Security in Prediction of Private Information on Social Network

R. Shiny Jenita
PG Scholar/Dept of CSE,
Karunya University, Tamil Nadu, India,

J. A. M. Rexie
Assistant Professor/Dept. of CSE,
Karunya University, Tamil Nadu, India,

Abstract

Social networks are online applications that allow the users to connect by various links. The user can bring out the information of their friends in the networks and some are private information. This private information has higher possibilities to predict the private information from the user's information using some learning algorithms. Both friendship links and details together gives better predictability than details alone. The usage of inference attacks are engaged to predict private information using the social networking data. Three refinement techniques are created which is used in various situations and the effectiveness of these techniques are explored. In the process, the results from the collective inference implications are combined with the individual results. It removes the details and friendship links together. It is the best way to reduce classifier accuracy. This method is probably infeasible in maintaining the use of social networks. However, removing the details only algorithm, greatly reduces the accuracy of local classifiers. Naive Bayes algorithm gives us the maximum accuracy that is able to achieve through any combination of classifiers. The objective is to reduce the classifier accuracy, while the details and the friendship links are removed together.

Index Terms – Privacy definition, anonymization network, data mining, social network analysis.

1. Introduction

The online applications allow their users to connect by means of various link types. For example, Facebook is a general-use social network, so individual users list their favourite activities, games, books, and movies. Likewise, LinkedIn is a professional network; because, users specify details which are related to their professional life.

The main objective is to develop a technique for prevent the inference attacks on privacy information in Social Networks. Social networks are online applications that allow the users to connect by various links. This type of networks allows the users to share the details to their friends in the network. The user can bring out the information of them in the networks and some are private information. This private information has higher possibilities to predict the private information from the user's information using some learning algorithms. A data mining technique like inference attack performed by analysing data for criminally gain knowledge about database. It may lead to predict the user's private information. So the privacy information is leaked[1].

In this process, the collective inference does not improve on using a simple local classification method to identify nodes. When combine the results from the collective inference implications with the individual results, that removing details and friendship links together is the best way to reduce classifier accuracy.

2. Degree anonymization

To reduce the degree of anonymization problems, develop a set of algorithms under graph construction and its relaxed graph construction version.

2.1 Graph construction

The Supergraph algorithm, which is an extension of the Construct Graph algorithm. The algorithm operates on the sequence of additional degrees in a manner similar to the one the Construct Graph algorithm operates on the degrees. The supergraph inputs are the original graph G and the desired k-anonymous degree. In each iteration it picks an arbitrary vertex v and adds edges

from v to a(v) vertices of highest residual additional degree, ignoring nodes v0 that are already connected to v in G. The a(v0) is decreased by 1, for every new edge (v; v0). If the algorithm terminates and out-puts a graph. If the algorithm does not terminate, then it outputs \Unknown". Though supergraph is similar to Construct Graph, it is not an oracle. This method is simple and efficient because these algorithms are based on princilples. But it is difficult to measure the utility of a graph.

2.2 Relaxed Graph Construction

The greedy swap algorithm and the priority algorithm are used in relaxed graph construction problem. This algorithm halts when there are no more valid swaps that can increase the size of the edge intersection. And also additionally show a simple modification of the Construct Graph algorithm that allows the construction of degree anonymous graphs with similar high edge intersection with the original graph directly, without using Greedy Swap. This algorithm is also known as the Priority algorithm because during the graph construction phase, it gives priority to already existing edges in the input graph *G* (*V; E*). The Priority algorithm is less computationally demanding than the naive implementation of the Greedy Swap procedure. This Priority algorithm is similar to the Construct Graph. In the case where Priority fails to construct a graph by reaching a dead-end in the edge allocation process, the Probing scheme is employed; and random noise addition is enforced until the Priority algorithm outputs a valid graph [4].

3. Graph anonymization

The process of anonymization involves taking the unanonymized graph data, making some modifications, and constructing a new released graph which will be made available to the adversary. The modifications include changes to both the nodes and edges of the graph.

3.1 Node anonymization

Assume that the nodes have been anonymized with one of the techniques introduced for single table data. This anonymization provides a clustering of the nodes into m equivalence classes (C1, . . . , Cm) such that each node is indistinguishable in its quasi-identifying attributes from some minimum number of other nodes. Using the notation C(vi) = Ck to specify that a node vi belongs to equivalence class Ck. The anonymization of nodes creates equivalent classes of nodes. Note, however, that these equivalent classes are based on node attributes only, there may be nodes with different identifying structural properties and edges[10].

3.2 Edge anonymization

For the relational part of the graph five possible anonymization approaches are described. The range from one which removes the least amount of information to a very restrictive one, which removes the greatest amount of relational data. Figure.1. (a) shows a simple data graph in which there are ten nodes and eight observed edges. The first (trivial) edge anonymization option is to only remove the sensitive edges, leaving all other observational edges. Figure.1.(b) shows an illustration of this technique applied to the original data graph of Figure.1. (a). In this running example, remove the friendship relationships, since they are the sensitive relationships. But leave the information about students taking classes together and being members of the same research group which have low privacy preservation.

Another anonymization option is to remove some portion of the relational observations. Either remove a particular type of observation which contributes to the overall likelihood of a sensitive relationship, or remove a certain percentage of observations that meet some pre-specified criteria (e.g., at random, connecting high-degree nodes, etc.). Figure.1. (c) shows an illustration of this technique when the edges are removed at random and Figure.1. (d) shows an illustration of the result from applying the algorithm. In order to determine the number of edges of a particular type connecting two equivalence classes, and also the anonymization algorithm picks the maximum of the number of edges of that type between any two nodes of the original graph. The maximum number of common classes that any pair of students from the two equivalence classes takes is one class together, then the equivalence classes are connected by one class edge. Figure 1(e) shows an illustration of this technique[10].

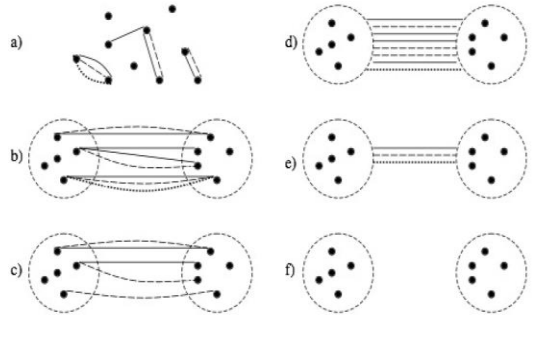

Figure 1: The original data graph (a)) and the output from five anonymization approaches to graph data:b) revealing the observations between nodes, c) removing 50% of the observations , d) revealing all the observations between equivalence classes of nodes (cluster-edge anonymization), e) constrained revealing of the observations between equivalence classes of nodes (cluster-edge anonymization with constraints), f) removing all relational observations.

4. Link-based classification

In this model is learned from a fully labeled training data set. In order to make use of it for prediction, in a more challenging situation than classical prediction. The object attributes and the links are observed only the categories are unobserved. To be able to predict the categories of all of the objects at once clearly each of these predictions depends on neighboring predictions. In proposed model, to predict the category for one object, the categories of object's neighbors, which will also be unlabelled. The iterative classification algorithm terminates when it converges or a maximum number of steps has been reached.
Step 1: Assign an initial category for each object in the test set.
Step 2: Iteratively apply the full model to classify each object until the termination criterion have been satisfied. (Iteration) For each object,
1. Compute the link statistics, based on the current assignments to linked objects
2. Compute the posterior probability for the category of this object.
3. The category with the largest posterior probability is chosen as a new category for current object.
In the iterative step there are many possible orderings for objects. Then evaluate the effectiveness of different ordering schemes based on link diversity. The logistic classifier built over the object attributes & link statistics outperforms simple content-only classifier[5]. But the convergence rate is slower in the outgoing link diversity.

5. Privacy attacks using links and groups

Link-based privacy attacks take advantage of autocorrelation, so that the property that the attribute values of linked objects are correlated. Example of autocorrelation is that people who are friends often share common characteristics.
In addition to friendship or link information, the social networks offer a very rich structure through the group memberships of users. Every individual users in a group are bound together by some observed or hidden interest(s) that they share, and every individuals often belong to more than one group. Likewise groups offer a broad perspective on a person, and it may be possible to use them for sensitive attribute inference[11]. This problem becomes more complex, and their distributions suggest different values for the sensitive attribute.

It is possible to construct a method which uses both links and groups to predict the sensitive attributes of users. Use a simple method which combines the flat-link and the group-based classification models into one: LINK-GROUP. It uses all links and groups as features Thus utilizing the full power of available data. Like LINK and GROUP, LINK-GROUP can used in any traditional classifier[11]. The advantage of this method is able to discover the sensitive attribute values of some users with surprisingly high accuracy on the real-world social-media datasets.

6. Attacks against k-anonymity

Even when sufficient care is taken to identify the quasi-identifier, and a solution that adheres to *k*-anonymity can still be vulnerable to attacks. They are described below,

6.1 Unsorted matching attack against *k*-anonymity

This attack is based on the order in which records appear in the released table. The release of a related table can leak sensitive information[7]. While maintained the use of a relational model, in real-world use this is often a problem, the order of tuples cannot be assumed. It can be corrected by randomly sorting the tuples of the solution table.

6.2 Complementary release attack against *k*-anonymity

As a result, when a table T, which adheres to *k*-anonymity, which is released. It should be considered as joining other external information. If all the attributes were in the quasi-identifier. It is more common that the attributes that constitute the quasi-identifier. Therefore, subsequent releases of the same privately held information must consider all of the released attributes of T a quasi-identifier to prohibit linking on and based on T [8].

6.3 Temporal attack against *k*-anonymity

Data collections are dynamic. Records are added, changed, and removed constantly. As a result, the generalized data which is released over time can be subject to a temporal inference attack[8].

7. Bayes error estimation

Parametric Estimates of the Bayes Error is one of the simplest bounds for the Bayes error is provided by the Mahalanobis distance measure (Devijver and Kittler 1982). The main advantage of this bound is the lack of restriction on the class distributions. Furthermore, it is easy to calculate using only sample mean and sample covariance matrices. It provides a quick way of obtaining an approximation for the Bayes error. But, it is not a particularly tight bound[9].

A nonparametric estimation method that provides an estimate for the Bayes error without requiring knowledge of the class distributions is based on the nearest neighbor classifier. The NN classifier assigns a test pattern to the same class as the pattern in the training set to which it is closest (defined in terms of predetermined distance metric)[11].Bayes error estimation based on decision boundaries there are many ways of combining the outputs of multiple classifiers. For example, if each classifier only provides the class label, then majority vote can be used. If suppose the outputs of the individual classifiers approximate the corresponding class posteriors, then the simple averaging of the posteriors and then picking the maximum of these averages typically proves to be an effective combining strategy[9].

8. Attacks on anonymized social network

In this method present both active and passive attacks on anonymized social networks, showing that both types of attacks can be used to reveal the true identities of targeted users, even from just a single anonymized copy of the network.

The active attacks will make use of the following two types of operations. In the first operation, an individual can create a new user account on the system; this adds a new node to G. Second, a node u can decide to communicate with a node v; this adds the undirected edge (u, v) to G.

The passive attack is based on the observation that most nodes in real social network data already belong to a small uniquely identifiable subgraph. If a user u is able to collide with a coalition of k − 1 friends after the release of the network, he or she will be able to identify additional nodes that are connected to this coalition, and thereby learn the edge relations among them [3].

9. Anonymization techniques

In privacy preserving data publishing, to prevent privacy attacks, data should be anonymized properly before it is released. Generalization and perturbation are the two popular anonymization approaches for relational data.

Anonymization methods should take into account the privacy models of the data and the utility of the data. Although privacy preservation in social network data is a relatively new problem, several privacy preserving methods have been developed. Like privacy preservation methods in relational data, the specific anonymization methods are developed for specific privacy models of social networks and specific utility goals of anonymized data[12]. As social network data is much more complicated than relational data, and the privacy preserving in social networks is much more challenging and needs many serious efforts in the near future.

10. Learning methods

10.1 Sanitizing technique

Data Sanitization is the technique which is used to disguising sensitive information and developed databases by overwriting it with looking realistic but false data of a similar type. The data in testing environments should be sanitized in order to protect valuable business information. Basically there are two types of security. The first type is concerned data integrity. In this type the modification of the records is strictly controlled. For example, it is not allowed to be credited or debited without specific controls. This type of security is not a major concern in test and development databases. This data can be modified without any business impact. The second type of security is the protection of the information from inappropriate visibility. Examples of this type of data are Names, addresses, phone numbers and credit card details[2]. This type of security requires that access to the information content is controlled in every environment.

10.2 Anonymization social network

Advances in technology have made it possible to collect data about individuals and the connections between them. Researchers who have collected such

social network data often have a compelling interest in allowing others to analysed data. In many cases the data describes relationships that are private and sharing the data in full can result in unacceptable disclosures.
Social network analysis is concerned with uncovering patterns in the connections between entities and widely applied to organizational networks to classify the influence or popularity of individuals and to detect collusion and fraud. Technological advances have made it easier than ever to collect the electronic records that describe the social networks. The agencies or the researchers who collect such data are often faced with a choice between two undesirable outcomes, can publish data for others to analyze, even though that analysis will create severe privacy threats, or withhold data because of privacy concerns, even though that makes further analysis is not possible [6].

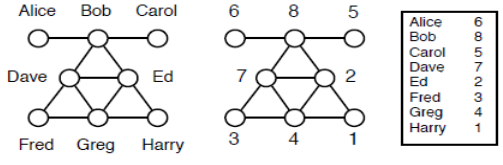

Figure 2: The naive anonymization of social network *G*; the anonymization mapping.

A graph shown in Figure 2 along with its naive anonymization, in which synthetic identifiers have replaced names. The anonymization mapping, shown in (c), is a random, protected mapping. Naive anonymization is a common practice. For example, an adversary may learn that Bob has at least three neighbours. And it follows that the node corresponding to Bob in the published graph must be 2, 4, 7 or 8 nodes. Thus, an entity's position in the graph of relationships acts as a quasi-identifier attribute. The extent to which an individual can be distinguished using graphical position depends on the structural similarity of nodes in the graph [3].

10.3 Naive bayes algorithm

Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. It provides a useful perspective for understanding and evaluating many learning algorithms. And also calculates explicit probabilities for hypothesis and it is robust to noise in input data. Bayesian reasoning is particularly suited when the dimensionality of the inputs is high. It is applied to decision making and inferential statistics that deals with probability inference, used to predict future events. Naive Bayes models parameter estimation uses the method of maximum likelihood. It requires a small amount of training data to estimate the parameters is the advantage. Determining an individual's political affiliation using a node n_i with m details and p potential classification labels, $C_1, \ldots, C_p$, C^i_x , the probability of ni being in class C_x, is given by the equation

$$\underset{1\le x\le p}{\text{argmax}} \; [\, P(C^i_x) \,|D^1_i,\ldots,D^m_i)] \qquad (1)$$

where arg max1 $\le x \le$ p represents the possible class label. But this is difficult to calculate, when x is not known. Then applying Bayes' theorem,

$$\underset{1\le x\le p}{\text{argmax}} \; \frac{[P(C^i_x) \times P(D^1_i,\ldots,D^m_i|C^i_x)\,]}{[\, P(D^1_i,\ldots,D^m_i\,]} \qquad (2)$$

However, that $P(D^1_x,\ldots.D^m_i)$ is equivalent for all values of C^i_x . In that case the probability of seeing any particular detail without consideration of any particular class x is equivalent for all x. To determine a new class label for n_i[9].

$$\underset{1\le x\le p}{\text{argmax}}[P(C^i_x) \times P(D^1_i|\, C^i_x) \times\ldots \times P(D^m_i|\, C^i_x)] \quad (3)$$

10.3.1 Friendship Links

Calculating $P(\, C^i_x \mid N_i\,)$ is to determining the class detail value of person n_i given their friendship links using a Naive Bayes model. If there are few people in the training set that have a friendship link to n_i, the calculations become extremely inaccurate. Instead of this, to decompose this relationship, link from person n_i to n_j, consider the probability of having a link from n_i to someone with n_j's details[9].

10.3.2 Weighing Friendships

The last step is to calculating $P(C^i_x\,|N_i\,)$. There are many ways to weigh friendship links, the method using is very easy to calculate and is based on the assumption that the more public details two people share, the concealed details they are likely to share. In this specific case of social networks, any two friends can be anything from acquaintances to family members or close friends. The following formula for $W_{i;j,}$ which represents the weight of a friendship link from n_i to node n_j:

$$W_{i;j} = \frac{|(D^1_i,\ldots,\, D^n_i\,) \cap (D^1_j,\ldots,\, D^n{}_j)\,|}{|D_i|} \qquad (4)$$

There are four algorithms are used to predict the political affiliation of each user on social network. The

initial algorithm is called "Details Only" algorithm which is used to predict political affiliation and ignores friendship links. Another algorithm is called "Links Only" algorithm which is used to predict political affiliation using friendship links and does not consider the details of a person. The third algorithm is called "Average". This algorithm predicts a node's class value based on the following equation:

$$P_A(C^i_a) = 0.5 * P_D(C^i_a) + 0.5 * P_L(C^i_a) \quad (5)$$

where P_D and P_L are the numerical probabilities assigned by the Details Only and Links Only algorithms, respectively[9]. The traditional naive Bayes classifier is the final algorithm, which is used as a basis of comparison for above proposed algorithms.

11. Conclusion

The proposed system uses learning methods in anonymization and classification tasks for hiding private information. After removal of details, test the removed details as an anonymization technique by using variety of different classification algorithms to test the effectiveness of proposed method. The effect of removing details and links is preventing sensitive information leakage. Compared with most previous studies, the proposed system overcomes the drawbacks of previous techniques. In the process, combine the results from the collective inference implications with the individual results, that removing details and friendship links together is the best way to reduce classifier accuracy. This method is probably infeasible in maintaining the use of social networks. However, removing only details greatly reduces the accuracy of local classifier, it gives the maximum accuracy were able to achieve through any combination of classifiers.

References

[1] L. Backstrom,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.

[2] *"Data Sanitization Techniques,"* A Net 2000 Ltd. White Paper.

[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] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, *"L-Diversity: Privacy Beyond K-Anonymity,"* ACM Trans.Knowledge Discovery from Data, vol. 1, no. 1, p. 3, 2007.

[5] 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.

[6] A. McCallum, A. Corrada-Emmanuel, and X. Wang. *"Topic and role discovery in social networks."* In IJCAI, 2005.

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

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

[9] IEEE Transactions On Knowledge And Data Engineering, VOL. 25, NO. 8, AUGUST 2013 *"Preventing Private Information Inference Attacks On Social Networks"*, Raymond Heatherly, Murat Kantarcioglu, And Bhavani Thuraisingham, Fellow, IEEE

[10] Van Eecke, Maarten Truyens *"Privacy and social networks".*

[11] D.J. Watts K. Liu and E. Terzi, *"Towards Identity Anonymization onGraphs,"* Proc. ACM SIGMOD Int'l Conf. Management of Data(SIGMOD '08), pp. 93-106, 2008.

[12] Zhou, Jian Pei, WoShun Luk School of Computing Science Simon Fraser University, Canada woshun@cs.sfu.ca*," A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data".*