ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Data Sharing in Participatory Social Sensing

Master Thesis

Ramapriya Sridharan

September 3, 2016

Advisors: Prof. Dr. Dirk Helbing, Dr. Pournaras Evangelos
Department of Computational Social Sciences , ETH Zürich

Contents

Abstract

Data from citizens needs to be collected and analyzed to create or improve current services in society. Data collected from them, in general, reveals information about their behavior and choices. In addition, it can also reveal sensitive information, that they might not be comfortable with. To preserve the privacy of citizens is where data privacy comes into play. There are various methods to maintain data privacy and different levels of privacy to maintain. The higher the privacy level, the more concealed the data is. Given the choice, citizens would generally choose the highest privacy level. At times, less concealed data is needed while solving problems that need data with less errors. To help citizens reduce the level of privacy of the data when needed, different kinds of incentives can be used, such as monetary incentives. From a fixed budget on the demand side, rewards(incentives) are handed out to citizens to incite them to give less privatized data, yet maintaining a minimum level of privacy. The goal of the Thesis is to understand the social dynamics of privacy and information sharing. Existing data can be used or data can be collected for the purpose of the analysis.

Chapter 1

Introduction

Chapter 2

Related Work

Chapter 3

Computational Model

3.1 Introduction

Why the model is needed, add some previous paper about incentives and what we do differently Our aim is to create a computational model that is able to collect useful data about the influence of incentives on mobile data sharing. (Quote some studies that have done similar studies with no data incentives). In the model, what we try to do is first create a user profile by asking the user some preliminary questions. Then, we proceed to assign each sensor data request with a maximum achievable credit using the formulated user profile. The idea of the model is that it assigns requests where the user least desires to share mobile sensor data with higher incentive costs. Respectively, the data requests where the user desires to share data more are assigned lower incentive costs. This permits us to see whether incentives do indeed make a difference in data sharing, since assigning more credit to data requests where the user would anyway give mobile sensor data is futile to our goals. The model aims to collect useful data that examine the relationship between incentives and mobile data sharing, mainly when the user least desires to share data.

3.2 Model Intricacies

The sections below explain the various building blocks of the computational model. The Figure 3.1 provides an overview of how the model works. To begin with the model, each user is asked to enter various non-intrusive personal information that can help in analysing the user's behavior. For example, this can consist of the age, gender, country of residence and employment.

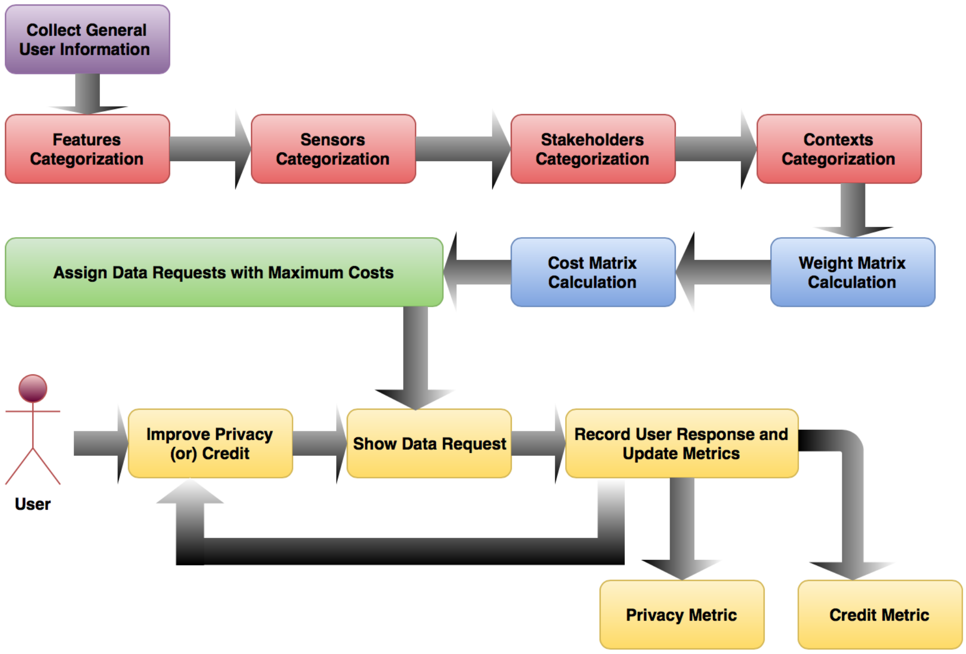


Figure 3.1: Computational Model Flow Chart

3.2.1 Categorization of the Features

After we are done recording the user's personal information, we go into the categorization of the features. In this model, features consist of Sensors, Stakeholders and Contexts. The Sensors consist of the sensors in the mobile phone which the user's can trade. Let the category assigned to the Sensors be represented by S. The Stakeholders consist of any entity that can request the user for mobile sensor data. Let the category assigned to the Sensors be represented by DC. The Contexts consist of the purpose for which a Stakeholder would like to obtain the user's mobile sensor data. Let the category assigned to the Sensors be represented by C. Categorization of the features means that the user places each of the features into predefined categories, and 2 or more features can be placed into the same category. The user is asked to categorize each feature in the available cat number of categories. The first category indicates that the feature does not contribute much to the data sharing decision. Respectively, the category cat indicates that the feature contributes a lot to the data sharing decision. The categories are linearly scaled and equally spaced. The reason categorization was chosen is as to not rule out the possibility to that two features may be considered equally important in the data sharing decision and this may be missed by ranking the features. Once the user has categorized the Sensors, Stakeholders and the Contexts, the weights of each feature in the data sharing decision is calculated. The category feature Sensors has been placed into be represented by S, the category feature Stakeholders has been placed into be represented by DC and the category feature Context has been placed into be represented by C. Hence the respective weights $weights_S$, $weight_{DC}$ and $weight_C$ are calculated as follows :

$$weights_S = \frac{S}{S + DC + C} \tag{3.1}$$

$$weight_{DC} = \frac{DC}{S + DC + C} \tag{3.2}$$

$$weight_C = \frac{C}{S + DC + C} \tag{3.3}$$

3.2.2 Categorization of the Sub-Features

Once the features have been categorized and their weights calculated, the sub-features need to be categorized. In this model, sub-features consist of the various types of Sensors available on the mobile phone, the various types Stakeholders that request mobile data from users and the different types of Contexts for which mobile data is requested. In other words, sub-features are the different kinds of features that appear during data request to the user. The following are examples of sub-features for each feature :

- Sensors : Accelerometer, Battery and Gyroscope
- Stakeholders : Company, Non-Governmental Organization and Government.
- Contexts : Education, Entertainment and Navigation.

For each of the features, the respective sub-features are assigned a unique identifier ranging from one to the length of sub features. Now for each of the features, the respective sub-features need to be in turn categorized in a similar fashion to section 3.2.1. Each sub-feature can be placed in the available *cat* categories. The first category indicates that the user finds the sub-feature very non privacy intrusive. This means that the user would not be worried trading data for this sub-feature. The last category indicates that the user finds this sub-feature very privacy intrusive. This in turn means that the user would be reluctant giving data for this sub-feature. All the categories in between are linearly scaled and equally spaced. The user then places for each feature, the respective sub-features in the *cat* available categories according to the perceived intrusion level. A conceptual diagram is shown in figure 3.2. For the sub-features of Sensors, categories they are placed in are represented by S_i, where S is the feature Sensors and i is the id of the

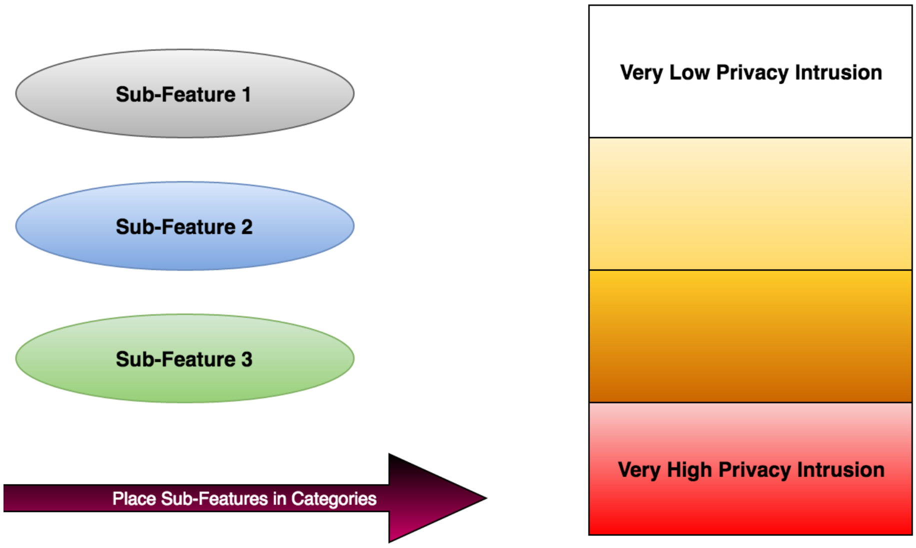


Figure 3.2: Categorizing Sub-Features according to the perceived Intrusion Level

sub-feature. Similarly, categories assigned to sub features of features Stakeholders and Contexts respectively are represented by DC_j and C_k, where j and k are the id's of the sub-features categorized.

3.2.3 Weight Matrix Calculation

Each data request to the user consists of the 3 features in them. Each of those features has a number of sub-features that can appear in turns in a data request, that is in a factorial form. Let $count(feature)$ be function that gives the number of sub-features given a feature. The total number of data requests is :

$$N_{DR} = count(Sensors) * count(Stakeholders) * count(Contexts) \qquad (3.4)$$

Let WM be a matrix with dimensions $count(Sensors)xcount(Stakeholders)xcount(Contexts)$. We call this the weight matrix. The cell $WM_{i,j,k}$ represents the data request which involves the Sensor's sub-feature with identifier i, Stakeholder's sub-feature with identifier j, and the Context's sub-feature with identifier k. That is, each cell of WM represents a data request to the user. The aim of the weight matrix is to use the information collected from the user profiling to assign various weights to each data requests. Intuitively, the process examines the data requests where the user is least likely to trade data and assigns higher weights to those data requests. This process can be seen in section 3.3 in more detail. As mentioned before, each cell of the matrix WM represents the weight of a data request with a unique Sensors sub-feature i, Stakeholders sub-feature j and Contexts sub-feature k. To calculate the weight of a data request :

$$WM_{i,j,k} = (S * S_i) + (DC * DC_j) + (C * C_k) \tag{3.5}$$

Applying this formula to every cell gives the weight matrix WM.

3.2.4 Cost Matrix Calculation

Now that the weights for every data request has been calculated, we need to calculate the exact amount of money the users can receive for a data request. Let CM be the cost matrix with dimensions $count(Sensors)xcount(Stakeholders)xcount(Contexts)$. Let us assume to have a budget of B for a day, where B can be in an actual currency or any sorts of virtual credits. For now, the budget will be referred to as credits. Each cell of the cost matrix will represent the amount of credits allocated for a particular data request for one day. To begin with, we calculate the sum of all the cells of the weight matrix WM:

$$sum(WM) = \sum_{i=1}^{count(Sensors)} \sum_{j=1}^{count(Stakeholders)} \sum_{k=1}^{count(Contexts)} wm_{i,j,k} \tag{3.6}$$

where the function $sum(matrix)$ gives the sum of a matrix, in this case the weight matrix. Let $CM_{i,j,k}$ represent the credit allocated for the data request which involves the Sensor's sub-feature with identifier i, Stakeholder's sub-feature with identifier j, and the Context's sub-feature with identifier k. To calculate one cell of the cost matrix :

$$CM_{i,j,k} = \frac{WM_{i,j,k} * B}{sum(WM)} \tag{3.7}$$

Doing this for every cell of CM, the whole cost matrix can be calculated. Now, we have the credits allocated per day for every data request.

3.2.5 Cost and Privacy Metrics

Every data request now has an associated cost. This is the maximum cost that a user can obtain for that data request. The Cost metric is the total amount of credits the user has obtained for one day. Similarly, the Privacy metric is the amount of privacy percentage the user has maintained. That is, it intuitively quantifies the amount of data the user has refused to share hence implying privacy. The Cost and Privacy are inversely proportional to each other, in the sense that when the Cost goes up and Privacy goes down and vice versa. For each data request, the user can choose how much data is to be shared, from the maximum amount of data to no data at all. Each option corresponds to a summarization level explained in detail in section 3.2.7. The cost assignment to each option is linearly scaled according to the cost assigned to each data request. Let us assume there are options for a data request ranging from 1 to m (numeric options), where 1 corresponds to where the user gives all the data requested and m to where the user chooses not give any data at all. Therefore there are a total of m options for a data request. While assigning costs there are two scenarios:

- Assigning option costs without a participation cost.
- Assigning option costs inclusive of a participation cost.

Let us examine the first scenario. Let us assume that we are calculating the option costs for data request with Sensors sub-feature i, stakeholders sub-feature j and contexts sub-feature k. Let us calculate the assigned cost for option number h of this data request:

$$cost_h = \frac{CM_{i,j,k} * (m - h)}{m - 1} \tag{3.8}$$

Applying this formula by replacing h by the options from 1 to m gives the cost the user receives for each option. Similarly, if you would like to assign a participation cost to each option, it would mean that even tough the user does not share data, they still receive some money for answering the data request. This concept can be implemented to ensure user participation. (Quote some paper with participation of users in PSS). Let x be a fraction of the total budget B that is dedicated for user participation. Using a geometric progression with $a = 1$ and $r = \sqrt[(m-1)]{x}$, we can calculate the fraction of the cost $frac_h$ an option numbered h gets:

$$frac_h = a * r^{h-1} \tag{3.9}$$

Now that we know the fraction of the cost option f can be assigned, to get the cost $cost_h$ of option h for the data request with Sensors sub-feature i, stakeholders sub-feature j and contexts sub-feature k :

$$cost_h = frac_h * CM_{i,j,k} \tag{3.10}$$

This assigns costs to each option, taking into consideration a participation cost that the user gets even if data is not shared for that data request.

Privacy percentage pri_h is linearly scaled between the first to the mth option between 0 and 100 as follows:

$$pri_h = \frac{(h - 1) * 100}{m - 1} \tag{3.11}$$

The total cost and privacy is the arithmetic average of all the costs and privacy obtained from every answered data request. If a data request is left unanswered, maximum privacy and minimum cost is assumed.

3.2.6 Improving the Metrics

Before the user answers a question, it is useful to know what the user interest lies in. Would the user like to improve the privacy metric, or would the user would like to increase the credit revenue. In addition, if we know what the user is looking to improve, we can retrieve the question that can improve the that particular metric the most. For example if the user wishes to improve his privacy further, we look at the questions where the user has given the most amount of data. We then put forth this question to answer, which indicating all the options that can improve the privacy. Similarly, if the user chooses to obtain more credit, the question where the user has given least amount of data is retrieved. Options that can improve the user credit are also indicated.

3.2.7 Summarization of Collected Data

As mentioned before, each data request can have options m number of options the user can choose from. These options range from 1, which indicates that the user would like to give all his data, to option number m, which indicates when the user does not want to give any data to this data request. Even tough all data is encrypted these days, it is still not enough as encryptions might be cracked. Summarization is a privacy algorithms that aggregates data to provide less information than in its original form. The higher the summarization level gives less data than than in its original form. The lower the summarization level gives data closer to its original form. In this model, data is collected for a period of 24 hours every y seconds for every data request.If the data is summarized, according to the option chosen, the data is collected either every y seconds or lesser.

Data is collected for the whole day, and at the end of the day according to the option chosen by the user, it is summarized. Summarization can be linearly assigned to each option starting with the highest privacy corresponding to highest summarization level , that is no data sharing to the lowest summarization level, that is no summarization at all. An example of assigning the summatization level $summ_h$ for option h can be the following :

$$summ_h = y * h \text{where} h \neq m \tag{3.12}$$

This gives the frequency of sensor data collection for every option of a data request.

3.3 Analysis of the Model

In this section, we take a scenario of the computational model and show how exactly the model works. In particular, the focus is on how the model varies the weights to questions according to the user input.

3.3.1 Setup

the sensors, stakeholders, and contexts and other special parameters such as number of options and all To explain the model using examples, we take into consideration the following sub-features for each feature:

1. Sensors
   a) Accelerometer -1

   b) Noise -2

   c) Location -3
2. Stakeholders
   a) Corporation -1

   b) Government -2

   c) education -3
3. Contexts
   a) Navigation -1

   b) Environment -2

   c) Social Media -3

The numbers indicated next to the sub-features is the sub-feature identifier. This uniquely identifies a sub-feature within a feature category. Each user will receive an amount of

$$count(Sensors) * count(Stakeholders) * count(Contexts) = 27$$

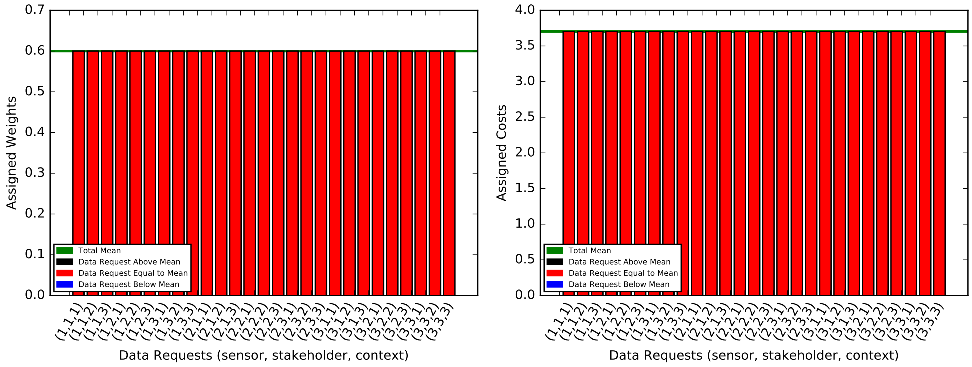
data requests in total. Each data request has five privacy options ranging from one to five. the option one indicates the users would like to to trade all their data, and option five indicates the users refuse to share data their for this data request. Additionally, it is assumed that the core phase has a Budget $B = 100$ per day. The input to the model are the user choices during the categorization of the features and sub-features.

3.3.2 Results

In this section, five user scenarios will be introduced and explained in order to explore the properties of the weight and cost matrices. First, we will begin by introducing the way the user has categorized the features and sub-features. This will be followed by an explanation of the generated matrices.

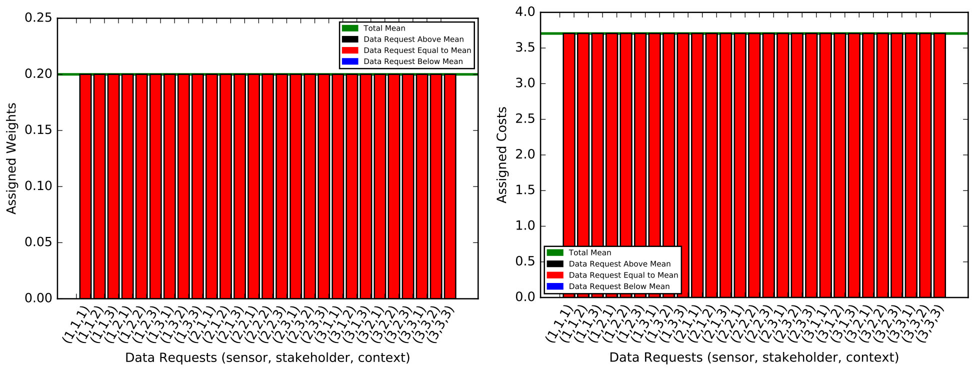
Scenario 1 and 2

In this scenario we assume that the user has entered the f



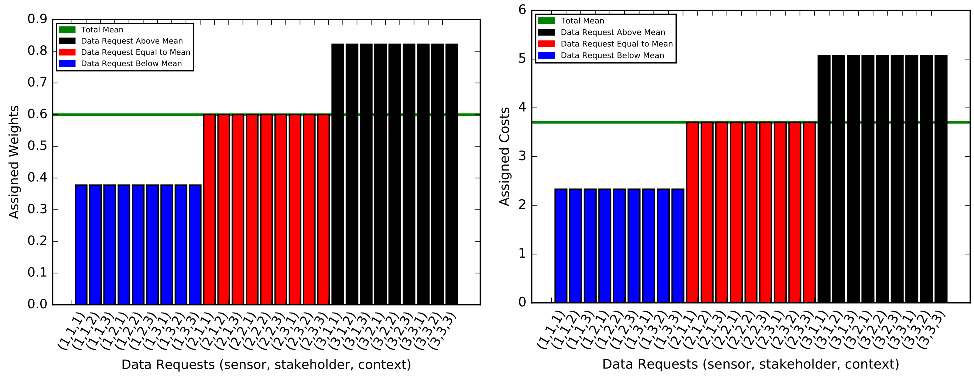
(a) Values of the Weight Matrix (b) Values of the Cost Matrix

Figure 3.3: Examining Scenario 1



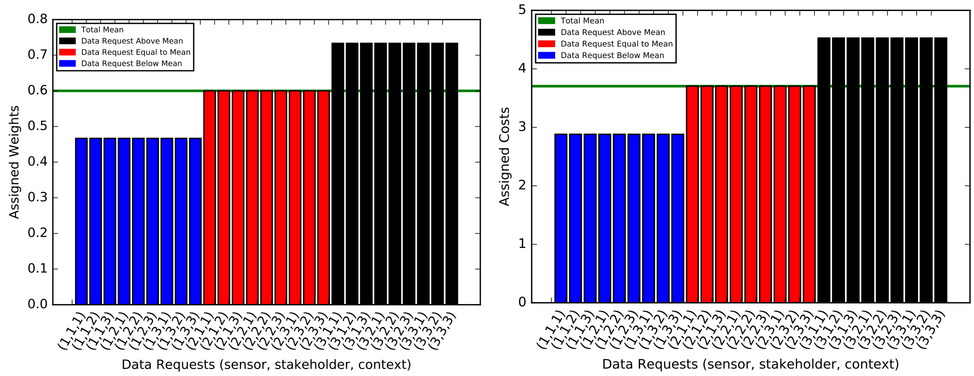
(a) Values of the Weight Matrix (b) Values of the Cost Matrix

Figure 3.4: Examining Scenario 1



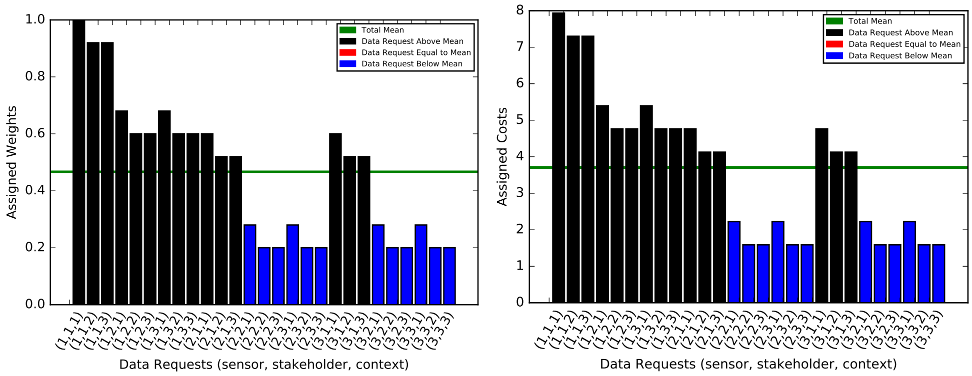
(a) Values of the Weight Matrix (b) Values of the Cost Matrix

Figure 3.5: Examining Scenario 2



(a) Values of the Weight Matrix (b) Values of the Cost Matrix

Figure 3.6: Examining Scenario 3



(a) Values of the Weight Matrix

(b) Values of the Cost Matrix

Figure 3.7: Examining Scenario 4

Chapter 4

Experiment Methodology

4.1 Preparatory Phase

4.1.1 Pre-Survey

(For each sensor,dc and context show graphs from the pre-survey why each of them was chosen)

The pre-survey is the survey created that runs before the deployment of the actual social experiment. This survey was made in order to study the perception of users on the three features to be studied:

1. Sensors - The mobile mobile sensor data shared
2. Stakeholders - The entity to whom mobile sensor data is shared
3. Context - The purpose for which mobile sensor data is shared

From the above meantioned features, there were a lot of sub-features to choose from. Increasing the number of sub-features in the experiment in turn increases the number of data requests posed to the user. In addition, we wanted to gain insight into the perception of users on the three features. Hence the survey was prepared to understand the above(link to appendix). Additionally, it can help us redesign some of the aspect of the experiment based on the ambiguities. Participants pool consist of both people who are aware and unaware of data privacy and sensors. Participants were not paid for filling out the survey.

4.1.2 Sub-Features

194 pre-survey entries have been filled out by participants. Using this data, a fewer number of sub-features were chosen.

4.1.3 Privacy Options

Each data request accompagnied with privacy options ranging from 1 to 5. Option 1 indicates that the user would like to share his raw data without any sort of summarization. 5 indicated that the user would not like to share his data for this data request. The options in between have linearly scaled summarization levels assigned to them. For more information refer to 3.2.7.

4.1.4 Question Structure

A data request is when a stakeholder asks a user for mobile sensor data for a particular context or purpose. Each data request to the user is posed in the form of a question with the following template :

"PleasechoosetheamountofXsensortypedatasharedwithYstakeholderforuseinaZcontextapp"

where Sensors X can be :

1. Accelerometer
2. Noise
3. Location
4. Light

where Stakeholders Y can be:

1. Corporation
2. Educational Institution
3. Non Governmental Organization
4. Government

and where Context Z can be:

1. Environment
2. Health/Fitness
3. Navigation
4. Social Networking

In total this makes 64 data requests to the user.

4.1.5 Budget and Experiment Duration

The experiment is set to run for a total of two days, excluding the time taken for the entry phase and exit phase. The budget set for the core phase of the experiment is $B = 35$ Chf and is excluding the cost of participation in the entry and exit phase. Participants are paid 10 Chf for coming to the Entry Phase, and 15 Chf for participating in it. Similarly for the Exit Phase, participants are given 10 Chf for showing up, and 5 Chf for participating in it. Out of the budget B, $\frac{1}{7}$ is given away for the participation of the users.

4.2 Entry Phase

explanation of screen shots

4.2.1 Collecting General User Information

As the figure **??** shows, the user is asked to answer some personal non-intrusive questions. The following is asked from the user:

1. Gender
2. Employment Status
3. Education Level
4. Year of birth
5. Country where user has lived most of his life
6. How many time a day do you check your Mobile phone per day.
7. Kind of applications the user has in the mobile phone.

The user may go back and re-answer the questions, but once the last questions answer is confirmed the data is sent to the server and hence cannot be changed. Users cannot navigate to the next page without filling out all the questions.

4.2.2 Categorization of Features

As described in chapter 3, the users need to categorize the features Sensors, Stakeholders and Contexts. As shown in figure **??**, the features are indicated followed by 5 options of privacy ranging from "very low privacy intrusion" to "very privacy high privacy intrusion". the option "very low privacy intrusion" means that the feature does not affect your mobile sensor data sharing decision, whereas "very privacy high privacy intrusion" refers to a feature that really affect the sharing of mobile sensor data. Users need to click on the drop down menu to choose one of the privacy intrusion options. All options are compulsory, a no default option is provided. Users cannot navigate to the next page without filling out all the questions.

4.2.3 Categorization of Sub-Features

For each of the features categorized in the previous sub-section, each of the sub-features need to be categorized in a similar fashion. Again, the privacy options range from very low privacy intrusion" to "very privacy high privacy intrusion" like in section 4.2.2 . The user is first presented with the categorization of Sensors sub-features. Below each sensor is a drop down menu where the user can choose how much each of the sensors would affect the mobile sensor data sharing. Once all the sensors have been associated with a intrusion level, the user can click the green submit button and is directed to the next page where the sub-features of stakeholders need to be in turn categorized in a similar fashion. Each stakeholder type has a drop down menu each where the user can again classify how much each of them affect data sharing. Once the user has finished filling up the intrusion level for stakeholders, the user can click the green submit button and is directed to the next page. In this page, the user will need to categorize how much each of the Context's sub-features affect mobile sensor data sharing. Each context has a drop down menu below where the user can rate each context from low privacy intrusion" to "very privacy high privacy intrusion". Once this has been done the user can click on the submit button.

The user will be redirected to the next page only if all the drop down boxes have been filled out. All questions are compulsory there is not default choice.

4.2.4 Answering Questions with No Incentives

After the categorization questions are answered and recorded, the user will be presented with 64 questions. Each question is a data request to the user. The user can choose from the available 5 privacy options mentioned in section 4.1.3. The higher the privacy, the less data is given away for that request. Users can change the answers to the data request until the green submit button on top of the options is clicked. After that the answer is locked. No indications of credit gain or privacy improvements are indicated in this round. The "i" button at the bottom of the screen is click-able. This takes the user to the FairDataShare portal. The user needs to then click on the data generator section of the website where the user can signup with his:

1. Username
2. Password
3. Email
4. Unique Identifier

The unique identifier can be located at the bottom of the page. If it is clicked the user can select it and then copy and paste it on the website. The users can use this website to see all the data collected from them for all the mobile sensors. More details about the FairDataShare portal refer section 4.5.

In the task-bar, the user can see the bidding day number and how many questions have been answered from the total available. Once all the questions have been answered, the user goes to the core phase of the experiment.

4.3 Core Phase

Once the entry phase is done, the user is presented with the screen shown in figure **??**. The first presented screen after the entry phase is over is what is called the "improvement screen". The button numbered 1 represents "improve privacy" and the button numbered 2 represents "improve credit" respectively. The items numbered 3 and 4 represent the privacy and credit obtained by the user respectively. Privacy is measured in terms of the percentage of mobile sensor data not traded to the stakeholders. Credit is measured in terms of the currency Swiss Francs. The button numbered 5 is the button that takes the user to the FairDataShare portal. The user can login into the portal after a minimum of 24 hours after the start of the core phase to see the data that has been collected and shared with the stakeholders. The item numbered 6 is the unique identifier of the user. This can be selected and copied by long pressing the unique identifier for one second. The item numbered 7 is the round number which indicates the number of times the user has answered all the data requests. The item numbered 8 is the number of questions the user has answered in the current round.

There are a total of 64 data requests, hence when all the 64 have been answered, the number of questions answered is reset and the number of round answered increases by one. This indicates all the data requests that have been answered and how many are left unanswered.

Each question will have 5 options to choose from ranging from maximum data sharing to least data sharing.

From the starting time of the core phase till 24 hours later marks one bidding day. Once 24 hours is over, another bidding day starts where the privacy and credit metrics are reset. The day number in the task bar is incremented by one. The user has to answer all the data requests again for this new bidding day. Previous responses to data requests are not carried over to the next day. If a data requests is not answered, it is considered that the user does not want to trade mobile sensor data for that request. Additionally, each data request carries a participation fee, this is irrespective of the amount of mobile sensor data shared, by not participating in a data request the user foregoes this credit gain.

The core phase goes on for a period of 48 hours.

4.3.1 Improve Privacy or Credit

The improvement screen is where the user can choose whether he would like to improve the privacy or the credit that has been obtained. The elements of this screen have been explained in the previous section 4.3. The improve credit button should be chosen if the user is interested in maximizing the credit already obtained further. This uses algorithm that uses the previous user answers to put forth a data request that can increase the credit to the maximum. The credit improvement button is represented by the number 4 in figure **??**. Similarly, the improve privacy button is used to further improve the privacy that has been obtained. This puts forth a data request that can further increase the user privacy. Then again, the ultimate change in the privacy or credit metrics depends on the option chosen by the user for the data request. The privacy improvement button is represented by the number 3 in figure **??**.

For example, if a user chooses to improve the privacy, then clicks on improve privacy button and gets a data request, but still chooses option 1 for that data request, this may not improve his privacy but decrease it. This is because option 1 indicates that the user trades all the data for this request. Trading all data gives the user more credit, but decreases the privacy metric.

Similarly, if a user chooses to improve the credit, then clicks on the improve credit button and gets a data request. If the user then chooses the option 5 which indicates that no data is traded for this request, this is counters the initial desire to improve the credit obtained. Trading no data increases one's privacy, but does not increase the credit to the maximum.

An actual improvement in the chosen metric depends on the chosen improvement button chosen and the choice of the appropriate option for that data request.

4.3.2 Answering Questions with Incentives

After choosing a metric to improve, the a screen is presented as shown in figure **??**. This screen is called the "bidding screen". This screen is very similar to the screen presented in the entry phase, except that the user is aware of the amount of privacy and credit obtained. Additionally, the user can see information about how the privacy and credit will increase or decrease for each data request, according to the chosen option.

The items numbered 9 are the improvement in privacy for each possible option of the current data request. The items numbered 10 are the improvements in credit for each possible options of the current question. Once the user decides on which options to choose according to how much data wants to be traded, the user can click on the radio option as explained in section **??** and then click again on the green submit button to confirm the answer.

Once the green button has been clicked on, answers cannot be changed. The user has the possibility to go back to the improve screen from the bidding screen using the back button. Using the back button in the improve screen leads the user out of the application.

Additionally, for every question there is an orange recommendation box that appears around the radio button options showcased at item numbered 12. This gives an indication to the user as to which options can improve the privacy or the credit compared to the previous time the user has answered this data request. For example, if the user has previously answered option 3 to a data request and has clicked on improve credit, the system puts an orange box around options 1,2 and 3. Similarly, if the user clicked on improve privacy button, the system would recommend the options 3,4 and 5. Two examples of this are provided in figures **??**.

4.4 Exit Phase

After the end of the core phase, the participants are asked to fill up a survey on their experience of the experiment. Some questions are about the rewards received, the privacy and credit metrics, design of the application, how the experiment was conducted, etc. The survey is linked to the user using the unique identifier assigned in the application. Once the survey is filled, the users receive their money for the entry phase, core phase and exit phase together, but only if they did not have their phones switched off throughout the experiment and participated in the core phase. This is done by checking the data collected on the server.

4.5 FairDataShare Web Portal

The FairDataShare portal is a website where users can view the data collected from them during the core phase of the experiment. Below is an explanation of how users ans stakeholders can view mobile sensor data.

4.5.1 Data Generator's Portal

Users first register as data generators as indicated in section 4.2.4. Once the users are registered, they can come back to the portal after 24 hours period or later to view their mobile sensor data collected in the server. The data portal login page is shown in figure **??**. Since the users are already registered from the mobile phone in the entry phase, they can go to the portal from their computers and this time login instead of register. Users should enter their:

1. Username

2. Password

Once the users do that, they will be redirected to the data collection page shown in figure **??** with the following options in the task-bar :

1. Accelerometer
2. Light
3. Noise
4. Location

Users can choose the sensor from the task-bar whose data they want to see by clicking on it. An example is given in figure **??** of what kind of data can be seen. The data is displayed in the following fashion :

1. Timestamp
2. Bidding day
3. Sensor Values

In the experiment, day number one is the entry phase, the core phase is day number two and three.

4.5.2 Stakeholder's Portal

For a stakeholder to view data, they need to register in the portal shown in figure **??** by clicking register. Once that is done, the page shown in figure **??** is shown asking for :

1. Company Name
2. Email
3. Stakeholder Category
4. Company Website

Stakeholder category is the type the stakeholder comes under such as :

1. Corporation
2. Educational Institution
3. Government
4. Non-Governmental Organization

After this, the stakeholder can click on the register button. He then gets access to data that is being collected according to the user's wishes. The stakeholder can then choose the bidding day, an anonymous user and a sensor and get data he is permitted to see. Figure **??** shows screenshots where the stakeholder can choose what data he wants and the page with the user data.

Chapter 5

Explanation of the Mobile Application

5.1 The Building Blocks

explain with diagram what interacts with what

5.2 The Mobile Application

5.2.1 Local Storage

5.2.2 Alarms

Going to the Next Data Sharing Day

Notifications

5.2.3 Privacy and Credit Improvement

5.2.4 Recommendations

how privacy credit is imporved

5.2.5 Recording User Choices

5.2.6 Sensor Data Collection and Summarization

5.2.7 Server Synchronization

5.3 The Server

5.3.1 Kinvey Data Storage

Kinvey is a mobile backend as a service which provides a platform for mobile phones to link applications at a backend cloud storage. This backend has been used to store data and for some business logic implementations.

Security

Table Store

All the data collected form the user's phones is stored in Kinvey Data Store's collections. Data is segregated in to the appropriate collections. Data is stored in the collections is done so in the following form :

1. Data collected in the form of radio buttons on the phone are stored as integers.
2. Data collected as check-box entries on the phone each have a column in the collection with entries zero or one.
3. Data collected as drop down lists on the phone are stored either by integer position in the list or by the entry name in the list itself.

This way of storing data applies to all the collections explained below. Each record of every collection is with a unique user identifier and with the timestamp whenever necessary. This is done to identify data that belongs to the same user accross different collections. The timestamp is collected in order to examine temporal relationships. The first collection is the GetUserInformation collection and is used to store all the basic non intrusive user information collected in the entry phase. The collection is shown in figure **??**. Next, is the UserResponse collection which is used to store all the responses of the users to data requests in the entry phase and the core phase. The collection is shown in figure **??**. Collection AccelerometerStore, LightStore, NoiseStore and LocationStore are used to store the mobile sensor data collected from the user at the end of a bidding day after local summarization on the mobile phone. Theses collections are shown in figures **?? ?? ?? ??**. The StorePoints collection is used to store the privacy and credit metrics obtained at the end of each bidding day for each user and is shown in figure **??**.

Bussiness Logic

?? Most of the bussiness logic used in the FairDataShare portal is present on Kinvey. Two things are done here:

1. Finding privacy for a user
2. Summarization

Data collected from the user consists of mobile sensor data with the least amount of summarization. This prevents for repeatedly collecting data from users and saves space and mobile data. Therefore, this mobile sensor data needs to be further summarized before being given to the stakeholder. To do this, we first have to find the most recent privacy setting from the UserResponseCollection. This is done due to the fact users can answer a data request more than once, hence we need to fish out the latest response for a particular sensor. Once this is done, using the recorded privacy level, we feed this input into another script that performs the summarization which has been explained in chapter 3.

5.3.2 FairDataShare Web Portal

The FairDataShare portal makes use of a server other Kinvey to safely store the usernames, passwords of users and the stakeholders. The database technology used is MongoDB. The username and passwords are both stored in a collection. The language used to interact with Kinvey is Express.js, which is based on Node.js. Most of the data portal business logic is on Kinvey described in section **??**. The webpage was constructed using Html and css. Screenshots of the portal are provided in chapter 4.

Chapter 6

Pre-Survey and Experiment Findings

6.1 Overview of the Pre-Survey Data

basic statistics about the data, plots to explain the data

6.2 Pre-Survey Methodology and Findings

6.3 Overview of the Experiment Data

6.4 Findings from the Experiment Data

Chapter 7

Conclusion

Appendix A

Appendix

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s): **First name(s):**

With my signature I confirm that
- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date **Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.