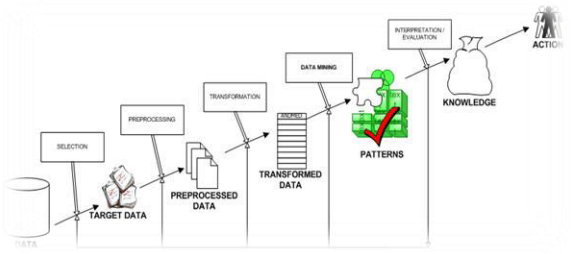
Preventing Private Information Inference Attacks On Social Networks

Pasala Pushpa Srivalli , Sri.G.Ramesh Kumar, Sri.V.Bhaskara Murthy
MCA Student, Assistant Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

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. We then devise three possible sanitization techniques that could be used in various situations. 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.

I. INTRODUCTION

What is Data Mining?



Structure of Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k=1). Sometimes called the k-nearest neighbor technique.

- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Characteristics of Data Mining:

- Large quantities of data: The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.

- Noisy, incomplete data: Imprecise data is the characteristic of all data collection.
- Complex data structure: conventional statistical analysis not possible
- Heterogeneous data stored in legacy systems

Benefits of Data Mining:

1) It's one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them
2) An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers
3) An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)
4) Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors
5) Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

Advantages of Data Mining:

1. Marketing / Retail:

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

2. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

3. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of

golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

4. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

5. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

6. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

II. 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.

The existing work could model and analyze access control requirements with respect to collaborative authorization management of shared data in OSNs. The need of joint management for data sharing, especially photo sharing, in OSNs has been recognized by the recent work provided a solution for collective privacy management in OSNs. Their work considered access control policies of a content that is co-owned by multiple users in an OSN, such that each co-owner may separately specify her/his own privacy preference for the shared content.

DISADVANTAGES OF EXISTING SYSTEM:

This problem of private information leakage could be an important issue in some cases.

III. PROPOSED SYSTEM

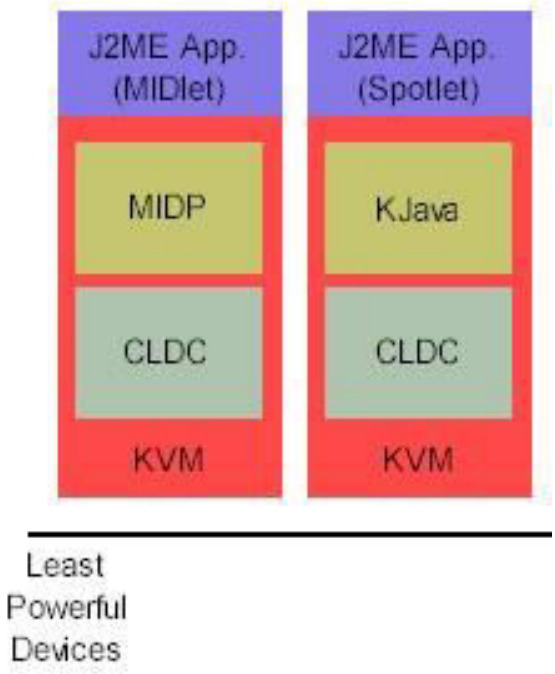
This paper focuses on the problem of private information leakage for individuals as a direct result of their actions as being part of an online social network. We model an attack scenario as follows: Suppose Facebook wishes to release data to electronic arts for their use in advertising games to interested people. However, once electronic arts has 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—through inference—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 (e.g., political or religious affiliation, sexual orientation) 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 non-private details.

In Proposed System we implemented a proof-of-concept Facebook application for the collaborative management of shared data, called MController. Our prototype application enables multiple associated users to specify their authorization policies and privacy preferences to co-control a shared data item.

ADVANTAGES OF PROPOSED SYSTEM:

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.

IV. SYSTEM ARCHITECTURE



V. IMPLEMENTATION

MODULES:

1. Privacy clarity for Formal data
2. Control of data's
3. Choosing of details Module
4. Operate Link Information
5. Generalization Module

MODULES DESCRIPTION:

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

VI. 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 of 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.

REFERENCES

[1] Facebook Beacon, 2007.

[2] T. Zeller, “AOL Executive Quits After Posting of Search Data,” The New York Times, no. 22, http://www.nytimes.com/2006/08/22/technology/22iht-aol.2558731.html?pagewanted=all&_r=0, Aug. 2006.

[3] K.M. Heussner, “‘Gaydar’ n Facebook: Can Your Friends Reveal Sexual Orientation?” ABC News, http://abcnews.go.com/Technology/gaydar-facebook-friends/story?id=8633224#. UZ939UqheOs, Sept. 2009.

[4] C. Johnson, “Project Gaydar,” The Boston Globe, Sept. 2009.

[5] 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.

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

[7] 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.

[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] R. Gross, A. Acquisti, 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.

[11] H. Jones and J.H. Soltren, "Facebook: Threats to Privacy," technical report, Massachusetts Inst. of Technology, 2005.

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

[13] B. Tasker, P. Abbeel, and K. Daphne, "Discriminative Probabilistic Models for Relational Data," Proc. 18th Ann. Conf. Uncertainty in Artificial Intelligence (UAI '02), pp. 485-492, 2002.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-based Systems, pp. 557-570, 2002.

[20] 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.

[21] C. Dwork, "Differential Privacy," Automata, Languages and Programming, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., vol. 4052, pp. 1-12, Springer, 2006.

[22] A. Friedman and A. Schuster, "Data Mining with Differential Privacy," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 493-502, 2010.