PRIVACY PRESERVED SOCIAL NETWORK WITH SECURITY

[1]ASWATHY RAJEEV K, [2]SUNDER R

[1]M.Tech, [2]Assistant Professor, Dept. of Information Technology, NCERC, Pampady-680597, Thrissur, Kerala
E-mail: aswathyrajeevk1991@gmail.com, sunder.mtech@yahoo.co.in

Abstract- One of the most important terms we hear on these days will be the social networks. Everyone will be a member of any of those social networking sites. These websites gain huge profit just by providing a platform for others to communicate. So far we have seen both merits and demerits of online social networking sites. They collect huge personal data and we take risk of trusting them. It is easy to collect these data by using learning algorithms to predict more private data. This paper explains the possibility of various inference attacks by these private data. These attacks can be minimized by sanitization methods put forwarded in this paper. This paper also comes with the security features which are essential for a online social networking site.

Keywords- Inference attacks, Privacy, Online social networks, Security.

I. INTRODUCTION

Online Social Networks gained large popularity in today's world. It becomes a medium for people to communicate, share in this busy world. Social Networks are the gathering of people who share some common things such as hobbies, likes, community, interests, religion, views etc. It is also a medium to a group of internet users who are willing to share their thoughts. It helps to unite families virtually. They provide great enthusiastic when loved ones are gathered in their easiest way.

There are more than 300 social networking sites all over the world. The features of all these sites are the ultimate social networking and they are all same to each other. These online social networking sites are ranked first among the visited websites than any other websites. There are both merits and demerits to these websites. Popular social networking sites include Facebook, LinkedIn, Twitter, Pinterest, Google Plus etc. All these collect enormous personal data from their users.Eventhough online social networking sites ensure the protection of data, there are many incidents regarding the private information drip of these social networking sites.

Many online social network (OSN) owners regularly publish data collected from their users' online activities to third parties such as sociologists or commercial companies. These third parties further mine the data and extract knowledge to serve their diverse purposes. In the process of publishing data to these third parties, network owners face a nontrivial challenge: how to preserve users' privacy while keeping the information useful to third parties. Failure to protect users' privacy may result in severely undermining the popularity of OSNs as well as restricting the amount of data that the OSN owners are willing to share with third parties.

Although these OSN provide various features to interact with the people they lack security features. The main advantage of these online social networks other than communication is the marketing and research fields.OSN now become a medium of major business center. Many companies can post ads on these websites and gain a huge profit. By using data mining algorithms on these networks able to analyze data and arriving on conclusions. These predictions and conclusion may violate others privacy. Social network is a graph consisting of nodes and links used to represent social relations on social network sites. The nodes represent the entities and links forms the relationships between them. Social network anonymization is adopted in order to provide privacy. Even though they guarantee privacy ensured, these networks can have various privacy issues. This paper discusses the various privacy measures to avoid prediction of online social network data and various security measures to provide better feasibility to users

II. SOCIAL NETWORKS

A. *Social Network Defined*

One of the important aspect of these social networking is that they can be used for research process and analyze various issues related to social networking.
For the purpose of understanding a social network can be considered as a graph where it contains vertices, edges and details.it is represented as

$$G= \{V, E, D\}$$

Detail type D, is a string defined over a $\sum$ that identifies a specific class. All detail types is denoted by H .A detail type is a pair consisting of detail type and detail value in which they are represented by a identifier J_k .It is the privilege of the user to input the details, it can be single valued or multi valued. A user can list various hobbies in his profile but can identify only one country. Another identifier is used for private details *I* and any detail is private if it belongs to H.

H= {favorite places,
favoriteactors,favouritefooditems,
religion}
I= {religion, political views}

Consider two nodes n_1 and n_2 be the two users and they are friends. Among the details described, two details are choosing to be private. The details chooses to be private is determined by the user's privilege.

These details meant to be private are considered as classification set *C*, in which they have separate values. That is for religious views various options can be assigned. Hindu, Christians, Muslims etc. The main aim of this paper is to identify the possible inference attacks and helpful sanitization techniques. The need to predict particular sensitive information for a particular user by analyzing the social network data.

For better predictability using of naïve Bayes classification algorithm is adopted. Naive Bayes algorithm is very useful in large scale data set.

B. Naïve Bayes classification Algorithm

Predicting a user's private details is a activity in graph classification. Suppose a user,ie; a node n with m details and p potential classification labels,$C_{1,}C_{2,}$……$C_{p,}$the probability of node n in class C,gives the Naïve Bayes classification algorithm.

$$arg \max_{1\leq x\leq p}[P(Cx^{i}|D_i^1,\ldots.D_m^i)]$$

Here arg max represents the possible private details class label s .But this is difficult to calculate since any value can be assigned to, so calculation can be done

$$arg \max_{1\leq x\leq p}\left[\frac{P(Cx^{i})\times P(D_i^1,\ldots D_m^i|C_x^i)}{P(D_i^1,\ldots D_i^m)}\right]$$

much more simplified form of this equation, considering that all details are independent to each other, then

$$arg \max_{1\leq x\leq p}\left[\frac{P(C_x^i)\times P(D_i^1|C_x^i)\ldots P(D_i^m|C_x^i)}{P(D_i^1,\ldots D_i^m)}\right]$$

Since the details are same for all classifications, only comparison of Is needed.

$$arg \max_{1\leq x\leq p}[\,P(C_x^i)\times P(D_i^1|C_x^i)\ldots P(D_i^m|C_x^i)]$$

C. Probability of Friendship Links

As Naïve Bayes classification algorithm can be used to predict the sensitive data attributes, while considering the case of friendship links, comparing a node with another node is done and training set is given these comparisons.

By considering the weightage of friendship links, there can be more than one type of relation between two friends. A relation of sister and brother to close friends. Based on the private details they share rather than public details. So weightage can be calculated using this equation,

$$W_{i,j}=\frac{|(D_i^1,\ldots\ldots,D_i^m)\cap(D_j^1,\ldots.D_j^m)|}{|D_i|}$$

The weight of friendship between two user's will be different.

III. SOCIAL NETWORK CLASSIFICATION

Social network data is a huge data set and it is very difficult to sort out these data by using normal classification algorithms. So in order to classify these huge dataset collective inference classification is used, which contains node details and links in the social graph. Each of these classifiers contain three modules a local classifier, a relational classifier and a collective inference algorithm.

Local classifiers are the first step in this classification algorithm. It examines the nodes and collect the details contained in it.This classifier builds a model based on the details of nodes,and apply this model to the training set to classify them.

Relational classifiers are separate type of learning algorithm, that identifies the link structure of the graph and uses the labels to build the model and apply to training set. In this paper ,all weights are considered as 1.

All these algorithms consider single entity only, local classifiers consider the details and relational classifiers links only. Real world data set cannot be classified with the relational classifiers. For the demerits of these classifiers, collective inference algorithm put up with these. By using local classifiers as first step, all nodes are identified. This is referred to as prior. Second step is that relational classifiers reclassify the nodes.It fully uses the graph. Collective inference controls the length of time the algorithm runs. Each step, algorithm calculates the probability.

By using NetKit, effectiveness of these classifiers can be calculated.NetKit-SRL, or NetKit for short, is an open-source Network Learning toolkit for statistical relational learning. It is written in Java 1.5 and was designed with a plug-and-play architecture to enable the mix-and-match between different components in the relational learning process. It integrates seamlessly with the weka machine learning toolkit, making it possible to use any of weka's learning classifiers in the context of relational learning.

IV. PRIVATE DETAILS HIDDEN

Privacy defined so far cannot protect from inference attacks and they are defined only for relational data only. A K-anonymization technique considers that an individual cannot be pointed out but privacy details can be identified. Differential privacy definition ensures that change in one person’s details cannot alter the whole content.

To implement a perfect privacy definition, there are two possibilities of an inference attack. If the attacker has some background knowledge, that is he is related to all the information with the user, no one can stop from the attack. Another possibility is that the attacker is known to only disclose network data then he can build a classifier to predict other sensitive information. In the first scenario privacy definitions are a failure.

The privacy definition for a social network is the protection of a user’s personal details which are sensitive, private even though an attacker has some background knowledge and released social network data. For this a best classifier algorithm which is developed from the released social network data and from the attacker’s background knowledge.

A relative privacy definition is developed due to the accuracy of classification algorithms that developed from the social network data.

V. PRIVACY IN SOCIAL NETWORK

D. Privacy Defined

The need to define a perfect privacy definition for the safety of user’s profile from the inference attack. Privacy is/was defined by each individual’s words. Various privacy definitions can be put forwarded. An attacker can formulate attacks by various possible methods. It can be passport number, voter id etc. The attacker can acquire background knowledge by various methods. We need to hide the sensitive data, decrease the classification accuracy and probability while preserving the essence of social networking.

Any set of classifiers, *C*, the classification accurateness of any random classifier *c’ € C* when trained on *K* and this is used to classify data set *G* to predict sensitive data $P_{c(K)}^{j}$.

$P_{c\ (G,K)}$ denote the prediction accuracy .another identifier is used to identify the additional accuracy of attacker. Possible attacks like predicting the death rate, predicting disease types by analyzing the revealed private details along with the public details.

E. Perturbation and Anonymization

For protecting against inference attacks on online social networks, we need to modify and remove certain parameters. The details residing on social network data can be deployed in three ways: adding details, modifying details and removing details from nodes. These methods are called Perturbation and Anonymization. By introducing noise into *D* to decrease classification can be considered as Perturbation. Removing nodes is considered as anonymization. Here consider two graphs which are sanitized versions of original graph. When artificial details are added to data set, the accuracy is minimized and sanitized versions of graphs cannot reveal the favorite activity. So classification accuracy is decreased. In short removing details to decrease the classification accuracy on sensitive attributes.

F. Details on the Social network

The main component of online social network is the details that are linked to each individual. We need to choose which details are to be removed. A social network data set *G* and list of sensitive details *I*, we need to determine which of the details should be removed which helps to decrease the classification accuracy.

Consider a user has the class value C_2 out of the set of classes of C and this person has the public details D_i

$$\underset{1\le x\le p}{arg\ \max}\left(P(C_x^i) \times P(D_i^1|C_x^i) \ldots P(D_i^m|C_x^i)\right.$$

We are removing th most correlated details with that of private. When we remove these details the building of a classifier to predict the sensitive details accuracy is decreased.

G. Links on Social Network

Another method to decrease the classification is the anonymization technique. That is the process of manipulating link information.Anonymization technique left with only two choices adding or removing links. We are considering the case of removing friendship links. Suppose a user belongs to two classes of friendship links and the true link is C_1.This technique helps to remove links in C_1 to decrease the classification accuracy. A node to be in class C_2 if the below equation is positive.

$$B(i)=p(C_2,N_i)-p(C_1,N_i)$$

We need to maximize the value of B(i) by removing links.

H. Detail Generalization Hierarchy and Detail Value Decomposition

Another important feature of detail anonymization is that the implementation of detail generalization hierarchy (DGH) is the process of generating a hierarchical ordering of the details within a given category.It is represented as tree structure.For example a user likes kathakali.we can replace it with art forms .we can also specify and replaces with traditional.

Another technique is that detail value decomposition in which a process of dividing a attribute into a series of representative tags. At each step we generalize each detail type.

VI. SECURITY CONCERNS FOR ONLINE SOCIAL NETWORK

A.Two-Factor Authentication

The two factor authentication consist of three steps verifying username and password, authenticate security token, User doesn't have a security token: logs user in right after he passes Step 1 only. Most of the online social networks fail to keep trust on the users. If a attacker uses a brute force attack and he can take the password of the user. Here comes the relevance of two factor authentication. Two-factor authentication is a security process in which the user provides two means of identification, one of which is typically a physical token, such as a card, and the other of which is typically something memorized, such as a security code. According to proponents, two-factor authentication could drastically reduce the incidence of online identity theft, phishing expeditions, and other online fraud, because the victim's password would no longer be enough to give a thief access to their information.

B. Blow Fish Cipher encryption

Blowfish is a fast block cipher, except when changing keys. Each new key requires pre-processing equivalent to encrypting about 4 kilobytes of text, which is very slow compared to other block ciphers. This prevents its use in certain applications, but is not a problem in others. In one application Blowfish's slow key changing is actually a benefit: the password-hashing method used in OpenBSD uses an algorithm derived from Blowfish that makes use of the slow key schedule; the idea is that the extra computational effort required gives protection against dictionary attacks. Blowfish has a memory footprint of just over 4 kilobytes of RAM.

This constraint is not a problem even for older desktop and laptop computers, though it does prevent use in the smallest embedded systems such as early smartcards. Blowfish was one of the first secure block ciphers not subject to any patents and therefore freely available for anyone to use. This benefit has contributed to its popularity in cryptographic software. Users uploaded images are encrypted with this algorithm and stored with the server. So pictures are securely stored in the server.

Blowfish has a 64-bit block size and a variable key length from 32 bits up to 448 bits. It is a 16-round Feistel cipher and uses large key-dependent S-boxes. In blowfish algorithm a 64-bit plaintext message is first divided into 32 bits. Each line represents 32 bits.

The algorithm keeps two sub key arrays: the 18-entry P-array and four 256-entry S-boxes. The S-boxes accept 8-bit input and produce 32-bit output. One entry of the P-array is used every round, and after the final round, each half of the data block is XOR ed with one of the two remaining unused P-entries.

C. Filter Unwanted Messages

This is the first proposal of a system to automatically filter unwanted messages from OSN user walls on the basis of both message content and the message creator relationships and characteristics.

D. GROUP priority

We assign priority to each group created by the user, so each group has unique priority with one another. This priority helps to view the posts and and allowed them to appear on the user walls.

E. Trusted Friends

We will be assigning our friends as trusted entity and this helps to recover the account whenever is necessary. Only these friends can help when the account is hacked.

CONCLUSION

Online social networking is one of the emerging trends in today’s world. Most of the people are involved in surfing the OSN.But most of the people are unaware of the attacks from social networks. People usually will publish their details in these sites and causes privacy issues.

Social networking also becomes the marketing media. When companies rely on social networking site, they have various intentions.Users need to take while publishing the details inside the social network. Security features of these sites should be improved in order to protect from various attacks.

REFERENCES

[1] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham, “Preventing Private Information Inference Attacks on Social Networks,” IEEE Trans. Knowledge And Data Engineering, vol. 25, no. 8, Aug 2013, pp.1849-1861.

[2] L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden

Patterns, and Structural Steganography,” Proc. 16th Int’l Conf. World Wide Web (WWW ’07), pp. 181-190, 2007

[3] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, “Anonymizing Social Networks,” Technical Report 07-19, Univ. of Massachusetts Amherst, 2007.

[4] J. He, W. Chu, and V. Liu, “Inferring Privacy Information from Social Networks,” Proc. Intelligence and Security Informatics, 2006.

[5] K. Liu and E. Terzi, “Towards Identity Anonymization on Graphs,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’08), pp. 93-106, 2008.

[6] S.A. Macskassy and F. Provost, “Classification in Networked Data: A Toolkit and a Univariate Case Study,” J. Machine Learning Research, vol. 8, pp. 935-983, 2007.

[7] C. Zhang, J. Sun, X. Zhu, and Y. Fang, Privacy and security for online social networks: Challenges and opportunities, *IEEE Netw.*, vol. 24, no. 4, pp. 13–18,Jul./Aug. 2010.

[8] L. A. Cutillo and R.Molva, ―Safebook: A privacy preserving online social network leveraging on real-life trust, *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 94–101, Dec. 2009.

[9] Nicole B Ellison, "Social network sites: Definition,history, and scholarship." *Journal of Computer-Mediated Communication*, Vol. 13(1), pp. 210-230, 2007.

[10] Ratan Dey, Cong Tang, Keith Ross And Nitesh Saxena(2009). “Estimating Age Privacy Leakage In Online Social Networks”

[11] E. Zheleva And L. Getoor(2008), “Preserving The Privacy Of Sensitive Relationships In Graph Data,” Proc. First Acm Sigkdd Int’l Conf.Privacy, Security, And Trust In Kdd, Pp. 153-171.

★★★