Wright State University

CORE Scholar

[Browse all Theses and Dissertations](#)

[Theses and Dissertations](#)

2014

Mining Privacy Settings to Find Optimal Privacy-Utility Tradeoffs for Social Network Services

Shumin Guo

Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the [Computer Sciences Commons](#)

Repository Citation

Guo, Shumin, "Mining Privacy Settings to Find Optimal Privacy-Utility Tradeoffs for Social Network Services" (2014). *Browse all Theses and Dissertations*. 1421.

https://corescholar.libraries.wright.edu/etd_all/1421

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

Mining Privacy Settings to Find Optimal Privacy-Utility Tradeoffs for Social Network Services

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

by

Shumin Guo
M.E., Xidian University, 2008

2014
Wright State University

Wright State University
SCHOOL OF GRADUATE STUDIES

April 17, 2014

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Shumin Guo ENTITLED Mining Privacy Settings to Find Optimal Privacy-Utility Tradeoffs for Social Network Services BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master in Computer Science.

Keke Chen
Thesis Director

Mateen Rizki, Ph.D
Chair, Department of Computer Science
and Engineering

Committee on
Final Examination

Keke Chen Ph.D

Guozhu Dong Ph.D

Krishnaprasad Thirunarayan Ph.D

Robert E.W. Fyffe, Ph.D.,
Vice President for Research and
Dean of the Graduate School

ABSTRACT

Guo, Shumin. M.S., Department of Engineering and Computer Science, Wright State University, 2014. *Mining Privacy Settings to Find Optimal Privacy-Utility Tradeoffs for Social Network Services*.

Privacy has been a big concern for users of social network services (SNS). On recent criticism about privacy protection, most SNS now provide fine privacy controls, allowing users to set visibility levels for almost every profile item. However, this also creates a number of difficulties for users. First, SNS providers often set most items by default to the highest visibility to improve the utility of social network, which may conflict with users' intention. It is often formidable for a user to fine-tune tens of privacy settings towards the user desired settings. Second, tuning privacy settings involves an intricate tradeoff between privacy and utility. When you turn off the visibility of one item to protect your privacy, the social utility of that item is turned off as well. It is challenging for users to make a tradeoff between privacy and utility for each privacy setting. We propose a framework for users to conveniently tune the privacy settings towards the user desired privacy level and social utilities. It mines the privacy settings of a large number of users in a SNS, e.g., Facebook, to generate latent trait models for the level of privacy concern and the level of utility preference. A tradeoff algorithm is developed for helping users find the optimal privacy settings for a specified level of privacy concern and a personalized utility preference. We crawl a large number of Facebook accounts and derive the privacy settings with a novel method. These privacy setting data are used to validate and showcase the proposed approach.

List of Symbols

Chapter 2

n	Number of persons in a test
k	Number of questions in a test
s_{ij}	The grade for person i on question j
θ	Trait of a person
θ_i	Trait of person i
α	Discrimination of a question
α_j	Discrimination of a question j
β	Level of difficulty of a question
β_j	Level of difficulty of question j
I_j	Profile item j of Facebook
U_i	User i of Facebook
ϕ	Utility expectation of a user
λ_j	Discrimination level of utility for item j
μ_j	Level of utility for item j
ω_j	Weight of profile item j
$\vec{w}_i$	Utility weight of user i
$\mathcal{W}$	Utility weight matrix for n users on k profile items
S	Profile item configurations for n users on k items

Contents

1	Introduction	1
1.1	Online Social Network	1
1.2	The Online Social Network Privacy Problem	2
1.3	Contributions	5
2	Modeling the Facebook Privacy Problem	7
2.1	Latent Response Theory (LRT)	7
2.2	Modeling Privacy and Utility for User Profiles	10
2.2.1	Definitions and Notations	10
2.2.2	Modeling User's Privacy Concern	11
2.2.3	Finding Optimal Privacy Configuration.	13
2.2.4	Modeling User's Utility Preference	14
2.3	Mining the Tradeoff between Privacy and Personalized Utility	17
3	Experiments	19
3.1	Preparing Data	19
3.1.1	Modeling User's Privacy Concern	23
3.1.2	Characteristics of Our Sample Data	26
3.1.3	Modeling User's Utility Expectations	29
4	Related Work	32
5	Conclusions and Future Work	34
	Bibliography	40

List of Figures

2.1	Sample logistic functions for modeling the probability of giving correct answers.	9
2.2	A screen shot of facebook's privacy control.	11
2.3	β of the item model can be explained as how sensitive the item is perceived by users. The lower the β is, the more users want to hide the item.	12
2.4	Illustration of finding the best tradeoff cases. The scatter plot of privacy vs. utility forms a best tradeoff curve. With any given privacy concern level θ , the best tradeoff utility level is the crossing of the horizontal line and the best tradeoff curve.	18
3.1	Histogram of the number of changed profile items.	21
3.2	The percentage of users sets the profile item to "disclosed".	22
3.3	The sketch of the item models for privacy modeling.	23
3.4	The number of users at different levels of privacy concern: θ	25
3.5	Comparing the privacy rating and the weighted number of hidden items. The optimal privacy configuration can be found from the blue 'x' points that are close to the red 'o' points.	26
3.6	Top: the the number of users at different levels of utility preference for un-weighted utility model. Bottom: for weighted utility model (overweighting "phone", "email", and "employers" by a factor 1.1)	27
3.7	The tradeoff between the privacy and utility for unweighted utility model.	28
3.8	Each point in the tradeoff graph represents one user. The points around the upper bound contain the optimal tradeoffs.	29
A.1	Data Flow of Experiments	35

List of Tables

3.1	The number of friends and friends of friends for three accounts.	20
3.2	Based on the item visibility from the real account and the fake account, we can derive the privacy setting. "Y" for Yes, "N" for No and "N/A" for impossible.	20
3.3	Top items with the lowest and highest β in the range $[-4, 4]$ for privacy modeling.	24
3.4	Crossing-model testing results represented with RMSE. Model i is trained with TestData i . The close RMSEs in each row means the corresponding model has similar goodness-of-fit for the three sets of data.	28
3.5	Mean and variance of users' θ values for different models.	29

Acknowledgment

I would like to thank my advisor Keke Chen. Dr. Chen dedicated lots of time and effort in the years of my graduate studies. I also would like to thank my classmates and friends who has helped my during the process.

Besides, I would like to thank the rest of my thesis committee members: Dr. Guozhu Dong and Dr. T.K. Prasad, for their insightful comments, and questions.

Dedicated to
my wife Min Han and my son Shawn Guo
without whom it would not have been accomplished

1. Introduction

1.1 Online Social Network

One of the most popular topics today on the internet is about social network. A social network is a social structure made up of a set of social actors and a set of ties between these actors [28]. These online networking sites resemble the structure of people and their relationships in a social community and make the interaction of people much easier and a lot faster.

In reality online social network is not a new idea. Early internet based system such as the bulletin board, mailing list and discussion groups etc. are the social networks of the early internet age. The property of such systems is that they are highly oriented towards people with enough computer skills. Things are changing with web 2.0 [29] become the trend of internet. The technologies for Web 2.0 makes it possible to design social networking sites that are accessible for people without enough computer skills. It features friendly user interface, much easier operations and much better media supports including text, images and videos. One other important reason for the popularity of social network is contributed by mobile devices. With mobile devices and online social networks, it becomes possible for people to keep connect any time and from any where. To this sense, online social networking is reforming the way people live everyday. And it has and will enforce tremendous effects on our society.

As social network services (SNS) are becoming a popular phenomenon, people use so-

cial network services as effective means to do different kinds of things. For example, some social networks such as Facebook and Google Plus are helping people to get connected with friends, sharing photos, daily activities and so on. The social network site LinkedIn [15] is helping people to get connected with employers and their colleagues as well as finding jobs and interesting topics of their interests. The dating site match.com [20] is helping singles to find their loved ones. Other technical social network sites such as github [7] and stackoverflow [25] are greatly facilitating collaboration among software developers. For example, github users can collaborate on software projects with people all around the world and stackoverflow users can ask and answer technical questions, discussion technical ideas and so on.

Online social networks are breaking the boundaries of countries and continents. They are helping people in an unprecedented way to work together, to enjoy together and even to live together. It promote the proliferation of information, and promotes the knowledge among people. It is reasonable to believe that online social networks will be an important part in our networked society and more and more online social network site will be build to help people better their life and work.

1.2 The Online Social Network Privacy Problem

When online social networks become popular, many people are joining in one or even a few online social network sites. But in order to join these sites, you actually need to have yourself registered with your personal information. For example, a lot of social networking sites require an user to provide a user name, the real name, birth date, address, phone number etc. For many people, these requirements are reasonable. And from a technical perspective, the pieces of information help the online systems identify and index a specific user. For example, if you want to find someone from New York city, the location information New York City becomes an critical piece of information. If the user you are looking

Visibility Level	Potential Audiences
Account Owner	Account Owner Only (private)
Friends	Account Owner and his/her friends
Friends of Friends	Account owner, friends and friends of friends
Everyone	Everyone that can access Facebook (Public)

for did not provide his/her address information, it is unlikely that you can find such a person. Another example for you to provide personal information is when you are using the job seeking web sites such as LinkedIn[15]. If the information you provided to LinkedIn is not complete or not valid, you will probably will miss out important job opportunities. The purpose for most people to use the social network is to benefit from it, in this thesis we use the word **utility** to describe such benefits.

Although it is necessary for the online social networking sites to collect personal private information, the information will be out of control by a user once it is provided to the social networking site. In such a case, some consequences might happen. For example, the social network service provider can use the personal sensitive information in an unexpected way such as selling it to a third party etc. Another example is that some curious users might use the web search tools to find the information one shared on the social network site. These raise the personal privacy concern on the social networking sites. In this thesis, we will use the word **privacy** to describe users' privacy concern of disclosing their private information when using a social network. The disclosing of private information can raise certain risks such as information spam [19, 6], insurance discrimination [30], and financial fraud [13]. More and more privacy incidents [3, 24] have educated people the risk of disclosing private information. As a result, social network services have provided privacy configuration options for users to tune their privacy preference. For example, Facebook provides four visibility options (the account owner only, friends only, friends of friends, and everyone) for 27 profile items. Table 1.2 summarizes the possible audiences who can

view the item for these four visibility level settings.

However, the fine-grained privacy tuning also brings great challenges to users. First, many users may have no sufficient knowledge and patience to tune them and choose to use the default privacy settings. SNS providers often set the visibility to the most public level "everyone" for most items, to maximize the utility of the social network as a whole. Users need to check item by item manually to tune the settings, which is very inconvenient. Second, when users tune privacy settings, the utility of SNS can be implicitly affected, although the users may not be aware of it. For example, hiding contact information such as email and phone number, the user may miss potential employers who want to contact the user. Therefore, privacy settings also reflect an intrinsic tradeoff between privacy and utility.

Users, who are confused with many privacy settings or unaware of the tradeoff between privacy and utility, will need an assistant tool to help them understand the settings and recommend candidate settings aligned with their utility preference. New users of SNS may not be aware of privacy risks [1], but they have much clearer intention on using the SNS, i.e., getting connected with friends, and reaching out to potential employers, etc. From the utility-centered view, it seems quite simple to set the visibility of the profile items. For example, a user looking for jobs via the social network can simply set all the related profile items visible, such as address, phone, email, and education, for potential employers evaluating and contacting her/him. However, when the user knows the potential privacy risk for each item, she/he may not be so decisive - how should an "acceptable" privacy level be specified? does item X have higher privacy risk than item Y? And which items can be selectively exposed so that the overall privacy risk is still acceptable?

1.3 Contributions

In this research thesis, we aim to develop a framework to help a new user to understand the potential privacy risks based on the fellow users' view, and achieve balanced and personalized privacy settings for a social network service. This approach tries to answer the following questions: can existing users' privacy settings be modeled to provide guidance for a new user to choose desired optimal settings? Can we also derive implicit utility preference from privacy settings? How to find the optimal tradeoff between privacy and a user-specified utility?

We mine real privacy settings crawled from a large number of user accounts in Facebook to answer the above questions. Although individual user's settings might be noisy, we observed surprisingly stable patterns with a large number of users. Based on the crawled data, we employ the latent trait theory [4] to model the relationship between the privacy settings and the existing users' latent levels of privacy concern. With this model, the new users can clearly understand the relationship between each profile item and the privacy risk perceived by other users, and use this knowledge to appropriately specify his/her own privacy concern. The same set of data, with an appropriate transformation, can also be used to model the users' implicit utility preference. The utility modeling algorithm is further revised to incorporate the new user's utility preference and results in a *utility preference model* biased towards this user's preference.

Based on the privacy and utility models, we design an algorithm to find the best settings on a new user's privacy and utility requirements. Once the privacy and utility models are determined, each user's privacy settings in the training data can be mapped to a trade-off between privacy and the personalized utility. As the new user's utility preference is already reflected in the utility model, we can easily find the tradeoff cases that meet the user's privacy concern and maximize the preferred utility. Therefore, the framework can satisfactorily address the aforementioned challenges of personalized privacy settings for a social network.

Our approach has a number of unique contributions.

- We introduce the utility aspect into privacy settings, and allow users to find optimal settings based on the tradeoff between privacy and utility.
- We develop a method to derive implicit utility preference from privacy settings, and allow the utility modeling biased towards a user-specified utility.
- The experimental studies are based on a large amount of real privacy settings (from over ten thousands of user accounts) obtained from the Facebook social network. These data will be made available to the public with appropriate data sanitization.

In Chapter 2, we first review the latent trait theory that we will use to model the privacy concerns and utility preference, then propose a best tradeoff model for finding optimal privacy settings. In Chapter 3, we focus on the experimental studies on real Facebook data. The related work is presented in Chapter 4 and we summarize this thesis in Chapter 5.

2. Modeling the Facebook Privacy

Problem

In this section, we first present the Latent Response Theory. Then, we model the social network privacy problem using the LRT model. Lastly, we present a tradeoff model for user to find optimal tradeoff between privacy and utility.

2.1 Latent Response Theory (LRT)

Latent trait theory (or item response theory) [4] is widely accepted and used in psychometrics to design tests to measure human subjects' abilities (or attitudes). The basic idea is to build a mathematical model based on the subjects' answers to a questionnaire. The difficulty level of question is determined by the number of people who answer the question correctly. Intuitively, the person who has higher ability should get more correct answers. The latent trait model tries to find the probability relationship between the person's ability and his/her answers to the questions. This model can be used to understand the properties of the questions, and thus determine whether these questions can be used to properly evaluate subjects' certain trait. Latent trait theory can also be applied to understand subjects' strength of an attitude. For example, in our study, we try to model the SNS user's level of privacy concern or implicit utility preference as the latent trait, using the users' privacy

settings (equivalent to the questionnaire).

The latent trait model assumes there exists a function modeling the relationship between the person's ability and the probability of giving the correct answer to a question, which is often modeled as a logistic function [10] as shown in Figure 2.1. Let the person's trait be θ . The popular two-parameter model [4] is described in equation 2.1.

$$Pr(\theta; \alpha, \beta) = \frac{1}{1 + \exp\{-\alpha(\theta - \beta)\}} \quad (2.1)$$

where α and β are model parameters to be learned from the data. Figure 2.1 illustrates some sample distributions. X-axis represents the level of ability and Y-axis is the probability of giving the correct answer to the item. When $\theta = \beta$, $Pr(\theta) = 0.5$, thus, β indicates the level of difficulty of the question for the ability test. β can have different meanings when the model is used for different purposes (we will explain it for our purpose of modeling). The slope of the curve $Pr(\theta)$ at β is $\alpha/4$, which indicates the discrimination level of the question. If the slope is small, the probabilities of giving correct answers for the persons of ability $\theta > \beta$ and for those of ability $\theta < \beta$ have small difference, implying low discrimination.

It is impossible to use one question to model the latent trait. In practice, we use multiple questions (e.g., k questions) for a number of persons (e.g., n persons), which result in a $n \times k$ grading matrix $\{s_{ij}\}$. This matrix is used to learn the item models: $Pr(\theta; \alpha_j, \beta_j)$, $j = 1 \dots k$. Maximum likelihood estimation [4] is the most popular method for estimating the parameters α_j , β_j , and also each person's ability θ_i from the data. The goodness of fit measure [4] checks the difference between the probability $Pr(\theta; \alpha_j, \beta_j)$ and the frequency of answering the question j correctly by the users at each specific level θ_i . If the sum of differences for all items and users follows a χ^2 distribution with p-value < 0.05 , the model fits the data well.

The learning process also finds an ability level θ_i for each person in the test. Often, the

Figure 2.1: Sample logistic functions for modeling the probability of giving correct answers.

sum $\sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$ is used to represent the adjusted average score for the answers to the questionnaire at ability level θ_i . Correspondingly, the answers of a specific person of ability θ_i are evaluated with the weighted original score: $\sum_{j=1}^k \alpha_j s_{ij}$, which should be close to $\sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$ for a well-fitting model. We will use the adjusted average score to represent the privacy rating at a certain level of privacy concern.

There are a few unique benefits for using latent trait modeling for our work. First, we can consider the profile items are the "questions" that the user need to answer, which reflects the level of privacy concern. Correspondingly, we can map the "ability" axis to the level of privacy concern. Second, the item-related parameter α_i can tell us how important this item works to describe the level of privacy concern. β_i tells how people perceives this item in terms of privacy. In next section, we will come to detailed modeling and explain the rationale behind that.

2.2 Modeling Privacy and Utility for User Profiles

Our purpose of this research is two-fold. First, we want to understand the relationship between privacy settings and user’s levels of privacy concern. Second, we want to study the tradeoff between privacy and utility in different privacy settings. In this section, we will focus on the modeling methods and then present the experimental results on real data sets in next chapter.

2.2.1 Definitions and Notations

Each person’s account has a number of profile items that can be set to different levels of visibility. These items can be email, birthdate, phone number, religious view, etc., which are denoted as $\{I_1, I_2, \dots, I_k\}$. In Facebook¹, these items can be set to be visible by ”everyone”, ”friends of friends”, ”friends only”, or ”only me”, as shown in Figure 2.2. For simplicity, we assume the disclosure setting is binary by merging the settings. It is not difficult to extend the model to multiple settings [18]. Considering ”friends of friends” are still strangers, we use ”disclosed” to represent ”everyone” or ”friends of friends”, and ”hidden” to other cases. Assume there are n users. The setting for item I_j of user U_i is denoted as s_{ij} . Thus, n users will generate a $n \times k$ matrix of privacy setting.

In studying the level of privacy concern, we set ”hidden” to 1 to represent the preservation of privacy, and ”disclosed” to 0 to represent the loss of privacy. We call the binary setting vector for a specific user *a privacy configuration*. One of the problems for a new user of SNS is to find a right privacy configuration for his/her account. As disclosing or hiding one item is also directly related to the social utility of that item, privacy settings can also be used to represent the user’s implicit utility preference. When we study users’ utility preference, we can set ”disclosed” to 1 to represent preserved utility, and ”hidden” to 0.

¹Facebook’s privacy control interface has been changed since we collected data, one additional option ”Custom” has been added to allow users to specify a list of users who can view an item

App Settings

On Facebook, your name, profile picture, cover photo, gender, networks, username, and user id are always publicly available, including to apps ([Learn Why](#)). Apps also have access to your friends list and any information you choose to make public.

Figure 2.2: A screen shot of facebook’s privacy control.

Correspondingly, we call the flipped binary setting vector *a utility configuration*. We will mine these data sets to discover the intrinsic relationships.

2.2.2 Modeling User’s Privacy Concern

Modeling user’s privacy concern is the first step to understand the tradeoff between privacy and utility. This task can help understand how the privacy settings are linked to the SNS user’s level of privacy concern, which will be extended to include the utility preference. Intuitively, the more profile items the user tries to hide, the higher level of privacy concern the user may have. However, items may have different weights in evaluating the privacy concern. For example, the items such as phone number have well-perceived privacy risk, and thus most users tend to hide these items. However, highly privacy conscious users may also try to hide some items that most other users think OK to disclose. As a result, two users with the same number of hidden items may have very different levels of privacy concern, and the total number of hidden items may not accurately reflect the user’s privacy concern. The latent trait theory can be applied to exactly address this kind of problem.

In our study, we define the latent trait as the level of user’s privacy concern, denoted

as θ , and each profile item as the question in the latent trait model. The privacy concern of user U_i is a global variable consistent for the privacy settings of different items. We use θ_i to describe user U_i 's privacy concern. The item parameters α_j and β_j , $j = 1 \dots k$, and the user's privacy concern θ_i , $i = 1 \dots n$, can be jointly learned from the privacy setting data.

Figure 2.3: β of the item model can be explained as how sensitive the item is perceived by users. The lower the β is, the more users want to hide the item.

Figure 2.3 shows the probability distributions of three items among 27 items learned from our experimental data sets. The result of modeling gives a quantitative description on users' privacy concern. Although the values on X-axis have no absolute meaning, they represent a relative measurement to the levels of privacy concern. The item "relationships" has the smallest β . It implies that even users with low level of privacy concern will set this item to "hidden". In other words, the privacy risk of disclosing it is well perceived by most users. In contrast, the item "current_city" is only set to "hidden" by some users who have high level of privacy concerns. These users are more concerned with their privacy than most other users, while most users may consider that disclosing this item has very low privacy risk and may bring more utility, such as getting connected with unknown people in

the same city.

Different from the other two, the item "networks" has a very low discrimination factor, meaning it is not an important factor for describing the level of user's privacy concern.

2.2.3 Finding Optimal Privacy Configuration.

With the latent trait model, each user will get an estimated level of privacy concern, θ_i . The weighted expected number of hidden items for the level θ_i is uniquely defined as $\sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$, where α_j serve as the weights for the items. We define this measure as *the privacy rating* at the privacy concern level θ_i . There is a relationship between the actual weighted number of hidden items and the privacy rating, according to the theory of latent trait model [4].

Theorem 1 *For an ideally fitting model, the actual weighted number of hidden items $\sum_{j=1}^k \alpha_j s_{ij}$ for the user U_i is the privacy rating, i.e.,*

$$\sum_{j=1}^k \alpha_j s_{ij} = \sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j) \quad (2.2)$$

where θ_i is the user U_i 's level of privacy concern.

This theorem gives the basis for finding the optimal privacy configuration for a desired level of privacy concern. A new user can specify his/her level of privacy concern θ based on the relative measure in the range [-4,4]: we can map the level of privacy concern to the nine grades, from the lowest level of concern -4 to the highest 4. With the specific θ , Algorithm 1 can be used to find the optimal privacy configuration based on the fellow users' privacy concerns. Algorithm 1 assumes the item models $Pr(\theta; \alpha_j, \beta_j)$, $j = 1 \dots k$ have been learned from the training data. Also, the level of privacy concern for each user in the training data has been calculated: θ_i , $i = 1 \dots n$. The algorithm first calculates the

privacy rating for the user-specified level of privacy concern θ , and then searches over the training dataset to find out the users having the similar level of privacy concern $|\theta - \theta_i| < \epsilon$, where $\epsilon > 0$ is a very small value, and these users' privacy configurations form a candidate set, from which the optimal privacy configuration is found so that the weighted number of hidden items $\sum_{j=1}^k \alpha_j s_{ij}$ is close to the privacy rating $\sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$, according to Theorem 1.

Algorithm 1 OptPrivacyConfig($\theta, \{\theta_i\}, \{(\alpha_j, \beta_j)\}, S$)

```

1: Input:  $\theta$  the desired level of privacy concern;  $\{\theta_i\}$ : the levels of privacy concern for the users
   in the training data;  $\{(\alpha_j, \beta_j)\}$ : the parameters for the item models;  $S$ : training data
2:  $T \leftarrow$  find the set of  $\theta_i$  so that  $|\theta - \theta_i| < \epsilon$ ;
3:  $p \leftarrow \sum_{j=1}^k \alpha_j Pr_{\alpha_j, \beta_j}(\theta_i)$ ;
4:  $d_{opt} \leftarrow \sum_{j=1}^k \alpha_j$ ;
5: OptConfig  $\leftarrow \emptyset$ ;
6: for each user  $U_i$  having  $\theta_i \in T$  do
7:    $p1 \leftarrow \sum_{j=1}^k \alpha_j s_{ij}$ ;
8:   if  $d_{opt} > |p - p1|$  then
9:      $d_{opt} \leftarrow |p - p1|$ ;
10:    OptConfig  $\leftarrow U_i$ 's privacy configuration;
11:   end if
12: end for
13: return OptConfig;

```

2.2.4 Modeling User's Utility Preference

Note that Algorithm 1 only gives the way to find the privacy configuration for a specific level of privacy concern, without considering the factor of utility. The other important goal of our research is to understand the tradeoff between privacy and utility, which gives a comprehensive guidance for users to set the disclosure level of their private information.

A Simple Utility Model

The next task is to model user's utility preference. We notice that utility and privacy is a pair of complementary factors - once one decides not to disclosure an item, any social utility with this item is disabled as well. Based on this understanding, we propose the

simple modeling method.

We transform the data to reflect users' *implicit* consideration on utility preference. A privacy setting of "disclosed" now means preserving the utility of that item, while "hidden" means a loss of utility. They are mapped to 1 and 0, respectively, which is a flipped version of the data for privacy modeling. Let $\bar{s}_{ij}$ be such a setting used for utility modeling, i.e., $\bar{s}_{ij} = 1 - s_{ij}$.

We use the same latent trait modeling method to derive the simple model for utility preference. For clear presentation, we use different symbols to represent the utility models. Let ϕ be the level of utility preference and (λ_j, μ_j) be the pair of parameters for item I_j . The utility item model is

$$Pr(\phi; \lambda_j, \mu_j) = \frac{1}{1 + \exp\{-\lambda_j(\phi - \mu_j)\}}. \quad (2.3)$$

The smaller the value of μ_i is, the more people choose to set the item "disclosed". Similarly, λ_j serves as the discrimination factor as α_j does for privacy models. Because of the use of the flipped version of data, it is easy to derive that $\lambda_j = \alpha_j$ and $\mu_j = -\beta_j$.

We can also similarly define the *utility rating* for the ideal utility configuration at the level of utility preference ϕ_i , which is $\sum_{j=1}^k \lambda_j Pr(\phi_i; \lambda_j, \mu_j)$. According to Theorem 1, we have

$$\sum_{j=1}^k \lambda_j Pr(\phi_i; \lambda_j, \mu_j) = \sum_{j=1}^k \lambda_j \bar{s}_{ij} \quad (2.4)$$

for ideal model fitting. Because $\bar{s}_{ij} = 1 - s_{ij}$ and $\lambda_i = \alpha_i$, it follows

$$\begin{aligned} & \sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j) + \sum_{j=1}^k \lambda_j Pr(\phi_i; \lambda_j, \mu_j) \\ & \approx \sum_{j=1}^k \lambda_j + \sum_{j=1}^k (\alpha_j - \lambda_j) s_{ij} = \sum_{j=1}^k \lambda_j, \end{aligned} \quad (2.5)$$

which implies that the privacy and utility ratings are linearly related under the simple utility

modeling. This relationship is critical for making the tradeoff between privacy and utility.

A Personalized Utility Model

Users often have clear intentions for using social network services, while having less knowledge on privacy risks. Among the set of well-known social utilities, a new user may want to specify his/her utility preference, and wish it is considered in utility modeling. For simplicity, we describe the utility preference as the importance weights on the items, denoted as a vector $w = (w_1, w_2, \dots, w_k)$. For the items less relevant to the utility, the user can assign 1 to them, while for a few relevant items, the user can assign a value larger than 1 to overweigh them. Note that the user can also demote certain category of utility by setting a weight less than 1 for the related items.

We develop an algorithm to incorporate these weights in learning the personalized utility model. Intuitively, we want to overweight those samples consistent with the target user's preference. The key is to appropriately revise the likelihood function to incorporate the weights for learning. Let's start with the formulation of the likelihood function. Since the privacy setting is a binary variable, the probability to set item I_j to $\bar{s}_{ij}$ follows the Bernoulli distribution:

$$\begin{aligned}
& Pr(S = \bar{s}_{ij} | \phi_i, \lambda_j, \mu_j) \\
&= Pr(\phi_i; \lambda_j, \mu_j)^{\bar{s}_{ij}} (1 - Pr(\phi_i; \lambda_j, \mu_j))^{1-\bar{s}_{ij}} \\
&= \frac{\exp\{-\lambda_j(\phi_i - \mu_j)(1 - \bar{s}_{ij})\}}{1 + \exp\{-\lambda_j(\phi_i - \mu_j)\}}.
\end{aligned} \tag{2.6}$$

Normally, the training of latent trait model holds the assumption to simplify model learning that a user's setting for an item is independent of other users and other items [4]. With this

assumption, we have the likelihood function

$$\begin{aligned}
L(\lambda, \mu, \phi|S) &= \prod_{i=1}^n \prod_{j=1}^k Pr(S = \bar{s}_{ij}|\phi_i, \lambda_j, \mu_j) \\
&= \frac{\exp\{\sum_{i=1}^n \sum_{j=1}^k (-\lambda_j(\phi_i - \mu_j)(1 - \bar{s}_{ij}))\}}{\prod_{i=1}^n \prod_{j=1}^k (1 + \exp\{-\lambda_j(\phi_i - \mu_j)\})}
\end{aligned} \tag{2.7}$$

where λ , μ , and ϕ represent $\{\lambda_j\}$, $\{\mu_j\}$, $\{\phi_i\}$, $j = 1 \dots k$, $i = 1 \dots n$, respectively. Over weighting the item I_j by w_j ($w_j > 1$) is equivalent to adding more samples of I_j setting to the training data, which can be directly reflected in the following likelihood function.

$$L(\lambda, \mu, \phi|S, w) = \frac{\exp\{\sum_{i=1}^n \sum_{j=1}^k (-\lambda_j(\phi_i - \mu_j)w_j(1 - \bar{s}_{ij}))\}}{\prod_{i=1}^n \prod_{j=1}^k (1 + \exp\{-\lambda_j(\phi_i - \mu_j)\})} \tag{2.8}$$

We revise the original MLE learning algorithm to use this likelihood function. We will skip the details of learning algorithm, and will show some experimental results later.

2.3 Mining the Tradeoff between Privacy and Personalized Utility

Note that for each user in the training data, we can derive a pair of privacy concern and utility preference (θ, ϕ) , which corresponds to a pair of privacy rating and utility rating (p, u) . These pairs represent the tradeoffs between the privacy and utility. We want to find the ones that satisfy the specific level privacy concern and, meanwhile, maximizing the utility.

Figure 2.4 illustrates a sample distribution of (p, u) pairs as we observed in experiments. A new user can specify a desired level of privacy concern θ . With the (p, u) pair distribution, we can find the pair and the corresponding privacy configuration. The

Figure 2.4: Illustration of finding the best tradeoff cases. The scatter plot of privacy vs. utility forms a best tradeoff curve. With any given privacy concern level θ , the best tradeoff utility level is the crossing of the horizontal line and the best tradeoff curve.

following Algorithm 2 can be used to find the optimal privacy configuration with user's specification on utility preference.

Algorithm 2 OptTradeoff(θ, S, w)

- 1: Input: θ : desired level of privacy concern; S : n users' privacy settings on k items; w : the weight vector for utility preference: $(w_1, w_2, \dots, w_k)$.
 - 2: $\{\alpha_j, \beta_j, \theta_i\} \leftarrow$ Learn the privacy model with S ;
 - 3: $\{\lambda_j, \mu_j, \phi_i\} \leftarrow$ Learn the personalized utility model with S and w ;
 - 4: **for** each user U_i **do**
 - 5: $p_i \leftarrow \sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$;
 - 6: $u_i \leftarrow \sum_{j=1}^k \lambda_j Pr(\phi_i; \lambda_j, \mu_j)$;
 - 7: $t_i \leftarrow (U_i, p_i, u_i, \theta_i, \phi_i)$;
 - 8: **end for**
 - 9: $T \leftarrow$ find the cases t_i that $|\theta - \theta_i| < \epsilon$;
 - 10: $t_{opt} \leftarrow$ find the case in T that has the maximum u_i ;
 - 11: return the user U_{opt} 's privacy configuration;
-

3. Experiments

The experimental study will be focused on three aspects: (1) preparing the real data and study the characteristics of privacy settings, (2) using the data to train the proposed privacy models and studying the properties of the item models; and (3) understanding the tradeoff between privacy and utility with simple and weighted utility models.

3.1 Preparing Data

The first challenging problem is to get real privacy settings for a sufficiently large number of users. Asking users to do survey is a time consuming process, and is often limited to very small scale [17]. We need a large number of samples to reduce the model bias. We design a novel approach to crawl and derive real privacy settings from a sufficiently large number of Facebook accounts. This approach utilizes the friends of friends (FoF) networks of Facebook accounts and derives the item privacy settings based on the visibility of the item crawled using the normal account and a fake account with no friend.

Concretely, the method is described as follows. First, we login to Facebook with a normal account and record all the FoF users of this login user and at the same time crawl the visible profile items of these FoF accounts. In order to reduce sampling bias, we have selected three independent user accounts for this data crawling purpose, and in the data cleaning process, FoF accounts are selected for each account such that the percentage of the overlap is less than 5%. The number of FoF accounts of one Facebook user often

reaches thousands. Table 3.1 shows the number of FoF accounts reached by the three normal Facebook accounts after cleaning the crawled data. Second, we login to the fake user's account that has no friend, and crawl the profile items of the recorded FoF accounts of normal users again.

	RealUser1	RealUser2	RealUser3
# of friends	99	83	73
Original # of FoF	4104	5287	4376
Cleaned # of FoF	3798	4215	3552

Table 3.1: The number of friends and friends of friends for three accounts.

Visible to Real User?	Visible to Fake User?	Privacy Setting
Y	Y	E
Y	N	FoF
N	Y	N/A
N	N	O/F

Table 3.2: Based on the item visibility from the real account and the fake account, we can derive the privacy setting. "Y" for Yes, "N" for No and "N/A" for impossible.

We can derive the privacy settings of the FoF accounts based on the visibility from the real user and the fake user's accounts. At the time we crawl the data, Facebook's privacy setting for each profile item has four levels: only the account owner who can see the item (O), only the friends (F), friends of friends (FoF), and everybody in the social network (E). If the real user can see a profile item of his/her friend of friend, we can simply infer that the friend of friend has set this item's visibility to FoF or E. Otherwise, the friend of friend may set the visibility to O or F (or simply does not provide). The following inference matrix (Table 3.2) can be used to derive the real privacy setting for the real user's FoF accounts.

Note that for items invisible to both the real user and fake user, the observed user may not even set the item values. This happens normally for items such as "political views"

and "graduate school" (e.g., the user did not attend any graduate school). However, it is consistent to treat them as "hidden" cases.

According to the inference matrix, we can surely derive three levels of privacy setting: O/F, FoF and E for each item. For simplicity, we merge two of the settings to derive the binary settings: "disclosed" and "hidden". There are two possible merging schemes: (1) $E \rightarrow \text{"disclosed"}; O/F \text{ or } FoF \rightarrow \text{"hidden"}$; or (2) $E \text{ or } FoF \rightarrow \text{"disclosed"}; O/F \rightarrow \text{"hidden"}$. We consider friends-of-friends still strangers, and thus use scheme (2) in our experiments.

Figure 3.1: Histogram of the number of changed profile items.

We can crawl 27 profile items, including bio, birthday, family, relationships, religion views, political views, websites, photos, videos, links, notes, hometown, current city, education (high school, college, and graduate school), employers, activity, interests, phone number, sex, facebookuri, and networks. In the default settings, except for "interests", "political views", "religious views", and "phone number" set to "friends only", all other items are set to "everyone".

Figure 3.2: The percentage of users sets the profile item to "disclosed".

The collected dataset is then cleaned and analyzed to find the statistics of the change of privacy settings. To determine how many profile items are changed for each account, we compare the derived settings with the default settings. Figure 3.1 shows the distribution of the number of items changed by the users crawled from Account 1. Since we consider no-value items as "hidden" items, Figure 3.1 shows different distribution from previous study [17]. Among 3798 friends of friends of Account 1, everyone has changed some of the profile items, while previous study shows about 20-30% users not changing any privacy setting [9]. As we have mentioned, it is consistent to treat no-value items as "hidden" items. Thus, it will not affect our modeling.

Figure 3.2 shows the list of items sorted by the percentage of users setting the item to "disclosed" (E or FOF). The figure shows that almost all the crawled users have disclosed "friend list", indicating users' lack of knowledge on the implications or little privacy concern caused by this item, and on the contrary, a very small percentage of users have disclosed address and phone number, possibly because their associated privacy risks are well perceived by most users.

3.1.1 Modeling User’s Privacy Concern

In this set of experiments, we want to explore the relationship between the level of user’s privacy concern and the item privacy settings.

The proposed modeling method (described in Section 2.2.2) is applied to generate the item probability models. We pool the friends-of-friends records crawled from the three user accounts and perform a five-fold cross validation to check the model validity. Concretely, the data are randomly partitioned into five shares of equal size. In each round, we use four of the five shares for training and one share for testing. The testing result is evaluated by the goodness of fit measure [4]. The p-value is used to check whether the model fits the testing data well (p-value < 0.05 suggests good fit). The testing results on the five folds show the models are in high quality, with all p-values < 0.02.

Figure 3.3: The sketch of the item models for privacy modeling.

According to the latent trait theory, the items with β values in $[-4, 4]$ are the most

informative ones. Figure 3.3 shows the sketch of the learned item distribution models with β in $[-4, 4]$. Small β means many users tend to hide the item - the privacy risk is highly perceived. Large β indicates only highly privacy conscious users tend to hide the item and the privacy risk of that item might not be well perceived. In the item models trained with the pooled data, we found the items "family" and "phone" have β values smaller than -4, which means most users tend to hide these items. "college" has a β value larger than 4, which makes sense since Facebook is originally designed for college students and thus "college" is normally disclosed to distinguish where users are from. We select the top items with the lowest β and the highest β in $[-4, 4]$, as shown in Table 3.3. "favorite_quotations" appears in the lowest five, partially because most people do not have favorite quotations. It appears very few people try to hide interests, education, and employers, as they are also the common things that people like to share with other people in real life.

Group	Profile Items
The lowest five	sex, photos_album, myposts, favorite_quotations, birthday
The highest five	interests, interested_in, grad_school, high_school, employers

Table 3.3: Top items with the lowest and highest β in the range $[-4, 4]$ for privacy modeling.

The other parameter α serves as a discrimination factor, showing how much the item can contribute to distinguishing the level of privacy concerns. They are important in calculating the privacy rating of a privacy configuration. We found most items have α parameters around $[1, 2.5]$. Only "family", "college", "networks", and "phone number" have α less than 1. This makes sense because most people hide "family" and "phone number", and most people disclose "college" according to their β values. Although "networks" has an appropriate β ($\beta = 1.79$), its setting seems a random choice, not very related to user's privacy concern.

We can also learn each user's level of privacy concern θ_i from the data. Figure 3.4

Figure 3.4: The number of users at different levels of privacy concern: θ .

shows the distribution of the number of users according to the level of privacy concern. There is no user with $\theta < -2$. Most users' levels of privacy concern are around the median $\theta = 0$ and in the range $[-1.5, 2]$. Interestingly, there is a spike around $\theta = -1.2$, and there are also very few persons having $\theta > 2$. Overall the mean θ is -0.063 .

Figure 3.5 shows the relationship between the number of hidden items and the privacy rating. Since each user has a unique θ_i , we can plot the weighted number of hidden items, $(\theta_i, \sum_{j=1}^k \alpha_j s_{ij})$, for the user's privacy configuration. Meanwhile, for each θ_i , there is a unique privacy rating

$\sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$ (the red curve in the figure). The figure shows the weighted number of hidden items are around the corresponding privacy rating. We can find the optimal

Figure 3.5: Comparing the privacy rating and the weighted number of hidden items. The optimal privacy configuration can be found from the blue 'x' points that are close to the red 'o' points.

privacy configurations around the points on or close to the curve of the privacy rating.

3.1.2 Characteristics of Our Sample Data

One may have questions on the bias of our data collection method, i.e., sampling the FoF data for only three accounts is not representative enough. We argue that we only use this dataset to demonstrate our methods, not intent to generate global models for the whole Facebook community. However, the FoF data collected from different accounts indeed show some interesting characteristics, which suggest data at this scale (e.g., thousands of examples) might be sufficient to capture the global privacy-concern model.

Figure 3.6: Top: the the number of users at different levels of utility preference for un-weighted utility model. Bottom: for weighted utility model (overweighting "phone", "e-mail", and "employers" by a factor 1.1)

We use the following method to understand whether the similarity between the three sets of FoF data. First, we train the privacy concern model for each of the three accounts. Then, we test each model with the three sets of FoF data, to find the $(\theta_i, \sum_{j=1}^k \alpha_j s_{ij}, \sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j))$ for each user U_i in the dataset. $r_{\theta_i} = \sum_{j=1}^k \alpha_j Pr(\theta_i; \alpha_j, \beta_j)$ is the ideal privacy rating for this user's level of privacy concern θ_i , and $p_i = \sum_{j=1}^k \alpha_j s_{ij}$ is the real privacy rating according to this user's privacy configuration. If a model perfectly fits the data, then these two values will be identical. Assume there are n instances in the testing dataset. We define the *root-mean-square-error* (RMSE) as $\sqrt{1/n \sum_{i=1}^n (p_i - r_{\theta_i})^2}$. Table 3.4 shows that for each model the three test datasets result in very close RMSE. We also check the distributions of resultant users' θ values and find their mean and variance values are also very close (Table 3.5, where "combined" means the model learned from the

Figure 3.7: The tradeoff between the privacy and utility for unweighted utility model.

	TestData1	TestData2	TestData3
Model1	2.624510	2.470485	2.416606
Model2	1.990507	1.725497	1.674528
Model3	2.701275	2.311214	2.122907

Table 3.4: Crossing-model testing results represented with RMSE. Model i is trained with TestData i . The close RMSEs in each row means the corresponding model has similar goodness-of-fit for the three sets of data.

combined data.). These results all suggest a certain level of similarity between the three datasets.

In addition, since we have carefully selected the three accounts to make their FoFs have low overlapping ($< 6\%$ overlapping for each pair of accounts), it seems that each dataset is sufficiently large to capture some shared global model. We will conduct more

Figure 3.8: Each point in the tradeoff graph represents one user. The points around the upper bound contain the optimal tradeoffs.

	Model1	Model2	Model3	Combined
Mean of θ	-0.0625	-0.0680	-0.0521	-0.0659
Variance of θ	0.7087	0.6765	0.7031	0.6884

Table 3.5: Mean and variance of users' θ values for different models.

experiments to verify this observation.

3.1.3 Modeling User's Utility Expectations

Users often consider privacy settings based on an implicit tradeoff between privacy and utility. As we have discussed in Section 2.2.4, users' privacy settings can also be used to model their utility preference. We use the same datasets with flipped values ("disclosed"

is set to 1 and "hidden" to 0) to model utility preference. We train the simple utility model with this flipped dataset. As we have expected, the resultant item models satisfy the relationship $\lambda_j = \alpha_j$ and $\mu_j = -\beta_j$. Due to the similarity with the privacy model, we skip the details for the simple utility model.

We can tune the utility model towards a specific user's preference of utility. A user may specify a special use of social network services, such as job hunting. We then provide a tradeoff function that is biased to the specific utility. This biased utility model is achieved by overweighting the items related to the specified utility and learning the weighted utility model as we described in Section 2.2.4. In the experiment, we overweight "employers", "phone" and "email" by 1.1 to make the privacy settings friendlier to job hunting. The weight 1.1 can be understood as adding 10% more training samples that have the overweighted items disclosed. The resultant weighted model well fits the data with p-value < 0.02 . How to pick weight is a complicated issue, which will be addressed by our future work. Here, we just want to show how the result looks by introducing the weights.

By comparing the weighted model and the original model, we found the β parameters for the weighted items are slightly reduced and the α parameters are significantly increased. Figure 3.6 shows the change of the number of users at different levels of utility preference.

Tradeoffs Without using the weighted scheme, we have shown that the privacy rating is linearly correlated with the utility rating. Figure 3.7 shows the relationship derived from the real dataset. As we expected, the relationship between these two factors is approximately linear. Overall, the distribution is biased to the utility side. However, it may not be necessary that the users care utility more than privacy. More likely, many users are not aware of the privacy risks and thus leave the default settings of most items unchanged. As Barnes [1] mentioned, most young people care less about their privacy because they are unaware of privacy risks. Because most Facebook users are young, it explains the overall pattern. Other studies [21, 27] show that different groups of people may have different preferences over privacy and utility of social network services, which will be explored in

our future work.

Figure 3.8 shows the relationship between privacy and the weighted utility (with the specified items overweighted by 1.1, and other items by 1), which is not linear anymore. Note the points around the upper bound of the distribution represent the most interesting cases. The tradeoff algorithm (Algorithm 1) takes these cases and generates optimal trade-offs for a specific level of privacy concern.

4. Related Work

Privacy leaking may result in serious attacks to the SNS users. Spamming and phishing [8, 12, 11, 19, 2], are probably the most popular attacks that utilize private information, which may cause significant financial loss of the targeted users. As more and more people are becoming SNS users, the privacy issues of SNS have great impacts on the whole society.

Fang et al. [5] consider the difficulty of tuning privacy settings for many profile items, especially for a new SNS user. They propose to learn classification models from existing users of the SNS. The classification models try to find the correlation between the user’s account features, such as age, sex, and education level, and the privacy setting for each profile item. New users can use the classification models to automatically determine their privacy settings. There are several problems with this approach. First, the classification models depend on and are biased by the training data, which do not incorporate the new user’s privacy preference. Second, it is difficult to evaluate whether the classification models suggest appropriate settings, when the training data does not support high-accuracy classification models. Third, this approach does not consider a privacy setting is a tradeoff between privacy and utility.

Liu et al. [16] proposed a framework for computing privacy score of users based on users’ privacy settings. The latent trait model is also used to model the privacy score. Similarly, the privacy setting data is used to generate the item models. The sensitivity of an item is defined as the item model parameter β_j . The visibility concept is introduced, which is the privacy setting s_{ij} for item I_j by user U_i . The privacy score is defined as $\sum_{j=1}^k \beta_j s_{ij}$, which

looks like the weighted number of hidden items in our approach. However, the formula is difficult to explain with the latent trait modeling theory. In addition, this approach does not consider the utility factor as well.

Yabing Liu et al. [17] studies privacy settings of SNS users by using survey data. The survey dataset is collected from 200 Facebook users recruited via Amazon Mechanical Turk. Their results show that most users use default settings provided by the SNS, and users have trouble to correct the default privacy settings, which supports the importance of our study.

SNS users' attitudes to the privacy issues of SNS were studied by Barnes [1]. The author observed the privacy paradox problem of online social network: adult people are concerned about the invasion of privacy while teenage people freely give up their personal information. The main reason for the paradox is that teens are usually inexperienced and thus unaware of the problems. Our study shows that overall people care more about utility of Facebook than the privacy issues, which is probably due to the fact that most users of Facebook are young people.

The tradeoff between utility and privacy is also studied by Norberg et al. [21]. Their survey data reveals that people indeed provide significantly greater amounts of personal information than they expected. They find that perception of risks has a negative effect on peoples' information sharing behavior, while on the other hand perceived trust or benefit doesn't have a clear positive effect on people's sharing behavior. Utz et al. [27] also studied the personal factors such as personal impression, narcissism and their relation to users' social network information sharing behaviors.

5. Conclusions and Future Work

In this thesis, we propose a framework to address the tradeoff between privacy and utility for users' privacy settings of a social network service. We use latent trait theory to model users' levels of privacy concern and levels of utility preference. We develop an algorithm to incorporate users' personalized utility preference in learning the utility model. Finally, a tradeoff algorithm is provided for users to find the optimal privacy configuration based on the desired privacy level and utility preference. The whole approach is validated and showcased with a large real dataset crawled from Facebook.

The future work will be conducted along several directions. First, we will continue to work on the problem of learning weighted utility model. In this thesis, we did not answer how to set item weights for a user's utility preference, which can be intricate. Further study is needed to find (or learn) the appropriate weighting scheme for any specific type of utility, so that users do not need to worry about the utility weights. Second, the current personalized utility modeling considers the users having mixed utility preference. If we can cluster the users by their utility preference, we can probably derive better privacy setting. However, clustering users by their utility preference is another challenging problem. We also consider collaborative filtering (CF) [26, 22, 14, 23] might be a relevant technique for recommending privacy settings because our problem seems to have some similar settings to the recommending systems.

Appendix A

We are using generic web scraping techniques to collect data from Facebook web pages¹. The data flow graph for the experiments are described in Figure A.1.

Figure A.1: Data Flow of Experiments

Java Code to Crawl Facebook Profile Data

It is a challenging task to get user profile data from Facebook like social networking sites. One reason is that a large portion of the web page contents are generated dynamically

¹The code shown in this part will be available on github at <https://github.com/gsm1011/puts>.

Listing A.1: Code Snippet for Crawling Facebook User Profile Raw HTML Data

```
1 private static void scrollToBottom(WebDriver driver) throws ...
   InterruptedException {
2     assert(driver != null);
3
4     // scroll page to the end.
5     while (true) {
6
7         Long y = (Long) ((JavascriptExecutor) ...
            driver).executeScript("return window.pageYOffset");
8         long y1 = y.longValue();
9
10        ((JavascriptExecutor) driver).executeScript("window.scrollTo(0, ...
            document.body.scrollHeight);");
11        Thread.sleep(1800);
12        y = (Long) ((JavascriptExecutor) driver).executeScript("return ...
            window.pageYOffset");
13        if (y1 - y.longValue() == 0) break;
14    }
15 }
```

using Ajax. This makes the simple web page downloading tools not working as expected. Because such tools typically download a web page by sending simple HTML requests, and a large portion of the dynamic content can not be obtained. For example, if one user has a large number of friends, this method will download only a small portion of his/her list of friends. Another reason is that social networking sites are enforcing restrictions on high frequency HTTP/HTTPS requests. This means that we need to respect this restriction on our crawling software to ensure our crawling tool work continuously.

The code snippet in listing [A.1](#) shows how to trigger the Ajax code in order to download all the web page content from Facebook.

Java Code to Generate XML Data from Raw HTML Data

The goal of this data cleaning step is to remove the HTML tags and unrelated contents. The result will be an XML data file for each user. The code in listing [A.2](#) shows how to get

clean data for the "About" information for a user and save the information into an XML file.

R Code to Plot the Graphs

The R programming language is a popular statistical language for exploratory data analysis as well as machine learning analysis. We use the "ltn" package to model our data using the LRT model. The code snippet in listing [A.3](#) shows how to plot the histogram of the users' setting changes for each profile item.

Listing A.2: Code Snippet for Cleaning the About Information of a User from HTML Data

```
1  try {
2      we = ...
        driver.findElement(By.xpath("//div[@id='pagelet_timeline_medley_info']"));
3  } catch (Exception e) {
4      return ;
5  }
6
7  out.write("<about>");
8  //// Work and education
9  we1 = we.findElement(By.xpath("//div[@id='eduwork']"));
10 wes = we.findElements(By.xpath("//div[@id='eduwork']/table/tbody"));
11
12 out.write("<edu_work>");
13
14 for (int i = 1; i ≤ wes.size(); i++) {
15     // the group label such as "graduate school", is hidden, so a ...
        normal xpath search can't find it.
16
17     we1 = ...
        we.findElement(By.xpath("//div[@id='pagelet_eduwork']/div/table/tbody[" ..
        + i + "]/tr[1]/th"));
18     label = ...
        (String) ((JavascriptExecutor) driver).executeScript("return ...
        arguments[0].innerHTML;", we1);
19     label = label.replaceAll("<[^>]*>", " " ..
        ").trim().toLowerCase().replace(' ', '_');
20
21     we1 = ...
        we.findElement(By.xpath("//div[@id='pagelet_eduwork']/div/table/tbody[" ..
        + i + "]/tr[1]/td"));
22
23     if (we1 == null) txt = "";
24     // txt = we1.getText().replaceAll("<[^>]*>", " ").trim();
25     txt = cleanTxt(we1.getText());
26
27     out.write("<" + label + ">");
28     out.write(txt);
29     out.write("</" + label + ">");
30 }
31 out.write("</edu_work>");
```


Listing A.3: Code Snippet for Plotting the Privacy Setting Histogram

```
1 postscript(file="priv.theta.hist.eps");
2 num <- 40
3 maximum <- 500
4 # par(mfrow=c(3,1));
5 hist(all.ltm.theta, num, main=NULL, xlab=(expression(theta)), ...
      xlim=c(-3,3), ylim=c(0,maximum), col="red");
6 # legend(1.8, 600, c("Utility", "Weighted Utility", "Privacy"),
7 #       lty=1:1, col=c("red", "green", "blue"));
8 hist(all.wgt.ltm.theta, num, main=NULL, ...
      xlab=(expression(theta[w])), xlim=c(-3,3), ylim=c(0,maximum), ...
      col="green");
9 hist(all.priv.ltm.theta, num, main=NULL, xlab=(expression(phi)), ...
      xlim=c(-3,3), ylim=c(0,maximum), col="blue");
10 dev.off();
```

Bibliography

- [1] Susan B. Barnes. A privacy paradox: Social networking in the United States. *First Monday*, 11(9), September 2006.
- [2] Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. Identifying video spammers in online social networks. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, pages 45–52, New York, NY, USA, 2008. ACM.
- [3] B.Schiffman. Facebook ceo apologizes, lets users turn off beacon, 2007.
- [4] R J de Ayala. *The Theory and Practice of Item Response Theory*. The Guilford Press, 2008.
- [5] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 351–360, New York, NY, USA, 2010. ACM.
- [6] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th annual conference on Internet measurement*, IMC '10, pages 35–47, New York, NY, USA, 2010. ACM.
- [7] github. <https://github.com/>.

- [8] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security, CCS '10*, pages 27–37, New York, NY, USA, 2010. ACM.
- [9] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, New York, NY, USA, 2005. ACM.
- [10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.
- [11] Markus Huber, Martin Mulazzani, Sebastian Schrittwieser, and Edgar Weippl. Cheap and automated socio-technical attacks based on social networking sites. In *Proceedings of the 3rd ACM workshop on Artificial intelligence and security, AISec '10*, pages 61–64, New York, NY, USA, 2010. ACM.
- [12] Markus Huber, Martin Mulazzani, Edgar Weippl, Gerhard Kitzler, and Sigrun Goluch. Exploiting social networking sites for spam. In *Proceedings of the 17th ACM conference on Computer and communications security, CCS '10*, pages 693–695, New York, NY, USA, 2010. ACM.
- [13] Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, 2007.
- [14] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. Grouplens: Applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):77–87, March 1997.
- [15] LinkedIn. <http://www.linkedin.com>.

- [16] Kun Liu and Evimaria Terzi. A framework for computing the privacy scores of users in online social networks. *ACM Trans. Knowl. Discov. Data*, 5:6:1–6:30, December 2010.
- [17] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11, pages 61–70, New York, NY, USA, 2011. ACM.
- [18] N.T. Longford. *Random Coefficient Models*. Clarendon Press, Oxford, 1993.
- [19] Benjamin Markines, Ciro Cattuto, and Filippo Menczer. Social spam detection. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '09, pages 41–48, New York, NY, USA, 2009. ACM.
- [20] match.com. www.match.com.
- [21] Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1), 2007.
- [22] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.
- [23] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

- [24] Brennon Slattery. Google hit with lawsuit over google buzz. <http://bit.ly/1dkHTw5>, Feb. 2010.
- [25] stackoverflow. <http://stackoverflow.com/>.
- [26] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.
- [27] Sonja Utz and Nicole C. Kramer. The privacy paradox on social network sites revisited: The role of individual characteristics and group norms. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 3(2), 2009.
- [28] Wikipedia. http://en.wikipedia.org/wiki/Social_network.
- [29] Wikipedia. http://en.wikipedia.org/wiki/Web_2.0.
- [30] James Williams. Social networking applications in health care: threats to the privacy and security of health information. In *Proceedings of the 2010 ICSE Workshop on Software Engineering in Health Care*, SEHC '10, pages 39–49, New York, NY, USA, 2010. ACM.