ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Fair Data Sharing in Participatory Social Sensing

Master Thesis

Ramapriya Sridharan

September 3, 2016

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

Contents

Abstract

Today, there are unlimited opportunities for people to generate and share their data in real time with the advent of Internet of Things, ubiquitous and pervasive computing [18]. Due to the unstructured nature, variability and rate of growth of this data, analysing and processing this efficiently can help to understand one's business or competitors that leads to better products and services. Today's Big Data systems use discriminatory, unfair and non-transparent ways to collect data where people are uninformed about the data collection process [22][20].

This thesis builds upon the paper by Pournaras et al [19] in which data sharing is modelled as a supply-demand system. The concept of a computational market is introduced where data aggregators who analyse data can incentivize people to share their data with a certain accuracy and this is performed using summarization functions that regulate the amount of information in the data shared. There is lack of information on how people make decisions in the sharing of mobile sensor data. This information is needed in order to design a platform for a computational market that rewards users fairly and increases the awareness of people about their privacy. This thesis fills this gap.

In this thesis, a social experiment and an Android application are designed to examine the relationship between mobile sensor data sharing and how incentives affect this. Users are requested for their data using data requests that are governed by three aspects: (a) sensor type, (b) the stakeholder to whom sensor data is shared, (c) the context or the application type for which data is shared. A computational model is introduced that assigns rewards to the possible data requests based on the user profiles formed by questioning users on the three aspects of data sharing. Data sharing decisions during the experiment are then tracked with and without incentives and the user data is recorded anonymously. The anonymous data recorded is then analysed and the findings can help in the making of platforms with fairer and more transparent data collection systems that make people more privacy aware.

Acknowledgement

Foremost, I would like to thank Prof. Dirk Helbing for giving me the opportunity to work under him. I would like to thank Dr. Evangelos Pournaras for being a pillar of support with his continuous mentoring, guidance, timely and constructive feedback through this six month period. Without him this thesis would have not being possible. In addition, I would like to thank Athina Voulgari for all her help with the pre-survey and Lewin Konneman for his help with the user interface of the mobile application.

Chapter 1

Introduction

Big Data systems today often collect data in ways that are unfair, non transparent and privacy intrusive to people. Most of the times, people are unaware that data collection is taking place. For this reason, new ways of acquiring data need to be designed. People should have control over their data and be given all the necessary information before data collection such as: (i) information about the data sharing process, (ii) who will access the data, (iii) what data is collected, (iv) time and duration of data collection, (v) what purpose the data collected is for. Additionally to the above mentioned points, users should be given the choice to control the privacy or accuracy of the data they share. In addition to security threats to data, there are also threats to people's privacies due to the information content in the data shared.

This thesis builds upon the earlier work on self-regulatory information sharing in participatory social sensing by Pournaras et al [19]. In this concept introduced, citizens who are suppliers are the ones sharing their data and they have the choice to choose how much of their data to share. On the other side, the data aggregators are the consumers who use the sensor data shared by citizens for some data analytics task. Data aggregators require data to be of some accuracy for their task and incentivize users accordingly to share data with a certain level of accuracy. For example, if data aggregators would like data with a higher accuracy, they can incentivize citizens with higher incentives for sharing their data. Similarly, if the analysis task does not require data with high accuracy, lower incentives can be awarded to citizens for sharing data. Information content or accuracy of data can be regulated using summarization functions concerning algorithms from simple arithmetic functions to clustering algorithms. A high summarization level filters out most of the information in the data, hence preserving the privacy of the citizen. This can cause higher errors for the data aggregator. Similarly, a lower summarization level preserves a higher information content

and causes lower errors for the data aggregator.

To design a computational market platform where users can share data in a more transparent and fair manner, there is a lack of information on the choices users make in sharing their data. Additionally, there is a need to know the perception of citizens on the privacy of their mobile sensor data and this data will help understand the supply-demand system for data sharing. The findings from the above can help to create a fairer system for data collection.

The contribution of this thesis is to bridge this gap. A social experiment is designed where citizens are approached with data requests governed by the three following aspects:

- The sensor type
- The stakeholder or entity requesting for citizen's sensor data
- The context or purpose of sensor data collection

Each data request is assigned rewards individually for every citizen using a computational model. This model first frames a citizen profile by asking citizens questions about the three aspects of a data request mentioned above. The citizen profile is then used to assign rewards to data requests.

A social experiment is then carried out with an Android application that has an inbuilt computational model. Citizen decisions are recorded anonymously for later analysis. A pre survey is launched before the social experiment is carried out to understand the perception of users on the three aspects studied. An exit survey is also launched after the completion of the social experiment to obtain feedback from participating citizens.

From the above, it is seen that 77.5% of citizens are moderately or more concerned about their mobile sensor data. Also, it is found that citizens are not motivated to share their data for no incentives or because their friends do so but are open to accept incentives other than money for data sharing. Findings from the social experiment indicate that even tough citizens are more interested in obtaining more rewards, ultimately the decision lies on the data request being asked. There is also an increased amount of data sharing and reduction in privacy when rewards are awarded to citizens. It is also found that citizens understood the value and privacy of their data through the rewards awarded to data requests. Additionally, it is found that 75% of citizens visited the FairDataShare portal to view the data collected from them.

Below is an outline of the thesis :

- **Chapter 2** - This chapter gives an overview of the previous work and elaborates on the drawbacks addressed in the thesis

- **Chapter 3** - This chapter introduces the working and mathematics behind the computational model
- **Chapter 4** - This chapter explains in detail the procedure and science behind the social experiment designed including the preliminary work
- **Chapter 5** - This chapter explains the implementation and algorithms used in the implementation of the Android application for the social experiment designed
- **Chapter 6** - This chapter presents the findings obtained from the surveys and the social experiment deployed
- **Chapter 7** - This chapter concludes the thesis presenting an overview of the work done, some important findings and possible future work

Chapter 2

Related Work

Participatory social sensing is the active participation of users with their mobile phones and any other sensory devices to form a network that enables the collection and analysis of data. The ubiquity of mobile devices and cellular infrastructure makes it a possibility to obtain data over a large area with minimal incremental cost. Burke et al talk about the concept [4] and further highlight the potential benefits and propose an architecture. Collecting data from various sensors is important for Big Data analysis and to find answers to complex social questions such as sentiment evolution and the spread of epidemics [12]. Lei Song et al [21] performs an extensive survey on the sensory devices for the purpose of health sensing and finds that different sensor combinations is the pillar to obtaining meaningful signals. Giannotti et al [12] propose the *Planetary Nervous System* to collect data from connected sensors and use that data to do big data analysis with privacy awareness.

Some applications of participatory sensing are LiveCompare, TraficSense and CenseMe mobile applications. LiveCompare is introduced by Linda Deng et al [11] where the widespread availability of mobile phones is made use of to find cheap groceries making use of the camera for barcode decoding and location to find the stores. TraficSense by Prashanth Mohan et al [16] is a concept aimed to keep track of traffic on the road with a mixture of traffic and vehicle types. It collects a variety of sensor data such as the accelerometer and the location but not limited to them. CenseMe created by Emiliano Miluzzol et al [15] where friends in social networks can share their status in terms of their mood, activity, surrounding and habit. This includes physical and virtual sensors that can capture the online life of a person.

Participatory sensing is needed for a fairer system to trade data due to the fact that many mobile applications take data away from users without their knowledge. Jinyan Zang et al [22] does a study using 110 Android and iOS apps to find the ones that share personal information, behavioural information and location data with third parties. This also reveals that collecting

user information does not require a notification from the application. They also find that paid applications still share sensitive information to third parties.

Ashwini Rao et al [20] examine the behavioural profiles formed by Google and Yahoo. Participants were surprised and concerned that data has been collected from them. Additionally, the profiles formed are found to be in some aspects inaccurate and have excess information, that the profiles did not seem to be anonymous anymore. Further, a survey is created asking participants questions about the behavioural profiles and was launched on Amazon Turk. Participants find the profiles formed to be not easily accessible and also complained that they wanted to know more about who is going to use their data and for what it will be used. Overall, the impression of participants is that the whole process of collecting data lacked transparency. This shows that there is a need for more privacy and control of data from the user side.

Studies have been done to investigate the relationship of users and their data. Alesaandro Acquisti et al [1] create a survey to observe the privacy concerns in e-commerce preferences and masking of location data. They find that users do not make reckless decisions, rather they make decisions based on what information they have, how much they care and what they believe the effect of their actions will be. This leans on the fact that with sufficient information, users can make rational choices about the privacy of their data.

Rebecca Balebacco et al [2] study through surveys and an experiment that users do not remember the sensors accessed by each application, shown during installation of the mobile application and proposes to inform users during the use of the application itself before collecting the sensor data. Similarly, Lin Jialiu et al [14] examines through crowdsourcing the perception of users to the data collection from mobile applications. The main takeaway from here is that users feel more comfortable if the purpose of a resource access was stated.

Additionally, studies have been done on assessing sensor data sharing in mobile phones. George Danezis et al [10] does a study to assess how much people value their location data using auction technique. They find that the median bid was 43$ for a period of one month, but this varies a lot on whether the person was a student, the relationship status and their travelling habits. Dan Cvreck et al [9] also examines the value of location privacy with over 1200 people and varied demographics. In this case the users are told fake goals in order not to be biased about their data privacy. Contradictions are found against the study [10] about the change in value due to travelling habits, but the median bid is found to be the same. There are also differences in results among the uses with different demographics. This shows that one

incentive does not fit all users.

Delphine Christine et al [8] perform an extention of the study in the paper of George Danezis et al [10]. It is attempted to analyse how various factors can affect data sharing such as demographics, incentives and spatio-temporal elements vary the importance users have on their data for various sensors. Other aspects such as the purpose for which data is shared and to whom the data is shared is also studied. It is found that younger people and people with affiliations with buyers of their data tend to share more information. They also find that users claim more rewards to corporations. The work by Camp Jean [5] mentions that the participants of the surveys may not tell the truth despite financial rewards. The later only ensures that the users successfully complete the survey.

Other than the surveys, there are studies done on the mobile phones themselves. Brush et al [3] collect the location data of 32 users for a period of 2 months. Users have five privacy options they can choose from:

- Deleting near home
- Mixing to provide k anonymity
- Randomizing
- Discretizing
- Subsampling

At the end of the two months, users are shown visualizations of their data. The authors mention that the user interface is not intuitive and that users might be biased to the location data due to the experimental setup. Additionally, it is found that users are not consistent with privacy decisions and with whom they share data. It is concluded that users need to be properly informed about every detail to enable them to make rational choices.

Haksoo Choi et al [6] propose a framework that provides sharing of sensor data based on rules along with the possibility of applying obfuscation algorithms. They find that users share data with a purpose and hence the purpose of data sharing should be included in the rules. Additionally, Eiji Hayashi et al [13] with 20 participants examine the sensor data sharing with all or no options. It is found that all or no options are a poor fit for user preferences and sharing of partial sensor information should also be provided.

Various algorithms can be used to protect the privacy of users while providing various amounts of data sharing possibilities. Pournaras et al [19] propose a scheme where users have the possibility to share various amounts of data. Users supply data to the data aggregators who buy data. The incentives that are received depend on the quality of the data that is shared and the quality required by the data aggregators. If the data shared is of

lower quality then the errors in data processing increase. Similarly, with higher quality the data processing tasks give lesser errors. The process of manipulating the quality of sensor data is called summarization. The errors in the data have the possibility to be mitigated if there is a large population of users participating in the data sharing tasks. Summarization concerns algorithms from simple arithmetic functions to clustering algorithms.

Delphine Christine et al [7] conclude their paper on the challenges for the future, the following points have been addressed [18]:

- Including the participants in the privacy equation
- Providing composable privacy solutions
- Trade-offs between privacy, performance and data fidelity
- Making privacy measurable
- Defining standards for privacy research
- Holistic architecture blueprints

Below is a collection of the drawbacks from the above mentioned papers:

- **No transparency in data collection** - Users are not informed about the process, time and duration of data collection.
- **Poor data accessibility** - Users do not have easy access to the data collected from them
- **Need for more information about data collection** - Users are not informed about what data is being collected, who is collecting their data and the purpose of collection
- **Users lack control over their data** - Users lack the possibility to share or not share their data. Furthermore they cannot decide on the privacy(quality) of data to share
- **No personalized incentives** - Incentives to data requests are not tailored to the user profile

Users need a platform where they have control over the sharing of their data. Furthermore, they should be informed about the process of data collection in a transparent manner by being given all the information needed to make a decision, whether to share their data or not. This includes the duration and time of data collection. Additionally, users should have the option to choose the privacy (quality) of their data to share instead of being restricted to all or no options. Users should know who is collecting their data, what data is being collected and for what purpose this data is used. Users should also have easy accessibility to see the data collected from them.

Chapter 3

Computational Model

3.1 Introduction

A data request is a request to users to trade their mobile sensor data. The aim is to create a computational model that is able to assign personalized rewards to each data request for every user. Each user is associated with a privacy intrusion profile. The model uses the user profiles to assign each data request a maximum achievable reward. The model attempts to identify the data requests where users might not be inclined to share mobile sensor data. These data requests are assigned a higher maximum obtainable reward. Similarly, the data requests where the users would want to share more mobile sensor data are assigned a lower maximum obtainable reward. This permits to see whether incentives do indeed make a difference in mobile sensor data sharing. The model aims to identify the amount of data each user would share for a data request and assign maximum obtainable rewards accordingly.

3.2 Model

The sections below explain the various building blocks of the computational model. Figure 3.1 provides an overview of the flow of the model.

3.2.1 Collecting User Information

To begin with, user information is collected. The information collected consists of but is not limited to :

- Gender
- Year of birth
- Country

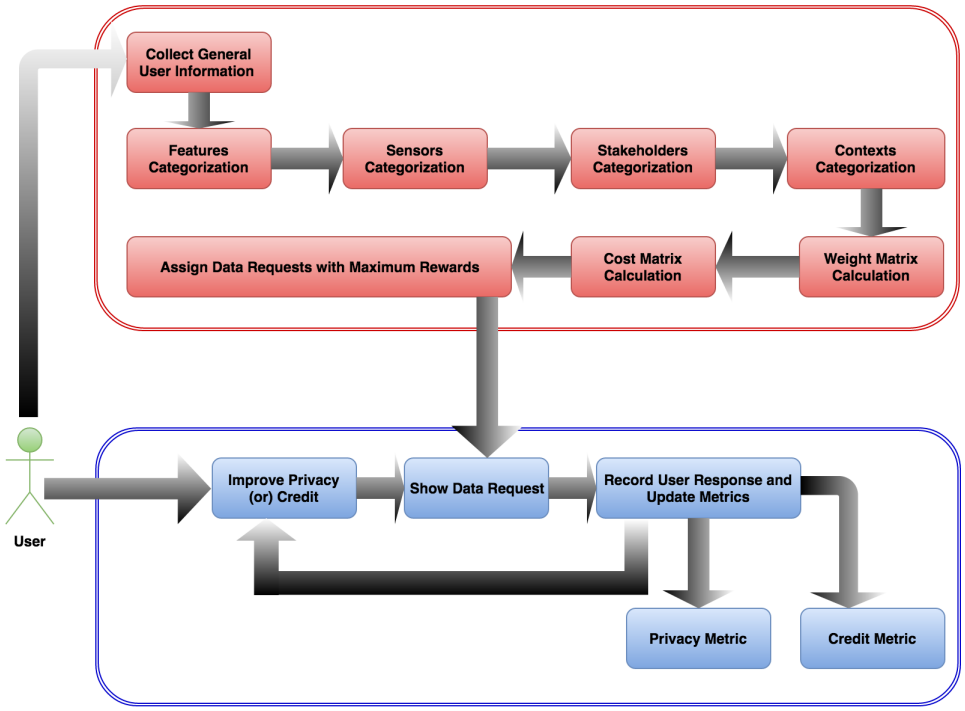


Figure 3.1: Computational model flow chart

- Education Level
- Occupation
- Frequency of mobile phone use per day
- List of different mobile applications present on the users phones

3.2.2 Categorization of the Features

Features are the aspects that govern the data sharing decision. A feature can be one of the following:

- **Sensors** : They consist of the data obtained from sensors in the mobile phone that users can trade for a data request
- **Stakeholders** : They consist of any entity that requests users for mobile sensor data
- **Contexts** : They consist of the purpose for which a stakeholder would like to obtain the user's mobile sensor data

Features are placed in categories according to how privacy intrusive they are. Features are the three dimensions that form a data request. A data

request is defined as a stakeholder asking users to share their mobile sensor data for a particular context. Users are asked to categorize the features into one of the five categories:

1. Very low privacy intrusion
2. Low privacy intrusion
3. Medium privacy intrusion
4. High privacy intrusion
5. Very high privacy intrusion

Categories are linearly scaled and equally spaced. As indicated by the numbers on the left of the categories, these range from 1 to 5 and users can place each of the features in a category according to their perceived intrusion level. Category 1 represents that the feature does not at all contribute to the data sharing decision. Similarly, category 5 represents that the feature contributes to the maximum possible for the user's data sharing decision. It represents that users are reluctant to give away their sensor data for this feature. More than one feature can be placed in the same category.

Let the variable n_{cat} represent the number of categories, which here are five. Additionally, let the category assigned to the sensors be represented by the variable a, the category assigned to the stakeholders be represented by the variable b and the category assigned to the contexts be represented by the variable c.

Once users have categorized the sensors, stakeholders and the contexts into the respective categories reflecting the importance of each of the features in the data sharing decision, each feature is assigned a weight. Let the respective weights of sensors, stakeholders and contexts be represented by the variables, w_a, w_b and w_c and calculated as follows:

$$w_a = \frac{a}{a+b+c} \tag{3.1}$$

$$w_b = \frac{b}{a+b+c} \tag{3.2}$$

$$w_c = \frac{c}{a+b+c} \tag{3.3}$$

3.2.3 Categorization of the Sub-Features

Once the features have been categorized and their weights calculated as above, sub-features are to be categorized. A sub-feature is defined as a type of a feature. In other words, sub-features are the different types of features that appear during data request to the user. The following are examples of sub-features for each feature :

- Sensors :
  - Accelerometer
  - Battery
  - Gyroscope
- Stakeholders :
  - Corporation
  - Government
  - Educational Institution
- Contexts :
  - Education
  - Navigation
  - Gaming

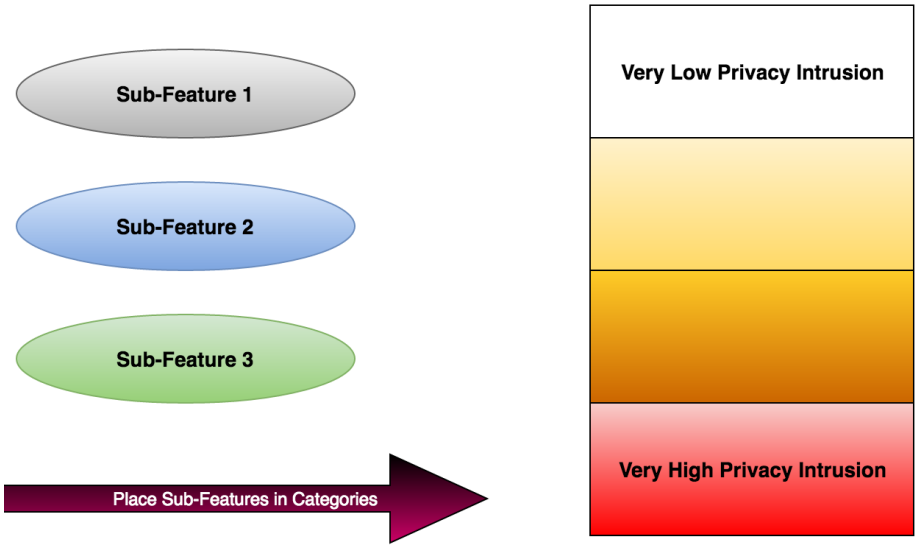


Figure 3.2: Categorizing sub-features according to the perceived intrusion level

Each of the above are different kinds or sub-features of the respective features. For each of the available features, the respective sub-features need to be in turn categorized in the same categories mentioned in Section 3.2.2.

The categories are the same as mentioned in the previous section. Let n_{sf} be the number of sub-features each feature has. It is assumed that each feature has the same number of sub-features.

As seen in the conceptual diagram which is shown in the Figure 3.2, users place each of the sub-features available for every feature in the given categories.

Let every sub-feature be represented by unique indices within its feature. For example, in the list of sub features provided above, accelerometer is the first sub-feature of sensors, corporation is the first sub-feature of stakeholders and education is the first sub-feature of contexts. For each of the sub-features of sensors, categories they are placed in by users is represented by a_i and i is the index of the sub-feature. Similarly, categories assigned to sub features of feature stakeholders and contexts respectively are represented by b_j and c_k, where j and k are the identifiers of the sub-features categorized.

3.2.4 Weight Matrix Calculation

Each data request consists of the three above mentioned features in them. Each of the features each have n_{sf} sub-features that can appear in turns in a data request assuming that each feature has the same number of sub-features. The total number of possible data requests are $n_{dr} = n_{sf}^3$.

Let W be a matrix with three dimensions n_{sf} x n_{sf} x n_{sf}. We call this the weight matrix. Each cell of W, that is $W_{i,j,k}$ represents a data request which involves the sensors sub-feature with identifier i, stakeholders sub-feature with identifier j, and the contexts sub-feature with identifier k. That is, each cell of W represents the weight of a data request to the users. The aim of the weight matrix is to use the information collected from the user categorizations, to assign 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 is seen in Section 3.3 with examples. As mentioned before, each cell of the matrix W represents the weight of a data request with a unique sensor sub-feature i, stakeholder sub-feature j and context sub-feature k. To calculate the weight of a data request :

$$W_{i,j,k} = (a * a_i) + (b * b_j) + (c * c_k) \tag{3.4}$$

Applying this formula for every possible values of i, j and k gives the weight matrix W.

3.2.5 Cost Matrix Calculation

The aim is now to assign a maximum obtainable rewards to each data request. This reward is the maximum credit users can receive for a particular data request. Let C be the cost matrix with the three dimensions n_{sf} x n_{sf} x n_{sf}. Let it be assumed to have a budget of B for a day, where B can be in an actual currency or any form of virtual credits. In this thesis the budget will be referred to with the unit credits. Each cell of the cost matrix will represent the amount of rewards allocated for a particular data request for one day. To begin with, we calculate the sum of all the cells of the weight matrix W:

$$s_W = \sum_{i=1}^{n_{sf}} \sum_{j=1}^{n_{sf}} \sum_{k=1}^{n_{sf}} W_{i,j,k} \tag{3.5}$$

Where the function s_W gives the sum of a matrix, in this case the weight matrix. Let $C_{i,j,k}$ 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 :

$$C_{i,j,k} = \frac{W_{i,j,k} * b}{s_W} \tag{3.6}$$

Repeating the above for every cell of C, the entire cost matrix can be calculated. Now, all the maximum obtainable rewards have been allocated per day for every data request.

3.2.6 Cost and Privacy Metrics

Every data request has been assigned a reward. This is the maximum reward that a user can obtain for that data request. The cost metric is the amount of rewards the user has obtained by trading sensor data for each data request. Similarly, the privacy metric is the amount of privacy percentage obtained while trading data for requests. It intuitively quantifies the amount of data the user has refused to share hence implying privacy. The cost and privacy metrics are inversely proportional to each other, in the sense that when the cost increases the privacy decreases and vice versa.

In each data request, one chooses how much data is to be shared, from the maximum amount of data to no data at all. The possible responses to a data request are called options. Each option corresponds to a summarization level explained in detail in Section 3.2.8. The reward assignment to each option is linearly scaled according to the reward 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 the option where the users give all their data for a request and m to where the users choose not to give any data for a request. Therefore there are a total of m options for every data request. Each option in a data request has the following:

- The amount of credit change if this particular option for a data request is chosen
- The amount of privacy change if this particular option for a data request is chosen

While assigning rewards to data requests there are two scenarios to consider:

- Assigning option rewards without a participation reward. Users are not rewarded for responding to data requests
- Assigning option rewards inclusive of a participation reward. Users are rewarded for responding to data requests irrespective of the option chosen

Let us examine the first scenario. Let the option rewards be calculated for the data request with sensor sub-feature i, stakeholder sub-feature j and context sub-feature k. The assigned reward for any option numbered h of this data request is calculated as follows:

$$r_h = \frac{C_{i,j,k} * (m - h)}{m - 1} \tag{3.7}$$

Applying this formula by replacing h by the option numbers from 1 to m gives the reward that the user can receive for each option. Similarly, if a participation reward is assigned to each option, it means that even tough the user does not share data, they still receive rewards for answering the data request. Let x be a fraction of the total budget B that is dedicated for user participation. Using a geometric progression with $a = 1$ and $z = \sqrt[(m-1)]{x}$, we can calculate the fraction of the maximum reward obtainable from a data request f_h, an option numbered h gets:

$$f_h = a * z^{h-1} \tag{3.8}$$

The fraction of the rewards an option h can be assigned has been calculated, to get the rewards r_h of option h for the data request with sensor sub-feature i, stakeholder sub-feature j and context sub-feature k :

$$r_h = f_h * C_{i,j,k} \tag{3.9}$$

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

Privacy percentage p_h is linearly scaled between the first to the m^{th} option between 0 and 1 as follows:

$$p_h = \frac{(h-1)}{m-1} \tag{3.10}$$

The total cost and privacy is the sum and arithmetic average of all the rewards and privacy respectively, obtained from every answered data request. If a data request is left unanswered, a maximum privacy of 1 and minimum cost of 0 credit is assumed.

3.2.7 Improving the Metrics

Users can choose to improve the privacy or cost metric. If one chooses to improve their cost, the data request where maximum rewards can be obtained. Otherwise, if users want to improve their privacy the data request which can maximize the privacy is obtained. Additionally, options that can improve either metrics are indicated.

3.2.8 Summarization of Collected Data

Each data request has the possibility to have m number of options the user can choose from for every data request. These options range from 1, which indicates that the user would like to give all his data, to option number m, which indicates that the user does not want to give any data to this data request.

Summarization is a privacy algorithm that modifies the quality of data to provide less information than in its original form [19]. A higher summarization level gives data with a lower quality. A lower summarization level gives data with a higher quality. In this model, sensor data is collected for a period of d 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 less.

Data is collected for a period of d hours, and at the end of this period according to the option chosen by the user, it is summarized. Summarization can be linearly assigned to each option. The highest privacy option m corresponds to the highest summarization level. The first option corresponds to the lowest summarization level. An example of assigning the summarization level l_h for an option h for a data request has the possibility of the following :

$$l_h = y * h \text{ where } h \neq m \tag{3.11}$$

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

3.3 Analysis of the Model

In this section, three different examples are explained in order to witness some properties of the weight and cost matrices.

3.3.1 Setup

In the following examples, the following features and sub-feature are considered:

1. Sensors
    a) Accelerometer
    b) Noise
    c) Location
2. Stakeholders
    a) Corporation
    b) Government
    c) Educational Institution
3. Contexts
    a) Navigation
    b) Environment
    c) Social Media

Numbers indicated to the left of the sub-features are the corresponding unique indices. This uniquely identifies a sub-feature of a feature. There are in total $num_{sf} = 3$ sub-features for each feature. Each user will receive a number of $num_{sf}^3 = 27$ data requests in total. The number of categories available to categorize is $n_{cat} = 5$ as explained in Section 3.2.2. Additionally, it is assumed that a budget $B = 100$ Chf per day is available. The input to the model are the user choices during the categorization of the features and sub-features.

3.3.2 Examples

To make referencing easier to the graphs, instead of sub-feature names, numeric indices are used. From now on each feature and sub-feature will be referred to by its index such as feature 1 for sensors and sub-feature 2 of feature 1 for the noise sensor. The tuple (a,b,c) represents a data request with:

1. a - Sensor's sub-feature a
2. b - Stakeholder's sub-feature b
3. c - Context's sub-feature c

where a, b and c are all numbers from one to three.

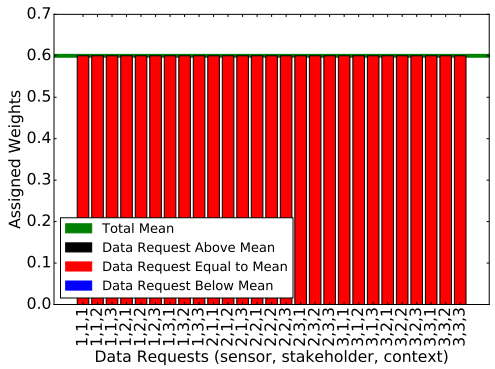
If features and sub-features have all been given the same categories by users respectively, then all data requests should be assigned equal weights and rewards. In example 1, users choose categories for the features and sub-features as shown in Table 3.1. From this input, the formulation of the weight matrix is shown in Figure 3.3a, and the cost matrix is shown in Figure 3.3b. For each data request indicated as a tuple (sensors, stakeholders, contexts) in the x-axis of Figures 3.3a and 3.3b, all have identical weights and rewards. This is due to the fact that the users find all the features and sub-features equally intrusive so all the data requests are weighed equally.

Table 3.1: Categorization for example 1

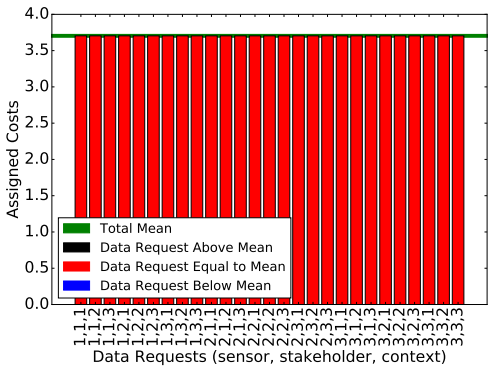
Feature	Sub-Feature ID = 1	Sub-Feature ID = 2	Sub-Feature ID = 3
Sensors	Accelerometer	Noise	Location
1	3	3	3
Stakeholders	Corporation	Government	Educational Institution
1	3	3	3
Contexts	Navigation	Environment	Social Media
1	3	3	3

It is concluded that if the users perceive the features and respective sub-features in an equally intrusive way, then all the data requests will receive the same weight and reward assignments.

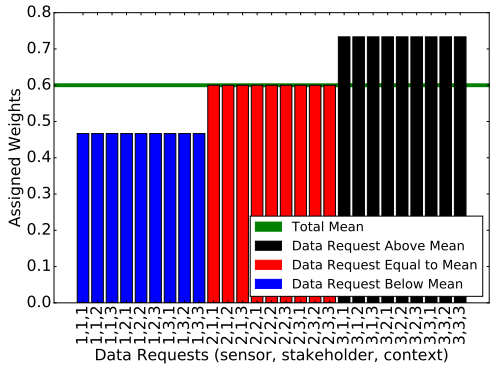
For example 2, it is attempted to test if data requests containing sub-features with higher intrusion levels are assigned higher weights and rewards. Table 3.2 indicates the user input. As it can be seen, all features have equal categories, and all sub-features have the categories of 3 with an exception of the sensor sub-features. The sensor sub-features with indices 1,2 and 3 have respectively categories 1,3 and 5. This means that requests with sensor sub-



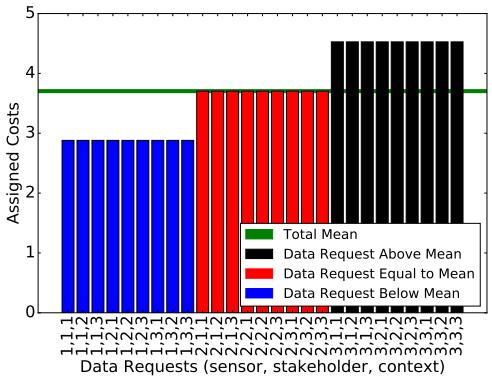
(a) Weight matrix example 1



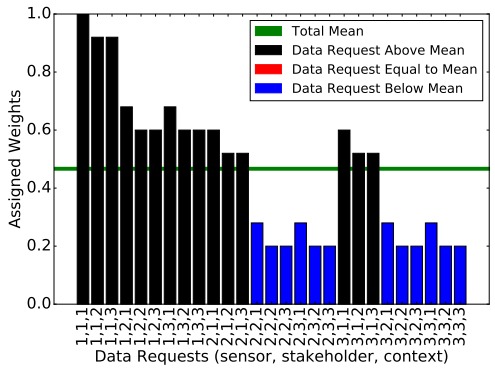
(b) Cost matrix example 1



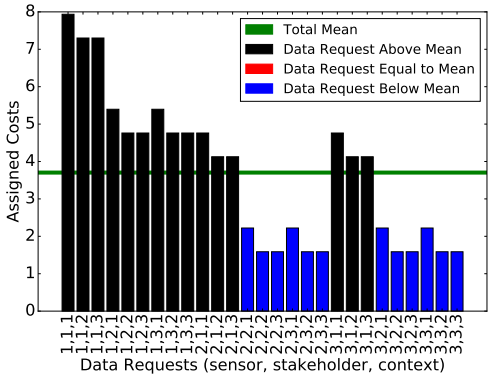
(c) Weight matrix example 2



(d) Cost matrix example 2



(e) Weight matrix example 3



(f) Cost matrix example 3

Figure 3.3: Values assigned to matrices

feature 1 will be assigned a lesser weight in comparison to the other sensor sub-features.

Similarly, the data requests with sensors sub-feature 2 will have a higher weight assigned than sensor sub-feature 1 because of its higher category, but lower than sensor sub-feature 3. Lastly, data requests with sensor sub-feature 3 will have the highest weight compared to others, due to the category being 5. The weight and cost matrices can be seen in Figures 3.3c and 3.3d respectively.

Table 3.2: Categorization for example 2

Feature	Sub-Feature ID = 1	Sub-Feature ID = 2	Sub-Feature ID = 3
Sensors	Accelerometer	Noise	Location
3	1	3	5
Stakeholders	Corporation	Government	Educational Institution
3	3	3	3
Contexts	Navigation	Environment	Social Media
3	3	3	3

From the above inputs and graphs, it is concluded that the model assigns a higher weight to data requests with sub-features that users find more intrusive compared to the others.

For example 3, the feature and sub-feature categories are both assigned different values, to show how varying their values together affects the assignments of the weight and cost matrix. Table 3.3 is the user input. All the features have different categories assigned from 3 to 5. Additionally, the sub-feature 1 of each feature has a category of 5, higher than the other sub-features which are all categorized as 1. The weight and cost matrices generated are shown in Figures 3.3e and 3.3f respectively.

Table 3.3: Categorization for example 3

Feature	Sub-Feature ID = 1	Sub-Feature ID = 2	Sub-Feature ID = 3
Sensors	Accelerometer	Noise	Location
5	5	1	1
Stakeholders	Corporation	Government	Educational Institution
4	5	1	1
Contexts	Navigation	Environment	Social Media
3	5	1	1

As it is observed in both figures, the data request with the highest weight is the one with the tuple (1,1,1). This tuple indicates that the data request involves all sub-features 1 of each feature. This happens because all of the sub-features 1 are assigned a category of 5. The feature sensor and its sub-feature 1 are categorized as 5, so all the data requests with tuple (1,*,*), where * represents all the other possible sub-features from other features, are all above average as seen in Figure 3.3e, irrespective of the categories of the other feature sub-features. This shows that assigning a higher category to a feature can lead to higher data request rewards.

The green horizontal line in the graph indicates the mean value of the weights and rewards. In general due to sub-features categorized as 5, those data requests receive a higher weight and reward. In some cases, the data requests still receive a lower weight such as tuple (2,2,1), (2,3,1),(3,2,1) and (3,3,1) tough context sub-feature 1 has a category of 5. This is due to the fact that sensor and stakeholder feature have a higher category of 5 and 4 respectively than the context feature. Since their sub-features are assigned a lower privacy intrusion category than the context sub-features, the weight of the data requests is lower. This shows that even tough a sub-feature may be regarded as very intrusive, it's weight increasing changing ability depends on the category of the feature it belongs to.

Additionally, it is noted that data requests with at least two sub-features 1 are all above average. We can witness the property of the model, which puts more emphasis on the perception of the features than the sub-features themselves. As seen in the figure, all the features with higher intrusion categorizations have weights and rewards that are well above average.

It can be concluded that the model assigns weights to data requests, by emphasizing on the feature's weights. A feature with a high category has the ability to assign higher rewards with a highly categorized sub-feature. It also has the ability to lower the weight of a data request with a sub-feature lowly categorized. Features with lower categories contribute lesser to the weight assignments, irrespective of their sub-feature categories.

Chapter 4

Experiment Methodology

In the previous chapter, the computational model has been explained in detail. This model has been implemented as a mobile application for the Android platform and is used to collect real data from users. This application will help collect information that will help see the influence of incentives on the data sharing decision. In this chapter some of the work and decisions that are taken before and after the start of the experiment are explained. It is then proceeded to explain how the experiment is carried out along with detailed instructions of the usage of the mobile application created.

4.1 Overview of The Experiment

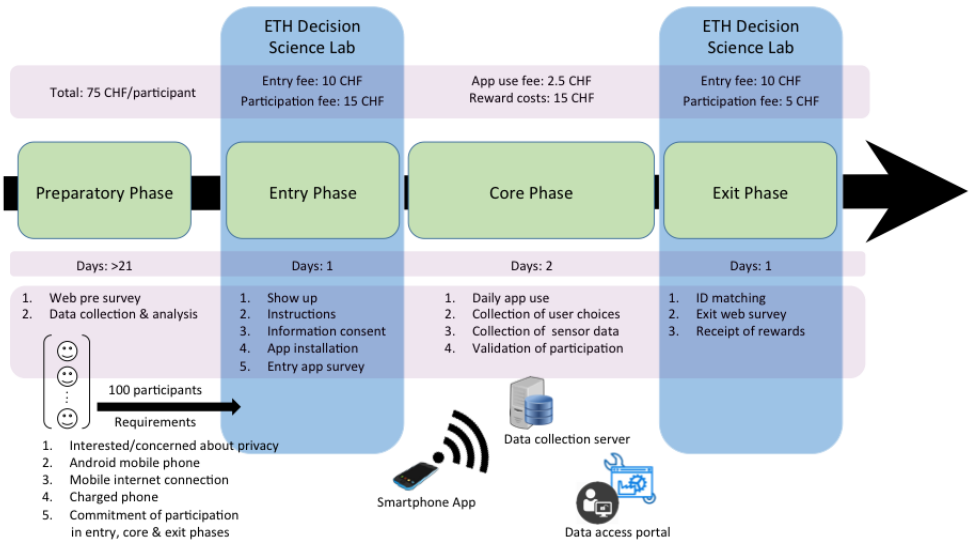


Figure 4.1: Experiment phases

The goal of the experiment is to examine the perception of privacy of users on mobile sensor data and to see how this affects their decision to share mobile sensor data. Furthermore, the experiment aims to understand the influence of monetary incentives on mobile sensor data sharing.

Figure 4.1[1] describes an overview of the social experiment. The first phase is the preparatory phase during which a pre-survey is launched to examine the perception of users on the three features which are explained further in Section 4.2. They are used to reduce the possible number of data requests to a manageable number and help make design decisions for the experiment.

The entry phase will take place at the ETH Decision Science Lab and a pool of users from the ETH and UZH are invited randomly by email. To participate, users have to fulfill the following requirements [18]:

- Interests or concerns about privacy
- Should own an Android smartphone version 4.4 and above
- Should have an internet connection on their phone
- Should appear at the entry phase with a fully charged phone
- Should be committed to participate in the next phases to come

Users are briefed about the experiment using oral and written instructions shown in Appendix A. Appendix A also shows the consent form that needs to be signed before their participation in the experiment. Users should then download the mobile application[2] from the Play Store and are each automatically assigned a unique identifier . Additionally, users will have to answer a survey that permits us to collect user data and configure parameters for the core phase. The questions (data requests) are similar to the ones asked during the core phase except that no incentives are awarded for each data request answered. These questions are used to validate whether users change their attitude to privacy when incentives are awarded.

During the core phase, users answer data requests to share their sensor data in exchange for monetary rewards for a number of days. Data sharing is performed through questions that are automatically formulated as explained in Section 4.2.5. In this phase, users can respond to questions by choosing from the five available amounts of sensor data to share for each question. Users are also informed about the amount of privacy and cost they have obtained by responding to questions, which are initially zero. The impact of each decision on the privacy and cost is indicated for every question.

[1] Figure created by Dr. Evangelos Pournaras

[2] https://play.google.com/store/apps/details?id=ch.ethz.nervousnet.trialapp04&hl=en

A few days after the end of the core phase, users are invited back to the ETH Decision Science Lab. Using the unique identifier assigned in the application, users are requested to fill an exit-survey to get feedback of their experience. Data received on the server from users will be evaluated before handing out money for participation and rewards for each phase. The following sections will give more details about each phase and detailed instructions of how to use the mobile application.

4.2 Preparatory Phase

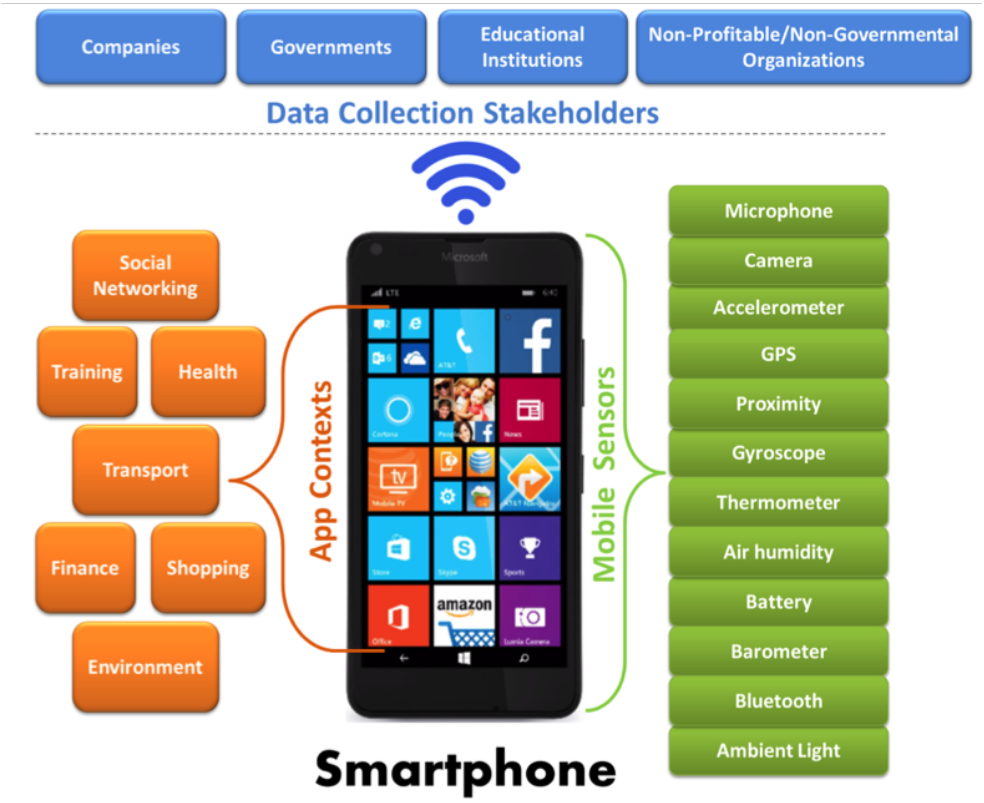


Figure 4.2: The three features examined

4.2.1 Pre-Survey

The pre-survey [3] shown in Appendix A is a survey created that runs before the deployment of the social experiment. The survey was made in order to study the perception of users on the three features to be studied which are

[3] https://descil.eu.qualtrics.com/SE/?SID=SV_0xGS6kfmr8GtQd7

explained in detail in Section 4.3.2. Figure 4.2[4] depicts the features and their sub-features visually.

As shown in the figure, there were a lot of sub-features to choose from each feature. Increasing the number of sub-features for each feature in the experiment in turn increases the number of possible data requests posed to the user. Additionally, we wanted to gain insight into the perception of the privacy intrusion of users on the different sensors, stakeholders and contexts. Hence the survey was prepared to understand all of the above. Additionally, the survey can help redesign some of the aspects of the experiment based on the ambiguities found and user feedback. The participants pool consists of people who are aware and unaware of data privacy and sensors. Participants were not paid for filling out the survey. Till now, 199 entries have been recorded.

4.2.2 Sub-Features

Figures 6.4a, 6.4c and 6.4e, each show the average intrusion level of each possible sub-feature for the sensors, stakeholders and contexts. For the experiment, it was decided to choose for each feature two non-intrusive and two intrusive sub-features. The reason is that if all sub-features sensors, stakeholders and contexts are chosen to be in the experiment, the number of possible data requests would be high. This is done in order to restrict the number of possible data requests. The minimum privacy intrusion level is one which indicates this sub-feature to not be intrusive, and the maximum is five which means that the sub-feature is very privacy intrusive.

For the sensors feature, it can be observed that the sub-features GPS and microphone are found to have an intrusion of 4.2 and 3.8 on a scale of five, which means users find these sensors on average very intrusive. On the other hand, sub-features light and accelerometer are found to be lower in intrusion with values of 2.2 and 2.3 on a scale of five, which means that users find these sensors non-intrusive in general. The average of all sensors intrusion values is 2.8 as indicated by the horizontal line.

Similarly, looking at the stakeholder feature graph 6.4c, it is seen that sub-features corporation and government are found to be intrusive by the users with levels 3.8 and 3.6 on a scale of five. On the other hand, sub-features educational institution and non governmental organization are found to be relatively less intrusive by the users with values of 3.2 and 2.95. For intrusion levels of contexts feature in graph 6.4e, it is observed that sub-features social-networking and health are found to be intrusive by the users with values 3.8 and 3.6 on a scale of five. Sub-features environment and transportation are regarded as less intrusive by users with values of 2.9 and 3.3 on a scale of

[4]Figure made by Athina Voulgari

five. The above mentioned sub-features for every feature have been chosen for the experiment.

4.2.3 Stakeholders

To make the experiment more realistic and to be transparent as to which stakeholder will be viewing user data, real stakeholders were approached to view shared user data. In this way, user choices to data requests will also be more realistic. In this experiment for each type of data collector, a stakeholder was chosen:

1. Tagesanzeiger as a Corporation
2. Swiss Made Software as a Non Governmental Organization
3. State Secretariat For Education Research and Innovation as a Government Organization
4. ETH Zurich as an Educational Institution

The letters stating their support to participate in the experiment are in Appendix A.

4.2.4 Privacy Options

Each data request is accompanied with privacy options ranging from 1 to 5 as explained in section 3.2.6. Option one indicates that the users would like to share their raw data without any sort of summarization or reduction in information. Option 5 indicates that the users would not like to share their data for this data request. The options in between have linearly scaled summarization levels assigned to them ranging from least privacy (1) to most privacy (5). For more information on the summarization levels for each option please refer to Section 3.2.8.

4.2.5 Question Structure

A data request is when a stakeholder asks for the users 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 :

"Please choose the amount of X data shared with Y to be used in the context of Z"

where Sensors X can be :

1. Accelerometer
2. Light
3. Noise

4. Location

where Stakeholders Y can be:

1. Corporation, Tagesanzeiger
2. Non Governmental Organization, Swiss Made Software
3. Government, State Secretariat For Education Research and Innovation
4. Educational Institution, ETH Zurich

and where Contexts Z can be:

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

In total this makes 64 data requests to the user. From now on, we will refer to mobile sensor data as just data.

4.2.6 Budget and Experiment Duration

The experiment is set to run for a total of two days with the help of the ETH Decision Science Lab, excluding the time taken for the entry phase and exit phase. From the total budget available for running the experiment, the budget is spread among the various experiment phases according to the number of participants estimated. 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 showing up 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. The budget is spread this way among all phases so that users are motivated to participate through out the experiment. Out of the budget B, $\frac{1}{7}$ is given away for the participation of the users in the core phase to motivate them to respond to data requests even if they do not want to give their data for data requests.

4.3 Entry Phase

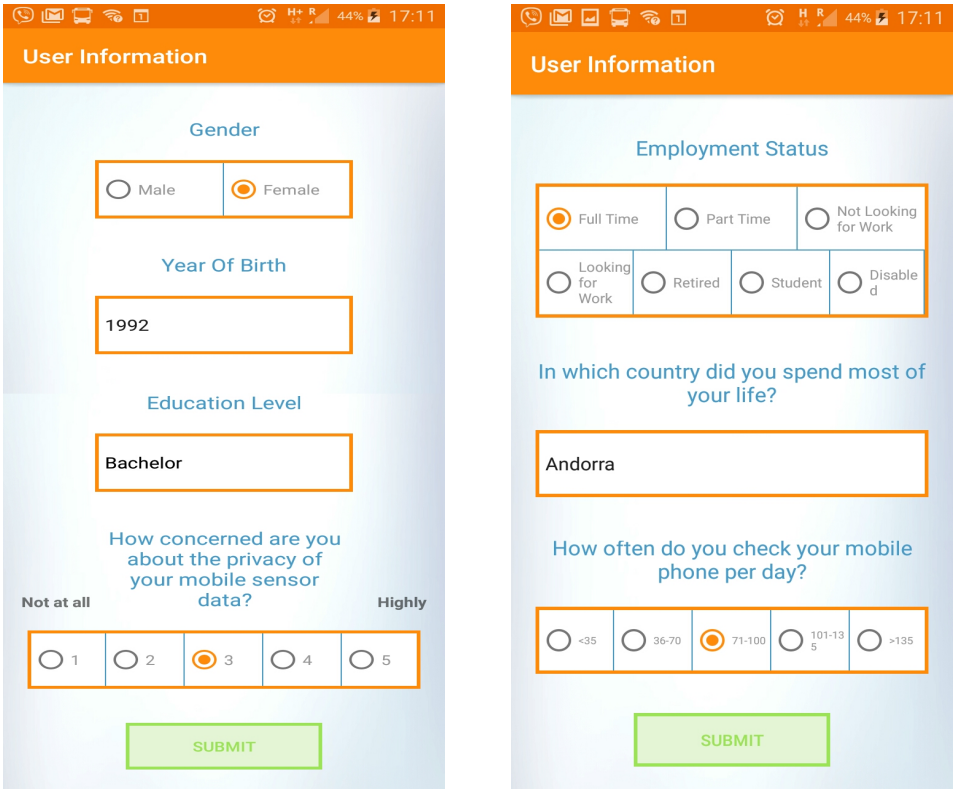
The entry phase denotes the first day of the experiment. Users are asked to install the Fair Data Share mobile application[5] from the PlayStore. Each application installed is assigned a unique identifier and remains the same throughout the experiment.

[5] https://play.google.com/store/apps/details?id=ch.ethz.nervousnet.trialapp04&hl=en

4.3.1 Collecting General User Information

Once the application is installed as the Figures 4.3 and 4.4b show, users are asked to answer some personal non-intrusive questions. The following is asked from the users:

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 times a day do you check your mobile phone per day.
7. Kind of applications the user has in the mobile phone.



(a) User information screen 1 (b) User information screen 2

Figure 4.3: User information screens

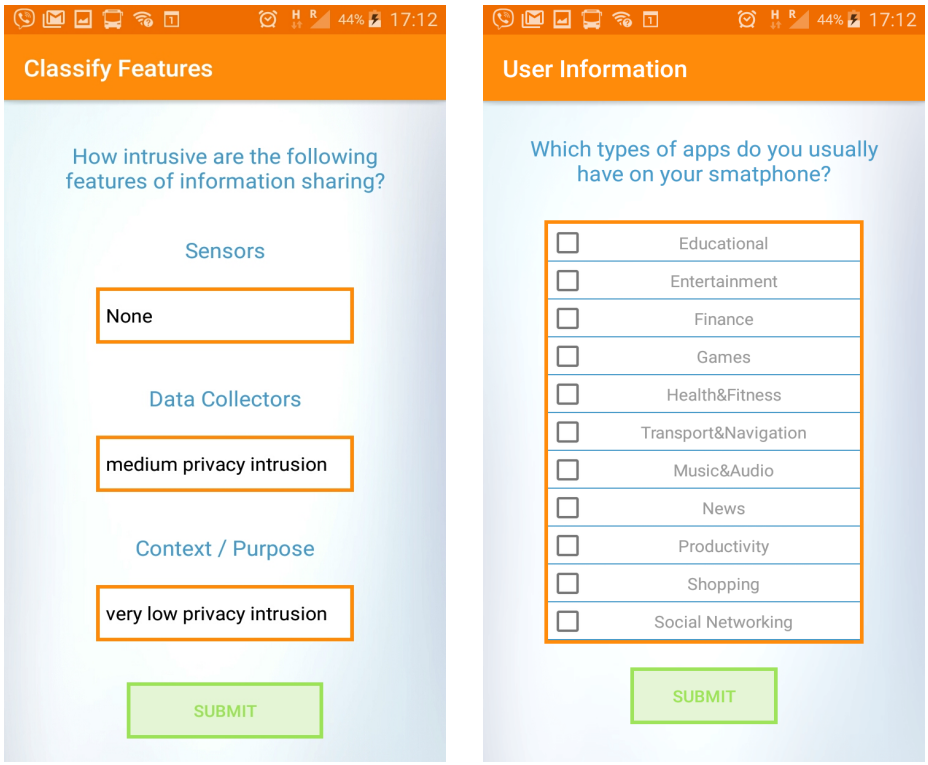
The users may go back and re-answer the questions, but once the submit

button is pressed on the screen 4.4b, the data is sent to the server and hence cannot be changed. Users cannot navigate to the next pages without filling out all the questions.

4.3.2 Categorization of Features

As described in chapter 3, users are asked to categorize the features sensors, stakeholders and contexts. As shown in Figure 4.4a, each of the features are indicated followed by a drop down list of privacy options ranging from *"very low privacy intrusion"* to *"very high privacy intrusion"*. The option *"very low privacy intrusion"* indicates that the feature does not affect the mobile sensor data sharing decision at all, whereas *"very high privacy intrusion"* indicates that the feature affects the sharing of mobile sensor data to the maximum.

Users need to click on the drop down menu to choose one of the privacy intrusion options. All the options are compulsory, and no default option is provided. Users cannot navigate to the next page without filling out all of the questions.



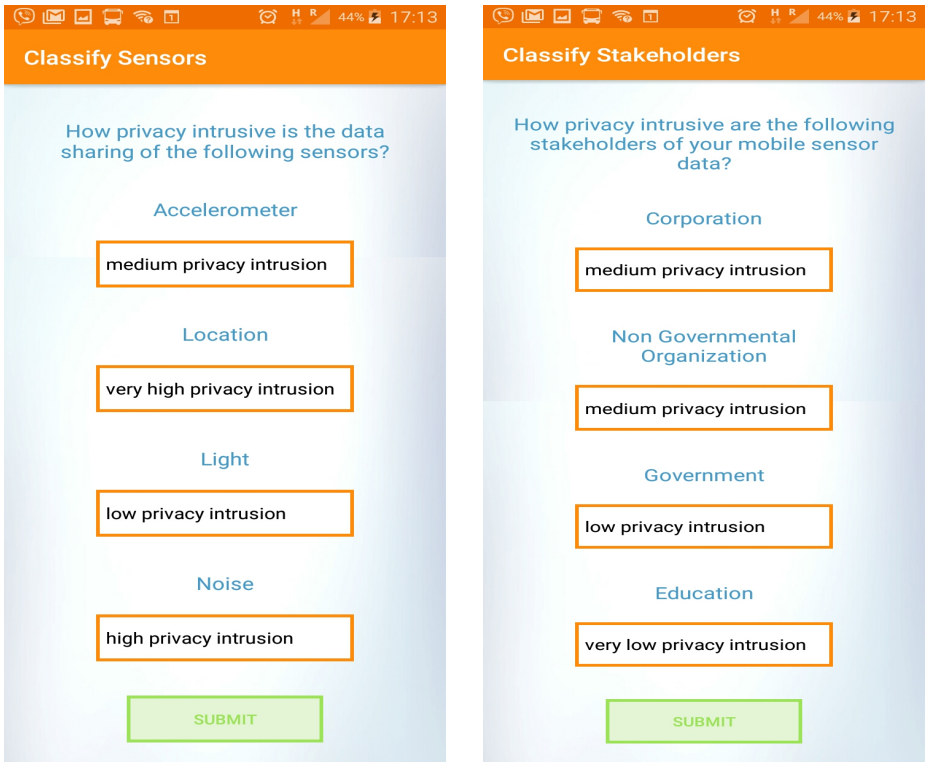
(a) Categorizing features (b) User information screen 3

Figure 4.4: Categorization and user information screens

4.3.3 Categorization of Sub-Features

For each of the features categorized in the previous sub-section, their sub-features need to be categorized in a similar fashion. Once again, the privacy options range from *"very low privacy intrusion"* to *"very high privacy intrusion"* like in section 4.3.2 . The users are first presented with the categorization of Sensors sub-features as shown in Figure 4.5a.

Below each sensor is a drop down menu where the user has the possibility to choose how much each of the sensors would affect the mobile sensor data sharing. Once all the sensors have been associated with a privacy intrusion level, the user can click the green submit button and is directed to the next page where various stakeholders need to be categorized. This is depicted in Figure 4.5b.



(a) Categorizing Sensors (b) Categorizing Stakeholders

Figure 4.5: Categorization of sensors and stakeholders screen

Each stakeholder has a drop down menu where users can classify how much each of them affect their data sharing decision. Once the user has finished entering the privacy intrusion level for stakeholders sub-features, the user has the possibility to click the green submit button and is directed to the

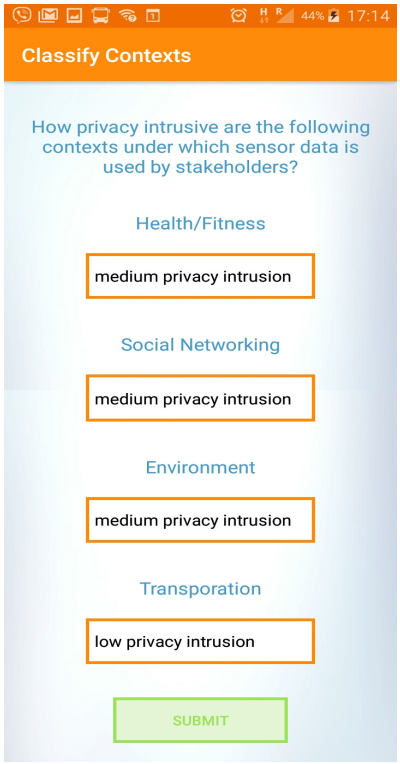


Figure 4.6: Categorization of contexts screen

next page.

On this page, the users are asked to categorize how much each of the contexts sub-features affect mobile sensor data sharing. This is depicted in Figure 4.6. Each context has a drop down menu below, where the user can rate each context. Once this has been done the user can click on the green 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 no default choice.

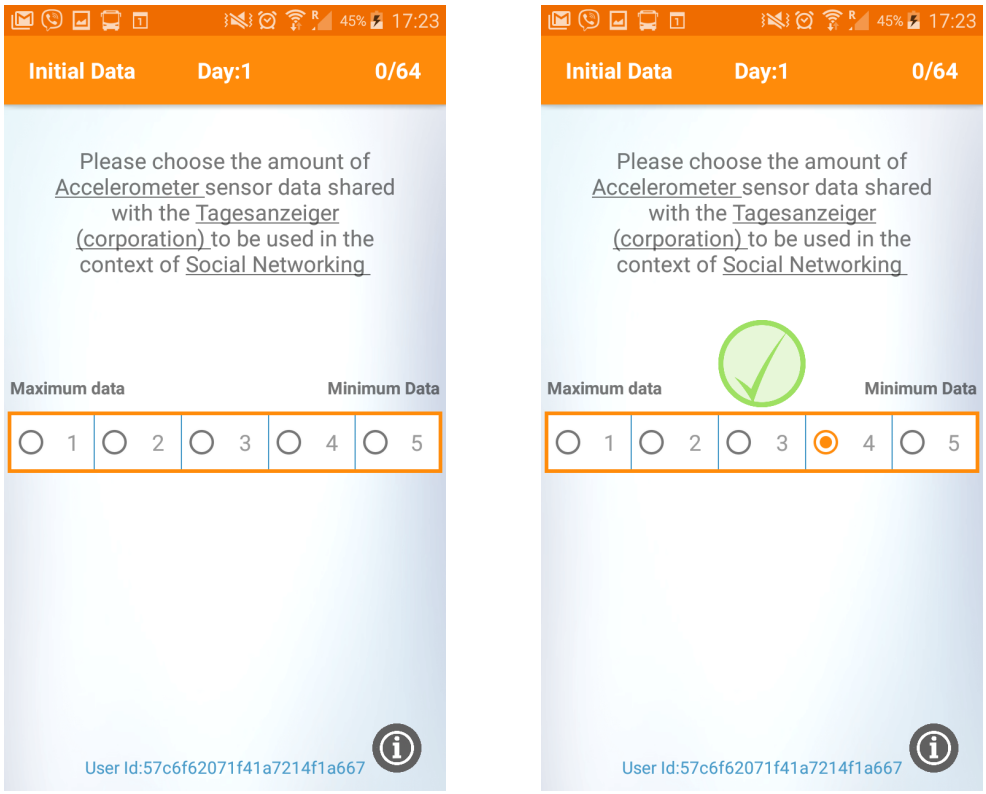
4.3.4 Answering Questions with No Incentives

After users are done categorizing the various sensors, stakeholders and contexts, they are presented with 64 data requests. Each of these requests are a stakeholder requesting the user for their mobile sensor data for a particular purpose or context. Users have the possibility to choose from the five available privacy options mentioned in Section 4.2.4.

The options are indicated as a measure of how much data users can give, ranging from maximum data to minimum data. The higher the privacy for the option, the less information about the sensor data is given away for that

request and vice versa. Users can change the answers for a data request until the green submit button that appears is clicked. The screen with the data request is shown in Figure 4.7a.

After the users choose an option for the data request, a green submit button appears which is shown in figure 4.7b. Clicking on the submit button sends the response of the data request to the server and cannot be further changed. At this stage, no indications of credit gained or privacy improvements are indicated and this is used to control the behaviour of the user to see whether they change their attitudes to privacy when they receive monetary incentives.

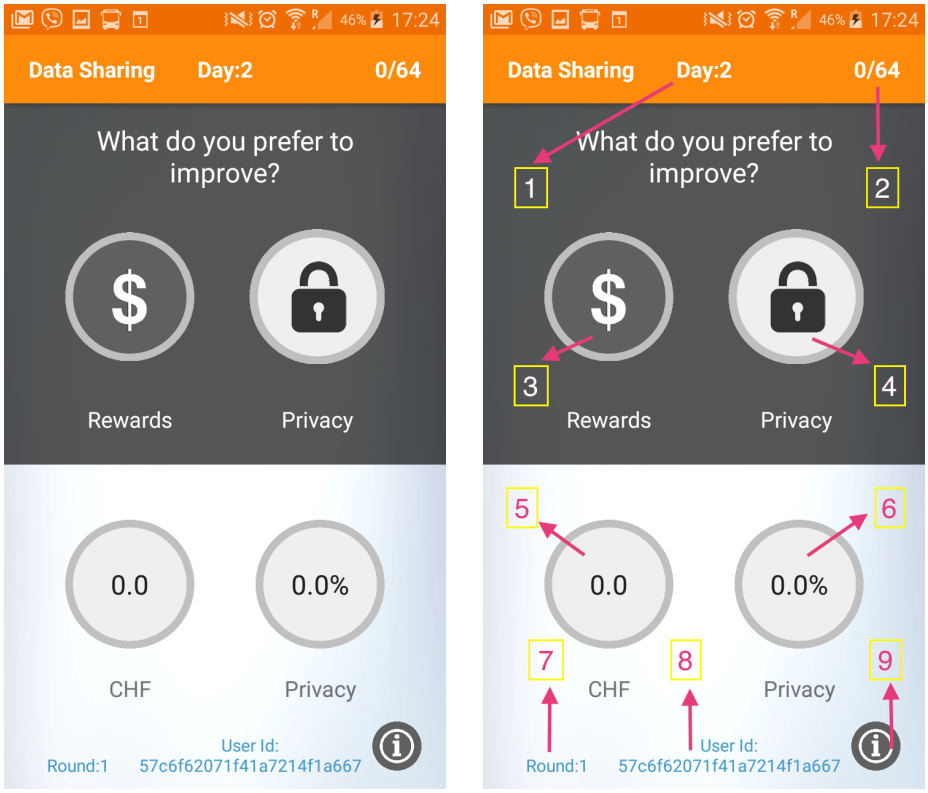


(a) Question screen

(b) Question screen with submit button

Figure 4.7: First Day Screen

Once all the questions have been answered, the user goes to the core phase of the experiment, which starts on day number two. In the experiment, day number one is the entry phase, the core phase is day number two and three.



(a) Original (b) Explained

Figure 4.8: Improvement screen

4.4 Core Phase

Once the entry phase is done, the user is presented with the screen shown in Figure 4.8. The *"i"* button at the bottom right of the screen denoted by the item number 9 shown in Figure 4.8b which is clickable. This takes the users to the FairDataShare portal. Figure 4.9 shows the homepage of the portal. Users can then click on the data generator registration section of the website where they can signup with their:

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

The unique participant identifier is located at the bottom of the application screen and is an alphanumeric sequence denoted by item number 8. If it is long pressed the user can select the identifier, then copy and paste it in the

textbox asking for the unique identifier in the portal. Figure 4.10 shows how the registration page looks like. The users can use this website to see all the data collected from them for all the mobile sensors. For more details about the FairDataShare portal refer to the Section 4.6.

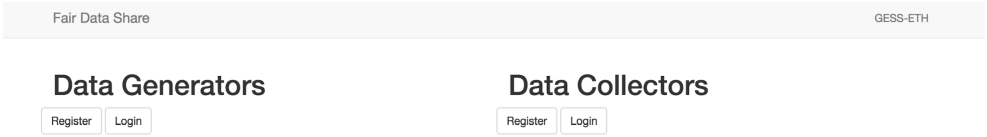


Figure 4.9: FairDataShare homepage

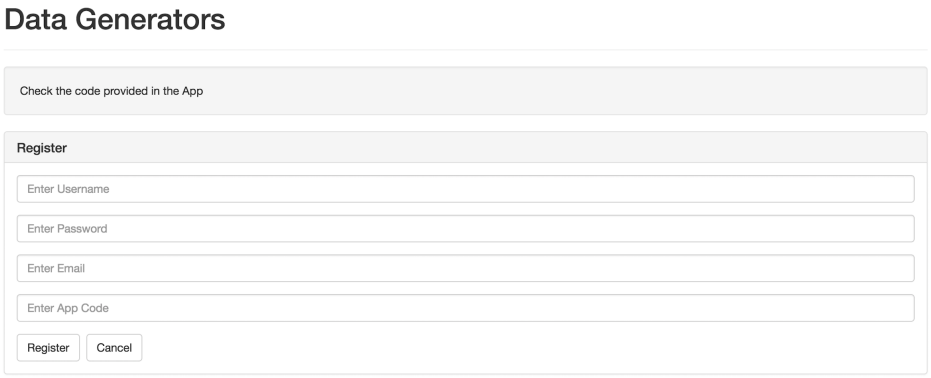
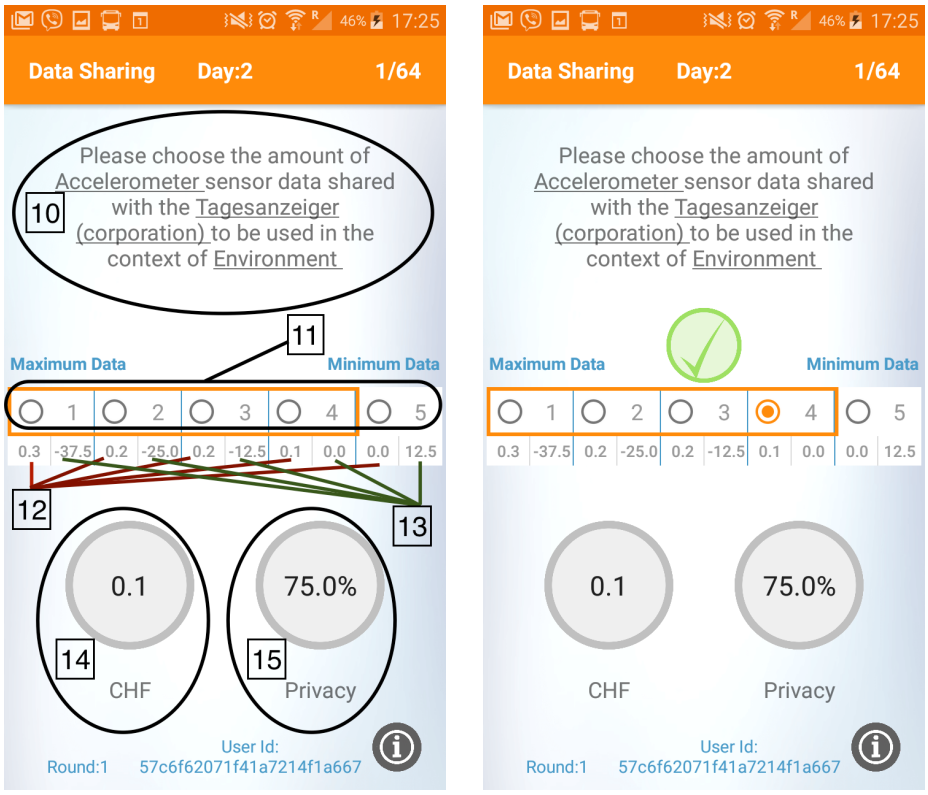


Figure 4.10: Data generators registration page

Users have the possibility to log onto the portal in 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.

In the task-bar, the user can see the bidding day number and the number of questions that have been answered from the total available shown by item number 1 and item number 2 in the Figure 4.8a. Day number one corresponds to the day where users answer questions with no incentives of any kind and was presented in the previous sub-section. The screen presented after the end of the entry phase is what is called the *"improvement screen"*. The button numbered 3 represents "improve privacy" and the button numbered 4 represents "improve credit" respectively. The items numbered 5 and 6 represent the credit and privacy percentage 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 obtained for trading data to the stakeholders.

The item numbered 7 is the number of times the user has answered all the data requests. The item numbered 2 is the number of questions the user has answered in the current round. Item number 1 indicates the experiment day number.



(a) Bidding screen

(b) Bidding screen with submit button

Figure 4.11: FairDataShare portal

There are a total of 64 data requests, hence after all the 64 have been answered, the number of questions answered is reset and the number of rounds 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 request 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.

4.4.1 Improve Privacy or Credit

The improvement screen shown in Figure 4.8 is where users can choose whether they would like to improve the privacy or the credit. The elements of this screen have been explained in the previous section 4.4. The improve credit button should be chosen if the user is interested in maximizing the amount of credit obtained. This calls an algorithm that uses the previous user answers to put forth a data request that can increase the credit to the maximum possible as explained in Section 5.2.3. The credit improvement button is represented by the item number 5. 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. It needs to be noted that the ultimate change in the privacy or credit metrics depends on the option chosen by the user for that data request. The privacy improvement button is represented by the number 6.

Scenario examples for each button are given in the next section after introducing the next screen in the application. For example, if a user chooses to improve the privacy, then the user clicks on the improve privacy button and gets a data request. If the user still chooses option 1 with maximum data sharing (least privacy) for this 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 without filtering the sensor information. Trading all data gives the user more credit, but decreases the privacy metric.

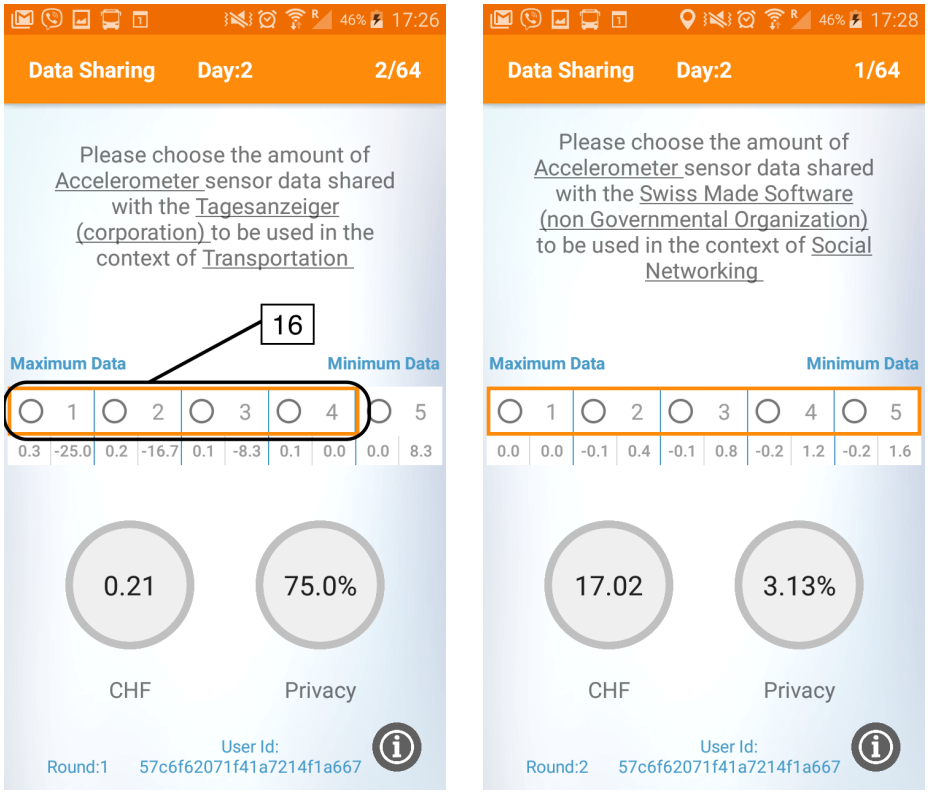
Similarly, if a user chooses to improve the credit obtainable, the user clicks on the improve credit button and gets a data request. Then the user chooses the option 5 with minimum data sharing (maximum privacy) which indicates that no data is traded for this request. This response counters the initial desire to improve the credit obtainable. Trading no data increases one's privacy, but does not increase the credit to the maximum. Therefore, an actual improvement in the chosen metric depends on the chosen improvement button and the choice of the appropriate option for that data request.

4.4.2 Answering Questions with Incentives

After choosing a metric to improve, a screen is presented as shown in Figure 4.11a. This screen is called the "bidding screen". This screen is very similar to the screen 4.8 presented in the entry phase, except that the user is aware

of the amount of privacy and credit obtained as indicated by items 14 and 15 respectively. Additionally, the user can see information about how the privacy and credit will increase or decrease for each privacy option of a data request. The items numbered 11 are the privacy options ranging from one to five.

The items numbered 12 are the improvement in privacy for each possible option of the current data request shown as item numbered 10. The items numbered 13 are the improvements in credit for each possible option of the current data request. Once the user decides on which options to choose according to how much data is intended to be traded, users can click on the options as explained in Section 4.2.4 and then click again on the green submit button that pops up shown in Figure 4.11b 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.



(a) Example 1 (b) Example 2

Figure 4.12: Recommendation box

Additionally, for every question there is an orange recommendation box surrounding some options. This recommendation is highlighted by item number 16 in Figure 4.12a. 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 4 to a data request and has clicked on improve credit, the system puts an orange box around options 1,2,3 and 4. Similarly, if the user clicked on improve privacy button, and the users previous answer was option 1, the system would recommend the options 1,2,3,4 and 5. Two examples of this are provided in Figures 4.12.

It needs to be noted that the orange box does not necessarily provide an improvement of the particular metric chosen, it is meant to indicate improvements compared to the last time the data request was answered.

4.5 Exit Phase

After the end of the core phase, the participants are asked to fill up a survey based on their experience in the experiment. Some questions are about the rewards received, the privacy and credit metrics, design of the application, and how the experiment was conducted. The purpose is to obtain feedback from users on their experience of participating in the experiment. Additionally, users are asked questions about the performance of the application and battery drain in order to make improvements to the mobile application. From this, problems faced by participants can be noted and improved upon. The survey [6] also shown in Appendix A is linked to the user using the unique identifier assigned in the application. Once the survey is filled, 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.

4.6 FairDataShare Web Portal

The FairDataShare portal [7] 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 and stakeholders can view mobile sensor data.

4.6.1 Data Generator's Portal

Once the users are registered as explained in Section 4.4, they can come back to the portal after a 24 hour period or later to view their mobile sensor data

[6] https://descil.eu.qualtrics.com/SE/?SID=SV_3POySMqNeOO6v5j
[7] http://fair-data-share.inn.ac/

collected in the server. The data portal login page is shown in Figure 4.13. 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 this is done, users will be redirected to the data collection page shown in Figure 4.14 with the following options in the task-bar to choose from:

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

Users can choose the sensor from the task-bar that they want to see by clicking on it. The data displayed includes the following columns :

1. Timestamp
2. Bidding day
3. Sensor values

Figures 4.15, 4.16, 4.17 and 4.18 show examples of the data that can be seen for the location, light, accelerometer and noise sensor.

Figure 4.13: Login page

Users first register as data generators as indicated in Section 4.3.4.

4.6.2 Stakeholder's Portal

For a stakeholder to view data, they need to register in the portal shown in Figure 4.9 by clicking register. Once that is done, the page in Figure 4.19 is shown asking for the following details :

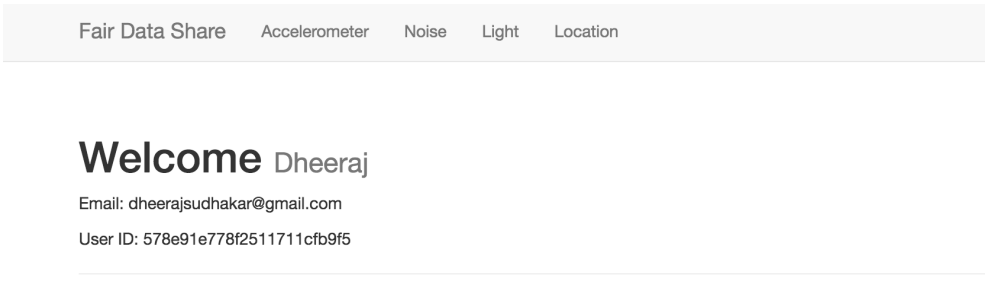


Figure 4.14: Welcome page

Figure 4.15: Location data

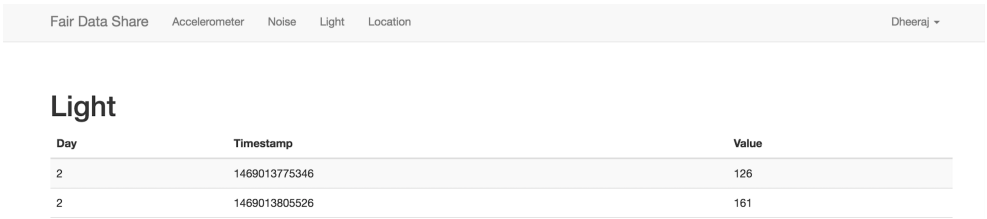


Figure 4.16: Light data

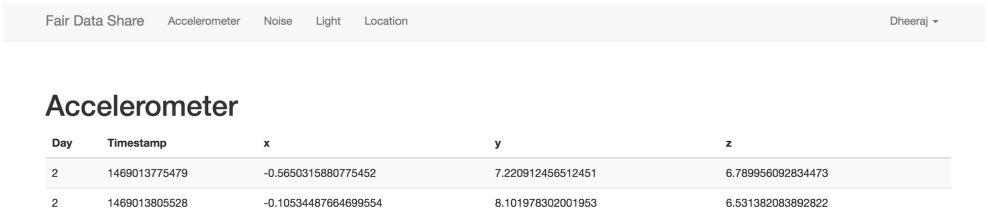


Figure 4.17: Accelerometer data

1. Company name
2. Email
3. Stakeholder category
4. Company website

Figure 4.18: Noise data

The stakeholder category is classified as follows:

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

Once these details have been filled in, the stakeholder can click on the register button. Once registered, the stakeholder can login as shown in Figure 4.20. When access is granted the stakeholder is redirected to the page shown in Figure 4.21. The stakeholder can choose from each of the available drop down lists :

1. A sensor
2. A context
3. An anonymous user
4. A bidding day number

Once this is entered, the stakeholder can see the data for that user with the privacy level decided by the anonymous user. If the stakeholder does not see any data, it means the user did not share data for that particular request. Stakeholders can view the sensor data in a similar fashion to users as seen in the previous figures. Data is available to the stakeholders 24 hours after the start of the core phase.

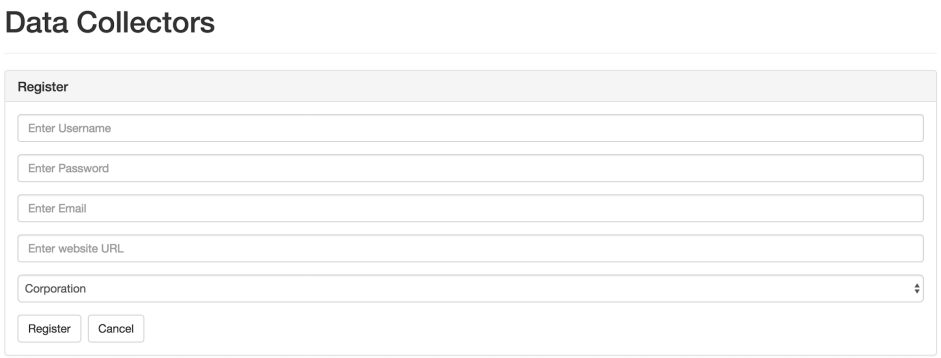


Figure 4.19: Registration page

Figure 4.20: Login page

Figure 4.21: Data collectors data retreiving page

Chapter 5

The Fair Data Share Mobile Application

This chapter illustrates the mobile application environment. First an overview is given, followed by a detailed explanation of the main components of the Android mobile application. This includes the architecture, database schemas and algorithms. Next, the server business logic and storage of the application is presented.

5.1 The Building Blocks

The following sections explain the integral parts of the server and client of the mobile application. A summary of the architecture is shown in Figure 5.1.

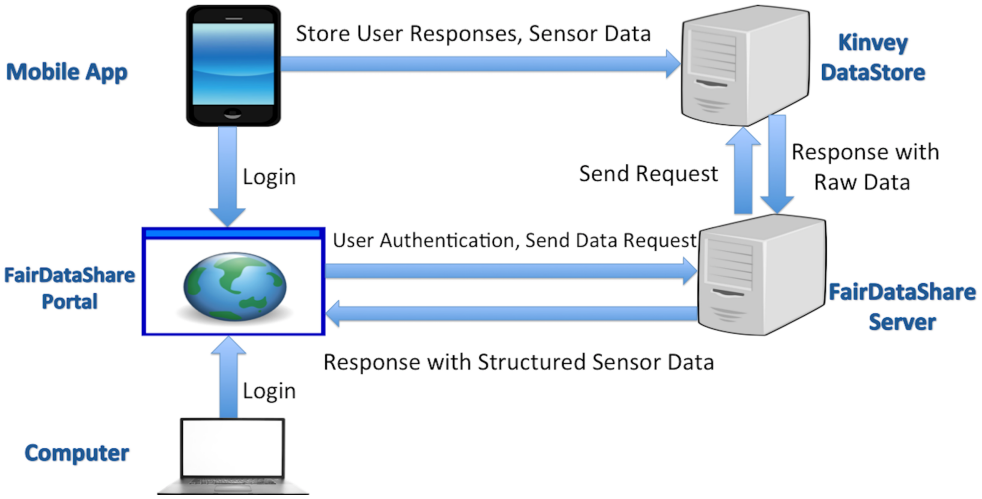


Figure 5.1: Conceptual diagram of mobile application architecture

Users have the possibility to login onto the FairDataShare Portal from their computer or the mobile application. Once the user is authenticated, the user requests are sent from the FairDataShare server to the Kinvey Data Store[1]. Kinvey in turn fetches the appropriate data and gives it back to the FairDataShare Server. This structures the data so it is readable, and sends the data to the user to see on the portal.

5.2 The Mobile Application

The mobile application is developed for the Android platform with phones having API above level 17[2]. Phones are assumed to have internet connectivity and sufficient storage space of at least 100 Mb. Below is an explanation of some of the tasks that take place in the application. A block diagram of the interaction of each of the components in the application is depicted in Figure 5.2.

5.2.1 Local Storage

The local storage is an integral part of the application. The database used is SQLite[3] and is the default database for the Android environment. Small sized unrelated data pieces are stored in preference files (as key value pairs), and larger related data are stored in the database. The following paragraphs explain each table present in this application followed by their function and schema. All tables explained here are pertaining to the user using the mobile application and not the server.

Figure 5.3a shows the `QUESTION_STORE` table schema. This table stores each possible data request with its features such as with the sensor *SENSOR*, stakeholder *STAKEHOLDER* and context *CONTEXT*. Each of these are represented by an integer, for example sensor 0 stands for accelerometer sensor. Each data request is accompanied by an unique question identifier *QID*, weight assigned *WEIGHT* and the cost assigned *COST*. This data is not sent to the server.

Figure 5.3b depicts the table `WHICH_ANSWERS`'s table schema. This stores the questions identifier *QID* of each data request that is answered by the user for each round. This is helpful while fetching data requests, so as not to fetch the request twice in the same round. This ensures that all questions are answered before answering them for a second time. This data is not sent to the server.

[1] https://kinvey.com

[2] https://developer.android.com/guide/topics/manifest/uses-sdk-element.html

[3] https://developer.android.com/reference/android/database/sqlite/package-summary.html

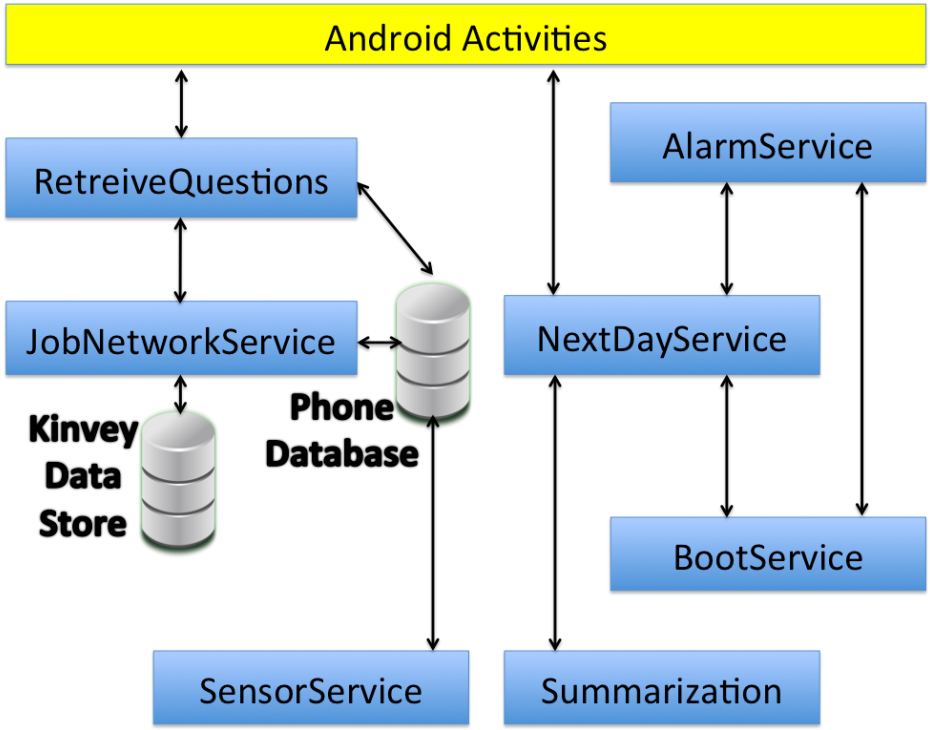


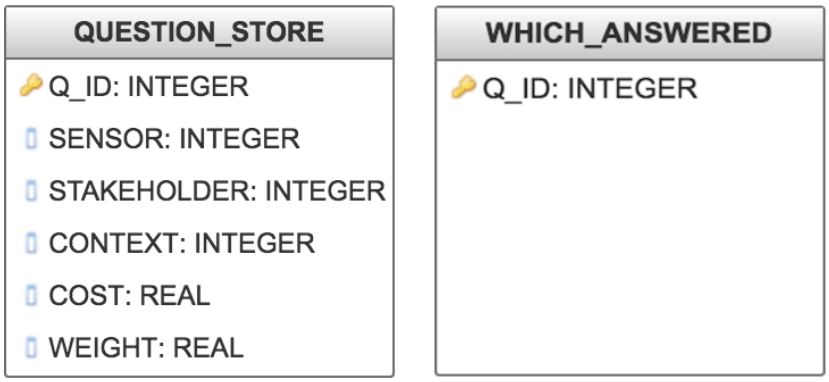
Figure 5.2: Interaction of components of mobile application

Figure 5.4a explains the schema of `STORE_ANSWERS` table. This table is used to store the data request identifier *QID* with the corresponding user responses *LEVEL*, along with the increase or decrease in credit obtained *COST_OBT*. The total cost is calculated by adding all the costs in this table. Similarly, the total privacy is calculated by averaging all the user responses stored in `STORE_ANSWERS` table. Only the most recent responses to a data request are stored in this table, since data requests can be answered more than once. The content of the table is not sent to the server.

Figure 5.4b denotes the schema of `STORE_POINTS` table. This table is used to store the credit and privacy obtained for each bidding day. This information is sent to the server as soon one bidding day is over.

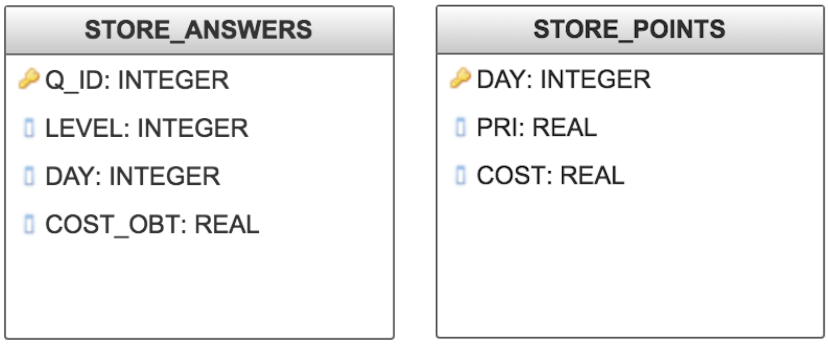
Figure 5.5 depicts the `USERRESPONSE_CACHE` table schema. This table stores a unique key *KEY* for each user response, followed by a flag *ISSENT*, which is 1 if the response is not sent to the server, and 0 if it is sent. The user response saved consists of the following entries :

1. User identifier
2. Timestamp of the response



(a) Table schema of `QUESTION_STORE` (b) Table schema of `WHICH_ANSWERS`

Figure 5.3: Table schemas



(a) Table schema of `STORE_ANSWERS` (b) Table schema of `STORE_POINTS`

Figure 5.4: Table schemas

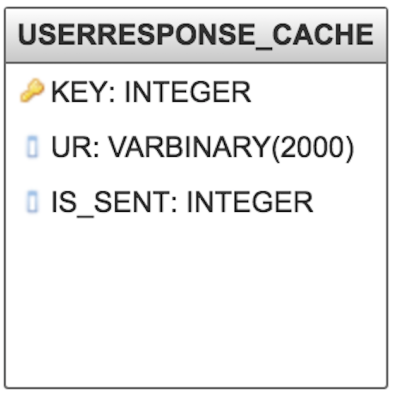


Figure 5.5: Table `USERRESPONSE_CACHE` schema

3. Sensor identifier
4. Stakeholder identifier
5. Context identifier
6. Privacy level response for this data request
7. Cost obtained for this data request
8. Current total privacy of the user
9. Current total credit of the user
10. Maximum obtainable credit for this data request in this round
11. Metric chosen to improve (Improve privacy or improve credit)

All of the above fields are packed into the field *ur* shown in Figure 5.5. The data in this table is sent to the server. Once the entry is sent to the server, the *ISSENT* field is changed to 0 and deleted locally. The unique keys *KEY* are useful for deleting sent entries. Figures 5.6 and 5.7 show the table schemas for data storage of the following sensors:

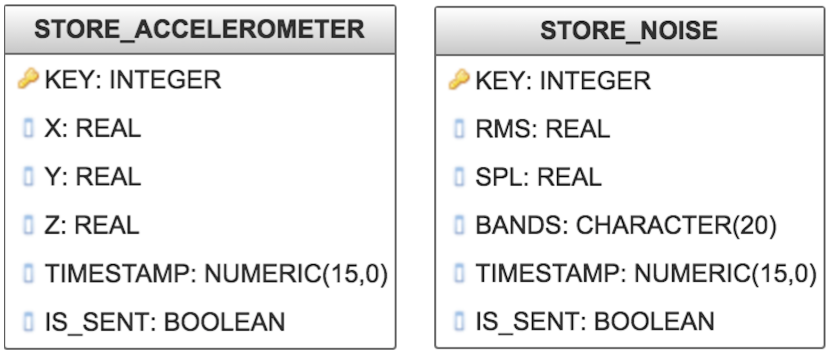
1. Accelerometer in the `STORE_ACCELEROMETER`
2. Noise in the `STORE_NOISE`
3. Location in the `STORE_LOCATION`
4. Light in the `STORE_LIGHT`

The general schema for all the sensor tables is the following :

1. *KEY* - Uniquely identifies each sensor entry
2. *TIMESTAMP* - The time when the sensor value is collected
3. *ISSENT* - Denotes whether the sensor entry is already sent to the server or not
4. The other columns are specific to each sensor and represent the actual sensor values collected

5.2.2 Alarms and Notifications

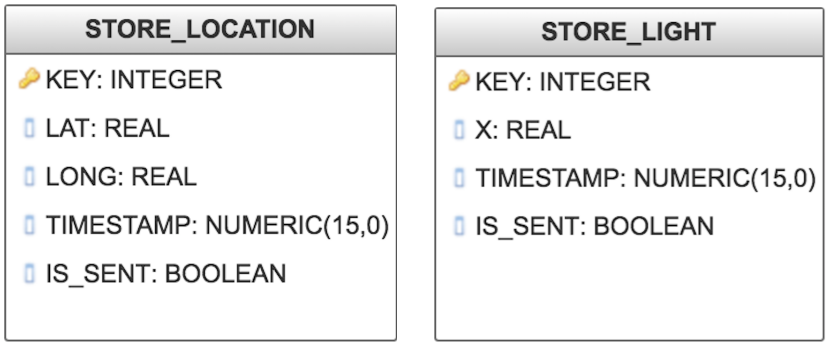
Every bidding day lasts 24 hours, and in this period the user answers data requests. After the completion of one bidding day, the system needs to be informed in a timely manner to perform some application critical functions such as going to the next bidding day. To inform the system of such an event, Android provides this functionality in the form of alarms.



(a) Table schema of `STORE_ACCELEROMETER`

(b) Table schema of `STORE_NOISE`

Figure 5.6: Table schemas for sensor data



(a) Table schema of `STORE_LOCATION`

(b) Table schema of `STORE_LIGHT`

Figure 5.7: Table schemas for sensor data

Alarms can be set to go off just once or repeatedly to trigger tasks. The built-in repetitive alarms[4] provided are not uniform across different Android versions because they are optimized to save battery. This can cause a delay of upto 24 hours in triggering the alarm. Hence, the application is programmed to set repeating alarms manually.

The first time the application opens, the alarm is set to ring in exactly 24 hours, but the timing changes when the phone is switched off. One of the conditions of the experiment is not to have the phone switched off at any time. Nevertheless, it is taken into account the scenario where the phone is kept switched off for a period of time. There are various things that can happen:

1. The phone is rebooted.

[4] https://developer.android.com/training/scheduling/alarms.html

2. The phone is switched off, during this time an alarm is missed.

3. The phone is switched off for a period greater than 24 hours. One or more alarms can be missed.

Once the phone is switched off, all alarms are erased from memory[5] and alarms do not execute when the phone is switched off. Hence, when the phone switches on the BootReceiver service of the application is triggered with pseudocode shown in Algorithm 1. This checks whether an alarm has been missed, 200 seconds are given for the phone to stabilize after booting before triggering tasks. Otherwise, a new alarm is set using the pseudocode shown in Algorithm 2. To set an alarm the time needed is the difference between the current time and when the alarm should be triggered. After that is calculated, the alarm is set. 86400 in the pseudocode indicates the number of seconds in a 24 hour period and is used to check if it has been more than 24 hours since an alarm has been triggered.

Algorithm 1 BootService Algorithm

```
1: procedure BootService
2:     now ← current timestamp
3:     i ← timestamp of last triggered alarm
4:     if now − i < 86400 then
5:         Call SetAlarmLater()
6:     else
7:         Set alarm in 200 seconds
```

The alarm Algorithm 2 is used to set the next alarm after the phone boots or when an alarm has just rung. It sets the alarm exactly 24 hours after the last alarm rang.

Algorithm 2 Alarm Algorithm

```
1: procedure SetAlarmLater
2:     now ← current timestamp
3:     i ← timestamp of last triggered alarm
4:     latertime ← i + 86400
5:     latergap ← latertime − now
6:     Set Alarm in latergap seconds
```

[5] https://developer.android.com/reference/android/app/AlarmManager.html

Going to the Next Data Sharing Day

Once the alarm rings, it marks the end of a bidding day. Once a bidding day ends, a number of tasks need to be executed and for this the NextDayService is triggered, which is described in pseudocode shown in Algorithm 5. Firstly, the privacy and credit is sent to the server and stored locally in the `STORE_POINTS` table. *Privacy* which is the total privacy obtained, *Credit* is the total credit obtained, *Round* which is the number of times the user answers all the questions and *CurrentQuestion* which is the current question the user is answering is all reset to zero. The *Day* corresponds to the current bidding day which increments by one to denote the next bidding day.

Algorithm 3 NextDayService Algorithm

```
1:  procedure NextDayService
2:      Store Privacy, Credit, Day in STOREPOINTS
3:      Send Privacy, Credit, Day to Server
4:      Privacy, Credit, Round, CurrentQuestion ← 0
5:      Day ← Day + 1
6:      Store current time
7:      Call Summarization()
8:      if Day > End then
9:          End experiment
10:     else
11:         Update user interface elements
```

The current time of executing the alarm is saved in case the phone is rebooted or switched off as mentioned in Section 5.2.2. After that, the sensor data which is saved locally needs to be summarized, the corresponding method is called and is explained in pseudocode shown in 4. A final check is done to see if the experiment is done and the user interface is updated accordingly. This means either the various metrics on the improvement and bidding screens (which ever is currently active) are updated, or the end of experiment screen is shown.

5.2.3 Fetching Data Requests

A data request needs to be fetched from the mobile phone database in two scenarios :

1. After a question is answered in the first bidding day (entry phase)
2. After the privacy or credit improvement button is clicked (core phase)

In the first bidding day, once a data request is answered the next is fetched sequentially from the mobile phone database. This requires knowing the

current data request number and fetching the next data request from table QUESTION_STORE. For the other bidding days, fetching of the data requests depends on the improvement button chosen. According to the choice, the following is done:

1. **Improve Privacy** - Obtain the data request from table STORE_ANSWERS where the user can maximize the privacy metric
2. **Improve Credit** - Obtain the data request from table STORE_ANSWERS where the user can maximize the cost metric

In addition to sending the data request to the user interface, it is needed to show how choosing each option of the data request will affect the total privacy and total credit metrics. To do this for the total cost, the computation $last - possible$ is output, where *last* stands for the credit obtained the last time the data request is answered. *possible* stands for the maximum amount of credit that can be obtained for this option (each data request has five privacy options 3.2.6). The possible total cost changes are shown under the options. For more detail on how credits are split among options in a data request refer to Section 4.2.4.

Every option of a data request has an associated percentage of data that is shared as described in Section 3.2.6. According to the percentage of data shared, the total privacy is calculated for each possible option. The difference between the current privacy and each potential privacy (the privacy if the user chooses to click a particular option) is calculated and indicated under each option. This gives an indication to the user as to how each option will affect the privacy and cost metrics.

5.2.4 Recording User Choices

Figure 5.5 describes the table USERRESPONSE_CACHE. Each time a user enters a response to a data request, all the fields mentioned in Section 5.2.1 are recorded and stored in a class object. This object is transformed into a byte array. When the JobNetworkService described in Section 5.2.6 is called, the byte array is converted back into a class object and sent to the Kinvey Data Store.

5.2.5 Sensor Data Collection and Summarization

Sensor data is collected from the following sensors :

1. Accelerometer sensor
2. Noise sensor
3. Location sensor (GPS)
4. Light sensor

A sensor service is triggered when the application is installed and is stopped when the experiment is over. The sensor service collects data from all sensors every 30 seconds and stores it in the appropriate tables mentioned in Section 5.2.1. At the end of a bidding day, sensor data needs to be summarized according to the privacy level chosen by the user. This starts by first finding out the lowest privacy level for each sensor. Privacy levels range from one to five, that is from the lowest to highest privacy levels. Using this level, summarization is done as shown in pseudocode 4. Every privacy level corresponds to an action:

1. 1- Send all data to the server
2. 2- Send 75% of the data
3. 3- Send 50% of the data
4. 4- Send 25% of the data
5. 5- Do not send any data

Initially all the sensor data has a field *ISSENT* with value of zero. Data to be sent to the server is set with $ISSENT = 1$, and all others that have value $ISSENT = 0$ are ignored. In the pseudocode 4, first the privacy level to which summarization should be performed is obtained. If the privacy level is 5, then no records are sent to the server and hence all data remains marked with $ISSENT = 0$. If the privacy level is 1, all records should be sent to the server and they are all hence marked with $ISSENT = 1$ without any summarization performed. Similarly, if the privacy level is 2, 3 or 4 the amount of records to be sent to the server are 75%, 50% and 25% respectively. To do this the, amount of records(mobile sensor data) sent to the server is 3 records for every 4 records, 1 record for every 2 records and 1 records for every 4 records respectively and they are marked with $ISSENT = 1$.

5.2.6 Server Synchronization

User responses and sensor data are sent to the server. This is done periodically every 5000 seconds in order to empty the space on the phone whenever internet is available. It is triggered first when the application is started for the first time. Data is fetched from the tables in the database. Data with fields marked as $ISSENT = 1$ is data that is ready and that has not been sent to the server. Such data is sent, and when an acknowledgement is received from the server, this data is deleted from the table.

Algorithm 4 Summarization Algorithm

```
1:  procedure SUMMARIZATION
2:      for each sensor do
3:          Fetch sensor data from sensor table
4:          level ← Fetch user privacy level
5:          if level ← 1 then
6:              Set all ISSENT ← 1
7:          else if level ← 2 then
8:              for 3 out of every 4 records do
9:                  ISSENT ← 1
10:         else if level ← 3 then
11:             for 1 out of every 2 records do
12:                 ISSENT ← 1
13:         else if level ← 4 then
14:             for 1 out of every 4 records do
15:                 ISSENT ← 1
16:         Delete all entries with ISSENT ← 0
17:         Update Database
```

Algorithm 5 JobNetworkService Algorithm

```
1:  procedure NETWORKSERVICE
2:      Fecth data from USERRESPONSECACHE
3:      for each record do
4:          if ISSENT == 1 then
5:              Send record to Server
6:              if SUCCESS then
7:                  Delete record
8:      for each sensor do
9:          Fecth data from sensor table
10:         for each record do
11:             if ISSENT == 1 then
12:                 Send record to Server
13:                 if SUCCESS then
14:                     Delete record
```

5.3 The Server

5.3.1 Kinvey Data Storage

Kinvey[6] is a mobile backend as a service which provides a platform for mobile phones to link applications to a backend cloud storage. For the purpose of this application, the backend has is used to store data and for some business logic implementations in javascript.

Security

All communications from the application to the server are encrypted using TLS/SSL encryption[7], to communicate with the backend service. This is automatically provided and done by the Kinvey SDK.

Collection Store

Locally, all information is stored in SQLite which is a relational database. The database used by the Kinvey Data Store is MongoDB. The schema of the collections in Kinvey are the same as described in Section 5.2.1 for the mobile phone databases. When the user starts the application, general personal information is entered as explained in Section 5.2.1. This data is stored in the collection USERINFORMATION with the screenshots shown in the Figures A.1 and A.2 and schema in Figure 5.8.

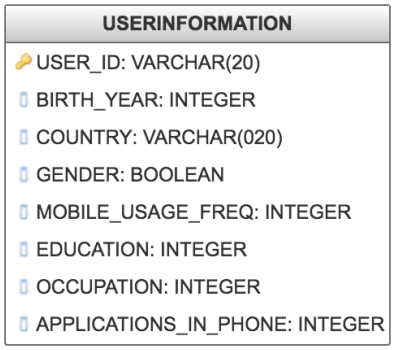
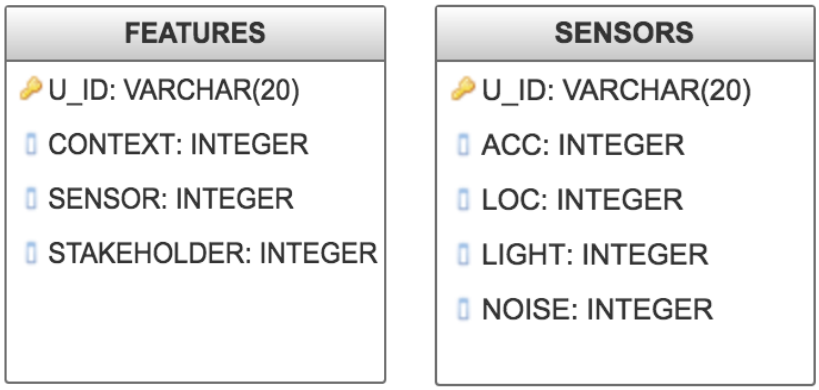


Figure 5.8: Table `USERINFORMATION` schema

Once this is done, users have to categorize the various features, sensors, stakeholders and then the various contexts. This information is sent to the server in collections named FEATURES, SENSORS, STAKEHOLDERS and CONTEXTS. Screenshots are shown in Figures A.3, A.4, A.5 and A.6 respectively and the schemas in Figure 5.9. User responses are stored in the

[6] http://kinvey.com/
[7] Kinvey white paper : KINVEY CLOUD SERVICE: SECURITY OVERVIEW 2014

collection UserResponse shown in Figures A.7 and A.8 and their schema is shown in Section 5.2.1.



(a) Table schema of `FEATURES`

(b) Table schema of `SENSORS`

(c) Table schema of `STAKEHOLDERS`

(d) Table schema of `CONTEXTS`

Figure 5.9: Table schemas for categorizations

The sensor data sent by the JobNetworkService is stored in collections named after the sensors themselves. The screenshots of the tables are shown in Figures A.9, A.10, A.12 and A.11 and their schema is shown in Section 5.2.1.

To keep track of all the existing users in the experiment, the collection USERS stores all unique user identification strings of participants. The screenshot is shown in Figure A.13.

Finally, the collection SCORE shown in Figure A.14 and the schema is shown in Figure 5.4b stores the total privacy and total credit obtained by the user for each bidding day.

Bussiness Logic

Most of the bussiness logic used for the FairDataShare portal is present in Kinvey. There are two main scripts stored in Kinvey:

1. Script to find the privacy preferences of users
2. Script to perform data summarization

The stakeholders make a request for data on the FairDataShare portal giving the following details:

1. Bidding day number
2. Anonymous user
3. Sensor
4. Context

Given this input and the category of the stakeholder (which is known from their registration), we look into the UserResponse Collection trying to find the most recent record that fits these criteria and extract the privacy level (this is the level to which sensor data should be summarized for this request).

Once the privacy level is known, summarization on user data is performed. Data is sent from the user phones with a certain summarization level from each sensor in order to avoid sending sensor data several times to the server.

If the summarization level with which data is collected from the user is lower than the privacy level needed, further summarization needs to be performed. When further summarization needs to be done, the difference between the privacy level and summarization level times hundred is the percentage of data that needs to be removed for this request. The pseudocode is shown in Algorithm 6.

5.3.2 FairDataShare Web Portal

The FairDataShare portal makes use of a server at ETH Zurich other than the Kinvey Data Store to store the usernames, passwords of the users and the stakeholders in a collection. The database technology used is MongoDB. 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 as described in section 5.3.1. The webpage is constructed using simple Html and css. All screenshots of the portal including detailed information are provided in Chapter 4.

Algorithm 6 Server Summarization Algorithm

```
1:  procedure SUMMARIZATION
2:      data ← sensor data from collection
3:      sl ← level with which data was summarized on the phone
4:      pl ← level to which data needs to be summarized
5:      if sl == pl then
6:          Return data
7:      else
8:          skip ← sl − pl + 1
9:          for every skip number records out of 4 do
10:             Delete record from data
11:     Return data to portal
```

Chapter 6

Experimental Findings

The following chapter gives an overview of the data obtained from the survey, which was conducted before running the experiment. Later, an overview of the data obtained from the experiment is explained along with some feedback received from the participants.

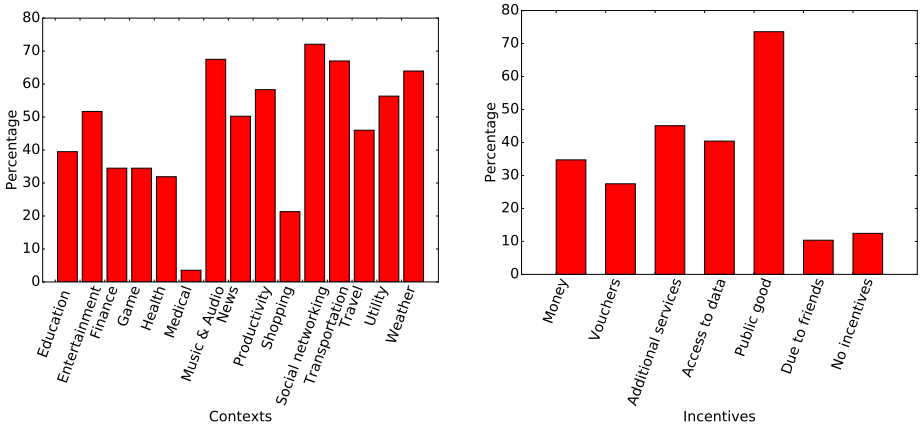
6.1 Findings from the Pre-Survey

The survey has 199 participants. Participants are not given any incentives to participate. After filtering out spurious and half-filled entries, 189 entries are used for the data analysis. In the following paragraphs, information obtained in the survey is introduced.

Out of the total participants 63.64% are male and 36.36% are female. The mean year of birth is found to be 1985. The demographics of the participants is illustrated in Table A.1. On the education level 2.53% have not completed high school, 9.60% have completed high school, 5.05% have gone to some college, 28.79% have obtained their bachelors degree, 39.90% have obtained their masters degrees and 14.14% have obtained their PHDs. About the employment of the participants, 51.52% are full time employees, 6.06% are part time employed, 6.06% are unemployed and looking for work, 1.52% are unemployed and not looking for work, 0.51% are retired and 41.92% are students. None of the participants are disabled.

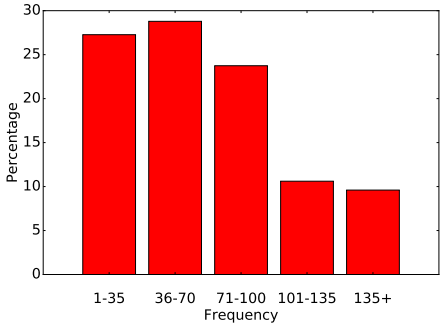
Figure 6.1c shows the frequency of mobile usage among the population. It is observed that the majority of the people use their phones 36-70 times a day.

Figure 6.1a depicts the percentage of users who have different applications on their mobile phones. Is it observed that the applications which are installed the most are music and audio, social networking, transportation and



(a) Applications in the mobile phone

(b) Incentives



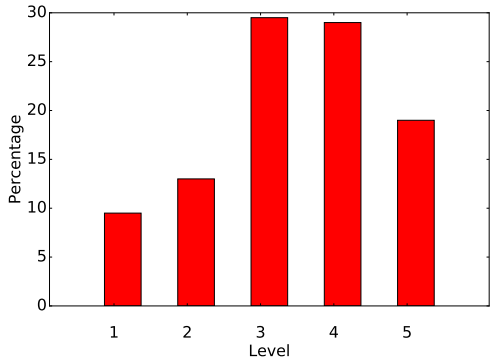
(c) Frequency of mobile phone usage

Figure 6.1: Figures of applications in the mobile phone, incentives and frequency of mobile phone usage

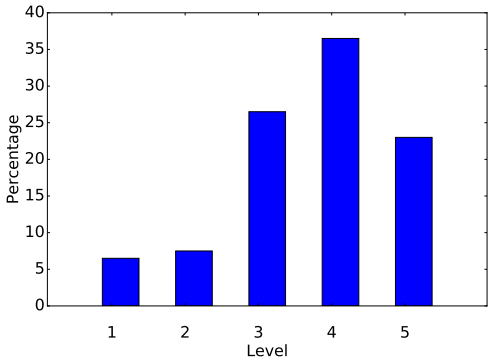
weather applications with 67.51%, 72.08%, 67.01% and 63.96% of users having them installed respectively.

Figure 6.1b shows the percentage of users who would give data for each incentive indicated. As it can be observed, 73.58% of users would give data for public good which is the most chosen option. The least popular reasons to share data are *"Due to friends"* and *"No incentives"*. 34.72% of users would accept *"Money"* as an incentive, 27.46% would accept *"Vouchers"*, 45.08% would accept *"Additional services"*, 40.41% would accept *"Free access to data"*.

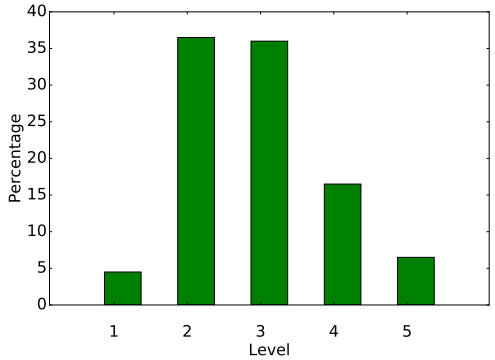
Figure 6.2a depicts the percentage of people who have different levels of concern for the privacy of their mobile sensor data. Level 1 corresponds to *"Not at all concerned"* and level 5 to *"Extremely concerned"*. As seen, 77.5% of users are concerned to a level of 3 and above and 22.5% of users are concerned to a level of 1 and 2 together. This shows that most users are at least moderately concerned about the privacy of their mobile sensor data.



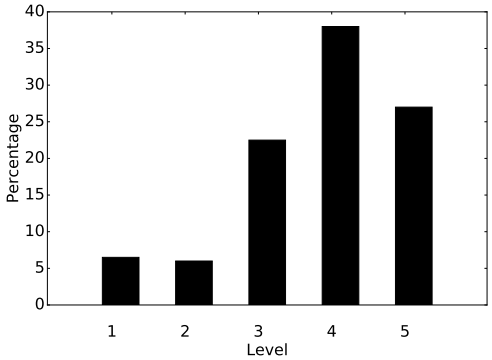
(a) Level of concern of privacy for mobile sensor data



(b) Importance of sensor type for data sharing



(c) Importance of stakeholder type for data sharing

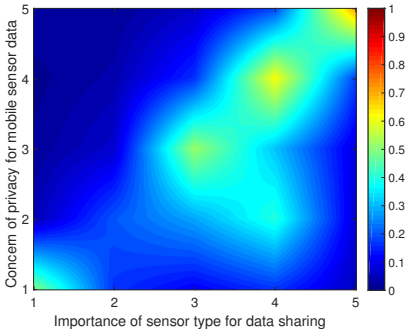


(d) Importance of context type for data sharing

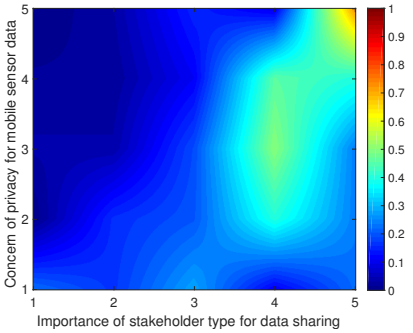
Figure 6.2: Figures for mobile privacy concern, importance of sensors, importance of stakeholders and importance of contexts for data sharing with levels from 1 - *"Not at all concerned"* to 5 - *"Extremely concerned"*

Figure 6.2b depicts the importance of the sensor type for whom mobile sensor data is shared. Level 1 corresponds to *"Not at all important"* and level 5 to *"Extremely important"*. As seen, 36.5% of users care to a level of 4 and 86% of users care to a level of 3 and above. Similarly, Figure 6.2c depicts the importance of the stakeholder type to which data is shared. 36% and 36.50% of users find the stakeholder important to a level of 4 and 5 respectively and 89% of users find the importance of stakeholder to a level 3 and above. Figure 6.2d shows the importance of the context of application, for which mobile sensor data is shared. 38% of users find the importance of the context of application to a level of 4 and 87.5% of users find the importance of the context of application to be of level 3 and above.

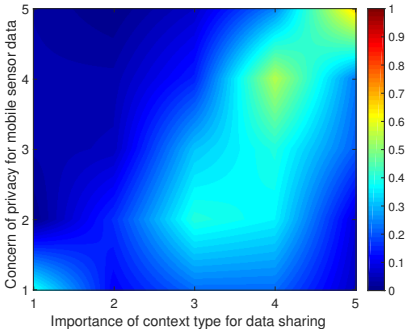
Figure 6.3a shows the probability of the *"importance of sensor type in data*



(a) Given the level of concern of mobile sensor data the probability of the level of importance of sensors



(b) Given the level of concern of mobile sensor data the probability of the level of importance of stakeholders



(c) Given the level of concern of mobile sensor data the probability of the level of importance of contexts

Figure 6.3: Given the level of concern of mobile sensor data the probability of the level of importance of sensors, stakeholders and contexts

sharing" for each level of the *"concern of privacy for mobile sensor data"*. Level 1 corresponds to *"Not at all important"* and level 5 to *"Extremely important"*. Probability values closer to 1 indicate a higher likelihood of a particular level of *"importance of sensor type in data sharing"* for each level of *"concern of privacy for mobile sensor data"* and vice versa. As observed, users with levels 1,3,4 and 5 of *"concern of privacy for mobile sensor data"* have a higher probability to have levels 1,3,4 and 5 respectively for the *"importance of sensor type in data sharing"*. For level 2 of the *"importance of sensor type in data sharing"*, the probability is the highest at level 4 for the *"importance of sensor type in data sharing"* but is also spread across levels 2 and 3.

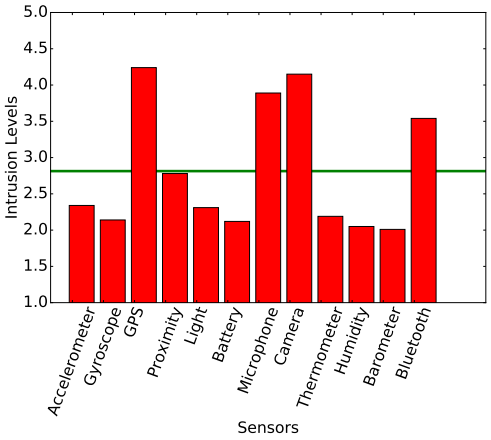
Figure 6.3b shows the probability of the *"importance of stakeholder type in data sharing"* for each level of the *"concern of privacy for mobile sensor data"*. Level 1 corresponds to *"Not at all important"* and level 5 to *"Extremely important"*.

Probability values closer to 1 indicate a higher likelihood of a particular level of *"importance of stakeholder type in data sharing"* for each level of *"concern of privacy for mobile sensor data"* and vice versa. As observed, users with levels 2,3,4 and 5 of *"concern of privacy for mobile sensor data"* have a higher probability to have levels 4,4,4 and 5 respectively for the *"importance of stakeholder type in data sharing"*. This means that for levels 2,3,4 and 5 of *"concern of privacy for mobile sensor data"* users view stakeholders as important for data sharing. For level 1 of *"concern of privacy for mobile sensor data"*, probabilities of levels of *"importance of stakeholder type in data sharing"* are spread around level 3.

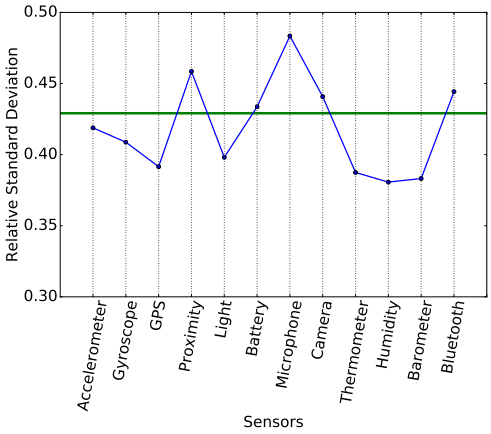
Figure 6.3c shows the probability of the *"importance of context type in data sharing"* for each level of the *"concern of privacy for mobile sensor data"*. Level 1 corresponds to *"Not at all important"* and level 5 to *"Extremely important"*. Probability values closer to 1 indicate a higher likelihood of a particular level of *"importance of context type in data sharing"* for each level of *"concern of privacy for mobile sensor data"* and vice versa. As observed, users with levels 1,3,4 and 5 of *"concern of privacy for mobile sensor data"* have a higher likelihood to have levels of 1,4,4 and 5 for the *"importance of context type in data sharing"*. For level 2 of *"concern of privacy for mobile sensor data"*, probabilities of levels of *"importance of context type in data sharing"* are spread around level 3 and 4.

Figure 6.4a indicates the level of privacy intrusion for all sensors. Privacy intrusion level 1 corresponds to *"Very low privacy intrusion"* and level 5 to *"Very high privacy intrusion"*. As seen in the figure, the location, camera, microphone and bluetooth sensors are found to be most privacy intrusive with levels of 4.23, 4.14, 3.95 and 3.52 respectively. The gyroscope, battery, humidity and barometer are found to be least privacy intrusive with privacy intrusion levels of 2.13, 2.11, 2.04 and 2.00 respectively. The accelerometer, proximity, light and thermometer are found moderately intrusive with privacy intrusion levels of 2.33, 2.79, 2.31 and 2.19 respectively. Figure 6.4b shows the relative standard deviation around the mean privacy intrusion of each sensor. This indicates the variability for opinions of privacy intrusion of sensors in each level. Among all sensors the privacy intrusion level spread around the mean is the highest for the proximity, battery, microphone, camera and bluetooth sensors.

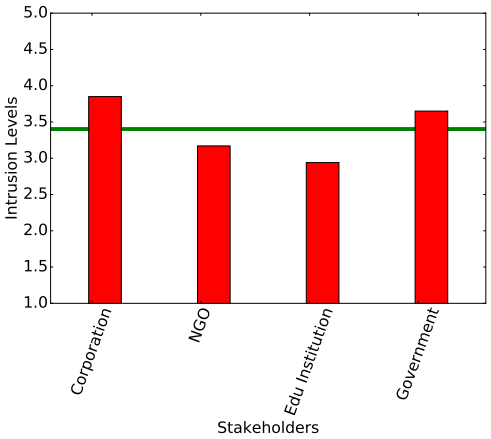
Figure 6.4c indicates the level of privacy intrusion for all the stakeholders. Privacy intrusion level 1 corresponds to *"Very low privacy intrusion"* and level 5 to *"Very high privacy intrusion"*. As seen in the figure, the most intrusive stakeholders are corporation and government with privacy intrusion levels of 3.84 and 3.64 respectively. Stakeholders NGO and educational institution have privacy intrusion levels of 3.17 and 2.94 respectively and have privacy intrusion levels lower than the mean depicted in the figure by the horizontal



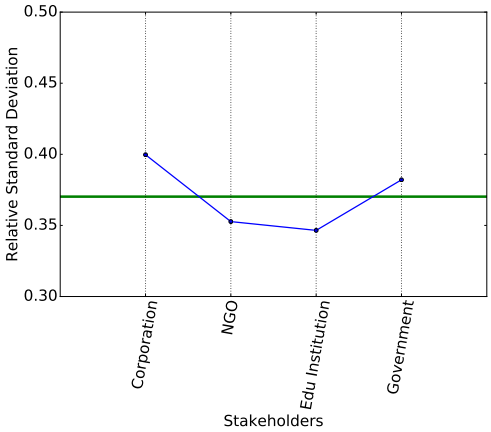
(a) Average intrusion of sensors



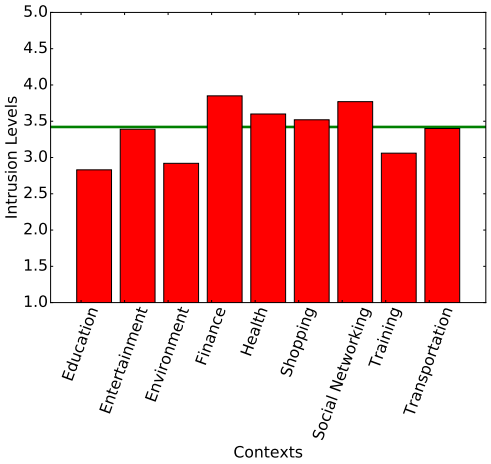
(b) Standard deviation of sensor intrusions



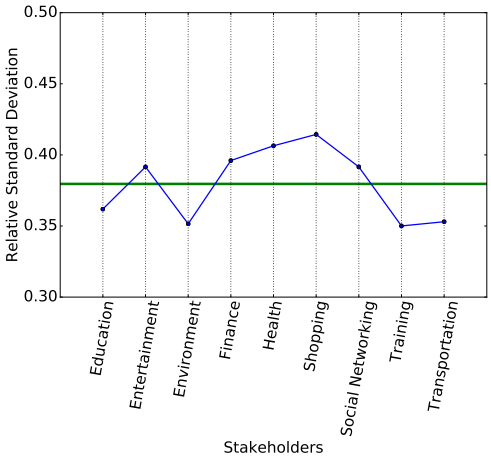
(c) Average intrusion of stakeholders



(d) Standard deviation of stakeholder intrusions



(e) Average intrusion of contexts



(f) Standard deviation of context intrusions

Figure 6.4: Average and relative standard deviation for intrusion of sensors, stakeholders and contexts

line. Figure 6.4d shows the relative standard deviation around the mean privacy intrusion level for each stakeholder. This indicates the variability for opinions of privacy intrusion of stakeholders in each level. As observed, stakeholder corporation and stakeholder government have a higher spread of privacy intrusion levels around the mean compared to the NGO and educational institution.

Figure 6.4e indicates the level of privacy intrusion for all the contexts of applications. Privacy intrusion level 1 corresponds to *"Very low privacy intrusion"* and level 5 to *"Very high privacy intrusion"*. As seen in the figure, the most privacy intrusive contexts are health, finance, shopping and social networking with levels of 3.60, 3.85, 3.50 and 3.75 respectively. Contexts whose privacy intrusion levels are less than the average are training, environment, entertainment, transportation and education with privacy intrusion levels of 3.06, 2.92, 3.39, 3.38 and 2.83 respectively. Figure 6.4f shows the relative standard deviation of the privacy intrusion levels for every context of application. This indicates the variability for opinions of privacy intrusion of contexts in each level. As observed, there is a higher spread of privacy intrusion levels around the mean for contexts entertainment, finance, health, shopping and social networking.

6.2 Findings from the Experiment

An emulation of the social experiment explained in Chapter 4 was held with 9 participants. This was done in order to test the working of the mobile application and receive user feedback before the actual experiment, that will be officially held with the ETH Decision Science Laboratory.

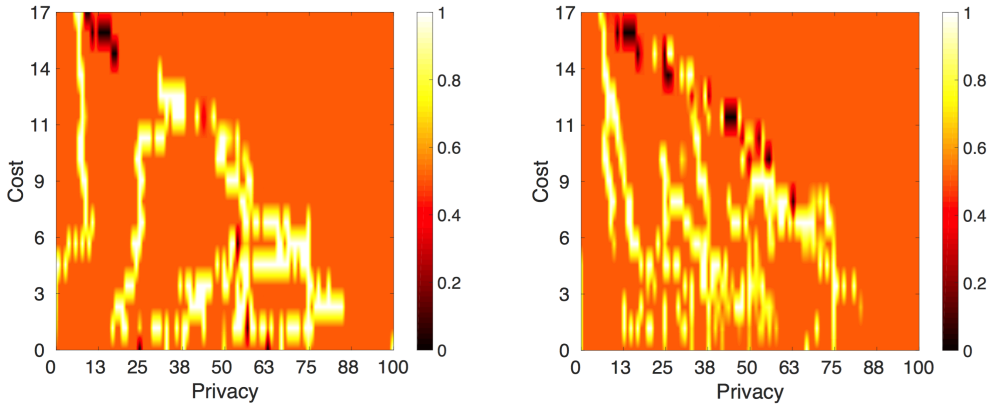
The experiment was held for a period of 3 days. Out of the total number of days, 3 participants did not answer requests on day 2 and day 3. Participants are not monetarily incentivized during the emulation of the experiment, but are asked to think of the incentives indicated in the application as real incentives they receive. This might cause a deviation from the ideal scenario where users are paid for their participation while examining the results. The mobile application ran successfully on all participating phones even after being switched off. All data was successfully recorded on the server. Using the data collected on the server, the relationship between data sharing and incentives is examined.

Table 6.1 shows the average privacy and rewards obtained by the users for each day of the experiment. The privacy and rewards are not shown to users in day 1 of the experiment, but is calculated and saved in the background for the purpose of analysis. It is observed that the privacy metric is higher on the first day than on day 2 and 3. Furthermore it is also observed that the rewards obtained are higher on day 2 and day 3 than day 1. It can

Table 6.1: Average Scores Obtained in the Experiment

Day	Privacy	Rewards
1	56.43%	7.75
2	40.17%	9.29
3	47.97%	8.80

be inferred that users have decreased their privacy in order to obtain more rewards on day 2 and 3.



(a) User click on the improve privacy or credit button probabilities

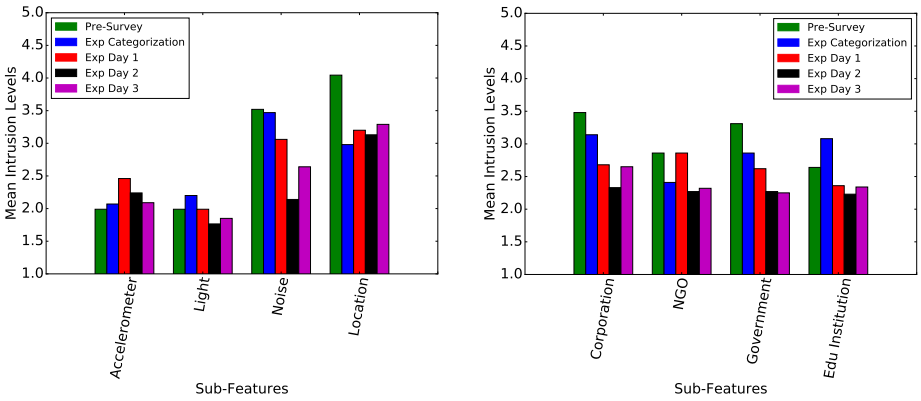
(b) Improvement in cost metric or privacy metric probabilities

Figure 6.5: Privacy and cost metric

Figure 6.5a depicts the of probability of the user clicking on the improve privacy or improve credit button for every possible cost and privacy metric value obtained in the experiment from all participants. In other words, this figure depicts the probability of whether the user wants to obtain more rewards or more privacy. Probability values closer to 1 depicts that the user probability of clicking on the improve credit button is highest. Similarly, probability value of 0 depicts that the user's likelihood of clicking on the improve privacy button is highest. The figure depicts that users have higher probabilities of clicking on the improve privacy button when they have a high cost metric and a low privacy metric.

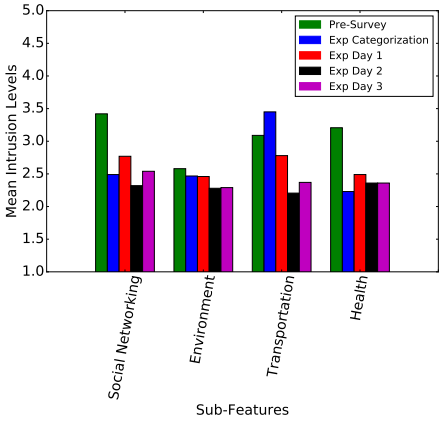
Figure 6.5b depicts the probability of an increment in the cost or privacy metric for every possible cost and privacy metric value obtained in the experiment. In other words, this figure depicts the probability that the user clicks in an option for a data request which increases the cost metric or increases

the privacy metric. Probability values closer to 1 depicts the user's likelihood of choosing an option for a data request that increases the cost metric. Similarly, probability value of 0 depicts the user's probability of choosing an option for a data request that increases the privacy metric. When Figures 6.5a and 6.5b are observed together, it is seen that when the probability that users click on the improve credit button is high, users clicking on an option for a data request that improves their cost metric is also high. It is also observed that in some areas where the probability of clicking on the improve button is high, Figure 6.5b shows that users have a high probability of clicking on an option for a data request that improves their privacy metric. It could be due to the fact that users have more intentions to improve their cost metric as seen before, but the ultimate decision could possibly lie on the data request presented whether they click on an option that increases the cost or privacy metric.



(a) Sensors

(b) Stakeholders



(c) Contexts

Figure 6.6: Mean privacy intrusion levels

Figure 6.6 depicts the mean of the privacy intrusion levels assigned to the sub-features in the pre-survey and experiment categorization. It also depicts how much data was shared for each sub-feature during the experiment on day 1, day 2 and day 3. As it can be seen, there is a difference in the mean privacy intrusion levels assigned during the pre-survey and during the experiment categorization for some sub-features. This is perhaps due to the low number of participants in the experiment.

Additionally, it is observed that for all sub-features, the privacy level chosen for data requests during day 1 of the experiment is much higher than on day 2 and day 3. This shows that users have chosen to improve their cost metric on day 2 and day 3.

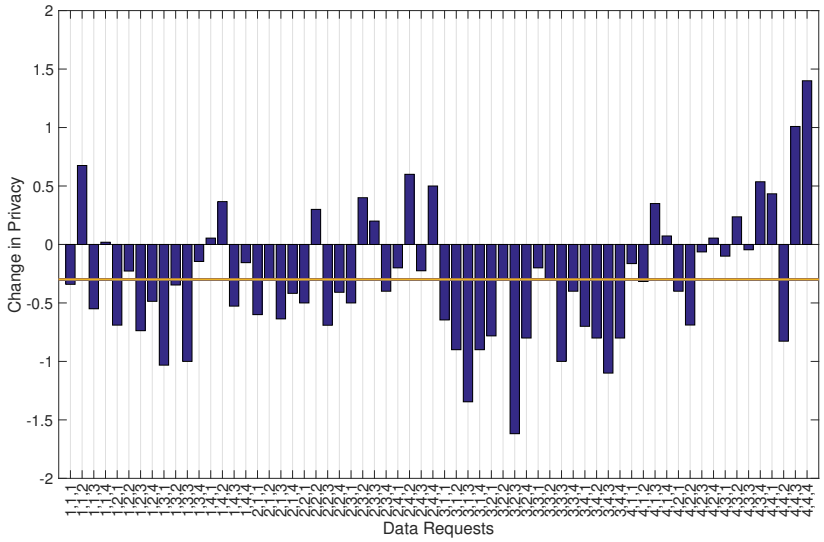


Figure 6.7: Gain in privacy between day 2 and day 1

Figure 6.7 depicts the gain in privacy for every data request in the experiment between day 2 and day 1. This was obtained by subtracting the responses to data requests on day 2 from the responses to data requests on day 1. If the bars are on the positive side, it indicates that users chose a higher privacy option for that data request on day 2 than on day 1. If the bars are on the negative side, it means that users chose a lower privacy option for that data request on day 2 than on day 1.

It is observed that there are more bars on the negative side of the graph, indicating that users in general have chosen to decrease their privacy to obtain more rewards. The horizontal line in the graph indicates the mean gain in privacy between day 2 and day 1 for all data requests. The line is on the negative side indicating that users have overall chosen to decrease their

privacy and opt to obtain more rewards. The average privacy option chosen for data requests on day 1 is 2.95 and on day 2 is 2.65. The decrease in the average privacy option chosen is 0.297 from day 1 to day 2.

There are some data requests for which the user has not decreased the privacy such as the data request involving the location sensor, stakeholder education and context transportation. This is perhaps because location sensor is categorized with a privacy intrusion level of 3.42 which is the second most intrusive sensor. Additionally, the context transportation is also categorized with a privacy intrusion level of 3.65 which is the most intrusive context. The stakeholder education is categorized with a privacy intrusion level of 3.29. From the Figure 6.6 it can be seen that for day 1, users have already given more data on average for educational institutions. Hence it could be that they are not incentivized enough to give even more data for this stakeholder than they already have. Putting all the points above together could be the reason why data request (4,4,4) has a bar on the positive side of the figure. Similar reasonings can be applied to other data request with an increase in privacy rather than decrease.

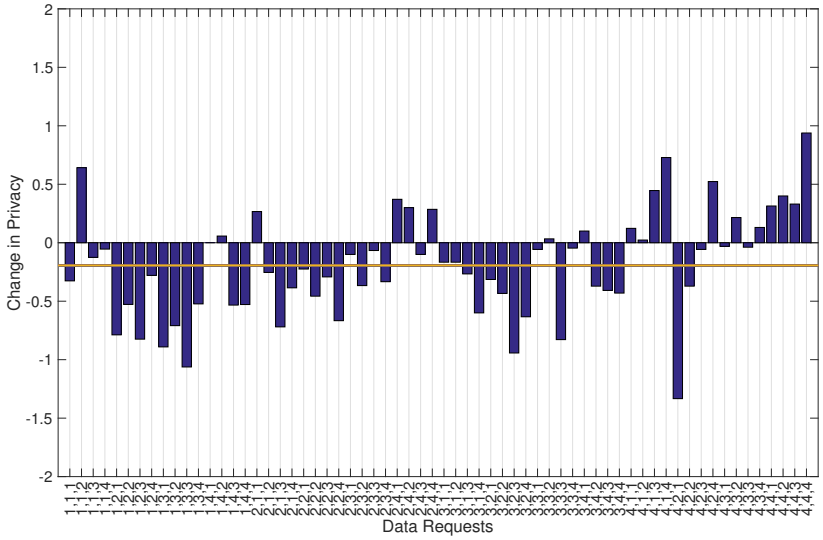


Figure 6.8: Gain in privacy between day 3 and day 1

Figure 6.8 depicts the gain in privacy for every data request in the experiment between day 3 and day 1. This was obtained by subtracting the responses to data requests on day 3 from the responses to data requests on day 1. If the bars are on the positive side, it indicates that users chose a higher privacy option for that data request on day 3 than on day 1. If the bars are on the negative side, it means that users chose a lower privacy

option for that data request on day 3 than on day 1.

It is observed that there are more bars on the negative side than the positive side hence this means that users have shared more data on day 3 than day 1 which means they chose to improve their cost metric over the privacy metric. The horizontal line shown in the figure depicts the average gain in privacy which is on the negative side. This shows that overall they have decreased their privacy level for data requests on day 3 compared to day 1. The average privacy option chosen for data requests on day 1 is 2.95 and on day 3 is 2.75. The decrease in the average privacy option chosen is 0.202 from day 1 to day 3.There are some data requests for which users have shared less data than on day 1, this could be due to the fact that users are not incentivized enough for these data requests.

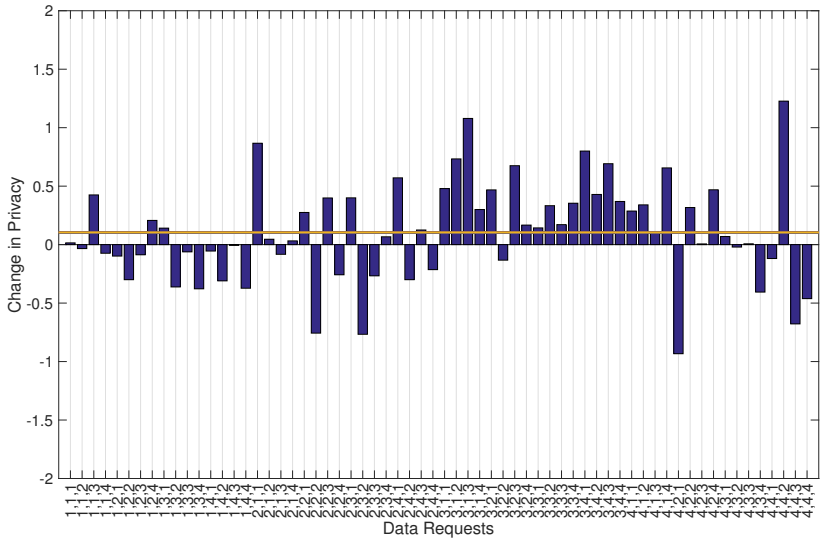


Figure 6.9: Gain in privacy between day 3 and day 2

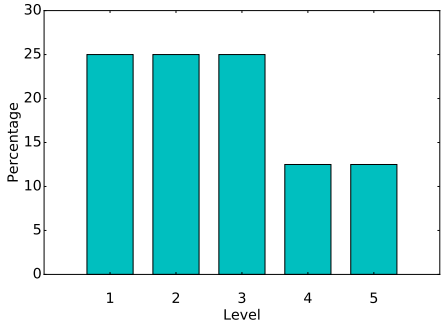
Figure 6.9 depicts the gain in privacy for every data request in the experiment between day 3 and day 2. This was obtained by subtracting the responses to data requests on day 3 from the responses to data requests on day 2. If the bars are on the positive side, it indicates that users chose a higher privacy option for that data request on day 3 than on day 2. If the bars are on the negative side, it means that users chose a lower privacy option for that data request on day 3 than on day 2.

As it is observed, the horizontal line which indicates the average gain in privacy is on the positive side which means that users have on average increased their privacy on day 3 compared to day 2. Additionally, it can be seen that there are more bars on the positive side. This could be due to the

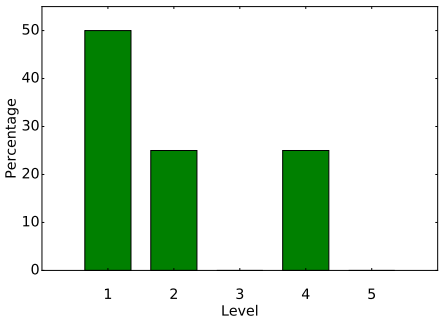
fact that users expected more rewards on day 3, or that they became more privacy aware as they used the application due to the privacy metric. The average privacy option chosen for data requests on day 3 is 2.75 and on day 2 is 2.65. The increase in the average privacy option chosen is 0.095 from day 2 to day 3.

6.3 Findings from the Exit Survey

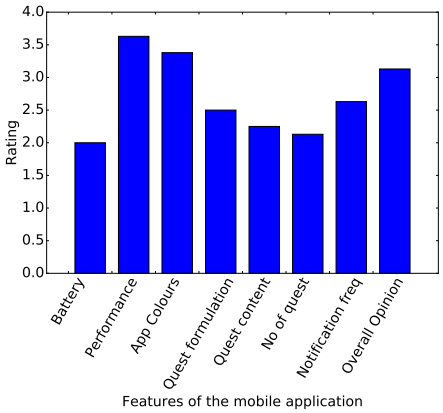
Eight fully filled entries are recorded from the exit survey. No incentives are awarded to participate in this survey. The following paragraphs present the findings obtained from the survey.



(a) User evaluation of application quality



(b) User evaluation of the ease of use of the application



(c) User evaluation of various features of the application

Figure 6.10: General user evaluation of the application with levels from 1 - ”Extremely good” to 5 - ”Extremely bad”

Figure 6.10a depicts the user ratings for the quality of the application. 1

stands for ”Extremely good” and 5 stands for ”Extremely bad”. As observed in the figure, 75% of users think of the application quality to a level of neither good or bad and above (level 1,2 and 3).

Figure 6.10b depicts the user ratings for the ease of use of the application. 1 stands for ”Extremely easy” and 5 stands for ”Extremely difficult”. It is observed that 75% of users find the application to be at least somewhat easy to use (level 1 and 2).

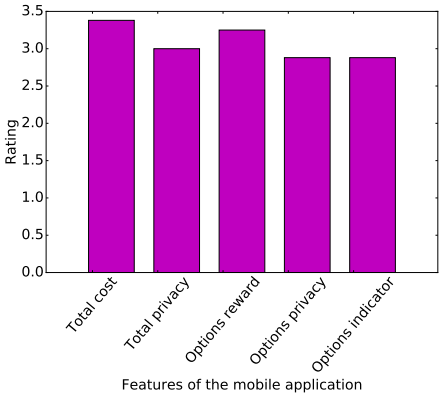
Figure 6.10c[1] depicts the user rating of the battery life, performance and speed, colors of the application, formulation of the data requests, content of the data requests, number of data requests, frequency of notifications and their overall opinion. A rating of 1 indicates that the user is ”Extremely dissatisfied” and a rating of 5 means that the user is ”Extremely satisfied”. It is seen that users are most satisfied with the performance, application colours and with the application overall with ratings of 3.63, 3.38 and 3.13 respectively. Users are less satisfied with the battery life, questions formulation, content of the questions and the number of questions with ratings of 2, 2.5, 2.25 and 2.13 respectively.

Figure 6.11a depicts the user ratings for the comprehension of the total cost, total privacy, rewards for each option of a data request, privacy for each option of a data request and the indicator of options (orange recommendation box). A rating of 1 indicates that the user is ”Extremely dissatisfied” and a rating of 5 means that the user is ”Extremely satisfied”. It is seen that users have a better understanding of the total cost and rewards for each option of a data request with ratings of 3.38 and 3.25 respectively than the total privacy, privacy for each option of a data request and the indicator of options (orange recommendation box) with ratings of 2.88, 3.13 and 3 respectively.

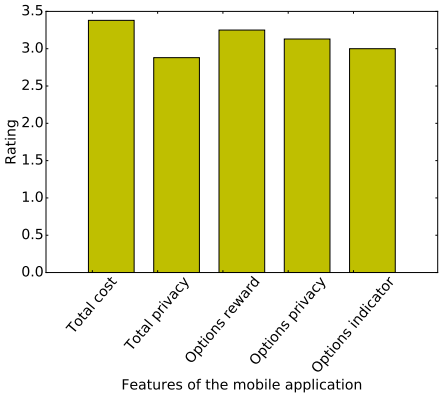
Figure 6.11b depicts the user ratings for the usefulness of the total cost, total privacy, rewards for each option of a data request, privacy options for a data request and the indicator of options (orange recommendation box). A rating of 1 indicates that the user is ”Extremely dissatisfied” and a rating of 5 means that the user is ”Extremely satisfied”. It is seen that users find the total cost and rewards for each option of a data request most useful with ratings of 3.38 and 3.25 respectively than the total privacy, privacy for each option of a data request and the indicator of options (orange recommendation box) with usefulness ratings of 3, 2.88 and 2.88 respectively.

Figure 6.11c depicts the user responses to ”if the experiment makes them more aware about the privacy of their data”, and ”if the privacy-preservation of their data deserves the sacrifice of their rewards”. A rating of 1 indicates ”Definitely not” and a rating of 5 means ”Definitely yes”. Experiment mak-

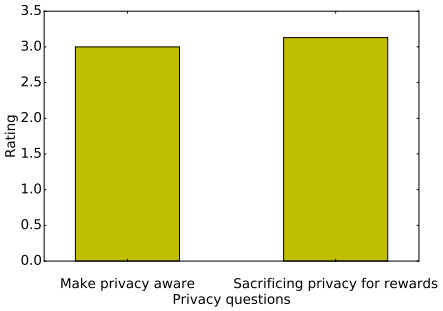
[1]Using the feedback obtained from this question, improvements are made in the user interface of the mobile application



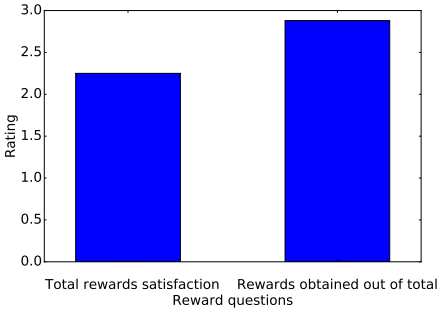
(a) User evaluation of the comprehension of features



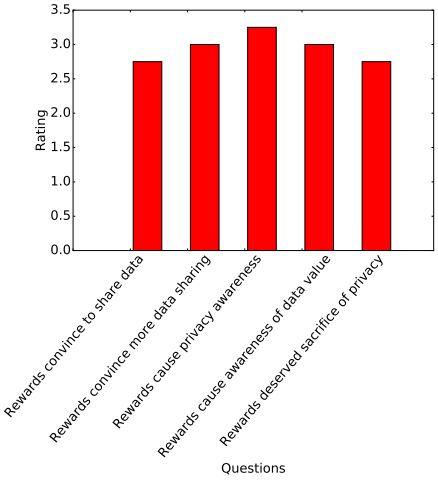
(b) User evaluation of the usefulness of features



(c) User evaluation of the privacy



(d) User evaluation of rewards satisfaction



(e) User evaluation of the cost

Figure 6.11: User evaluation of cost and privacy

ing users more aware about the privacy of their data receives a rating of 3 and the privacy-preservation of data deserving the sacrifice of their rewards receives a rating of 3.13. This shows that users are willing to sacrifice their privacy.

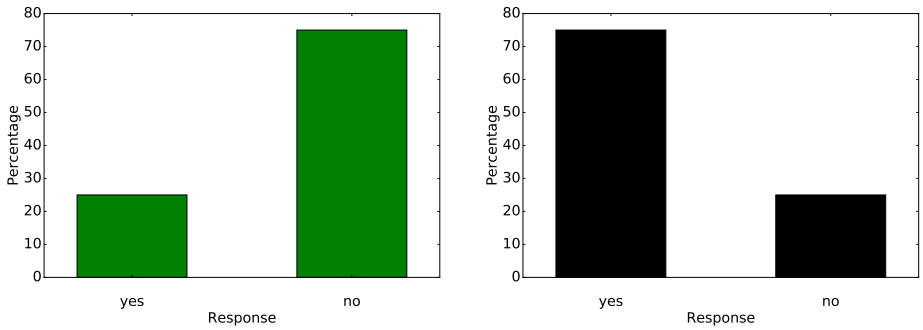
Figure 6.11d depicts the user responses to the satisfaction of users to the total obtainable rewards (30 CHF) and the satisfaction of users to the rewards they obtained out of the total obtainable rewards. A rating of 1 indicates ”Definitely not” and a rating of 5 means ”Definitely yes”. It is seen that users are not satisfied with the total rewards of 30 CHF for 2 bidding days indicated by a rating of 2.25. Users indicate that they are more satisfied about the rewards they obtained out of the total possible rewards with a rating of 2.88.

Figure 6.11e depicts the user responses to if rewards convince users to share data, if rewards convince users to share more data than without rewards, if rewards make users more privacy aware, if rewards make users aware about the value of their data and if rewards deserve the sacrifice of the privacy of their data. A rating of 1 indicates ”Definitely not” and a rating of 5 means ”Definitely yes”. Users agree more that rewards convince them to share more of their data, rewards make them more aware of the privacy of their data and that rewards make them aware about the value of their data to a rating of 3, 3.25 and 3. Users gave a rating of 2.75 and 2.75 to rewards convincing them to share data and rewards deserving the sacrifice of the privacy of their data.

Figure 6.12a shows whether users wanted to drop out of the experiment at any point of time. As it is seen, 75% of users said ”no” and 25% of users said ”yes”. Some of the reasons stated to drop out of the experiment are *”battery drain”* and *”it got boring after a point”*.

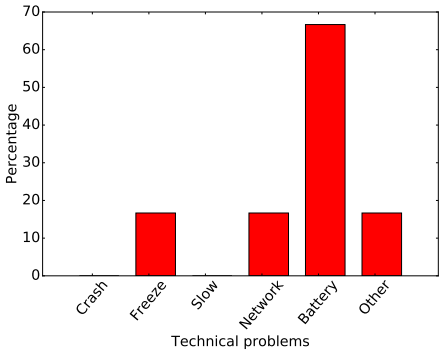
Figure 6.12b shows whether users visited the FairDataShare portal at any point of time in order to see what sensor data has been collected from them. As observed, 75% of users said ”yes” and 25% said ”no”.

Figure 6.12c shows the percentage of users who suffered from technical problems. Users did not face problems of the application crashing or problems of the application being slow. The problems 66.67% of users faced is the drain of their battery charge. This is due to the fact that sensor data is constantly collected from them as a background service in the application and this is a battery intensive task. 16.67% of users faced problems of network connectivity. There is no indication that the application is the cause of the network connectivity problem. 16.67% faced problems of the application interface freezing. Finally, 16.67% of users faced other problems. When looking in to this further, the ”other” problem is due the fact that the privacy and cost metrics do not change as the user re-answered data requests. This is due to the fact that if the user repeatedly answers the same data requests



(a) User experiment dropout

(b) User FairDataShare portal visit



(c) User evaluation of the technical problems faced

Figure 6.12: User responses to experiment dropout, FairDataShare portal visit and technical problems faced

again with the same responses, the metrics do not change. The metrics only change when the user re-answers a data request with a difference response, otherwise same rewards and privacy are obtained for this data request and this does not change the privacy and cost metrics.

Chapter 7

Conclusion and Future Work

A computational model is introduced to be able to assign incentives to mobile sensor data requests. A user profile is formed and using this profile rewards are assigned to data requests in a personalized manner. The model additionally includes the option for users to choose data requests according to the metric (privacy or cost) they wish to improve. This model is incorporated in a mobile application and launched on Google Play Store. The application shows data requests where stakeholders can request users to share their mobile sensor data for a particular context. Users can then choose from five different privacy options ranging from giving all of their data (level 1) to giving none at all (level 5).

All inputs to the application and sensor data shared by users are recorded and sent to the server by mobile application background services. Additionally, a pre-survey is deployed to understand the perception of users on mobile sensors, stakeholders and contexts. This is also used to reduce the number of sensors, stakeholders and contexts in data requests and choose which ones to examine in more detail. Users can access the FairDataShare website to see the data that has been collected from them. Furthermore, stakeholders can also view the data shared by users with the appropriate privacy level chosen by the users themselves. This makes data sharing a more transparent process where users are requested for data and have the ability to choose from various privacy options rather than an all or none option. Furthermore, users can view the data that is collected from them on the FairDataShare portal which makes the whole data sharing process more transparent and gives users more control of their data.

From the data obtained in the pre survey, it is seen that 77.5% of users are at least moderately concerned about their mobile sensor data. It is also observed that users have lower motivation to share their data for no incentives, which means that incentives play an important role in the data sharing process. Additionally, it is revealed by users that money is not the only incen-

tive that would be accepted. The GPS, camera, microphone and bluetooth sensors are found to be privacy intrusive. Corporation and government stakeholders are found to be privacy intrusive. Finance, health, shopping and social networking contexts are found to be most privacy intrusive.

An emulation of the social experiment with 9 participants where initially no incentives are awarded, is held and it is seen that the mobile application and the FairDataShare web portal are fully functional. Additionally, it is observed that there is an increase in data sharing on the days where incentives are given compared to the days where no incentives are given. It is also observed that for most cost and privacy metric values, users tend to click on the "improve credit" button. It is also seen that even though users click on the "improve credit" button, the ultimate decision to improve the cost or privacy metric depends on the data request itself.

From the exit survey, it is seen that 75% of users find the quality of the application to be moderate (not good or bad) and above. 75% of users find the application easy to use. Users rate the performance of the application to a level of 3.63 on 5, but rate the number of questions and battery life to be 2.13 and 2 respectively, which means that users find that there are too many questions in the experiment and their phone battery is affected. Users find the total cost and each option reward of a data request as the most useful and comprehensible features of the application. Additionally, users are not as satisfied with the total available budget of 30 CHF but they are more satisfied with the amount of rewards obtained out of the total which means they could serve their goal.

It is also seen that users are willing to sacrifice their privacy for rewards. Another outcome is that the experiment makes users aware about the privacy of their data. The rewards also cause privacy awareness and also cause awareness about the value of users data. During the experiment 25% of users wanted to drop out at some point whereas 75% of users were still interested to continue. 75% of users visited the FairDataShare portal. 66.67% of users also faced problems with battery drain.

In the future, the social experiment will be held with a larger sample population who are awarded the incentives indicated in the mobile application with the help of the ETH Decision Laboratory. More work can be done to analyse the data obtained to find inter-relationships between features and relate the data to the user information provided. Deeper comparisons of the pre survey and experiment data can also be done. Furthermore, the model could incorporate machine learning algorithms to predict the sequence of user choices based on previous ones to assign appropriate incentives for each data request. Additionally, problems mentioned by users during the exit survey will be addressed. Another interesting addition would be to increase the duration of the data collection and see the behavioural change in

users.

Appendix A

Appendix

birth_year	check_mobile_frequency	country	education	education_backgrou…	education_level	employment_status	entertainment	finance
1994	3	"France"	0	0	3	6	0	1
1924	3	"Arménie"	0	0	4	4	0	0

Figure A.1: Screenshot of Collection UserInformation Part 1

gender	health	medical	mobile_sensor_priv…	music	user_id	navigation	news	productivity	shopping	social_network
2	1	0	3	1	"57a8f8f1848532cf7…	0	0	0	1	1
2	0	0	3	0	"579a148f352257bc0…	0	0	0	0	0

Figure A.2: Screenshot of Collection UserInformation Part 2

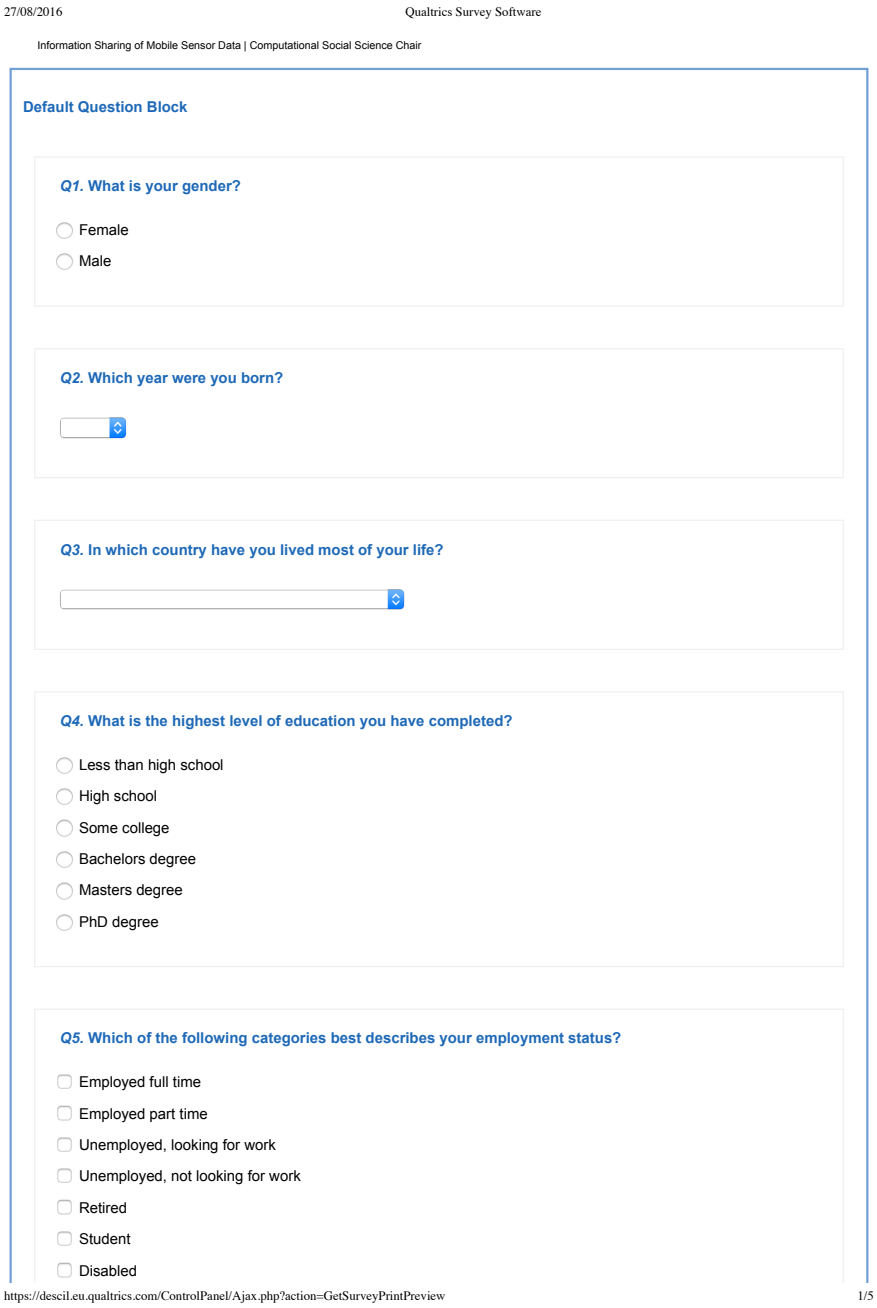
user_id	context	data_collector	sensor
"57a8f8f1848532cf7…	1	3	5
"579a148f352257bc0…	2	3	3

Figure A.3: Screenshot of Collection Features

user_id	acc	gps	light	noise
"57a8f8f1848532cf76b0836f"	3	5	2	4
"57a8f8f1848532cf76b0836f"	3	5	2	4

Figure A.4: Screenshot of Collection Sensors

Pre-Survey



27/08/2016 Qualtrics Survey Software

Information Sharing of Mobile Sensor Data | Computational Social Science Chair

Default Question Block

Q1. **What is your gender?**

- ○ Female
- ○ Male

Q2. **Which year were you born?**

Q3. **In which country have you lived most of your life?**

Q4. **What is the highest level of education you have completed?**

- ○ Less than high school
- ○ High school
- ○ Some college
- ○ Bachelors degree
- ○ Masters degree
- ○ PhD degree

Q5. **Which of the following categories best describes your employment status?**

- ☐ Employed full time
- ☐ Employed part time
- ☐ Unemployed, looking for work
- ☐ Unemployed, not looking for work
- ☐ Retired
- ☐ Student
- ☐ Disabled

https://descil.eu.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview 1/5

Q6.
Which types of apps do you usually have on your smartphone?

- [ ] Education apps (Exam preparations, study-aids, vocabulary, language learning, etc.)
- [ ] Entertainment apps (Streaming video, movies, TV, interactive entertainment, etc.)
- [ ] Finance apps (Banking, payment, ATM finders, financial news, insurance, taxes, portfolio/trading, tip calculators, etc.)
- [ ] Game apps (puzzles, charades, etc.)
- [ ] Health & Fitness apps (Personal fitness, workout tracking, diet and nutritional tips, health & safety, etc.)
- [ ] Medical apps (Drug & clinical references, handbooks for health-care providers, medical journals, etc.)
- [ ] Music & Audio apps (Music services, radios, music players, etc.)
- [ ] News apps (local news, national headlines, technology announcements, etc.)
- [ ] Productivity apps (calendar, to do list, price checker, etc.)
- [ ] Shopping apps (Online shopping, auctions, coupons, price comparison, grocery lists, product reviews, etc.)
- [ ] Social networking apps (location check-ins, friend status updates, messaging etc.)
- [ ] Transportation apps (Public transportation, navigation tools, driving, etc.)
- [ ] Travel apps (airplane tickets, tourist guides, etc.)
- [ ] Utility apps (calculate, convert, translate, etc.)
- [ ] Weather apps (local forecasts, natural disaster updates, etc.)

Q7.
How many times do you check your mobile phone during the day (e.g. check notifications/time, open apps, etc.)?

1-35	36-70	71-100	101-135	135 +
○	○	○	○	○

Q8.
How concerned are you about the privacy of your mobile sensor data?

1 (Not at all concerned)	2	3	4	5 (Extremely concerned)
○	○	○	○	○

Q9.
Which level of privacy-intrusion would you assign to the following mobile sensors?

	1 (Very low)	2	3	4	5 (Very high)

Accelerometer (It measures the changes of the velocity of the smartphone)	○	○	○	○	○
Gyroscope (It measures the rotation/ twist of the smartphone)	○	○	○	○	○
GPS (It measures the geographical location of the smartphone)	○	○	○	○	○
Proximity Sensor (It measures the physical distance of your smartphone from your body)	○	○	○	○	○
Ambient Light Sensor (It measures the ambient light level)	○	○	○	○	○
Battery sensor (It measures the battery level)	○	○	○	○	○
Microphone (It measures several sound features, e.g. level of sound frequencies)	○	○	○	○	○
Camera	○	○	○	○	○
Thermometer (It measures the temperature of the device)	○	○	○	○	○
Air humidity sensor (It measures the relative humidity in a range 0-100%)	○	○	○	○	○
Barometer (It measures the atmospheric pressure)	○	○	○	○	○
Bluetooth (It measures the proximity of your device with other devices)	○	○	○	○	○

Q10. **How important for your privacy is the type of sensor from which you share data?**

1 (Not at all important)	2	3	4	5 (Extremely important)
○	○	○	○	○

Q11. **Which level of privacy-intrusion would you assign to the following stakeholders if you had to share your mobile sensor data with them?**

	1 (Very low)	2	3	4	5 (Very high)
Corporations/companies	○	○	○	○	○
Non-profitable/non-governmental organizations	○	○	○	○	○
Educational institutions (Public)	○	○	○	○	○
Governments & governmental organizations	○	○	○	○	○

Q12.
How important for your privacy is which stakeholder you share your mobile sensor data with?

1 (Not at all important)	2	3	4	5 (Extremely important)
○	○	○	○	○

Q13. Which level of privacy-intrusion would you assign to the following contexts of apps with access to your mobile sensor data?

	1 (Very low)	2	3	4	5 (Very high)
Education	○	○	○	○	○
Entertainment	○	○	○	○	○
Environment	○	○	○	○	○
Finance	○	○	○	○	○
Health	○	○	○	○	○
Shopping	○	○	○	○	○
Social networking	○	○	○	○	○
Training	○	○	○	○	○
Transportation/Traveling	○	○	○	○	○

Q14. How important is for your privacy the context of apps in which you share your mobile sensor data?

1 (Not at all important)	2	3	4	5 (Extremely important)
○	○	○	○	○

Q15. Select one or more incentives which would motivate you to share your mobile sensor data

- ☐ Money
- ☐ Vouchers/discounts on services and stores
- ☐ Free access to additional services (maps, recommended apps, etc.)
- ☐ Free access to data
- ☐ Contributing to public good
- ☐ Contributing data if my friends did
- ☐ Contributing data without incentives

Block 1

Exit-Survey

Default Question Block

Dear Participant,

Thank you very much for participating at the social experiment "Information Sharing of Mobile Sensor Data".

Before the end of this experiment, you should fill in this survey that summarizes your overall feedback on the experiment and participation.

Please double check and make sure you have filled in your unique ID number.

Kind regards,

Dr. Evangelos Pournaras

Please enter your unique ID number.

Which mobile phone model have you used for this experiment?

User Interface & Mobile App Functionality

How easy was it to use the mobile app of the experiment?

Extremely easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Extremely difficult
○	○	○	○	○

How would you rate the quality of the app?

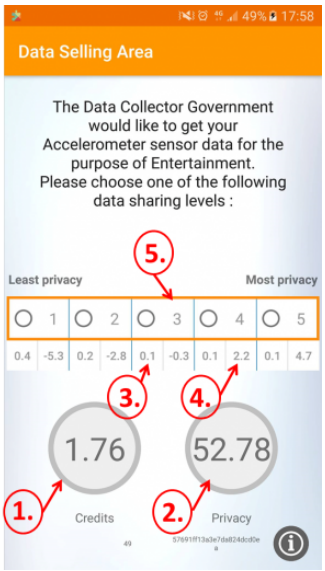
Extremely bad	Somewhat bad	Neither good nor bad	Somewhat good	Extremely good
○	○	○	○	○

How satisfied are you with each of the following features of the mobile app?

	Extremely dissatisfied				Extremely satisfied
Battery consumption	○	○	○	○	○
Performance speed	○	○	○	○	○
Colors	○	○	○	○	○
Formulation of questions	○	○	○	○	○
Content of questions	○	○	○	○	○
Number of different questions	○	○	○	○	○
Frequency of the notifications	○	○	○	○	○
The application as an overall	○	○	○	○	○

Please take a careful look at the following screenshots of the mobile app. The enumerated red arrows point to certain features of the mobile app:

1. Total rewards
2. Total privacy
3. Rewards for a certain choice
4. Privacy for a certain choice
5. Options



Please evaluate the following features of the mobile app.

	Very little				Very much
How comprehensible was the indicator of the total rewards? (Arrow 1)	○	○	○	○	○
How useful was the indicator of the total rewards in order to make a choice? (Arrow 1)	○	○	○	○	○
How comprehensible was the indicator of the total privacy? (Arrow 2)	○	○	○	○	○
How useful was the indicator of the total privacy in order to make a choice? (Arrow 2)	○	○	○	○	○
How comprehensible was the indicator of rewards for each data sharing level? (Arrow 3)	○	○	○	○	○
How useful was the indicator of rewards for each data sharing level in order to make a choice? (Arrow 3)	○	○	○	○	○
How comprehensible was the indicator of privacy for each data sharing level? (Arrow 4)	○	○	○	○	○
How useful was the indicator of privacy for each data sharing level in order to make a choice? (Arrow 4)	○	○	○	○	○
How comprehensible was the indicator of the options? (Arrow 5)	○	○	○	○	○
How useful was the indicator of the options in order to make a choice? (Arrow 5)	○	○	○	○	○

Do you have any other comments regarding the indicators?

Privacy & Rewards

Please evaluate the following questions about privacy.

	Definitely not				Definitely yes
Did this experiment make you feel more aware of the privacy of mobile sensor data?	○	○	○	○	○
Did the values of privacy represent well your choices of privacy-preservation?	○	○	○	○	○
Could you easily adjust your total privacy when it was not satisfactory?	○	○	○	○	○
Did your privacy-preservation					

choices deserved the sacrifice of rewards?	○	○	○	○	○

How satisfied are you with the following:

	Extremely dissatisfied				Extremely satisfied
The total available amount of rewards (30 CHF).	○	○	○	○	○
The amount of rewards you gained during the experiment out of the total available amount of rewards.	○	○	○	○	○

Please answer the following questions about rewards.

	Definitely not				Definitely yes
Did rewards convince you to share mobile sensor data?	○	○	○	○	○
Did rewards convince you to share more mobile sensor data than without rewards?	○	○	○	○	○
Did rewards make you more aware about the privacy of mobile sensor data?	○	○	○	○	○
Did rewards make you more aware about the value of mobile sensor data?	○	○	○	○	○
Did your rewards choices deserved the sacrifice of privacy?	○	○	○	○	○

Evaluate the change in rewards among the different data sharing options (Arrow 3).

Very low	Moderately low	Reasonable	Moderately high	Very high
○	○	○	○	○

Evaluate the change in privacy level among the different data sharing options (Arrow 4).

Very low	Moderately low	Reasonable	Moderately high	Very high
○	○	○	○	○

Experiment

Have you participated before in the following:

	Yes	No
An experiment at ETH Decision Science Lab?	○	○
A social experiment elsewhere?	○	○
An experiment that requires the use of a mobile app?	○	○

How interesting was the experiment?

Not interesting at all	Slightly interesting	Moderately interesting	Very interesting	Extremely interesting
○	○	○	○	○

Would you participate in a similar experiment again?

Definitely not	Probably not	Might or might not	Probably yes	Definitely yes
○	○	○	○	○

How satisfied are you with the following:

	Extremely dissatisfied				Extremely satisfied
The instruction about the experimental process.	○	○	○	○	○
Your participation in the entry phase	○	○	○	○	○
Your participation in the core phase	○	○	○	○	○
Your participation in the exit phase	○	○	○	○	○
Your participation in the overall experiment	○	○	○	○	○
The technical support of the staff members moderating the experiment	○	○	○	○	○

Have you run out of the battery during the experiment?

○ Yes

○ No

If yes, provide some more information (e.g. how long, how many times, at what time of the day)

Which of the following reasons prevented you from answering more questions?

- [ ] I was not interested anymore
- [ ] I was not enough motivated
- [ ] I faced technical problems
- [ ] I ran out of battery
- [ ] I was busy
- [ ] I was not satisfied by the experiment
- [ ] I was concerned about my privacy
- [ ] Other

Did you think at any time to drop out the experiment?

- ( ) Yes
- ( ) No

If yes, what was the reason?

Did you experience any of the following technical problems?

- [ ] Application crashed
- [ ] Application froze
- [ ] Application was too slow
- [ ] Network connection problems
- [ ] Battery drain
- [ ] Other

Have you been aware of the 'fair-data-share' web portal?

- ( ) Yes
- ( ) No

Have you ever visited the 'fair-data-share' portal?

- ◯ Yes
- ◯ No

Support Letters

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Wirtschaft, Bildung und Forschung WBF

**Staatssekretariat für Bildung,
Forschung und Innovation SBFI**
Nationale Forschung und Innovation

CH-3003 Bern, SBFI, frs

A-Post
Fair Data Sharing Portal
The Information Sharing Social Experiment

electronic

Ihr Zeichen:
Sachbearbeiter/in: frs
Bern, June 27 2016

Letter of intent

Dear Dr. Pournaras,

I am writing to confirm that our division expresses interest in accessing data collected throughout the *Information Sharing Social Experiment* via the so-called *Fair Data Sharing Portal*. The experiment is being conducted by the *Chair for Computational Social Science* at ETH Zurich in Switzerland.

We are the division of national research and innovation policy at the Federal State Secretariat of Education, Research and Innovation. I am leading the Swiss Innovation Policy Unit and through the present promotion letter, I try to advance towards a deeper understanding of data and its specific privacy requirements.

I have been told that the scope of the Information Sharing Social Experiment is to gather information about decision-making in information sharing of sensor data collected from the sensors of smartphones. This field is of great interest to us. Therefore, we express our support for this project and will eventually apply for access to the database of the project. We are aware that the collected data is anonymous, can only be used as stated in the data portal's terms of reference and will not be used in any other way or shared with third parties. Access to the data will comply with informed consent of participants and with the decisions that participants will make throughout the experiment.

Yours sincerely,

Sebastian Friess PhD
Head of Innovation Policy Unit
Deputy Division Head

Staatssekretariat für Bildung,
Forschung und Innovation SBFI
Dr. Sebastian Friess
Einsteinstrasse 2, 3003 Bern
Tel. +41 58 464 94 04
sebastian.friess@sbfi.admin.ch
www.sbfi.admin.ch

swiss made
software

Basel, 27. Juni 2016

Letter of support

Dear Dr. Pournaras,

I am writing to confirm that swiss made software expresses interest in accessing data collected throughout the Information Sharing Social Experiment via the so-called Fair Data Sharing Portal. The experiment is being conducted by the Chair for Computational Social Science at ETH Zurich in Switzerland.

The swiss sade software label is the leading label for software made in Switzerland and a monitor for the Swiss software industry. As such we have an interest in data science and in a deeper understanding of data and its specific privacy requirements.

The scope of the Information Sharing Social Experiment – to gather information about decision-making in information sharing of sensor data collected from the sensors of smartphones – is of great interest to us. Therefore, we express our support for this project and apply for access to the database of the project.

swiss made sofware is aware that the collected data is anonymous, can only be used as stated in the data portal's terms of reference and will not be used in any other way or shared with third parties. Access to the data will comply with informed consent of participants and with the decisions that participants will make throughout the experiment.

Yours sincerely,

Ihr swiss made software-Team

Christian Walter
Managing Partner

swiss made software gmbh
Dufourstrasse 9
CH-4052 Basel

www.swissmadesoftware.org
contact@swissmadesoftware.org
Tel.: 061 690 20 52

UID: CHE-144.409.352 MWST
Basler Kantonalbank 4002 Basel
IBAN: CH12 0077 0252 8756 5200 1

New York, 27. June 2016

Tamedia AG
Tagesanzeiger
Werdstrasse 21
8004 Zürich
Tel. 044 248 41 11

Fair Data Sharing Portal
The Information Sharing Social Experiment
Letter of Interest

Dear Dr. Pournaras,

I am writing to confirm that the Tages-Anzeiger (published by Tamedia AG) expresses interest in accessing data collected throughout the *Information Sharing Social Experiment* via the so-called *Fair Data Sharing Portal*. The experiment is being conducted by the *Chair for Computational Social Science* at ETH Zurich in Switzerland.

The Tages-Anzeiger is a leading Swiss daily newspaper published by Tamedia AG. Tamedia AG is expanding it's data analysis capabilities as it regards such journalistic skills as an important feature of future pubishing. Across it's portfolio Tamedia AG helps to store, manage, protect and analyze its most valuable asset - information - in a more agile, trusted and cost-efficient way.

The scope of the Information Sharing Social Experiment – to gather information about decision-making in information sharing of sensor data collected from the sensors of smartphones – is of great interest to us. Therefore, we express our support for this project and apply for access to the database of the project.

Tages-Anzeiger and Tamedia AG have well proven its capability to process very large data sets. We are aware that the collected data is anonymous, can only be used as stated in the data portal's terms of reference and will not be used in any other way or shared with third parties. Access to the data will comply with informed consent of participants and with the decisions that participants will make throughout the experiment.

Sincerely,

Barnaby Skinner

--

Datenjournalist
SonntagsZeitung & Tages-Anzeiger
Office +41 44 248 52 26
Mobile +41 79 640 98 1
barnaby.skinner@sonntagszeitung.ch
barnaby.skinner@tages-anzeiger.ch
www.barnabyskinner.com

Information Sheet

General points:

This study aims at studying the perception people have about privacy and information sharing in the context of mobile sensor data. Nowadays, smartphones are equipped with sensors that can collect real-time information such as our GPS location, the acceleration of motion or even environmental information, for instance temperature and humidity. Smartphones run applications (apps) that are pieces of software with potential access to sensor data. This data can be shared with remote stakeholders such as companies, governments, educational institutions, and others.

The analysis of mobile sensor data by these stakeholders may put privacy at risk, especially when the sensor data is collected with a fine-grained frequency. Therefore, the amount of mobile sensor data that a user chooses to share with a certain stakeholder for a certain purpose (app context) indicates his/her preferred privacy settings. Participants can assume that any security requirement is met and privacy is entirely governed by their decisions.

Stakeholders can access the sensor data of participants via the web portal fair-data-share.inn.ac. Access to the data complies to the decisions that participants make during the experiment. Stakeholders agree to neither share the data that they can potentially access via the web portal nor infer any individual from the values of sensor data.

Mandatory components:	**Annotation:**
a) Goals of the study	Understand human perception on privacy of mobile sensor data and how this perception influences online decision-making about sharing sensor data. Moreover, this study aims at understanding how decision-making is influenced when incentives, e.g. monetary ones, are given to citizens in order to share a higher/lower amount of sensor data at a cost of lower/higher privacy-preservation respectively.
b) Research procedure (methods)	A social experiment requiring a 2-day participation at the ETH Decision Science Lab and 2-day usage of a mobile app.
c) Schedule	The social experiment is outlined in 3 phases: 1. Entry phase (45 mins work): Show up at the ETH Decision Science Lab, instructions, sign of information consent, app installation, entry app survey 2. Core phase (45 mins work): A two-day app usage. 3. Exit phase (30 mins work): Show up at the ETH Decision Science Lab, exit web survey, receipt of rewards.
d) Conditions to be met for participation in the study	Participation in this study requires the following: 1. having a general interest and concerns about privacy 2. having a smartphone running Android 3. having a mobile internet connection 4. coming up to the entry phase with a fully charged phone

	5. committed to participate in all following three phases of the experiment.
e) Advantages and disadvantages for participants / Risks	This research is dedicated to a better understanding of privacy. The findings can be potentially used to improve privacy awareness and preservation of citizens. All risks related to participants' anonymity, information leak, malfunctions and data loss are minimized.
f) Source of funding	This study is supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem' (http://www.sobigdata.eu)
g) Compensation/Reimbursement	Participants will be reimbursed with a maximum of 75 CHF as follows: • 20 CHF entry fee for the entry and exit phase • 20 CHF participation fee for the entry and exit phase • 5 CHF app use fee for the core phase • 30 CHF of max rewards based on the received answers
h) Right of withdrawal	As a participant, you have the right to withdraw from the study at any time without needing to specify any reasons or facing negative consequences.
i) Data protection	The anonymity of the participants is guaranteed throughout the experimental process and later on during the data analysis. Participants will be have a unique identifier. The participants' responses and sensor data collected are bound to this unique identifier and there is no other link to any personal information. The data leaving the phone are encrypted and stored in a secure server. The data collected can be processed and analyzed within the Computational Social Science group for research purposes. The stakeholders can potentially access the sensor data that each participant permitted during the experiment and they are not allowed to share any further these data.
j) Insurance coverage	Possible damages to your health, which are directly related to the study and are demonstrably the fault of ETH Zurich, are covered by the general liability insurance of ETH Zurich (insurance policy no. 100.001 of the Swiss Mobiliar insurance company). However, beyond the before mentioned, the health insurance and the accident insurance (e.g. for the way to or back from the study location) is in the responsibility of the participant.
k) Contact person(s)	Dr. Evangelos Pournaras – epournaras@ethz.ch

Consent Form

⇒ ***Please read this form carefully.***
⇒ ***Please ask the investigator or the contact person if you have any questions.***

Study title: Inventivized and privacy-preserving sharing of mobile sensor data – A social experiment

Study location: ETH Zurich, Professorship of Computational Social Science, Clausiusstrasse 50, 8092, Zurich, Switzerland

Principal Investigator's Name and First Name*: Prof. Dr. Dirk Helbing and Dr. Evangelos Pournaras*

Participant's Name and First Name:

Participant:

⇒ I participate in this study on a voluntary basis and can withdraw from the study at any time without giving reasons and without any negative consequences.
⇒ I have been informed orally and in writing about the aims and the procedures of the study, the advantages and disadvantages as well as potential risks.
⇒ I have read the written information for the volunteers. My questions related to the study participation have been answered satisfactorily. I have been given a copy of the information for the volunteers and the consent form.
⇒ I was given sufficient time to make a decision about participating in the study.
⇒ With my signature I certify that I fulfill the requirements for the study participation mentioned in the information for the volunteers.
⇒ I have been informed that possible damages to my health which are directly related to the study and are demonstrably the fault of ETH Zurich, are covered by the general liability insurance of ETH Zurich (insurance policy no. 100.001 of the Swiss Mobiliar insurance company). However, beyond the before mentioned, my health- and/or accident insurance (e.g. for the way to or back from the study location) will apply.
⇒ I agree that the responsible investigators and/or the members of the ethical committee have access to the original data under strict confidentiality.
⇒ I am aware that during the study I have to comply with the requirements and limitations described in the information for the volunteers. In my own health interest the investigators can, without mutual consent, exclude me from the study.
⇒ I agree the involved stakeholders of this study to have access to the mobile sensor data I will share according to the choices I will make during this study.

Location, date . . . . . Signature volunteer . . . . . .

Location, date . . . . . Signature investigator . . . . . .

user_id	corp	edu	gov	ngo
"57a8f8f1848532cf76b0836f"	3	1	4	3
"579a148f352257bc0612c70b"	3	3	4	2

Figure A.5: Screenshot of Collection Stakeholders

user_id	environment	health	social_networking	transportation
"57a8f8f1848532cf76b0836f"	3	3	5	3
"579a148f352257bc0612c70b"	4	2	2	2

Figure A.6: Screenshot of Collection Contexts

contexts	credit	credit_can_be	credit_gain	credit_question	data_collectors	timestamp	day_no
3	9.75761217948…	0.2804487179487179	0.21033653846153844	0.3271901709401709	2	"2016-08-08 23:45:24.581"	2
0	9.05048076923…	0.3084935897435897	0.17628205128205127	0.3084935897435897	3	"2016-08-08 23:44:45.49"	2

Figure A.7: Screenshot of Collection UserResponse Part 1

user_id	improve	privacy_can_be	privacy_gain	privacy_level	privacy_percentage	sensors
"57935b55a67b0ba32f81eeac"	1	1.5625	0.78125	3	63.28125	1
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2

Figure A.8: Screenshot of Collection UserResponse Part 2

day_no	user_id	lat	long	summarization	timestamp
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186498206
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186468203

Figure A.9: Screenshot of Collection Location

day_no	summarization	timestamp	user_id	x	y	z
3	1	1469186493918	"578e91e778f251171…	0.0191536135971546…	-0.143652096390724…	10.15141487121582
3	1	1469186463719	"578e91e778f251171…	0.0191536135971546…	-0.143652096390724…	10.15141487121582

Figure A.10: Screenshot of Collection Accelerometer

bands	user_id	day_no	rms	spl	summarization	timestamp
"0.0,1.9080862E-5,…	"578e91e778f2511711cfb9f5"	3	107.31494140625	62.65440368652344	1	1469186468205
"0.0,1.5665331E-5,…	"578e91e778f2511711cfb9f5"	3	88.882568359375	61.01753234863281	1	1469186498214

Figure A.11: Screenshot of Collection Noise

day_no	summarization	timestamp	user_id	x
3	3	1469447239362	"57935b55a67b0ba32…	47
3	3	1469447109437	"57935b55a67b0ba32…	54

Figure A.12: Screenshot of Collection Light

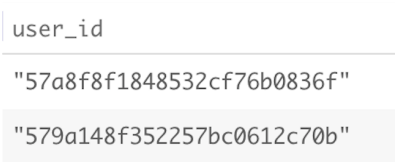


Figure A.13: Screenshot of Collection Users

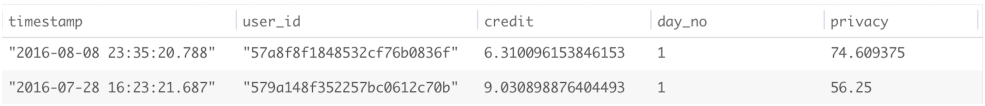


Figure A.14: Screenshot of Collection Score

Table A.1: Demographics of Population in the Survey

Country	Percentage
United States of America	1.01%
United Arab Emirates	0.51%
The former Yugoslav Republic of Macedonia	0.51%
Syrian Arab Republic	0.51%
Switzerland	20.71%
Spain	1.01%
Slovakia	0.51%
Serbia	5.05%
Russian Federation	0.51%
Netherlands	1.52%
Italy	2.02%
Iran	1.01%
India	14.65%
Hungary	0.51%
Greece	29.29%
Germany	10.61%
France	1.52%
Czech Republic	1.01%
Costa Rica	0.51%
China	0.51%
Columbia	0.51%
Canada	0.51%
Bolivia	0.51%
Brazil	1.52%
Bahrain	0.51%
Argentina	0.51%
Austria	2.02%

Bibliography

[1] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 2(2005):24–30, 2005.

[2] Rebecca Balebako, Florian Schaub, Idris Adjerid, Alessandro Acquisti, and Lorrie Cranor. The impact of timing on the salience of smartphone app privacy notices. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 63–74. ACM, 2015.

[3] AJ Brush, John Krumm, and James Scott. Exploring end user preferences for location obfuscation, location-based services, and the value of location. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 95–104. ACM, 2010.

[4] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. Participatory sensing. *Center for Embedded Network Sensing*, 2006.

[5] L Jean Camp. State of economics of information security, the. *ISJLP*, 2:189, 2005.

[6] Haksoo Choi, Supriyo Chakraborty, Zainul M Charbiwala, and Mani B Srivastava. Sensorsafe: a framework for privacy-preserving management of personal sensory information. In *Workshop on Secure Data Management*, pages 85–100. Springer, 2011.

[7] Delphine Christin. Privacy in mobile participatory sensing: current trends and future challenges. *Journal of Systems and Software*, 116:57–68, 2016.

[8] Delphine Christin, Christian Büchner, and Niklas Leibecke. What's the value of your privacy? exploring factors that influence privacy-

sensitive contributions to participatory sensing applications. In *Local Computer Networks Workshops (LCN Workshops), 2013 IEEE 38th Conference on*, pages 918–923. IEEE, 2013.

[9] Dan Cvrcek, Marek Kumpost, Vashek Matyas, and George Danezis. A study on the value of location privacy. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 109–118. ACM, 2006.

[10] George Danezis, Stephen Lewis, and Ross J Anderson. How much is location privacy worth? In *WEIS*, volume 5. Citeseer, 2005.

[11] Linda Deng and Landon P Cox. Livecompare: grocery bargain hunting through participatory sensing. In *Proceedings of the 10th workshop on Mobile Computing Systems and Applications*, page 4. ACM, 2009.

[12] Fosca Giannotti, Dino Pedreschi, Alex Pentland, Paul Lukowicz, Donald Kossmann, James Crowley, and Dirk Helbing. A planetary nervous system for social mining and collective awareness. *The European Physical Journal Special Topics*, 214(1):49–75, 2012.

[13] Eiji Hayashi, Oriana Riva, Karin Strauss, AJ Brush, and Stuart Schechter. Goldilocks and the two mobile devices: going beyond all-or-nothing access to a device's applications. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 2. ACM, 2012.

[14] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 501–510. ACM, 2012.

[15] Emiliano Miluzzo, Nicholas D Lane, Shane B Eisenman, and Andrew T Campbell. Cenceme–injecting sensing presence into social networking applications. In *European Conference on Smart Sensing and Context*, pages 1–28. Springer, 2007.

[16] Prashanth Mohan, Venkata N Padmanabhan, and Ramachandran Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 323–336. ACM, 2008.

[17] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA law review*, 57:1701, 2010.

[18] Evangelos Pournaras. Application form to the research ethics committee of eth zurich.

[19] Evangelos Pournaras, Jovan Nikolic, Pablo Velásquez, Marcello Trovati, Nik Bessis, and Dirk Helbing. Self-regulatory information sharing in participatory social sensing. *EPJ Data Science*, 5(1):1, 2016.

[20] Ashwini Rao, Florian Schaub, and Norman Sadeh. What do they know about me? contents and concerns of online behavioral profiles. *arXiv preprint arXiv:1506.01675*, 2015.

[21] Lei Song, Yongcai Wang, Ji-Jiang Yang, and Jianqiang Li. Health sensing by wearable sensors and mobile phones: a survey. In *e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on*, pages 453–459. IEEE, 2014.

[22] Jinyan Zang, Krysta Dummit, James Graves, Paul Lisker, and Latanya Sweeney. Who knows what about me? a survey of behind the scenes personal data sharing to third parties by mobile apps. *Proceeding of Technology Science*, 2015.

ETH

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):

Fair Data Sharing in Participatory Social Sensing

Authored by (in block letters):

For papers written by groups the names of all authors are required.

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

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)**
Zurich, 27-08-2016	Ramapriya

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