Tracking topics on revision graphs of Wikipedia edit history

Lin GU Basilisa MVUNGI Mizuho IWAIHARA
Graduate School of Information, Production and Systems, Waseda University,
Fukuoka 808-0135, Japan

Abstract

Private information is sometimes unintentionally disclosed to the public in Social Network Services (SNSs), leading to great concern on users' privacy settings. In order to guide SNS users for their appropriate privacy settings, measures for evaluating privacy attitude of users have been proposed. However, the existing measure of privacy score, which utilizes the item response theory (IRT), only considers whether each privacy attribute is disclosed or not and does not consider the activity level of the attribute, which is a quantitative indicator of how much amount the user published on the attribute. In this paper, we first analyze the relationships between the original privacy score and user activity levels. After confirming the relationships between them, we propose activity privacy score using user activity levels. With the activity privacy score, we can group SNS users by their behavior and improve the original privacy score model.

Keywords
Wikipedia, delta summary, supergram

1. Introduction

In recent years, the number of social networking services (SNSs) such as Facebook, Twitter etc. has been increasing rapidly. People spend a great amount of time to keep in touch with their friends and relatives, and identify themselves among others. According to the survey of [1], nearly half of the people accessing to the Internet are members of one or more SNSs. In spite of the usage, people disclose private information, resulting in unforeseen future threats. Security and privacy mechanisms have been adopted in most of the SNSs [2], nevertheless, their mechanism is still inadequate to ensure the security of the users' private information.

As a measure to quantify SNSs risks, privacy score has been proposed by [3], which takes into account visibility (disclosure scope) and sensitivity (weight) of profile attributes. In spite of its efficiency and practical utility, activity factor of each attribute, such as the numbers of photos and friends, was not considered.

In this paper, we evaluate the influence of user activity on user profile privacy setting behavior. We begin by looking at linear and causative relationships between them. The results of the relationship will give us an insight of how to incorporate user activities into the existing privacy score model. We believe that user activity is important factor to consider because it indicates how much amount a user publishes his/her personal information to the public. Such information can identify a user which leads to a user to be subjective to threats. One interesting research question is in what situations and what types of users disclose their personal information, even though certain privacy risk exists.

We find that most of the user activity levels have influence on the user profile privacy. Among them, public photos have the highest influence. Also, there are activity attributes that have negative influence: nonpublic photos and groups. We propose new scores to measure user's activity level and group the users by these scores.

The rest of the paper is organized as follows. In Section 2, we survey related work. In

Section 3, we will discuss our data collection method. In Section 4, we will discuss our detailed analysis. In Section 5, we conclude our work and mention our future work.

2. Related Work

Liu [4] has found that the actual privacy settings in Facebook do not match the users' desired privacy settings expectation. In [4], Liu found that photos are the most sensitive private information in Facebook. Privacy recommendation is a measure of resolving inadequate privacy settings. Semantics-based privacy configuration system (SPAC) is designed to give recommendation to SNS users automatically [5]. The system diagnoses user's data, friends' features, user's configuration history and ontology, and outputs recommendation based on classification algorithms.

Liu [2] introduced privacy score for quantifying privacy risk of a user. This indicator can guide a user into proper privacy settings. Two factors were used for calculating privacy score: visibility and sensitivity. Visibility means the openness level of the attributes in user profiles. The more the open the private information is, the higher the privacy risk. Sensitivity gives weight to each attribute. The more sensitive private information user discloses, the higher the privacy risk.

Pergament has also proposed privacy score called Friends-Oriented Reputation Privacy Score (FORPS) [6]. This privacy score does not consider private information itself. It calculates the chances the users' friends will propagate their private information, which is ignored in most of existing research. In this paper, we have not put it into the privacy score model. We will try this factor in the future work.

3. Data Collection method

We made a Facebook ID generator and randomly collected 6993 users' data.

Facebook allows only attributes disclosed to the public to be collected. For the purpose of our study we collected the following data.

- **Profile attributes:** address, bio and favorite quotations, birthday, current work, email, interested in and looking for, location, religious and political views, relationship, websites, mobile phone
- **Gender** of each user.
- **User activity levels:** Friends: the number of friends user added, Public photos: the number of photos user disclosed to public, Nonpublic photos: the number of photos user does not disclose to public, Music, Movies, TV shows, Books, Games, Likes, Groups: the number of groups user join in.

Uses can opt whether or not these profile attributes and activity attributes are opened to the public. If a user chooses not to disclose his/her friend list, then we cannot know how many friends this user has. In this case, such an activity attribute should be treated as undisclosed. On the other hand, activity attributes like Movies and Games may be left blank by a user. In this case, the blank can be caused by either (1) the user's privacy setting, (2) the user does not enter data, or (3) the user simply does not have activities on the attribute.

We cannot distinguish these three causes of blank; we can only observe that a blank occurred but its cause is unknown.

Regarding visibility, Facebook and other SNSs provide five or more levels of visibility, for example: (1) only me, (2) specific friends, (3) friends, (4) friends of friends, and (5) public. But since visibility levels below public are not available, we focus on investigating patterns of public disclosure and their relationship to user activity levels.

4. Experimental Evaluation

In this section, we present several measures to evaluate the influence of user activities in privacy setting behavior.

4.1 Relationship between User Activity and Profile Privacy Setting

In order to analyze the relationship between user profile privacy and user activity level, we use score of profile attributes to compare with each user activity and try to find relationships or influence between them.

In [3], user profile privacy is aggregated into one parameter: profile privacy score. Also, profile openness score shows user profile privacy disclosure.

Profile openness score: The number of profile attributes open to everyone. This score only considers the visibility and gives an equal weight (sensitivity) to each attribute.

Profile privacy score: The privacy score defined in [3]. In this paper we call it profile privacy score. There are two kinds of privacy score in [3]: basic privacy score and IRT-Based privacy score. We adopt the IRT-based privacy score which is advanced. Privacy score is measured by the visibility and sensitivity of profile attributes. In IRT-based privacy score, visibility is mapped to the item discrimination in item response theory and sensitivity is mapped to item difficulty.

We used Pearson correlation analysis to observe the linear relationship between user activity levels and either profile privacy score or profile openness score. Here we transform each activity level by logarithm of base 10. The results are shown in Table 1.

From the correlation table, we find that the highest correlation value is about 0.3 which means slightly correlated. Pearson correlation was used as a basic step to observe whether there is linear relationship between user profile privacy and user activity. We observe weak relationship and therefore we use binary logistic regression analysis to observe non-linear causative relationship between the variables. This time the criterion variable is profile openness score and gender is added as a factor. The results are shown in Table 2.

Table 1. Pearson correlation between profile scores and user activity levels

Pearson Correlation		
	Profile privacy score	Profile openness score
Friends	.104**	.089**
Public photos	**.306****	**.315****
Nonpublic photos	*-.102***	*-.087***
Music	.275**	.288**
Movies	.109**	.119**
TV Shows	.079**	.092**
Books	.088**	.092**
Games	.036**	.044**
Likes	.030**	.046**
Groups	*-.086***	*-.080***
** p<0.01		

In table 2, Exp(B) is the indicator meaning that one unit increase of the attribute (such as log friends), probability of having profile privacy score higher than x (x equals to the threshold) open attributes raises by Exp(B) times. For example, being female reduces the chance of opening one attribute by 29.7 percent (1-0.703). Sig. shows the significance on the null hypothesis of the value Exp(B).

Tufekci [7] revealed that gender causes some of profile information to be disclosed. However our focus was on the trend of profile openness score which is more diverse. We found that regression analysis gives better evaluation of the influence of user activities on privacy setting behavior.

Our results reveal that public photos, number of friends or music interest disclosed causes more attributes in profile, while private photos or groups cause less attributes disclosed.

Table 2. Binary logistic regression between profile privacy score and user activity levels

Profile Privacy Score	≥1	≥2	≥3	≥4	≥5
Gender	.703 (.000/36.219)	.613 (.000/39.448)	.522 (.000/24.599)		
Friends	1.174 (.000/29.683)	1.164 (.000/16.918)	1.350 (.000/29.342)		
Public Photos	**1.979 (.000/382.319)**	**2.086 (.000/259.055)**	**2.173 (.000/105.972)**	**2.177 (.000/22.254)**	**2.652 (.002/9.859)**
Nonpublic Photos	*.634 (.000/103.234)*	*.610 (.000/63.330)*	*.680 (.000/14.183)*		
Music	1.900 (.000/100.299)	1.908 (.000/77.814)	1.589 (.000/17.973)		
Books		2.001 (.007/7.382)			
Likes			2.001 (.007/7.385)		
Groups	*.471 (.000/48.304)*	*.366 (.00039.787)*	*.384 (.000/12.329)*		
Numbers are exp(B) values and number in brackets are corresponding sig. and Wald’s statistics values					

Therefore, we summarize three main findings in our result. 1) Because activity levels are also influencing profile privacy, the attributes calculating current privacy score may be inadequate. The current privacy score can be improved by considering the activity attributes. 2) The privacy influence by each activity is also different. Some activities are highly influencing profile privacy, while some are slightly influencing. The influence of activity on profile privacy can be ranked by the Wald’s statistics value Public photos have the most significant influence on user’s private information because the Wald’s statistics value of public photos is the highest. However, because over half of users do not disclose their friend list, the number of friends needs to be analyzed without users disclosing friend lists. 3) There are activities negatively correlated with profile privacy. Attribute nonpublic photo is negatively related with profile privacy, because Facebook users who set their personal photos not visible to the public also tend not to open their profiles. We also find that the number of groups is difficult for users to hide in Facebook, it is indicating activity levels cannot be hidden by users. One possible explanation is that activities in groups are not public unless the groups are public, so joining a large number of groups indicates that the user is interested in sharing activities only with group members – a sign of hiding tendency, and it is consistent with less profile disclosure.

Negatively influencing attributes can also be incorporated into our privacy score model.

These attributes indicate users are proactively hiding certain private information, so we can measure how much degree a user has intention to hide activities. This feature is remarkable, since other activity attributes like Games are unable to be discerned whether the user has no activity or is hiding.

From the regression results, we see the need to include the weight on each activity attribute. For simplicity we include the sign and amount of weight based on our results. We suggest the weight of public photos to be the highest, followed by weight of number of friends and music. For determine weights, we need to redesign the algorithm of privacy score [3] that employs two-parameter logistic model, to reflect magnitude of each activity level.

4.2 Ratio of Photos

In Section 4.1, we showed that public photos have highest influence on profile privacy and nonpublic photos have negative influence. Therefore, we introduce the ratio of public photos to non-public photos as an indicator in analyzing user privacy. The definition of ratio of photos R is:

$$R = \frac{NumberofPublicPhotos}{NumberofNonpublicPhotos}$$

Table 3 shows the results of regression between ratio of photos and profile privacy score.

Table 3. Binary logistic regression between profile privacy score and user activity levels

Profile Privacy Score	≥1	≥2	≥3	≥4	≥5
Ratio of Photos	1.938 (.000)	2.073 (.000)	2.069 (.000)	2.105 (.000)	2.603 (.000)
Numbers are exp(B) values and number in brackets are corresponding sig. values					

The results reveal that as profile openness score increases, the odds of ratio of photos increases. This can be explained as users who disclose more profile attributes to the public tends to also disclose more photos to the public and hide few photos. We call this type of users as *extrovert user*.

4.3 Hobby Activity and Interaction Activity

In this section, we analyze the relationship on each pair of activity attributes. Pearson correlations are shown in Table 4.

From the correlation between activity attributes, we find that Music, Movies, TV shows, Books and Likes are highly correlated with each other. All these attributes are entered by users on their profile page. Games, Music, Movies, TV shows, and Books are related to hobbies. However, Games has a different feature such that it will be seen when a user plays an online game on Facebook. Number of groups shows only the number of groups the user joins. Games and Groups are interactive indicators similar to photos and friends.

Therefore, we classify Music, Movies, TV shows, and Books as **Hobby Activities** and other activity attributes as **Interaction Activities**. Interaction activities are relatively well correlated with profile privacy, while hobby activities are slightly correlated with profile privacy.

4.4 Extension to privacy score and new scores

From Sections 4.1 and 4.2, we have confirmed the relationship between user activity levels and profile privacy. Therefore, we can also use one indicator to represent the activity privacy of each user. We call this indicator **activity privacy score** and obtain this score by extending the current privacy score model.

Table 4. Pearson correlation between activities

	Friends	Public Photos	Nonpublic Photos	Music	Movies	TV Shows	Books	Likes	Games	Groups
Friends	1	-.017	-.052**	.012	-.023	.001	-.011	.048**	-.033**	-.047**
Public Photos		1	.181**	.322**	.141**	.119**	.084**	.082**	.102**	.204**
Nonpublic Photos			1	.113**	.181**	.209**	.130**	.226**	.103**	.381**
Music				1	**.509****	**.447****	**.339****	**.350****	.134**	.065**
Movies					1	**.638**	**.528**	**.481**	.256**	.192**
TV Shows						1	**.589**	**.642**	.336**	.291**
Books							1	**.503**	.289**	.233**
Likes								1	.344**	.342**
Games									1	.136
Groups										1

In the original privacy score model, disclosure levels of attributes are integers. User activities as collected as continuous values. To categorize user activity levels into dichotomous values, we set a threshold for each activity value. For simplicity, we set thresholds between zero and non-zero, and at the peak in the distribution of activity levels.

In Section 4.3, activities are separated into hobby activities and interaction activities. Scores for measuring user behavior on SNS can also be separated into **hobby activity score** and **interaction activity score**. We obtain these two scores by hobby activities and interaction activities individually and use them in Section 4.5.

4.5 High and Low Disclosure-Rate Profile Privacy Score and User Grouping

User profile attributes are listed in Section 3. The disclosure rate of each profile attribute is shown in Table 5. Here, only the disclosure rate of current work, interested in and looking for, location and relationship are higher than 15%. The other seven attributes have disclosure rates lower than 5%. We define high/low profile openness scores by checking whether the disclosure rate of the attribute is higher than 15% or lower.

Table 5. Disclosure rate of profile attributes

Profile attributes	**Disclosure rate**
current work	26.0%
location	24.6%
relationship	21.9%
interested in and looking for	19.6%
bio and favorite quotations	3.4%
birthday	3.3%
website	2.6%
religious and political views	2.3%
email	1.5%
address	0.5%
mobile phone	0.3%

High profile openness score: The number of high disclosure-rate profile attributes open to everyone. High disclosure-rate profile attributes are: current work, interested in and looking for, location and relationship.

Low profile openness score: The number of the other seven low disclosure-rate profile attributes open to everyone.

We find user activities are much higher correlated with high profile openness score than low one. Also, the correlation between profile openness score and activities are mostly caused by high disclosure rate attributes. The correlation is shown in Table 6. The relationship between activity indicators and high profile openness score are more obviously increasing or decreasing, compared with profile openness score. For example, as Figure 1 shows, ratio of photos is more clearly decreasing when high profile openness score increases.

Table 6. Pearson correlation between high/low profile openness score and activities

Pearson Correlation			
	High profile openness score	Low profile openness score	Profile openness score
Friends	.089**	.039**	.089**
Public photos	.299**	.176**	.315**
Nonpublic photos	-.099**	-.010	-.087**
Music	.276**	.150**	.288**
Movies	.110**	.071**	.119**
TV Shows	.089**	.045**	.092**
Books	.089**	.045**	.092**
Games	.050**	.006	.044**
Likes	.058**	-.005	.046**
Groups	-.099**	.007	-.080**
**.Correlation is significant at the 0.01 level			

Therefore, profile privacy score is suitable to separate into high disclosure rate profile privacy score and low profile privacy score.

Finally, we introduced hobby/interaction activity score and high/low disclosure rate profile privacy score.

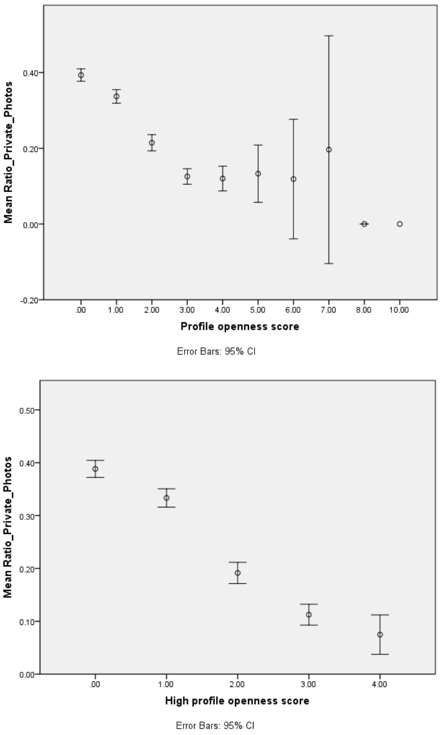

Figure 1. Distribution between ratio of photos and profile scores.

5. Conclusion and Future Work

In this paper we investigated the relationships and influence between user activities and profile privacy. We examined published data of 6993 Facebook users and extracted profile attributes and activity levels.

From the results, we find most activities are influencing user profile disclosure. Among them, public photos have the highest influence and should be given high weights when estimating the overall influence of activity attributes to profile privacy. Nonpublic photos and groups have negative correlations, indicating hiding intentions of the users. After separating user profile attributes by disclosure rates, and dividing activity attributes into hobby and interaction, we can define four distinct scores measuring user behavior on SNSs. In future, we plan to examine clustering of users by these scores.

In future work, we will improve the privacy score model which has not been finished yet;

the calculation of scores also need improve, which can group the user more accurately.

Reference

[1]. "Global Publics Embrace Social Networking, PewResearchCenter, 2010, " http://pewglobal.org/2010/12/15/global-publics-embrace-social-networking/.

[2]. H.X. Hu, G.J. Ahn, and J. Jorgensen, "Multiparty Access Control for Online Social Networks: Model and Mechanisms," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 7, pp. 1614-1627, July 2013, doi:10.1109/TKDE.2012.97

[3]. K. Liu and E. Terzi, "A framework for computing the privacy score of users in online social networks," Proc. Int. Conf. Data Mining, 2009.

[4]. Y.B. Liu, K.P. Gummadi, B. Krishnamurthy, and A. Mislove, "Analyzing Facebook Privacy Settings: User Expectations vs. Reality," IMC'11, 2011

[5]. Q.R. Li, J. Li, H. Wang, and A. Ginjala, "Semantics-enhanced Privacy Recommendation for Social Networking Sites," International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11, 2011

[6]. D. Pergament, A. Aghasaryan, J.G. Ganascia, and S.B. Brezetz, "FORPS: Friends-Oriented Reputation Privacy Score," Baabda, Lebanon, SeceS'11, June, 2011.

[7]. C. Tufekci, "Can You See Me Now? Audience and Disclosure Regulation in Online Social Network Sites," Journal of Bulletin of Science, Technology, and Society, vol. 28, no. 1, pp.20-36, 2008.