A Framework to Customize Privacy Settings of Online Social Network Users

Agrima Srivastava
Department of Computer Science and
Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad,India
Email: agrimasrivastava1@gmail.com

G Geethakumari
Department of Computer Science and
Information Systems
BITS Pilani, Hyderabad Campus
Hyderabad,India
Email: geetha@hyderabad.bits-pilani.ac.in

***Abstract*—Privacy is one of the most important concerns in an online social network. Online social network data is big data as millions of users are a part of it. The personal information of every user in an online social network is an asset that may be traded by a third party for its benefits. Individuals should be aware of how much of their personal information could be shared without risk. Different people have different requirements to share a profile item hence measuring privacy of such huge and diverse population is a challenging and complicated task in itself. In this paper we have proposed a framework that ensures privacy of individuals by allowing them to measure their privacy with respect to some specific people of their choice rather than measuring it with the entire population on online social networks. We have suggested a method to choose the best model to fit the real world data and to calculate the sensitivities of various profile items. The framework gives specific labels to users that indicates their profile privacy strength and enable them to customize their privacy settings so as to improve the privacy quotient. The users can also act as advisers to their online friends whose privacy quotients are low and thus spread privacy awareness in social networks.**

***Keywords*—customized privacy;privacy strength;item response theory**

I. Introduction

Online Social Networks (OSNs) have become the most popular Internet service to be used to date. It helps the users to stay connected with the world [1]. The users of OSNs like Facebook, Twitter, LinkedIn, MySpace etc have increased exponentially in the recent years. Using OSN people can interact with each other by sharing their personal information such as photos, likes, dislikes, interests, relationship status, job details, current town details, political views, religious views etc. [2]. There is an information explosion in the quantity and diversity of high frequency online social network data. This greatest advancement came with a price of privacy. With so much of personally identifiable information (PII) available online privacy concerns are bound to arise. Our PII is an asset to many third parties and can be mined and utilized for benefits. [3].

People are unintentionally and intentionally sharing their personal details without understanding the cost that they are paying for it. If personal data about individuals are collected, processed, stored and retrieved without their consent, their privacy is under threat. Exposure of one's private details can lead to identity theft, market benefits and embarrassment etc. OSN users are not aware of their online privacy leaks. They should know what and how much information they should share such that they are not at risk.

Privacy greatly depends upon the context. What is private to a particular individual may not be private to others. To measure a person's intelligence we need the intelligence quotient, in the similar fashion to measure the privacy of the user we need to know and understand the privacy quotient (PQ). There has to be a privacy measuring scale in which the individuals can be ranked according to their PQs and should know where they stand in comparison to the rest of the world. Calculating the privacy quotient for the entire social network population is a Big Data problem and a difficult task to undertake.

The privacy requirements are a function of the demography of the users as well as the social ties they have. Customizing the privacy settings of users by looking at such a diversified data does not really help because different people have different requirements and understanding of privacy which fetches them different privacy quotients. Instead of measuring the privacy of the whole network our framework allows the users to measure the privacy with the user profiles of their choice which reduces the data set greatly. Calculation of privacy quotient is analogous to the standard classical test theory. We have used the existing privacy measuring models to formulate a framework to measure the privacy of the users in an OSN with respect to their circle in the OSN.

Our framework considers naive approach, one parameter constrained and unconstrained data model and two parameter data model and selects the best model out of the four adaptive test models [4]. There are various approaches to measure the privacy of an online social networking user. Every approach deals with a different model and every model may not necessarily fit the real world data. Selecting the best model out of all is an important step. We have used the concept of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log likelihood to select the suitable model.

After calculating the privacy quotient using a suitable model we have fixed a range to the privacy quotient using the k means clustering algorithm and have labelled a strength to each of the range. Understanding text is a lot easier than

978-1-4799-2178-2/13/$31.00 ©2013 IEEE

numbers, hence by looking at the label the users will know the strength of their profiles in comparison with the their list of selected friends.

Our paper is organized as follows; Part II describes the related work in the field, in Part III we have given the algorithm of our framework to measure the Privacy Quotient of the user's surroundings and know the privacy strength of the user. Various groups will have its own range of privacy quotients hence in Part IV we have explained in detail the various steps enlisted in Part III by working out an example. Finally at the end in Part V we have given the conclusion and discussed our future work.

II. Related Work

Rasch introduced a psychometric model that is used to analyze the data on the basis of the respondent's ability and the difficulty of the item [5]. The mathematical theory behind the rasch model is the special case of the item response theory. Hiding an item reduces its utility hence Guo et al have proposed a framework which caters the utility needs of a profile item while ensuring privacy of users by mining the facebook privacy settings of users. They have developed a tradeoff algorithm to help the users to know their optimal privacy settings for a particular level of privacy concern and their personalized utility preferences [6]. Wiese et al have carried out a study on 42 participants and calculated their frequency of collocation and communication, closeness, and social group and showed that self-reported closeness is the strongest indicator of willingness to share and individuals are more likely to share in scenarios with common information [7].

Liu et al have provided an intuitively and mathematically sound methodology for computing the privacy scores of users in the OSNs by making use of the Item Response Theory model. They have used the two parameter logistic model to calculate the privacy scores of individuals [8]. Fang et al have proposed a template for social networking privacy wizard. They have used the fact that the real users conceive their privacy settings based on some rules which are implicit in nature. They proved that using machine learning techniques and with a limited amount of information and knowing the user's preferences they could calculate the privacy settings of the user automatically [9]. Aiello et al have tackled the trade-off problem between security, privacy and services in distributed social networks by providing the users the possibility to tune their privacy settings through a very flexible and fine-grained access control system [10].

We use a concept of *Friend Set* and best model selection technique for measuring the privacy strength which not only improves the target user's privacy but also sends an indication to the ones having a lesser privacy strength than the target user, thereby increasing the privacy awareness amongst the users.

III. The Proposed Framework

We describe below the steps involved in the calculation of privacy strength for a user in our proposed framework :

A: Input to the framework: Friend Set i.e the list of people in the user's friend list with whom they want to compare their privacy.
B: Formation of a dichotomous response matrix.
C: Selection of the best model that will fit the response matrix.
D: Calculation of sensitivity of the profile items.
E: Calculation of probability matrix.
F: Calculation of privacy quotient.
G: Deciding the range of privacy quotients using k means clustering algorithm.
H: Setting up labels for each of the ranges.
I: Output of the framework: Privacy Strength of the user's profile with respect to the Friend Set.

IV. Detailed explanation of the framework

Fig 1 given below gives a complete flow of the framework. To explain our point effectively we will explain the process of measuring privacy using values from a real data set.

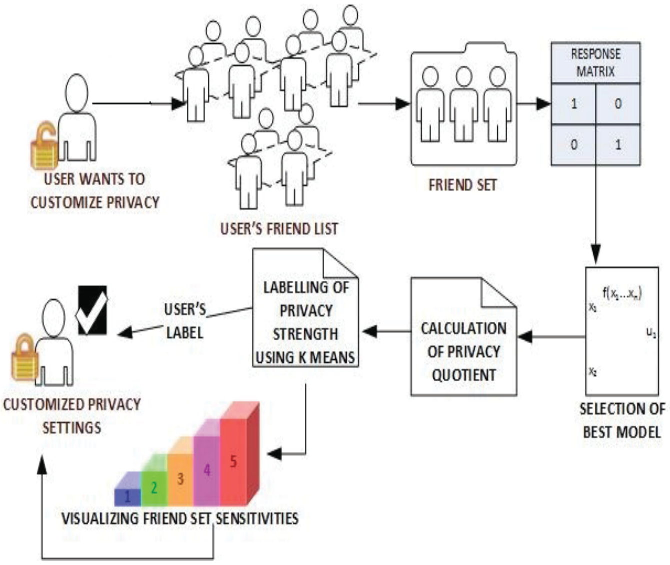
RESPONSE MATRIX
1
0
0
1
USER WANTS TO CUSTOMIZE PRIVACY
USER'S FRIEND LIST
FRIEND SET
$f(x_1...x_n)$
x_1
u_1
x_2
SELECTION OF BEST MODEL
CALCULATION OF PRIVACY QUOTIENT
LABELLING OF PRIVACY STRENGTH USING K MEANS
USER'S LABEL
CUSTOMIZED PRIVACY SETTINGS
1
2
3
4
5
VISUALIZING FRIEND SET SENSITIVITIES

Fig. 1: Proposed Privacy Measuring Framework.

To explain our proposed solution fully we will be using the data collected from a user having about 60 users in the Friend Set and then will collect the information for the following list of 11 profile items. Table I shows the list of 11 profile items. The cronbach's alpha for all these 11 items is 0.7228, which is ≥ 0.7 and hence will be able to measure the attitude of the user correctly. The accuracy of the model will be more if we use larger data set consisting of as many of our friends' profiles as possible that we want to include.

TABLE I: List of profile items

SNo	Profile Items
1	Contact Number
2	E mail
3	Address
4	Birthdate
5	Hometown
6	Current Town
7	Job Details
8	Relationship status
9	Interests
10	Religious Views
11	Political Views

A. Selecting the Friend Set

We have extracted the data sets from Facebook because Facebook provides an API for data collection. The users were asked to create the Friend Set i.e a set of users with whom they want to compare their privacy. Here we are working out the solution for a user having 60 friends in their Friend Set. The users can have a minimum of 10 and maximum of the number of digital friends they have in their profiles. We have assumed here that the user using this application will at least have 10 friends in their friend list.

B. Formation of response matrix

If we take n dichotomous variables for N users in the Friend Set then we can generate a Nxn dichotomous response matrix. If 1 ≤ j ≤ N and 1 ≤ i ≤ n then for an item i being shared by an individual j the value of *jth* row and *ith* column is marked as 1 otherwise is marked as 0.

C. Selecting the best model

The Item Response Theory (IRT) is widely used to measure the latent trait of an individual. It gives a relationship between the person's response and attitude. If there are j individuals and i profile items, and if the ability of each of the individual is θ_j then, P$_{ij}(\theta_j)$ is the probability that the jth individual with an ability θ_j will share an item having an index i [11]. A graph with θ_j on the x axis and P$_{ij}(\theta_j)$ on the y axis will give a S shaped curve that is known as the Item Characteristic Curve (ICC). Fig 2 shows an Item Characteristic Curve.

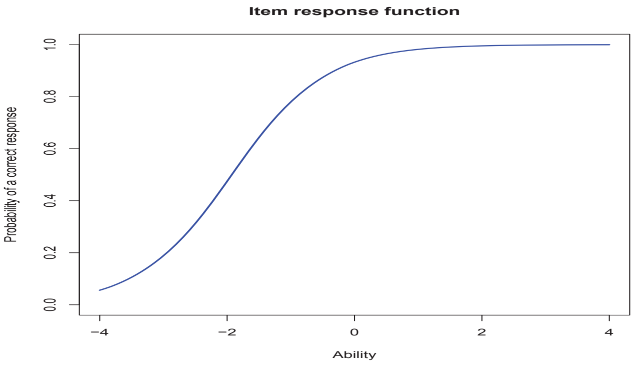

Fig. 2: The Item Characteristics Curve.

Each item characteristic curve has two properties namely β i.e the sensitivity and α i.e the discrimination constant. Sensitivity(β) denotes the difficulty of an item. The more is the sensitivity the less are the chances of it being shared. The item discrimination(α) is measured by the steepness of the ICC. A high value of discrimination constant can better differentiate between the individuals having a low and high value of P$_{ij}(\theta_j)$. In IRT each item has a set of item parameters and each individual has their ability i.e the attitude.

$$Prob(Sharing) = f(parameters, ability) \tag{1}$$

Equation 1 shows that the probability of an item being shared is a function of the model's item parameters and ability of the individual.

1) One Parameter Logistic Item Response Theory Model:

The one parameter logistic model is the simplest of all the item response models. This has a single parameter β_i that denotes the sensitivity of an item i. Having a constant α means that all the items are equally discriminating. The probability of a jth individual who is having an ability of θ_j for sharing an ith profile item is given by equation 2.

$$PR(\theta_{ij} = 1) = \frac{1}{1 + e^{\alpha_i(\theta_j - \beta_i)}} \tag{2}$$

where β_i is the sensitivity of the ith profile item, α_i is the discrimination constant of the ith profile item which is set to 1 for constrained model and set to a fixed value for an unconstrained model, θ_j is the ability of the jth user. One parameter logistic model can be categorized as constrained and unconstrained one parameter logistic model.

Calculation of sensitivity for constrained one parameter logistic model:
Table II gives us the values of sensitivity of all the 11 profile items and in Fig 3 we can see that at any given instance for a person with a specific ability the probability of sharing the date of birth (curve number 4) is much higher than the probability of sharing the address (curve no 3). In the one parameter logistic model the value of discrimination constant is constrained to 1 for all the items.

TABLE II: Discrimination and Difficulty Table

SNo	Profile Item	Discrimination	Sensitivity
1	Contact Number	1	2.9446946645
2	Email	1	0.6117859481
3	Address	1	4.5480914380
4	Birthdate	1	0.0000000000
5	Home Town	1	0.3316536983
6	Current Town	1	0.0003948577
7	Job Details	1	0.7386141987
8	Relationship Status	1	2.0182711543
9	Interests	1	1.3872733265
10	Religious Views	1	2.7736557816
11	Political Views	1	3.3946345724

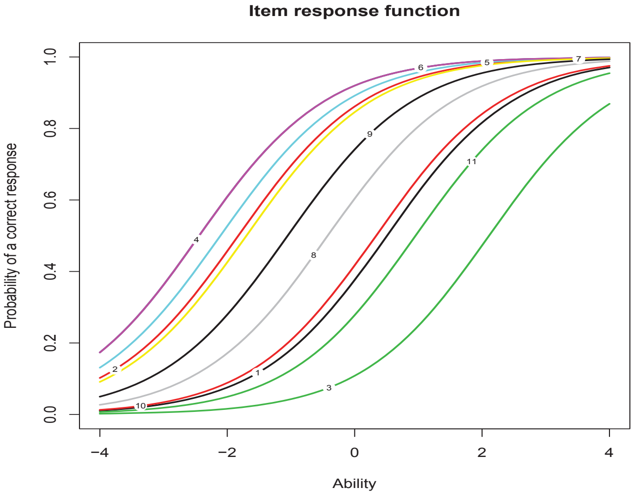

Fig. 3: An Item characteristic curve for one parameter constrained logistic model.

Calculation of sensitivity for unconstrained one parameter logistic model: Here the discrimination constant(α) for all the profile items is set to a single estimated value. Table III gives us the values of sensitivity and difficulty of all the 11 profile items. Here the estimated single value i.e the discrimination constant is 1.363163 for all the profile items.

TABLE III: Discrimination and Difficulty Table

SNo	Profile Item	Discrimination	Sensitivity
1	Contact Number	1.363163	2.323667e+00
2	Email	1.363163	4760522e-01
3	Address	1.363163	3.601252e+00
4	Birthdate	1.363163	3.453648e-06
5	Home Town	1.363163	2.573421e-01
6	Current Town	1.363163	0.000000e+00
7	Job Details	1.363163	5.753196e-01
8	Relationship Status	1.363163	1.585066e+00
9	Interests	1.363163	1.085306e+00
10	Religious Views	1.363163	2.187068e+00
11	Political Views	1.363163	2.683194e+00

Fig 4 is slightly steeper than Fig 3 and hence has got a better discriminating ability. The rate of change of probability from going to one ability level to another ability level increases faster than with what we see in the constrained one parameter logistic model.

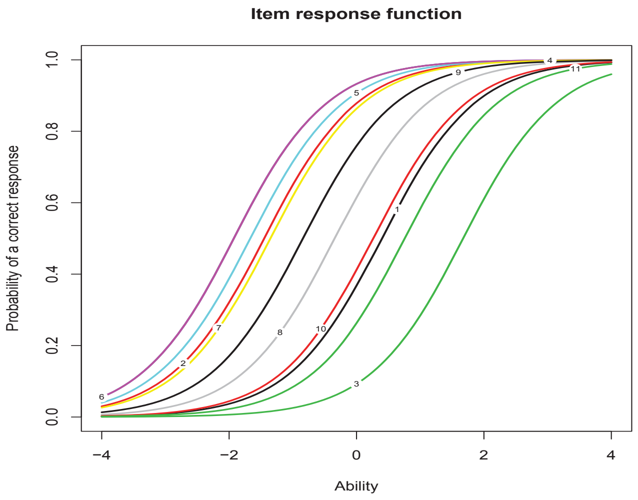

Fig. 4: An Item characteristic curve for one parameter unconstrained logistic model.

2) Two Parameter Model: The two parameter model calculates the probability of the user sharing an item based on the sensitivity as well as the difficulty.

$$Prob(Sharing) = f(\alpha, \beta, \theta) \tag{3}$$

Equation 3 shows that the probability of an item being shared in a two parameter model is a function of both the item parameters(α,β) and the ability i.e θ of the individual. The probability of a jth individual who is having an ability of θ_j for sharing an ith profile item is given by equation 4.

$$PR(\theta_{ij} = 1) = \frac{1}{1 + e^{\alpha_i(\theta_j - \beta_i)}} \tag{4}$$

where β_i is the sensitivity of the ith profile item, α_i is the discrimination constant of the ith profile item, θ_j is the ability of the jth user. Table IV gives us the sensitivity as well as the difficulty of all the 11 profile items. Unlike the one parameter model, the two parameter model has different discrimination values for each of the profile items.

TABLE IV: Discrimination and Difficulty Table

SNo	Profile Item	Discrimination	Difficulty
1	Contact Number	0.8744772	4.012698
2	Email	0.3523724	1.223004
3	Address	0.0000000	8.815682
4	Birthdate	0.2619243	0.000000
5	Home Town	0.7471443	1.658689
6	Current Town	0.8839744	1.527580
7	Job Details	1.1031453	2.283478
8	Relationship Status	0.5572550	3.144633
9	Interests	2.6932438	2.977822
10	Religious Views	1.9696065	3.783610
11	Political Views	10.5183067	4.009395

In Fig 5 we can see that each item has its own set of discrimination and sensitivity values. Political views (curve no 11) has got the highest discrimination value and address (curve no 3) has got the lowest of all.

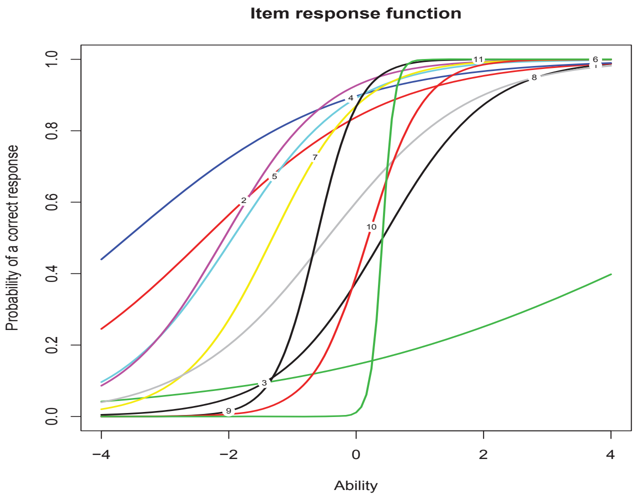

Fig. 5: An Item characteristic curve for two parameter logistic model.

3) Naive Model Approach: This approach does not seem to work well in comparison with the item response theory models. Complex IRT calculations can be avoided by making use of the naive approach if the number of profiles in the Friend Set is less than the number of profile items.

- *Calculation of sensitivity for naive model* : Sensitivity of an item i is the measure of the difficulty or sensitiveness of an item. The more an item is sensitive the least it is being shared. Sensitivity(β_i) of a profile item can be calculated as

$$\beta_i = \frac{N - |R_i|}{N} \tag{5}$$

where $|R_i|$ is the summation of the number of users who have shared an item i.

TABLE V: Sensitivity of the profile items using naive model

SNo	Profile Item	Sensitivity
1	Contact Number	.7
2	E mail	.2833
3	Address	.95
4	Birthdate	.2166
5	Hometown	.25
6	Current Town	.2166
7	Job Details	.3
8	Relationship status	.5166
9	Interests	.4
10	Religious Views	.6666
11	Political Views	.7833

Here in Table V though address is the most sensitive profile item but no discrimination can be made between birthdate and current town.

- *Calculation of visibility using the naive model* : Visibility is the popularity of an item in the network. Visibility of an ith item by the user j can be calculated as
$$V(i,j) = \frac{|R_i|}{N} X \frac{|R_j|}{n} \tag{6}$$
where $|R_i|$ is the summation of the number of times an item is shared by all the users and $|R_j|$ is the summation of the number of profile items shared by the user j.

D. Model Selection

Naive model is a population biased model and does not differentiate well between the sensitivities of two profile items. In table V birthdate and current town have the same sensitivity. Hence we will go for the naive model if and only if the number of users are less than the number of profile items. Most of the times the Friend Set is big enough and applying item response theory models is the best choice to go for. Our task is to select the best model out of constrained one parameter logistic model, unconstrained one parameter logistic model and two parameter model.

When we use a model to calculate the results we may not get the exact value. There is always a difference between the exact result and the results obtained. This happens because the selected model does not fit the data completely hence turns out to be erroneous. Our aim is to select the best model out of all such that the loss of information is minimized. In order to do that we should know the Information Criterion (IC) of the model. We will utilize the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for model selection. AIC and BIC values gives us the loss of information. AIC can be calculated as follows

$$AIC = -2(log - likelihood) + 2K \tag{7}$$

BIC can be calculated as follows

$$BIC = -2(log - likelihood) + Kln(N) \tag{8}$$

where K is the number of parameters used in the model. These values makes sense only when it is compared with the other models. As it gives us the information loss hence the model having their least value of AIC and BIC is preferred over all the models [12]. In table VI we have compared the constrained one parameter logistic model with the unconstrained one parameter logistic model and observed that the unconstrained one parameter logistic model gives the lowest AIC and BIC values. Hence at the end of the first comparison we select the unconstrained one parameter logistic model.

TABLE VI: Comparison of constrained and unconstrained models

Model	AIC	BIC	log.lik
Constrained 1PL	53498.80	53570.49	-26738.40
Unconstrained 1PL	53290.42	53368.63	-26633.21

In the second step we have made a comparison between the unconstrained one parameter logistic model and the two parameter model. In Table VII the AIC BIC values of the two parameter model were found to be minimum.

TABLE VII: Comparison of unconstrained and two parameter model

Model	AIC	BIC	log.lik
Unconstrained 1PL	53290.42	53368.63	-26633.21
2PL	51356.44	51499.81	-25656.22

Hence, the final model that could best fit our data is the two parameter model.

E. Calculation of Privacy Quotient and deciding the ranges

Selection of the best model is followed by the calculation of the privacy quotient. The privacy quotient can be calculated as

$$PQ(j) = \sum_i \beta_i * P_{ij}(\theta_j) \tag{9}$$

where β_i is the sensitivity and $P_{ij}(\theta_j)$ is the probability of sharing an item i by an individual j with an ability of θ_j. To decide the range of privacy quotients we have used the k means algorithm for 1000 iterations and have calculated five centers, sorting these centers gave us the upper limits of the five ranges.

TABLE VIII: No of users with the PQ in the given range

SNo	Range of Privacy Quotient	Percentage of users (out of 60%)
1	0.0 - 3.373010	6.61
2	3.373010 - 8.644307	12.84
3	8.644307 - 13.713378	14.14
4	13.713378 - 19.113382	10.14
5	19.113382 - 25.347364	16.36

The maximum and minimum of the privacy quotient obtained for the Friend Set of size 60 (hence 60%) were 29.2634 and 0.

F. Labelling the privacy strength

Understanding text is better than numbers and hence we have given labels to the range of privacy quotient. Each of the ranges have been categorized with various labels as ”High PQ”, ”Good PQ”, ”Average PQ”, ”Below Average PQ”, ”Poor PQ”. Table IX shows the mapping between the range of privacy quotients and the privacy strength

TABLE IX: Mapping of privacy quotient with privacy strength

SNo	Range of Privacy Quotient	Percentage of users (out of 60%)	Privacy strength
1	0.0 - 3.373010	6.61	High PQ
2	3.373010 - 8.644307	12.84	Good PQ
3	8.644307 - 13.713378	14.14	Average PQ
4	13.713378 - 19.113382	10.14	Below Average PQ
5	19.113382 - 25.347364	16.36	Poor PQ

In Fig 6 we can see that most of the users are having the privacy quotient in the range of 19.113382 - 25.347364 which implies Poor Privacy Quotient. This means that most of the users in the Friend Set of the user have a poor privacy strength. Here all the percentages are out of 60% and the lower the privacy quotient the better is the privacy strength.

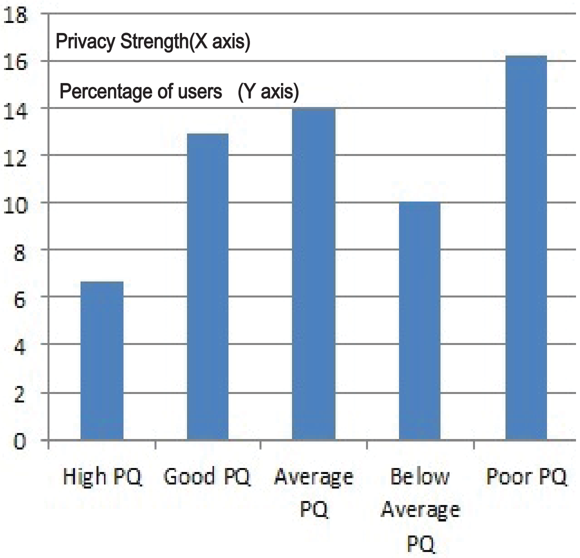

Fig. 6: A bar graph showing the percentage of users having a given privacy strength.

If the user has a privacy quotient below 13.713378 then his privacy settings will fall in any of the three categories i.e ”High PQ”,”Good PQ” and ”Average PQ” which is an acceptable privacy setting.

The user can now view the sensitivity of the various profile items in the group which will give an indication as what are the list of items that are extensively shared and what are the list of items that are being shared the least. The user can customize their privacy settings and can recheck their privacy strength. As most of the users in the Friend Set have got a poor privacy quotient the user can as well send an alert to those users alerting them of their privacy leaks.

V. Conclusion and Future Work

In this paper we have described a framework that would ensure privacy of a user by comparing with the privacy of other users in their friend list. We have made the use of various models like the Naive, one parameter logistic model (constrained and unconstrained) and two parameter model and have suggested a method to select the best fit model out of all using the AIC and BIC values. The framework utilizes the best model selected to calculate the privacy quotient and then determine the privacy strength of the user’s profile. In future we will be solving the problems of privacy in unstructured data and will be working on group privacy settings to measure loss of privacy through the members of the group and hence prevent privacy leaks due to group memberships.

Acknowledgment

We would like to thank KP Krishna Kumar for his help.

References

[1] J. Heer and D. Boyd, “Vizster: Visualizing online social networks,” in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, 2005, pp. 32–39.

[2] L. A. Cutillo, R. Molva, and T. Strufe, “Safebook: A privacy-preserving online social network leveraging on real-life trust,” *Communications Magazine, IEEE*, vol. 47, no. 12, pp. 94–101, 2009.

[3] R. Gross and A. Acquisti, “Information revelation and privacy in online social networks,” in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 2005, pp. 71–80.

[4] B. Krishnamurthy and C. E. Wills, “Characterizing privacy in online social networks,” in *Proceedings of the first workshop on Online social networks*. ACM, 2008, pp. 37–42.

[5] G. Rasch, “Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.” 1960.

[6] S. Guo and K. Chen, “Mining privacy settings to find optimal privacy-utility tradeoffs for social network services,” in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 656–665.

[7] J. Wiese, P. G. Kelley, L. F. Cranor, L. Dabbish, J. I. Hong, and J. Zimmerman, “Are you close with me? are you nearby?: investigating social groups, closeness, and willingness to share.” in *UbiComp*, 2011, pp. 197–206.

[8] K. Liu and E. Terzi, “A framework for computing the privacy scores of users in online social networks,” in *Data Mining, 2009. ICDM’09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 288–297.

[9] L. Fang and K. LeFevre, “Privacy wizards for social networking sites,” in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 351–360.

[10] L. M. Aiello and G. Ruffo, “Lotusnet: tunable privacy for distributed online social network services,” *Computer Communications*, vol. 35, no. 1, pp. 75–88, 2012.

[11] F. Baker and S.-H. Kim, *Item response theory: Parameter estimation techniques*. CRC Press, 2004, vol. 176.

[12] M. J. Mazerolle, “Appendix 1: Making sense out of akaikes information criterion (aic): its use and interpretation in model selection and inference from ecological data.”