Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Data Sharing in Participatory Social Sensing

Master Thesis

Ramapriya Sridharan

September 3, 2016

Advisors: Prof. Dr. Dirk Helbing, Dr. Pournaras Evangelos

Department of Computational Social Sciences , ETH Zürich

Contents

Abstract

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

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 his continuous mentoring, guidance and support through this six month period without who 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

In today's world, almost everybody owns a smartphone. Information collected from a large number of interconnected smartphones equipped with multiple number of sensors each can aid to accumulate large volumes of data which is heterogeneous and autonomous in nature. This large amounts of data in any form is called Big Data[1] and it is the stepping stone to large scale data analytics and Deep learning to understand the complexities in our society.

Currently, a lot of the data collected in mobile applications[2] and on the web[3] is done in a manner where users are unaware of the collection of their data. This process not only lacks transparency but also does not give users control over their data. Users should be able to make informed decisions about what data to share or not. Additionally, since no privacy algorithms are implemented on the data collected, user privacies are at risk. Paul Ohm in his paper [17] explains that it can be shown that anonymized data can be de-anonymized surprisingly easily. In this dissertation, focus is not given on the attacks such as hacking and any other methods but concentrated on the threats to the information in the data collected.

The aim is to setup a fair way to collect data by setting up a platform over a participatory sensing network where users can trade their data for some incentives to stakeholders who approach them. Incentives can be money, vouchers or anything else. Stakeholders inform users about the date, duration for which data is collected, who will use the data and what will it be used for. Additionally, users have the possibility to share data with some added levels of privacy. To achieve this goal, it is first essential to understand the relationship between incentives and mobile sensor data sharing. In this

[1]Date:23-08-2016 https://en.wikipedia.org/wiki/Big_data

[2]Date:23-08-2016 http://www.theregister.co.uk/2014/02/21/appthority_app_privacy_study

[3]Date:23-08-2016 http://www.techlicious.com/blog/whos-gathering-your-personal-information

dissertation, a survey and social experiment is designed where a platform is created and users can trade and view their data in a transparent manner. Data from user inputs and decisions is collected and later analysed to throw light on user decision of their sensor data.

Chapter 2

Related Work

Participatory social sensing is the active participation of users with their mobile phones to form a network that enables the collection and analyzation of large amounts of data. The concept was first introduced by Burke et al in 2006 [4], which talks about the potential benefits and propose an initial architecture. Collecting data from various sensors is important for Big data analysis and to find answers to complex social questions. Lei Song et al [21] performed an extensive survey on the sensor devices and their applications. It was found that different sensor combinations is the foundation to grasp important information. 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] did a study using 110 Android and iOS apps to find the ones that share personal information, behavioural information and location data with third parties. Doing this revealed that it does not require a notification from the application. They also found that paid applications still shared sensitive information to third parties.

Ashwini Rao et al [20] examined the behavioural profiles formed by Google and Yahoo. Participants were surprised and concerned that data has been collected from them. Additionally, the profiles formed were found to be in some aspects inaccurate and had excess of information, so much that the profiles did not seem to be anonymous anymore. Further, a survey was created asking participants questions about the behavioural profiles and was launched on Amazon Turk. Participants dint find the profiles formed to be accessible enough and they also complained that they wanted to know more about who and where the data 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] created a survey to observe the privacy concerns in e-commerce preferences and masking of location data. They found 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 felt 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] did a study to assess how much people value their location data using auction technique. They found that the median bid was 43$ a period of one month, but this varied 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 were told fake goals in order not to be biased about their data privacy. Contradictions were found to the study [10] about the change in value due to travelling habits, but the median bid was found to be the same. There was also differences in results among the uses with different demographics. This goes to show that one incentive does not fit all the users.

Delphine Christine et al [8] perform an extention of the study in the paper by 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 found that users claimed more rewards to corporations. The work by Camp Jean [5] has mentioned that the participants of the surveys may not tell the truth inspite of 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 were shown visualizations of their data. The authors mentioned that the user interface was not intuitive and that users might have been biased to the location data due to the experimental setup. Additionally, it was found that users were not consistent with privacy decisions and with whom they shared data. It was concluded that users need to be properly informed about every detail to enable them to make rational choices.

Haksoo Choi et al [6] is a framework that provides sharing of sensor data based on rules along with the possibility of applying obfuscation algorithms. They found 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 nothing options. It is found that all or nothing 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 accuracy of the data that is shared and the accuracy required by the data aggregators. If the data shared is of lower accuracy then the errors in data processing increase. Similarly, with higher accuracy the data processing tasks will give lesser errors. The process of manipulating the accuracy of sensor data is called summarization. The errors in the data has the possibility to be mitigated if there is a large population of users participating in the data sharing tasks. Summarization is any algorithm from simple arithmetic functions to clustering algorithms.

In this dissertation, it is attempted to address some of the drawbacks of the aforementioned studies. A social experiment is created and an Android application developed to study the relationship between incentives and data sharing for a wide range of sensors. Additional factors that can affect data sharing are also included in the equation such as potential data buyers and the purpose for which the data collected is used. Users are allowed to share data in five levels each corresponding to a summarization level.

The user interface has been designed to be intuitive and easy to use. Users are approached to share their data and are incentivized with some credits. The amount of credit received is proportional to the amount of data shared. The amount of credit per data request is dynamically allocated according to the user profile created. High amounts are allocated to data requests that are considered privacy intrusive by the user. To keep the users motivated a participation fee is included, which means users will receive credits for answering questions irrespective of the amount of data shared. Additionally, a survey is deployed before the experiment to gain insight into the perception of users on various features and to fine tune the mobile application details.

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

Chapter 3

Computational Model

3.1 Introduction

The aim is to create a computational model that is able to collect useful information about the influence of monetary incentives on mobile data sharing. The users are first asked some preliminary questions to form a profile about them. The model proceeds to use the user profiles formed to assign each sensor data request with a maximum achievable credit. The model attempts to identify the data requests where users might not be inclined to share mobile sensor data willingly. These data requests are assigned higher maximum obtainable costs. Similarly, the data requests where the users would want to share more mobile sensor data is assigned a lower maximum obtainable cost. 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 costs accordingly.

3.2 Model Intricacies

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

3.2.1 Collecting User Information

To begin with the model, each user is asked to enter various non-intrusive personal information. The information collected consists of but is not limited to :

- Gender
- Year of birth

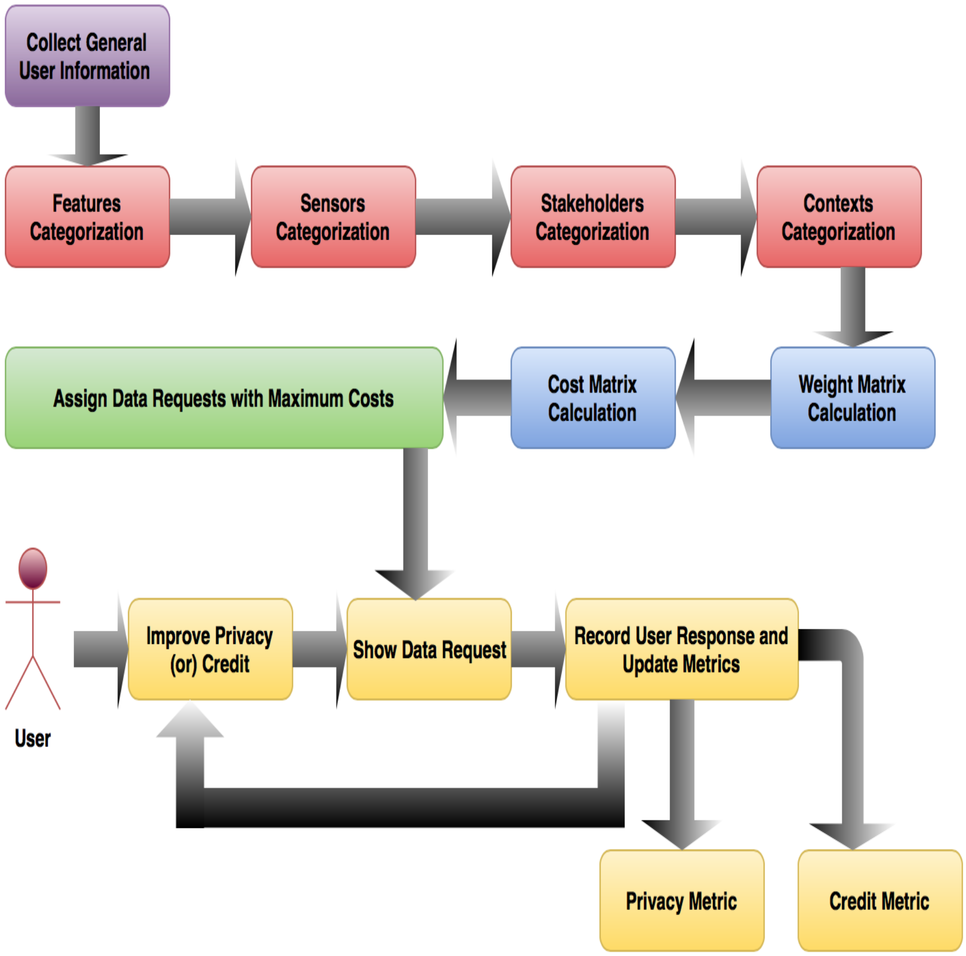


Figure 3.1: Computational Model Flow Chart

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

Collecting this information is crucial to the data analysis that is conducted in the later stages.

3.2.2 Categorization of the Features

After the users personal information is collected, users are asked to place the features in categories according to how privacy intrusive they are to the user. A feature can be one of the following:

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

Features are the three dimensions that form a unique data request. A data request is defined as a Stakeholder asking users to share their mobile Sensor data for a particular Context.

As mentioned before, 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 one to five and users can place each of the Features in a category according to their perceived intrusion level. Category one represents that the Feature does not contribute a lot to the data sharing decision. Similarly, category five represents that the Feature contributes a lot to the user's data sharing decision. Similarly, category five 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, which makes it a more powerful tool than the ranking mechanism.

Let the variable *cat* represent the number of categories, which here is five. Additionally, let the category assigned to the Sensors be represented by the variable *se*, the category assigned to the Stakeholders be represented by the variable *st* and the category assigned to the Contexts be represented by the variable *co*.

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_{se}, w_{dc} and w_{co} they are calculated as follows :

$$w_{se} = \frac{se}{se + st + co} \tag{3.1}$$

$$w_{st} = \frac{st}{se + st + co} \tag{3.2}$$

$$w_{co} = \frac{co}{se + st + co} \tag{3.3}$$

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 one 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

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 a similar fashion to section 3.2.2. The categories are the same as mentioned in the previous section. Let num_{sf} be the number of sub-features each feature has.

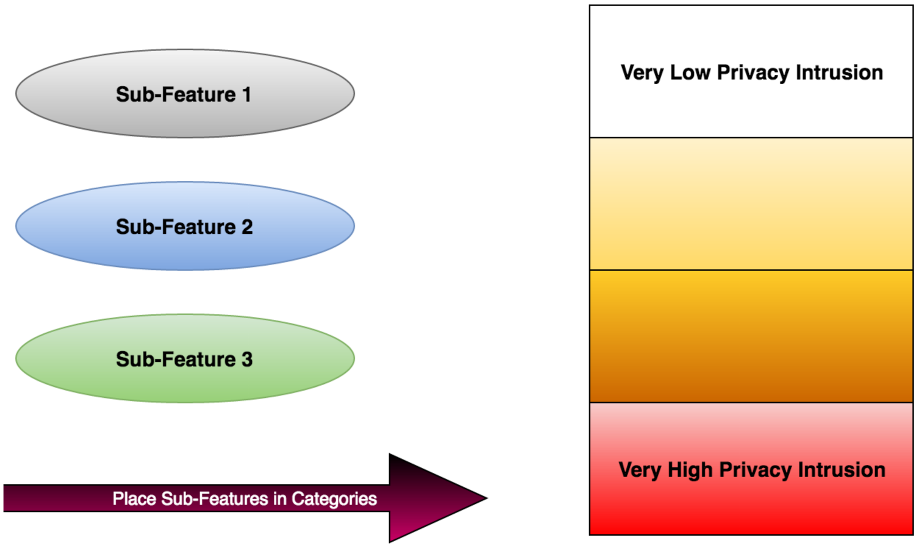


Figure 3.2: Categorizing Sub-Features according to the perceived Intrusion Level

The first category indicates that users find the sub-feature would not hinder the data sharing decision. This means that the user would not be worried trading data for a data request involving this sub-feature. The last category indicates that users find this sub-feature would hinder the data sharing decision . This means that users would be reluctant of giving data for a data request involving this sub-feature. 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 a unique identifier 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 se_i and i is the identifier of the sub-feature. Similarly, categories assigned to sub features of features Stakeholders and Contexts respectively are represented by st_j and co_k, where j and k are the identifiers of the sub-features categorized.

3.2.4 Weight Matrix Calculation

Each data request to users consists of the three above mentioned features in them. Each of the features each have num_{sf} sub-features that can appear in turns in a data request. So the total number of data requests are :

$$num_{dr} = num_{sf} * num_{sf} * num_{sf} \tag{3.4}$$

Let WM be a matrix with three dimensions num_{sf} x num_{sf} x num_{sf}. We call this the weight matrix. Each cell of WM, that is $WM_{i,j,k}$ 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 WM 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 WM represents the weight of a data request with a unique Sensors sub-feature i, Stakeholders sub-feature j and Contexts sub-feature k. To calculate the weight of a data request :

$$WM_{i,j,k} = (se * se_i) + (st * st_j) + (co * co_k) \tag{3.5}$$

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

3.2.5 Cost Matrix Calculation

Once WM has been calculated, it can give an idea of the weight each data request receives. The aim is now to assign a maximum obtainable cost to each data request. This cost is the maximum credit users can receive for a particular data request. Let CM be the cost matrix with the three dimensions num_{sf} x num_{sf} x num_{sf}. Let it be assumed to have a budget of b for a day, where b can be in an actual currency or any virtual credits. In this literature the budget will be referred to with the unit credits. Each cell of the cost matrix will represent the amount of credits allocated for a particular data request for one day. To begin with, we calculate the sum of all the cells of the weight matrix WM:

$$sum_{WM} = \sum_{i=1}^{num_{sf}} \sum_{j=1}^{num_{sf}} \sum_{k=1}^{num_{sf}} wm_{i,j,k} \tag{3.6}$$

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

$$CM_{i,j,k} = \frac{WM_{i,j,k} * b}{sum_{WM}} \tag{3.7}$$

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

3.2.6 Cost and Privacy Metrics

Every data request has been assigned a cost. This is the maximum cost that a user can obtain for that data request. The Cost metric is the total amount of credits the user has obtained by trading data for data requests for one day. 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.

Each data request the user 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 cost assignment to each option is linearly scaled according to the cost assigned to each data request. Let us assume there are options for a data request ranging from 1 to m (numeric options), where 1 corresponds to 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 costs to data requests there are two scenarios to consider:

- Assigning option costs without a participation cost. Users are not rewarded for their participation
- Assigning option costs inclusive of a participation cost. Users are rewarded for responding to data requests irrespective of how much data they choose to share

Let us examine the first scenario. Let the option costs be calculated for the data request with Sensors sub-feature i, stakeholders sub-feature j and contexts sub-feature k. The assigned cost for any option numbered h of this data request is calculated as follows:

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

Applying this formula by replacing h by the option numbers from 1 to m gives the cost the user can receive for each option.

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

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

The fraction of the cost an option h can be assigned has been calculated, to get the cost $cost_h$ of option h for the data request with Sensors sub-feature i, stakeholders sub-feature j and contexts sub-feature k :

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

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

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

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

The total cost and privacy is the sum and arithmetic average of all the costs and privacy respectively, obtained from every answered data request. If a data request is left unanswered, a maximum privacy of 100% and minimum cost of 0 credit is assumed.

3.2.7 Improving the Metrics

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

3.2.8 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 algorithms that aggregates data to provide less information than in its original form [19]. A higher summarization level gives less data than than in its original form. A lower the summarization level gives data closer to its original form. In this model, sensor data is collected for a period of 24 hours every y seconds for every data request. If the data is summarized, according to the option chosen, the data is collected either every y seconds or lesser.

Data is collected for the 24 hour period, 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 $summ_h$ for an option h for a data request has the possibility of the following :

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

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

3.3 Analysis of the Model

In this section, three different examples are explained in order to test some of the properties that the model assigns to the weight and cost matrix.

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

The numbers indicated to the left of the sub-features is the sub-feature unique identifier. 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

$$num_{sf}.num_{sf}.num_{sf} = 27$$

data requests in total. The number of categories available to categorize is $cat = 5$ as explained in 3.2.2. Additionally, it is assumed that a budget $b = 100$ 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 Results

To begin with, the way the user has categorized the features and sub-features is introduced. This will be followed by an explanation of the generated matrices. To make reference easier to the graphs, instead of sub-feature names, numeric identifiers are used. From now on each feature and sub-feature will be referred to by its identifier 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.

Scenario One

If features and sub-features have all been given the same categories respectively by users, then all data requests should be assigned equal weights and costs. In scenario 1, the users choose categories for the Features and sub-features as shown in the table 3.1. As the table shows, each feature receives the category 1, and all their sub-features are categorized as 3. In short, all the features have the same categorization and their respective sub-features all have the same categorization as well. 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 of (sensors, stakeholders, contexts) in the x-axis of figures 3.3, all have identical weights and costs. This is due to the fact that the users find all the features and sub-features equally intrusive so all the data requests are weighted equally.

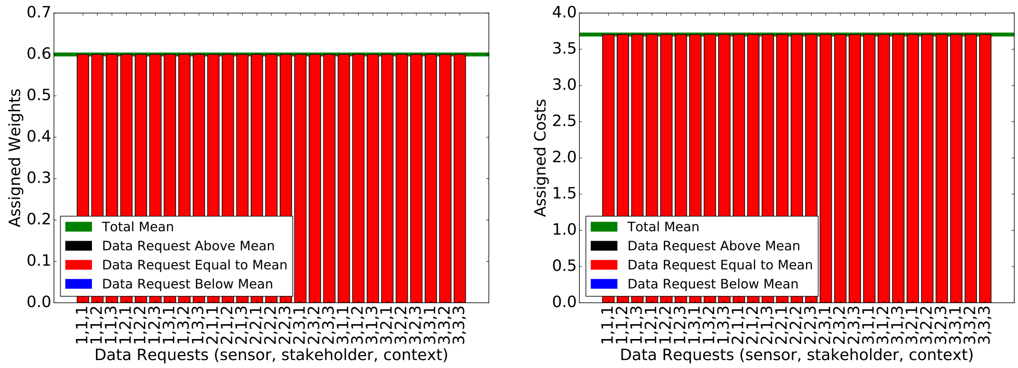
Table 3.1: Categorization for Scenario 1

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 feature and respective sub-features in an equally intrusive way, then all the data requests will receive the same weights and costs assignments.

Scenario 2

In this section it is attempted to test if data requests containing sub-features with higher intrusion levels are assigned higher weights and costs. Table 3.2 indicates the user input to scenario 2two. As it can be seen, all features have equal categories, and all sub-features have the categories of 3 with an exception the Sensor's sub-features. The Sensors sub-features with identifiers 1,2 and 3 have respectively categories 1,3 and 5. This means that requests with



(a) Values of the Weight Matrix

(b) Values of the Cost Matrix

Figure 3.3: Examining Scenario 1

Sensor's sub-feature 1 will be assigned a lesser weight in comparison to the other Sensors sub-features. Similarly, the data requests with Sensors sub-feature 2 will have a higher weightage assigned than Sensor's sub-feature 1 because of its higher category, but lesser than Sensor's sub-feature 3. Lastly, data requests with Sensor's sub-feature 3 will have a highest weight compared to the others, due to its category being 5. The weight and cost matrices can be seen in figures 3.4a and 3.4b respectively.

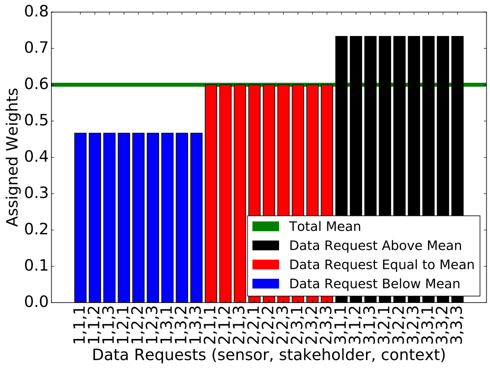
Table 3.2: Categorization for Scenario 2

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

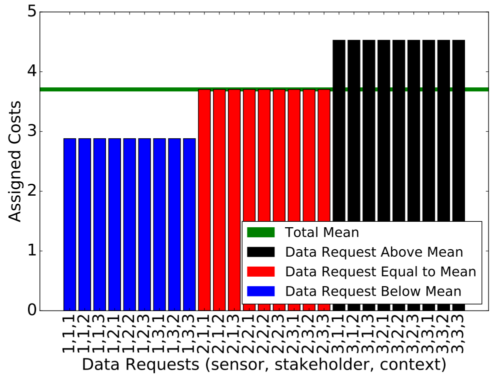
From the above inputs and graphs, it is concluded that the model assigns a higher weightage to data requests with sub-features that the user finds more intrusive compared to the others.

Scenario 3

The feature and sub-feature categories are both assigned different values in this section, to show how varying their values together affects the assignments of the weight and cost matrix. Table 3.3 is the user input to the



(a) Values of the Weight Matrix



(b) Values of the Cost Matrix

Figure 3.4: Examining Scenario 2

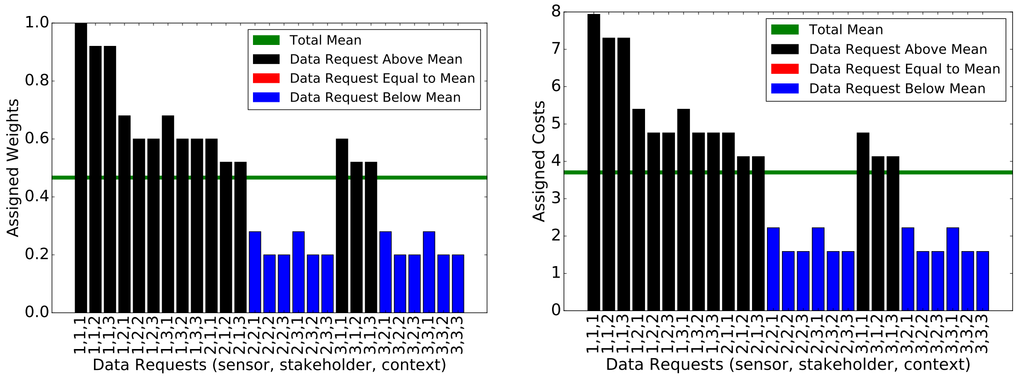
scenario 3. All the features have different categories assigned from 3 to 5. Additionally, the sub-feature 1 of each feature has a category of 5, higher than the other sub-features which are all categorized as 1. The weight and cost matrices generated for this scenario are shown in figures 3.5a and 3.5b respectively.

Table 3.3: Categorization for Scenario 3

Feature	Sub-Feature ID = 1	Sub-Feature ID = 2	Sub-Feature ID = 3
Sensors 5	Accelerometer 5	Noise 1	Location 1
Stakeholders 4	Corporation 5	Government 1	Educational Institution 1
Contexts 3	Navigation 5	Environment 1	Social Media 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-feature 1 of each feature. This happens because all of the sub-feature 1 are assigned a category of 5. The feature sensors and its sub-feature 1 are categorized as 5, so all the data requests with tuple (1,*,*), where * is all the other possible sub-features from other features, are all above average as seen in figures 3.5a, irrespective of the categories of the other feature's sub-features. This shows that assigning a higher category to a feature can lead to higher data request costs.

The green horizontal line in the graph indicates the mean value of the



(a) Values of the Weight Matrix

(b) Values of the Cost Matrix

Figure 3.5: Table Schemas

weights and costs. In general due to sub-features categorized as 5, those data requests receive a higher weight and cost. 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) even tough Contexts sub-feature 1 has a category of 5. This is due to the fact that Sensors and Stakeholders feature have a higher category of 5 and 4 respectively than the contexts feature. Since their sub-features are assigned a lower privacy intrusion category than the context's 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 still depends on the category of the feature it belongs to.

Additionally, it can be 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 costs that are well above average.

It can be concluded that the model assigns weights to data requests, by putting more emphasis on the feature's weights. A feature with high category has the ability to assign higher costs 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 aid to 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 to the usage of the mobile application created.

4.1 Preparatory Phase

4.1.1 Pre-Survey

The pre-survey [1] is a survey created that runs before the deployment of the social experiment. This survey was made in order to study the perception of users on the three features to be studied which are explained in detail in section 4.2.2. Figure 4.1[2] 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 data requests posed to the user. Additionally, we wanted to gain insight into the perception of users on the three features. Hence the survey was prepared to understand all of the above. Additionally, it can help redesign some of the aspects of the experiment based on the ambiguities found and user feedback. The participants pool consists of both 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.

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

[2] Figure made by Athina Voulgari

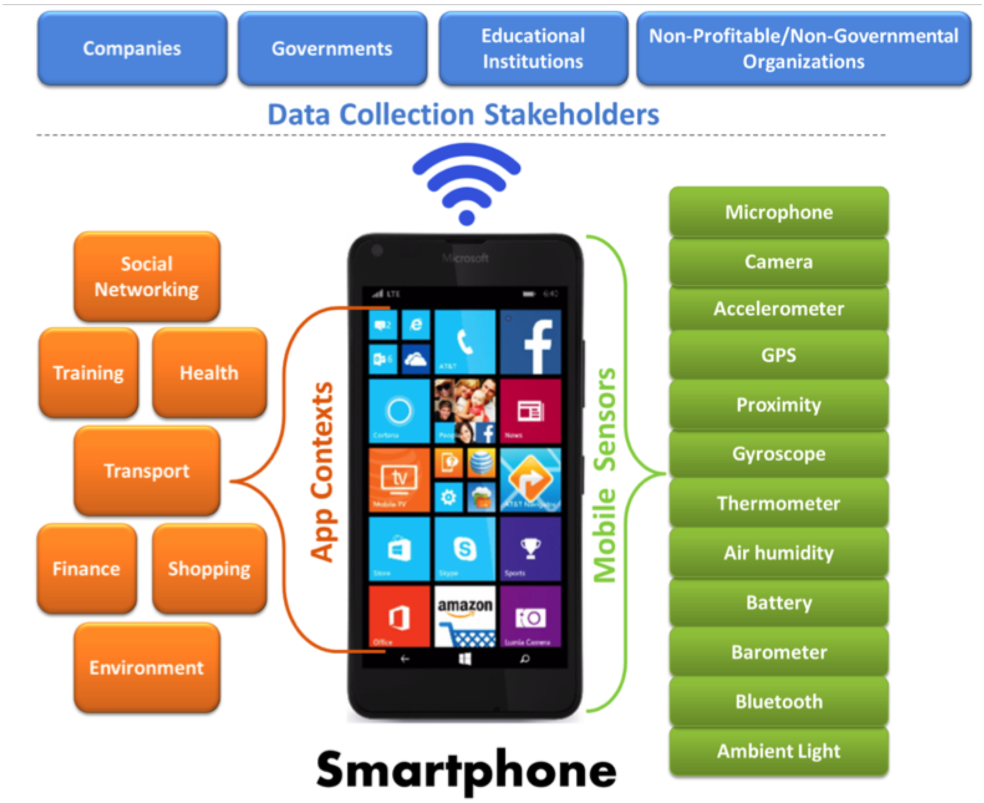


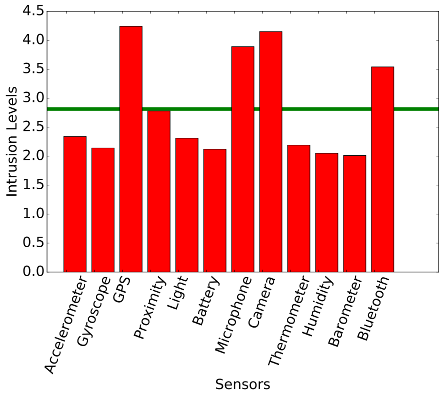
Figure 4.1: The Three Features Examined

4.1.2 Sub-Features

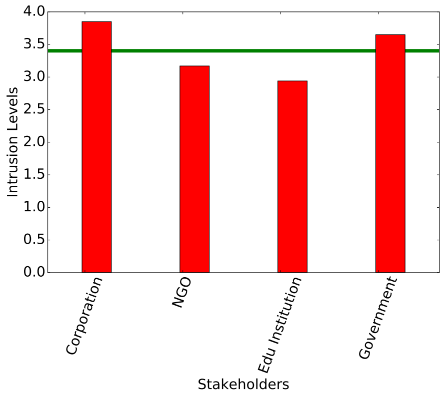
Figures 4.2a, 4.2b and 4.2c, each show the average intrusion level of each possible sub-feature for the sensors, stakeholders and contexts. For the experiment, it was decided to choose for each feature two non-intrusive and two intrusive sub-features each. The minimum privacy intrusion level is one which 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 be 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 blue line.

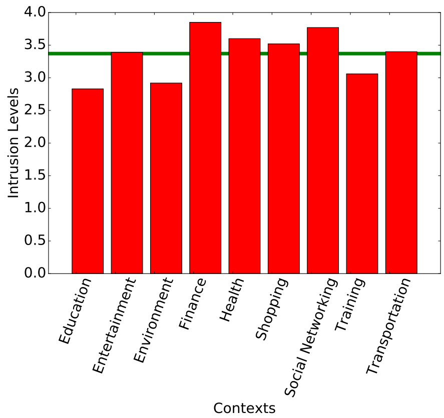
Similarly, looking at the stakeholder feature graph 4.2b, it is seen that sub-features corporation and government are found to be intrusive by the users with levels 3.8 and 3.6 on a scale of five. On the other hand, sub-features



(a) Average Intrusion of Sensors Sub-Features



(b) Average Intrusion of Stakeholders Sub-Features



(c) Average Intrusion of Contexts Sub-Features

Figure 4.2

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 4.2c, 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 user with values of 2.9 and 3.3 on a scale of five. The above mentioned sub-features for every feature have been chosen for the experiment.

4.1.3 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 number 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.

4.1.4 Question Structure

A data request is when a stakeholder asks 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 sensor type data shared with Y stakeholder for use in a Z context app"

where Sensors X can be :

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

where Stakeholders Y can be:

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

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.

4.1.5 Budget and Experiment Duration

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

4.2 Entry Phase

The entry phase denotes the first day of the experiment. Users are asked to install the application from the PlayStore.

4.2.1 Collecting General User Information

As the figure 4.3 shows, the users are asked to answer some personal non-intrusive questions. The following is asked from the users:

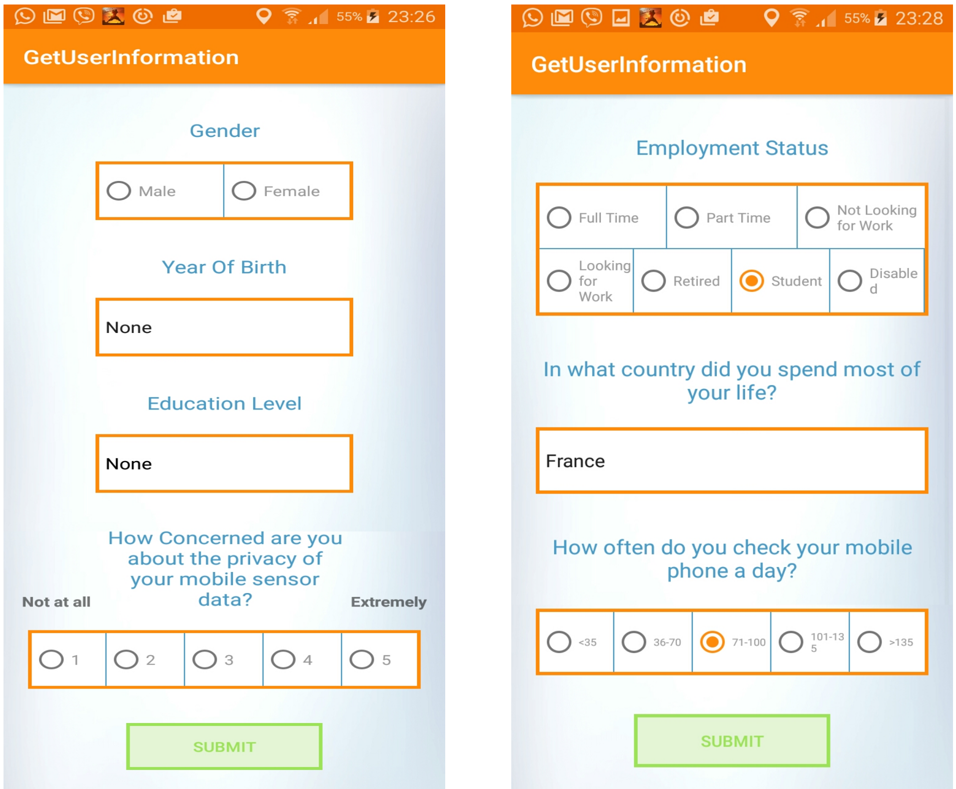
1. Gender
2. Employment Status
3. Education Level
4. Year of birth
5. Country where user has lived most of his life
6. How many time a day do you check your Mobile phone per day.
7. Kind of applications the user has in the mobile phone.

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.

4.2.2 Categorization of Features

As described in chapter 3, the 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"* means that the feature does not affect the users mobile sensor data sharing decision, whereas *"very high privacy intrusion"* refers to a feature that very much affects the sharing of mobile sensor data.

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



(a) User Information Screen 1

(b) User Information Screen 2

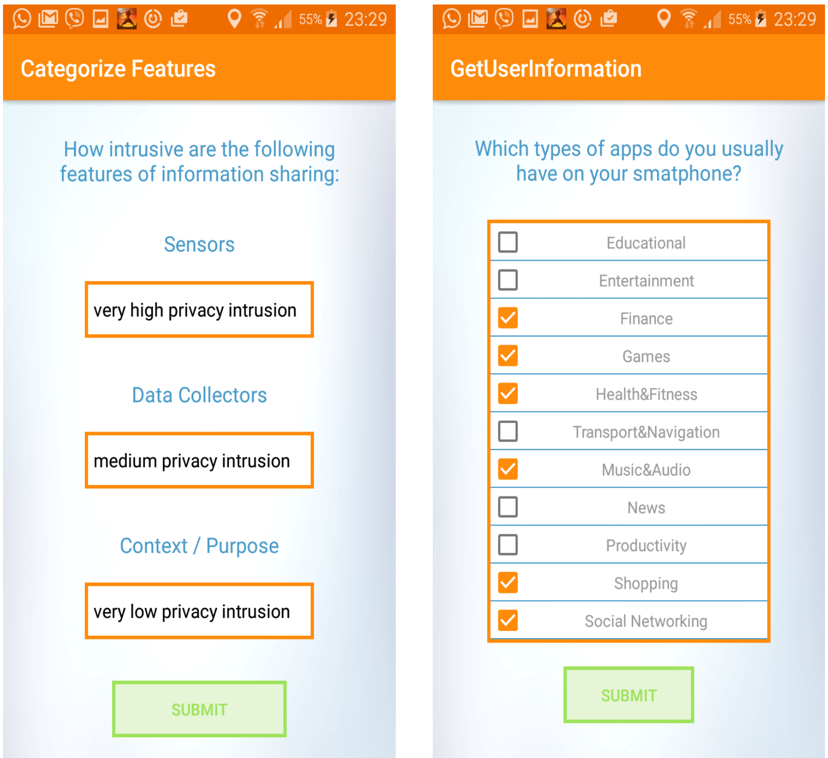
Figure 4.3: User Information Screens

provided. Users cannot navigate to the next page without filling out all of the questions.

4.2.3 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 4.2.2 . 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 the sub-features of stakeholders need to be in turn categorized in a similar fashion. This is depicted in figure 4.5b.



(a) Categorizing Features (b) User Information Screen 3

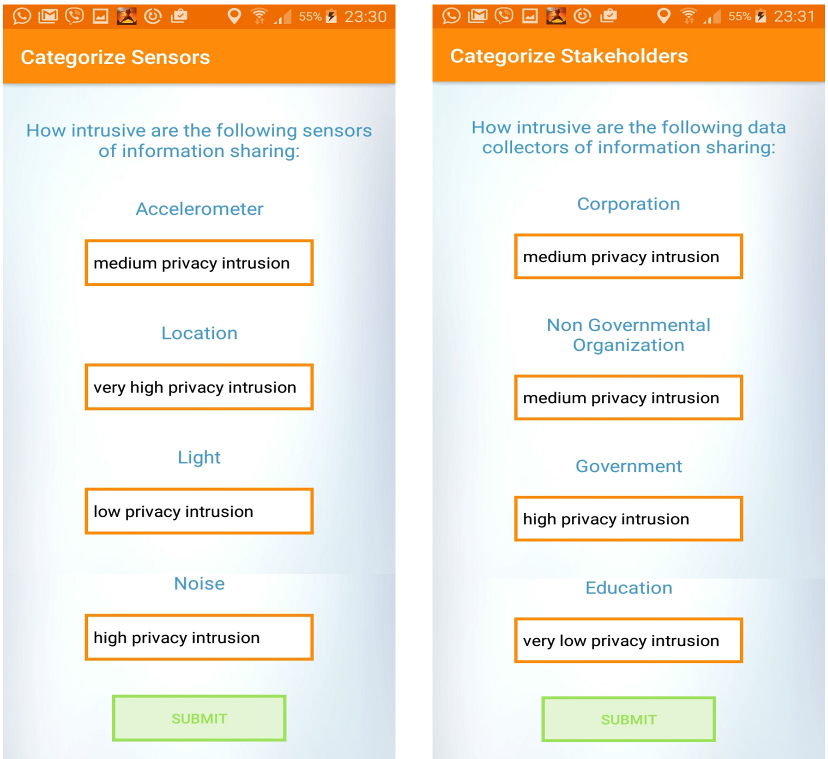
Figure 4.4: Categorization and User Information Screens

Each stakeholder type has a drop down menu each where the user can once again classify how much each of them affect data sharing. 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 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.

4.2.4 Answering Questions with No Incentives

After the categorization questions are answered and user answers are recorded, users will be presented with 64 questions. Each of these questions is a mo-



(a) Categorizing Sensors (b) Categorizing Stakeholders

Figure 4.5: Categorization of Sensors and Stakeholders Screen

bile sensor data request to the users. Users have the possibility to choose from the available five privacy options mentioned in section 4.1.3. The options are indicated as a measure of how much data users can give, ranging from maximum data to least data. The higher the privacy of 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 on top of the options 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 to the data request to the server and cannot be changed. At this stage, no indications of credit gained or privacy improvements are indicated.

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

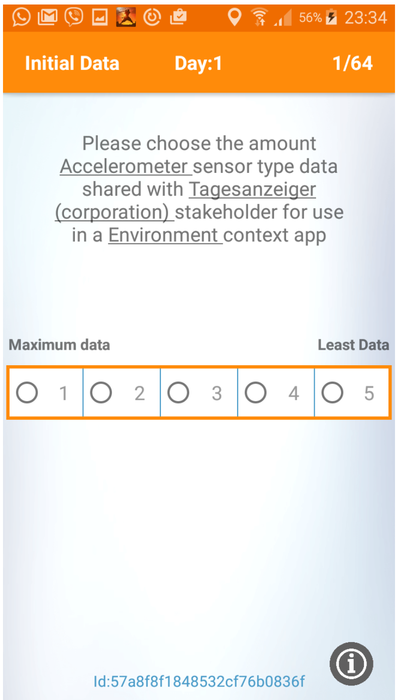
Figure 4.6: Categorization of Contexts Screen

4.3 Core Phase

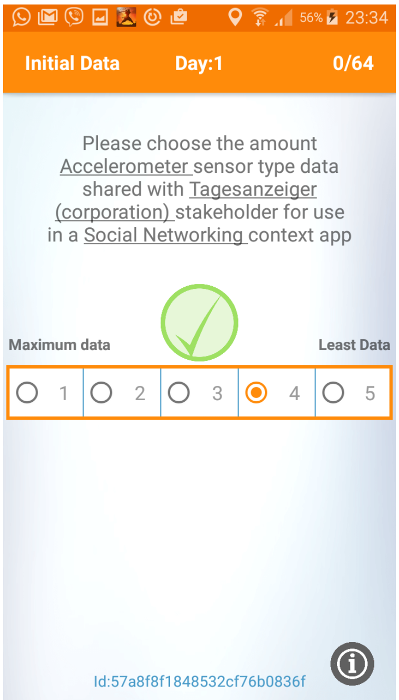
Once the entry phase is done, the user is presented with the screen shown in figure 4.8. The *"i"* button at the bottom right of the screen denoted by the number 9 is clickable. This takes the users to the FairDataShare portal. Figure 4.9 shows the homepage of the portal. Users can then click on the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

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

The unique identifier is located at the bottom of the application screen is an alphanumeric sequence denoted by 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 **??** shows what the registration page looks like. The users can use this website to see all the data collected from



(a) Question Screen



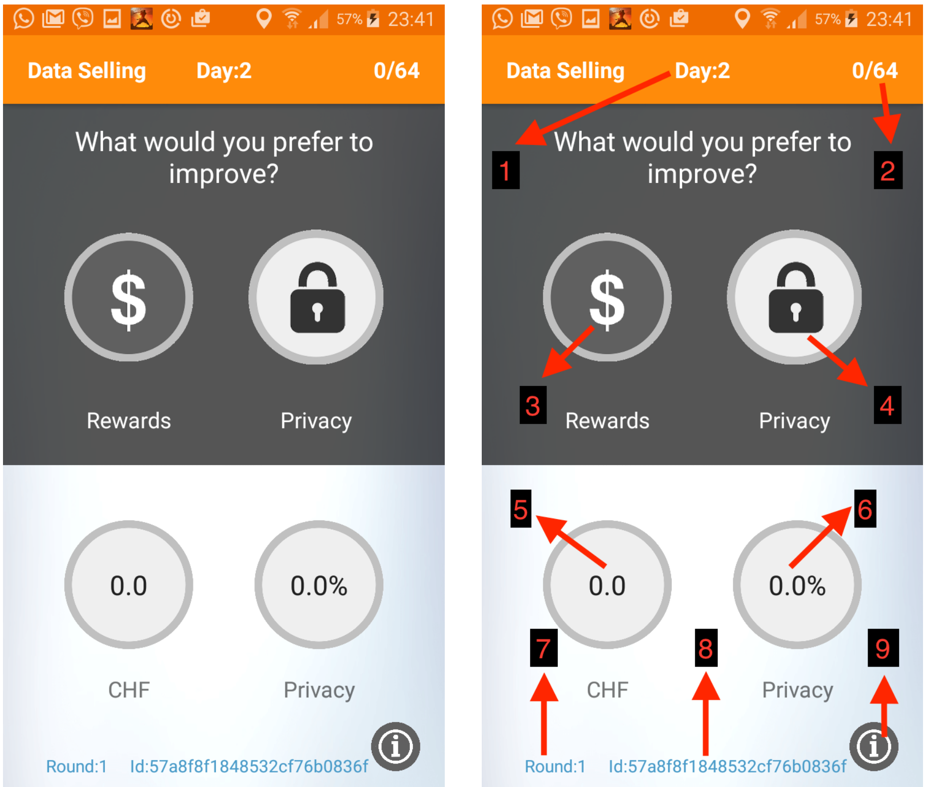
(b) Question Screen with Submit button

Figure 4.7: First Day Screen

them for all the mobile sensors. More details about the FairDataShare portal refer to the section 4.5.

The user have the possibility to login into the portal after a minimum of 24 hours after the start of the core phase to see the data that has been collected and shared with the stakeholders.

In the task-bar, the user can see the bidding day number and how many questions have been answered from the total available shown by numbers 1 and 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 entry phase is over 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 6 and 5 represent the privacy percentage and credit obtained by the user respectively. Privacy is measured in terms of the percentage of mobile sensor data not traded to the stakeholders. Credit is measured in terms of the currency Swiss Francs obtained for trading data to the stakeholders.



(a) (b)

Figure 4.8: Improvement screen

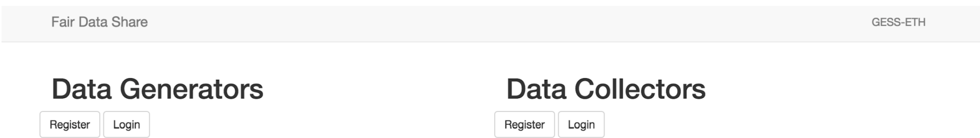


Figure 4.9: FairDataShare Homepage

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

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

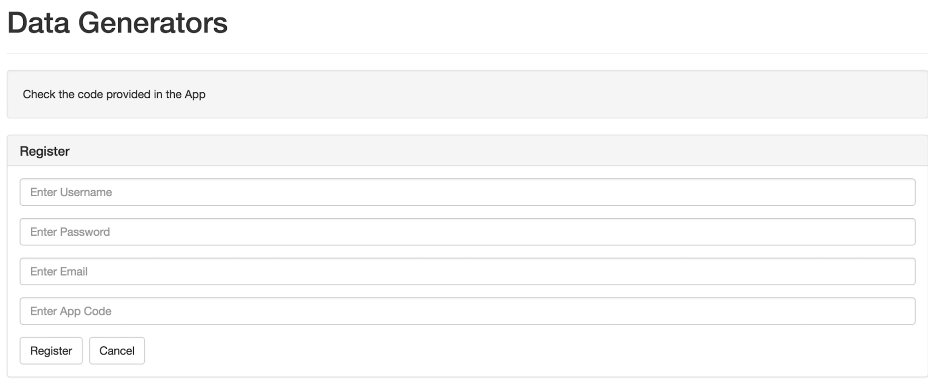
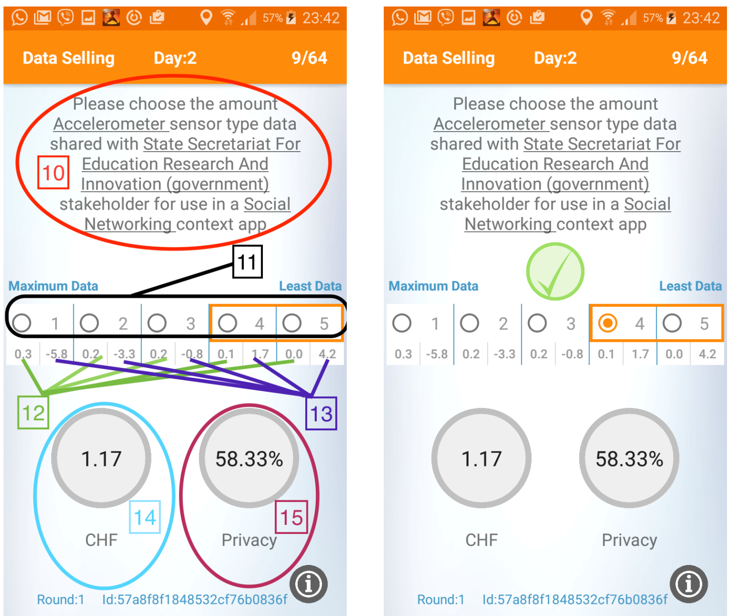


Figure 4.10: Data Generators Registration Page



(a) Bidding Screen

(b) Bidding Screen with Submit Button

Figure 4.11: FairDataShare Portal

sharing.

From the starting time of the core phase till 24 hours later marks one bidding day. Once 24 hours is over, another bidding day starts where the privacy and credit metrics are reset. The day number in the task bar is incremented by one. The user has to answer all the data requests again for this new bidding day. Previous responses to data requests are not carried over to the next day. If a data requests is not answered, it is considered that the user does not want to trade mobile sensor data for that request. Additionally, each data request carries a participation fee, this is irrespective of the amount of mobile sensor data shared, by not participating in a data request the user foregoes this credit gain. The core phase goes on for a period of 48 hours.

4.3.1 Improve Privacy or Credit

The improvement screen 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 4.3. The improve credit button should be chosen if the user is interested in maximizing the amount of credit obtainable. This uses an algorithm that uses the previous user answers to put forth a data request that can increase the credit to the maximum 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 the data request. The privacy improvement button is represented by the number 6.

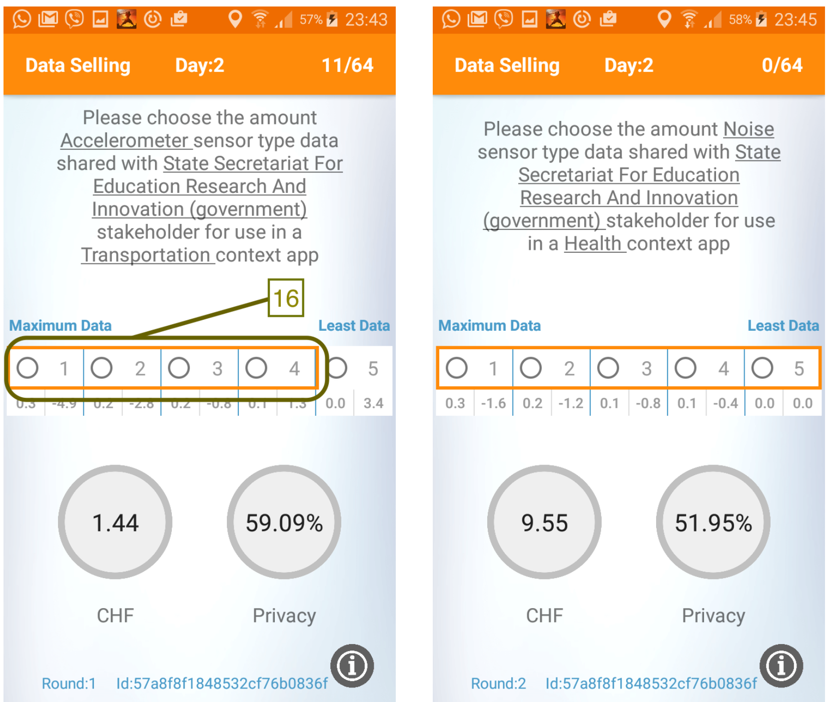
Scenario examples for each button is given in the next section after introducing the next screen in the application. For example, if a user chooses to improve the privacy, then clicks on improve privacy button and gets a data request. The user still chooses option one with maximum data sharing (least privacy) for the 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 five with least 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 chosen and the choice of the appropriate option for that data request.

4.3.2 Answering Questions with Incentives

After choosing a metric to improve, 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 options of the current data request. Once the user decides on which options to choose according to how much data wants to be traded, the users can click on the radio option as explained in section 4.1.3 and then click again on the green submit button that pops up shown in 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 the 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 previous time the data request was answered to.

4.4 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 survey [3] is linked to the user using the unique identifier assigned in the application. Once the survey is filled, the users receive their money for the entry phase, core phase and exit phase together, but only if they did not have their phones switched off throughout the experiment and participated in the core phase. This is done by checking the data collected on the server.

4.5 FairDataShare Web Portal

The FairDataShare portal [4] 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.

4.5.1 Data Generator's Portal

Once the users are registered which was explained in section 4.3, they can come back to the portal after a 24 hours period or later to view their mobile sensor data 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

[3] https://descil.eu.qualtrics.com/SE/?SID=SV_3P0ySMqNeO06v5j

[4] http://fair-data-share.inn.ac/

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 whose data 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.

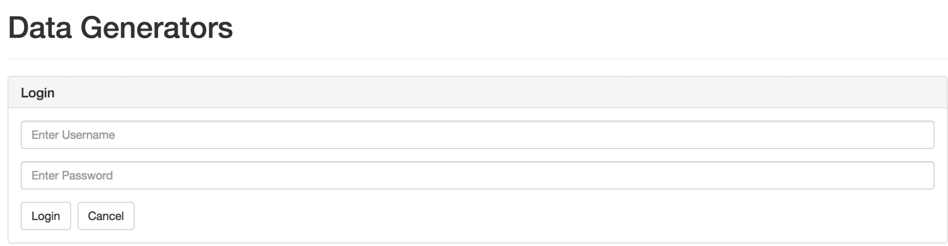


Figure 4.13: Login Page

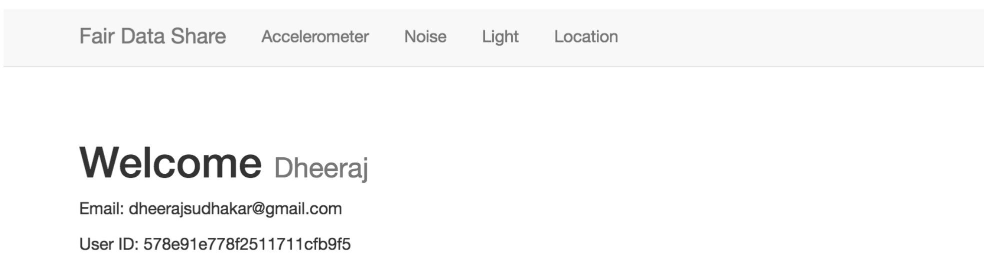
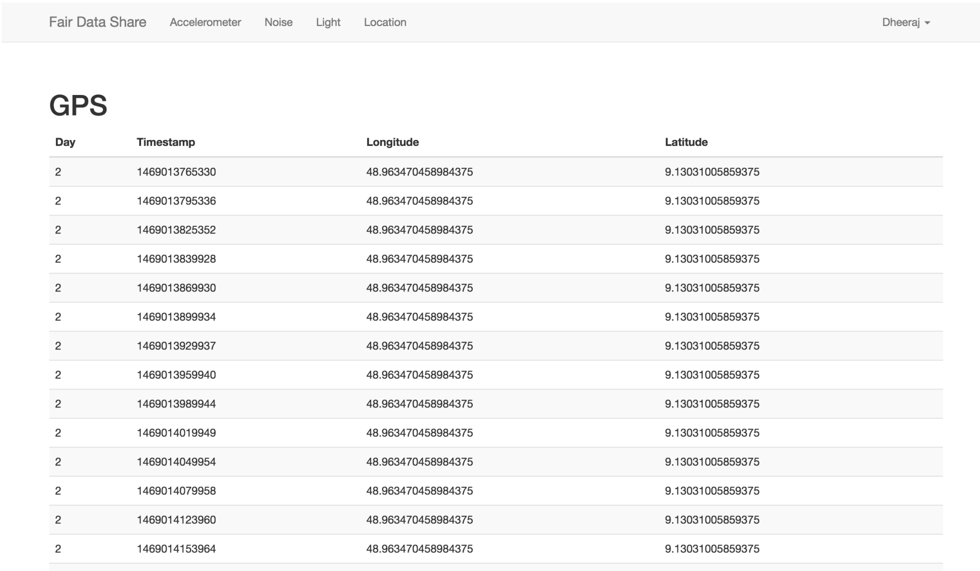


Figure 4.14: Welcome Page

Users first register as data generators as indicated in the section 4.2.4.

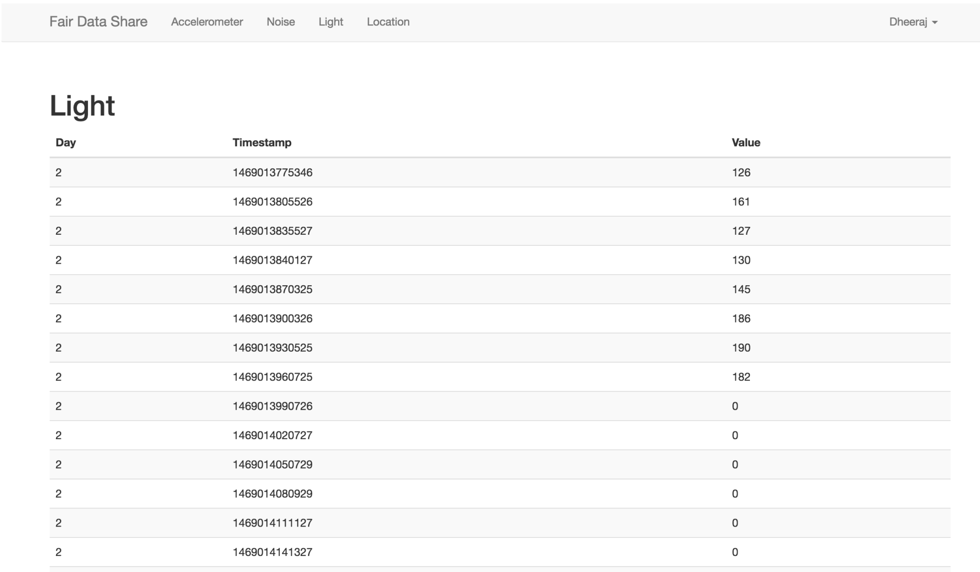


Fair Data Share Accelerometer Noise Light Location Dheeraj

GPS

Day	Timestamp	Longitude	Latitude
2	1469013765330	48.963470458984375	9.13031005859375
2	1469013795336	48.963470458984375	9.13031005859375
2	1469013825352	48.963470458984375	9.13031005859375
2	1469013839928	48.963470458984375	9.13031005859375
2	1469013869930	48.963470458984375	9.13031005859375
2	1469013899934	48.963470458984375	9.13031005859375
2	1469013929937	48.963470458984375	9.13031005859375
2	1469013959940	48.963470458984375	9.13031005859375
2	1469013989944	48.963470458984375	9.13031005859375
2	1469014019949	48.963470458984375	9.13031005859375
2	1469014049954	48.963470458984375	9.13031005859375
2	1469014079958	48.963470458984375	9.13031005859375
2	1469014123960	48.963470458984375	9.13031005859375
2	1469014153964	48.963470458984375	9.13031005859375

Figure 4.15: Location Data



Fair Data Share Accelerometer Noise Light Location Dheeraj

Light

Day	Timestamp	Value
2	1469013775346	126
2	1469013805526	161
2	1469013835527	127
2	1469013840127	130
2	1469013870325	145
2	1469013900326	186
2	1469013930525	190
2	1469013960725	182
2	1469013990726	0
2	1469014020727	0
2	1469014050729	0
2	1469014080929	0
2	1469014111127	0
2	1469014141327	0

Figure 4.16: Light Data

4.5.2 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 :

1. Company Name

Fair Data Share Accelerometer Noise Light Location Dheeraj

Accelerometer

Day	Timestamp	x	y	z
2	1469013775479	-0.5650315880775452	7.220912456512451	6.789956092834473
2	1469013805528	-0.10534487664699554	8.101978302001953	6.531382083892822
2	1469013835557	0.38307225704193115	6.368576526641846	7.5848307609558105
2	1469013839977	0.3734954595565796	6.196194171905518	7.785943984985352
2	1469013870007	-0.038307227194309235	6.339846134185791	7.422025203704834
2	1469013900047	-0.05746084079146385	10.141838073730469	-0.13407529890537262
2	1469013930077	0.009576806798577309	10.141838073730469	-0.15322890877723694
2	1469013960107	0.028730420395731926	10.170568466186523	-0.14365209639072418
2	1469013990180	-6.847416877746582	-6.119579315185547	-3.246537446975708
2	1469014020378	-6.89530086517334	-6.100425720214844	-3.265691041946411
2	1469014050578	-7.268796443939209	-6.244078159332275	-1.6376339197158813
2	1469014080579	-7.393294811248779	-6.205770969390869	-1.3503297567367554
2	1469014110779	-7.2113356590271	-6.358999729156494	-1.1300631761550903
2	1469014140798	-7.364564418792725	-6.224924564361572	-0.6129156351089478

Figure 4.17: Accelerometer Data

Fair Data Share Accelerometer Noise Light Location Dheeraj

Noise

Day	Timestamp	Rms	Spl	Bands
2	1469013765685	146.42578125	65.35355377197266	0.0,2.4057037E-5,6.9086714E-6,6.0710937E-7,2.9846692E-7,3.0724346E-5,2.7675502E-5,6.437194E-4,6.671106E-5 6,
2	1469013795661	126.37060546875	64.0741195678711	0.0,1.7532275E-5,7.1945624E-6,5.1027865E-7,1.262046E-7,1.1799247E-5,3.324563E-5,5.093394E-4,4.5384477E-5, 7,
2	1469013825662	241.427001953125	69.69691467285156	0.0,1.9904654E-5,6.4675687E-6,5.0418254E-7,3.6831852E-7,4.132275E-5,1.20294404E-4,2.49995E-4,2.0823021E-4 4,2.1346188E-5,
2	1469013840254	503.76220703125	76.08570861816406	0.0,2.13898E-5,2.8374447E-5,3.0414352E-5,3.4566263E-5,9.557794E-5,0.0050782408,0.0015603136,0.0013168416, 5,3.2593807E-6,
2	1469013885667	524.6171875	76.43804931640625	0.0,4.706202E-5,7.744441E-6,1.4033919E-6,1.2413806E-6,1.418927E-4,0.0063855834,0.0031244773,0.0013166494, 5,1.3288992E-5,
2	1469013915660	172.947021484375	66.7994613647461	0.0,1.6703209E-5,6.2209865E-6,3.9988538E-7,4.921786E-7,1.0719621E-5,3.601787E-5,5.6616304E-4,1.1179377E-4
2	1469013945661	278.451171875	70.9361801147461	0.0,3.054911E-5,5.686226E-6,4.36027E-7,1.0158309E-6,4.9987982E-5,4.676138E-5,0.0017947021,9.102744E-4,2.33
2	1469013975678	1768.630615234375	86.99394226074219	0.0,2.2745946E-5,2.0635262E-5,2.6642087E-5,2.901498E-5,9.767513E-5,3.8286814E-4,0.002663158,0.13813451,0.0 5,
2	1469014005674	182.28271484375	67.25611114501953	0.0,2.0691217E-5,5.0631206E-6,5.293276E-7,5.824385E-7,3.2161366E-5,2.3359193E-5,1.09764966E-4,2.6371961E- 5,
2	1469014035678	116.080810546875	63.33640670776367	0.0,1.3490896E-5,6.9372954E-6,5.145818E-7,7.5139013E-7,1.004016E-5,2.0504498E-4,2.4968915E-4,3.118082E-5, 6.

Figure 4.18: Noise Data

2. Email
3. Stakeholder Category
4. Company Website

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

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

Once these details have been filled in, the stakeholder can click on the register button. Once registered, the stakeholder can login like 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 this 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.

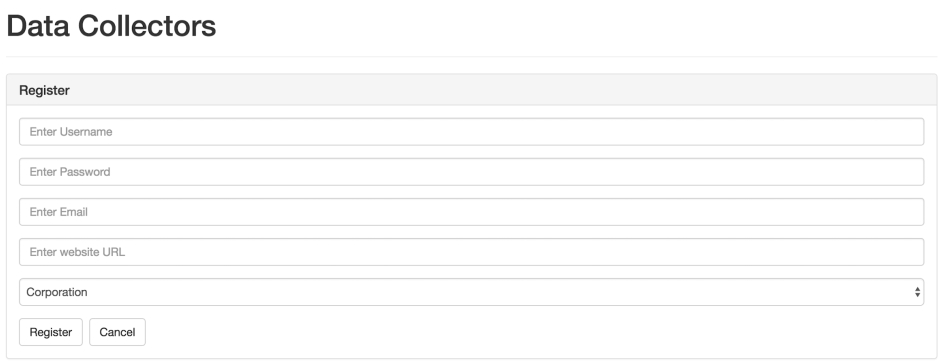


Figure 4.19: Registration Page

Figure 4.20: Login Page

Figure 4.21: Data Collectors Data Retreiving Page

Chapter 5

Explanation of the Mobile Application

This chapter explains the details behind the making of 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 will explain integral parts of the server and client of the mobile application. A gist of the architecture is shown in the figure 5.1. It shows the mobile application represents the user participating in the experiment. As the experiment goes on, mobile sensor data and responses to the data requests which are collected are periodically sent to the Kinvey Data Store.

The users can login into 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 in turn structures the data so it can be readable, and pushes it to the user to see on the portal. The concept is similar for the Stakeholders, except they can only access the portal through the computer and not the mobile application.

[1] https://kinvey.com

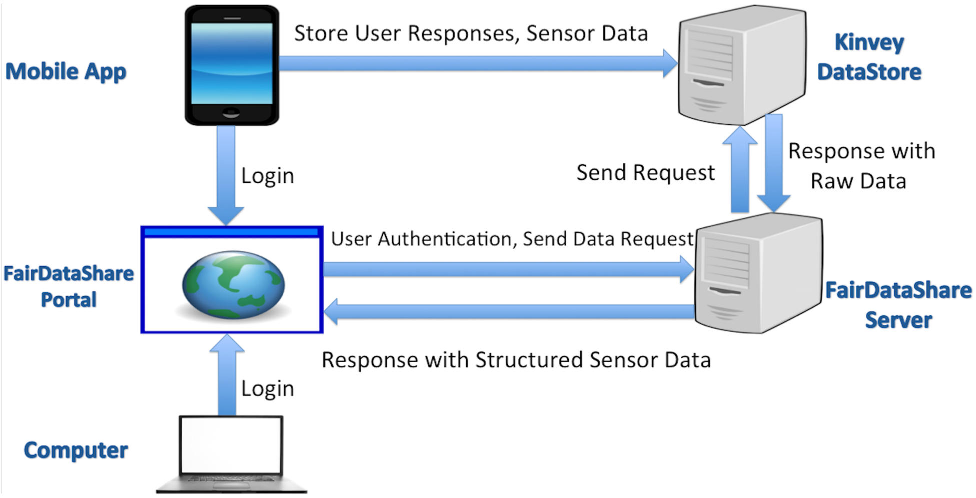


Figure 5.1: Conceptual Diagram of Mobile Application Architecture

5.2 The Mobile Application

The mobile application was 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), whereas larger related data is stored in the database. The following paragraphs will explain each table present in this application followed with 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`'s table schema. This table stores each possible data request with its features such as with its 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*,

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

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

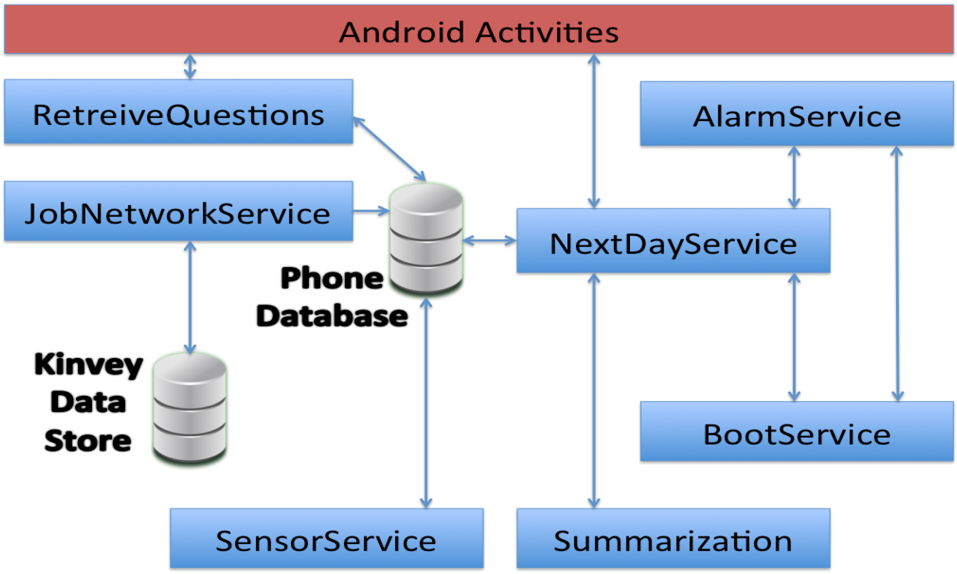
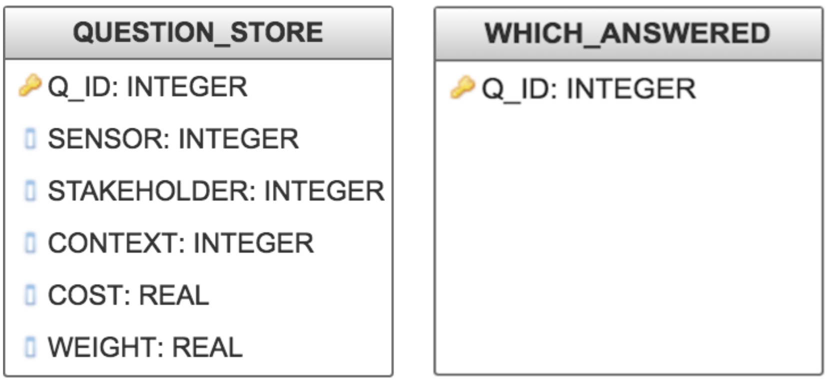


Figure 5.2: Interaction of Components of Mobile Application

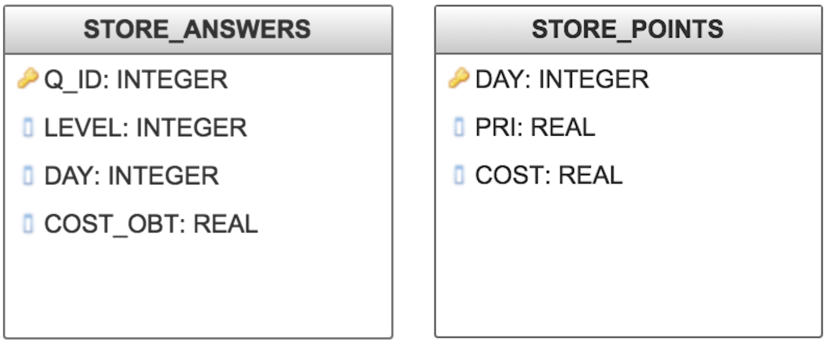


(a) Table Schema of `QUESTION_STORE` (b) Table Schema of `WHICH_ANSWERS`

Figure 5.3: Table Schemas

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 has been 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. It makes sure that all questions are answered before answering them for a second time. This data is not sent to the server.



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

Figure 5.4: Table Schemas

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 of all the user responses stored in this table. Only the most recent responses are stored in this table. The content of the table is not sent over 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.

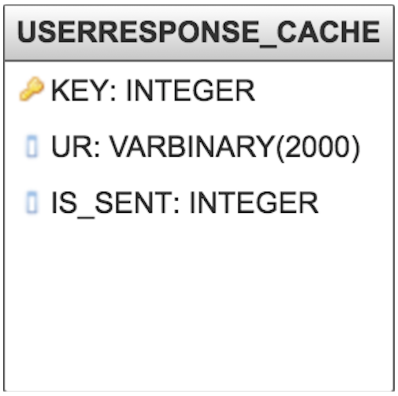


Figure 5.5: Table `USERRESPONSE_CACHE` Schema

Figure 5.5 depicts the `USERRESPONSE_CACHE` tables's 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
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 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. Figure 5.6 and 5.7 show the table schemas for data storage of the following sensors:

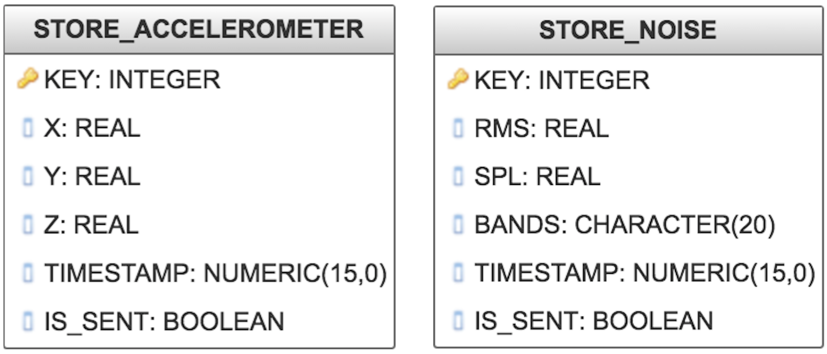
1. Accelerometer in the `STORE_ACCELEROMETER` table
2. Noise in the `STORE_NOISE`
3. Location in the `STORE_LOCATION`
4. Light in the `STORE_LIGHT`

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

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

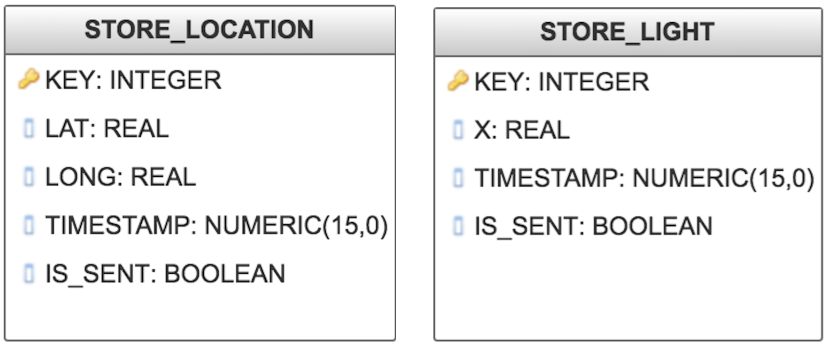
5.2.2 Alarms and Notifications

Every bidding day where the user answers data requests lasts for a period of 24 hours. After one bidding day is over, the system needs to be informed in a timely manner to perform some application critical functions. The functions



(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

performed are explained in detail in section 5.2.2. To inform the system of such an event Android provides the functionality in the form of alarms.

Alarms can be set to go off just once or in a repeated fashion to trigger tasks. Unfortunately, the alarms provided by Android are not exact for some versions [4], in the sense that they are triggered around that time set but not exactly at that time to optimize the battery, and can be delayed upto 24 hours. Hence, it is decided to set the repeating alarms manually.

The first time the application opens the alarm is set to ring in exactly 24 hours, but things change 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:

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

1. The phone is rebooted.
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]. Alarms do not execute when the phone is switched off. Hence, when the phone switches on, BootReceiver service of the application is triggered with pseudocode shown in 1. This checks whether an alarm has been missed, if it has been missed 200 seconds is given for the phone to stabilize after boot before triggering tasks. Otherwise, a new alarm is set using the pseudocode shown in 2. To set an alarm we need the time difference between now and when the alarm should ring. After that is calculated, the alarm is set.

Algorithm 1 BootService Algorithm

```
procedure BootService
    now ← current timestamp
    i ← timestamp of last triggered alarm
    if now − i < 86400 then
        Call SetAlarmLater()
    else
        Set alarm in 200 seconds
```

Algorithm 2 Alarm Algorithm

```
procedure SetAlarmLater
    now ← current timestamp
    i ← timestamp of last triggered alarm
    latertime ← i + 86400
    latergap ← latertime − now
    Set Alarm in latergap seconds
```

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 5. To start with the 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 answered

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

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 day number is incremented by one to denote the next bidding day.

Algorithm 3 NextDayService Algorithm

procedure NextDayService
    Store *Privacy, Credit, Day* in *STOREPOINTS*
    Send *Privacy, Credit, Day* to Server
    $Privacy, Credit, Round, CurrentQuestion \leftarrow 0$
    $Day \leftarrow Day + 1$
    Store current time
    Call *Summarization()*
    if $Day > End$ **then**
        End experiment
    else
        Update user interface elements

The current time of executing the alarm is saved in case the phone is rebooted or switched off. 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. Finally it needs to be to checked if the experiment is over or not and update the user interface 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 requests need to be fetched from the database in two scenarios :

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

In the first bidding day, once a data request has been answered the next one is fetched sequentially from the database. This just 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 data request from table `STORE_ANSWERS` where user has answered with lowest privacy

2. **Improve Credit** - Obtain data request from table STORE_ANSWERS where user has answered with highest privacy

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 was 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 4.1.3.

Every option of a data request has an associated percentage of data that is given away as described in 3.2.6. According to the percentage of data given away, the total privacy is calculated for each possible option. The difference between the current privacy and each possible total privacy is calculated and indicated under each option. This gives an indication to the user as to what each option will do to the metrics.

5.2.4 Recording User Choices

The 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 so as to be stored easily in the table as is without transformation. When the JobNetworkService described in 5.2.6 is called, the class object is sent as it is to the server after converting it back to an object.

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
4. Light sensor

A sensor service is triggered when the application is installed and is stopped when the experiment is over. This collects data from every sensor 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 wishes of 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- All data is sent 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 that should be sent to the server is set with $ISSENT = 1$, and all others that have value $ISSENT = 0$ are ignored.

Algorithm 4 Summarization Algorithm

```
procedure SUMMARIZATION
    for each sensor do
        Fetch sensor data from sensor table
        level ← Fetch user privacy level
        if level ← 1 then
            Set all ISSENT ← 1
        else if level ← 2 then
            for 3 out of every 4 records do
                ISSENT ← 1
        else if level ← 3 then
            for 1 out of every 2 records do
                ISSENT ← 1
        else if level ← 4 then
            for 1 out of every 4 records do
                ISSENT ← 1
        Delete all entries with ISSENT ← 0
        Update Database
```

5.2.6 Server Synchronization

User responses and sensor data need to be sent to the server. This is done periodically every 5000 seconds in order to free up space on the phone whenever the 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 yet to the server. Such data is sent, and when an acknowledgement is received from the server, this data is deleted from the table.

Algorithm 5 JobNetworkService Algorithm

```
procedure NetworkService
    Fecth data from USERRESPONSECACHE
    for each record do
        if ISSENT == 1 then
            Send record to Server
            if SUCCESS then
                Delete record
    for each sensor do
        Fecth data from sensor table
        for each record do
            if ISSENT == 1 then
                Send record to Server
                if SUCCESS then
                    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 [7]. For the purpose of this application the backend has been used to store data and for some business logic implementations in javascript.

Security

All communications from the application to the server is encrypted using TLS/SSL encryption [8] 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 in Kinvey is MongoDB so instead we have collections on the server. When the user starts the application, general personal information is entered as explained in 5.2.1. This data is stored in the collection UserInformation with the schema shown in the screen shots A.1 and A.2.

Once this is done, users have to categorize the various Features, Sensors, Stakeholders and then the various Contexts. This information is sent to the

[6] http://kinvey.com/

[7] https://en.wikipedia.org/wiki/Mobile_backend_as_a_service

[8] Kinvey white paper : KINVEY CLOUD SERVICE: SECURITY OVERVIEW 2014

server in collections named Features, Sensors, Stakeholders and Contexts. Schema is shown in A.3, A.4, A.5 and A.6 respectively.

All the data stored locally on the mobile phone which is sent by the JobNetworkService explained in section 5.2.6 is received by Kinvey. User responses are stored in the collection UserResponse shown in A.7 and A.8.

The sensor data sent by the JobNetworkService is stored in collections named after the sensors themselves. The schema of the tables is shown in figures A.9, A.10, A.12 and A.11.

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

Finally, the collection Score shown in A.14 stores the total privacy, 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 are:

1. Script to find the privacy preferences of the users
2. Script to perform data summarization 4

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 plus 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 this criteria and extract the privacy level.

Once the privacy level is known, summarization on user data is performed. Data has been taken from the user with a certain summarization, and if the summarization level is lower than the privacy level extracted, further summarization needs to be done. The pseudocode is shown in 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 safely store the usernames, passwords of the users and

Algorithm 6 Server Summarization Algorithm

```
procedure SUMMARIZATION
    data ← sensor data from collection
    if summarizationlevel == privacylevel then
        Return data
    else
        skip ← summarizationlevel − privacylevel + 1
        for every skipnumber records out of 4 do
            Delete record from data
    Return data to portal
```

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 was constructed using simple Html and css. All screenshots of the portal including detailed information is provided in chapter 4.

Chapter 6

Pre-Survey and Experiment Findings

The following chapter will give 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 feedback received from the participants. This chapter puts forth all that was learnt from the above mentioned.

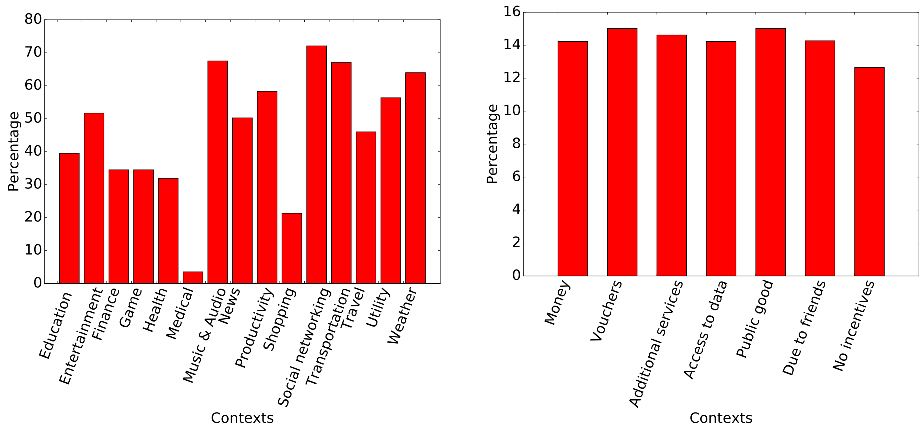
6.1 Overview of the Pre-Survey Data

The survey had 199 participants. Participants were 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 will be introduced.

Out of the total participants 63.64% are male and 36.36% are female. The mean birth year was found to be 1985. The demographics of the participants is explained in table B.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 6.1a depicts the various kinds of applications the population has on their mobile phones. As seen, the most popular applications are social networking, transportation and music applications. Plot 6.2b 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.

In figure 6.2a, the bars with the color dark blue represents for each privacy



(a) Applications in the Mobile Phone (b) Mobile concern and Incentives

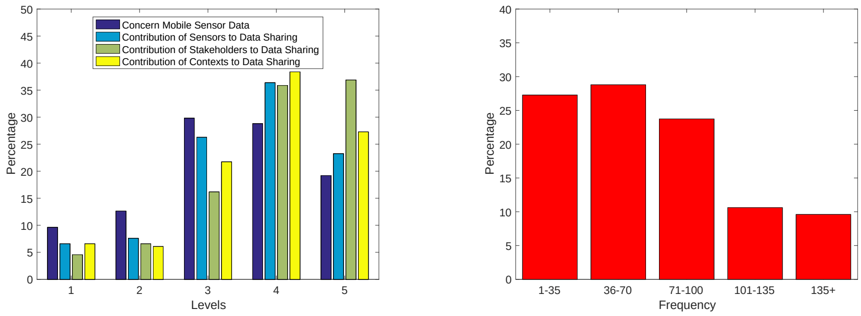
Figure 6.1: Table Schemas

intrusion level the percentage of participants. The scale of the levels are from one to five, one indicates that people do not care to five indicating that people care a lot about mobile sensor data. As observed, most people have answered level 3, meaning they care to a medium level and 77.28% care to a level of 3 and above.

In the same figure, the color light blue shows the amount of contribution that sensors in general have in the data sharing decision. Level one indicates that it contributes none, level 5 means that it contributes a lot. It is observed that most people find that sensors contribute to a level of 4. Similarly, the color green shows the amount of contribution stakeholders in general have in the data sharing decision. It can be observed that most people find that it contributed to a level of 4. Lastly, the color yellow depicts the level of contribution contexts have in the data sharing decision. Most people feel that it contributes to a level of 5.

From the above, it is understood that in general, more than the sensor, the more important thing in a data request is for what purpose the data is being collected followed by the entity to whom the data is traded with. The privacy intrusion levels of the participants for the individual different sensors, stakeholders and contexts has been explained in chapter 4.

The aim is to study the relationship between mobile sensor data sharing and incentives. Figure 6.1b shows the percentage of people that would give data for the corresponding incentives inspite of a certain percentage of mobile concern. The higher the value of the bar, higher the mobile concern and the probability that users will share data for that incentive. As seen, ”No incentives” has the lowest value of around 12%, indicating that even tough users are less concerned about their data, lesser number of users would share data for free.



(a) Graph depicting the concern of Mobile Sensor Data and the contribution of various features to the data sharing decision

(b) Frequency of Mobile Phone Usage

Figure 6.2: Table Schemas

6.2 Pre-Survey Methodology and Findings

All the results presented below were performed on the data by performing the following changes to the data:

1. Rows with empty fields were removed
2. Rows with spurious data were removed
3. Data was scaled or normalized when necessary

Other than the above, the data was not manipulated. Outliers were not excluded either.

Perception of Individual Sensor Grouped on the Intrusion of Sensors in General

We try to examine here if the perception of intrusion of Sensors in general can affect the way a person views the individual sensors themselves. In other words, we try to examine if there is a significant difference in perception of each sensor depending on the perception of the Sensors as a whole. For this, we grouped the survey data based on the responses to question 10, which examines the contribution of sensors in general for a data request. Since there are 5 possible responses to this question, this makes 5 individual groups from 1 to 5.

The reason why this is necessary is because, if users who view Sensors in general view in the same manner all the individual sensor sub-features, there would be no need to ask users to individually categorize each sub-feature, a profile can be formed just based on feature categorizations.

Five groups of users are formed from one to five, who view Sensors as a whole with different intrusion levels and their perception of each of the individual sensors is now to be compared. Before going into the comparison, we try to understand the properties of the data to analyze. Essentially, we try to examine if each of the groups formed come different populations and for this the most popular statistical tests is the one-way ANOVA. To perform a one-way ANOVA test, the data should be[1]:

1. Normally distributed
2. Homoscedastic
3. Ordinal or continuous

Since the data of the survey is discrete and follows the Likert Scale with options from 1 to 5, it gives a skewed normal distribution. The scale used to collect data is in the ordinal form. Additionally, the variances of values within the groups formed are not similar. The one-way ANOVA test is quite robust to heteroscedacity, as long as the maximum variance among all groups is less than four times the group with the lowest variance. Accounting for all the violations, we instead opt for non-parametric tests such as the Kruskal-Wallis H-test and the Dunn's test which only assumes the following[2] :

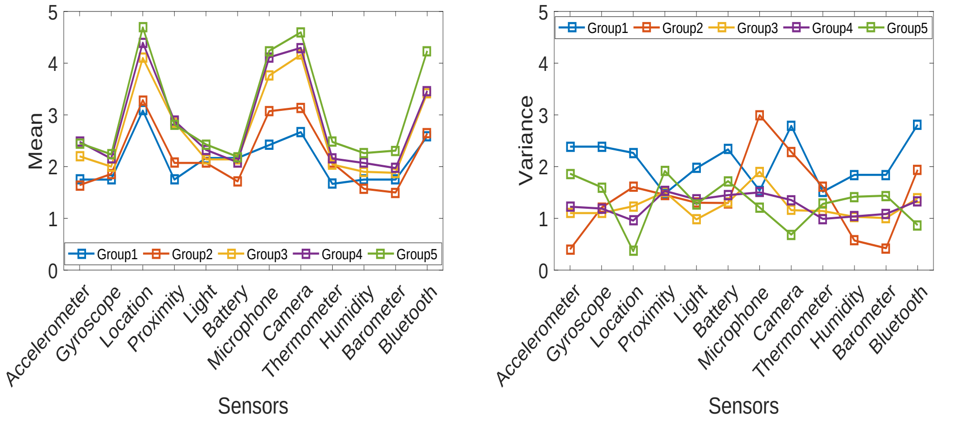
1. Groups are independant from one another
2. All observations are independant
3. The dependant variables should be in the ordinal scale or continuous

The above tests do not make any assumptions about the distribution of the data and are robust to heteroscedastic data.

Group 1 to 5 which consist of users with different perception of Sensors have 13, 14, 50, 71 and 42 people in each group respectively. For in depth information of the composition of each group in terms of employment, education, gender and birth year please refer to tables B.2,B.5,B.3,B.4. Figures 6.3a and 6.3b depict the mean and variances for each of the individual groups.

We start by performing the non paramteric Kruskal-Wallis test for each individual sensor. The value of alpha assumed here is 0.05. The null hypothesis states that all the groups perceive all the sensors in the similar way. This

[1] https://fr.wikipedia.org/wiki/Analyse_de_la_variance
[2] https://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance



(a) Mean of each Group for each Sensor (b) Variance of each Group for each Sensor

Figure 6.3: Table Schemas

means they come from the same population. The alternative hypothesis is that the groups perceive each sensor in a significantly different way and hence their populations are distinct. The table 6.1 depicts the p-values obtained from the statistical test.

Table 6.1: Kuskal-Wallis Test For Sensors

Sensor	p-value
Accelerometer	0.0151
Gyroscope	0.2959
Location	1.0664e-05
Proximity	0.0147
Light	0.6933
Battery	0.6950
Microphone	3.0070e-04
Camera	2.1191e-05
Thermometer	0.0693
Air Humidity	0.1292
Barometer	0.0949
Bluetooth	3.4877e-05

Accelerometer, location, proximity, microphone, camera and bluetooth are the sensors for which the groups opinions vary significantly. On these sensors, we proceed with a post hoc test for further investigation by performing a pariwise Dunn's test to examine if there is an actual significant difference between the groups and if so between which groups. The sensors with p-values with less than 0.05 are examined in more detail and the p-values are presented in table 6.2. The table shows the results for each pairwise test done, with the p-values adjusted using the Bonferroni Method[3]. The reason for choosing to adjust the p-values is that repeated experiments can increase the chances of accepting the alternative hypothesis so p-values are adjusted according to the number of experiments performed. 10 experiments are performed per sensor.

For the accelerometer and proximity sensors, it is seen none of the pairwise groups have a significant difference from each other. This means that even tough the groups perceive sensors differently in general, they all view accelerometers and proximity sensors in a similar way. This is reinforced by figures 6.4a and 6.4c, where it is seen that there is significant overlap between the population of the groups.

For the location sensor, it is observed that groups (1,4), (1,5), (2,4), (2,5) have a significant difference. This can be attributed to the fact that since the groups are formed from the perception of people of the sensors feature, the difference in perception between group 1 and group 5 will be larger than between group 1, group 2 and group 1, group 3 since they are not much apart in the scale. Figure 6.4b depicts the spread of the group populations for the location sensor. As found groups (1,4), (1,5), (2,4), (2,5) have the least amount of overlap.

For the microphone sensor, it can be seen that groups (1,3), (1,4) and (1,5) are significantly different from each other. This goes to show that if people rate sensors as even a little intrusive, they all rate the microphone's in a significantly different way than the people who rate sensors as non-intrusive. Figure 6.4d depicts the spread of the population of the groups.

For the camera sensor, it can be observed that groups (1,4), (1,5), (2,4), (2,5) have a significant difference in their perception of the intrusion. Similar to the location sensor, people with perception of sensors in general with a lower intrusion level have significantly different responses to the camera intrusion than the people who rate sensors with more intrusion. Figure 6.4e depicts the spread of the population of the groups for the camera sensor.

For the Bluetooth sensor, there is a significant difference between groups (1,5), (2,5), (3,5) and (4,5). This shows that responses by people who find sensors extremely intrusive is different from the rest of the groups. This fact is reinforced by figure 6.4f where it is seen that group 5 has least overlap with other groups.

It can be concluded that some sensors such as the accelerometer, proximity, gyroscope, light, battery, thermometer, humidity, barometer are sensors

[3] https://en.wikipedia.org/wiki/Bonferroni_correction

where each of the groups do not have significantly different opinions of intrusion. Sensors such as the camera, microphone, location and bluetooth are where opinions differ between groups. These sensors are found to be on average more intrusive with intrusion values of 4.23, 3.87, 4.14 and 3.52 respectively on a scale of five. Groups with larger differences between them (such as groups 1 and 4) tend to have larger differences in opinions for some sensors generally considered intrusive. Groups with smaller differences between them (such as group 1 and 2) tend to have similar opinions.

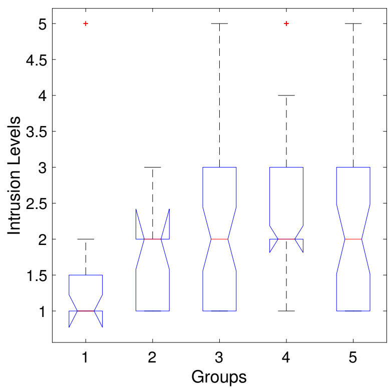
Even tough users tend to see sensors as a whole in a different light, all groups can be incentivized similarly for lower intrusion sensor. For higher intrusion sensors, groups need to be incentivized differently, with higher intrusion groups receiving more incentives.

Table 6.2: Dunn's Test For Sensors

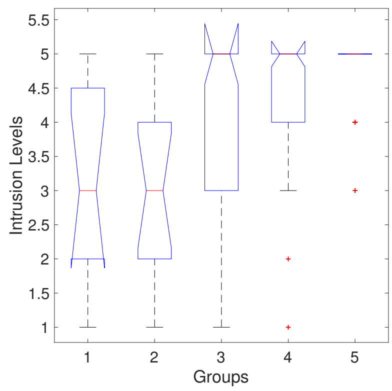
Groups	Accelerometer	Location	Proximity	Microphone	Camera	Bluetooh
(1,2)	1.0000	1.0000	0.9992	0.8365	1.0000	1.0000
(1,3)	0.4207	0.2084	0.0699	0.0365	0.0732	0.7825
(1,4)	0.0595	0.0125	0.0513	0.0012	0.0048	0.6442
(1,5)	0.2054	0.0010	0.1191	0.0009	0.0007	0.0029
(2,3)	0.6548	0.1774	0.3713	0.8927	0.1694	0.5921
(2,4)	0.1185	0.0077	0.2617	0.2287	0.0123	0.4270
(2,5)	0.3659	0.0005	0.5184	0.1597	0.0018	0.0007
(3,4)	0.8989	0.7642	1.0000	0.8052	0.9040	1.0000
(3,5)	0.9997	0.0869	1.0000	0.6390	0.2947	0.0066
(4,5)	0.9998	0.8360	1.0000	1.0000	0.9617	0.0059

Perception of Individual Stakeholders Grouped on the Intrusion of Stakeholders in General

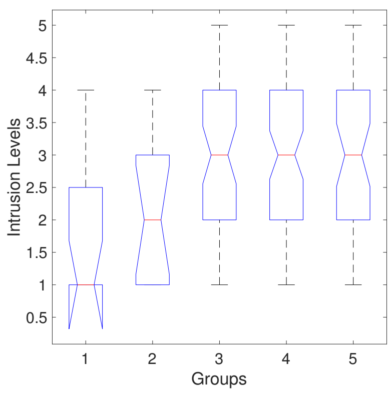
In this section, we try to see if the intrusion level perception by people of Stakeholders in general is linked to the intrusion of the individual stakeholders. In other words we examine the significant differences between groups formed by using question 12's responses (which looks into the perception of users for stakeholders in general) on the perception of each individual stakeholder. Since there are 5 different responses to question 12 ,this results in five independent groups. The groups 1 to 5 have 7, 11, 32, 69 and 70 people in each group respectively. More detailed information about the employment, education, gender and age distribution in the groups is given in tables B.6, B.9, B.7 and B.8. The mean and variances of each group formed is depicted in figures 6.5a and 6.5b.



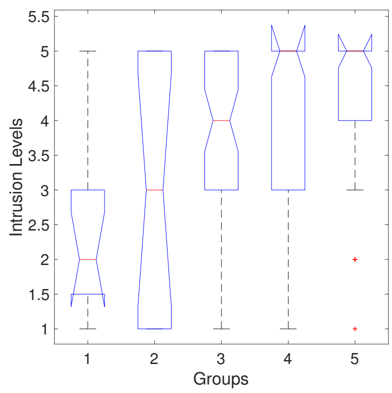
(a) Accelerometer



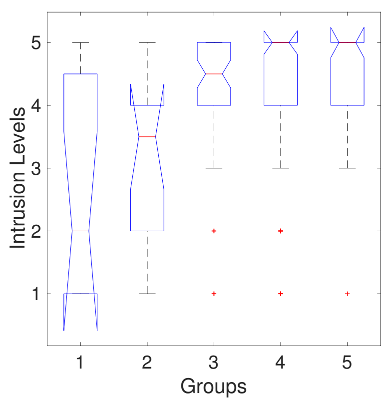
(b) Location



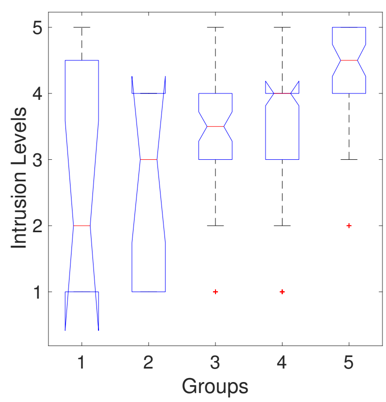
(c) Proximity



(d) Microphone

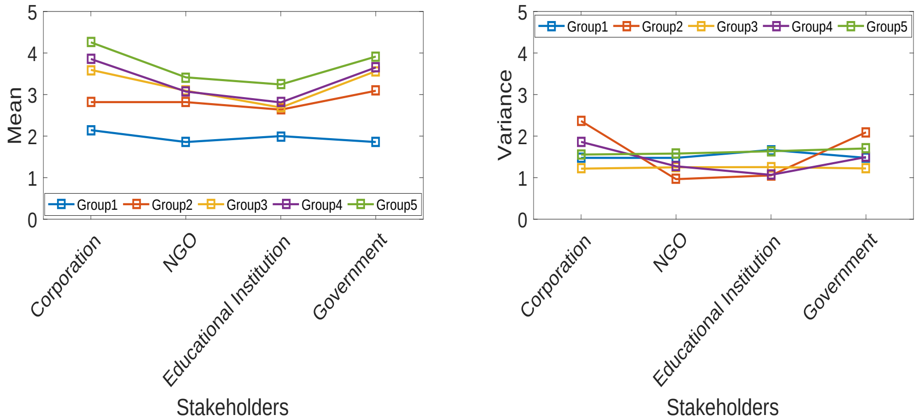


(e) Camera



(f) Bluetooth

Figure 6.4: Box Plots Indicating the Population Spread of each Group



(a) Mean of each Group for each Stakeholder (b) Variance of each Group for each Stakeholder

Figure 6.5: Mean and Variance of Groups

To start, we examine all the groups simultaneously for all the individual stakeholders.The Kruskal-Wallis H-test since the data is discrete and not normally distributed. Further detail about the reason the test is chosen is provided in the previous section 6.2. The null hypothesis states that all the groups rate the intrusion of each stakeholder in a similar way. The alternative hypothesis is that the groups rate the intrusion of each stakeholder in a significantly different way.

The resulting p-values of this test are displayed in table 6.3. As it is seen, the test pronounces that all the groups are significantly different at an alpha with 0.05 for all stakeholders.

Table 6.3: Kuskal-Wallis Test for Stakeholders

Stakeholder	p-value
Corporation	2.1432e-05
Non-Governmental Organization	0.0221
Educational Institution	0.0396
Government	0.0024

This prompts us to take a closer look at which of the groups are significantly different from each other for each stakeholder. For this, we continue the experiment by performing a Dunn's Test with p-values adjusted by the Bonferroni Method. The results from the tests are shown in table 6.4. The test was done for all stakeholders since the Kruskal-Wallis test denoted that all groups differ significantly for all stakeholders.

Looking at the stakeholder corporation, it is seen that groups (1,4), (1,5) and (2,5) differ significantly. This goes to show that groups with larger differences in their outlook to stakeholders as a whole view corporations in a significantly different way. Figure 6.6a shows the spread of each group for corporations. It is observed that adjacent groups have a large overlap.

For the non-governmental organizations, only the groups (1,5) differ significantly. This goes to show that groups that do not find stakeholders intrusive and groups that find stakeholders very intrusive rate the intrusion of Non-Governmental Organization in significantly different ways. Additionally, it shows that groups with large differences in the perception of stakeholders have significant differences in their perceptions on NGOs. Figure 6.6b reinforces this claim.

For Educational Institutions, none of pairwise comparisons have p-values below 0.05. This goes to show that the intrusion of Educational Institutions by all groups does not differ significantly, that is all groups have similar opinions. Figure 6.6c shows that all groups have significant overlap.

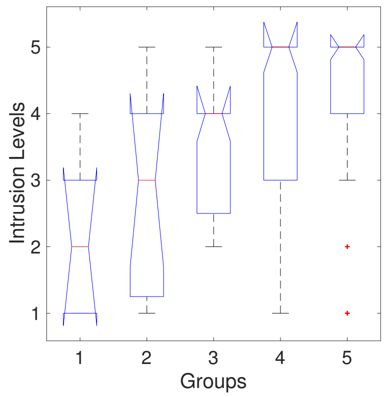
Lastly, for the stakeholder government, the groups (1,4) and (1,5) differ significantly. This goes to show that groups that view the intrusion of stakeholders with a larger difference view Government in a significantly different way. Figure 6.6d shows the spread of the group's opinions on governments.

The trend observed above is that there is a significant difference in the outlook of individual stakeholders in between groups with larger differences in their outlook to stakeholders as a whole, with the exception of Educational Institutions where the alternative hypothesis was rejected because on average, users find it lesser intrusive to a level of 2.94 on a scale of five which is least intrusive compared to the others as shown in 4.

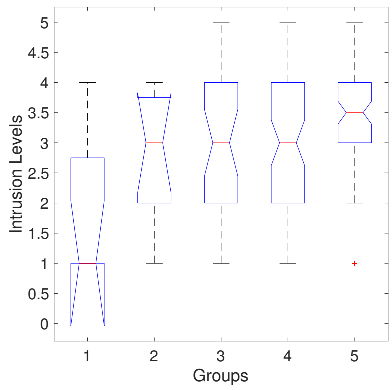
Hence we can conclude that for generally perceived non intrusive stakeholders, there is a similar opinion across various groups. For the rest, there is a significant difference in opinion between groups with larger differences in their outlook to stakeholders as whole. More incentives should be given out to people for generally high intrusive stakeholders and even more so to users from high intrusion groups (such as 4 and 5) to motivate them to give their data to generally perceived intrusive stakeholders. For the lower intrusion stakeholders, there is lesser need of higher rewards to different groups.

Perception of Individual Contexts Grouped on the Intrusion of Contexts in General

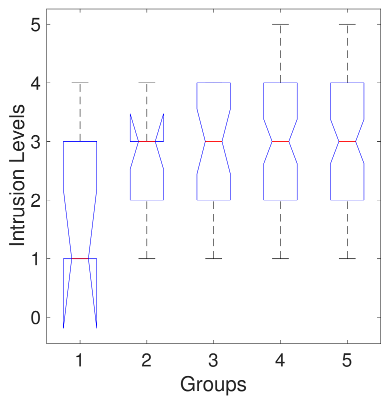
In this section, the relationship between the perception of contexts as a whole and the individual contexts is studied. To do this, like the above



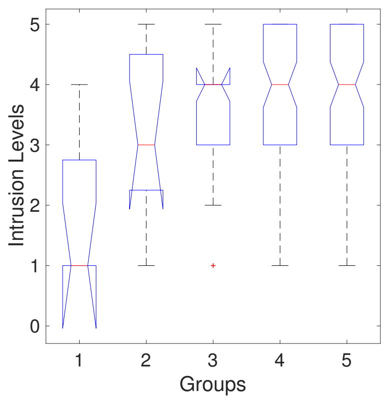
(a) Corporations



(b) NGO



(c) Educational Institutions



(d) Government

Figure 6.6: Box Plots Indicating the Population Spread of each Group

Table 6.4: Dunn's Test for Stakeholders

Groups	Corporation	NGO	Educational Institution	Government
(1,2)	0.9839	0.8042	0.9565	0.6615
(1,3)	0.3282	0.2351	0.8540	0.0986
(1,4)	0.0240	0.1962	0.6867	0.0289
(1,5)	0.0012	0.0243	0.1254	0.0028
(2,3)	0.9467	0.9992	1.0000	0.9958
(2,4)	0.2047	0.9991	1.0000	0.9252
(2,5)	0.0110	0.7192	0.8415	0.3670
(3,4)	0.6893	1.0000	1.0000	0.9999
(3,5)	0.0197	0.8906	0.4112	0.5794
(4,5)	0.4640	0.6027	0.3427	0.7378

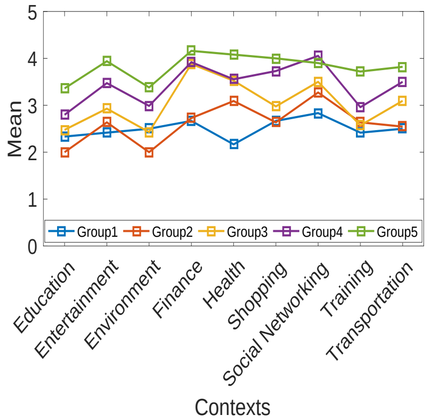
sections the data is partitioned into groups based on the answers given question 14, which asks the user the perception of intrusion of contexts in a data request. There are five groups in total partitioned using the responses given to question 14. Group one to five have each 12, 11, 42, 74 and 50 people respectively. Additional information about the groups on employment, education , gender and year of birth is given in tables B.10, B.13, B.11 and B.12. The mean and variance of each group is shown in figures 6.7a and 6.7b.

Like in the previous sections, since the data is discrete and not normal, we use the Kruskal-Wallis test to compare the groups perceptions on various contexts 6.2. The alpha value is considered to be 0.05. The results of the test are presented in table 6.5. As it is seen, the test says that there is a significant difference between all the groups for all contexts.

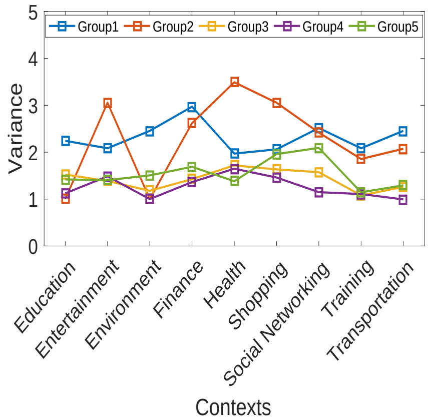
Table 6.5: Kuskal-Wallis Test for Contexts

Context	p-value
Education	6.4694e-04
Entertainment	1.0660e-04
Environment	1.3079e-04
Finance	0.0021
Health	0.0011
Shopping	5.4227e-05
Social Network	0.0120
Training	1.2071e-05
Transportation	4.9043e-04

Dunn's Test is performed as a post hoc test for all the contexts to observe



(a) Mean of each Group for each Context



(b) Variance of each Group for each Context

Figure 6.7: Mean and Variance of Groups

the exact group pairs that might be significantly different. The results are presented in tables 6.6 and 6.7.

For the context education, the groups (2,5) and (3,5) are significantly different from each other. The figure 6.8a shows the spread of the groups. It is seen that groups with larger difference in opinion about contexts have larger differences in opinion about the context education. The reason why the group 1 was not significantly different from group 5 can be attributed to the fact that the population of the survey might not be a representative sample. Secondly, the number of participants in the survey is limited to 189. Thirdly, due to the previous statement the number of people in group 1 is just 12 and the population sample can be unrepresentative of the actual distribution.

For the context entertainment, further investigation shows that groups (1,5), (2,5) and (3,5) are significantly different in each others responses. The figure 6.8b shows the spread of the groups. Therefore, we infer that groups with higher differences in opinions about contexts (such as groups 1 and 5) have varying opinions of intrusion on entertainment.

For the context environment, groups (2,5) and (3,5) are significantly different from each other. The figure 6.8c shows the spread of the groups. We can come to a similar conclusion as to the context education as to why group 1 is not significantly different from group 5.

For the context finance, the groups (1,5), and (2,5) are significantly different from each other. The figure 6.8d shows the spread of the groups and we can infer that groups with bigger differences in opinions on contexts also have significant differences in their opinions on finance context.

In the context health, except for groups (1,5) all the other groups are not significantly different from each other. The figure 6.8e shows the spread of the groups and we can infer people who find contexts non intrusive and very intrusive have significant difference in viewpoint of the health context. As seen, all the other groups have significant overlap with groups 1 and 5.

For the context shopping and social network, none of the groups are significantly different from each other as seen in 6.8f. This show that irrespective of their viewpoint about contexts, all groups have similar opinions with the context shopping and social network.

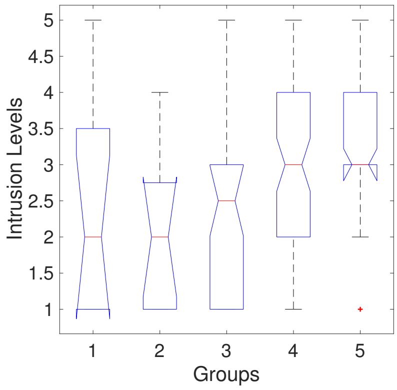
In the context training, the groups (1,5),(3,5) and (4,5) are significantly different from each other. The figure 6.9b shows the spread of the groups. This show that all groups have similar opinions about training contexts except for group 5. The reason for group 2 not being different from group 5 is due to the small population size of group 2. Additional points are mentioned for the context education.

Finally for the context transportation, the groups (1,5), (2,5) and (3,5) are significantly different from each other. It is seen in figure 6.9c that groups 1,2,3 are significantly different form group 5. Group 4 is not significantly different from group 5 due to the fact that the users of group 4 and 5 have closer opinions of contexts than the rest. Hence users who find contexts intrusive will have a significant difference in opinion with sufficiently different groups.

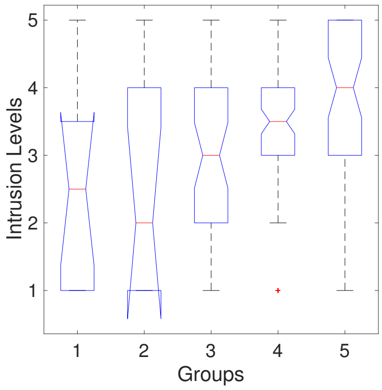
This goes to show that groups that perceive contexts in a more different light tend to have different ways of viewing the individual contexts. We do observe that in some cases (2,5) are significantly different, but (1,5) is not. This can be attributed to the low number of responses and the noise in the data.

Shopping and social networking are considered highly intrusive with values 3.75 and 3.50 and every group had similar opinions. Finance and health are also very intrusive with levels 3.85 and 3.60 on a scale of five. Here a significant difference in opinions was observed between groups with largely different opinions about contexts in general (such as group 1 and 5).

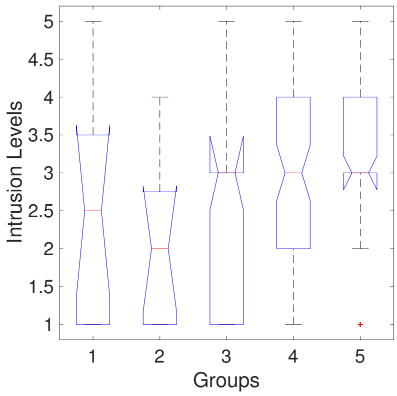
Education and environment are considered 2.83 and 2.92 intrusive on a scale of five, which are the lowest. For these two we observe that there is a significant difference between groups 2 with 5 and 3 with 5. For entertainment and transportation which have medium intrusion levels of 3.39 and 3.38 on a scale of five and it is observed that groups 1, 2, 3 differ significantly from group 5. It is still observed that groups with higher intrusion perception of contexts (such as group 5) view some contexts in a more intrusive light and should be incentivized more for those.



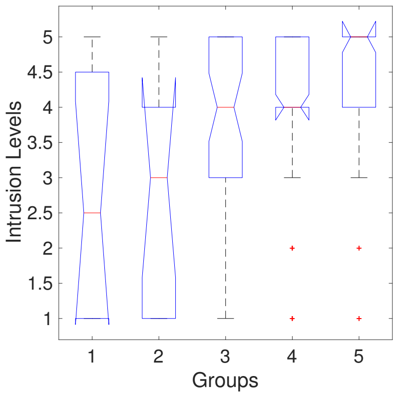
(a) Education



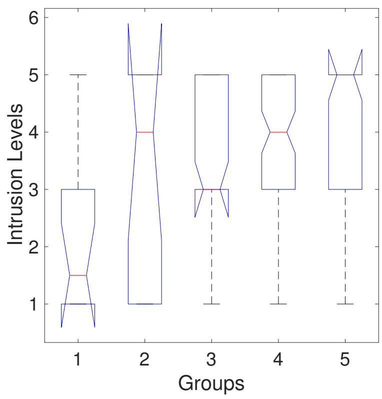
(b) Entertainment



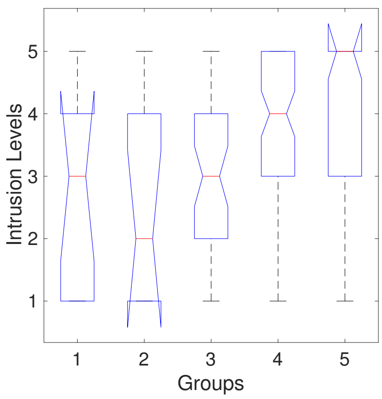
(c) Environment



(d) Finance

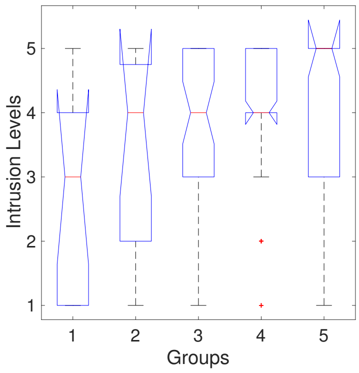


(e) Health

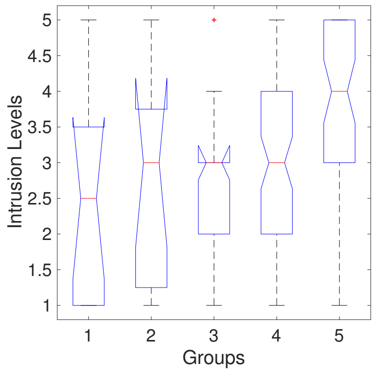


(f) Shopping

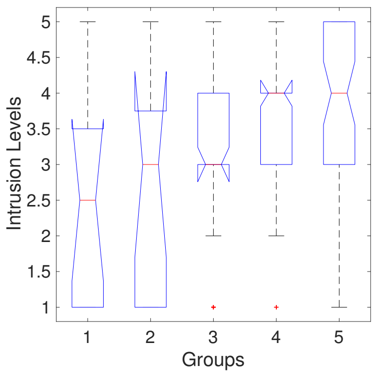
Figure 6.8: Mean and Variance of Groups



(a) Social Networking



(b) Training



(c) Navigation

Figure 6.9: Mean and Variance of Groups

Table 6.6: Dunn's Test for Contexts Part 1

Groups	Education	Entertainment	Environment	Finance	Health
(1,2)	0.99796	0.99982	0.97055	1.0000	0.63908
(1,3)	1.0000	0.99174	1	0.32963	0.076147
(1,4)	0.94148	0.19262	0.86998	0.22077	0.042547
(1,5)	0.11471	0.0056283	0.21553	0.015364	0.00051115
(2,3)	0.9342	1	0.96665	0.31604	0.9999
(2,4)	0.32744	0.76121	0.083389	0.21508	0.99963
(2,5)	0.0082212	0.082023	0.0048936	0.016365	0.50246
(3,4)	0.83617	0.23034	0.10875	1.0000	1.0000
(3,5)	0.0064191	0.00086418	0.0012905	0.65747	0.32713
(4,5)	0.13825	0.28231	0.60016	0.57519	0.21258

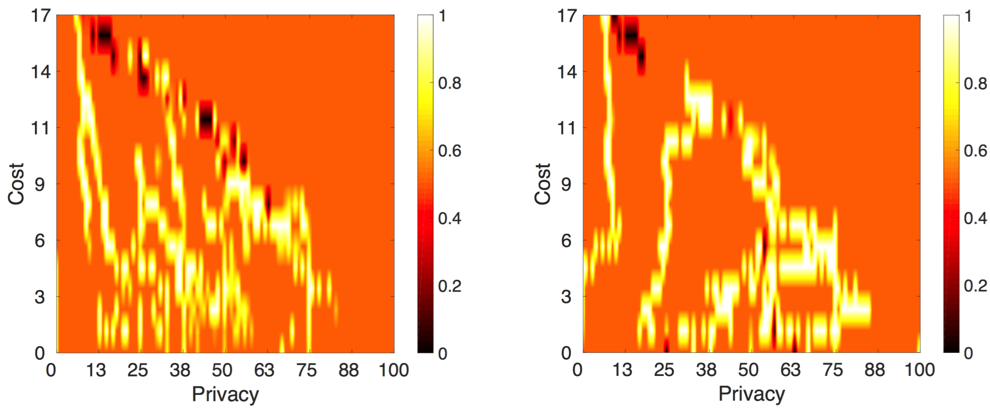
Table 6.7: Dunn's Test for Contexts Part 2

Groups	Shopping	Social Network	Training	Transportation
(1,2)	1.0000	0.99831	1.0000	1.0000
(1,3)	0.99992	0.94767	1.0000	0.9722
(1,4)	0.18309	0.089135	0.90197	0.23809
(1,5)	0.015842	0.10636	0.010906	0.016376
(2,3)	1.0000	1.0000	1	0.98253
(2,4)	0.39426	0.70552	0.99778	0.30509
(2,5)	0.052334	0.7272	0.068368	0.026144
(3,4)	0.038351	0.21259	0.58123	0.51397
(3,5)	0.00050454	0.29374	2.1697e-05	0.012924
(4,5)	0.69535	1.0000	0.0033189	0.55629

In the above, the various intrusions and their relationships with their sub-features was examined. If there was a constant relationship between both, categorization of sub-features in the computational model would be redundant. The above shows that categorizations of sub-features is still essential to the computational model since the data above is not sufficient to make claims about the relationships. In the future, deploying this survey with a larger more representative population can help reduce the number of categorization questions in the model and perhaps even eliminate the categorizations all together. It could also aid to construct better incentivizing mechanisms.

6.3 Findings from the Experiment Data

The social experiment described in chapter 4 took place using 9 participants. Out of the total 3 days of the experiment, 3 of the participants did not answer data requests on day 2 and 3. The mobile application ran successfully on all participants phones even after being switched off. All data was successfully recorded on the server. Participants were not paid but were asked to behave like they were paid, so results might skew from the ideal scenario. A trial was run to see if the application created works and to get user feedback on the usability. Additionally, using the data collected on the server, the relationship between data sharing and incentives is examined.



(a) Actual improvements in Credit and Privacy (b) Improve Privacy and Credit Button

Figure 6.10: Table Schemas

Figure 6.10b depicts for every possible cost and privacy the probability of whether users clicked on the privacy button or credit button. The darker color indicates high probability for clicking on the privacy button (0), a light yellow indicates a higher probability for for clicking on the credit button (1) and the orange depicts the probability of clicking on both buttons equally (0.5). It is observed that probability of users to click on the improve privacy button is very close to when they have no privacy at all and a very high cost. This shows that users are more interested to improve their cost than the privacy metric.

Similarly, figure 6.10a depicts for every possible cost and privacy the probability of whether users actually improved their privacy or their credit rather than just wanting to. The darker color indicates high probability for improvement in privacy (0), a light yellow indicates a higher probability for improvement in credit (1) and the orange depicts the probability of an equal improvement in both metrics (0.5). It is observed that the probability of users incrementing their credit is much higher than their privacy just like in the last figure. Users are interested in improving their privacy when they have sufficient cost and their privacy is becoming lower.

Figures 6.10b and 6.10a are now compared. As observed in 6.10b, users click on the privacy button close to when they have almost no privacy but it is also observed that they improve their privacy even after clicking on the improve credit button as seen in 6.10a. This can be attributed to the fact that users were not comfortable giving data for the data requests that appeared, even tough the intentions were to obtain more credit. This means that they were not incentivized enough for those data requests.

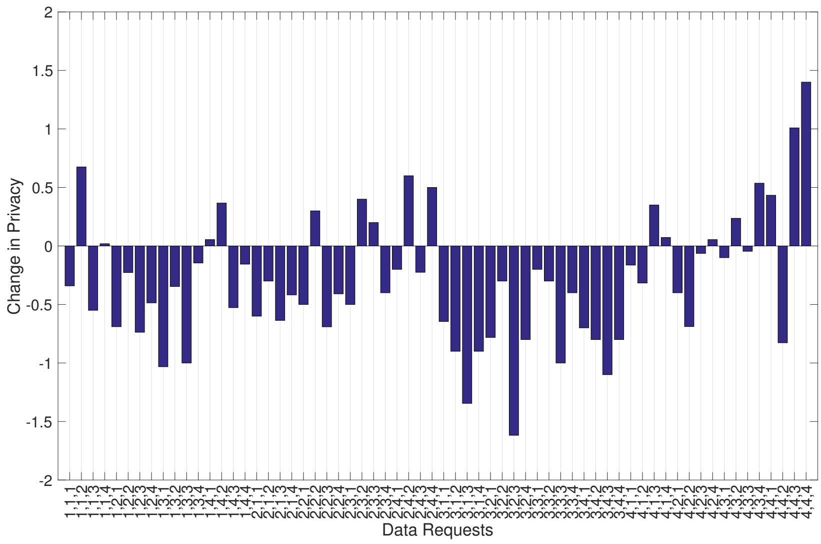


Figure 6.11: Gain in Privacy Between Day 2 and Day 1

Figure 6.11 depicts the gain in privacy for day 2 compared to day 1. In day 2, incentives were given for data requests whereas in day 1 no incentives were given. The x-aixs depicts each data request with a sensor identifier, a stakeholder identifier and a context identifier. For detailed explanation as to which feature belongs to which unique identifier please refer to the section 4.1.4. Since on an average, the contexts were categorized as 3.26 and more intrusive than the rest, they play a bigger role in model in the incentive assignments than sensors and stakeholders which are categorized as low intrusion sub-features.

For the accelerometer, it is observed that for corporations for environment and health contexts users did not share more data than day 1. This can be due to the fact that they were categorized on average as 2.68 and 2.52

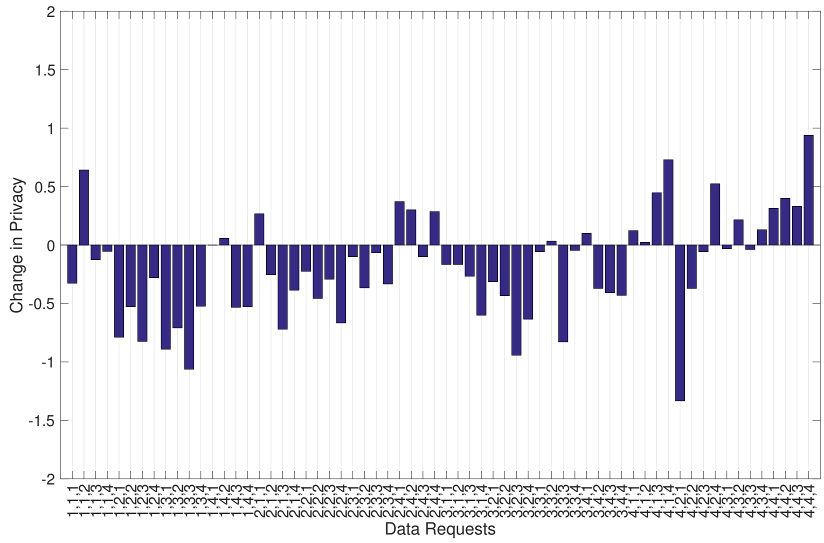


Figure 6.12: Gain in Privacy Between Day 3 and Day 1

respectively, hence these requests were not assigned enough incentives to share more data. Educational institutions were rated second most intrusive with value of 3.29, the contexts social networking and environment were assigned intrusions of 2.97 and 2.68 respectively. For social networking the data request associated with it could have been assigned insufficient incentives whereas environment context was assigned a lower categorization level could have had very low incentives compared to the rest. Furthermore, since the accelerometer was assigned a category of 2.27 which is low in intrusion, the data requests were assigned lower

For the light sensor, it is observed that users gave more data for corporations due to its high categorization compared to the others of 3.37, it was assigned much higher cost and users were incentivized. For the context environment we see that users did not share their data which is categorized as 2.68. But for the health context categorized as 2.52 they shared more data. This can be attributed to the fact that they find environment more intrusive than the health context and users were not incentivized enough. For the noise sensor, which rated categorized the most intrusive of all as 3.65, users shared more data than on day 1. This can be due to the fact that it was assigned much higher incentives than the other sensors.

The location sensor is assigned a category of 3.42 which is the second most intrusive sensor there are a lot of request for which more data was not shared. This can be due to insufficient incentives for these requests. Health was categorized as 2.52 as the least intrusive context and users were not incentivized enough to share location data for lower incentives. For the stakeholder education users have shared less data except for the context environment. The reasons can be that users did not understand how this stakeholder would use their location data for those contexts and needed more explanation, or due to the small sample size of 6 people one person has mistakenly given a very skewed response on the first day.

As observed, it is seen that for most data requests the privacy is negative which means users chose lower summarization levels to gain more credit. Similarly, figure 6.12 depicts the gain in privacy for day 3 compared to day 1. In day 3, incentives were given for data requests whereas in day 1 no incentives were given. Here as well it is seen that users have chosen to improve their credit compared to day 1.

In day 3 6.12, more of the users are responding to incentives as it is seen that there are lesser bars on the positive side of privacy compared to day 2. This can be because users were getting used to the application interface on day 2 and were more explorative to see changes in privacy and credit metrics. Additionally, it is also observed that the bars in the negative side in day 3 are smaller than day 2. This can be because users may be expecting even higher rewards for day 3 of the experiment than day 2, but still wanted to improve their credit.

For the accelerometer, it is seen that the environment context did not provide sufficient incentives for users to share to educational institutions and corporations which were rated high intrusion stakeholders. For the light sensor, the context social networking was categorized as intrusive but still did not incentivize users to share more data than day 1. For the stakeholder educational institution, except for the context transportation users shared less data than on the first day. In the light sensor, for the corporation and educational institutions users did not share more data for the social networking context.

For the sensor noise, users shared more data than day one for all stakeholders and contexts one except for the educational institution stakeholder for the purpose of social networking. For the location sensor the data requests containing the health context more data was not shared. This can be due to the fact that health context is assigned a lower incentive due to its lower categorization of 2.52. The very intrusive transportation context assigned a 3.65 categorization of intrusion, users did not share more data to the intrusive stakeholders corporation and educational institutions. For the context environment, since it is categorized as lowly intrusive with 2.68, enough incentives were not given and did not motivate users to share more of their sensor data. Lastly, for stakeholder educational institution and corporations which considered intrusive intrusive, users did not share more data for the context social networking because they found the sensor, context and stakeholder combination intrusive.

To summarize, we found that users did give away more data for day 2 and day 3 than on day 1. On day 2, it is seen that some data requests for which features are considered non-intrusive during categorization are being given lesser with incentives than without. This could be due to the fact that users could have been explorative during the first day, with the metrics and the application itself non-intrusive were. On day 3, it is observed that responses to data requests are more consistent with the categorization of features. Small inconsistencies of the responses with the categorizations can be attributed to the fact that the population of the experiment was 6 people for day 2 and day 3, hence even if one person answers dramatically differently it affects the overall mean.

Additionally, it was seen that after stakeholder corporations, educational institutions were the data requests where people did not change their minds about. It can be attributed to the fact that most users were affiliated with the department and did not want the department to see their data.

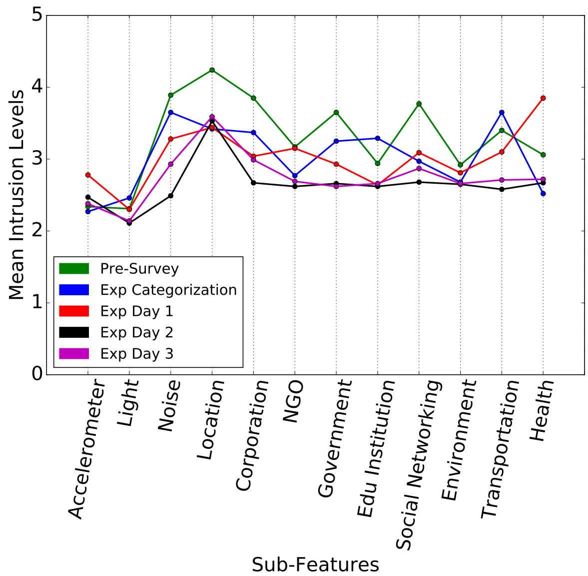


Figure 6.13: Mean Intrusion Levels for Sub-Features

Chapter 7

Conclusion and Future Work

In this dissertation, a survey was deployed to understand the perception of users on mobile sensors, stakeholders and contexts. This information was analysed and used in fine tuning the parameters of the mobile application. The mobile application developed has a computational model inbuilt that automatically assigns incentives to data requests according to the user profiles. A data request consists various features such as the sensor type to collect, stakeholder who is collecting the data and the purpose of data collection. The model assigns higher incentives to data requests which are considered intrusive and lesser incentives to data requests that are found less intrusive. The application collects all user input data plus mobile sensor data shared with the summarization chosen by the user. This data is sent to the cloud server and is accessed by both the user and the stakeholder in a transparent manner. Interrelationships within features are examined and the data obtained from the mobile application is analysed. Results to reinforce the need for the computational model are obtained. Additionally, an increased amount of data sharing is obtained after users are compensated with incentives.

In future, the social experiment can be held with a larger more representative pool of people who are actually awarded the incentives indicated in the mobile application. 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 put up appropriate incentives for each data request.

Appendix A

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
1923	3	"Armenia"	0	0	4	2	0	0
1922	3	"Aruba"	0	0	3	2	0	0
1992	1	"France"	0	0	4	6	0	0
1991	1	"France"	0	0	5	6	0	0
1924	3	"Argentin…	0	0	3	2	0	0
1921	3	"Andorre"	0	0	2	1	0	0
1923	3	"Argentin…	0	0	2	2	0	0
1922	3	"Antigua-…	0	0	2	1	0	0
1922	3	"Anguilla"	0	0	2	2	0	0
1922	3	"Anguilla"	0	0	2	1	0	0
1923	3	"Angola"	0	0	2	2	0	0
1926	2	"Angola"	0	0	4	6	0	0
1924	3	"Andorre"	0	0	3	5	0	0
1924	3	"Angola"	0	0	4	6	0	0
1920	5	"Fiji"	0	0	1	3	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
2	0	0	3	1	"57975541e813f9973...	0	0	0	0	0
2	0	0	2	0	"57975159890927b61...	0	0	0	0	0
2	0	0	3	1	"57935b55a67b0ba32...	1	0	0	0	0
2	1	0	3	1	"579357faa67b0ba32...	1	0	0	0	1
2	1	0	3	0	"579357faa67b0ba32...	0	0	0	0	0
1	0	0	3	0	"57931cad866a46bd5...	1	0	0	0	0
2	1	0	3	0	"57930d55837af5db6...	0	0	0	0	0
1	1	0	3	0	"5792471c493006891...	0	0	0	0	0
2	1	0	3	0	"57923fbfc3d7cee30...	0	0	0	0	0
2	1	0	3	0	"57923946c3d7cee30...	0	0	0	0	0
2	0	0	3	0	"5792373d3692318e3...	1	0	0	0	0
2	0	0	4	1	"57922e1afb5591741...	0	0	0	0	0
2	1	0	3	0	"57922a329bb316492...	0	1	0	0	0
2	0	0	3	1	"579224ffba4636590...	0	0	0	0	0
2	0	0	1	0	"5791d082bb71b5202...	0	0	0	0	0
1	1	0	4	1	"578e91e778f251171...	1	1	0	0	1

Figure A.2: Screenshot of Collection UserInformation Part 2

user_id	context	data_collector	sensor
"57a8f8f1848532cf7...	1	3	5
"579a148f352257bc0...	2	3	3
"57975541e813f9973...	4	5	1
"57975159890927b61...	4	3	1
"57935b55a67b0ba32...	1	3	5
"579357faa67b0ba32...	1	3	5
"57931cad866a46bd5...	3	2	2
"57930d55837af5db6...	4	1	2
"5792471c493006891...	4	2	2
"57923fbfc3d7cee30...	4	1	2
"57923946c3d7cee30...	3	1	2
"5792373d3692318e3...	4	1	2
"57922e1afb5591741...	4	2	2
"57922a329bb316492...	3	2	4
"579224ffba4636590...	4	2	2
"578e91e778f251171...	3	5	4
"578e28687d1cdd1b6...	4	3	2

Figure A.3: Screenshot of Collection Features

user_id	acc	gps	light	noise
"57a8f8f1848532cf76b0836f"	3	5	2	4
"57a8f8f1848532cf76b0836f"	3	5	2	4
"579a148f352257bc0612c70b"	2	2	4	3
"57975541e813f99735dd0598"	2	4	3	3
"57975159890927b613d0f49f"	1	3	5	3
"57935b55a67b0ba32f81eeac"	3	5	1	5
"579357faa67b0ba32f81e724"	3	5	1	5
"57931cad866a46bd554a1896"	2	2	3	4
"57930d55837af5db6734a438"	2	1	2	4
"5792471c493006891e4e0c2b"	2	2	3	4
"57923fbfc3d7cee306b50957"	2	3	3	4
"57923946c3d7cee306b4fb44"	2	4	2	4
"5792373d3692318e3ba9e929"	2	1	2	4
"57922e1afb55917415770d44"	2	3	4	4
"57922a329bb316492f70242b"	2	1	2	4
"579224ffba46365901e66e78"	2	2	2	4
"578e91e778f2511711cfb9f5"	1	5	1	3

Figure A.4: Screenshot of Collection Sensors

user_id	corp	edu	gov	ngo
"57a8f8f1848532cf76b0836f"	3	1	4	3
"579a148f352257bc0612c70b"	3	3	4	2
"57975541e813f99735dd0598"	1	4	1	3
"57975159890927b613d0f49f"	2	4	4	5
"57935b55a67b0ba32f81eeac"	5	3	5	5
"579357faa67b0ba32f81e724"	5	3	5	5
"57931cad866a46bd554a1896"	2	4	4	1
"57930d55837af5db6734a438"	3	3	1	2
"5792471c493006891e4e0c2b"	2	4	1	1
"57923fbfc3d7cee306b50957"	3	4	3	2
"57923946c3d7cee306b4fb44"	3	3	3	3
"5792373d3692318e3ba9e929"	3	3	2	2
"5792373d3692318e3ba9e929"	3	3	2	2
"57922e1afb55917415770d44"	4	3	2	1
"57922a329bb316492f70242b"	2	4	2	1
"579224ffba46365901e66e78"	2	4	2	1
"578e91e778f2511711cfb9f5"	5	2	5	4

Figure A.5: Screenshot of Collection Stakeholders

user_id	environment	health	social_networking	transportation
"57a8f8f1848532cf76b0836f"	3	3	5	3
"579a148f352257bc0612c70b"	4	2	2	2
"57975541e813f99735dd0598"	2	2	4	2
"57975159890927b613d0f49f"	3	1	5	5
"57975159890927b613d0f49f"	3	1	5	5
"57975159890927b613d0f49f"	3	1	5	5
"57935b55a67b0ba32f81eeac"	1	1	1	1
"579357faa67b0ba32f81e724"	3	1	5	3
"57931cad866a46bd554a1896"	3	3	3	5
"57930d55837af5db6734a438"	3	2	1	3
"5792471c493006891e4e0c2b"	1	2	1	3
"57923fbfc3d7cee306b50957"	3	3	2	5
"57923946c3d7cee306b4fb44"	3	2	3	3
"5792373d3692318e3ba9e929"	2	3	1	3
"5792373d3692318e3ba9e929"	2	3	1	3
"5792373d3692318e3ba9e929"	2	3	1	3
"5792373d3692318e3ba9e929"	3	3	2	4

Figure A.6: Screenshot of Collection Contexts

contexts	credit	credit_can_be	credit_gain	credit_question	data_collectors	timestamp	day_no
0	6.510009765625	0	-0.076904296874999…	0.17944335937499994	0	"2016-07-25 11:51:57.234"	3
1	6.436767578125	0.30029296874999994	0	0.35034179687499994	2	"2016-07-25 10:19:00.938"	3
0	6.436767578125	0.30029296874999994	0	0.35034179687499994	2	"2016-07-24 14:53:15.508"	3
2	6.436767578125	0.30029296874999994	0	0.35034179687499994	1	"2016-07-24 14:53:04.694"	3
3	6.436767578125	0.30029296874999994	0	0.35034179687499994	1	"2016-07-24 14:53:13.376"	3
1	6.436767578125	0.30029296874999994	0	0.35034179687499994	1	"2016-07-24 14:52:55.396"	3
0	6.436767578125	0.30029296874999994	0	0.35034179687499994	1	"2016-07-24 14:52:41.651"	3
2	6.436767578125	0.30029296874999994	0	0.35034179687499994	0	"2016-07-24 14:52:37.05"	3
3	6.436767578125	0.30029296874999994	0	0.35034179687499994	0	"2016-07-24 14:52:39.684"	3
1	6.25732421875	0.17944335937499994	0.17944335937499994	0.17944335937499994	2	"2016-07-24 14:23:08.071"	3
3	6.436767578125	0.17944335937499994	0.17944335937499994	0.17944335937499994	2	"2016-07-24 14:23:09.255"	3
1	6.436767578125	0.30029296874999994	0	0.35034179687499994	0	"2016-07-24 14:52:34.802"	3
0	6.436767578125	0.30029296874999994	0	0.35034179687499994	0	"2016-07-24 14:52:20.038"	3
0	6.077880859375	0.17944335937499994	0.17944335937499994	0.17944335937499994	2	"2016-07-24 14:23:07.005"	3
1	5.53955078125	0.17944335937499994	0.17944335937499994	0.17944335937499994	1	"2016-07-24 14:23:02.056"	3
3	5.8984375	0.17944335937499994	0.17944335937499994	0.17944335937499994	1	"2016-07-24 14:23:04.974"	3
2	5.718994140625	0.17944335937499994	0.17944335937499994	0.17944335937499994	1	"2016-07-24 14:23:03.856"	3

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
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0.5504352278545781	-1.036866359…	1	64.28571428571429	1
"57935b55a67b0ba32f81eeac"	2	0.5580357142857082	-1.004464285…	1	63.28125	1
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0	0	5	63.28125	2
"57935b55a67b0ba32f81eeac"	2	0.542041248016929	-1.070861977…	1	65.3225806451613	1
"57935b55a67b0ba32f81eeac"	2	0.5113968439509051	-1.183518410…	1	68.64406779661017	1
"57935b55a67b0ba32f81eeac"	2	0.5327868852458977	-1.106557377…	1	66.39344262295081	1
"57935b55a67b0ba32f81eeac"	2	0.5225988700564983	-1.144067796…	1	67.5	1

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
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186408196
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186438200
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186378192
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186288183
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186348189
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186258180
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186228177
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186318186
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186198171
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186168166
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186138163
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186108160
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186078154
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186018145
3	"578e91e778f2511711cfb9f5"	47.419864654541016	8.502890586853027	1	1469186048150

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
3	1	1469186373308	"578e91e778f251171…	0.0191536135971546…	-0.124498486518859…	10.180145263671875
3	1	1469186433709	"578e91e778f251171…	0.0287304203957319…	-0.134075298905372…	10.15141487121582
3	1	1469186343108	"578e91e778f251171…	0.0287304203957319…	-0.134075298905372…	10.15141487121582
3	1	1469186403508	"578e91e778f251171…	0.0191536135971546…	-0.134075298905372…	10.15141487121582
3	1	1469186282709	"578e91e778f251171…	0.0191536135971546…	-0.143652096390724…	10.15141487121582
3	1	1469186222308	"578e91e778f251171…	0.0287304203957319…	-0.105344876646995…	10.20887565612793
3	1	1469186252509	"578e91e778f251171…	0.0287304203957319…	-0.143652096390724…	10.160991668701172
3	1	1469186312909	"578e91e778f251171…	0.0287304203957319…	-0.143652096390724…	10.15141487121582
3	1	1469186131909	"578e91e778f251171…	0.0191536135971546…	-0.143652096390724…	10.160991668701172
3	1	1469186192307	"578e91e778f251171…	0.0287304203957319…	-0.134075298905372…	10.160991668701172
3	1	1469186162108	"578e91e778f251171…	0.0191536135971546…	-0.143652096390724…	10.160991668701172
3	1	1469186101709	"578e91e778f251171…	0.0287304203957319…	-0.143652096390724…	10.15141487121582
3	1	1469186011488	"578e91e778f251171…	0.0191536135971546…	-0.143652096390724…	10.132261276245117
3	1	1469185981478	"578e91e778f251171…	0.0287304203957319…	-0.143652096390724…	10.160991668701172
3	1	1469186071708	"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
"0.0,1.952634E-5,5…	"578e91e778f2511711cfb9f5"	3	100.1298828125	62.05247497558594	1	1469186438208
"0.0,2.1573624E-5,…	"578e91e778f2511711cfb9f5"	3	47.744140625	55.61960220336914	1	1469186408210
"0.0,2.0647063E-5,…	"578e91e778f2511711cfb9f5"	3	73.727294921875	59.39376449584961	1	1469186378210
"0.0,2.1016425E-5,…	"578e91e778f2511711cfb9f5"	3	71.015380859375	59.0682487487793	1	1469186348202
"0.0,2.062715E-5,6…	"578e91e778f2511711cfb9f5"	3	96.7724609375	61.7562370300293	1	1469186258193
"0.0,2.1374477E-5,…	"578e91e778f2511711cfb9f5"	3	67.723876953125	58.656036376953125	1	1469186168186
"0.0,1.7658736E-5,…	"578e91e778f2511711cfb9f5"	3	106.538818359375	62.59135818481445	1	1469186198197
"0.0,1.8755187E-5,…	"578e91e778f2511711cfb9f5"	3	40.89111328125	54.27377700805664	1	1469186318192
"0.0,1.429928E-5,4…	"578e91e778f2511711cfb9f5"	3	21.7080078125	48.773597717285156	1	1469186078184
"0.0,1.9742341E-5,…	"578e91e778f2511711cfb9f5"	3	85.58349609375	60.68899917602539	1	1469186138189
"0.0,2.134739E-5,5…	"578e91e778f2511711cfb9f5"	3	78.437744140625	59.93170166015625	1	1469186108185
"0.0,2.007436E-5,5…	"578e91e778f2511711cfb9f5"	3	82.403076171875	60.360069274902344	1	1469186048176
"0.0,2.0650641E-5,…	"578e91e778f2511711cfb9f5"	3	76.4638671875	59.710323333740234	1	1469185958181
"0.0,1.4806586E-5,…	"578e91e778f2511711cfb9f5"	3	21.33837890625	48.624427795410156	1	1469186018178
"0.0,2.0179677E-5,…	"578e91e778f2511711cfb9f5"	3	101.021484375	62.12947463989258	1	1469185988177

Figure A.11: Screenshot of Collection Noise

day_no	summarization	timestamp	user_id	x
3	3	1469447239362	"57935b55a67b0ba32…	47
3	3	1469447109437	"57935b55a67b0ba32…	54
3	3	1469447319071	"57935b55a67b0ba32…	39
3	3	1469446323286	"57935b55a67b0ba32…	109
3	3	1469446998112	"57935b55a67b0ba32…	180
3	3	1469446812605	"57935b55a67b0ba32…	165
3	3	1469446228120	"57935b55a67b0ba32…	83
3	3	1469445977205	"57935b55a67b0ba32…	96
3	3	1469445805362	"57935b55a67b0ba32…	156
3	3	1469445621109	"57935b55a67b0ba32…	157
3	3	1469445343373	"57935b55a67b0ba32…	136
3	3	1469445541953	"57935b55a67b0ba32…	143
3	3	1469445255903	"57935b55a67b0ba32…	150
3	3	1469444855996	"57935b55a67b0ba32…	127
3	3	1469444963549	"57935b55a67b0ba32…	127
3	3	1469445171668	"57935b55a67b0ba32…	100

Figure A.12: Screenshot of Collection Light

user_id
"57a8f8f1848532cf76b0836f"
"579a148f352257bc0612c70b"
"57975541e813f99735dd0598"
"57975159890927b613d0f49f"
"57935b55a67b0ba32f81eeac"
"579357faa67b0ba32f81e724"
"57931cad866a46bd554a1896"
"57930d55837af5db6734a438"
"5792471c493006891e4e0c2b"
"57923fbfc3d7cee306b50957"
"57923946c3d7cee306b4fb44"
"5792373d3692318e3ba9e929"
"5792373d3692318e3ba9e929"
"57922e1afb55917415770d44"
"57922a329bb316492f70242b"
"579224ffba46365901e66e78"
"578e91e778f2511711cfb9f5"

Figure A.13: Screenshot of Collection Users

timestamp	user_id	credit	day_no	privacy
"2016-08-08 23:35:20.788"	"57a8f8f1848532cf76b0836f"	6.310096153846153	1	74.609375
"2016-07-28 16:23:21.687"	"579a148f352257bc0612c70b"	9.030898876404493	1	56.25
"2016-07-26 14:22:12.236"	"57975541e813f99735dd0598"	9.01900773195876	1	57.03125
"2016-07-26 14:09:18.367"	"57975159890927b613d0f49f"	8.850940265486726	1	58.203125
"2016-07-26 14:00:00.565"	"57935b55a67b0ba32f81eeac"	0	4	0
"2016-07-25 13:59:00.084"	"57935b55a67b0ba32f81eeac"	6.510009765625	3	63.28125
"2016-07-24 13:59:31.599"	"57935b55a67b0ba32f81eeac"	8.8885498046875	2	50.390625
"2016-07-23 13:59:31.382"	"57935b55a67b0ba32f81eeac"	5.90576171875	1	67.1875
"2016-07-23 09:31:59.106"	"57931cad866a46bd554a1896"	1.2409156976744176	1	55
"2016-07-23 08:36:40.373"	"57930d55837af5db6734a438"	8.348214285714286	1	60.9375
"2016-07-22 18:18:33.16"	"5792471c493006891e4e0c2b"	1.2428977272727268	1	52.5
"2016-07-22 17:48:06.151"	"57923fbfc3d7cee306b50957"	1.253551136363636	1	55
"2016-07-22 17:19:40.016"	"57923946c3d7cee306b4fb44"	9.385190217391308	1	53.125
"2016-07-22 16:32:46.545"	"57922e1afb55917415770d44"	7.999999999999998	1	63.28125
"2016-07-22 16:22:07.321"	"57922a329bb316492f70242b"	8.459821428571429	1	60.9375
"2016-07-22 13:22:00.152"	"578e91e778f2511711cfb9f5"	14.11313657407408	3	18.359375

Figure A.14: Screenshot of Collection Score

Appendix B

Table B.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%

Table B.2: Employment Classification of Groups For Sensors

Occupation	1	2	3	4	5
Employed full time	4.90%	6.86%	26.47%	38.24%	23.53%
Employed part time	8.33%	16.67%	33.33%	16.67%	25.00%
Unemployed and looking for work	8.33%	16.67%	16.67%	25.00%	33.33%
Unemployed and not looking for work	0.00%	0.00%	0.00%	66.67%	33.33%
Retired	0.00%	0.00%	100.00%	0.00%	0.00%
Student	7.23%	4.82%	26.51%	39.76%	21.69%
Disabled	0.00%	0.00%	0.00%	0.00%	0.00%

Table B.3: Gender Classification of Groups For Sensors

Gender	1	2	3	4	5
Female	5.56%	4.17%	33.33%	30.56%	26.39%
Male	7.14%	9.52%	22.22%	39.68%	21.43%

Table B.4: Average Birth Year of Groups For Sensors

1	2	3	4	5
1989	1979	1986	1986	1983

Table B.5: Education Classification of Groups For Sensors

Education	1	2	3	4	5
Less than high school	20.00%	20.00%	20.00%	40.00%	0.00%
High school	10.53%	0.00%	52.63%	31.58%	5.26%
Some college	10.00%	20.00%	30.00%	20.00%	20.00%
Bachelors degree	7.02%	10.53%	24.56%	42.11%	15.79%
Masters degree	3.80%	5.06%	24.05%	35.44%	31.65%
PhD degree	7.14%	7.14%	17.86%	35.71%	32.14%

Table B.6: Employment Classification of Groups For Stakeholders

Occupation	1	2	3	4	5
Employed full time	3.92%	4.90%	16.67%	33.33%	41.18%
Employed part time	16.67%	16.67%	8.33%	50.00%	8.33%
Unemployed, looking for work	8.33%	8.33%	16.67%	16.67%	50.00%
Unemployed, not looking for work	0.00%	0.00%	0.00%	33.33%	66.67%
Retired	0.00%	0.00%	0.00%	100.00%	0.00%
Student	4.82%	7.23%	15.66%	37.35%	34.94%
Disabled	0.00%	0.00%	0.00%	0.00%	0.00%

Table B.7: Gender Classification of Groups For Stakeholders

Gender	1	2	3	4	5
Female	5.56%	6.94%	23.61%	36.11%	27.78%
Male	3.97%	6.35%	11.90%	35.71%	42.06%

Table B.8: Average Birth Year of Groups For Stakeholders

1	2	3	4	5
1986	1989	1984	1985	1984

Table B.9: Education Classification of Groups For Stakeholders

Education	1	2	3	4	5
Less than high school	20.00%	20.00%	0.00%	60.00%	0.00%
High school	15.79%	5.26%	21.05%	31.58%	26.32%
Some college	0.00%	10.00%	20.00%	40.00%	30.00%
Bachelors degree	1.75%	12.28%	15.79%	35.09%	35.09%
Masters degree	5.06%	1.27%	13.92%	37.97%	41.77%
PhD degree	0.00%	7.14%	21.43%	28.57%	42.86%

Table B.10: Employment Classification of Groups For Contexts

Occupation	1	2	3	4	5
Employed full time	4.90%	3.92%	23.53%	36.27%	31.37%
Employed part time	16.67%	8.33%	25.00%	25.00%	25.00%
Unemployed, looking for work	8.33%	16.67%	25.00%	25.00%	25.00%
Unemployed, not looking for work	0.00%	0.00%	0.00%	66.67%	33.33%
Retired	0.00%	0.00%	0.00%	100.00%	0.00%
Student	7.23%	6.02%	20.48%	44.58%	21.69%
Disabled	0.00%	0.00%	0.00%	0.00%	0.00%

Table B.11: Gender Classification of Groups For Contexts

Gender	1	2	3	4	5
Male	7.14%	6.35%	23.02%	38.10%	25.40%
Female	5.56%	5.56%	19.44%	38.89%	30.56%

Table B.12: Average Birth Year of Groups For Contexts

1	2	3	4	5
1986	1986	1986	1985	1983

Table B.13: Education Classification of Groups For Contexts

Education	1	2	3	4	5
Less than high school	20.00%	0.00%	0.00%	80.00%	0.00%
High school	10.53%	10.53%	31.58%	36.84%	10.53%
Some college 1	0.00%	0.00%	40.00%	30.00%	20.00%
Bachelors degree	7.02%	8.77%	24.56%	35.09%	24.56%
Masters degree	6.33%	3.80%	17.72%	40.51%	31.65%
PhD degree	0.00%	7.14%	17.86%	35.71% 3	9.29%

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):

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

Zurich, 25-08-2016

Signature(s)

Ramapriya

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