SindhUniv. Res. Jour. (Sci. Ser.) Vol.49(1) 125-128 (2017)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

User Profiling: A Privacy Issue in Online Public Network

S. ALI[++], A. RAUF*, N. ISLAM, H. FARMAN, S. KHAN**

Department of Computer Science, Islamia College Peshawar, Pakistan

Received 12th June 2016 and Revised 2nd September 2016

Abstract-The internet users, especially those who are using the Online Public Network (OPN) have concern about their private data. The OPNs are collecting user information from their profiles which are provided during the account creation and linking these attributes (for example age, gender and location) to the user explicitly posted data. In this way the OPNs collect and analyze the user's psychological and behavioral characteristics to create user profiles. These user profiles are used for advertisement purposes and can be provided to third party for financial profit, but these profiles have private information about users which could be used for malicious purposes. In this paper, a relationship hiding technique is proposed to protect the user from profiling. The relationship between the user account profile data and the user explicit posted data has been hidden using the data encryption to protect user privacy, but the encryption hides the information even from the legitimate users. In order to allow the legitimate users to see the shared contents, a collaborative sharing technique is proposed. The collaborative sharing technique is implemented and tested as a proof of concept which gives satisfactory results. The proposed and the state of the art sharing techniques are taking almost same processing time.

Keywords: Profiling, Privacy, Online Public Networks

1. INTRODUCTION

Privacy is one the hot issue for public media research community. Public media has great impact over the society, it gives information on different topics to the public media users, it brings people close together, the users can find old friends and make new friendships through Online public networks (OPNs), but it has the problem of personal privacy (Erlandsson *et. al.* 2012). There are different types of privacy issues in public media one of them is the user profiling. The American Heritage Dictionary defines the user profiling as "The recording and analysis of a person's psychological and behavioral characteristics, so as to assess or predict their capabilities in a certain sphere or to assist in identifying categories of people". User profiling has some advantages as it provides the recommendations of needed product to users. The user can accurately search the required information, but user profiling has security and privacy issues because it reflects the user itself. The user profiles are made through the user routine activities on internet or OSNs, these activities are pseudo form of user (Atote *et. al.* 2016). The user profiling has personal information about individual which are provided to third party for profit, these information are used for personal advertisement but it may use for malicious activities (Hassan *et. al.* 2013).

Profiling or categorization of persons can be done through different parameters, such as, gender, age and location, but the accurate profiling need some additional information along with these attributes (Baddelet. 2011; Pennacchiotti and Popescu. 2011). These attributes may be the activities of user on public media. The activities of user represent the behavioral characteristics, which can help the data collector or any third party for malicious profiling of user. It means some other data is also needed to link it with these attributes and accurately group the people. In this paper, the hiding of relationship between such data has been proposed to stop profiling of public media users.

Rest of the paper is organized as: Section 2 is about related work regarding user profiling, Proposed technique is discussed in section 3, Implementation is carried out in section 4 and at the end, paper is concluded.

2. RELATED WORK

While using the internet and in particular when using an online public networks (OPN), users must have expectation about their private data which can reveal him in future, and can be susceptible to identity theft. It is because, mostly data collectors (OPNs) are creating the users profiles for target advertisements and these information may be used for malicious purpose as well.

[++]Correspondence shaukat@icp.edu.pk, sahibkhan@uetpeshawar.edu.pk

*Department of Computer Science, University of Peshawar, Pakistan

**Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan

The personalized web search based on user profile is more accurate, effective and it is improving the quality of web based searches. In personalized search approach requires the user personal and behavior information to create the user profile, such type of data may gather through query history (Speretta and Gauch. 2005; Teevan *et. al.* 2005), browsing history (Sugiyama *et. al.* 2004), click stream (Dou *et. al.* 2007) and other user activities using online public network and web searches. The exposure of such information can expose the user privacy. Generalization is one technique that can be used for user privacy but it losing the useful information while applying the generalization technique (Xiao and Tao. 2006).

A lot of research has been done regarding the user privacy while using OSNs (Carminati *et. al.* 2007; Strater *et. al.* 2007; Beato *et. al. 2013*), they are appreciating OSNs for bringing people close together but at the same time they blame data collector about the users profiling which could lead to privacy leakages. Some of the researchers consider the current centralized OSNs as security and privacy risk for OSNs users and suggest a new decentralized OSNs for bringing people close together (Cutillo *et. al.* 2009; Jahid *et. al.* 2012), but has some disadvantages such as; users and data availability, the storage location etc.

3. PROPOSED PRIVACY TECHNIQUE

The data collectors or any malicious user can create user profiles through the user routine activities on the public network while combining these activities with the given information of user in his public profile for example age, gender, location etc. The user provides two types of data to public media server, one at the time of user account creation, in which the user gives basic information to public media server, for example, age, area of interest, location etc. The other type of data is provided through the routine activities by the user, while posting a message or tweets some explicit thoughts. The data collectors can use and link the data provided at the time of account creation and now the routine explicit activities to behaviorally and psychologically study the user and create his profile or group this user in some special group of interest. Moreover it leads to the violation of privacy because the data is not provided by users to the data collectors for classification or profiling. When data is collected for one purposes and used it for some other purposes such use of data can lead to user privacy violation. If the user posted information is used for user profiling that will be privacy violation, because it was not provided for user profiling.

In this paper, the encryption of explicit data of user during the routine activities are proposed in order to hide the relationship of user explicitly posted data and the data provided during creation of user account. The hiding of this relationship will ensure user protection from data collector while study or analyze him behaviorally and psychologically. Generally, cryptography based techniques are used for security and privacy of such data. In these types of security techniques, the reader cannot read the cipher text and thus cannot disseminate it further through OPNs. It is difficult to distribute the encrypted data among public media users, while distribution of information among the legitimate users is the main goal of public media. To distribute information among the legitimate users, the collaborative contents sharing is proposed in this paper.

In the proposed contents sharing scheme, whenever the data owner shares some content with his friends (tagged users) on public media and allow them to share it further collaboratively. The tagged users are those friends to whom data is directly shared by the data owner. The data owner secures its data using encryption technique from the public media server so that they cannot profile him using his publically posted contents. The collaboration is used to distribute contents among legitimate users. Collaboration means, if the number of total friends (tagged users) to whom data owner shares the contents directly are represented by n, then the collaborative users would be t for collaborative sharing for his contents, where t is user defined threshold and $t \leq n$. The data owner publishes those contents with tag friends but want to protect the contents from public media server or even friend of friends (or other friends) until the threshold $t \leq n$ of tagged users collaborate to share with any legitimate user. The term viewer is used here for friends of friends or any public user to whom the data owner didn't share the contents directly.

The tagged users can see the contents without collaborating with other tagged users, because the data owner has already shared the encryption key ($\mathcal{K}$) with them through a secure channel or the tagged users can also request key ($\mathcal{K}$) directly from the data owner. The contents are showed to tagged users directly because he will then decide whether to share it further or not. The tagged user will sent his secret share (S_i) for further collaboration with other users. Share is the part of secret key which could then be used to reconstruct the encryption key. The share will be sent to users in encrypted format through any secure channel. Figure 1

shows the data flow diagram for the proposed collaborative secret sharing model.

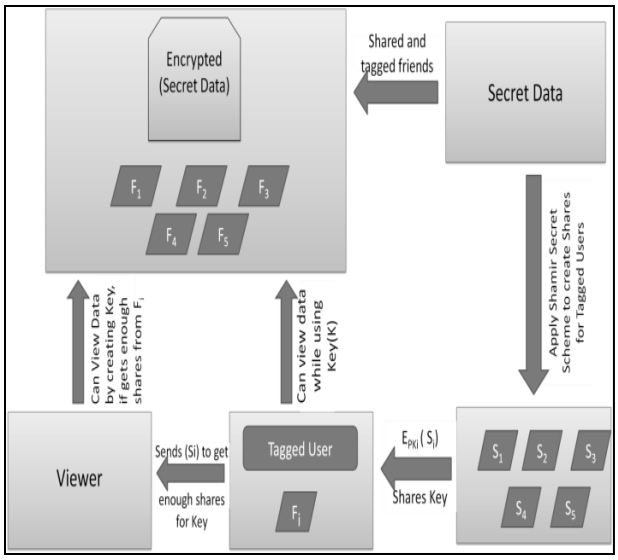

Fig. 1. Data flow diagram for the proposed collaborative contents sharing

The data owner takes secret key and will apply any secret sharing scheme for example, Shamir secret scheme (Shamir. 1979) to create n shares for all tagged users.

Suppose the encryption key $\mathcal{K}$ can be divided in n parts (secret shares) S_1, S_2, …S_n for each n tagged user in such a way that:

1. The collaboration of any t or more users can reconstruct the key $\mathcal{K}$.
2.The collaboration of $t - 1$ or fewer users cannot determine or guess the encryption key $\mathcal{K}$.

This sharing scheme is known as (t, n) sharing scheme, where n is the total number of tagged users and t $(t \leq n)$ is the minimum number of users to coordinate and reconstruct the encryption key $\mathcal{K}$.

Let the (t, n) model is used for secret sharing of encryption key, where $0 < t \leq n$, choose $t - 1$any positive integer $a_i = a_0 + a_1 + \ldots + a_{t-1}$ such that a_i is any random number and $a_0 = \mathcal{K}$. The following polynomial will be used for the sharing scheme (t, n).

$f(x) = a_0 + a_1x + a_0x^2 + \ldots + a_{t-1}x^{t-1}$ Where $a_0 = \mathcal{K}$

Suppose the function $(i, f(i))$ is used to construct n shares of key for every user $i = 1, 2, \ldots n$ using the polynomial.

Upon receiving the secret shares of key by every tagged users and want to reconstruct the encryption key $\mathcal{K}$ at receiving side by the collaboration of t users, the following Lagrange polynomial is used (Liu 1968).

$$f(x) = \sum_{i=0}^{t-1} y_i \cdot \prod_{\substack{0 \leq m \leq t-1 \\ m \neq i}} \frac{x - x_m}{x_i - x_m}$$

This equation creates a polynomial and the constant value produced in this Lagrange polynomial is the encryption key.

4. IMPLEMENTATION

The idea is implemented and tested just as proof of concept. It is implementable and it has very low overhead in terms of time, because all the encryption/decryption is performed on local computer system. It is also applicable to protect the data from unauthorized users, because the contents are encrypted and only the authorized user can see the contents to which the data owner shares the encryption key. Sample data were taken as shown in figure 2 and were encrypted with encryption key $\mathcal{K}$, the data is not in readable format and shared on public media. Since the data is encrypted and not readable for public media server, therefore, it cannot link this information to other data to create user profile. Similarly, the contents are protected from the unauthorized users, only those users can view the contents to whom data owner shared the encryption key $\mathcal{K}$.

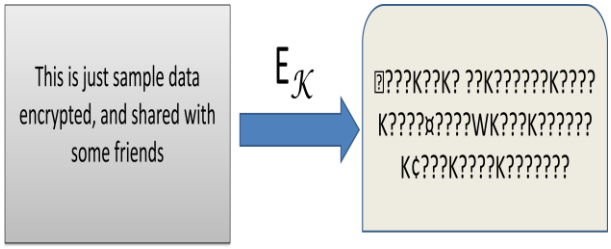

Fig. 2. The contents and its cipher text

5. CONCLUSION

The user profiling has advantages if the profiling is done correctly, but there are issues of privacy and security while doing profiling of users. In order to protect user privacy raised due to unwilling user profiling, a technique is proposed to hide the relationship of data which is needed for the creation of user profile in public media. The user explicitly shared data is encrypted to hide its relationship with the data provided at the time of account creation for protecting user privacy. The dissemination of data to the legitimate users, which is the main goal of public media providers, is assured through the collaborative contents sharing.

REFERENCES:

Atote, B., S. Zahoor, B. Dangra, and M. Bedekar. (2016). Personalization in user profiling: Privacy and security issues. In IEEE International Conference on Internet of Things and Applications (IOTA): 415-417. Pune, India.

Baddeley, M. (2011). A Behavioural Analysis of Online Privacy and Security.

Beato, F., I. Ion, S. Čapkun, B. Preneel, and M. Langheinrich. (2013). For some eyes only: protecting online information sharing. In proceedings of the third ACM conference on Data and application security and privacy, ACM: 1-12. Texas, USA.

Carminati, B., E. Ferrari, and A. Perego. (2007). Private relationships in social networks. In IEEE 23rd International Conference on Data Engineering Workshop : 163-171. Istanbul, Turkey.

Cutillo, L. A., R. Molva, T. Sturfe. (2009). Safebook: Feasibility of transitive cooperation for privacy on a decentralized network. In International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops, IEEE: 1-6. Kos Greece.

Dou, Z., R. Song, J. R. Wen. (2007). A large-scale evaluation and analysis of personalized search strategies. In Proceedings of the 16th international conference on World Wide Web, ACM: 281-590. Alberta Canada.

Erlandsson, F., M. Boldt, H. Johnson. (2012). Privacy Threats Related to User Profiling in Online Social Networks. International Conference on Privacy, Security, Risk and Trust (PASSAT) and International Conference on Social Computing (Social Com): 838-842. Boston, Massachusetts, USA.

Hasan, O., B. Habegger, L. Brunie, N. Bennani, and E. Damiani. (2013). A discussion of privacy challenges in user profiling with big data techniques: The excess use case. In IEEE International Congress on Big Data (BigData Congress): 25-30. California, USA.

Jahid, S., S. Nilizadeh, P. Mittal, N. Borisov, and A. Kapadia. (2012). DECENT: A decentralized architecture for enforcing privacy in online social networks. IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops): 326-332. Lugano, Switzerland.

Liu, C. L. (1968). Introduction to combinatorial mathematics, New York (NY) : McGraw-Hill.

Pennacchiotti, M. and A. M. Popescu. (2011). A Machine Learning Approach to Twitter User Classification. ICWSM, vol. 11(1): 281-288.

Shamir, A. (1979). How to share a secret. Communications of the ACM, vol. 22 (11): 612-613.

Speretta, M. and S. Gauch (2005). Personalized search based on user search histories. In Proceedings of International Conference on IEEE/WIC/ACM Web Intelligence: 622-628. France.

Strater, K. and H. Richter. (2007). Examining privacy and disclosure in a social networking community. In Proceedings of the 3rd symposium on usable privacy and security, ACM: 157-157. Pittsburgh, PA, USA

Sugiyama, K., K. Hatano, and M. Yoshikawa. (2004). Adaptive web search based on user profile constructed without any effort from users. In roceedings of the 13th international conference on World Wide Web, ACM: 675-684. New York, NY, USA.

Teevan, J., S. T. Dumais, and E. Horvitz. (2005). Personalizing search via automated analysis of interests and activities. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM: 449-456. Salvador, Brazil.

Xiao, X. and Y. Tao (2006). Personalized privacy preservation. In Proceedings of the ACM SIGMOD international conference on Management of data, ACM: 229-240. Chicago, Illinois, USA.