ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Preventing Private Information Leakage on Social mining

A.Krishna Kumar[1] , M.Suriya[2]

M.Tech, Department of IT,SNS College of Engineering,Coimbatore, Tamilnadu, India[1]

Asst.Professor ,Department of IT,SNS College of Engineering,Coimbatore, Tamilnadu, India[2]

Abstract : Online social networks, such as facebook are increasingly used by many users and these networks allow people to publish and share their data to their friends. The problem is user privacy information can be inferred via social relations. Hence managing those confidential information leakage is an challenging issue in social networks. It is possible to use learning methods on user released data to predict private information. Since our goal is to distribute social network data while preventing sensitive data disclosure, it can be achieved through sanitization techniques. Then the effectiveness of those techniques are explored and use methods of collective inference to discover sensitive attributes of the user profile data set. Hence sanitization methods can be used efficiently to decrease the accuracy of both local and relational classifiers and allow secure information sharing by maintaining user privacy.**Keywords**: social networking, social network privacy, sanitization, collective inference

I. INTRODUCTION

A social networking service is an online service, platform, site that focuses on facilitating the building of social networks or social relations among people share their interests, activities, backgrounds, or real-life connections. A social network service consists of a representation of each user, his/her social links, and a variety of additional services. Most social network services are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging. Online community services are sometimes considered as a social network service usually means an individual-centered service whereas online community services are group-centered.Social networking sites allow users to share ideas, activities, events, and interests within their individual networks.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

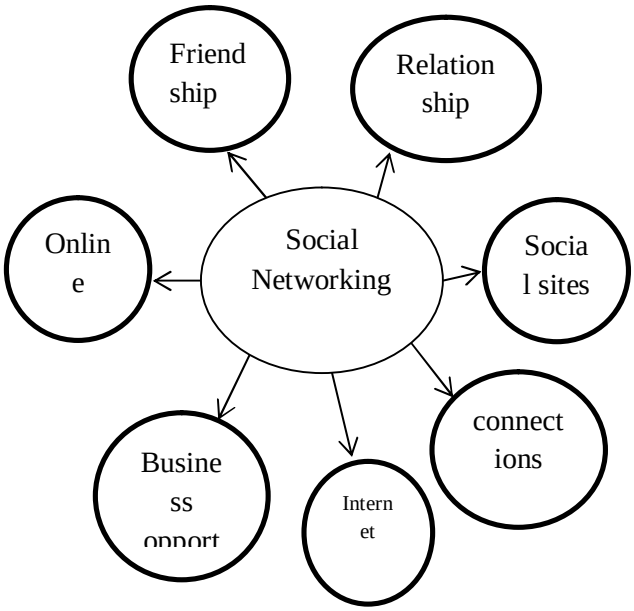

Fig . Social Networking

This means that although you are in the India, you could develop an online friendship with someone in United States. social networking often involves grouping specific individuals or organizations together. While there are a number of social networking websites that focus on particular interests, there are others that do not. The websites without a main focus are often referred to as traditional social networking websites and usually have open memberships. This means that anyone can become a member, no matter what their hobbies, beliefs, or views. However, once you are inside this online community, you can begin to create your own network of friends and eliminate members that do not share common interests or goals. Social networks are allow their users to connect by means of various link types. As part of their offerings, these networks allow people to list details about themselves that are relevant to the nature of the network.

Privacy concerns of individuals in a social network can be classified into two categories: Privacy after data release,Private information leakage. Instances of privacy after data release involve the identification of specific individuals in a data set subsequent to its release to the general public or to paying customers for a specific usage. Perhaps the most illustrative example of this type of privacy breach and significant number of "vanity" searches—searches on an individual's name, social security number, or address that could then be tied back to a specific individual. the problem of private information leakage for individuals as a direct result of their actions as being part of an online social network. Model an attack scenario as follows: Suppose Face book wishes to release data to electronic arts for their use in advertising games to interested people. However, once an electronic art has this data, they want to identify the political affiliation of users in their data for lobbying efforts. Because they would not only use the names of those individuals who explicitly list their affiliation, but also through inference could determine the affiliation of other users in their data, this would obviously be a privacy violation of hidden details. So explore how the online social network data could be used to predict some individual private detail that a user is not willing to disclose such as political or religious affiliation, sexual orientation and explore the effect of possible data sanitization approaches on preventing such private information leakage, while allowing the recipient

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

of the sanitized data to do inference on non-private details.In this work address two issues with respect to an inference attack. First, need to have some understanding of the potential prior information (i.e. background knowledge) the adversary can use to launch an inference attack. Second, need to analyze the potential success of inference attack given the adversary's background information.

II . RELATED WORK

In social network focus on the problem of private information leakage for individuals as a direct result of their actions .L.Backstrom et al [11] consider an attack against an anonymized network. In their model, the network consists of only nodes and edges. The goal of the attacker is simply to identify people. Further, their problem is very different than the one considered they ignore details and do not consider the effect of the existence of details on privacy. S.A.Macskassy et al[17] includes classification in network data based on a node-centric frame work in which classifiers comprise a Local classifier,a relational classifier and a collective inference procedure.
Lisegetoor [18] specifies thetraditional machine learning classification algorithms attempt to classify data organized as a collection of independent and identically distributed samples. Most real world data, on the other hand, is relational where di®erent samples are related to each other. Classification in the presence of such relationships requires that we exploit correlations present in them. Link-based classificationis the task of classifying samples using the relations or links present amongst them. Jianming He et al[13].specifies the causal relations among friends in social networks can be modeled by a Bayesian network, and personal attribute values can be inferred with high accuracy from close friends in the social network. so they propose schemes to protect private information by selectively hiding or falsifying information based on the characteristics of the social network. Ralph Gross et al[7] proposed patterns of information revelation in online social networks and their privacy implications. Patterns of Personal Information revelation is the information available depends on the purpose of the network.

III . METHODS

3.1 Local Classifiers

Local classifiers are a type of learning method that are applied in the initial step of collective inference. Typically, it is a classification technique that examines details of a node and constructs a classification scheme based on the details that it finds there. For instance, the naı¨ve Bayes classifier we discussed previously is a standard example of Bayes classification. This classifier builds a model based on the details of nodes in the training set. It then applies this model to nodes in the testing set to classify them.

3.2 Relational Classifiers

The relational classifier is a separate type of learning algorithm that looks at the link structure of the graph, and uses the labels of nodes in the training set to develop a model which it uses to classify the nodes in the test set. Specifically, in [18], Macskassy and Provost examine four relational classifiers: class-distribution relational neighbor (cdRN),weighted-vote relational neighbor (wvRN), network-only Bayes classifier (nBC), and network-only link-based classification(nLB).

3.3 Data Sanitization

Data Sanitization is the process of making sensitive information in non-production databases safe for wider visibility[11].Data sanitization techniques includes

Nulling out: Simply deleting a column of data by replacing it with NULL values is an effective way of ensuring that it is not inappropriately visible in test environments. Unfortunately it is also one of the least desirable options from a test database standpoint. Usually the test teams need to work on the data or at least a realistic approximation of it.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Masking data :Masking data means replacing certain fields with a Mask character (such as an X). This effectively disguises the data content while preserving the same formatting on front end screens and reports. For example, a column of credit card numbers might look like:
4346 6454 0020 5379
4493 9238 7315 5787
4297 8296 7496 8724
after the masking operation the information would appear as:
4346 XXXX XXXX 5379
4493 XXXX XXXX 5787
4297 XXXX XXXX 8724
The masking characters effectively remove much of the sensitive content from the record while still preserving the look and feel. Take care to ensure that enough of the data is masked to preserve security. It would not be hard to regenerate the original credit card number from a masking operation such as: 4297 8296 7496 87XX since the numbers are generated with a specific and well known checksum algorithm. Also care must be taken not to mask out potentially required information. A masking operation such as XXXX XXXXXXXX 5379 would strip the card issuer details from the credit card number.
Substitution : This technique consists of randomly replacing the contents of a column of data with information that looks similar but is completely unrelated to the real details. For example, the surnames in a customer database could be sanitized by replacing the real last names with surnames drawn from a largish random list. Substitution is very effective in terms of preserving the look and feel of the existing data. The downside is that a large store of substitutable information must be maintained for each column to be substituted. For example, to sanitize surnames by substitution, a list of random last names must be available. Then to sanitize telephone numbers, a list of phone numbers must be available. Frequently, the ability to generate known invalid data (phone numbers that will never work) is a nice-to-have feature. Substitution data can sometimes be very hard to find in large quantities. For example, if a million random street addresses are rrequired, then just obtaining the substitution data can be a major exercise in itself.
Shuffling records :Shuffling is similar to substitution except that the substitution data is derived from the column itself. Essentially the data in a column is randomly moved between rows until there is no longer any reasonable correlation with the remaining information in the row.

IV . CONCLUSION

In order to protect private information leakage using both friendship links and details together gives better predictability than details alone. By using collective inference the identifying simple classification nodes can be reduced. Then combine the results from the collective inference implications with the individual results, removing details and friendship links together is the best way to reduce classifier accuracy. However, it also show that by removing only details, and hide the images downloading process in social network and also filter the commenting text in wall and greatly reduce the accuracy of local and relational classifiers. In Future, the information that are to be shared can be verified for original and sensitive information before sharing it.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

REFERENCES

[1] Backstrom. L, Dwork. C, and Kleinberg. J, "Wherefore ArtThour3579x?:Anonymized Social Networks, Hidden Patterns, and Structural Steganography," Proc. 16th Int'l Conf. World Wide Web (WWW '07), pp. 181-190, 2007.
[2] Clifton. C, "Using Sample Size to Limit Exposure to Data Mining,"J. Computer Security, vol. 8, pp. 281-307, http://portal.acm.org/citation.cfm?id=371090.371092, Dec. 2000.
[3] Chawla. s, Mcsherry. F.Data privacy through optimal k-anonymization. In IEEE 21st International Conference on Data Engineering,April 2005.
[4] Dwork.C, "Differential Privacy," Automata, Languages and Programming, Bugliesi.M, Preneel.B, Sassone.V, and I. Wegener,eds., vol. 4052, pp. 1-12, Springer, 2006.
[5] Friedman .A and Schuster. A, "Data Mining with Differential Privacy," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 493-502, 2010.
[6] Fukunaga .K and Hummels. D.M, "Bayes Error Estimation Using Parzen and K-nn Procedures," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-9, no. 5, pp. 634-643, http:// portal.acm.org/citation.cfm?id=28809.28814, Sept. 1987.
[7] Gross .R, Acquisti. A, and Heinz. H, "Information Revelation and Privacy in Online Social Networks," Proc. ACM Workshop Privacy in the Electronic Soc. (WPES '05), pp. 71-80, http:// dx.doi.org/10.1145/1102199.1102214, 2005.
[8] Hay. M, Miklau.G, Jensen.D, Weis.P, and Srivastava .S, "Anonymizing Social Networks," Technical Report 07-19, Univ.of Massachusetts Amherst, 2007.
[9] http://en.wikipedia.org/wiki/social networks.
[10] http://money.cnn.com/technology/types of social networks.
[11] http://www.datamasker.com
[12] Heussne. K.M,"Gaydar" nFacebook: Can Your FriendsReveal Sexual Orientation?" ABC News, http://abcnews.go.com/Technology/gaydar-facebook, Sept. 2009.
[13] He. J, Chu.W, and Liu.V, "Inferring Privacy Information from Social Networks," Proc. Intelligence and Security Informatics,2006.
[14] Lindamood, Heatherly .R, Kantarcioglu .M, and Thuraisingham. B, "Inferring Private Information Using Social Network Data,"Proc. 18th Int'l Conf. World Wide Web (WWW), 2009.