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

Privacy Preservation for User Profiles In Social Networks

Ms. U.P. Umasree[1], Mr. V. Bhaskar, ME, [2]

II ME (CSE), K.L.N. College of Engineering, Madurai, India[1]

Associate Professor (CSE Department), K.L.N. College of Engineering, Madurai, India[2]

Abstract: A social network describes entities and connections between them. The entities are often individuals; they are connected by personal relationships, interactions, or flows of information. Social network analysis is concerned with uncovering patterns in the connections between entities. It has been widely applied to organizational networks to classify the influence or popularity of individuals and to detect collusion and fraud. Social network analysis can also be applied to study disease transmission in communities, the functioning of computer networks, and emergent behavior of physical and biological systems. Class learning algorithms are used on released data to predict private information. Inference attacks are initiated using released social networking data to predict private information. Collective inferences are used to discover sensitive attributes of the data set. Social network data classification is carried out with the combination of node details and connecting links in the social graph. Navie bayes classification algorithm is tuned to classify friendship links in a network. Local classifier, a relational classifier, and a collective inference algorithm are the three components used in the social network analysis. Local classifiers are a type of learning method that are applied in the initial step of collective inference. The relational classifier analyzes the link structure and labels of the node to identify a model for classification. Collective inference algorithm is used to increase the classification accuracy from the local and relational data values. The privacy preservation model is designed to protect sensitive attribute in user profiles from social networks. The user can select attribute for hiding process. Local classification and relational classification are applied to estimate anonymity levels. The system manages the key node alter and remove operation.

I. INTRODUCTION

With the increasing popularity of Social Network Services (SNS), more and more online societies such as Friendster, Livejournal, Blogger and Orkurt have emerged. Unlike traditional personal homepages, people in these societies publish not only their personal attributes, but also their relationships with friends. As social networks grow rapidly, many interesting research topics arise. Unfortunately, among these topics, privacy has not been fully addressed yet. Given the huge amount of personal data and social relations available in online social networks, it is foreseeable that privacy may be compromised if people are not careful in releasing their personal information. Information privacy has become one of the most urgent research issues in building next-generation information systems. A great deal of research effort has been devoted to protecting people's privacy. Aside from recent developments in cryptography and security protocols that provide secure data transfer capabilities, there has been work on enforcing industry standards and government policies to grant individuals control over their own privacy.

These existing techniques and policies aim to effectively block *direct* disclosure of sensitive personal information. However, to the best of our knowledge, none of the existing techniques handle *indirect* disclosure which can often be achieved by intelligently combining pieces of seemingly innocuous or unrelated information. Specifically, in scenarios like social networks, we realize that individuals connected in social networks often share common attributes. For instance, in a dance club, people come together due to their common interest; in an office, people connect to each other because of similar professions. Therefore, it is possible that one may be able to infer someone's attribute from the attributes of his/her friends. In such cases, privacy is indirectly disclosed by their social relations rather than from the owner directly. We want to analyze under what conditions and to what extent privacy might be disclosed by social relations. In order to perform privacy inference, we propose an approach to map Bayesian networks to social networks. We discuss prior probability, influence strength and society openness which might affect the inference, and conduct extensive experiments on a real online social network structure.

II. RELATED WORK

In this paper, we touch on many areas of research that have been heavily studied. The area of privacy inside a social network encompasses a large breadth, based on how privacy is defined. Backstrom et al. consider an attack against an anonymized network. Other papers have tried to infer private information inside social networks. In [8], He et al. 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, LiveJournal, they use hypothetical attributes to analyze their learning algorithm. Also, we provide techniques that can help with choosing the most effective details or links that need to be removed for protecting privacy. Finally, we explore the effect of collective inference techniques in possible inference attacks. In [9], Zheleva and Getoor propose several methods of social graph anonymization, focusing mainly on the idea that by anonymizing both the nodes in the group and the link structure, that one thereby anonymizes the graph as a whole. However, their methods all focus on anonymity in the structure itself. For example, through the use of k anonymity or t-closeness, depending on the quasi-identifiers which are chosen, much of the uniqueness in the data may be lost. Through our method of anonymity preservation, we maintain the full uniqueness in each node, which allows more information in the data post release. Gross et al. examine specific usage instances at Carnegie Mellon. They also note potential attacks, such as node reidentification or stalking, that easily accessible data on Facebook could assist with.

They further note that while privacy controls may exist on the user's end of the social networking site, many individuals do not take advantage of this tool. This finding coincides very well with the amount of data that we were able to crawl using a very simple crawler on a Facebook network. We extend on their work by experimentally examining the accuracy of some types of the demographic reidentification that they propose before and after sanitization. The Facebook platform's data has been considered in some other research as well. Jones and Soltren crawl Facebook's data and analyze usage trends among Facebook users, employing both profile postings and survey information. However, their paper focuses mostly on faults inside the Facebook platform. They do not discuss attempting to learn unrevealed details of Facebook users, and do no analysis of the details of Facebook users. Their crawl consisted of around 70,000 Facebook accounts. The area of link-based classification is well studied. In [6], Sen and Getoor compare various methods of link-based classification including loopy belief propagation, mean field relaxation labeling, and iterative classification. However, their comparisons do not consider present an alternative classification method where they build on Markov networks. However, none of these papers consider ways to combat their classification methods.

In [10], Menon and Elkan use dyadic data methods to predict class labels. We show later that while we do not examine the effects of this type of analysis, the choice of technique is arbitrary for anonymization and utility. In [4], Zheleva and Getoor attempt to predict the private attributes of users in four real-world data sets: Facebook, Flickr, Dogster, and BibSonomy. 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 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. In fact, Zheleva and Getoor work provides a substantial motivation for the need of the solution proposed in our work. In [7], Talukder et al. propose a method of measuring the amount of information that a user reveals to the outside world and which automatically determines which information should be removed to increase the privacy of an individual. Finally, in [5], we do 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. PREVENTING INFERENCE ATTACKS ON SOCIAL NETWORKS

The conflict between the desired use of data and individual privacy presents an opportunity for privacy-preserving social network data mining—that is, the discovery of information and relationships from social network data without violating privacy. Privacy concerns of individuals in a social network can be classified into two categories: privacy after data release, and 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 is the AOL search data scandal. In 2006, AOL released the search results from 650,000 users for research purposes. However, these results had a significant number of "vanity" searches on an individual's name, social security number, or address that could then be tied back to a specific individual [2]. Private information leakage, conversely, 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. A trivial example of this type of information leakage is a scenario where a user, say John, does not enter his political affiliation because of privacy concerns. However, it is publicly available that he is a member of the "legalize the same sex marriage." Using this publicly available information regarding a general group membership, it is easily guessable what John's political affiliation is. Somewhat less obvious is the favorite movie "The End of the Spear." We note that this is an issue both in live data and in any released data.

This paper focuses on the problem of private information leakage for individuals as a direct result of their actions as being part of an online social network. We model an attack scenario as follows: Suppose Facebook wishes to release data to electronic arts for their use in advertising games to interested people. However, once electronic arts have 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 could determine the affiliation of other users in their data, this would obviously be a privacy violation of hidden details. We explore how the online social network data could be used to predict some individual private detail that a user is not willing to disclose and explore the effect of possible data sanitization approaches on preventing such private information leakage, while allowing the recipient of the sanitized data to do inference on nonprivate details. This problem of private information leakage could be an important issue in some cases. Recently, both ABC News [3] and the Boston Globe [4] published reports indicating that it is possible to determine a user's sexual orientation by obtaining a relatively small subgraph from Facebook that includes only the user's gender, the gender they are interested in, and their friends in that subgraph. Predicting an individual's sexual orientation or some other personal detail may seem like inconsequential, but in some cases, it may create negative repercussions. For example, using the disclosed social network data, predicting an individual's likelihood of getting Alzheimer disease for health insurance and employment purposes could be problematic.

To the best of our knowledge, this is the first paper that discusses the problem of sanitizing a social network to prevent inference of social network data and then examines the effectiveness of those approaches on a real-world data set. In order to protect privacy, we sanitize both details and the underlying link structure of the graph. That is, we delete some information from a user's profile and remove some links between friends. We also examine the effects of generalizing detail values to more generic values. We then study the effect these methods have on combating possible inference attacks and how they may be used to guide sanitization. We further show that this sanitization still allows the use of other data in the system for further tasks. In addition, we discuss the notion of "perfect privacy" in social networks and give a formal privacy definition that is applicable to inference attacks discussed in this paper.

Class learning algorithms are used on released data to predict private information. Inference attacks are initiated using released social networking data to predict private information. Collective inferences are used to discover sensitive attributes of the data set. Social network data classification is carried out with the combination of node details and connecting links in the social graph. Navie bayes classification algorithm is tuned to classify friendship links in a network. Local classifier, a relational classifier, and a collective inference algorithm are the three components used in the social network analysis. Local classifiers are a type of learning method that are applied in the initial step of collective inference.

The relational classifier analyzes the link structure and labels of the node to identify a model for classification. Collective inference algorithm is used to increase the classification accuracy from the local and relational data values. The following problems are identified in the social network privacy preservation methods. They are static sensitive attribute selection, key node identification is not optimized, key node remove and alter operations are not supported limited privacy model.

IV. CLASSIFICATION OF SOCIAL NETWORK DATA

Collective inference is a method of classifying social network data using a combination of node details and connecting links in the social graph. Each of these classifiers consists of three components: a local classifier, a relational classifier, and a collective inference algorithm.

4.1. 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 naive 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.

4.2. 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 [1], 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). That is, it defines

$$P(C_x^i | N_i) = \frac{P(N_i|C_x^i) \times P(C_x^i)}{P(N_i)} = \prod_{n_j \in N_i} \frac{P(C_a^j|C_x^i) \times P(C_x^i)}{P(n_i)}$$

where N_i are the neighbors of n_i, and then uses these probabilities to classify n_i. The nLB classifier collects the labels of the neighboring nodes and by means of logistic regression, uses these vectors to build a model. In the wvRN relational classifier, to classify a node n_i, each of its neighbors, n_j, is given a weight. The probability of n_i being in class C_x is the weighted mean of the class probabilities of n_i's neighbors. That is,

$$P(C_x^i | N_i) = \frac{1}{Z} \sum_{n_j \in N_i} [W_{i,j} \times P(C_x^i)];$$

where N_i is the set of neighbors of n_i and $w_{i,j}$ is a link weight parameter given to the wvRN classifier. For our experiments, we assume that all link weights are 1.

4.3. Collective Inference Methods

Unfortunately, there are issues with each of the methods described above. Local classifiers consider only the details of the node it is classifying. Conversely, relational classifiers consider only the link structure of a node. Specifically, a major problem with relational classifiers is that while we may cleverly divide fully labeled test sets so that we ensure every node is connected to at least one node in the training set, real-world data may not satisfy this strict requirement. If this requirement is not met, then relational classification will be unable to classify nodes which have no neighbors in the training set. Collective inference attempts to make up for these deficiencies by using both local and relational classifiers in a precise manner to attempt to increase the classification accuracy of nodes in the network. By using a local classifier in the first iteration, collective inference ensures that every node will have an initial probabilistic classification, referred to as a prior. The algorithm then uses a relational classifier to reclassify nodes. At each of these steps i > 2, the relational classifier uses the fully labeled graph from step i - 1 to classify each node in the graph.

The collective inference method also controls the length of time the algorithm runs. Some algorithms specify a number of iterations to run, while others converge after a general length of time. That is, at each step i, the algorithm uses the probability estimates, not a single classified label, from step i - 1 to calculate new probability estimates. Further, to account for the possibility that there may not be a convergence, there is a decay rate, called α set to 0.99 that discounts the weight of each subsequent iteration compared to the previous iterations. We chose to use relaxation labeling because in the experiments conducted by Macskassy and Provost [1], relaxation labeling tended to be the best of the three collective inference methods. Each of these classifiers, including a relaxation labeling implementation, is included in NetKit-SRL. As such, after we perform our sanitization techniques, we allow NetKit to classify the nodes to examine the effectiveness of our approaches.

V. PRESERVATION OF USER PROFILES IN SOCIAL NETWORKS

The privacy preservation model is designed to protect sensitive attribute in user profiles from social networks. The user can select attribute for hiding process. Local classification and relational classification are applied to estimate anonymity levels. The system manages the key node alter and remove operation. The social network data classification is performed with privacy preservation model. Profile data values are analyzed in tree structure. Incremental mining is performed with privacy update features. The system is divided into five major modules. They are profile analysis, initial classification, inference analysis, keynode management and privacy preservation. The profile analysis module is designed to collect user profile information. Initial classification is designed to perform local and relational classification process. Inference analysis module is designed to classify the profile information. Keynode alter and delete operations are carried out under the keynode management module. Privacy preservation module is designed to anonymize the sensitive data values.

Social network account details are collected from the users. User profile data values are fetched from the account information. Friendship link details are collected from the account information. User profile attributes are data values are updated into the database. Local and relational classification operations are performed in the initial classification process. Learning process is performed in the local classification process. Link structure and their relationships are analyzed in the relational classification process. Classification model is learned from the initial classification process. The social graph is used in the inference analysis process. Social graph is constructed with the friendship link values. Node details and connecting links are used in the inference analysis process. Classification process assigns the labels for the users.

Profile attributes are referred as keynodes from the users. Social graph is updated with reference to the keynodes. Profile attribute changes are reflected in keynode alter process. Keynode delete operation removes the keynodes from the social graph environment. Privacy preservation module is used to protect sensitive attributes. Sensitive attribute details are collected from the users. Anonymization methods are used for the privacy preservation process. Dynamic threshold levels are used for the privacy preservation process.

VI. CONCLUSION

The social networks are constructed with user profiles. The sensitive attributes are protecting using privacy preservation methods. The local and relational classification methods are used to protect sensitive attributes. Keynode management mechanism is enhanced in the system. The system improves the privacy for sensitive attributes in social networks. User choice based attribute selection is provided in the privacy preservation process. Efficient graph management mechanism is provided in the social network profile classification system. The system supports incremental mining model with privacy ensured classification scheme.

REFERENCES

[1] S.A. Macskassy and F. Provost, "Classification in Networked Data: A Toolkit and a Univariate Case Study," J. Machine Learning Research, 2007.

[2] T. Zeller, "AOL Executive Quits After Posting of Search Data," The New York Times, no. 22, Aug.2006.

[3] K.M. Heussner, "'Gaydar' n Facebook: Can Your Friends Reveal Sexual Orientation?" ABC News, Sept. 2009.

[4] E. Zheleva and L. Getoor, "To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private user Profiles," Technical Report CS-TR-4926, Univ. of Maryland, College Park, July 2008.

[5] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring Private Information Using Social Network Data," Proc. 18th Int'l Conf. World Wide Web (WWW), 2009.

[6] P. Sen and L. Getoor, "Link-Based Classification," Technical Report CS-TR-4858, Univ. of Maryland, Feb. 2007.

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

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

[9] E. Zheleva and L. Getoor, "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, 2008.

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