ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Sanitations To Prevent Inference Attack On Social Network Data

Chandra D[1],Antony Rosewelt.L[2]

PG Student,ME CSE,Dhaanish Ahmed College of Engineering,Padappai,Chennai, India[1]

Assisstant Professor,DEPT OF CSE,Dhaanish Ahmed College of Engineering,Padappai,Chennai, India[2]

Abstract— Social Networks, like Facebook, are used by lots of people. These networks allow users to share specific of them and hook up with their friends. Many of the information revealed inside these network is designated to be private. Yet it's possible that corporations can use learning algorithms on released data to calculate undisclosed private data. This paper explore how you can launch inference attacks using released online community data to calculate undisclosed private information about individuals, including their political affiliation or sexual orientation. Then devise three possible sanitation techniques that may be found in various situations. Then, the potency of these methods by implementing them with a dataset from a certain nation-state in facebook online community application and looking to use ways of collective inference to learn sensitive features of the information set. And also this paper shows that have possibilities where the strength of both local and relational classification algorithm may be minimized with the sanitation methods as described.

Keywords— Social Networks, Inference Attack, Private Information, Sanitation Techniques

I. INTRODUCTION

Social networks are platforms that allow people to publish details about themselves and to connect to other members of the network through friendship links. Recently, the popularity of such on-line social networks is increasing significantly. For example, Face book now claims to have more than 110 million active users. The existence of on-line social networks that can be easily mined for various reasons creates both interesting opportunities and challenges. For example, social network data could be used for marketing products to the right customers. At the same time, privacy concerns can prevent such efforts in practice. Therefore, for future social network applications, privacy emerges as an important concern.

Social networking sites vary in the levels of privacy offered. For some social networking sites like Face book, providing real names and other personal information is encouraged by the site (onto a page known as a 'Profile'). This information usually consists of birth date, current address, and telephone number(s). Some sites also allow users to provide more information about themselves such as interests, hobbies, favorite books or films, and even relationship status. Where most people prefer to be anonymous. Thus, linking users to their real identity can sometimes be rather difficult. Nevertheless, individuals can sometimes be identified with face re-identification. Study has been done on two major social networking sites, and it is found that by overlapping 15% of the similar photographs, profile pictures with similar pictures over multiple sites can be matched to identify the users. It was revealed that 89% of the users gave genuine names, and 61% gave a photograph of themselves for easier identification. Majority of users also had not altered their privacy setting, allowed a large number of unknown users to have access to their personal information (the default setting originally allowed friends, friends of friends, and non friends of the same network to have full view of a user's profile).

As a way to protect privacy, we sanitize both trait (e.g., deleting good info coming from a user on-line profile) and link details (e.g., deleting links between friends) and explore the effects they've on combating possible inference attacks. Our initial results indicate which simply sanitizing trait information or link information might not be enough to prevent inference attacks and comprehensive sanitization techniques that entail both aspects should infer private information via friendship links by developing a Neural Network through the links within a social networking. An identical privacy problem for online social networks is discussed. Provide techniques which help when choosing the very best traits or links that ought to be removed for protecting privacy.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

In chapter six describes the experimental results. In Chapter seven the conclusion of the paper and suggests for the future improvements of the system.

II. RELATED WORK

The area of privacy inside a social network encompasses a large breadth, based on how privacy is defined. Consider an attack against an anonymized network. In their model, the network consists of only nodes and edges. Trait details are not included. The goal of the attacker is to simply identify people. Further, their problem is very different than the one considered in this paper because they ignore trait details and do not consider the effect of the existence of trait details on privacy. In several ways of anonymzing social networks. However, our work focuses on inferring details from nodes in the network, not individually identifying individuals. Other papers have tried to infer private information inside social networks. Consider ways to infer private information via friendship links by creating a Bayesian Network from the links inside a social network. While they crawl a real social network, Live journal, they use hypothetical attributes to analyze their learning algorithm. Provide techniques that can help with choosing the most effective traits or links that need to be removed for protecting privacy. Finally, we explore the effect of collective inference techniques in possible inference attacks.

Attempt to predict the private attributes of users in four real-world data sets: Facebook, Flickr Dogster, and Isonomy. They do not attempt to actually anonymize or sanitize any graph data. Instead, their focus is on how specific types of data, namely, that of declared and inferred group membership, may be used as a way to boost local and relational classification accuracy. Their defined method of group-based (as opposed to details-based or link-based) classification is an inherent part of our details-based classification, as we treat the group membership data as another detail, as we do favorite books or movies.

Get or work provides a substantial motivation for the need of the solution proposed in our work. Presenting a framework for analyzing privacy and anonymity in social networks and develop a new re-identification algorithm targeting anonymized social network graphs. To demonstrate its effectiveness on real world networks, we show that a third of the users who can be verified to have accounts on both Twitter, a popular micro blogging service, and Flickr, an online photo-sharing site, can be re-identified in the anonymous Twitter graph with only a 12% error rate.

The attributes attached to nodes, such as the user's interests, are usually far more sensitive. Social Security numbers can be predicted from Facebook profiles with higher accuracy than random guessing for other privacy breaches based on profile data. Even implicit attributes such as node degree can be highly sensitive, *e.g.*, in a sexual network.

In propose a method of measuring the amount of information that a user reveals to the outside world and which automatically determines which information (on a per-user basis) should be removed to increase the privacy of an individual.

Finally, preliminary work on the effectiveness of our Details, Links, and Average classifiers and examine their effectiveness after removing some details from the graph. Here, we expand further by evaluating their effectiveness after removing details and links.

III. PROPOSED SYSTEM

Implementing inference attack using Face book API. In order to access the profiles using face book API, one has to register on face book developer website. So it will register our profile and get token in our application and try to access our friends profile information, while creating access token and specify the information like to access and see, like status, date of birth, friends count etc. And store this information in database and use that information to feed our neural network classifier, again this will classify the data based on the constraints and provide us some data which is use to predict a person behavior or character. Access token which used to access the token to access the profile information, pass the access.

IV. PROBLEMS

The social network platforms such as Facebook allow third-party apps on their platform. Those third-party apps can access user profile data and can gather lots of information related to user. Although they promise to abide by the rules of the platform, it is possible that they may abuse the data.

First, 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.

Second, Private information leakage, is related to details about an individual that are not explicitly stated, but, rather, are inferred through other details released and or relationships to individuals who may express that detail.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Third, Inference attack is carried out by using crawler API, while information here is limited and not current in some aspects. And face book updated their API to their own way and provide more information for developers in Face book API.

Forth, character analyses is carried on two criteria, sexual relationship and political afflictions the person mentions in their profile and predict their character based on that information.

V. IMPLEMENTATION

Neural Network is a powerful tool used in modern intelligent systems. Many applications that involve pattern recognition, feature mapping, clustering, classification and etc. use Neural Networks as an essential component. Many types of neural networks have been developed. Back Error Propagation, Kohonen feature map and Hopfield network are some of basic networks that have been developed and are used in many applications. The most common technique used to train neural networks is called back-propagation. This technique is the subject of a large number of research articles—so many, in fact, that if you're new to the field of neural network classification, you could easily be led to believe that back-propagation is the only technique used for training. Estimating the best set of weights for a neural network is a numeric minimization problem. Two common alternatives to using back-propagation are using a real-valued genetic algorithm (also called an evolutionary optimization algorithm), and using particle swarm optimization. Each estimation technique has strengths and weaknesses.

5.1 USING SVM

SVMs as a classification technique. To obtain these results, use the SVM classifier packaged in WEKA, after representing details as a bit string. Here that when removing no details, the classification accuracy of the SVM has a classification accuracy between our links Only an Average/Details Only classifiers, with the exception of sets where the graph has a large percentage of unknowns (80 and 90 percent of the graph is unknown) where the SVMclassifier can actually output form the Details Only/Average classifier.

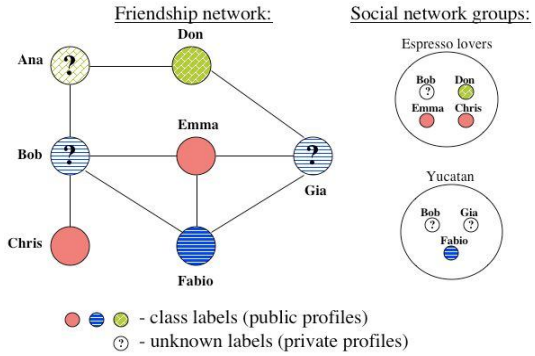

Fig.1 Network Group Members Classification

However, once we remove details (see Figs. 2d and 3d), the classification accuracy of the SVM drops much further than the Average/Details Only classifier, and even performs worse than the Links Only classification method.

Next, examine the effects of removing the links. Remove K links from each node, where K 2 ½0; 10, and again partition the nodes into a test set and training set of equal size. We then test the accuracy of the local classifier on this test set. Repeat this five times and then take the average of each accuracy for the overall accuracy of each classifier after K links are removed. For K 2 ½1; 6, each link removal steadily decreases the accuracy of the classifier. Removing the seventh link has no noticeable effect, and subsequent removals only slightly increase the accuracy of the links only classifier. Also, due to space limitations, for the remainder of experiments we show only the results of the average classifier, as it is generally the best of the three classifiers. Again examine the performance of the SVM, see similar results to what was seen with details only and average. Since the SVM does not include the link structure in its classification, there is no real affect from removing links on this classification method. However, show that by applying our technique, we routinely restrict classification accuracy to some arbitrary value below 95 percent. This means that graph is effectively ð_; C; G; KÞ-private because an attacker would be forced to use only K to determine classification labels.

VI. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

A.Experimental Results

Wrote a program to crawl the Facebook network to gather data for our research. Because of the size of Facebook's social network, we limited crawling to profiles inside the Dallas/ForthWorth (DFW) network.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

This means that if two people share a common friend that is outside the DFW network, this is not reflected inside the database. Also, some people have enabled privacy restrictions on their profile and prevented the crawler from seeing their profile details. Our total crawl resulted in over 167,000 profiles, almost 4.5 million profile details, and over 3 million friendship links. All but 22 of the people crawled were inside one, large component of diameter 16. For our experiments, consider only the subset of the graph for which we know the expressed political affiliation as either "Conservative" or "Liberal". This reduces our overall set size from approximately 160,000 to approximately 35,000 nodes.

To compare our methods to a traditional Naive Bayes classifier, we implemented our own version of a traditional Naive Bayes classifier. Then, we use the ideas discussed in to create a list of the most representative traits in the graph, which we use to remove the 10 most predictive traits from the graph. That is, when we say that we remove K traits, we calculate which K traits are globally the most likely to reveal your true political affiliation and then remove those traits from every node that originally had them. Similarly, we use the ideas discussed in to remove the 10 most telling links from every node in the graph. Unlike removing traits, which is done globally, removal of links is done locally. Finally, we combine the two methods and generate test sets with both 10 traits and 10 links removed from the graph.

Refer to these sets as 0t, 0l; 10t, 0l; 0t, 10l; 10t, 10l removed, respectively. Following this, we randomly divide our nodes to form sets of 50% of the nodes in the training and 50% in the test sets. We repeated the previous process five times, and run each experiment independently. We then take the average of each of these five runs as the overall accuracy. Our results, as shown in Table 1, indicate that the Average algorithm substantially outperformed traditional NaiveBayes and the Links algorithm. Additionally, the Average algorithm generally performed better than the Details Only algorithm with the exception of the (0 traits, 10 links) experiments. An examination of the Links results for that experiment shows that the drop in Average accuracy can be accounted for by the exceptionally low performance of the Links classifier and the consistent Details Only performance for that point.

Classifier	0t, 01	0t, 101	10t, 01	10t, 101
Neural Network	0.7533	0.7157	0.6838	0.6790
Links Only	0.7942	0.7942	0.7003	0.7003
Details Only	0.7163	0.5855	0.6977	0.6066
Average	0.7970	0.7799	0.7184	0.7069

Table 1 Performance of Classifier in links

B. Performance Evaluation and Analysis

The Details classification accuracy only decreased when we removed traits from nodes, and the (0t, *) accuracies are approximately equivalent. Similarly, the Links accuracies were mostly affected by the removal of links between nodes, and the (*, 0l) points of interest are approximately equal. The difference of in accuracy between (0t, 0l) and (10t, 0l) can be accounted for by the weighting portion of the Links calculations, that depend on the similarity between two nodes. Next, examine the specific affects of removing traits. We first test the local classification accuracies after removing K traits, where K 2 [0, 10]. After removing the K traits, we randomize our collection of nodes and create a test set of 50% of the nodes in the training and test sets. We then test the accuracy of the local classifier on this test set. We repeat 5 times and average the results for the overall accuracy for K, at each classifier. The results of this are shown in Figure 1. As evident from the results, after removing one trait, the classification accuracy immediately decreases significantly. The classification accuracy immediately decreases significantly.

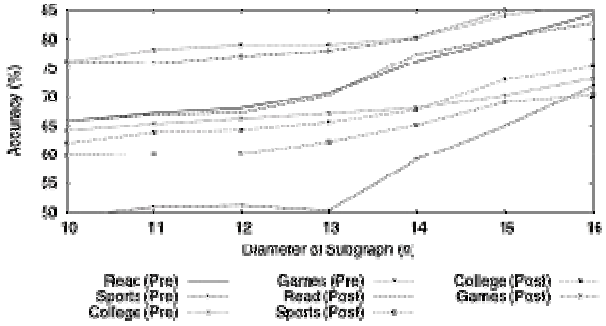

Fig .1 Classification of the Graph traits

After removing an additional trait, the classification returns to its prior accuracy, and for each subsequent trait removed we see a slight downward trend in classification accuracies. The sudden downward spike can be easily explained looking at the trait removal lists. The highest-ranked trait is evidence for the trait value of "Liberal". Removing this trait makes the probability of being"Conservative"outweigh the probability of a trait being "Liberal". This is why the Details accuracy is approximately the same as merely guessing the majority class for each node. However, when we remove the second trait, which is representative of being "Conservative" the probabilities again balance. None of the remaining traits are as highly indicative as the initial

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

two, so we instead see a gradual decrease in the accuracy over the tested parameters. Unsurprisingly, the Links Only classifier is only slightly affected by the removal of traits. We report additional experimental results that show the impact of link removal, collective inference and varying labeled vs unlabeled nodes ratios.

VII. CONCLUSION

Inference attack is an emerging engineering field which will be widely used by many companies and other organizations to predict a person behavior, since this current age, people use internet so widely and spent more time on social networks to share their mood, or get in touch with their family. Inference attack can provide some decent information on how to analyze his/her behavior. This information can be widely used by HR person or managers to analyze a person before committing them into their company. Inference attack can be used in agency can easily use inference attack and store some vital character analysis and store it private and head hunters can use this private information to filter candidates. This can also be used in matrimonial sites to fine tune searches. Recruitment websites, since now a day's companies started using open login system, a recruitment System.

REFERENCES

[1] A. Friedman and A. Schuster, "Data Mining with Differential Privacy," Proc. 16th ACM SIGKDD Int'l Conf. *Knowledge Discovery and Data Mining, pp. 493-502, 2010.*

[2]N.Talukder, M.Ouzzani, A.K. Elmagarmid, H. Elmeleegy, and.Yakut,"Privometer: Privacy Protection in Social Networks,"*Proc. IEEE 26th Int'l Conf. Data Eng*. Workshops (ICDE '10), pp. 266-269, 2010.

[3] A. Menon and C. Elkan, "Predicting Labels for Dyadic Data, "*Data Mining and Knowledge Discovery*, vol. 21, pp. 327-343,2010.

[4] K.M. Heussner, "'Gaydar' n Facebook: Can Your Friends Reveal Sexual Orientation?"ABCNews,http://abcnews.go.com/Technology/gaydarfacebookfriends/story?id=8633224#.UZ939UqheOs, Sept. 2009.

[5] C. Johnson, "Project Gaydar," The Boston Globe, Sept. 2009.

[6] 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.

[7] 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.

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

[9] J. He, W. Chu, and V. Liu, "Inferring Privacy Information from Social Networks,"*Proc.Intelligence and Security Informatics*,2006.

[10] T. Zeller, "AOL Executive Quits After Posting of Search Data," The New York Times, no. 22, http://www.nytimes.com/2006/08/22/technology/22iht-aol.2558731.html?pagewanted=all&_r=0, Aug. 2006.