Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(4):481-486

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Privacy quantification of microblog users and correlation analysis of privacy value and user attribute

Xia Chen*, Tingjie Lu, Longfei Guo and Lei Bai

School of Economics and Management, Beijing University of Posts and Telecommunications, China

---

ABSTRACT

Along with the popularization and development of Microblog, the privacy of Microblog users has been a crucial problem. The study, based on the actual data of user's privacy settings, takes Sina Microblog as its object, qualifies the privacy value of each user and then divides the users in groups of different privacy values to conducts correlation analysis on user attributes. Through quantitative analysis, it is found that the users with more attention to the privacy of geographical information enjoy a larger circle, while the users with little attention tend to be more active, thus indicating that users' behaviors are greatly influenced by their concerns on privacy.

Keywords: social network, privacy, privacy quantification, user attributes, network user behavior, Microblog

---

INTRODUCTION

With the rapid development of the Internet, Sina Microblog has become a new platform for information dissemination and exchange in recent years. Nevertheless, along with the explosive information dissemination, a lot of hidden dangers emerge .Some undesirable privacy disclosure events make more and more users pay more attention to private information andits protection[1].

Several papers have expounded the development of privacy and relevant definitions of privacy[2,3].For measure of privacy, some scholars have already put forward relevant measurement models of privacy[3,4],but rare research has been done on privacy quantification. Some scholars has put forward the system of PaaS (privacy as a service) to quantify the privacy disclosed by users[5], but it fails to quantify users' privacy attributes.

For user privacy in Microblog environment, the research has recently proposed thatdifferent privacy settings shall be done for different groups of friends[6]. Therefore, the modeling quantification and grouping based on privacy attributes of users will be of great significance for analysis on user's behavior.

For the behavior analysis of Microblog users, the study, with the whole user group as the object, mainly focuses on analyzing their behavior characteristics and relation characteristics from the perspective of human behavior dynamics and statistics[7-9]. For the study on Microblog users not as an entirety, some scholars just, according to the grouping based on interest, study the communication rules within the groups[10].

However, no scholar has conducted quantification modeling on privacy attributes of Microblog user currently, nor combined privacy quantification to group Microblog users according to the privacy value and conduct correlation analysis on it, as well as user attributes and network behaviors.Therefore, the paper has some theoretical and remarkable practical significance for discussion of user behavior rule from the view of the social network privacy.Besides, through the customized web crawler software in the study, the actual data of user's privacy settingshas been adopted, and thus ensure the scientificity of the study.

The rest of the paper is organized as follows. In the second part, the data will be discussed. In the third part, based on the actual data of user's privacy settings, the weights of different privacy attributes will be measured, so as to calculate the privacy value of user and then put forward a new quantitative model for privacy. In the fourth part, the Microblog users are grouped according to the different privacy values, and then the correlation analysis on groups with different privacy values and the user's behavior attributes are conducted respectively.

1. Data Acquisition

Through web crawler tool, 32,368 public data of efficient SinaMicroblog users is acquired in the API open interface of SinaMicroblog.The acquiring process starts from a random user, and breadth-first traversal algorithm is applied to acquire friends who build mutual friend relationship with the user; the process is repeated again and again, so as to acquire four layers of user data with this relationship.

2. Modeling and Analysis

3.1 Privacy Quantification Study of Microblog Users

In this part, specific quantification study and analysis are conducted. In the data acquisition, three main privacy settings of Sina Microblog users are found: (1)AllowMsg(A_c)：Whether all other users are allowed to send me private messages;(2)AllowComment(A_m)：Whether strangers are allowed to comment；(3)AllowGeo(A_g)：Whether geographical location is allowed to be marked.The above three privacy settings of Microblog users belong to the bool-type variable, which means its value is whether 1 or 0, namely, allowing or not allowing.

And, a vector can be employed to represent the privacy value of various Sina Microblog users for the specific privacy quantification of the users.

P=(A_m, A_c, A_g)In the formula, vector P (privacy) stands for the protection value of privacy of every user, which is called privacy value in this study.

In order to study the factors of different privacy settings of users, entropy processing technology of decision-making analysis theory is applied to realize the correlation analysis of different privacy attributes and measurement of weights. The specific method is as follows: (1)Establish a decision matrix; take the three privacy attributes of user data as indicators of measurement, and establish a decision matrix$D = \{X_{mn}\}$;(2) Standardize the above decision matrix, and get the matrix$R = \{r_{ij}\}_{mn}$;(3) Calculate the output entropy of the standardized matrix,$E_j = -k * \sum P_{ij} \ln ij$, Therein,$P_{ij} = {r_{ij}}/{\sum_{i=1}^{m} r_{ij}}$ $k = (\ln m)^{-1}$;(4). Calculate the degree of deviation$d_j = 1 - E_j$;(5)Get the weight coefficient influencing the relationship$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j}$.

Through the entropy processing technology, the weight coefficients influenced by the three privacy settings related to user privacy are calculated respectively: 42.211% AllowMsg, 33.977% Allow Comment and 23.812% Allow Geo. Privacy vector adopts the quantification summation of the three related factors as its value:

$$P = 42.211\% * q(A_m) + 33.977\% * q(A_c) + 23.812\% * q(A_g)$$

Therein, $q(A_m)$, $q(A_c)$ and $q(A_g)$ respectively stands for the probabilities of AllowMsg, Allow Comment and Allow Geo, whose values are whether 0 or 1. And privacy concern degree of various users can be described through the privacy value of user.

The privacy values in eight points from left to right in the above figure respectively represent different privacy setting combinations: (1) point 0 indicates the three settings are not allowing, which belongs to the user type with highest privacy settings; (2) 0.23812 signifies that $A_g$is allowed; (3) 0.33977 shows that only $A_c$is allowed; (4) 0.42211 represents that onlyA_m is allowed; (5) 0.57789 stands for $A_c + A_g$are allowed, and only private message $A_m$is not allowed, which is marked as $-A_m$in the following contrastive analysis graphic; (6) 0.66023 stands for $A_m + A_g$ are allowed, and only comment A_c is not allowed, which is marked as $-A_c$in the following contrastive analysis graphic; (7) 0.76188 stands for $A_c + A_m$are allowed, and only geographic mark $A_g$is not allowed, which is marked as $-A_g$in the following contrastive analysis graphic; (8) 1.1327 represents all the three are allowed, and illustrates that this type of users have the lowest privacy settings.

As shown in Figure 1: (1) user privacy value is in a very high distribution value at the point of 0.57789, which

signifies that $A_c + A_g$ is allowed. It can be known that, comparatively, Microblog users don't care much about the comment on Microblog information and acquisition of geographic information; (2) the distribution proportions of the group with low privacy concern and the group with high privacy concern in the whole are lower that 5%, which means that most users show different degrees of concern on privacy protection.

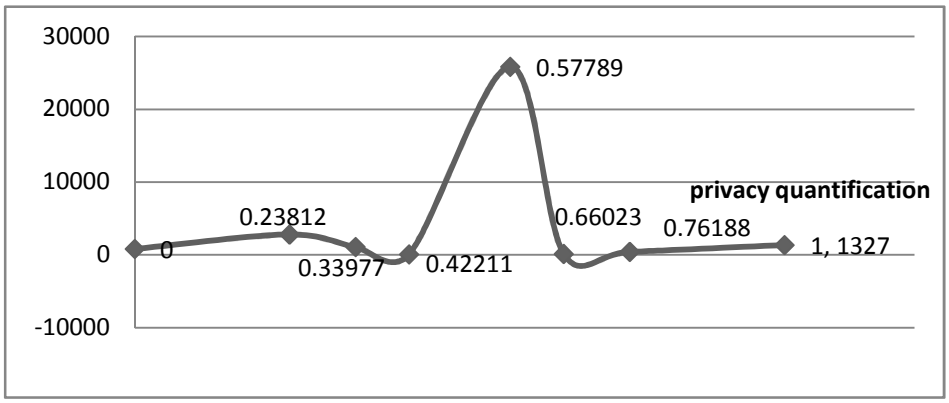

Fig.1. Distribution diagram of quantification of user privacy settings

3.2 Correlation Analysis of Privacy Value and User Correlative Attributes

(1)Analysis on user attribute of friends

Through user privacy quantification, it is available to analyze user's following, friends and followers of each kind of privacy attribute, which is considered to be capable of fully presenting the characteristics of user circles. Therefore, it can explain the relationship between characteristics of user's privacy attributes and scale of user circles to some extent.

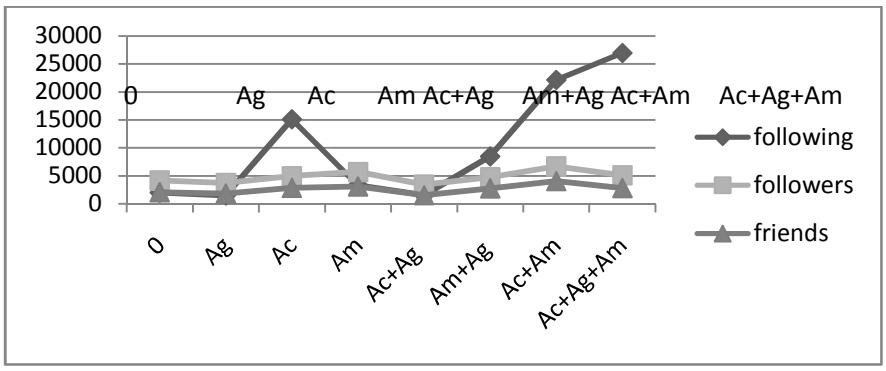

Fig. 2 Comparison Graph of Privacy Quantification of Commenting &Following, Friends and Followers

The vertical axis refers to the numbers of users of this privacy setting (unitary processing is done), the horizontal axis refers to the user quantity from small to large, and the combination of various privacy settings. In Figure 2, it is founded that:

1) The scale of their circle is closely relevant to the concern of privacy: (1) As for the typical users with high privacy concerns, their following, friends and followers are not the least. Such usershave a certain friend circle and communication group, which is relatively smaller, and likely to contain his/her closest and safest friends. (2)For the users with three allowable settings ($A_c + A_m + A_g$), their friend circles are obviously more than those with three unallowable settings, as well as the average friend circles, indicating that user circles with low concerns on privacy have a more frequent communication. (3) The users with only geographic indication allowable rank the second in privacy quantification distribution diagram, but their circle is very small, indicating that there are indeed many users like to use geological information service, but they have a strong sense of privacy protection. Meanwhile, the users who do not allow private messages just have a similarly small circle, but the reason is different. These users mostly do not want others to disturb their own living space and information space.

2) The variation trends of the curves of average following and average friends are nearly the same, which means that users make consistent perception and judgments of friends and following, but they are apparently different from that of average followers.

3) Apart from the same rules in the curves of average following and friends: (1) The user circles who allow geographic indication, comments and private messages are increasing (value area of 2, 3,4), as well as the user circles who allow two settings (value area of 5, 6, 7). It shows that judged by user's perception standard, the lower the privacy concern, the larger the circles will be, which is identical to practical experience. (2) The users who allow private messages and comments but not geographic indication enjoy the largest circle, indicating the communication function is the key factor to determine the scale of circle; at the same time, as long as geographic indication information is added, the scale of circle will be reduced, which also applies to the users with three allowable settings. From it, we can get that geographic information manifest the stronger privacy attribute compared with the other two settings. And privacy attribute is bound to influence the frequencies that users communicate with people.

4) It shows different rules for followers: (1)The users who allow comments have a significantly greater circle than the ones who allow geographic indication and private messages. It suggests that users actually expect to see their followers' comments, and meanwhile, for followers, they care more about their geographic indication and message service. (2) The user who allows three settings enjoys the largest circle, indicating that users who pay little attention to followers and have less awareness to privacy tend to have a larger circle.

(2)Analysis on user's attributes of favorites and rank

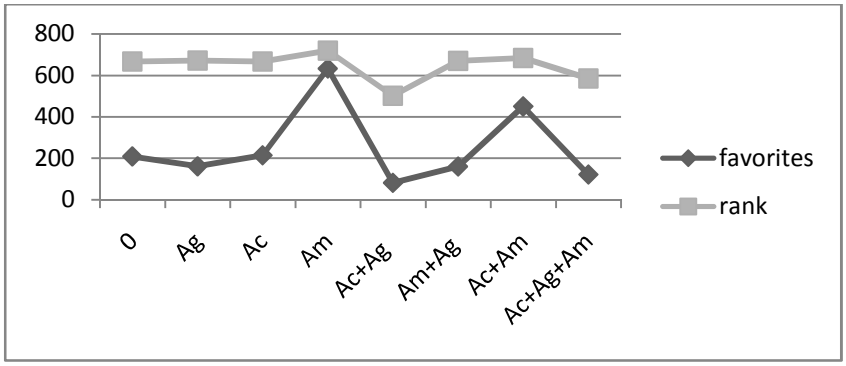

Fig. 3. Comparison Figure of Privacy Quantification of User's Favorites and Rank

We will, through favorite factors and user rank, do analysis and research on the correlation between them and user's privacy value.

1) In user rank distribution: (1) The group with highest rank generally only allows private messages. It is thus clear that the top Microblog users attach great importance to privacy protection. (2) The user group with the lowest rank is just contrary to the group with highest rank, and they may want to reveal their geographic location and understand other's evaluation to accustom themselves to a new circle or a new environment.

2) In user's favorite label distribution: (1) For users who only allows private messages, they accumulated far more labels than other users, which is consistent with the rules in rank distribution, further proving that the users with long-term use experience will put them in a fully protected environment, which is to say, they finally formed a safe and stable small circle. (2)The users who do not allow others to obtain geographic information are also a group accumulating many labels. Compare with the former group, the group just adds the settings of allowing comments. Relatively speaking, comments can be ignored but geographic informationwill be protected very well. (3) The result for the groups with the least and lesser accumulation of labels is identical to that in rank distribution. Respectively, they the users who expect to attract attention, or have no sufficient experience and consciousness in privacy settings.

(3)Analysis on user's attribute of geographic behavior

User's geographic information behavior reflects user's privacy information in the most direct way. The three kinds of user behavior are all related to user's geographic privacy information.

As shown in Figure 4,it is founded from privacy value that there is same distribution rules for numbers of Microblog with geographic information (LBS Microblog for short), pictures with geographic information (LBS pictures), and user's signing in.

(1) For the users whose geographic indication is not allowed (including 0, Ac, Am, -Ag), they basically refuse location-based service.
(2) For the users who allow others to obtain their geographic indication (including Ag, -Am, -Ac, Ac+Ag+Am), the

first group they send private messages to are those who allow geographic information extraction and private messages, and the second group are the ones who only allows the geographic information extraction. On one hand, it indicates that the users with great attention to privacy rarely use the geographic information. On the other hand, the two settings of AllowMsg and AllowGeo have the same influence on user's privacy perception.
(3) The users with low privacy concerns have more Microblog behavior related to geographic information than that of the ones with high privacy concerns.

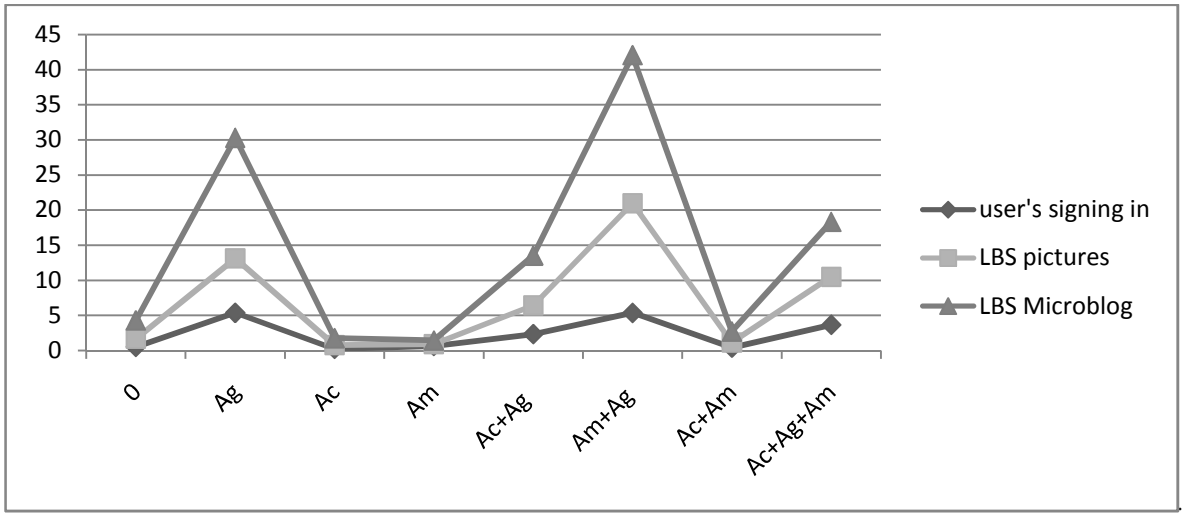

Fig. 4. Comparison &Analysis Figure of Privacy Quantification of Geographical Information Behavior

CONCLUSION

By privacy quantification, the apparent finding is that user's network behavior is obviously affected by privacy concerns in the current Microblog environment. Therefore, if Microblog service supplier could provide some efficient and reasonable personal information protection mechanisms to better protect users' privacy information, more users will join in to use and experience Microblog.

As for user's circles, friends and following mainly reflect user's judgment on the circle, from which, it is concluded that the circle of users who allows geographic information is smaller, indicating that most users pay great attention to geographic information. As the circle increases, the users who would like to make their strong privacy attribute information like geographic information known to others are decreasing.

Through rank and favorite label, it isfound that there is a clear relation between user's activeness and stickiness and privacy attributes. The excellent active users with high rank and most favorite labels do not publish their geographic information, but only allow private messages. Some of them allow comments. It is thus clear that, for users who have used Microblog for a long time, they attach great importance to privacy of geographic information. On the contrary, new users generally pay little attention to privacy.

Finally, the business behavior with geographic information reflects users' consistency in privacy concerns. That is to say, the users who pay great attention to LBS geographic information tend to rarely publish geographic information or use location-based service business.

Acknowledgments
The research is supported by National Basic Research Program of China (973 Program) (2012CB315805), Project of National Natural Science Foundation of China (71172135, 71231002), and the Fundamental Research Funds for the Central Universities(2013RC0603).

REFERENCES

[1] Hexun[OL]. http://tech.hexun.com/2012-03-03/138917028.html,**2012**.3.3.
[2] Smith H J, Milberg S J, Burke S.*MIS Quarterly*.**1996**, 20(2):167-196.
[3] Dinev T, Hart P. *International Journal of Electronic Commerce*. **2005**,10(2):7-29.
[4] Wang Bin,Duan Youxiang.*Computer Engineering and Applications*.**2011**,47(27).
[5] E. Michael Maximilien, Tyrone Grandison, Tony Sun, Dwayne Richardson, Sherry Guo, Kun Liu. P.*Workshop*

Program - W2SP **2009**,*Web 2.0 Security and Privacy*.**2009**.
[6] Subramanian S, March W.*Proceeding CHI 2010 Workshop on Microblogging*. **2010**: 54-60.
[7] Wang Xiaoguang.*Library And Information Service*. **2010**, 54(14):66-70.
[8] W Guan, H Gao, M Yang, Y Li, H Ma, W Qian, Z Cao, X Yang. *Physica A*. **2014**(395): 340-351.
[9] Qiang Yan, Lianren Wu, Lan Zheng. *Physica A*. **2013**(392):1712-1723.
[10] W Fan, K H Yeung, K Y Wong. *Physica A*. **2013**(392):1090-1099.