Different sanitization techniques to prevent inference attacks on social network data

Ms. Patel Madhuri
M.E in Information Technology
Parul institute of Engineering and Technology

Abstract:

In Current Time, Social Network has become one of the most Powerful Resources for people to connect to their friends and family members or etc. and also publish their sensitive Information. Some of the information in social network is meant to be private. For different work, the Social Network Data like Medical Dataset or etc was given by the social sites to the advisor for some advertisement of products or etc. But Sometime, these datasets was used to infer the private information using various inference attacks. Here, we present a brief review of the inference attacks on social network dataset. We also identify the new challenges in privacy preserving on social network data comparing to the extensively studied relational case, and examine the possible problem. We survey the sanitization methods which used to preventing inference attacks.

1 Introduction

Data mining in social Network can extend researchers' capability to understanding new phenomena due to the use of social network and also improve business intelligence to provide better services and develop innovative opportunities. Social networks are platforms that allow people to publish details about themselves and to connect to other members of the network through friendship links. Social networks model social relationships by graph structures using vertices and edges. Vertices model individual social actors in a network, while edges model relationships between social actors. Many different kinds of social networks present in our lives such as friendship networks, telephone call networks, and academia co-authorship networks. Recently, the popularity of such on-line social networks is increasing significantly. 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 [1]. We explore how the on-line 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) and explore the effect of possible data sanitization alternatives on preventing such private information leakage.

Disclosing private information means violating the rights of people to control who can access their private information. In order to prevent private information leakage, it is important to be aware of the ways in which an adversary can attack a social network to learn users' private attributes. Studies on the challenges of preserving the privacy of individuals in social networks have emerged only in the last few years, and they have concentrated on inferring the identity of nodes based on structural properties such as node degree. In contrast, we are interested in inferring sensitive attribute of nodes using approaches developed for relational learning, another active area of research in the last few years. The novelty of our work is that we study the implications of mixing private and public profiles in a social network. For example, in Facebook many users choose to set their profiles to private, so that no one but their friends can see their profile details. Yet, fewer people hide their friendship links and even if they do, their friendship links can be found through the back links from their public-profile friends. Similarly for group participation information – even if a user makes her profile private, her participation in a public group is shown on the group's membership list. Currently, neither Facebook nor Flicker allows users to hide their group memberships from public groups. Both commercial and governmental entities may employ privacy attacks for targeted marketing, health care screening or political monitoring – just to mention a few. Therefore, social media website providers need to protect their users against undesired eavesdropping and inform them of the possible privacy breaches and providing them with the means to be in full control of their private data.

1.1 Inference attacks on Social Network Data

Information appearing private means give the rights of people to control that who can access their private information. To stop private information leakage, it is very important to be well informing of the ways in which can attacks a social network to learn users' private sensitive information. Studies on the challenges of protecting the privacy of individuals in social networks have find out only in the few years, and they have focus on assuming the identity of nodes based on structural properties In addition, a social networking site has its business needs to encourage to users for easily find each other and expand their friendship networks as widely as possible. Hence, social media poses new security challenges to avoid security threats to users and organizations. With the variety of personal information assumed in user profiles (e.g., information about other users and user networks may be indirectly accessible), individuals may put themselves and members of their social networks at risk for a

variety of attacks. The goal is not to librated anonymized data but to illustrate how social network data can be exploited to predict hidden information essential information in the anonymization process.

There are many issues related to private information leakage in social network. In survey show that using details alone, one can predict class values more accurately than using friendship links alone. Further research will show that using both friendship links and details together gives better predictability than details alone. It explored the effect of removing traits and links in preventing sensitive information leakage. In the process, research will discovered situations in which collective inferencing does not improve on using a simple local classification method to identify nodes. To combine the results from the collective inference implications with the individual results, it will begin to see that removing details and friendship links together is the best way to reduce classifier accuracy. Users have strong expectations of privacy on such data. When social network data is made public in one way or another, it is far from sufficient to protect privacy by simply replacing the identifying attributes.

1.2 Objective of the Attack

An adversary attacking some data may have various objectives ranging from identifying the home of the target to reconstructing his social network, or even obtaining knowledge of his to:

- **Identify places**, called Points of Interests (POIs) which characterize the interests of an individual [17]. POI may be the school or place of work of an individual or story books or political party. Disclosing the POIs of a particular person is likely to cause a privacy as this data may be used for infer sensitive information such as hobbies, religious beliefs, political preferences or even potential diseases.
- **Assume the movement of an individual** such as his school, college or current working area [12]. From the movement patterns, it is possible to deduce other PII such as the mode of transport, the age or even the lifestyle.
- **Learn the behaviour of an individual** from the knowledge of his POIs and movement patterns. From this information, the adversary can derive a clearer understanding about the interests of an individual as well as his behaviour than simply from his movement location.
- **Link the records of the same individual**, which can be contained in different or same datasets, either anonymized or different pseudonyms. This is the private equivalent of the statistical disclosure risk in which privacy is measured according to the risk of linking the record of the same individual in two different databases.
- **Find out social relations between individuals** by considering that two individuals that are in contact during a non-negligible amount of time where they was share some kind of social link.

2 Inference Techniques

Here we present there are some algorithms and methods that can be used as inference techniques to infer private information:

- **Clustering** is a form of unsupervised learning in that tries to group different objects that are similar in the same cluster while putting objects that are dissimilar in different clusters. This algorithm can be used to find out the POIs of one particular individual if it is fed only with his data or the generic hotspots and if it is given the data of a whole population.
- **Data coming from social applications** is a most available source of information that the attackers might draw attack to the privacy of individuals. Example of social application is Google Latitude that offers the possibility of real-time on a map the movements of friends who have previously agreed to this service by confirming this on a SMS received on their phone.
- **Data coming from public sources** is a potential source of knowledge that can be exploited by the adversary. For instance, by using Google Maps and Yahoo! Maps the adversary can easily reconstruct the path followed by an individual between two consecutive mobility traces.

Now, we study the effect of sanitization methods have on combating possible inference attacks and how they may be used to guide sanitization. A sanitization procedure usually comes with some privacy guarantees. Data Sanitization is the process of disguising sensitive information in test and development databases by overwriting it with realistic looking but false data of a similar type.

Why Sanitize Information in Test and Development Databases?

The data in testing environments should be sanitized in order to protect valuable business information and also because there is, in most countries, a legal obligation to do so.

- **Protecting Valuable Information**
  Fundamentally there are two types of security. The first type is concerned with the integrity of the data. In this case the modification of the records is strictly controlled. The second type of security is the

protection of the information content from inappropriate visibility. Names, addresses, phone numbers and credit card details are
good examples of this type of data.

- **Legal Obligations**
  The legal requirements for Data Sanitization vary from country to country but most countries now have regulations of some form for which your organization is responsible gets loose and appropriate steps were not taken to prevent that release, then your organizations lawyers could well find themselves in court trying to put their best spin on the matter. However large the legal liabilities are, they could seem trivial in comparison to the losses associated with the catastrophic loss of business confidence caused by a large scale privacy breach.

Data Sanitization Techniques

Social network teams need to work with databases which are structurally correct functional copies of the live environments. However, they do not necessarily need to be able to view security sensitive information. For test and development purposes, as long as the data looks real, the actual record content is usually irrelevant.

- **Manipulating details**
  Clearly, details can be manipulated in three ways: adding details to nodes, modifying existing details and removing details from nodes. The goal in the first case is to add details that may prevent learning algorithms from being able to infer a person's private details. In the second case, the goal is to prevent leakage of "accurate" information by modifying profile details (e.g., anonymization techniques). In the third case, the goal is to remove those details that most help a learning algorithm to predict a person's private details. In the context of a social network, removing details does not introduce any misleading information. This follows from the implied nature of listed details inside a social network. If a detail is missing, it simply implies that the person failed to mention that detail. A missing detail does not imply that the detail does not describe the person. However, if a detail is mentioned, then it is implied that the detail does indeed describe the person. Unlike anonymization techniques such as k-anonymity, removing details could be easily done by each individual profile owner.

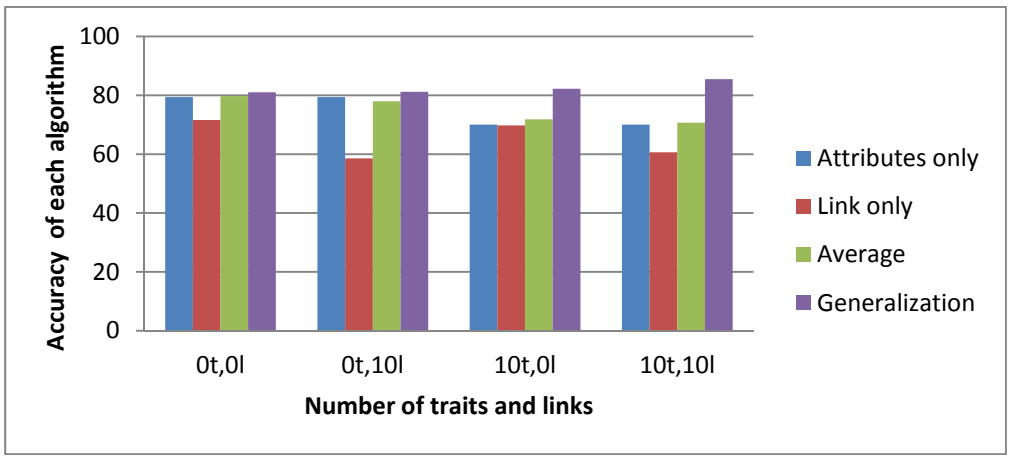

Fig.1 comparison all sanitize techniques

- **Manipulating Link Information**
  Links can be manipulated in the same way details can. For the same reasons given in above section, first choose to evaluate the effects of privacy on removing friendship links instead of adding fake links.
- **Detail Generalization**
  To combat inference attacks on privacy, we attempt to provide detail anonymization for social networks. Generalize each detail type by determining which attributes can be further generalized without complete removal and keep a list of the accuracy of this generalization.

There are sanitization techniques which are useful to combat the social network data from inference attack. According to [4] and [7], we compare all the techniques of sanitization in fig.1. In which we can group the different number of details (attributes or traits) and friendship links. We use four group of traits and link which are (0t,0l),(10t,0l),(0t,10l),and (10t,10l) and find the accuracy of sanitization techniques with each group and finally conclude that the generalization technique is more efficient than others techniques.

Conclusion:

Due to rapidly increasing use of social networking sites for various purposes, more people participate in social networks. In many cases, those social networking sites can serve as CRM tool for companies advertising

products and services. Few Companies can also use social networking site to identify customers or recruit candidate employees. Privacy attacks on relational data are re-identify individuals by joining a published table containing sensitive information with external tables modelling background knowledge of attackers. Various sanitize techniques are used to combat these type of inference attacks and in this paper we compare few techniques of sanitization.

References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] R. Gross, A. Acquits, and J.H. Heinz, "Information Revelation and Privacy in Online Social Networks," Proc. ACM Workshop Privacy in the Electronic Soc. (WPES '05), pp. 71-80, http:// dx.doi.org/10.1145/1102199.1102214, 2005.

[7] R. Heatherly, M. Kantarcioglu, J. Lindamood, and B. Thuraisingham. Preventing private information inference attacks on social networks. Technical Report UTDCS-03-09, University of Texas at Dallas, 2009.