Quantifying Privacy Risk for Web Based Social Networks

1 Introduction

Privacy protection of Web Based Social Network(WBSN) has attracted great concern in recent years. But how to quantify the privacy risks faced by WBSN users is still a challenging problem. Past work on this topic has been attempted by Justin et al.[1], Tran Hong et al. [2] and Kun et al. [3]. Kun et al. proposed a model to calculate the privacy score of WBSN users. The model uses two intuitive metrics, sensitivity and visibility, as the factors for privacy score calculation. But we argue that the model proposed by Kun et al. are problematic both in interpretation of sensitivity and visibility and experimental evaluation. And based on the arguments, we propose methods for better calculation and interpretation of these two factors.

2 Related Work

Kun et al. [3] proposed a framework to estimate the privacy score for each social network user. Intuitively, the privacy score increases with the sensitivity of the information being revealed and with the visibility of the information gets in the social network. And the privacy score calculated for each user is a combination of each partial privacy score of each one of the user's profile item settings e.g. name, address, hometown, ssn etc..

The primary input of this framework is an $n \times N$ response matrix, where n is the number of profile items user can set and N is number of users. Each element in the response matrix is assumed to be natural numbers, and higher values means that user is more willing to disclose the profile item. And based on the input from users represented as response matrix, the privacy score is calculated using theories from Item Response Theory (IRT). Besides, the structure of social network can affect the privacy score calculation.

Definition of Sensitivity

The sensitivity of item $i \in \{1, \dots, n\}$ is denoted by β_i, This property depends on the item itself. And it is common that some items are more sensitive than than other items.

Definition of Visibility

Visibility describes the scale of visibility of item i for user j, the more it spreads, the higher visibility. And it is defined as:

$$V(i,j) = P_{ij} \times 1 + (1 - P_{ij}) \times 0 \tag{1}$$

where P_{ij} =Prob$\{R(i,j) = 1\}$.

Privacy Score Calculation from Sensitivity and Visibility

The privacy score of individual j due to item i, denoted by $\textsc{Pr}(i,j)$, can be any combination of sensitivity and visibility. And for simplicity, we use:

$$\textsc{Pr}(i,j) = \beta_i \times V(i,j)$$

And the privacy score for user j can be calculated by:

$$\textsc{Pr}(j) \sum_{i}^{n} \beta_i \times V(i,j) \tag{2}$$

IRT-Based Privacy Score Calculation

In order to calculate the privacy score, we need to calculate β_i and the visibility which is represented using the IRT theory.

$$P_{ij} = \frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \tag{3}$$

3 Redefining Privacy Score Calculation Model

In this section, we argue that the interpretation and calculation of sensitivity and visibility are problematic and based on these arguement, we propose a new method for the calculation of sensitivity and visibility.

3.1 Sensitivity

According to Section 2, sensitivity is interpreted as parameter β_i w.r.t. the IRT model. And by estimating parameter β_i using maximum likelihood estimation, we can obtain this parameter and thus the sensitivity for each profile item i. In the IRT theoretical model, parameter β_i represents the difficulty of a certain question. So similar to the sensitivity of profile items, this parameter represents the property of a question itself, e.g. some questions are inherently harder while others are easier. But we argue that a direct reinterpretation of question difficulty in the IRT model to the sensitivity of the privacy score calculation model is not proper enough. And it is better to define sensitivity as marginal distribution of the response matrix. And with large number theory we use gaussian distribution, with parameters $(\overline{\beta_i}, \sigma_i^2)$, to represent the sensitivity of each profile item. Here $\overline{\beta_i}$ is the mean value of sensitivity for profile item i, which is an unbiased estimation of real sensitivity for profile item i. And σ_i^2 represents the confidence of the estimation of sensitivity, the larger its value, the less confident about the estimation. For complex scenarios when different groups of users make different settings for item i, we can use an attribute or a group of attributes such as religion and location to split sensitivity into different groups, but for this paper, we only consider a simple gaussian distribution for sensitivity calculation and we leave the complex scenario for later work.

Formal Definition

3.2 Redefining Visibility

In the original privacy score calculation model, visibility is interpreted with equation 1 which is calculated with equation 3. We argue that this definition does not hold w.r.t. the response matrix. Because the value R_{ij} from the response matrix are observed value, while probability are defined over a group of observations. And also, this interpretation does not take into consideration the structure of the user's social network as well the visibility contributed by services such as random user search and network traversal. So, we redefine visibility by taking into consideration the following factors: the social structure of user's network and the random search and traversal provided by the social network platform.

We observe that social networks such as facebook provider serveral levels of access control settings, private, friend, friend of friend and public. These settings can bring a network effect, which means the more friends the user has, the higher visibility. And the access control settings also have indirect impact on the visibility of this user.

Social networks such as Facebook can allow users to randomly search a person, this functionality also has effect on the visibility of user. But for simplicity, we denote this effect as a constant number r.

Formal Definition

3.3 Putting Everything together

4 Experiments

5 Conclusion and Future work.

References

[1] Justin Becker and Hao Chen. Measuring Privacy Risk in Online Social Networks.

[2] Tran Hong Ngoc, Isao Echizen, Kamiyama Komei, and Hiroshi Yoshiura. New approach to quantification of privacy on social network sites. In *Proceedings of the 2010 24th IEEE International Conference on Advanced Information Networking and Applications*, AINA '10, pages 556–564, Washington, DC, USA, 2010. IEEE Computer Society.

[3] 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.