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

Inferring Private Information from Social Network Using Collective Classification

A.Annapoorani[1], Ms.P.Indira Priya[2]

P.G student, Tagore Engineering College, Chennai, Tamilnadu, India[1]

Senior Asst. Professor, Tagore Engineering College, Chennai, Tamilnadu, India[2]

ABSTRACT: Online social networks are used by many people. These Social networks allow their users to connect by means of various link types in which the network gives an opportunity for people to list details about themselves that are relevant to the nature of the network. Here there is a chance of inference when user released some personal information in the network. Social network is represented as graph structure in which nodes and edges denotes user's of network and relationship links with friends. In this paper, the social network data has been classified with the help of collective classification (both node and link classification) method. Using the collective classification method the system could infer more sensitive information from the network with high accuracy. In collective classification method, it involves three components called local classifier, relational classifier and collective inference. From this experiments conducted in this research work, it is observed that the proposed work provide better classification accuracy due to the application of collective classification method in link analysis.

KEYWORDS: Social Network Analysis, Data Mining, Inference, machine learning methods, Collective Classification Algorithm.

I. INTRODUCTION

Social networking used to connect and share information with friends. People may use social networking services for different reasons: to network with new contacts, reconnect with former friends, maintain current relationships, build or promote a business or project, participate in discussions about a certain topic, or just have fun meeting and interacting with other users.

Facebook and Twitter, have a broad range of users. LinkedIn has positioned itself as a professional networking site—profiles include resume information, and groups are created to share questions and ideas with peers in similar fields. Unlike traditional personal homepages, people in these societies publish not only their personal attributes, but also their relationships with friends. It may causes the privacy violation in social networks. Information privacy is needed for users. Existing techniques are used to prevent direct disclosure of sensitive personal information.

This paper focuses on social network data classification and inferring the individual's private information. More private information are inferred by applying collective classification algorithm. The system explore how the online social network data could be used to predict some individual private trait that a user is not willing to disclose (e.g. political or religious affiliation).

For instance, 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. This is called personal information leakage from inference.

II. ORGANIZATION OF THE PAPER

This paper is organized as follows. In Section 1, it deals with introduction. Section 2 , describes the organization of the thesis. Section 3 , briefly describes the related work of the research. Section 4, describes the proposed system, system design of the proposed work and the system functions. the system design of the proposed work, Section 5, discuss the result and Section 6, describes the conclusion and future work for proposals.

III. RELATED WORK

Lars Backstrom, Cynthia Dwork and Jon Kleinberg consider an attack against an anonymized network. In their model, the network consists of only nodes and edges. Detail values are not included. The goal of the attacker is simply to identify people. Backstrom and Kleinberg consider a “communication graph,” in which nodes are e-mail addresses, and there is a directed edge (u, v) if u has sent at least a certain number of e-mail messages or instant messages to v, or if v is included in u’s address book.

Here they will be considering the “purest” form of social network data, in which there are simply nodes corresponding to individuals and edges indicating social interaction, without any further annotation such as time-stamps or textual data.

Michael Hay, Gerome Miklau, David Jensen, Philipp Weis, and Siddharth Srivastava consider several ways of anonymizing social networks. Advances in technology have
made it possible to collect data about individuals and the connections between them, such as email correspondence and friendships. Agencies and researchers who have collected such social network data often have a compelling interest in allowing others to analyze the data.

Hay et al. and Liu and Terzi consider several ways of anonymizing social networks. Our work focuses on inferring details from nodes in the network,not individually identifying individuals. 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.

Compared to Jianming He approch, provide techniques that can help with choosing the most effective details or links that need to be removed for protecting privacy. Sen and Getoor compare various methods of link-based classification including loopy belief propagation, mean field relaxation labeling, and iterative classification. They rate each algorithm in terms of its robustness to noise, both in attribute values and correlations across links. And also compare the performance of these classification methods &various types of correlations across links.

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. Zheleva and Getoor work provides a substantial motivation for the need of the solution proposed in our work.

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 (on a per-user basis) should be removed to increase the privacy of an individual.

For example, telephone accounts previously determined to be fraudulent may be linked, perhaps indirectly, to those for which no assessment yet has been made. Macskassy and Provost discuss various classification algorithms for social network classification and Such networked data present both complications and opportunities for classification and machine learning. Finally, the system infer the individuals private information by classifying the publically released social network user data.

IV. PROPOSED SYSTEM

The proposed system use collective classification algorithm for classifying the social network data. It has three components: local classifier, relational classifier and collective inference. Relaxation labeling is used as collective inference method. By applying the collective classification method the system could infer (indirect disclosure) the user private information using the released network data.

The advantage of the system: Collective classification used to improve the classifier accuracy. The collective inference method (relaxation labeling) runs 99 iterations for classifying the network data. It uses local classifier as first iteration and set as

a prior, and relational classifier as second iteration for trying more combinations with nodes and links to gain more user attributes which is used to infer the personal information.

4.1. SYSTEM ARCHITECTURE

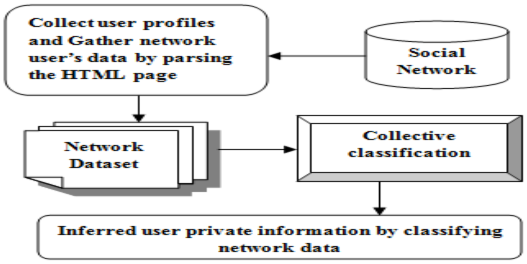

Collect user profiles and Gather network user's data by parsing the HTML page
Social Network
Network Dataset
Collective classification
Inferred user private information by classifying network data

Fig 4.1 System Architecture Diagram

Crawl the Social (Ex.Facebook) network to gather data for experiments. Here the crawler loaded a profile, parsed the details out of the HTML, and stored the details inside a MySQL database. Then, the crawler loaded all friends of the current profile and stored the friends inside the database both as friend- ship links and as possible profiles to later crawl.

By crawling the profile the dataset has been collected for the experiment. From the dataset, the user profiles and links are converted into the graph structure. Then use the collective classification method on social network user data to infer the user's private information.

a. SOCIAL NETWORK DATA GATHERING

For proposed work the details have been collected as follows.Username and password details of users in social network such as Facebook are collected. Log in to user accounts and download their profiles as .html files. Now apply html parser to that parses HTML files and collects attribute values of user profiles. Store the results in database. The records in database are exported into .csv format file for network classification. Model the dataset file as network graph.

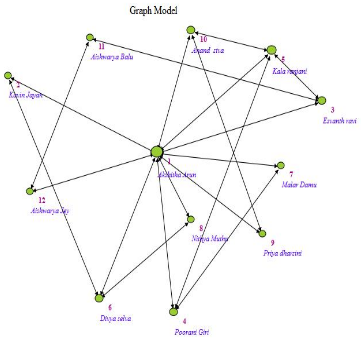


Fig 4.2 Social network graph structure

A Social network is represented a graph structure. The graph model contains vertex, edges and details, where each node represents a unique user of the social network. The set of edges in the graph, which are the links defined in the social network and the links used to establish the connection between the friends in the network.

b. NETWORK CLASSIFICATION

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.

Local classifiers are a type of learning method that are applied in the initial step of collective inference. Naive bayes algorithm is used as a local classsifier. This classifier builds a model based on the details of nodes in the training set. It then applies this model to nodes.

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. 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).

Local classifiers consider only the details of the node it is classifying. And relational classifiers consider only the link structure of a node. Collective inference uses both node and links in the network to improve the classifier accuracy. 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.

For collective inference, relaxation labeling was best when there are few known labels. For relational classification, the link-based classifier clearly was preferable when many labels were known. The lower-variance methods (wvRN and cdRN) dominated when fewer labels were known. Relaxation Labeling - repeatedly estimate class distributions on all unknowns, based on current estimates.

Steps involved in Collective classification:

Step 1: Assign initial label using local classifier. Use naïve bayes algorithm as local classifier.

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

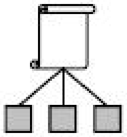


Fig.4.3 Node in a network

Step 2: In first iteration the Naïve Bayes classifier selects the most likely classification V_{nb} given the attribute value $a_{1,}a_{2,\ldots,}a_{n.}$ This result in,

$$V_{nb} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

Generally estimate P $(a_i |V_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

Where, n = the number of training examples for which $v = v_{j\,;}$ n_c = number of examples for which $v = v_j$ and $a = a_i$; p = a priori estimate for $P(a_i \mid v_J)$;
m = the equivalent sample size

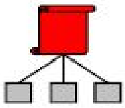


Fig 4.4 Assign Initial Label

Step 3: Assign Initial Label which has high probability.Set initial label as prior. Start the second iteration using relational classifier as weighted vote Relational Neighbor.

Step 4: 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,

$$P(C_x^i|\mathcal{N}_i) = \frac{1}{Z} \sum_{n_j \in \mathcal{N}_i} [w_{i,j} \times P(C_x^j)],$$

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. Assume that all link weights are 1.

Step 5: Learn a classifier from the labels or/and attributes of its neighbors to the label of one node. Here the network information is used.

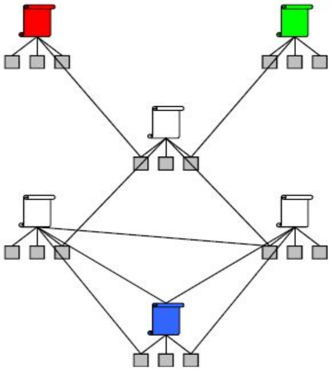


Fig.4.5 Use the attributes of related objects.

Step 6: Apply relational classifier to each node iteratively and reclassify the labels.

Step 7: Relaxation labeling is used to assign the number of iterations to run and Iterate until the inconsistency between neighboring labels is minimized.

V. RESULTS AND DISCUSSION

When classifying the social network data by collective classification method, it improves the classifier accuracy. By doing this, the proposed system could infer user private information with high accuracy. Consider the details and accuracy of the classifiers when infer the private information with various classification methods.

Table 5.1 Classifier Accuracy

Accuracy Private Data	Local Classifier Only	Relational Classifier Only	Collective Classification
Gender	0.7214	0.1672	0.8621
Religion	0.5134	0.4751	0.9519
Political Views	0.5541	0.2151	0.6273
Sexual Orientation	0.4023	0.2543	0.6979

In proposed system, local classifier uses the naïve bayes algorithm. Naïve bayes classifies the user nodes in the network and it finds the probability based on the node attributes. wvRN algorithm is used for relational classification. It used to infer the details from the friendship links. Both the algorithms are infer the data from node/links. In this the system first it classifies the node attributes and set as prior. So here some class labels are known. For collective inference, relaxation labeling and wvRN was the best when there are few known labels. Relational classifier is used as relational classifer and reassigns the class labels based on the link details. The table 5.1 shows that the calculation of Various classifier accuracy.

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

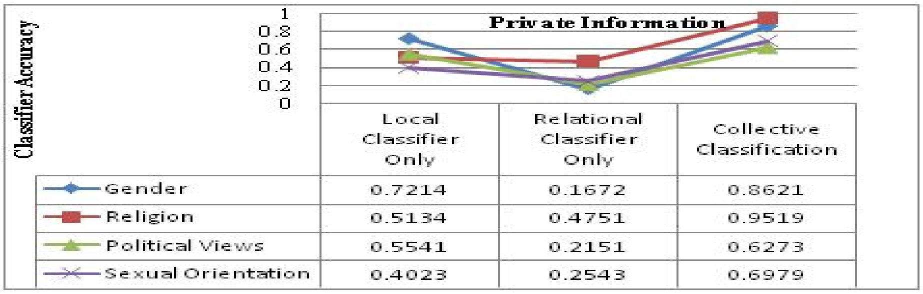
International Journal of Innovative Research in Computer and Communication Engineering

(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



	Local Classifier Only	Relational Classifier Only	Collective Classification
Gender	0.7214	0.1672	0.8621
Religion	0.5134	0.4751	0.9519
Political Views	0.5541	0.2151	0.6273
Sexual Orientation	0.4023	0.2543	0.6979

Fig.5.1 Calculating Classifier Accuracy

From this experiments conducted in this research work, it is observed that the proposed work provide better classification accuracy due to the application of collective classification method in link analysis.

VI. CONCLUSION AND FUTURE WORK

Here collective classification method used to infer the private information from the user nodes and related links. The system showed that, user's private information can be inferred via social relations and release of personal information in the social network.

To protect the individuals private information leakage in social networks, the system either hide our friendship relations or ask our friends to hide their attributes. For protecting the user's private information perform the sanitization process and suppression techniques on the network data. When sanitize the network data it reduces the chance of inferring the individuals private information.

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] 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.

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

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

[7] 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.

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

[9] 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.

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