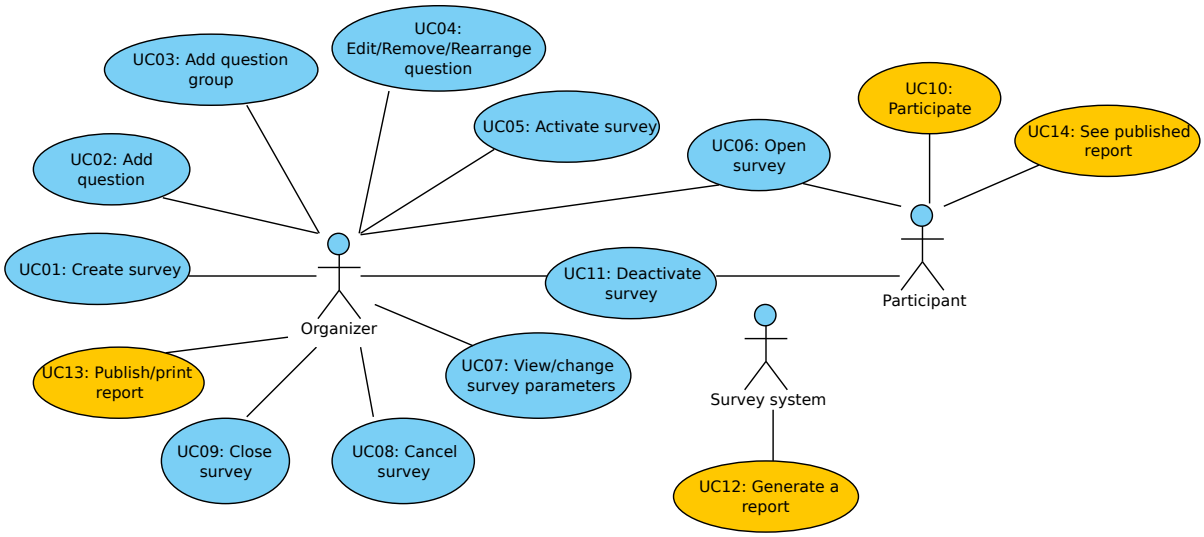

Figure 7: Survey system use cases

5.1 Design survey

This section describes the survey creation process in Figure 8.

Use cases that are describing this activity diagram:

- UC01: Create survey
- UC02: Add question
- UC03: Add question group
- UC04: Edit/remove/rearrange question(s).

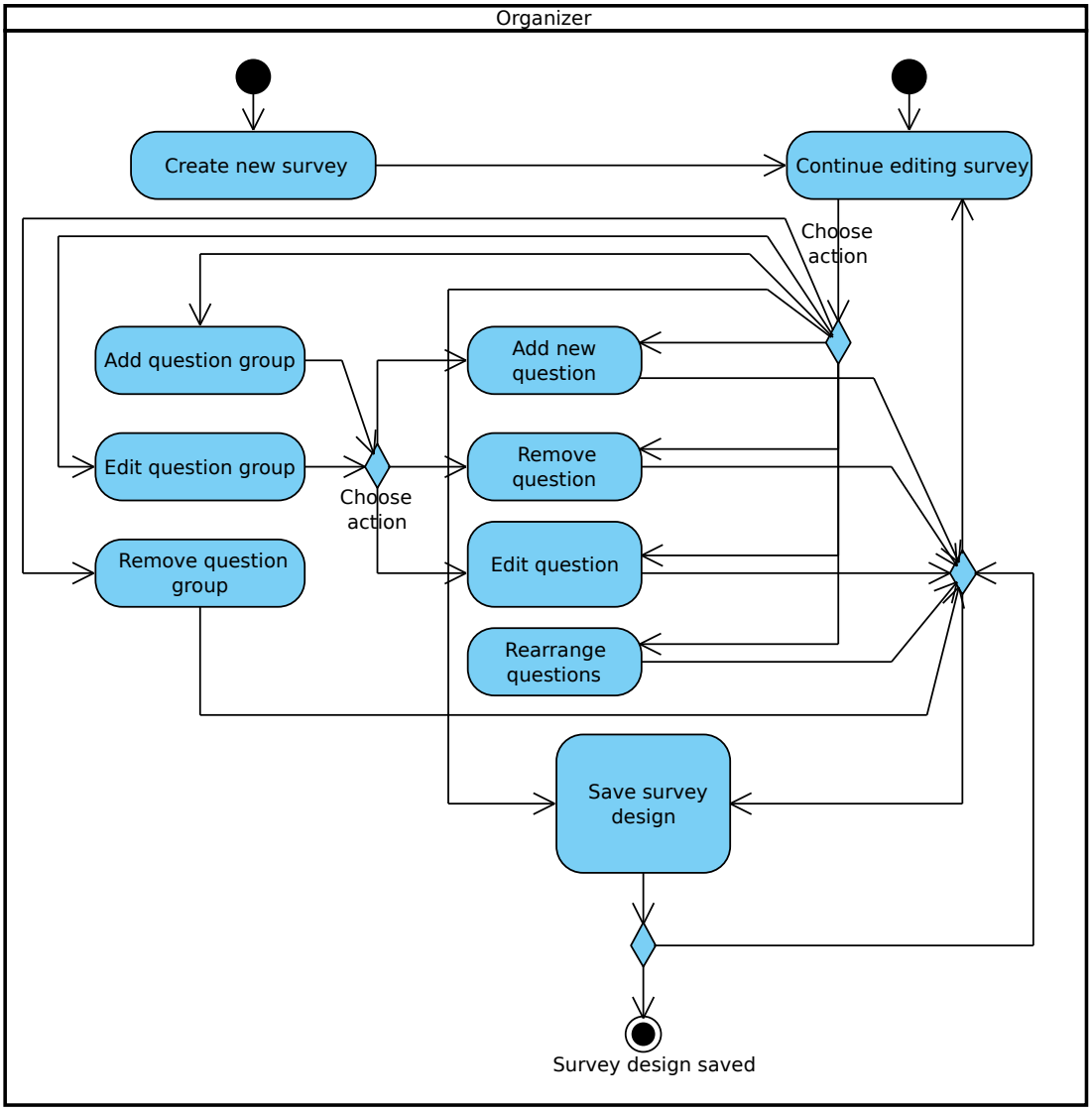

Figure 8: Activity diagram for survey design

UC01: Create survey

Brief description: Organizer will create a survey by creating a new blank questionnaire.
Involved actors: Organizer
Precondition: Organizer has logged in to the survey system.
Trigger: Organizer decided to start creating a new survey. *(Function createSurvey)*
Postcondition: Organizer has created a survey and the survey is displayed to organizer. *(Survey will have status 'Draft')*
Security requirements: Access control
Main success scenario
1. Organizer chooses to create a survey.
2. System displays the survey creation form.
3. Organizer inserts a name for the survey and selects creating a new survey.
4. System creates a new survey and displays the created survey to organizer.

Extensions
3a. Organizer cancels creating the survey.

3a.1. System cancels the process.

UC02: Add question

Brief description: Organizer will add new question to the survey by a choosing question type, adding question text and the answer list. The updated survey will be displayed to organizer.
Involved actors: Organizer
Precondition: Organizer has created a survey and opened it for editing. System has displayed the survey. *(Survey status is 'Draft')*
Trigger: Organizer decided to add new question to the questionnaire. *(Function amendQuestionnaire)*
Postcondition: Organizer has added question and answer options to the questionnaire and updated survey is displayed to organizer. *(Survey status will not change)*
Security requirements: Access control
Main success scenario
1. Organizer chooses to add a new question.
2. System displays an editable question form.
3. Organizer inserts the question text and a list of answers, chooses the type of the question and submits the question.
4. System adds the question to the questionnaire and displays updated survey.
Extensions
3a. Organizer inserts the question text and/or answer list and decides to leave the page.
    3a.1. System displays a warning about changes not saved.
    3a.2. Organizer decides to submit the question.
    3a.3. The use case continues from main success scenario step 4.
3b. Organizer inserts the question text and/or answer list and decides to leave the page.
    3b.1. System displays warning about changes not saved.
    3b.2. Organizer decides to leave the page.
    3b.3. System displays the page organizer navigated to.

UC03: Add question group

Brief description: Organizer will add new a question group to the survey. Editable question group form will be displayed to the organizer.
Involved actors: Organizer
Precondition: Organizer has created a survey and opened it for editing. System has displayed the survey. *(Survey status is 'Draft')*
Trigger: Organizer decided to add new question group to the questionnaire. *(Function amendQuestionnaire)*
Postcondition: Organizer has added question group to the questionnaire and editable question group form is displayed to organizer. *(Survey status will not change)*
Security requirements: Access control
Main success scenario
1. Organizer chooses to add a new question group.
2. System displays editable question group form.

3. Organizer continues with UC02: Add question and/or UC04: Edit/remove/rearrange question(s).
Extensions
3a. Organizer decides to leave the page.
    3a.1. System displays warning about changes not saved.
    3a.2. Organizer decides to leave the page.
    3a.3. System displays the page organizer navigated to.

UC04: Edit/remove/rearrange question(s)

Brief description: Organizer will edit the question or remove the question from questionnaire or rearrange questions. Updated survey will be displayed to organizer.
Involved actors: Organizer
Precondition: System has displayed the survey. If the organizer wants to edit the question or remove the question from the questionnaire, then the questionnaire should include at least one question. If organizer wants to rearrange questions then the questionnaire should include at least two questions. *(Survey status is 'Draft')*
Trigger: Organizer decided to edit question or remove question from questionnaire or rearrange questions. *(Function amendQuestionnaire)*
Postcondition: Organizer has edited question or removed question from the questionnaire or rearranged questions. The updated survey is displayed to organizer. *(Survey status will not change)*
Security requirements: Access control
Main success scenario
1. Organizer chooses to edit a question.
2. System presents the editable question form.
3. Organizer changes the question text and/or answer list and submits the question.
4. System saves the changes and displays the updated survey.
Extensions
1a. Organizer chooses to remove a question.
    1a.1. System asks for confirmation.
    1a.2. Organizer confirms the decision.
    1a.3. System removes the question and displays updated survey.
1b. Organizer chooses to rearrange questions and does it by using the drag-and-drop method.
    1b.1. System saves modified questionnaire and displays updated survey.

5.2 Activate survey

This section describes survey activating process in Figure 9.

Use cases that are describing this activity diagram:

- UC05: Activate survey
- UC06: Open survey

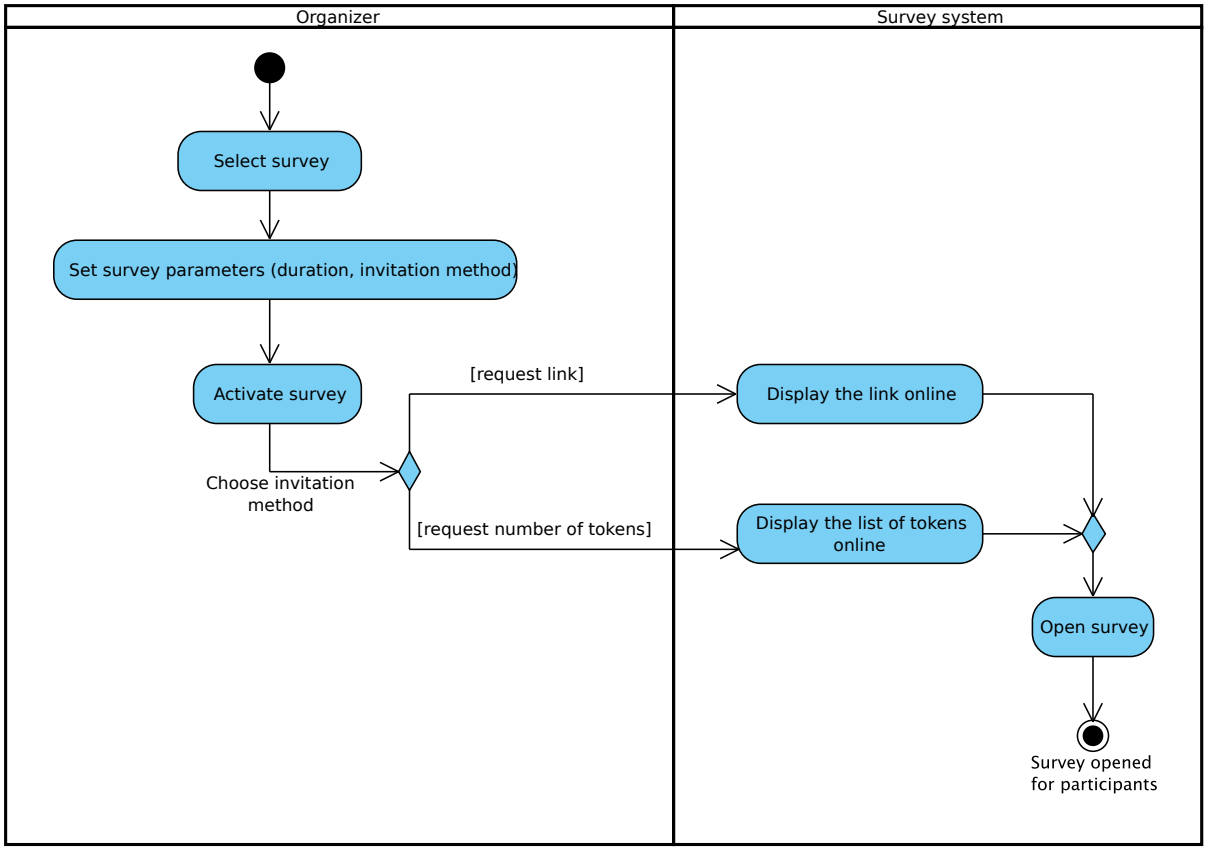

Figure 9: Activity diagram for survey activation

UC05: Activate survey

Brief description: Organizer has decided to activate the survey. Organizer will set the survey parameters (duration, invitation method). Survey parameters will be saved and survey will be activated.
Involved actors: Organizer
Precondition: Organizer has created a survey. At least one question is added to the survey. *(Survey status is 'Draft')*
Trigger: Organizer decided to activate the survey. *(Function activateSurvey)*
Postcondition: Survey parameters are saved. *(Survey will have status 'Active')*
Security requirements: Access control
Main success scenario
1. Organizer chooses to activate the survey.
2. System displays an editable form of survey parameters.
3. Organizer inserts survey parameters.
4. System asks for confirmation about activating the survey.
5. Organizer confirms his decision.
6. If the chosen invitation method is 'request link', the system activates the survey and displays the link to organizer online.

Extensions
3a. Organizer cancels activating the survey.
    3a.1. System displays warning about changes not saved.
    3a.2. Organizer decides to leave the page.

3a.3. System cancels the process and displays the page organizer navigated to.
4a. System has detected validation errors and displays an error message.
4a.1. The use case continues from the main success scenario step 3.
6a. If the chosen invitation method is 'request number of tokens', the system activates the survey and displays the list of tokens to the organizer online.

UC06: Open survey

Brief description: The survey start time has come and the first participant has navigated to the survey using the link. System will open the survey for participants.
Involved actors: Organizer, participant
Precondition: Organizer has activated the survey and survey start time has come. *(Survey status is 'Active')*
Trigger: Organizer or participant has navigated to the survey or organizer has required to see the data. *(Function openSurvey [startTimePassed AND surveyAccessed])*
Postcondition: Survey is open to participants. *(Survey will have status 'Open')*
Security requirements: Access control
Main success scenario
1. Survey starting time has come and organizer has navigated to the survey or required to see the data.
2. System opens survey for participants.
Extensions
1a. The survey starting time has come and the first participant has navigated to the survey using the link.
1a.1. System opens survey for participants.

5.3 Run survey

This section describes the functionality available to organizer while survey is activated in Figure 10.

Use cases that are describing this activity diagram:

- UC07: Change survey parameters/check survey status
- UC08: Cancel survey
- UC09: Close survey
- UC11: Deactivate survey.

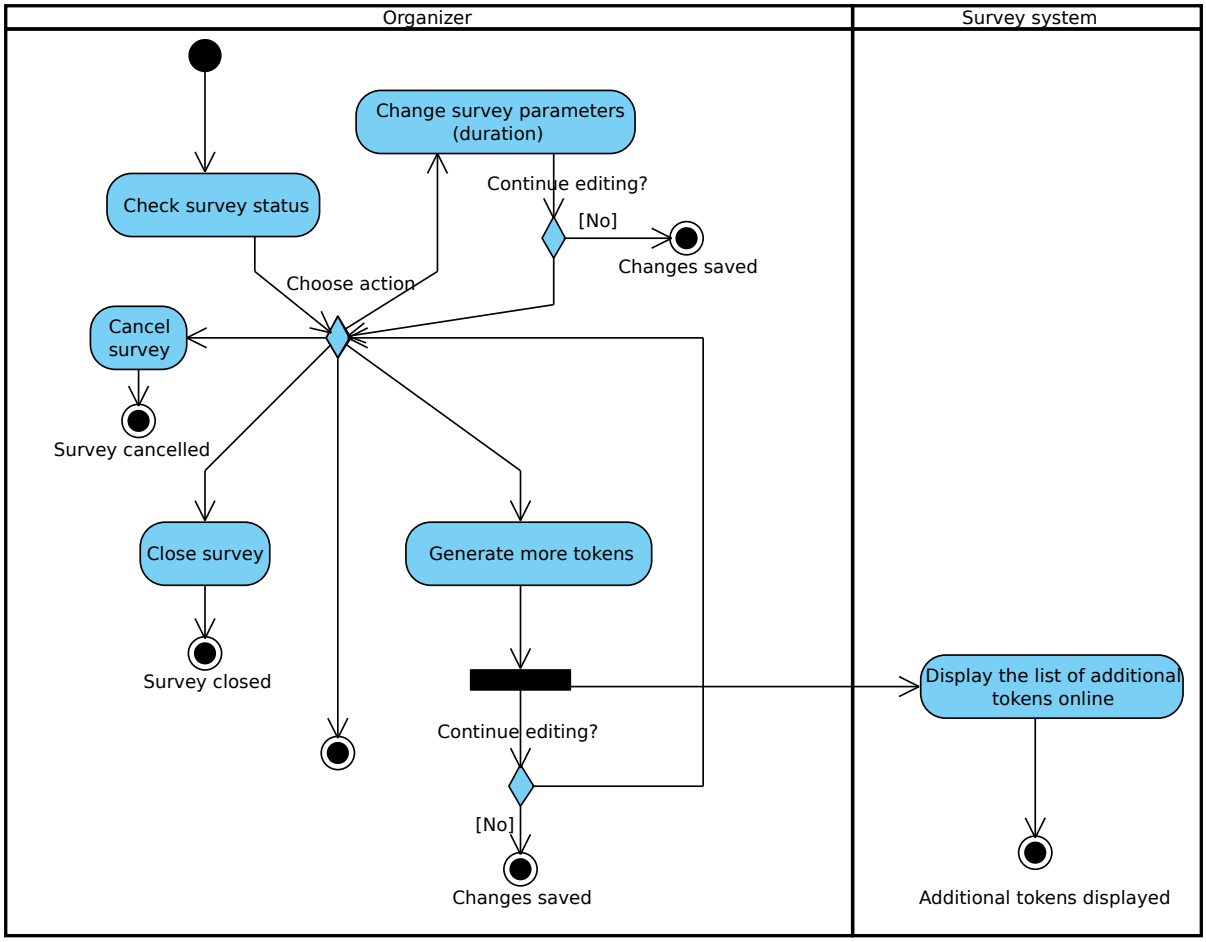

Figure 10: Activity diagram for running a survey

UC07: Change survey parameters/check survey status

Brief description: Organizer has activated the survey. While the survey is running, organizer can check survey status, change duration time or generate more tokens.
Involved actors: Organizer
Precondition: Organizer has activated the survey. *(Survey status is 'Active', 'Open' or 'RespondingDeadlinePassed')*
Trigger: Organizer decided to change survey parameters or check survey status. *(Function amendSurveyParameters)*
Postcondition: Survey parameters are changed or survey status is displayed to organizer. *(Survey status will not change)*
Security requirements: Access control
Main success scenario
1. Organizer chooses to change the survey parameters.
2. System displays an editable form.
3. Organizer changes the survey parameters.
4. System saves changes. If organizer has requested to generate more tokens system displays the list of additional tokens to organizer online.
Extensions
1a. Organizer chooses to check the survey status.

1a.1. System displays the survey status.
3a. Organizer decides to leave the page.
3a.1. System displays a warning about changes not saved.
3a.2. Organizer decides to leave the page.
3a.3. System displays the page that the organizer navigated to.
4a. System has detected validation errors and displays an error message.
4a.1. The use case continues from main success scenario step 3.

UC08: Cancel survey

Brief description: Organizer has created the survey, organizer can cancel the survey. Survey will be cancelled and all participant answers data will be deleted.
Involved actors: Organizer
Precondition: Organizer has created the survey. *(Survey status is 'Draft', 'Active' or 'Open')*
Trigger: Organizer decided to cancel the survey. *(Function cancelSurvey)*
Postcondition: Survey is cancelled. *(Survey will have status 'Cancelled')*
Security requirements: Access control
Main success scenario
1. Organizer chooses to cancel the survey.
2. System asks for confirmation about cancelling the survey.
3. Organizer confirms his/her decision.
4. System cancels the survey and deletes all participant answers data.
Extensions
3a. Organizer cancels cancelling the survey.
3a.1. System cancels the process.

UC09: Close survey

Brief description: Organizer has activated the survey. While the survey is running, organizer can close the survey before the duration has ended. Survey will be closed for participants and all data will be preserved.
Involved actors: Organizer
Precondition: Organizer has activated the survey. *(Survey status is 'Open' or 'RespondingDeadlinePassed')*
Trigger: Organizer decided to close the survey. *(Function closeSurvey)*
Postcondition: Survey is closed. *(Survey will have status 'Closed')*
Security requirements: Access control
Main success scenario
1. Organizer chooses to close the survey.
2. System asks for confirmation about closing the survey.
3. Organizer confirms his decision.
4. System closes the survey.
Extensions
3a. Organizer cancels closing the survey.

3a.1. System cancels the process.

UC11: Deactivate survey

Brief description: The responding deadline for the survey has passed. System will deactivate the survey.
Involved actors: Participant, organizer
Precondition: Survey is open for participants and survey responding deadline has passed. *(Survey status is 'Open')*
Trigger: Organizer or participant has navigated to the survey or organizer has required to see the data. *(Function deactivateSurvey [deadlinePassed AND surveyAccessed])*
Postcondition: Survey is not open for participants *(Survey will have status. 'RespondingDeadlinePassed')*
Security requirements: Access control (if tokens are used or organizer is interacting with the system)
Main success scenario
1. Survey responding deadline has passed and organizer has navigated to the survey or required to see the data.
2. System deactivates the survey.
Extensions
1a. Survey responding deadline has passed and participant has navigated to the survey.
1a.1. System deactivates the survey.

5.4 Participate

This section describes the process of participating in a survey in Figure 11.

Use cases that are describing this activity diagram:

- UC10: Participate

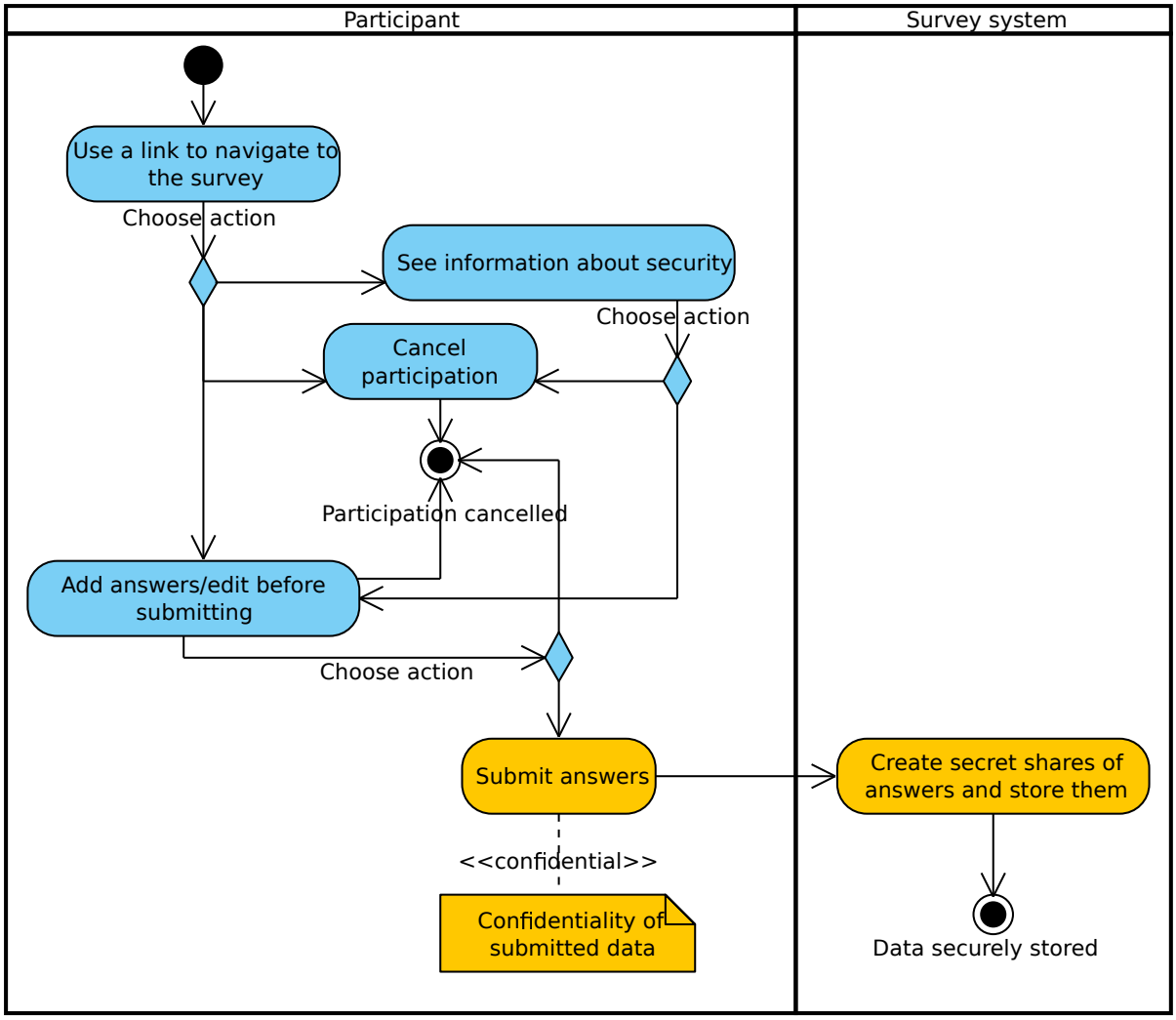

Figure 11: Activity diagram for participating in a survey

UC10: Participate

Brief description: Organizer has activated the survey and the survey is open for participants. Participant has received an invitation and navigated to first page of the survey. Participant can fill in the questionnaire as well as check out the information about the security guarantees. Participant will submit the answers and system will store the secret shares of the answers in three different servers.
Involved actors: Participant
Precondition: Organizer has activated the survey. Participant has received an invitation. The survey is open for participants. *(Survey status is 'Open')*
Trigger: Participant decided to start answering the survey by navigating to first page of the survey. *(Function participate)*
Postcondition: The secret shares of the survey data are securely stored in three different servers. *(Survey status will not change)*
Security requirements: Access control (if tokens are used), data confidentiality
Main success scenario
1. Participant chooses to participate in the survey and uses the link to start answering the questionnaire.

2. System displays the first page of the survey.
3. Participant answers the questions and submits the answers.
4. System stores the secret shares of the answers in three different servers.

Extensions

3a. Participant requires to see information about security guarantees.
    3a.1. System displays information about security guarantees.
    3a.2. Participant is satisfied with the information and decides to continue answering questions.
    3a.3. The use case continues from main success scenario step 2.
3b. Participant requires to see information about security guarantees.
    3b.1. System displays information about security guarantees.
    3b.2. Participant is not satisfied with the information and decides to terminate.
    3b.3. System will not store any answers.
3c. Participant terminates from the survey.
    3c.1. System will not store any answers.

5.5 Analyze

This section describes the process of analysing the outcome of the survey in Figure 12.

Use cases that are describing this activity diagram:

- UC12: Generate a report
- UC13: Publish/print report
- UC14: See published report

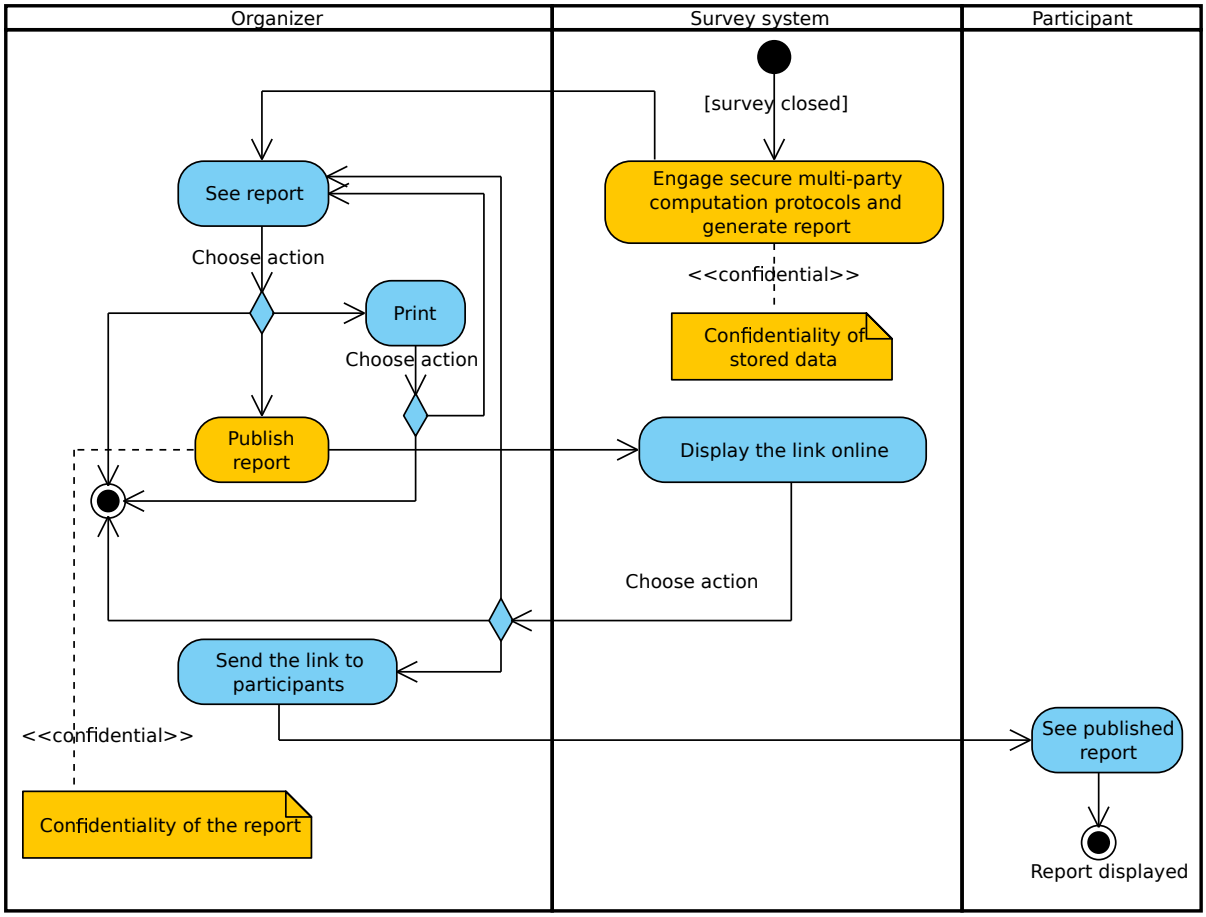

Figure 12: Activity diagram for analyzing the data

UC12: Generate a report

Brief description: The survey has been closed. System will compute the results using secure multi-party computation, generate the report and make the report available to organizer. Organizer can only see computed results and not the initial answers of participants.
Involved actors: Survey system
Precondition: The survey has been closed. *(Survey status is 'Closed')*
Trigger: The state of the survey shifts to 'Closed'. *(Function generateReport)*
Postcondition: Report is generated and available to organizer. *(Survey will have status 'ResultsAvailable')*
Security requirements: Access control, data confidentiality
Main success scenario
1. The state of the survey shifts to 'Closed'.
2. System shifts the survey to the state 'Computing', computes the results by using secure multi-party computation.
3. System generates the report, shifts the survey to the state 'ResultsAvailable' and makes the report available to organizer.
Extensions
2a. Less than five participants have submitted answers.

2a.1. System cancels the process and displays an explaining text.

UC13: Publish/print report

Brief description: System has generated a report and organizer can choose to print and/or publish it. When published, the report will be accessible for participants only through a special link (this functionality has not been completely implemented in the prototype by the time of submitting the thesis).
Involved actors: Organizer
Precondition: The report has been generated. *(Survey status is 'ResultsAvailable')*
Trigger: Organizer decided to print/publish the report. *(Function publishReport)*
Postcondition: Report is published/ready for printing. *(Survey will have status 'ResultsPublished')*
Security requirements: Access control, data confidentiality
Main success scenario
1. Organizer chooses to print the report.
2. System displays the print view of the report to organizer.
3. Organizer prints the report.
Extensions
1a. Organizer chooses to publish the report and the invitation method was 'request link' or 'request number of tokens'.
    1a.1. System publishes the report and generates a link for published report. System displays the link to organizer online.
    1a.2. Organizer sends the link to participants.

UC14: See published report

Brief description: Organizer has published the report. Participant has received the link and navigated to the report view (this functionality has not been completely implemented in the prototype by the time of submitting the thesis).
Involved actors: Participant
Precondition: Organizer has published the report. Participant has received the link. *(Survey status is 'ResultsPublished')*
Trigger: Participant decided to view the report by navigating to the report view.
Postcondition: The report is displayed to participant. *(Survey status will not change)*
Security requirements: Access control, data confidentiality
Main success scenario
1. Participant chooses to view the report and uses the link to display it.
2. System displays the report.

5.6 State machine diagram

To further provide an overview of the system the state machine is provided in Figure 13.

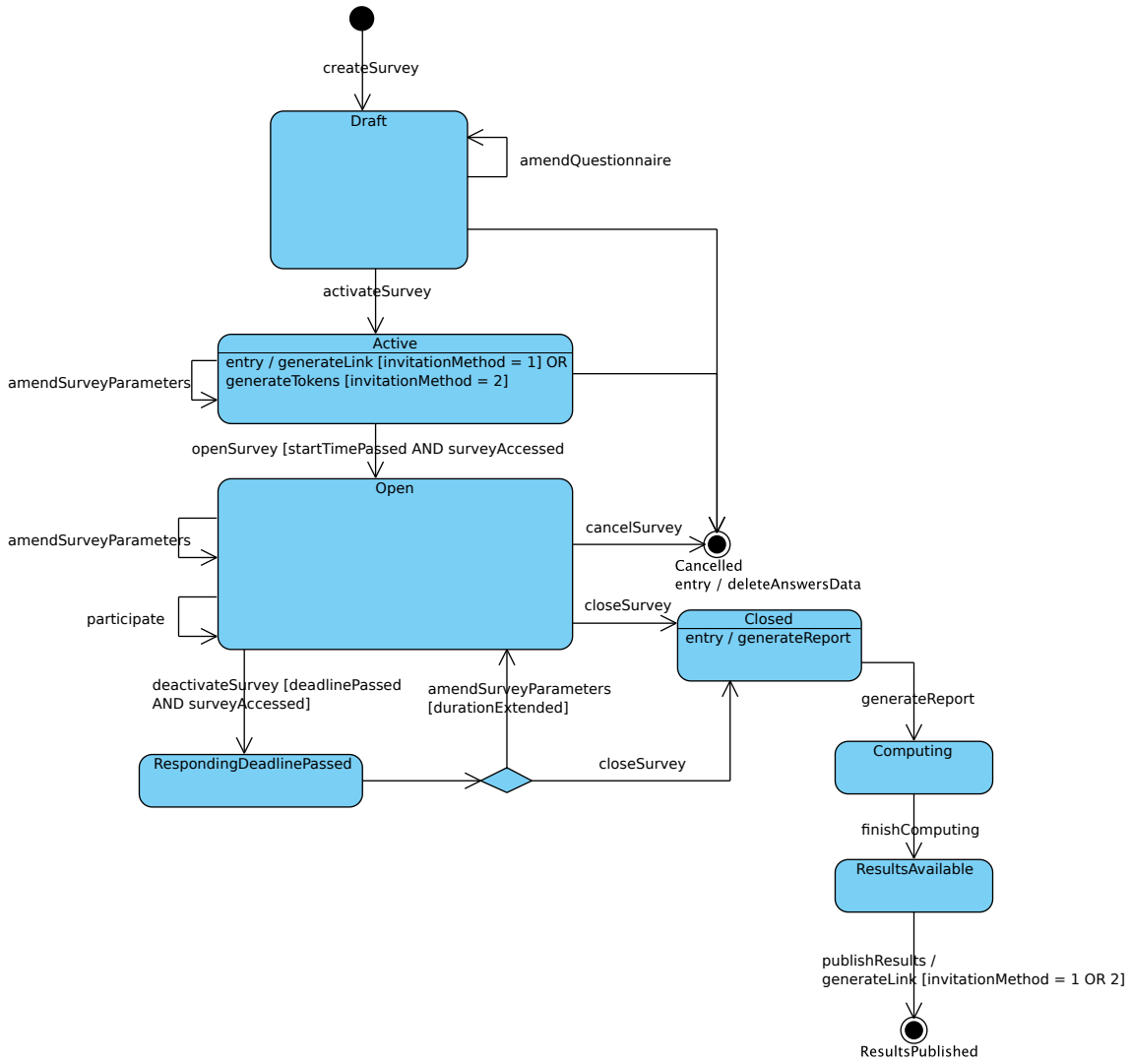

Figure 13: Survey system state machine diagram

6 Design of the secure survey system

The secure survey system consists of two main components: client software and server software (see Figures 14 and 15). The client software is a web application for a general web browser. Browser based web applications are built in HTML, CSS and JavaScript. The web application is deployed to the user over the internet. When requested by the user, the browser contacts the server, downloads the web application, and runs it. This process eliminates any need for installing custom software on the user's computer and can easily complete within a second without affecting the user experience.

The server side software of SHAREMIND is distributed (duplicated) across three servers. Each server software instance runs on different cloud providers' virtual machines using Linux as an operating system and is maintained by different PRACTICE partners so one single partner nor one single cloud service provider can gain access to all secret shares of participants' answers and hereby compromise the confidentiality of the secret-shared values.

Public data is stored in plain text form and in case of private data, the shares of data are stored. The secret shares of participants' answers can not be reconstructed by the server host, any of the involved cloud service providers nor by the survey organizer. To be able to reconstruct secret shares of the answers, a single party should know all the shares of data. As none of them control all of the servers, they do not have all of the shares and therefore can not reconstruct the answer data.

The secret shares of participants' answers are moved to the server over a secure channel and thereby also protected against eavesdropping. The result of a survey can be reconstructed only by the survey organizer and the organizer may then allow others to reconstruct the results e.g. the participants or the general public. Only organizer can reconstruct survey analysis results, because server checks if organizer is authenticated and only then sends secret shares of the results to the organizer browser, which then reconstructs the results and shows them to the organizer.

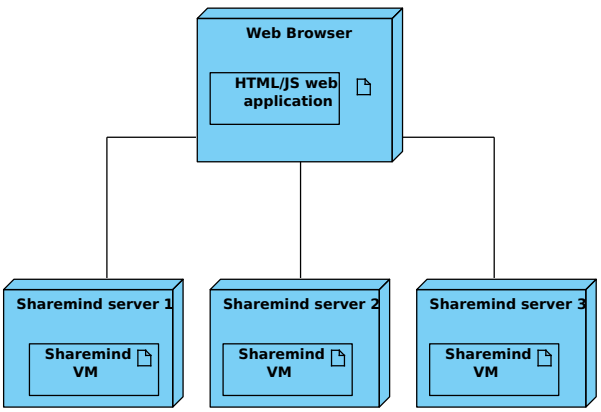

Figure 14: The simplified deployment diagram of secure survey system

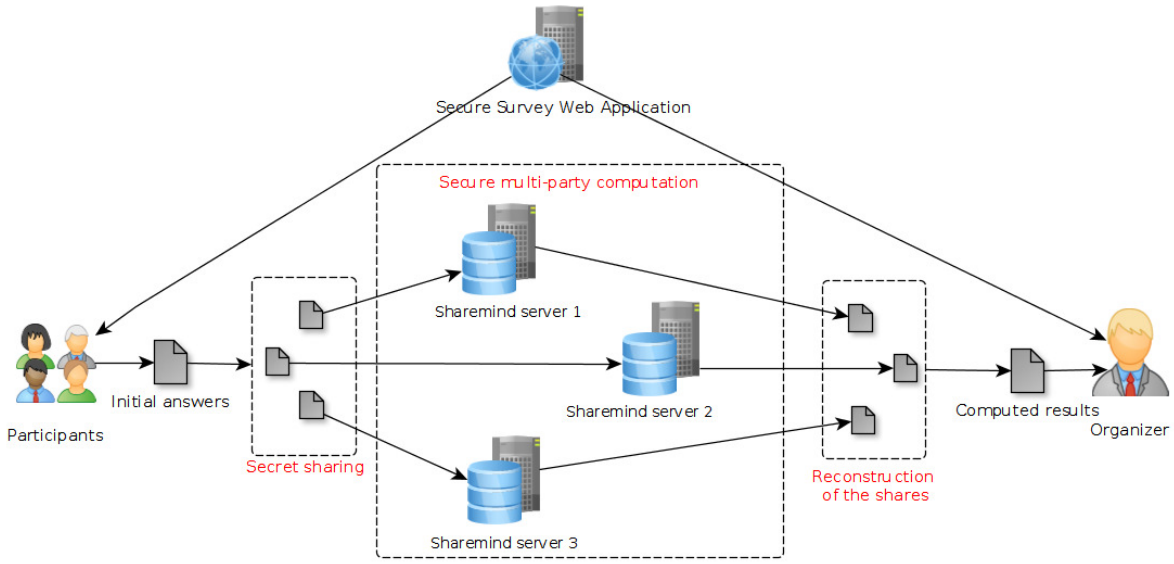

Figure 15: The deployment model of secure survey system

7 Aspects of analyzing and deploying SMC-based systems

7.1 Introduction

The author of this paper has an experience in analyzing the business processes of two prototypes that are based on secure multi-party computation. The first one is the secure survey system prototype that is described in this master's thesis. The second one is the prototype of the Estonian Tax and Customs Board (MTA) value-added tax (VAT) fraud detection system [16].

According to the law that came into force in December 2014, Estonian companies must report transactions with each partner with whom the monthly sum of transactions exceeds 1000 euros. MTA detects tax fraud by analyzing the financial records of the suspect company and its partners to determine the actual taxable sum. The system runs risk analysis algorithms to find cases where a company has incorrectly declared transactions (or not declared them at all).

The Estonian Traders association was concerned about the security of the "super database" of financial transactions. Moreover, as MTA has a significant employee turnover, a tax officer could copy the database to support his or her future business ambitions in the private sector. Agreeing with the significant privacy risk, the President of Estonia blocked the legislation at first [9].

Examining the problem, we saw the secure multi-party computation as a solution. We designed a prototype with a reduced scope that conducts the risk analysis while the transactions are in the encrypted domain. Only the risk scores will be published to the tax officer who can then request the detailed records for the at-risk companies. This protects the information and rights of the honest taxpayers, as their declaration annexes remain encrypted during the whole process.

This section describes how the business processes and the deployment of the systems that are based on secure multi-party computation are different from the usual solutions.

7.2 Case 1: a privacy-preserving survey system

The secure survey system is an example of a privacy-preserving cloud service that is based on secure multi-party computation. The business processes of the system are not remarkably different from the processes that would be implemented in a non-SMC system. Users' behaviour in the system is not affected by the fact that the data is secret-shared at the source and SMC is used for computing the results.

The difference is noticeable when we take a look at the architecture of the system. Instead of storing data in one server, the system is distributed across three servers and secret shares of the data are each stored in different server. This requires that each server is hosted by a different company who has capability to host the computing instance and also has no intentions to collude with other parties.

The positive side of the distributed system is that it significantly rises the level of the privacy of the system. The answer data of participants remains confidential at all times even during report preparation. The new trust model gives the participants a guarantee that their sensitive data remains confidential and encourages participants to give truthful answers.

7.3 Case 2: a privacy-preserving VAT fraud detection system

The Estonian Tax and Customs Board (MTA) value-added tax (VAT) fraud detection system is an example of a privacy-preserving software that processes highly confidential data. Similarly to the secure survey prototype, the business processes of the VAT fraud detection system are not different from the processes that are implemented in a non-SMC system and there are no changes in users' behaviour.

The changes had to be made in the architecture of the system and data was stored in three servers instead of one. However, the risk analysts of MTA were concerned with the required transparency. Today, MTA can perform risk analyses autonomously so that unauthorized parties have no knowledge of the kind of algorithms that are used. SMC would change this and MTA would have to agree on the algorithms with other hosts.

Based on the calculations from MTA, 80 000 companies will upload 50 million economic transactions every month. We estimated that our prototype can process one month of Estonian economy in ten days, using about 20 000 euros worth of hardware. This was met with some concern, as today, MTA processes VAT returns in three days. Nevertheless, MTA agreed to consider SMC as a technology for confidential data collection and analysis in future application, inspired by our prediction that the cost of deploying SMC will be further reduced in the coming years.

7.4 Key differences in the deployment of SMC-based software compared to the usual solutions

This subsection describes the key differences in the deployment of SMC-based software compared to the non-SMC solutions.

- Using SMC based on secret sharing, the architecture of the system has to be changed. Instead of storing data in one server the system is distributed across multiple servers to maintain the security guarantees of secret sharing.
- The distributed nature of the system increases the complexity of the maintenance and, also, administrative costs.
- Each server has to be hosted by a different company who has capability to host the computing instance and, also, has no intentions to collude with other parties.
- In the systems that deal with a large amount of data the computing process is significantly slower for a practical use with today's SMC techniques.
- If the analysis algorithms are confidential it is hard to hide the algorithms from other computing parties as all hosts would need to agree to the risk analysis algorithms.

- The use of SMC significantly rises the level of the confidentiality of the data. It protects the data from unauthorized access by both insiders and outside attackers.

8 Conclusion

In this thesis we describe the analysis and design of the secure survey prototype that is based on secure multi-party computation. The main goal was to implement a survey system that allows researchers to collect sensitive data without compromising the privacy of participants. This paper introduces the business processes as well as the design of the prototype of the secure survey system.

To communicate the business processes of the system, the author has modelled the activity diagrams using Unified Modelling Language (UML). In addition, there are the use cases and the state machine diagram used, to give a more detailed description of the system. The secure survey system prototype is designed to run on two different secure multi-party computation engines: Sharemind and Fresco/SPDZ. This master's thesis concentrates on the implementation using the Sharemind 3 framework.

Using SMC allows us to build applications that are used for collecting and analyzing confidential data. The privacy preservation techniques, like secret-sharing of the data, distributing the system across three different servers, setting the minimum number of answers required for each question and public data checks, are used to accomplish the level of security that corresponds to the actors' needs.

The design of the secure survey prototype is also introduced in this thesis. The system consists of two main components: client software and server software. The client software is a web application for a general web browser. The server side software of Sharemind is distributed across three servers. Each server software instance runs on different cloud providers' virtual machines and is maintained by different PRACTICE partners.

In addition to the secure survey prototype, the author of this thesis has an experience in analyzing the business processes of another prototype that is based on secure multi-party computation – the prototype of the Estonian Tax and Customs Board (MTA) value-added tax (VAT) fraud detection system. Based on the experience, there are the key differences in the deployment of SMC-based software compared to the usual solutions introduced in Section 7.

For example, the architecture of the system has to be changed irrespective of whether we deal with small or large amount of data. In addition, when the system is required to handle a large amount of data, there might be problems with the performance of computing the results. However, the use of SMC significantly rises the level of the confidentiality of the data and protects the data from unauthorized access by both insiders and outside attackers.

8.1 Future work

As a future work, the functionality of the secure survey system could be improved, for example, by adding functionalities like template system (possibility to copy existing survey or question), e-mail template design system, user creation system, etc. For building trust with participants, there could be added a possibility for organizer to upload his/her

company's logo and contact information on the survey template. For participants there could be added a possibility to preview how the final report would look like.

The prototype will probably be used to run real surveys in the near future – an employee satisfaction survey of city government of Tartu, a survey on "correct use of standards" within H2020 project, surveys within the Alexandra Institute and PRACTICE survey on security issues.

8.2 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement №609611 (PRACTICE). The author would like to acknowledge all the teams that participated in developing the secure survey system prototype – Partisia, Cybernetica AS and Alexandra Institute.

References

[1] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Published online at http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:31995L0046, 1995. Last accessed May 2015.

[2] HIPAA: Health Insurance Portablity and Accountability Act. Published online at http://health.state.tn.us/hipaa/, 1996. Last accessed May 2015.

[3] United States – European Union Safe Harbor Framework. Published online at http://export.gov/safeharbor, 2000. Last accessed May 2015.

[4] U.S.-EU Safe Harbor List. Published online at https://safeharbor.export.gov/list.aspx, 2000. Last accessed May 2015.

[5] Federal Information Security Management Act. Published online at http://www.dhs.gov/federal-information-security-management-act-fisma, 2002. Last accessed May 2015.

[6] Isikuandmete kaitse seadus. RT I 2007, 24, 127. Published online at https://www.riigiteataja.ee/akt/112072014051, 2007. Last accessed May 2015.

[7] International standard ISO/IEC 29100:2011. Published online at http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html, 2011. Last accessed May 2015.

[8] SOC 2 Report - Trust Services Principles. Published online at https://www.ssae-16.com/soc-2/, 2011. Last accessed May 2015.

[9] Otsus 348. Käibemaksuseaduse ja raamatupidamise seaduse muutmise seaduse väljakuulutamata jätmine. Published online at http://president.ee/et/ametitegevus/otsused/9726-2013-12-18-11-37-04/index.html, 2013. Last accessed May 2015.

[10] TRUSTe. Published online at https://www.truste.com/, 2013. Last accessed May 2015.

[11] Adi Shamir. How to Share a Secret. Communications of the ACM, 22(11):612-613, 1979.

[12] Andrew Chi-Chih Yao. Protocols for Secure Computations (Extended Abstract). In Proceedings of FOCS'82, pages 160-164. IEEE, 1982.

[13] CASRO. Code of Standards and Ethics for Market, Opinion, and Social Research. Published online at http://c.ymcdn.com/sites/www.casro.org/resource/resmgr/Media/Code_of_-Standards_and_Ethics.pdf, 2014. Last accessed May 2015.

[14] Dan Bogdanov. Sharemind: programmable secure computations with practical applications. PhD thesis, University of Tartu, 2013.

[15] Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht. Turvalist ühisarvutust kasutava käibemaksudeklaratsioonide riskianalüüsi süsteemi prototüüp. Cybernetica teadusaruanne, T-4-22, 2014.

[16] Dan Bogdanov, Marko Jõemets, Sander Siim, Meril Vaht. How the Estonian Tax and Customs Board Evaluated a Tax Fraud Detection System Based on Secure Multiparty Computation. Financial Cryptography and Data Security - 19th International Conference, 2015, San Juan, Puerto Rico, 2015.

[17] Dan Bogdanov, Peeter Laud, Jaak Randmets. Domain-Polymorphic Programming of Privacy-Preserving Applications. Ninth Workshop on Programming Languages and Analysis for Security (PLAS 2014), 2014.

[18] Dan Bogdanov, Riivo Talviste, Jan Willemson. Deploying secure multi-party computation for financial data analysis (Short Paper). In Angelos Keromytis, editor, Financial Cryptography and Data Security, LNCS 7397, pp. 57-64. Springer, 2012.

[19] Dan Meir. The Seven Stages of Effective Survey Research. American Marketing Association, 2002.

[20] Florian Hahn, Daniel Demmler, Hiva Mahmoodi, Thomas Schneider, Peter S. Nordholt, Michael Stausholm, Roman Jagomägis, Matthias Schunter, Meilof Veeningen, Niels de Vreede, Antonio Zilli, Johannes U Jensen, Kurt Nielsen. Deployment Models and Trust Analysis for Secure Computation Services and Applications, 2013.

[21] George Robert Blakley. Safeguarding cryptographic keys. In Proceedings of the 1979 AFIPS National Computer Conference, pages 313-317, Monval, NJ, USA, 1979. AFIPS Press.

[22] Google. Privaatsuseeskirjad. Published online at http://www.google.com/intl/et/policies/privacy/, Last updated: 2014. Last accessed May 2015.

[23] OECD. Good Practices in Survey Design Step-by-Step. Measuring Regulatory Performance: A Practitioner's Guide to Perception Surveys, 2013.

[24] Peter Bogetoft, Dan Lund Christensen, Ivan Damgård, Martin Geisler, Thomas Jakobsen, Mikkel Krøigaard, Janus Dam Nielsen, Jesper Buus Nielsen, Kurt Nielsen, Jakob Pagter, Michael Schwartzbach, Tomas Toft. Secure Multiparty Computation Goes Live. In Roger Dingledine and Philippe Golle, editors, Financial Cryptography and Data Security, volume 5628 of Lecture Notes in Computer Science, pages 325-343. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-642-03549-4 20.

[25] Peter S. Nordholt, Roman Jagomägis, Marko Jõemets, Reimo Rebane, Meril Vaht, Johannes Ulfkjær Jensen, Kurt Nielsen. Secure Survey Prototype - a supplementary report, 2015.

[26] Qualtrics. Security Statement. Published online at http://www.qualtrics.com/security-statement/. Last accessed May 2015.

[27] QuestionPro. Feature details: Respondent Anonymity Assurance. Published online at http://www.questionpro.com/features/respondent-anonymity-assurance.html. Last accessed May 2015.

[28] QuestionPro. QuestionPro Security Overview. Published online at http://www.questionpro.com/images/qphome/QuestionPro-Security-Policy-Procedures.pdf. Last accessed May 2015.

[29] Sherrie Mersdorf. Market Research Process: 6 Steps to Project Success. Published online at http://survey.cvent.com/blog/cvent-web-surveys-blog/market-research-process-6-steps-to-project-success, 2009. Last accessed May 2015.

[30] SurveyGizmo. Privacy Policy. Published online at http://www.surveygizmo.com/privacy/, Last updated: 2014. Last accessed May 2015.

[31] SurveyMonkey. Security Statement. Published online at https://www.surveymonkey.com/mp/policy/security/, Last updated: 2013. Last accessed May 2015.

Non-exclusive licence to reproduce thesis and make thesis public

I, Meril Vaht (date of birth: 27th of January 1990),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

The Analysis and Design of a Privacy-Preserving Survey System,

supervised by Dan Bogdanov.

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 21.05.2015