International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161
©2014 INPRESSCO®, All Rights Reserved
Available at http://inpressco.com/category/ijcet

Research Article

Data Implication Attacks on Social Networks with Data Sanitization

Divya.R[Á*], B.Mahesh[Á] and R.Ushasree[B]

[Á]Dr. K.V.Subbareddy College of Engineering for Women. Kurnool, India
[B]The Oxford College of Engineering, Bangalore, India

Accepted 05 June 2014, Available online 20 June 2014, **Vol.4, No.3 (June 2014)**

Abstract

Online social networks, such as Facebook, are increasingly utilized by many people. These networks allow users to publish details about themselves and to connect to their friends. Some of the information revealed inside these networks is meant to be private. Yet it is possible to use learning algorithms on released data to predict private information. In this paper, we explore how to launch inference attacks using released social networking data to predict private information. Then, we explore the effectiveness of these techniques and attempt to use methods of collective inference to discover sensitive attributes of the data set. We show that we can decrease the effectiveness of both local and relational classification algorithms by using the sanitization methods we described. Data Sanitization is the process of making sensitive information in non-production databases safe for wider visibility

Keywords: *Privacy clarity for Formal data, Generalization Module*

1. Introduction

AOL Inc. is a multinational mass media corporation based in New York City that develops, grows, and invests in brands and web sites. The company's business spans digital distribution of content, products, and services, which it offers to consumers, publishers, and advertisers. Founded in 1985 as Quantum Computer Services, an online services company by Jim Kimsey from the remnants of Control Video Corporation, AOL has franchised its services to companies in several nations around the world or set up international versions of its services. Data Sanitization is the process of making sensitive information in non-production databases safe for wider visibility. This White Paper is an overview of various techniques which can be used to sanitize sensitive production data in test and development databases.

An initial discussion of the primary motivations for data sanitization is given. The remainder of the paper is devoted to a generic survey of the various masking techniques and their individual benefits and drawbacks.
Facebook Beacon was a system that posted your activity on third-party websites - Blockbuster, Gamefly, Overstock.com and more - back to your Facebook profile. Privacy advocates rallied against it, however, arguing that data was being sent without the users' explicit permission

2. Literature Survey

Wherefore Art Thou?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography

Authors: L. Backstrom, C. Dwork, and J. Kleinberg

In a social network, nodes correspond topeople or other social entities, and edges correspond to social links between them. In an effort to preserve privacy, the practice of anonymization replaces names with meaningless unique identifiers. We describe a family of attacks such that even from a single anonymized copy of a social network, it is possible for an adversary to learn whether edges exist or not between specific targeted pairs of nodes.

Anonymizing Social Networks

Authors: M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava

Operators of online social networks are increasingly sharing potentially sensitive information about users and their relationships with advertisers, application developers, and data-mining researchers. Privacy is typically protected

*Corresponding author: **Divya.R**

by anonymization, i.e., removing names, addresses, etc.We present 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 microblogging service, and Flickr, an online photo-sharing site, can be re-identified in the anonymous Twitter graph with only a 12% error rate.Our de-anonymization algorithm is based purely on the network topology, does not require creation of a large number of dummy "sybil" nodes, is robust to noise and all existing defenses, and works even when the overlap between the target network and the adversary's auxiliary information is small.

Towards Identity Anonymization on Graphs

Authors: K. Liu and E. Terzi

The proliferation of network data in various application domains has raised privacy concerns for the individuals involved. Recent studies show that simply removing the identities of the nodes before publishing the graph/social network data does not guarantee privacy. The structure of the graph itself, and in its basic form the degree of the nodes, can be revealing the identities of individuals. To address this issue, we study a specific graph-anonymization problem. We call a graph *k*-degree anonymous if for every node *v*, there exist at least *k*-1 other nodes in the graph with the same degree as *v*. This definition of anonymity prevents the re-identification of individuals by adversaries with *a priori* knowledge of the degree of certain nodes. We formally define the graph-anonymization problem that, given a graph *G*, asks for the *k*-degree anonymous graph that stems from *G* with the minimum number of graph-modification operations. We devise simple and efficient algorithms for solving this problem. Our algorithms are based on principles related to the realizability of degree sequences. We apply our methods to a large spectrum of synthetic and real datasets and demonstrate their efficiency and practical utility.

Inferring Privacy Information from Social Networks

Authors: J. He, W. Chu, and V. Liu

Currently, millions of individuals are sharing personal information and building social relations with others, through online social network sites. Recent research has shown that those personal information could compromise owners' privacy. In this work, we are interested in the privacy of online social network users with missing personal information. We study the problem of inferring those users' personal information via their social relations. We present an iterative algorithm, by combining a Bayesian label classification method and discriminative social relation choosing, for inferring personal information. Our experimental results reveal that personal information of most users in an online social network could be inferred through mere social relations with high accuracy.

Preserving the Privacy of Sensitive Relationships in Graph Data

Authors: E. Zheleva and L. Getoor

In this paper, we focus on the problem of preserving the privacy of sensitive relationships in graph data. We refer to the problem of inferring sensitive relationships from anonymized graph data as *link re-identification*. We propose five different privacy preservation strategies, which vary in terms of the amount of data removed (and hence their utility) and the amount of privacy preserved. We assume the adversary has an accurate predictive model for links, and we show experimentally the success of different link re-identification strategies under varying structural characteristics of the data.

3. Overview of Project

Existing System: Other papers have tried to infer private information inside social networks. In, 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, Live Journal, they use hypothetical attributes to analyze their learning algorithm.

Proposed System: 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.

In Proposed System we implemented a proof-of-concept Facebook application for the collaborative management of shared data, called *MController*.

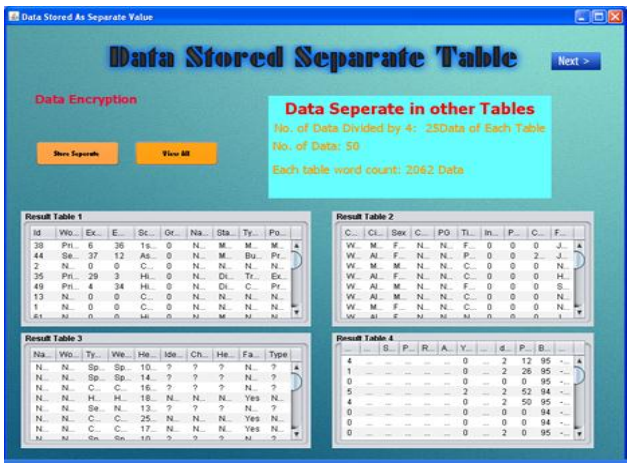


Figure: Data Stored

As we explained in the proposed system the private information will be stored and secured. Our goal is to release preventing rich social network data set .As can be seen from the results, our methods are generally successful at reducing the accuracy of classification tasks.

4. Methods of Social Networks

Privacy clarity for Formal data: In this module we develop the privacy clarity of formal data where, Privacy

definition could be applied to other domains. Consider the scenario where we want to decide whether to release some private information (e.g., eating habits, lifestyle), and combined with some public information (e.g., age, zip code, cause of death of ancestors) or not. We may be worried that whether the disclosed information could be used to build a data mining model to predict the likelihood of an individual getting an Alzheimer's disease. Most individuals would consider such information to be sensitive for example, when applying for health insurance or employment. Our privacy definition could be used to decide whether to disclose the data set or not due to potential inference issues.

Control of data's: Clearly, details can be manipulated in three ways: adding details to nodes, modifying existing details and removing details from nodes. However, we can broadly classify these three methods into two categories: perturbation and anonymization. Adding and modifying details can both be considered methods of perturbation—that is, introducing various types of "noise" into D to decrease classification accuracies. Removing nodes, however, can be considered an anonymization method.

Choosing of details Module: We must now choose which details to remove. Our choice is guided by the following problem statement. This allows us to find the single detail that is the most highly indicative of a class and remove it. Experimentally, we later show that this method of determining which details to remove provides a good method of detail selection.

Operate Link Information: The other option for anonymzing social networks is altering links. Unlike details, there are only two methods of altering the link structure: adding or removing links.

Generalization Module: To combat inference attacks on privacy, we attempt to provide detail anonymization for social networks. By doing this, we believe that we will be able to reduce the value of an acceptable threshold value that matches the desired utility/privacy tradeoff for a release of data.

5. Implementation

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

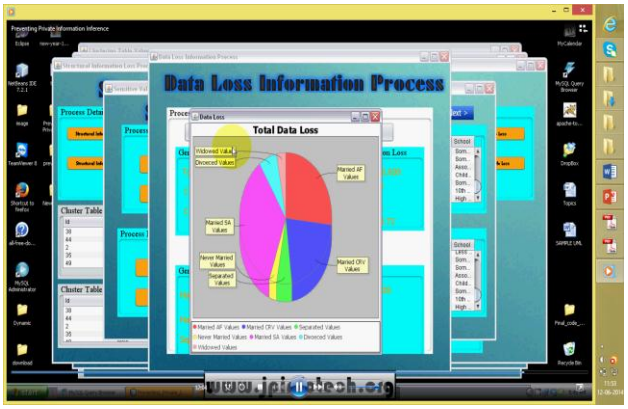
- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

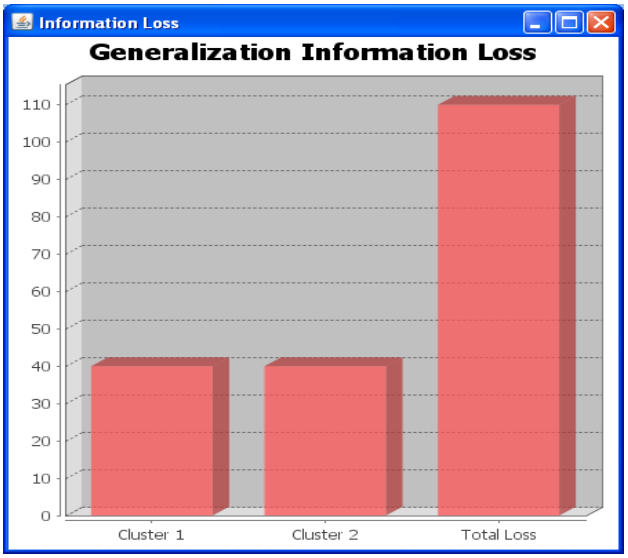
6. Results

As per we have come to the results part of the project analysis done. Web addressed various issues related to private information leakage in social networks by identifying key nodes of the graph structure to see if removing or altering these nodes can decrease information leakage.

The below information of the graphs explains that the data which was collected first would be clusterd. Then reveals all the data which was stored in the database. Generalization does indeed decrease the accuracy of classification on the data set. We see the most benefit here from generalizing only the activity details. We also see

that group is most consistently anonymized completely until the required privacy allowance.





Conclusion

We addressed various issues related to private information leakage in social networks. We show that using both friendship links and details together gives better predictability than details alone. In addition, we explored the effect of removing details and links in preventing sensitive information leakage. In the process, we discovered situations in which collective inferencing does not improve on using a simple local classification method to identify nodes. When we combine the results from the collective inference implications with the individual results, we begin to see that removing details and friendship links together is the best way to reduce classifier accuracy. This is probably infeasible in maintaining the use of social networks. However, we also show that by removing only details, we greatly reduce the accuracy local classifiers, which give us the maximum accuracy that we were able to achieve through any combination of classifiers. We also assumed full use of the graph information when deciding which details to hide. Useful research could be done on how individuals with limited access to the network could pick which details to hide. Similarly, future work could be conducted in identifying key nodes of the graph structure to see if removing or altering these nodes can decrease information leakage.

Authors



R.Divya B.Mahesh R.Ushasree

Reference

Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham (August 2013.), Fellow, IEEE, Preventing Private Information Inference Attacks on Social Networks, IEEE transactions on knowledge and data engineering, vol. 25, No. 8.