ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Title of Thesis

Master Thesis

Ramapriya Sridharan

September 3, 2016

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

Contents

Chapter 1

Details of the Model behind the Personalized Survey

In this Chapter, details behind how the Personalized Survey is performed with the reasoning behind it is presented. Here, the data of interest are the choices that users make , given certain factors with respect to privacy of their data. More than a survey, this is more like a game.

1.1 User Profiling

This part of the survey consists of drawing a profile of the user from some non intrusive personal details to personal choices about the factors that can affect their choice with respect to the privacy of their data. This can help in profiling the users using some of their personal traits. In addition, it is useful to ask personalising questions to the user.

1.1.1 Personal Information

Personal Information refers to basic information about the user such as age, country and so on...This personal information is collected so as to be able to make inferences on a group of users with similar information. Before the survey starts, each user is assigned an unique personal identification. This identification remains the same throughout the survey. This can help link various information that belongs to the user on the database, where the data is stored. Below, the kind of data collected is described. This is followed by the options that the user can choose from for that question. They are :

- What is your gender?
    1. *Male*
    2. *Female*

3. *None*

- What is your age?
  - *The user enters his (or) her age*
- What is your Relationship Status?
  1. *Single*
  2. *Married*
- Education Level
  1. *High School*
  2. *Bachelor Degree*
  3. *Graduate Degree*
- How active are you on Social Media?
  1. *Multiple times a day*
  2. *Once a day*
  3. *Few times a week*
  4. *Lesser than the above*
- What is your educational background?
  1. *Computer Science (or) Electrical Engineering*
  2. *Other*
- Country of residence
  - *The user enters his (or) her country of residence*

1.1.2 Ranking the parameters that affect data sharing

The users are asked to rank which parameters affect the sharing their data the most. That is, the ranking goes from the feature that affects data sharing the most, to the parameter that doesn't affect the data sharing decision much. An example could be that if the user feels that sharing of his data sharing can be affected the most by which sensor the user is sharing, then that feature is placed first. Then, if the user feels that his data sharing next can be affected by the entity that asks for the data, the Buyers of the data is placed second. Lastly, the user places the feature in the last position that affects the least his data sharing decision. The following list consists of the features the user should rank from one to three:

- *Sensors* - This represents the sensors from which the data is obtained. For example data from the Accelerometer sensor.
- *Buyers* - This represents the entity who would like to obtain the user's data. For example the user can give his data to a Non Governmental Organization.
- *Contexts* - This represents the purpose for which the data can be collected. For example, data can be collected for the purpose of entertainment.

1.1.3 Ranking the various Sensors

In this part of the user profiling, users are asked to rank the various sensors available on the mobile phone. Here, the ranking can be done from the sensor's data which the user is most comfortable sharing to the sensor's data which the user is least comfortable sharing. An example could be that if the user feels that data from Accelerometer sensor is least sensitive to him, it can be placed first. Again, if the user feels that the data from the Battery sensor is more sensitive than that of the Accelerometer sensor, yet less sensitive than the others it can be placed second. In a similar way, this can be done for all the other sensors. The sensor ranked last in the list is the sensor's data that the user feels is most sensitive of all.
Below are the sensors available in the phone that are included in the survey, followed by a short description:

- *Accelerometer* - An Accelerometer is a sensor that can detect the estimated acceleration in three directions which are the x, y and z axis. From this, the velocity and displacement of the device can be found as well [8]. This sensor can be used for a number of scenarios, such as detection of transportation [13] and the detection of driving behaviour [11].
- Gyroscope - A Gyroscope is a sensor that can detect orientation and rotation in three directions which are the x, y and z axis. This combination of the Gyroscope's rotation measurement and the Accelerometers linear movement measurement helps achieve more accurate movement recognition within the 3D space than the Accelerometer by itself. Hence it can usually be used along with the Accelerometer sensor.
- *Magnetometer* - A Magnetometer is a sensor that is used to measure the strength and the direction of the magnetic fields. The coordinate system used by the magnetometer is the coordinate system of the earth. This can be used to synchronize various devices in the same direction [10].

- *Battery* - The Battery sensor is one that measures level of the battery, *i.e.* the amount of capacity left in the mobile phone battery. It can also provide extra information such as the temperature of the battery, how the battery is being charged (*i.e. such as by the USB, wirelessly or using an AC charger*), the voltage level, the technology of the battery and the current health of the battery [1].

- *Light* - The Light sensor is the one that measures the brightness of the ambient light. The reason for its presence by default in phones is for automatic adjustment of the mobile phone's display screen. Some light sensors can even detect brightness of individual colors [3].

- *Location* - The Location sensor is the one that gives a person's exact location on a map. The coordinates consist of the longitude and the latitude values. The location of a person can be of 3 types:

  - Wifi Localization - The MAC address (*Media Access Control Address, generally called the physical address*) and the SSID (*Service Set Identifier, it is a unique identifier in a particular area*) are used to locate the Wifi hotspot. Once the hotspot is located, using the intensity of the signal received by the device, the distance from the Wifi hotspot is estimated [7].

  - Triangulation Localization - In this method, three cell phone towers are used to triangulate a mobile device. The accuracy depends on the concentration of the cell phone towers, so usually urban areas would achieve highest accuracy [9].

  - GPS (*Global Positioning System*) - This is a navigation system that is space based that gives the location of a device anywhere on the earth. To get an accurate location, at least four satellites are needed. If the altitude of the device is not needed, three satellites can be enough. Space ahs a total of 28 satellites [6]. The mobile has a receiver that receives time coded information, the location of the satellites and some more bits of information [5].

- *Proximity* - The Proximity Sensor is one that is comprised of an LED and IR light detector. It can be placed near the ear piece of the phone . The reason can be that when a person is on a phone call, this sensor can detect that the ear is against the phone and it can switch off the display screen [3]. This can help in not pressing any buttons while the user is on a call.

Two or more sensors cannot have the same rank.

1.1.4 Ranking the various Buyers

In the next part of the user profiling, the users are asked to rank the various buyers who can buy their data. The ranking is done from the buyer to which the user is most comfortable sharing the data to, to the buyer to whom the user is least comfortable sharing the data. An example being if the user feels most comfortable to share his data with a Charitable Organization compared to all the other options available, Charitable Organization is ranked first. In addition, if the user is less comfortable giving the data to another user compared to giving it to a Charitable Organization, but is more comfortable doing so compared to all the other entities, another user is ranked second.In the same way, all the other entities are ranked.The last entity is the buyer to which the user is most uncomfortable sharing the data to.
Below, buyers available in the survey are listed:

- *Employer* - An Employer is the person or a business that hires people mainly for wages or salary [4].
- *Corporation* - A Corporation is an company that is created by a set of individuals having some association. This can be done by the law or under the authority of the law. It lives on independent of the existence of its members. In addition, its power and liabilities are distinct from those of its members [2].
- *Educational Institution* - An Educational Institution is one that is dedicated to education. It is generally meant to educate for life, perform specific research and grant degrees.
- *Non Governmental Organization (NGO) (or) Charitable Organization* - A Non Governmental Organization is one that performs activities that might include human rights, environmental, improving health, or development work. An NGO's level of operation indicates the scale at which an organization works, such as local, regional, national, or international [12].
- *A friend* - A friend represents a friend in your circle whom you know.
- *A person in your Family* - A person in you Family can be anybody from your grand-parents to you nephew.
- *Another user (or) An individual Person* - This can be anybody from you friend, family member to a random person in the world.

Two or more buyers cannot have the same rank.

1.1.5 Categorize the various Contexts

Next part in the user profiling is the categorizing of the contexts. Instead of asking the user to rank the available contexts, users are asked to place them

in the available buckets,*i.e. users should place given contexts in pre existing slots*. Each of the slots given have a pre assigned rank to them. So all the contexts assigned to that slot is assigned the same rank. Each of the slots can be called categories. Therefore, each of the contexts need to be categorized.

- *Educational* - Data collected from the user, can be used for educational purposes. An example, the data can be used to make a mobile application that is used for educational purposes or, to be used for a lecture at a school.
- *Health* - Data collected from the user can be used for purposes of health and medicine. An example can be the data can used for creating an health application such as running, or an application for monitoring patients.
- *Navigation (or) Localization* - Data collected from the user can be used for Navigation or Localization. An example can be for constructing a map application for tourists.
- *Entertainment* - Data collected from the user, can be used for Entertainment purposes. An example can be for making an application for gaming.
- *Shopping* - Data collected from the user, can be used for for the purpose of shopping. An example could be an application that suggests things to buy using collected behaviour.
- *Social Media* - Data collected from the user, can be used for the purpose of social media. An example could be to make a new application that helps share your thoughts to people within a certain radius.

Above the various available contexts were described. Next, the user can place the various contexts in the pre-defined categories. The categories are:

- Very Useful *i.e.. Contexts users find useful, or for which they would be most likely to give their data*
- Useful *i.e.. Contexts which users might or might not give their data*
- Not Useful *(or)* Charitable Organization *i.e.. Contexts for which the user most likely will not give their data*

There can be more than one context assigned to one category.

1.2 Characterization of the model

1.2.1 Forming the questions

The questions for the survey are formed using the all the different and unique combinations of sensors, buyers and contexts. The template of the

questions for the survey is:
" Buyer B would like to obtain your S sensor data for the purpose of C "
where:

- *"B"* - stands for the Buyer who wants to buy the user's data
- *"S"* - stands for the Sensor whose data the Buyer is asking for
- *"C"* - stands for the context, which is the purpose for the data collection

Let n_s be the number of sensors, n_b can be the number of buyers and n_c can be the number of contexts. From this we can obtain that there can be a maximum of:

$$NQ = n_s * n_b * n_c$$

unique questions.

1.2.2 Formulating the Weight Matrix

Now that the user profiling has been done, the weight matrix can be formulated. Let wq be the representation of the weight matrix. The experiment has a budget TB, each time period of duration T can have a budget B. Let NT be the number of such time periods. Hence we infer:

$$TB = B * NT$$

The weight matrix wq is the one that holds the amount of money each questions gets with respect to the other. The weight matrix is designed such that:

$$\sum_{j,k,l}^{n_s,n_b,n_c} wq_{j,k,l} < NQ$$

The above equation represents that each element of the matrix wq cannot be more than one. The reason being weights can be one only for questions that get maximum share of the budget and in this model not all questions can get a maximum share. Hence the inequality.
In this personalized survey, it is attempted to give each question a different weight so that questions that the user is not comfortable giving data for can earn him more money. Doing so helps us see if this can attract users to give data to questions they would normally not so readily give.
The method to assign money to each question will be discussed in the later sections. For the following sections, the focus will be on the formulation of the weight matrix.

1.2.3 Calculating Weights for Parameters

In the above section, the users were asked to rank the features Sensors and Buyers and Contexts themselves before ranking the options of the features.

For the features, a higher rank *i.e. one*, means that this feature can carry a lot of importance when it comes to the data sharing decision. Similarly, a lower rank can signify that this feature contributes the least compared to the other available features to the data sharing decision. Ranking of the parameter can be used to give a certain weighting to each feature in the model. Instead of having a fixed a model, the model's parameters are themselves variable. To begin, each rank assigned to the features needs to be converted into a weight parameter in the model. Let f be the number of features available. In this case $f = 3$, because there are three features involved. Then, it is tried to obtain the weights corresponding to each rank. To do that the following equation can be solved:

$$\sum_{i=1}^{p} i * x = 1$$

After finding the variable x (*NOTE :x is just a variable that can help solve the equation, it has no special meaning in this context*), the weight of the individual features w_{f_i} can be found using :

$$w_{f_i} = x * (f - r_{f_i} + 1)$$

Where p_i can represent the individual features, and r_{p_i} can represent the rank assigned to each of the features. The equation intuitively says that as the rank is lower, that feature contributes lesser to the data sharing decision. Hence, such a feature should weighted less than another ranked higher. In addition, from the above:

$$\sum_{i=1}^{f} w_{f_i} = 1$$

can be obtained *i.e. the weight of all the individual features adds up to one*. This is because

the maximum weight that a question can carry is one.

1.2.4 Calculating Weights of various Sensors

As a part of profiling, users were asked to rank the various sensors available according to their preferences. The sensors ranked high are less sensitive to the users, whereas sensors that are ranked low are the ones that are the most sensitive to the user.

Let w_{f_s} be the weight assigned to the feature Sensors, w_{s_j} be the weight of the individual sensor s_j and let n_s be the number of sensors. In this feature Sensors *i.e.. which is a group*, there are individual sensors *i.e.. feature options*. For each of these sensors, their rank is denoted by r_{s_j} where s_j is an individual sensor. The way the individual sensor weights are to be

calculated depends on the rank assigned to them. The sensor that the user is most comfortable sharing its data gets least weight, whereas the sensor's data the user is least comfortable sharing get all the weight out of w_{p_s}.
To calculate the weight of individual sensors the following expression can be used :

$$w_{s_j} = (w_{f_s}/n_s) * r_{s_j}$$

In the above way, individual sensor weights are calculated . For this experiment, n_s = 6 because there are 6 different feature options involved.

1.2.5 Calculating Weights of various Buyers

In previous section the users were asked to rank the buyers from the one to which they are most comfortable sharing their data, to the one they are least comfortable sharing their data to. Let w_{p_b} be the weight of the parameter Buyers *i.e.. which is a group*, w_{b_k} be the weight of the individual buyer k *i.e.. which are sub-parameters* and n_b be the number of buyers available. For each buyer in the Parameter Buyers, each buyers is a sub-parameter. Each of their ranks is denoted by r_{b_k} where b_k is the individual buyers. If a buyer is ranked high*i.e. the user is most comfortable sharing to data to the buyer*, intuitively this buyer can be assigned a lower weight. This is due to the fact that it is known that the user will most likely anyway share the data with this buyer. Similarly, if a buyer is ranked lower by the user, the weight of this buyer is higher. Again, this is because it is known that the user is less comfortable sharing the data with this buyer. Each buyer is assigned a weight, where the weight of the individual buyer is a maximum of w_{f_b}. To calculate the weight of the individual buyers the following expression can be used:

$$w_{b_k} = (w_{f_b}/n_b) * r_{b_k}$$

In the above way, the individual buyer weights are calculated. For this experiment, n_b = 4 because the different possible buyers of data is 4.

1.2.6 Calculating Weights of various Contexts

The next step after ranking the Buyers is the ranking of the contexts. Users are asked to categorize the available contexts. Categorizing is similar to ranking, but instead of ranking, the users place the existing contexts in predefined slots, which are ranked by us. More than one context can fall in one slot. The total amount of weight that can be obtained by any context is w_{f_c}, which is the weight assigned to the feature Context. Let the individual weights of each context be w_{c_l} where c_l is any context and the number of contexts is n_c. In addition, let the number of Categories be n_{cat}, r_{cat_k} be the rank assigned by us to each category and the weight of the the individual categories to be w_{cat_z}, where cat_z is the individual categories into which contexts can be placed *i.e.. the Categories are the slots*. In this experiment, there

are n_{cat} = 3 Categories and n_c = 6 contexts.
Each Category is assigned a rank by us beforehand. That rank assigned to each category is referred to as r_{cat_z}. The user can therefore place the contexts into these categories. The contexts can take the weight of the category it is assigned to. Hence the weights of the categories are calculated, contexts assigned to that category get that exact same weight. In our case we have the Categories: *1. Very Useful, 2. Useful, 3. Not useful* assigned the respective ranks one, two and three. The higher ranked Categories get a lower weighting, and lower ranked categories are assigned a higher weighting. Intuitively, it can be taken as the contexts ranked higher are the ones the user will give his data for more easily. Similarly, the contexts ranked lower are the contexts the user will be more reluctant to give his data for. The weights of the categories can be calculated using the following expression:

$$w_{cat_z} = (w_{f_c}/n_{cat}) * r_{cat_k}$$

Once each context is placed in a category, it takes the weight of that category. For example, if the context c_y was placed in category cat_g, then:

$$w_{c_y} = w_{cat_g}$$

.

1.2.7 Calculating the Weights of each questions

The weight matrix consists of 3 parameters, *Sensors, Buyers and Contexts*. As discussed before, the combination of all the 3 sub-parameters gives us the number of questions NQ. Each cell of the matrix had the weight of that particular question, and each questions has a unique combination of sensor s_j, buyer b_k and context c_l. The way to calculate the value of a cell of the weight matrix $wq_{j,k,l}$, where j, k, l represents the index of the sub-parameters is using the following expression:

$$wq_{j,k,l} = w_{s_j} + w_{b_k} + w_{c_l}$$

The current question number CQ can be obtained by the following expression:

$$CQ = j * k * l$$

Some questions get weighted higher than some others, which means that there is a potential to obtain more credit *i.e.. it can be money or vouchers etc...* by answering such questions. The whole wq matrix is generated this way.

1.3 Personalized Survey Explained

The previous section 1.2.2 explained how the weight matrix is formulated in detail. In this section, the whole picture of the personalized survey along with how the users are assigned credits and privacy percentage is explained.

1.3.1 Profiling Questioning

To be able to create the weight matrix which explains which questions can get more credit than others, profiling questions are asked to the users. As explained in section 1.1, there are two kinds of profiling questions. One is explained in section 1.1.1. These are the questions which can be asked to the users for grouping users according to their traits during post processing of user choices. The second kind of profiling is explained in section 1.1.2. This can be done to obtain the weight matrix wq.

Users who live in Zurich (Switzerland) are briefed at ETH Zurich. Users who do not live in Zurich can be briefed on the phone or via Skype. Users are to be briefed all on the same day. This marks the first day of the experiment. After explaining the rules, terms and conditions of the experiment users start answering all the profiling questions. Making sure that the users do this step in front of us can make sure that the matrix wq is not formed wrongly. In other words, users can directly ask questions if anything is unclear and the basis of the experiment being wrong can be avoided.

1.3.2 The Preliminary Questioning

As mentioned before, there is a maximum of NQ questions that can be asked in this survey. From the questions asked in section **??**, the matrix wq is formed. Now, users are asked to answer all the NQ questions present in the survey all at once. No incentives, total credit or total privacy percentages are indicated. Users answer questions without any details, which can help analyse how users answer questions without any incentives or information. In addition, another angle from which this step is important is to clear the doubts of users. While we may not be able to help them during the experiment phase, since the questions are repeated users can clear their doubts while filling out the preliminary round of questions with us. This can make sure that users are clear about what they are filling out.
Once the questions are answered, this marks the end of the first day of the experiment. Users can return to their schedule.

1.3.3 The Actual Questioning

Once the preliminary questioning round is done, the actual questioning round starts. From this time on the whole experiment is split into sale period. Each sales period lasts for time duration T and has a budget B. In this case, one sales period is equal to one day. When the user sells the data on day 2 for example, the data is collected on day 3. The user can choose the data he/she wishes to sell the next day. All the questions asked in the preliminary round are asked again for every sales period. This time, the

user is presented with the credit associated with each of the answer that can be chosen for a question, along with the cumulated credit and privacy percentage. From this round, the interaction of the user with credits as well as the privacy can be observed. The round can be repeated *ro* number of time, which makes the duration of the experiment $ro + 2$ days. This is because of the first day of briefing people and because the data is sold for the next day, one more day is required to collect the data from the last day of data sales.

Bibliography

[1] Battery manager android. http://developer.android.com/reference/android/os/BatteryManager.html. Last Accessed: 2016-04-9.

[2] Corporation meaning. http://www.dictionary.com/browse/corporation. Last Accessed: 2016-04-10.

[3] Did you know how many different kinds of sensors go inside a smartphone? http://www.phonearena.com/news/Did-you-know-how-many-different-kinds-of-sensors-go-inside-a-smartphone_id57885. Last Accessed: 2016-04-9.

[4] Employer meaning. http://www.dictionary.com/browse/employer. Last Accessed: 2016-04-10.

[5] How does gps in a mobile phone work exactly? http://stackoverflow.com/questions/33637/how-does-gps-in-a-mobile-phone-work-exactly. Last Accessed: 2016-04-9.

[6] What is a gps? how does it work? http://www.loc.gov/rr/scitech/mysteries/global.html. Last Accessed: 2016-04-9.

[7] Y. Chen and H. Kobayashi. Signal strength based indoor geolocation. *Proceedings of the IEEE International Conference on Communications*, 1:436–439, 2002.

[8] PNishkam Raviand Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. *IAAI'05 Proceedings of the 17th conference on Innovative applications of artificial intelligence*, 3:1541–1546, 2005.

[9] Fredrik Gustafsson and Fredrik Gunnarsson. Mobile positioning using wireless networks. *IEEE SIGNAL PROCESSING MAGAZINE*, 2005.

[10] Ming Liu. A study of mobile sensing using smartphones. *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks*, 2013(272916), 2013.

[11] Pushpendra Singh, Nikita Juneja, and Shruti Kapoor. Using mobile phone sensors to detect driving behavior. *ACM DEV '13 Proceedings of the 3rd ACM Symposium on Computing for Development*, (53), 2011.

[12] Anna C. Vakil. Confronting the classification problem: Toward a taxonomy of ngos. *World Development*, 25(12):2057–2070, 1997.

[13] Shuangquan Wang, Canfeng Chen, and Jian Ma. Accelerometer based transportation mode recognition on mobile phones. *Wearable Computing Systems, Asia-Pacific Conference on, Wearable Computing Systems, Asia-Pacific Conference*, pages 44–46, 2010.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

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

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

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

Title of work (in block letters):

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

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

With my signature I confirm that

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

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

Place, date	**Signature(s)**

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